
Summary: “This book is filled with the knowledge about our solar system that resulted from all this exploration, whether by spacecraft or by telescopes both in space and earth-bound. All of this new knowledge is based on discoveries made in the interim by scientist-explorers who have followed their inborn human imperative to explore and to understand. Many old mysteries, misunderstandings, and fears that existed 50 years ago about what lay beyond the Earth have been eliminated. We now know the major features of the landscape in our cosmic backyard and can look forward to the adventure, excitement, and new knowledge that will result from more in-depth exploration by today’s spacecraft, such as those actually exploring the surface of these faraway places, including the Huygens Titan lander and the Mars Exploration rovers, doing things that were unimaginable before the Space Age began. The Encyclopedia of the Solar System is filled with images, illustrations, and charts to aid in understanding. Every object in the solar system is covered by at least one chapter. Other chapters are devoted to the relationships among the objects in the solar system and with the galaxy beyond. The processes that operate on solar system objects, in their atmospheres, on their surfaces, in their interiors, and interactions with space itself are all described in detail. There are chapters on how we explore and learn about the solar system and about the investigations used to make new discoveries. And there are chapters on the history of solar system exploration and the missions that have carried out this enterprise. All written by an international set of world-class scientists using rigorous yet easy-to-understand prose”–Provided by publisher.

ISBN 978-0-12-415845-0


2014002257
The solar system has become humankind's new backyard. It is the playground of robotic planetary spacecraft that has surveyed just about every corner of this vast expanse in space. Nowadays, every schoolchild knows what even the farthest planets look like. Fifty years ago, these places could only be imagined, and traveling to them was the realm of fiction. In just this short time in the long history of the human species we have leapt off the surface of our home planet and sent robotic extensions of our eyes, ears, noses, arms, and legs to the far reaches of the solar system and beyond.

In the early twentieth century, we were using airplanes to extend our reach to the last unexplored surface regions of our own planet. Now 100 years later, at the beginning of the twenty-first century, we are using spacecraft to extend our reach from the innermost planet Mercury to the outmost planet Neptune, and we have a spacecraft on the way to Pluto and the Kuiper Belt. Today, there are telescopes beyond imagination 100 or even 50 years ago that can image Pluto and detect planets around other stars! Now, Sol’s planets can say “we are not alone”; there are objects just like us elsewhere in the universe. As humanity’s space technology improves, perhaps in the next 100 years or so human beings also may be able to say “we are not alone.”

When I was a kid more than 50 years ago, I was thrilled by the paintings of Chesley Bonestell and others who put their imagination on canvas to show us what it might be like “out there.” Werner Von Braun’s Collier’s magazine articles of 1952—1954 superbly illustrated how we would go to the Moon and Mars using new rocket technologies. Reading those fabulous articles crystallized thoughts in my young mind about what to do with my life. I wanted to be part of the adventure to find out what these places were like. Not so long after the Collier’s articles appeared, we did go to the Moon, and pretty much as illustrated, although perhaps not in such a grand manner. We have not sent humans to Mars—at least we have not yet—but we have sent our robots to Mars and to just about every other place in the solar system as well.

This book is filled with the knowledge about our solar system that resulted from all this exploration, whether by spacecraft or by telescopes both in space and earth-bound. It could not have been written 50 years ago as almost everything in this Encyclopedia was unknown back then. All of this new knowledge is based on discoveries made in the interim by scientist-explorers who have followed their inborn human imperative to explore and to understand. Many old mysteries, misunderstandings, and fears that existed 50 years ago about what lay beyond the Earth have been eliminated.

We now know the major features of the landscape in our cosmic backyard and can look forward to the adventure, excitement, and new knowledge that will result from more in-depth exploration by today’s spacecraft, such as those actually exploring the surface of these faraway places, including the Huygens Titan lander, the Mars Exploration and the Curiosity rovers, doing things that were unimaginable before the Space Age began.

The Encyclopedia of the Solar System is filled with images, illustrations, and charts to aid in understanding. Every object in the solar system is covered by at least one chapter. Other chapters are devoted to the relationships among the objects in the solar system and with the galaxy beyond. The processes that operate on solar system objects, in their atmospheres, on their surfaces, in their interiors, and interactions with space itself are all described in detail. There are chapters on how we explore and learn about the solar system and about the investigations used to make new discoveries. And there are chapters on the history of solar system exploration and the missions that have carried out this enterprise. All written by an international set of world-class scientists using rigorous yet easy-to-understand prose.

Everything you want to know about the solar system is here. This is your highway to the solar system. It is as much fun exploring this Encyclopedia as all the exploration it took to get the information that it contains. Let your fingers be the spacecraft as you thumb through this book visiting all the planets, moons, and other small objects in the solar system. Experience what it is like to look at our solar system with ultraviolet eyes, infrared eyes, radio eyes, and radar eyes.

* This foreword to the second edition has been editorially updated to be included in the present edition.
It has been almost 15 years since the first edition. The exploration of space has continued at a rapid pace since then, and many missions have flown in the interim. New discoveries are being made all the time. This third edition will catch you up on all that has happened since the previous editions, including several new chapters based on information from our latest missions.

I invite you to enjoy a virtual exploration of the solar system by flipping through the pages in this volume. This book deserves a place in any academic setting and wherever there is a need to understand the cosmos beyond our home planet. It is the perfect solar system reference book, lavishly illustrated and well written. The editors and authors have done a magnificent job.

We live in a wonderful time of exploration and discovery. Here is your window to the adventure.

Wesley T. Huntress
Geophysical Laboratory,
Carnegie Institution of Washington,
Washington, D.C.
The known is finite, the unknown infinite; intellectually we stand on an islet in the midst of an illimitable ocean of inexplicability. Our business in every generation is to reclaim a little more land.

Thomas Henry Huxley

It is now 15 years since the first publication of the Encyclopedia of the Solar System and 8 years since the second and revised edition. The book has been an undisputed jewel in every library of books in solar system science and a great success with readers.

When Elsevier approached us to prepare a third edition, with a largely new editorship we thought hard on how we would proceed. Elsevier had left us to decide whether we wanted a completely new concept or to simply update the existing book. We finally settled on a concept that tried to further evolve an outstandingly successful work.

The past decade has seen an increasing importance of geophysical tools for the exploration of planets. In addition, our theoretical knowledge of the inner workings of terrestrial planets has substantially increased. We have acknowledged this by adding a chapter on geophysical exploration tools, in general, and on exploration of the Moon and on potential landing sites on Mars. We further added a chapter on rotation of the planets and using its observation to constrain models of the interior of terrestrial planets. Chapters on the interiors of Mars and the Moon—the two planets we know best—have been added as well as theoretical chapters pertaining to the inner workings of terrestrial planets—the generation of their magnetic fields and the relation between their thermal evolution, convection in the interior and their tectonics.

High resolution and stereo imaging is another novel tool of planetary exploration that we serve by adding a chapter. For the outer solar system we have added a chapter on Enceladus. Finally, we have complemented the suite of chapters dealing with the history of space exploration with a chapter describing the strategies that the space fairing nations have jointly developed in the International Space Exploration Initiative to take us from robotic exploration to human exploration to permanent human outposts.

Most of the authors of the previous edition have stayed on the team. They have worked meticulously to bring their chapters up-to-date, reflecting our current state of knowledge. A few authors have been unable to contribute in which case we have found new (co)authors or—in just a couple of cases—have reprinted the chapters after editorial updates.

It is sad to note that Conway Leovy, author of the Mars atmosphere chapter for the previous editions and Don Hunten, author of the Venus atmosphere chapter have passed away. David Catling was so kind to provide us with a newly written chapter on the Martian atmosphere while Fred Taylor updated Don’s chapter on Venus.

There have been significant advances in our knowledge, many related to new missions (compare the table in the appendix). Since 2006, when most of the chapters of the previous edition went to press, six missions have been launched to the sun, among them, a Russian mission and a French microsatellite. Unfortunately, the Russian satellite failed. The French microsatellite is an example of a new tool to explore the solar system, dedicated small to very small and affordable missions. China, India, and Japan have sent their own missions to the Moon in these years in addition to three NASA missions, totaling eight new missions altogether. The Moon continues to be the reachable target of great scientific interest. But even the Apollo data and samples remain valuable as the new discovery of water in lunar rocks and the seismic confirmation of the core and the discovery of its layering show.

Among the new missions since 2006 is Messenger, a NASA orbiter mission that revolutionized our knowledge of Mercury, the innermost planet and one of the two mostly unexplored places up to then. (The other being Pluto to which the New Horizon mission is on its way.) Another new mission is Venus Express, the first European Venus orbiter. In the inner solar system where the earthlike planets are located, Mars continues to be the prime target next to the Moon with NASA continuing to launch a mission at almost every opportunity, not the least because of Mars’ astrobiological potential. Here, exploration has proceeded along the classical exploration path where flybys are followed by orbiters, orbiters by landers, and landers by rovers. Robotic exploration would culminate with sample return, planned for the next decade only to be topped by human landing and exploration.
To the outer solar system, New Horizons has been launched in 2006 to visit Pluto and the Kuiper Belt for the first time after almost a decade-long journey. The mission shares this long travel time with Rosetta, the ESA mission to orbit and land in the fall of 2014 on comet nucleus Churyumov–Gerasimenko. A new mission to Jupiter, Juno, has been launched by NASA in 2011 to revisit the king of the planets. Missions to small bodies launched in recent years have been ESA’s Rosetta mission, NASA’s Dawn mission to explore Vesta and Ceres, and Japan’s innovative Hayabusa mission that has brought back samples from Asteroid Itokawa.

Earthbound and space telescopes have pushed the frontiers of planetary sciences beyond the solar system. NASA’s Kepler and the European Corot missions together with earthbound telescopes have increased the count of confirmed planets to almost 1800, about an order magnitude more than were known when the previous edition went to print. Three thousand planet candidates have been identified many of which, unfortunately, will never be confirmed by backup observation in the foreseeable future because of their large orbital distances and periods. The first nearly earth-sized planet in the habitable zone has been discovered by Kepler just recently bringing renewed interest to the science of Astrobiology. The habitable zone is defined as the range of orbital distance from the central star in which temperature on a planet in equilibrium with stellar radiation would allow liquid water to exist on the surface. We may speculate that the next edition of this book will perhaps report on the discovery of biosignatures in spectroscopic data from an extrasolar planet.

Other missions have continued their work in orbit or on the surfaces of planets. Among these are NASA’s Mars Exploration Rover Spirit, ESA’s Mars Express, and NASA’s Cassini mission to Saturn all three of which have celebrated their 10 years anniversary at their target planet at the time of this writing.

The missions we cited above just like the missions that preceded them have been extremely helpful to “reclaim a little more land” as Thomas Henry Huxley has put it. More missions are on the horizon such as NASA’s InSight mission, a geophysical station with a seismometer and a heat flow probe from Europe, ESA’s JUICE mission to the Jovian moons, and ESA’s PLATO and NASA’s TESS exoplanet telescopes.

We, the editors are deeply thankful to our outstanding colleagues who authored the chapters in this book and to about as many friends and colleagues who gave us their time and thoroughly reviewed the chapters.

We are equally indebted to the people at Elsevier who helped the project along over the past 4 years. John Fedor has sewed the first seeds to get the new edition under way. Katy Morrisey, Jill Cetel, and Louisa Hutchins have been our Editorial Project Managers at Elsevier and Poulouse Joseph and Paul Chandramohan have succeeded each other as Project Managers at Elsevier Book Productions in Chennai, India overseeing the proof composition and corrections.

With the editors of the previous edition, we share the “hope that this Encyclopedia will help you, the reader, appreciate this ongoing process of discovery and change as much as we do.”

Tilman Spohn
Torrence V. Johnson
Doris Breuer
May 13, 2014
Knowledge is not static. Science is a process, not a product. Some of what is presented in this volume will inevitably be out of date by the time you read it.

From the Preface to the first edition, 1999.

Written on the eve of the new millennium, the statement above was our acknowledgment that we cannot simply ‘freeze’ our knowledge of the solar system we inhabit; we box it up and display it like a collection of rare butterflies in the nineteenth-century “cabinet of curiosities.” Rather our goal was to provide our readers with an introduction to understanding the solar system as an interacting system, shaped by its place in the universe, its history, and the chemical and physical processes that operate from the extreme pressures and temperatures of the Sun’s interior to the frigid realm of the Oort cloud. We aimed to provide a work that was useful to students, professionals, and serious amateurs at a variety of levels, containing both detailed technical material and clear expositions of general principles and findings. With the help of our extremely talented colleagues who agreed to author the chapters, we humbly believe we achieved at least some of these ambitious goals.

How to decide when to update a work whose subject matter is in a constant, exuberant state of flux? It is difficult. Waiting for our knowledge of the solar system to be “complete” was deemed impractical, since our thesis is that this will never happen. Picking an anniversary date (30 years since this, or 50 years after that) seemed arbitrary. We compromised on taking an informal inventory of major events and advances in knowledge since that last edition whenever we got together at conferences and meetings. When we realized that virtually every chapter in the first edition needed major revisions and that new chapters would be called for to properly reflect new material, we decided to undertake the task of preparing a second edition with the encouragement and help from our friends and colleagues at Academic Press.

Consider how much has happened in the relatively short time since the first edition, published in 1999. An international fleet of spacecraft is now in place around Mars and two rovers are roaming its surface, with more to follow. Galileo ended its mission of discovery at Jupiter with a spectacular fiery plunge into the giant planet’s atmosphere. We have reached out and touched one comet with the Deep Impact mission and brought back precious fragments from another with Stardust. Cassini is sending back incredible data from the Saturn system and the Huygens probe descended to the surface of the giant, smog-shrouded moon Titan, revealing an eerily earthlike landscape carved by methane rains. NEAR and Hayabusa each orbited and then touched down on the surface of near-earth asteroids Eros and Itokawa, respectively. Scientists on the earth are continually improving the capabilities of telescopes and instruments, while laboratory studies and advances in theory improve our ability to synthesize and understand the vast amounts of new data being returned.

What you have before you is far more than a minor tweak to add a few new items to a table here or a figure there. It is a complete revamping of the Encyclopedia to reflect the solar system as we understand it today. We have attempted to capture the excitement and breadth of all this new material in the layout of the new edition. The authors of existing chapters were eager to update them to reflect our current state of knowledge, and many new authors have been added to bring fresh perspectives to the work. To all of those authors who contributed to the second edition and to the army of reviewers who carefully checked each chapter, we offer our sincere thanks and gratitude.

The organization of the chapters remains based on the logic of combining individual surveys of objects and planets, reviews of common elements and processes, and discussions of the latest techniques used to observe the solar system. Within this context you will find old acquaintances and many new friends. The sections on our own home planet have been revised and a new chapter on the Sun–Earth connection added to reflect our growing understanding of the intimate relationship between our star and conditions here on Earth. The treatment of Mars has been updated and a new chapter included incorporating the knowledge gained from the rovers Spirit and Opportunity and new orbital exploration of the red planet. Galileo’s remarkable discovery of evidence for subsurface oceans on the icy Galilean satellites is treated fully in new chapters devoted to Europa and to Ganymede and Callisto. New information from the Deep Impact mission and the Stardust sample return is included as well. We continue to find out
more and more about the denizens of the most distant reaches of the solar system, and have expanded the discussion of the Kuiper belt with a new chapter on physical properties. The area of observational techniques and instrumentation has been expanded to include chapters covering the X-ray portion of the spectrum, new generation telescopes, and remote chemical analysis.

Finally, nothing exemplifies the dynamic character of our knowledge than the area of extrasolar planets, which completes the volume. In the first edition the chapter on extrasolar planets contained a section entitled, “What is a Planet?” which concluded with this: “The reader is cautioned that these definitions are not uniformly accepted.” The chapter included a table of 19 objects cautiously labeled “Discovered Substellar Companions.” As this work goes to press, more than 200 extrasolar planets are known, many in multiplanet systems, with more being discovered everyday. And at the 2006 General Assembly of the International Astronomical Union, the question of the definition of “planet” was still being hotly debated. The current IAU definition is discussed in the introductory chapter by one of us (PRW) and other views concerning the status of Pluto may be found in the chapter on that body.

In addition to the energy and hard work of all of our authors, this edition of the Encyclopedia is greatly enhanced by the vision and talents of our friends at Academic Press. Specifically, we wish to thank Jennifer Helé, our Publishing Editor, who oversaw the project and learned the hard truth that herding scientists and herding cats are the same thing. Jennifer was the task master who made us realize that we could not just keep adding exciting new results to the volume, but one day had to stop and actually publish it. Francine Ribeau was our very able Marketing Manager and Deena Burgess, our Publishing Services Manager in the U.K., handled all of the last minute loose ends and made certain that the book was published without a hitch yet on a very tight schedule. Frank Cynar was our Publishing Editor for the first edition and for the beginning of the second, assisted by Gail Rice who was the Developmental Editor early on for the second edition. At Techbooks, Frank Scott was the Project Manager who oversaw all the final chapter and figure submissions and proof checking. Finally, also at Techbooks, was Carol Field, our Developmental Editor, simply known as Fabulous Carol, who seemed to work 30-h days for more than a year to see the volume through to fruition, while still finding time to get married in the midst of it all. This Encyclopedia would not exist without the tireless efforts of all of these extremely talented and dedicated individuals. To all of them we offer our eternal thanks.

Extensive use of color and new graphic designs have made the Encyclopedia even more beautiful and enhanced its readability while at the same time allowing the authors to display their information more effectively. The Encyclopedia before you is the result of all these efforts and we sincerely hope you will enjoy reading it as much as we enjoyed the process of compiling it.

Which brings us back to the quotation at the start of the Preface. We sincerely hope that this edition of the Encyclopedia will indeed also be out of date by the time you read it. The New Horizons spacecraft is on its way to the Pluto/Charon system, MESSENGER is on its way to Mercury, Rosetta is en route to a rendezvous with periodic comet Churyumov–Gerasimenko, new spacecraft are probing Venus and Mars, many nations are refocusing on exploration of the Moon, plans are being laid to study the deep interior of Jupiter and return to Europa, while the results from the Saturn system, Titan, and Enceladus have sparked a multitude of ideas for future exploration. We hope this Encyclopedia will help you, the reader, appreciate and enjoy this ongoing process of discovery and change as much as we do.

Lucy-Ann McFadden, Paul R. Weissman, Torrence V. Johnson
November 1, 2006
This is what hydrogen atoms can accomplish after four billion years of evolution.


The quote above comes from the final episode of the public television series “Cosmos,” which was created by Carl Sagan and several colleagues in 1981. Carl was describing the incredible accomplishments of the scientists and engineers who made the Voyager 1 and 2 missions to Jupiter and Saturn possible. But he just as easily could have been describing the chapters in this book.

This Encyclopedia is the product of the many scientists, engineers, technicians, and managers who produced the spacecraft missions which have explored our solar system over the past four decades. It is our attempt to provide to you, the reader, a comprehensive view of all we have learned in that 40 years of exploration and discovery. But we cannot take credit for this work. It is the product of the efforts of thousands of very talented and hardworking individuals in a score of countries who have contributed to that exploration. And it includes not only those involved directly in space missions, but also the many ground-based telescopic observers (both professional and amateur), laboratory scientists, theorists, and computer specialists who have contributed to creating that body of knowledge called solar system science. To all of these individuals, we say thank you.

Our goal in creating this Encyclopedia is to provide an integrated view of all we have learned about the solar system, at a level that is useful to the advanced amateur or student, to teachers, to nonsolar system astronomers, and to professionals in other scientific and technical fields. What we present here is an introduction to the many different specialties that constitute solar system science, written by the world’s leading experts in each field. A reader can start at the beginning and follow the course we have laid out, or delve into the volume at almost any point and pursue his or her own personal interests. If the reader wishes to go further, the lists of recommended reading at the end of each article provide the next step in learning about any of the subjects covered.

Our approach is to have the reader understand the solar system not only as a collection of individual and distinct bodies, but also as an integrated, interacting system, shaped by its initial conditions and by a variety of physical and chemical processes. The Encyclopedia begins with an overview chapter which describes the general features of the solar system and its relationship to the Milky Way galaxy, followed by a chapter on the origin of the system. Next we proceed from the Sun outward. We present the terrestrial planets (Mercury, Venus, Earth, and Mars) individually with separate chapters on their atmospheres and satellites (where they exist). For the giant planets (Jupiter, Saturn, Uranus, and Neptune) our focus shifts to common areas of scientific knowledge: atmospheres, interiors, satellites, rings, and magnetospheres. In addition, we have singled out three amazing satellites for individual chapters: Io, Titan, and Triton. Next is a chapter on the planetary system’s most distant outpost, Pluto, and its icy satellite, Charon. From there we move into discussing the small bodies of the solar system: comets, asteroids, meteorites, and dust. Having looked at the individual members of the solar system, we next describe the different view of those members at a variety of wavelengths outside the normal visual region. From there we consider the important processes that have played such an important role in the formation and evolution of the system: celestial dynamics, chaos, impacts, and volcanism. Last, we look at three topics which are as much in our future as in our past: life on other planets, space exploration missions, and the search for planets around other stars.

A volume like this one does not come into being without the efforts of a great number of very dedicated people. We express our appreciation to the more than 50 colleagues who wrote chapters, sharing their expertise with you, the reader. In addition to providing chapters that captured the excitement of their individual fields, the authors have endured revisions, rewrites, endless questions, and unforeseen delays. For all of these we offer our humble apologies. To ensure the quality and accuracy of each contribution, at least two independent reviewers critiqued each chapter. The peer review process maintains its integrity through the anonymity of the reviewers. Although we cannot acknowledge them by name, we thank all the reviewers for their time and their conscientious efforts.

We are also deeply indebted to the team at Academic Press. Our Executive Editor, Frank Cynar, worked tirelessly with us to conceptualize and execute the encyclopedia,
while allowing us to maintain the highest intellectual and scientific standards. We thank him for his patience and perseverance in seeing this volume through to completion. Frank’s assistants, Daniela Dell’Orco, Della Grayson, Linda McAleer, Cathleen Ryan, and Suzanne Walters, kept the entire process moving and attended to the myriad of details and questions that arise with such a large and complex volume. Advice and valuable guidance came from Academic Press’ director of major reference works, Chris Morris. Lori Asbury masterfully oversaw the production and copyediting. To all of the people at Academic Press, we give our sincere thanks.

Knowledge is not static. Science is a process, not a product. Some of what is presented in this volume will inevitably be out of date by the time you read it. New discoveries seem to come everyday from our colleagues using earth-based and orbiting telescopes, and from the flotilla of new small spacecraft that are out there adding to our store of knowledge about the solar system. In this spirit we hope that you, the reader, will benefit from the knowledge and understanding compiled in the following pages. The new millennium will surely add to the legacy presented herein, and we will all be the better for it. Enjoy, wonder, and keep watching the sky.

Paul R. Weissman,
Lucy-Ann McFadden,
Torrence V. Johnson
Tilman Spohn is director of the Institute of Planetary Research of the German Aerospace Center (DLR) in Berlin, Germany and Professor of Planetology at the Westfälische Wilhelms-University Münster, Germany. He specializes in the Thermodynamics of Planetary Interiors and in physical problems of Astrobiology and has written 140 papers for scientific journals and books. He is a Principal Investigator (PI) for MUPUS on the Rosetta Lander Philae, for BELA on BepiColombo, and the instrument PI for HP3 on the InSight mission. Spohn has served as member and chairperson of ESA scientific working and advisory groups and is presently a member of ESA’s HISPAC. He is the chairman of the science committee of the international Space Science Institute and has served as editor for Earth and Planetary Science Letters, Reviews of Geophysics, and the Treatise on Geophysics. He is the recipient of the 2013 EGU Runcorn-Florensky Medal and a fellow of the American Geophysical Union.

Doris Breuer is Head of the Department of Planetary Physics at the Institute of Planetary Research of the German Aerospace Center (DLR) in Berlin, Germany and an Associate Professor at the Institute de Physique du Globe de Paris, France. She specializes in the interior dynamics, thermo-chemical evolution and interior structure of terrestrial bodies. She has published over 60 articles in refereed journals, has been coinvestigator on ESA’s BepiColombo mission and has served on several ESA science definition teams and on the ESA Solar System Working Group.

Torrence V. Johnson is a specialist on icy satellites in the solar system. He has written over 130 papers for scientific journals. He received a PhD in planetary science from the California Institute of Technology and is currently a Senior Research Scientist at the Jet Propulsion Laboratory. Johnson was on the Voyager camera team during its exploration of the outer solar system and was the Project Scientist for the Galileo mission. He is currently an active investigator on the Cassini mission exploring the Saturn system. He is the recipient of two NASA Exceptional Scientific Achievement Medals and the NASA Outstanding Leadership Medal and has an honorary doctorate from the University of Padua, where Galileo made his first observations of the solar system.
Mahesh Anand  Department of Physical Sciences, The Open University, UK
Markus J. Aschwanden  Lockheed Martin ATC Solar and Astrophysics Laboratory, Palo Alto, CA, USA
Fran Bagenal  Department of Astrophysical & Planetary Sciences, Laboratory for Atmospheric & Space Physics, University of Colorado, Boulder, Boulder, CO, USA
W. Bruce Banerdt  Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
James F. Bell, III  School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA
Anil Bhardwaj  Space Physics Laboratory, Vikram Sarabhai Space Centre, Trivandrum, Kerala, India
Richard P. Binzel  Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
John C. Brandt  Department of Astronomy, University of Washington, Seattle, Washington, USA
Doris Breuer  Institute for Planetary Research, German Aerospace Center (DLR), Berlin, Germany
Daniel T. Britt  University of Central Florida, Orlando, FL, USA
Bonnie J. Buratti  Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
James D. Burke  The Planetary Society, Pasadena, CA, USA
Michael H. Carr  U. S. Geological Survey, Menlo Park, CA, USA
Lynn M. Carter  Planetary Geodynamics Laboratory; NASA Goddard Space Flight Center, Greenbelt, MD, USA
David C. Catling  University of Washington, Dept. of Earth and Space Sciences/Astrobiology Program, Seattle, WA, USA
John E. Chambers  Department of Terrestrial Magnetism, Carnegie Institution of Washington, Washington, DC, USA
Geoffrey Collins  Physics and Astronomy Dept., Wheaton College, Norton, Massachusetts, USA
Athena Coustenis  LESIA - Observatoire de Paris, CNRS, UPMC Univ. Paris 06, Univ. Paris-Diderot – Meudon, France
Ian A. Crawford  Department of Earth and Planetary Sciences, Birkbeck College, University of London, London, UK
Wanda L. Davis  Space Science Division, NASA Ames Research Center, Moffett Field, CA, USA
Véronique Dehant  Royal Observatory of Belgium, Brussels, Belgium
Konrad Dennerl  Max-Planck-Institut für extraterrestrische Physik, Garching, Germany
Imke de Pater  Astronomy Department, University of California, Berkeley, CA, USA; Faculty of Aerospace Engineering, Delft University of Technology, Delft, NL; SRON Netherlands Institute for Space Research, Utrecht, The Netherlands
Deborah L. Domingue  Planetary Science Institute, Tucson, AZ, USA
Luke Dones  Southwest Research Institute, Boulder, CO, USA
Timothy E. Dowling  Department of Physics and Astronomy, University of Louisville, Louisville, KY, USA
Line Drube  German Aerospace Center (DLR), Institute of Planetary Research, 12489 Berlin, Germany
Adam M. Dziewonski  Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA
Michael Endl  McDonald Observatory, University of Texas at Austin, Austin, TX, USA
Carolyn M. Ernst  Space Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA
Berndt Feuerbacher  German Aerospace Center, Cologne, Germany (ret.)
Jonathan J. Fortney  Department of Astronomy and Astrophysics, University of California, Santa Cruz, CA, USA
Contributors

Matthew P. Golombek Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

J.T. Gosling Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, CO, USA

Richard A.F. Grieve Centre for Planetary Science and Exploration, University of Western Ontario, London, ON, Canada

Robert Grimm Southwest Research Institute, Boulder, CO, USA

Matthias Grott German Aerospace Center (DLR), Institute of Planetary Research, Berlin, Germany

Eberhard Grün Max-Planck-Institut für Kernphysik, Heidelberg, Germany and LASP, University of Colorado, Boulder, CO, USA

Guy Consolmagno, S. J. Specola Vaticana, Vatican City State

Klaus Gwinner Planetary Geodesy Department, German Aerospace Center, Institute of Planetary Research, Berlin, Germany

Alex N. Halliday Department of Earth Sciences, University of Oxford, Oxford, UK

Alan W. Harris German Aerospace Center (DLR), Institute of Planetary Research, 12489 Berlin, Germany

Matthew M. Hedman Center for Radiophysics and Space Research, Cornell University, Ithaca, New York, USA

Amanda R. Hendrix Planetary Science Institute, Tucson, AZ, USA

Harald Hiesinger Institut für Planetologie, Westfälische Wilhelms-Universität, Münster, Germany

Bernhard Hufenbach European Space Agency, Noordwijk, The Netherlands

Donald M. Hunt University of Arizona, AZ, USA

Ralf Jaumann Deutsches Zentrum für Luft- und Raumfahrt (DLR), Berlin, Germany; Freie Universität Berlin, Institut für Geologische Wissenschaften, Berlin, Germany

Torrence V. Johnson Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA

Katherine H. Joy School of Earth, Atmospheric and Environmental Sciences, University of Manchester, UK

Randolph L. Kirk U.S. Geological Survey, Flagstaff, AZ, USA

Margaret Galland Kivelson Department of Earth & Space Sciences, University of California, Los Angeles, CA, USA and Department of Atmospheric, Oceanic, and Space Sciences, University of Michigan, Ann Arbor, MI, USA

Harald Krüger Max-Planck-Institut für Sonnensystemforschung, Göttingen, Germany

William S. Kurth Department of Physics and Astronomy, University of Iowa, Iowa City, IA, USA

Larry Lebofsky Planetary Science Institute, Tucson, AZ, USA

David Leverington Stoke Lacy, Herefordshire, United Kingdom

Harold F. Levison Southwest Research Institute, Boulder, CO, USA

Michael E. Lipschutz Purdue University, West Lafayette, IN, USA (Professor Emeritus)

Jack J. Lissauer Space Science & Astrobiology Division, NASA Ames Research Center Moffett Field, CA, USA

Carey M. Lisse Applied Physics Laboratory, Johns Hopkins University, Laurel, Maryland

Philippe Lognonné Institut de Physique du Globe de Paris, Paris, France

Rosaly M.C. Lopes Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

J.G. Luhmann Space Sciences Laboratory, University of California, Berkeley, CA, USA

Mark S. Marley Space Science Division, NASA Ames Research Center, Moffett Field, California, USA

Lucy A. McFadden Planetary Systems Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD 20771 USA

Christopher P. McKay Space Science Division, NASA Ames Research Center, Moffett Field, CA, USA

William B. McKinnon Department of Earth and Planetary Sciences and McDonnell Center for the Space Sciences, Washington University, Saint Louis, MO, USA

Harry Y. McSween Department of Earth & Planetary Sciences, University of Tennessee, Knoxville, TN, USA

Alessandro Morbidelli Observatoire de la Côte d’Azur, Nice, France

Nils Mueller German Aerospace Center (DLR), Institute of Planetary Research, Berlin, Germany

Scott L. Murchie Space Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA

Carl D. Murray Astronomy Unit, Queen Mary University of London, London, England, UK
Catherine D. Neish  Department of Physics and Space Science; Florida Institute of Technology, Melbourne, FL, USA
Robert M. Nelson  Planetary Science Institute, Tucson, AZ, USA
Francis Nimmo  Dept. Earth and Planetary Sciences, University of California Santa Cruz, CA, USA
Jürgen Oberst  Planetary Geodesy Department, German Aerospace Center, Institute of Planetary Research, Berlin, Germany
Gordon R. Osinski  Centre for Planetary Science and Exploration, University of Western Ontario, London, ON, Canada
Robert T. Pappalardo  Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
David C. Pieri  Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
Carolyn Porco  CICLOPS, Space Science Institute, Boulder, CO, USA
Thomas H. Prettyman  Planetary Science Institute, Tucson, AZ, USA
Frank Preusker  Planetary Geodesy Department, German Aerospace Center, Institute of Planetary Research, Berlin, Germany
Louise M. Prockter  Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA
Attilio Rivoldini  Royal Observatory of Belgium, Brussels, Belgium
Ludolf Schultz  Max-Planck-Institut für Chemie, Mainz, Germany (retired)
Adam P. Showman  Department of Planetary Sciences, Lunar and Planetary Laboratory, University of Arizona, Tucson, AZ, USA
Suzanne E. Smrekar  Earth and Space Sciences, Jet Propulsion Laboratory/Caltech, Pasadena, CA, USA
Sue Smrekar  Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA
S.C. Solomon  High Altitude Observatory, National Center for Atmospheric Research, Boulder, CO, USA
Tilman Spohn  Institute for Planetary Research, German Aerospace Center (DLR), Berlin, Germany
Sabine Stanley  Department of Physics, University of Toronto, Toronto, ON, Canada
S. Alan Stern  Space Science and Engineering Division, Southwest Research Institute, Boulder, CO, USA
Ellen R. Stofan  National Aeronautics and Space Administration, Headquarters, Washington, DC, USA
Robert G. Strom  Department of Planetary Sciences, University of Arizona, Tucson, AZ, USA
Mark V. Sykes  Planetary Science Institute, Tucson, AZ, USA
Fredric W. Taylor  University of Oxford, Oxford, UK
Stephen C. Tegler  Northern Arizona University, Flagstaff, Arizona, USA
Peter C. Thomas  Department of Astronomy, Center for Radiophysics & Space Research, Cornell University, Ithaca, NY, USA
Matthew S. Tiscareno  Center for Radiophysics and Space Research, Cornell University, Ithaca, NY, USA
Alan T. Tokunaga  Institute for Astronomy, University of Hawaii, Honolulu, HI, USA
Livio L. Tornabene  Centre for Planetary Science and Exploration, University of Western Ontario, London, ON, Canada
Nicola Tosi  Institute for Planetary Research, German Aerospace Center (DLR), Berlin, Germany
Tim Van Hoolst  Instituut voor Sterrenkunde, KU Leuven, Celestijnenlaan 200D, B-3001 Leuven, Belgium; Royal Observatory of Belgium, Brussels, Belgium
Ronald J. Vervack, Jr.  Space Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA
Renee C. Weber  NASA Marshall Space Flight Center, Huntsville, AL, USA
Paul R. Weissman  Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
Robert A. West  Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA
Lionel Wilson  Lancaster Environment Centre, Lancaster University, Lancaster, UK
Chapter 1

The Solar System and Its Place in the Galaxy

Paul R. Weissman
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Chapter Outline

1. Introduction 3
2. The Definition of a Planet 4
3. The Architecture of the Solar System 5
   3.1. Dynamics 5
   3.2. Nature and Composition 9
   3.3. Satellites, Rings, and Things 15
   3.4. The Solar Wind and the Heliosphere 20
4. The Origin of the Solar System 22
5. The Solar System’s Place in the Galaxy 24
6. The Fate of the Solar System 27
7. Concluding Remarks 28
Bibliography 28

1. INTRODUCTION

The origins of modern astronomy lie with the study of our solar system. When ancient humans first gazed at the skies, they recognized the same patterns of fixed stars rotating over their heads each night. They identified these fixed patterns, now called constellations, with familiar objects or animals, or with stories from their mythologies and their culture. But along with the fixed stars, there were a few bright points of light that moved each night, slowly following similar paths through a belt of constellations around the sky. The Sun and Moon also appeared to move through the same belt of constellations. These wandering objects were the planets of our solar system. Indeed, the name “planet” derives from the Latin planta, meaning wanderer.

The ancients recognized five planets that they could see with their naked eyes. We now know that the solar system consists of eight planets, at least five dwarf planets, plus a myriad of smaller objects: satellites, asteroids, comets, rings, and dust. Discoveries of new objects and new classes of objects are continuing even today. Thus, our view of the solar system is constantly changing and evolving as new data and new theories to explain (or anticipate) the data become available.

The solar system we see today is the result of the complex interaction of physical, chemical, and dynamical processes that have shaped the planets and other bodies. By studying each of the planets and other bodies individually as well as collectively, we seek to gain an understanding of those processes and the steps that led to the current solar system. Many of those processes operated most intensely early in the solar system’s history, as the Sun and planets formed from an interstellar cloud of dust and gas, 4.567 billion years ago. The first billion years of the solar system’s history was a violent period as the planets cleared their orbital zones of much of the leftover debris from the process of planet formation, flinging small bodies into planet-crossing, and often planet-impacting, orbits or out to interstellar space. In comparison, the present-day solar system is a much quieter place, although many of these processes continue today on a lesser scale.

Our knowledge of the solar system has exploded in the past five decades as interplanetary exploration spacecraft have provided close-up views of all the planets, as well as of a diverse collection of satellites, rings, asteroids, and comets. Earth-orbiting telescopes have provided an unprecedented view of the solar system, often at wavelengths not accessible from the Earth’s surface. Ground-based observations have also continued to produce exciting new discoveries through the application of a variety of new technologies such as charge-coupled device cameras, infrared detector arrays, adaptive optics, and powerful planetary radars. Theoretical studies have also contributed
significantly to our understanding of the solar system, largely through the use of advanced computer codes and high-speed, dedicated computers. Serendipity has also played an important role in many new discoveries.

Along with this increased knowledge have come numerous additional questions as we attempt to explain the complexity and diversity that we observe on each newly encountered world. The increased spatial and spectral resolution of the observations, along with in situ measurements of atmospheres, surface materials, and magnetospheres, have revealed that each body is unique, the result of a different combination of the physical, chemical, and dynamical processes that formed and shaped it. Also, each body’s formation zone (i.e. distance from the Sun) and the different initial solar nebula composition at that distance play an important role. Yet, at the same time, there are broad systematic trends and similarities that are clues to the collective history of the solar system.

We have now begun an exciting new age of discovery with the detection of numerous planet-sized bodies around nearby stars. Although the properties and placement of many of these extrasolar planets appear to be very different from those in our solar system, they are likely the prelude to the discovery of planetary systems that may more closely resemble our own.

We may also be on the brink of discovering evidence for life on other planets, in particular, Mars. There is an ongoing debate as to whether biogenic materials have been discovered in meteorites that were blasted off the surface of Mars and have found their way to the Earth. Although still very controversial, this finding, if confirmed, would have profound implications for the existence of life elsewhere in the solar system and the galaxy.

The goal of this chapter is to provide the reader with an introduction to the solar system. It seeks to provide a broad overview of the solar system and its constituent parts, to note the location of the solar system in the galaxy and to describe the local galactic environment. Detailed discussions of each of the bodies that make up the solar system, as well as the processes that have shaped those bodies and the techniques for observing the planetary system, are provided in the following chapters of this book. The reader is referred to those chapters for more detailed discussions of each of the topics introduced herein.

Some brief notes about planetary nomenclature will be useful. The names of the planets are all taken from Greek and Roman mythology (with the exception of the Earth, which is named for a goddess from Norse mythology), as are the names of their satellites, with the exception of the Moon and the Uranian satellites, the latter being named after Shakespearean characters. The Earth is occasionally referred to as Terra and the Moon as Luna, each the Latin version of their names. The naming system for planetary rings is different at each planet and includes descriptive names of the structures (at Jupiter), letters of the Roman alphabet (at Saturn), Greek letters and Arabic numerals (at Uranus), and the names of scientists associated with the discovery of Neptune (at Neptune).

Asteroids were initially named after women in Greek and Roman mythology. As their numbers have increased, asteroids have been named after the family members of the discoverers, after observatories, universities, cities, provinces, historical figures, scientists, writers, artists, literary figures, and, in at least one case, the astronomer’s cat. Initial discoveries of asteroids are designated by the year of their discovery and a letter/number code. Once the orbits of the asteroids are firmly established, they are given official numbers in the asteroid catalog: about 632,000 asteroids have been discovered and 385,000 asteroids have been numbered (as of January 2014). The discoverer(s) of an asteroid are given the privilege of suggesting its name, if done so within 10 years from when it was officially numbered.

Comets are generally named for their discoverers, although in a few well-known cases such as comets Halley and Encke, they are named for the individuals who first computed their orbits and linked several apparitions. Because some astronomers have discovered more than one short-period comet, a number is added at the end of the name in order to differentiate them, although this system is not applied to long-period comets. Comets are also designated by the year of their discovery and a letter code (a recently abandoned system used lowercase Roman letters and Roman numerals in place of the letter codes). The naming of newly discovered comets, asteroids, and satellites, as well as surface features on solar system bodies, is overseen by several working groups of the International Astronomical Union (IAU).

2. THE DEFINITION OF A PLANET

No formal definition of a planet existed until very recently. Originally, the ancients recognized five planets that could be seen with the naked eye: Mercury, Venus, Mars, Jupiter and Saturn, plus the Earth. Two more giant planets, Uranus and Neptune, were discovered telescopically in 1781 and 1846, respectively.

The largest asteroid, Ceres, was discovered in 1801 in an orbit between Mars and Jupiter and was hailed as a new planet because it fit into Bode’s law (see discussion later in this chapter). However, it was soon recognized that Ceres was much smaller than any of the known planets. As more and more asteroids were discovered in similar orbits between Mars and Jupiter, it became evident that Ceres was simply the largest body of a huge swarm of bodies between Mars and Jupiter that we now call the Asteroid Belt. A new term was coined, “minor planet”, to describe these bodies.
Searches for planets beyond Neptune continued and culminated in the discovery of Pluto in 1930. As with Ceres, it was soon suspected that Pluto was much smaller than any of the neighboring giant planets. Later, measurements of Pluto’s diameter by stellar occultations showed that it was also smaller than any of the terrestrial planets, in fact, even smaller than the Earth’s Moon. As a result, Pluto’s status as a planet was called into question.

In the 1980s, dynamical calculations suggested the existence of a belt of many small objects in orbits beyond Neptune, left over from the formation of the solar system. In the early 1990s the first of these objects, 1992 QB₁, was discovered at a distance of 40.9 astronomical units (AU). More discoveries followed and over 1500 bodies have now been found in the trans-Neptunian region (as of September 2013). They are collectively known as the Kuiper belt.

The existence of the Kuiper belt suggested that Pluto, like Ceres in the asteroid belt, was simply the largest body among a huge swarm of bodies beyond Neptune, again calling Pluto’s status into question. Then came the discovery of 136199 Eris (2003 UB₃₁₃), a Kuiper belt object (KBO) in a distant orbit, which turned out to be comparable in size to Pluto and somewhat more massive.

In response, the IAU, the governing body for astronomers worldwide, formed a committee to create a formal definition of a planet. The definition was presented at the IAU’s triennial gathering in Prague in 2006, where it was revised several times by the astronomers at the meeting. Eventually the IAU voted and passed a resolution that defined a planet.

That resolution states that a planet must have three qualities: (1) it must be round, indicating its interior is in hydrostatic equilibrium; (2) it must orbit the Sun; and (3) it must have gravitationally cleared its zone of other debris. The last requirement means that a planet must be massive enough to be gravitationally dominant in its zone in the solar system. Any round body orbiting the Sun that fails condition (3) is labeled a “dwarf planet” by the IAU.

This outcome left the solar system with the eight major planets discovered through 1846, and reclassified Ceres, Pluto, and Eris as dwarf planets. Two other KBOs, 136108 Haumea and 136472 Makemake, have also been added to that list. Other large objects in the asteroid and Kuiper belts may be added to the list of dwarf planets if observations show that they too are large and round.

There are weaknesses in the definition, particularly in condition (3), which may be modified by an IAU committee tasked with improving the definition. However, the likelihood of the definition being changed sufficiently to again classify Pluto as a planet is small.

The IAU has a somewhat different definition for planets discovered around other stars, known as “extrasolar” planets. At some point the two definitions need to be reconciled. See the chapter on Extra-Solar Planets for more discussion of this matter.

3. THE ARCHITECTURE OF THE SOLAR SYSTEM

The solar system consists of the Sun at its center, eight planets, five dwarf planets, 173 known natural satellites (or moons) of planets (as of September 2013), four ring systems, approximately 1 million asteroids (greater than 1 km in diameter), perhaps a trillion comets (greater than 1 km in diameter), the solar wind, and a large cloud of interplanetary dust. The arrangement and nature of all these bodies are the result of physical and dynamical processes during their origin and subsequent evolution, and their complex interactions with one another.

At the center of the solar system is the Sun, a rather ordinary, main sequence star. The Sun is classified spectroscopically as a G2V dwarf, which means that it emits the bulk of its radiation in the visible region of the spectrum, peaking at yellow-green wavelengths. The Sun contains 99.86% of the mass in the solar system, but only about 0.5% of the angular momentum. The low angular momentum of the Sun results from the transfer of momentum to the accretion disk surrounding the Sun during the formation of the planetary system, and to a slow spin down due to angular momentum being carried away by the solar wind.

The Sun is composed of hydrogen (70% by mass), helium (28%), and heavier elements (2%). The Sun produces energy through nuclear fusion at its center, hydrogen atoms combining to form helium and releasing energy that eventually makes its way to the Sun’s surface as visible sunlight. The central temperature of the Sun where fusion takes place is 15.7 million K, while the temperature at the visible surface, the photosphere, is about 5800 K. The Sun has an outer atmosphere called the corona, which is only visible during solar eclipses, or through the use of specially designed telescopes called coronagraphs.

A star like the Sun is believed to have a typical lifetime of 9–10 billion years on the main sequence. The present age of the Sun (and the entire solar system) is estimated to be 4.567 billion years, so it is about halfway through its nominal lifetime. The age estimate comes from radioisotope dating of meteorites, as well as from theories of stellar evolution.

3.1. Dynamics

The planets all orbit the Sun in roughly the same plane, known as the ecliptic (the plane of the Earth’s orbit), and in the same direction, counterclockwise as viewed from the north ecliptic pole. Because of gravitational torques from the other planets, the ecliptic is not inertially fixed in space, and so dynamicists often use the invariable plane, which is
the plane defined by the summed angular momentum vectors of all the planets.

To first order, the motion of any body about the Sun is governed by Kepler’s Laws of Planetary Motion. These laws state that (1) each planet moves about the Sun in an orbit that is an ellipse, with the Sun at one focus of the ellipse; (2) the straight line joining a planet and the Sun sweeps out equal areas in space in equal intervals of time; and (3) the squares of the sidereal periods of the planets are in direct proportion to the cubes of the semimajor axes of their orbits. The laws of planetary motion, first set down by J. Kepler in 1609 and 1619, are easily shown to be the result of the inverse square law of gravity with the Sun as the central body, and the conservation of angular momentum and energy. Parameters for the orbits of the eight planets and five dwarf planets are listed in Table 1.1.

Because the planets themselves have finite masses, they exert small gravitational tugs on one another, which cause their orbits to depart from perfect ellipses. The major effects of these long-term or “secular” perturbations are to cause the perihelion direction of each orbit to precess (rotate counterclockwise) in space, and the line of nodes (the intersection between the planet’s orbital plane and the ecliptic plane) of each orbit to regress (rotate clockwise). Additional effects include slow oscillations in the eccentricity and inclination of each orbit, and the inclination of the planet’s rotation pole to the planet’s orbit plane (called the obliquity). For the Earth, these orbital oscillations have periods of 19,000–100,000 years. They have been identified with long-term variations in the Earth’s climate, known as Milankovitch cycles, although the linking physical mechanism is not well understood.

Relativistic effects also play a small but detectable role. They are most evident in the precession of the perihelion of the orbit of Mercury, the planet deepest in the Sun’s gravitational potential well. General relativity adds 43 arcsec/century to the precession rate of Mercury’s orbit, which is 574 arcsec/century. Prior to Einstein’s theory of general relativity in 1916, it was thought that the excess in the precession rate of Mercury was due to a planet orbiting interior to it. This hypothetical planet was given the name Vulcan, and extensive searches were conducted for it, primarily during solar eclipses. No planet was detected.

A more successful search for a new planet occurred in 1846. Two celestial mechanicians, U. J. J. Leverrier and J. G. Adams, independently used the observed deviations of Uranus from its predicted orbit to successfully predict the existence and position of Neptune. Neptune was found by J. G. Galle on September 23, 1846, using Leverrier’s prediction.

More complex dynamical interactions are also possible, in particular when the orbital period of one body is a small integer ratio of another’s orbital period. This is known as a “mean-motion resonance” and can have dramatic effects. For example, Pluto is locked in a 2:3 mean-motion resonance with Neptune, and although the orbits of the two bodies cross in space, the resonance prevents them from

<table>
<thead>
<tr>
<th>Name</th>
<th>Semimajor Axis (AU)</th>
<th>Eccentricity</th>
<th>Inclination (°)</th>
<th>Period (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
<td>0.38710</td>
<td>0.205631</td>
<td>7.0049</td>
<td>0.2408</td>
</tr>
<tr>
<td>Venus</td>
<td>0.72333</td>
<td>0.006773</td>
<td>3.3947</td>
<td>0.6152</td>
</tr>
<tr>
<td>Earth</td>
<td>1.00000</td>
<td>0.016710</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Mars</td>
<td>1.52366</td>
<td>0.093412</td>
<td>1.8506</td>
<td>1.8808</td>
</tr>
<tr>
<td>Ceres</td>
<td>2.7665</td>
<td>0.078375</td>
<td>10.5834</td>
<td>4.601</td>
</tr>
<tr>
<td>Jupiter</td>
<td>5.20336</td>
<td>0.048393</td>
<td>1.3053</td>
<td>11.862</td>
</tr>
<tr>
<td>Saturn</td>
<td>9.53707</td>
<td>0.054151</td>
<td>2.4845</td>
<td>29.457</td>
</tr>
<tr>
<td>Uranus</td>
<td>19.1913</td>
<td>0.047168</td>
<td>0.7699</td>
<td>84.018</td>
</tr>
<tr>
<td>Neptune</td>
<td>30.0690</td>
<td>0.008586</td>
<td>1.7692</td>
<td>164.78</td>
</tr>
<tr>
<td>Pluto</td>
<td>39.4817</td>
<td>0.248808</td>
<td>17.1417</td>
<td>248.4</td>
</tr>
<tr>
<td>Haumea</td>
<td>43.0127</td>
<td>0.196577</td>
<td>28.1976</td>
<td>282.1</td>
</tr>
<tr>
<td>Makemake</td>
<td>45.4904</td>
<td>0.161420</td>
<td>29.0130</td>
<td>306.82</td>
</tr>
<tr>
<td>Eris</td>
<td>68.1461</td>
<td>0.432439</td>
<td>43.7408</td>
<td>562.55</td>
</tr>
</tbody>
</table>

2Dwarf planet.
ever coming within 14 AU of each other. Also, when two bodies have identical perihelion precession rates or nodal regression rates, they are said to be in a “secular resonance”, and similarly interesting dynamical effects can result. In many cases, mean-motion and secular resonances can lead to chaotic motion, driving a body onto a planet-crossing orbit, which will then lead to it being dynamically scattered among the planets and eventually either ejected from the solar system, or impacted on the Sun or a planet. In other cases, such as Pluto and some asteroids, the mean-motion resonance is actually a stabilizing factor for the orbit.

Chaos has become a very exciting topic in solar system dynamics in the past 25 years and has been able to explain many features of the planetary system that were not previously understood. It should be noted that the dynamical definition of chaos is not always the same as the general dictionary definition. In celestial mechanics, the term “chaos” is applied to describe systems that are not perfectly predictable over time. That is, small variations in the initial conditions, or the inability to specify the initial conditions precisely, will lead to a growing error in predictions of the long-term behavior of the system. If the error grows exponentially, then the system is said to be chaotic. However, the chaotic zone, the allowed area in phase space over which an orbit may vary, may still be quite constrained. Thus, although studies have found that the orbits of the planets are chaotic, this does not mean that Jupiter may one day become Earth-crossing, or vice versa. It means that the precise position of the Earth or Jupiter in their orbits is not predictable over very long periods. Because this happens for all the planets, the long-term secular perturbations of the planets on one another are also not perfectly predictable and can vary.

On the other hand, chaos can result in some extreme changes in orbits, with sudden increases in eccentricity that can throw small bodies onto planet-crossing orbits. One well-recognized case occurs near mean-motion resonances in the asteroid belt, which causes small asteroids to be thrown onto Earth-crossing orbits, allowing for the delivery of meteoroids to the Earth.

The natural satellites of the planets and their ring systems (where they exist) are governed by the same dynamical laws of motion. Most major satellites and all ring systems are deep within their planets’ gravitational potential wells and so they move, to first order, on Keplerian ellipses. The Sun, planets, and other satellites all act as perturbers on the satellite and ring particle orbits. Additionally, the equatorial bulges of the planets, caused by the planets’ rotation, act as perturbers on the orbits. Finally, the satellites raise tides on the planets (and vice versa), and these result in yet another dynamical effect, causing the planets to transfer rotational angular momentum to the satellite orbits in the case of direct or prograde orbits (satellites in retrograde orbits lose angular momentum). As a result, satellites may slowly move away from their planets into larger orbits (or into smaller orbits in the case of retrograde satellites).

The mutual gravitational interactions can be quite complex, particularly in multisatellite systems. For example, the three innermost Galilean satellites of Jupiter (so named because they were discovered by Galileo in 1610)—Io, Europa, and Ganymede—are locked in a 4:2:1 mean-motion resonance with one another. In other words, Ganymede’s orbital period is twice that of Europa and four times that of Io. At the same time, the other Jovian satellites (primarily Callisto), the Sun, and Jupiter’s oblateness perturb the orbits, forcing them to be slightly eccentric and inclined to one another, while the tidal interaction with Jupiter forces the orbits to evolve outward. These competing dynamical processes result in considerable energy deposition in the satellites, which manifests itself as volcanic activity on Io, as a possible subsurface ocean on Europa, and as past tectonic activity on Ganymede.

This illustrates an important point in understanding the solar system. The bodies in the solar system do not exist as independent, isolated entities, with no physical interactions between them. Even these “action-at-a-distance” gravitational interactions can lead to profound physical and chemical changes in the bodies involved. To understand the solar system as a whole, one must recognize and understand the processes that were involved in its formation and its subsequent evolution, and that continue to act today.

An interesting feature of the planetary orbits is their regular spacing. This is described by Bode’s law, first discovered by J. B. Titius in 1766 and brought to prominence by J. E. Bode in 1772. The law states that the semimajor axes of the planets in astronomical units can be roughly approximated by taking the sequence 0, 3, 6, 12, 24, ..., adding 4, and dividing by 10. The values for Bode’s law and the actual semimajor axes of the planets and two dwarf planets are listed in Table 1.2. It can be seen that the law works very well for the planets as far as Uranus, but it then breaks down. It also predicts a planet between Mars and Jupiter, the current location of the asteroid belt. Yet Bode’s law predates the discovery of the first asteroid by 35 years, as well as the discovery of Uranus by 15 years.

The reason why Bode’s law works so well is not understood. H. Levison has recently suggested that, at least for the giant planets, it is a result of their spacing themselves at distances where they are equally likely to scatter a smaller body inward or outward to the next planet in either direction.

However, it has also been argued that Bode’s law may just be a case of numerology and not reflect any real physical principle at all. Since Bode’s law was formulated after the semimajor axes of the first six planets were known, Titius and Bode were free to fit the form of the equation to the known data. Computer-based dynamical simulations have shown that the spacing of the planets is such that a body placed in a
It is generally believed that comets originated as icy planetesimals in the outer regions of the solar nebula, at the orbit of Jupiter and beyond. Those protocomets with orbits between the giant planets were gravitationally ejected, mostly to interstellar space. However, a fraction of the protocomets, about 4%, were flung into distant but still bound orbits; the Sun’s gravitational sphere of influence extends \( \sim 2 \times 10^5 \text{ AU} \), or about 1 parsec (pc = 206,264.8 AU). These orbits were sufficiently distant from the Sun that they were perturbed by random passing stars and by the tidal perturbation from the galactic disk. The stellar and galactic perturbations raised the perihelia of the comet orbits out of the planetary region. Additionally, the stellar perturbations randomized the inclinations of the comet orbits, forming a spherical cloud of comets around the planetary system and extending halfway to the nearest stars. This region is now called the Oort cloud, after J. H. Oort who first suggested its existence in 1950.

The current population of the Oort cloud is estimated to be about \( 2 \times 10^{12} \) comets, with a total mass of about one Earth mass of material. Between 20 and 50% of the Oort cloud population is in a dynamically active shell between \( 10^5 \) and \( 2 \times 10^5 \) AU from the Sun. Comets in this shell are perturbed by random passing stars and the galactic tide. The perturbations can change the perihelion distances of comets, sending them back into the planetary region where they are observed as long-period comets (those with orbital periods greater than 200 years). Interior to the shell is a dense inner Oort cloud that contains 50–80% of the comets, extending as close as 1000 AU from the Sun. The inner Oort cloud is not dynamically active. It is too close to the Sun to be significantly perturbed by external perturbers, unless the latter come very close, such as stars passing through the Oort cloud.

A second reservoir of comets is the Kuiper belt beyond the orbit of Neptune, named after G. P. Kuiper who in 1951 was one of the first to suggest its existence. Because no large planet grew beyond Neptune, there was no body to scatter away the icy planetesimals formed in that region. (The failure of a large planet to grow beyond Neptune is generally attributed to the increasing timescale for planetary accretion and the decreasing density of solar nebula materials with increasing heliocentric distance.) This belt of remnant planetesimals may terminate at \( \sim 50 \text{ AU} \) or may extend out several 100 AU from the Sun, analogous to the disks of dust that have been discovered around main sequence stars such as Vega and Beta Pictoris (Figure 1.1).

The Kuiper belt actually consists of two different dynamical populations. The classical Kuiper belt is the population in low-inclination, low-eccentricity orbits beyond Neptune. Some of this population, including Pluto, is trapped in mean-motion resonances with Neptune at both the 3:2 and 2:1 resonances. The second population is objects in more eccentric and inclined orbits, typically with larger semimajor

<table>
<thead>
<tr>
<th>Planet</th>
<th>Semimajor Axis (AU)</th>
<th>( n )</th>
<th>Bode’s Law</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceres$^1$</td>
<td>2.767</td>
<td>5</td>
<td>2.8</td>
</tr>
<tr>
<td>Jupiter</td>
<td>5.203</td>
<td>6</td>
<td>5.2</td>
</tr>
<tr>
<td>Saturn</td>
<td>9.537</td>
<td>7</td>
<td>10.0</td>
</tr>
<tr>
<td>Uranus</td>
<td>19.19</td>
<td>8</td>
<td>19.6</td>
</tr>
<tr>
<td>Neptune</td>
<td>30.07</td>
<td>9</td>
<td>38.8</td>
</tr>
<tr>
<td>Pluto$^1$</td>
<td>39.48</td>
<td>10</td>
<td>77.2</td>
</tr>
</tbody>
</table>

$^1$Dwarf planet.
axes, called the scattered disk. These latter objects all have perihelia relatively close to Neptune’s orbit, such that they continue to gravitationally interact with Neptune.

The Kuiper belt may contain many tens of Earth masses of comets, although the mass within 50 AU is currently estimated as \( \sim 0.1 \) Earth mass. A slow gravitational erosion of comets from the Kuiper belt, in particular from the scattered disk, due to the perturbing effect of Neptune, causes these comets to “leak” into the planetary region. Eventually, some fraction of the comets evolves due to gravitational scattering by the giant planets into the terrestrial planets region where they are observed as short-period comets. Short-period comets from the Kuiper belt are often called Jupiter-family or ecliptic comets because most are in orbits that can have close encounters with Jupiter, and are also in orbits with low inclinations, close to the ecliptic plane. Based on the observed number of ecliptic comets, the number of comets in the Kuiper belt between 30 and 50 AU has been estimated at \( \sim 10^9 \) objects larger than 1 km diameter, with a roughly equal number in the scattered disk. Current studies suggest that the Kuiper belt has been collisionally eroded out to a distance of \( \sim 100 \) AU from the Sun, but that considerably more mass may still exist in orbits beyond that distance.

Although gravity is the dominant force in determining the motion of bodies in the solar system, other forces do come into play in special cases. Dust grains produced by asteroid collisions or liberated from the sublimating icy surfaces of comet nuclei are small enough to be affected by radiation pressure forces. For submicron grains, radiation pressure from sunlight is sufficient to blow the grains out of the solar system. For larger grains, radiation pressure causes the grains to depart from Keplerian orbits. Radiation effects can also cause centimeter-sized grains to slowly spiral in toward the Sun through the Poynting–Robertson effect, and meter- to kilometer-sized bodies to slowly spiral either inward or outward due to the Yarkovsky effect.

Electromagnetic forces play a role in planetary magnetospheres where ions are trapped and spiral back and forth along magnetic field lines, and in cometary Type I plasma tails where ions are accelerated away from the cometary coma by the solar wind. Dust grains trapped in planetary magnetospheres and in interplanetary space also respond to electromagnetic forces, although to a lesser extent than ions because of their much lower charge-to-mass ratios.

### 3.2. Nature and Composition

The solar nebula, the cloud of dust and gas out of which the planetary system formed, almost certainly exhibited a strong temperature gradient with heliocentric distance, hottest near the forming proto-Sun at its center, and cooler as one moved outward through the planetary region. This

---

**FIGURE 1.1** False-color images of the dust disk around the star Beta Pictoris, discovered by the *Infrared Astronomical Satellite* in 1983. The disk is viewed nearly edge on and is over 900 AU in diameter. The gaps in the center of each image are where the central star image has been removed. The top image shows the full disk as imaged with the Wide Field Planetary Camera 2 (WFPC2) onboard the *Hubble Space Telescope* (HST). The lower image shows the inner disk as viewed by the Space Telescope Imaging Spectrograph (STIS) instrument on HST. The orbits of the outer planets of our solar system, including the dwarf planet Pluto, are shown to scale for comparison. There is evidence of a warping of the Beta Pic disk, possibly caused by perturbations from a passing star. Infrared data show that the disk does not extend all the way in to the star, but that it has an inner edge at about 30 AU from Beta Pic. The disk interior to that distance may have been swept up by the accretion of planets in the nebula around the star. This disk is a possible analog for the Kuiper belt around our own solar system.
A temperature gradient is reflected in the compositional arrangement of the planets and their satellites vs heliocentric distance. Parts of the gradient are also preserved in the asteroid belt between Mars and Jupiter and possibly in the Kuiper belt beyond Neptune.

Physical parameters for the planets and dwarf planets are given in Table 1.3. The planets fall into two major compositional groups. The terrestrial or Earth-like planets are Mercury, Venus, Earth, and Mars and are shown in Figure 1.2. The terrestrial planets are characterized by predominantly silicate compositions with iron cores. Gravitational potential energy heated the terrestrial planets as they formed resulting in them melting and then chemically differentiating. Their volatile content, i.e. atmospheres and oceans, may have accreted directly with the solid matter or may have been added later by asteroid and comet bombardment. Also, the modest masses of the terrestrial planets and their closeness to the Sun did not allow them to capture and retain gas directly from the solar nebula. The terrestrial planets all have solid surfaces that are modified to varying degrees by both cratering and internal processes (tectonics, weather, etc.).

Mercury is the most heavily cratered planet because it has no appreciable atmosphere to protect it from impacts or weather to erode the cratered terrain, and also because encounter velocities with Mercury are very high that close to the Sun. Additionally, tectonic processes on Mercury appear to have played a role in modifying its surface, which is partially covered by lava flows, like the Earth’s Moon. Mars is next in the degree of cratering, in large part because of its proximity to the asteroid belt. Also, Mars’ thin atmosphere affords little protection against impactors. However, Mars also displays substantial volcanic and tectonic features, and evidence of erosion by wind and flowing water, the latter presumably having occurred early in the planet’s history.

The surface of Venus is dominated by a wide variety of volcanic terrains. The degree of cratering on Venus is less than that on Mercury or Mars for two reasons: (1) Venus’ thick CO$_2$ atmosphere (surface pressure = 93 bar) breaks up smaller asteroids and comets before they can reach the surface and (2) volcanism on the planet has covered over the older craters on the planet surface. The surface of Venus is estimated to be 300–600 million years in age.

The Earth’s surface is dominated by plate tectonics, in which large plates of the crust can move about the planet, and whose motions are reflected in such features as mountain ranges (where plates collide) and volcanic zones (where one plate dives under another). The Earth is the only planet with the right combination of atmospheric surface pressure and temperature to permit liquid water on its surface, and some 70% of the planet is covered by oceans.

<table>
<thead>
<tr>
<th>Name</th>
<th>Mass (kg)</th>
<th>Equatorial Radius (km)</th>
<th>Density (g/cm$^3$)</th>
<th>Rotation Period</th>
<th>Obliquity (°)</th>
<th>Escape Velocity (km/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun</td>
<td>1.989 × 10$^{20}$</td>
<td>695,508</td>
<td>1.41</td>
<td>25.4–35. d</td>
<td>7.25</td>
<td>617.7</td>
</tr>
<tr>
<td>Mercury</td>
<td>3.302 × 10$^{23}$</td>
<td>2440</td>
<td>5.43</td>
<td>56.646 d</td>
<td>0</td>
<td>4.25</td>
</tr>
<tr>
<td>Venus</td>
<td>4.869 × 10$^{24}$</td>
<td>6052</td>
<td>5.24</td>
<td>243.018 d</td>
<td>177.33</td>
<td>10.36</td>
</tr>
<tr>
<td>Earth</td>
<td>5.974 × 10$^{24}$</td>
<td>6378</td>
<td>5.52</td>
<td>23.934 h</td>
<td>23.45</td>
<td>11.18</td>
</tr>
<tr>
<td>Mars</td>
<td>6.419 × 10$^{23}$</td>
<td>3397</td>
<td>3.94</td>
<td>24.623 h</td>
<td>25.19</td>
<td>5.02</td>
</tr>
<tr>
<td>Ceres$^1$</td>
<td>9.47 × 10$^{20}$</td>
<td>474</td>
<td>2.1</td>
<td>9.075 h</td>
<td>3</td>
<td>0.52</td>
</tr>
<tr>
<td>Jupiter</td>
<td>1.899 × 10$^{27}$</td>
<td>71,492</td>
<td>1.33</td>
<td>9.925 h</td>
<td>3.08</td>
<td>59.54</td>
</tr>
<tr>
<td>Saturn</td>
<td>5.685 × 10$^{26}$</td>
<td>60,268</td>
<td>0.70</td>
<td>10.656 h</td>
<td>26.73</td>
<td>35.49</td>
</tr>
<tr>
<td>Uranus</td>
<td>8.662 × 10$^{25}$</td>
<td>25,559</td>
<td>1.30</td>
<td>17.24 h</td>
<td>97.92</td>
<td>21.26</td>
</tr>
<tr>
<td>Neptune</td>
<td>1.028 × 10$^{26}$</td>
<td>24,764</td>
<td>1.76</td>
<td>16.11 h</td>
<td>28.80</td>
<td>23.53</td>
</tr>
<tr>
<td>Pluto$^1$</td>
<td>1.314 × 10$^{22}$</td>
<td>1151</td>
<td>2.0</td>
<td>6.387 d</td>
<td>119.6</td>
<td>1.23</td>
</tr>
<tr>
<td>Haumea$^1$</td>
<td>4.006 × 10$^{21}$</td>
<td>575–718</td>
<td>2.6–3.3</td>
<td>3.915 h</td>
<td></td>
<td>0.84</td>
</tr>
<tr>
<td>Makemake$^1$</td>
<td>3 × 10$^{21}$</td>
<td>715</td>
<td>1.7</td>
<td>7.771 h</td>
<td></td>
<td>0.74</td>
</tr>
<tr>
<td>Eris$^1$</td>
<td>1.67 × 10$^{22}$</td>
<td>1163</td>
<td>2.52</td>
<td>25.9 h</td>
<td></td>
<td>1.38</td>
</tr>
</tbody>
</table>

$^1$Dwarf planet.
Craters on the Earth are rapidly erased by its active geology and weather, although the atmosphere only provides protection against very modest size impactors, on the order of 60 m diameter or less. Still, 181 impact craters or their remnants have been found on the Earth’s surface or under its oceans.

The terrestrial planets each have substantially different atmospheres. Mercury has a tenuous atmosphere arising from its interaction with the solar wind. Hydrogen and helium ions are captured directly from the solar wind, while oxygen, sodium, and potassium are likely the product of sputtering at the surface. In contrast, Venus has a dense CO₂ atmosphere with a surface pressure 93 times the pressure at the Earth’s surface. Nitrogen is also present in the Venus atmosphere at a few percent relative to CO₂. The dense atmosphere results in a massive greenhouse on the planet, heating the surface to a mean temperature of 735 K. The middle and upper atmosphere contain thick clouds composed of H₂SO₄ and H₂O, which shroud the surface from view. However, thermal radiation from the surface does penetrate the clouds, making it possible to view surface features through infrared “windows”.

The Earth’s atmosphere is unique because of its large abundance of free oxygen, which is normally tied up in oxidized surface materials on other planets. The reason for this unusual state is the presence of life on the planet, which traps and buries CO₂ as carbonates and also converts the CO₂ to free oxygen. Still, the bulk of the Earth’s atmosphere is nitrogen (78%), with oxygen making up 21% and argon about 1%. The water vapor content of the atmosphere varies from about 1% to 4%. Various lines of evidence suggest that the composition of the Earth’s atmosphere has evolved considerably over the history of the solar system and that the original atmosphere was denser than the present-day atmosphere and dominated by CO₂.

Mars has a relatively modest CO₂ atmosphere with a mean surface pressure of only 6 mbar. The atmosphere also contains a few percent of N₂ and argon. Mineralogic and isotopic evidence and geologic features suggest that the past atmosphere of Mars may have been much denser and warmer, allowing liquid water to flow across the surface in massive floods.

The volatiles in the terrestrial planets’ atmospheres and in the Earth’s oceans may have been contained in hydrated
minerals in the planetesimals that originally formed the planets, and/or may have been added later due to asteroid and comet bombardment as the planets dynamically cleared their individual zones of leftover planetesimals. It appears most likely that all these reservoirs contributed some fraction of the volatiles on the terrestrial planets.

The giant or Jupiter-like planets are Jupiter, Saturn, Uranus, and Neptune and are shown in Figure 1.3. The giant planets are also referred to as the gas giants. They are characterized by low mean densities and thick hydrogen–helium atmospheres, presumably captured directly from the solar nebula during the formation of these planets. The composition of the giant planets is similar to that of the Sun, although more enriched in heavier elements. Because of their primarily gaseous composition and their high internal temperatures and pressures, the giant planets do not have solid surfaces. However, they may each have silicate—iron cores of several to tens of Earth masses of material.

Because they formed at heliocentric distances where ices could condense, the giant planets may have initially had a much greater local density of solid material to grow from. This may, in fact, have allowed them to form before the terrestrial planets interior to them. Studies of the dissipation of nebula dust disks around nearby solar-type protostars suggest that the timescale for the formation of giant planets is on the order of 10 million years or less. This is very rapid as compared with the ~100 million year timescale currently estimated for the formation of the terrestrial planets (although questions have now been raised as to the correctness of that accretionary timescale). Additionally, the higher uncompressed densities of Uranus and Neptune (0.5 g/cm³) vs those of Jupiter and Saturn (0.3 g/cm³) suggest that the outer two giant planets contain a significantly lower fraction of gas captured from the nebula. This may mean that the outer pair formed later than the inner two giant planets, consistent with the increasing timescale for planetary accretion at larger heliocentric distances.

Because of their heliocentric arrangement, the terrestrial and giant planets are occasionally called the inner and outer planets, respectively, although sometimes the term “inner planets” is used only to denote Mercury and Venus, the planets interior to the Earth’s orbit.

There are currently five recognized dwarf planets, described below.

Ceres, discovered in 1801, is the largest body in the asteroid belt and the only main belt object classified as a dwarf planet. It has a surface composition and density...
The Pluto—Charon system is fully tidally evolved. This means that Pluto and Charon each rotate with the same period, 6.38723 days, which is also the revolution period of Charon in its orbit. As a result, Pluto and Charon always show the same faces to each other. It is suspected that the Pluto—Charon system was formed by a giant impact between two large KBOs.

Haumea, discovered in 2004, is in an orbit that ranges between 35 and 51 AU from the Sun, and inclined 28° to the ecliptic. It may be trapped in a 7:12 mean-motion resonance with Neptune. Spectra show that Haumea is covered with a layer of crystalline ice, much like Pluto’s satellite Charon. Haumea has two known satellites: Hi’iaka and Namaka.

Makemake, discovered in 2005, has an orbit that ranges between 38 and 53 AU with an inclination of 28°. Spectra of Makemake show the presence of methane ices on the surface, similar to Pluto. No satellites have been detected around Makemake.

The dwarf planet Eris was discovered in 2005 and is a scattered disk object in a distant orbit that ranges from 37.8 to 97.5 AU from the Sun, with an inclination of 43°. It is comparable in size to Pluto, has a somewhat higher bulk density, and also displays evidence for methane frost on its surface. Eris has one satellite, Dysnomia.

There has been considerable speculation as to the existence of a major planet beyond Neptune, often dubbed “Planet X”. The search program that found Pluto in 1930 was continued for many years afterward but failed to detect any other distant objects, even though the limiting magnitude was considerably fainter than Pluto’s visual magnitude of ~13.5. Other searches have been carried out, most notably by the Infrared Astronomical Satellite (IRAS) in 1983–1984. An automated algorithm was used to search for a distant planet in the IRAS data; it successfully “discovered” Neptune, but nothing else. More recently, the WISE (Widefield Infrared Survey Explorer) spacecraft surveyed the infrared sky in four wavelengths in 2010–2011 with much higher sensitivity than IRAS. Although nothing was found, analysis of the WISE data is continuing. As noted above, telescopic searches for KBOs have found objects comparable to Pluto in size, but none significantly larger.

Gravitational analyses of the orbits of Uranus and Neptune show no evidence of an additional perturber at greater heliocentric distances. Studies of the trajectories of the Pioneer 10 and 11 and Voyager 1 and 2 spacecraft have also yielded negative results. Analyses of the spacecraft trajectories do provide an upper limit on the unaccounted mass within the orbit of Neptune of $< 3 \times 10^{-6}$ solar masses ($M_\odot$), equal to about one Earth mass.

The compositional gradient in the solar system is perhaps best visible in the asteroid belt, whose members range from silicate-rich bodies in the inner belt (inside of ~2.6 AU), to volatile-rich carbonaceous bodies in the outer main belt (out to about 3.3 AU). (See Figure 1.5.)

---

**FIGURE 1.4** Hubble Space Telescope (HST) image of the dwarf planet Pluto (center) with its large moon Charon (just below and to the left of Pluto), and the four small satellites discovered with HST. The images of Pluto and Charon have been deliberately reduced in brightness so that the smaller satellites can be seen. A NASA spacecraft mission, New Horizons, was launched in 2006 and will fly by Pluto and Charon in 2015. Courtesy of NASA and the Space Telescope Science Institute.

---

similar to carbonaceous chondrite meteorites. This is a primitive class of meteorites that shows only limited processing during and since formation. Water frost has also been detected on the surface of Ceres. Because of its large size, the interior of Ceres is likely differentiated. A National Aeronautics and Space Administration (NASA) mission, Dawn, recently visited the large asteroid Vesta and is now on its way to Ceres, arriving in 2015.

Pluto, discovered in 1930, is the first object discovered in the Kuiper belt. It is classified as a dwarf planet, and has five satellites, the largest being Charon, which is about half the size of Pluto. Both are shown in Figure 1.4 along with the four smaller satellites. Pluto bears a strong resemblance to Triton, Neptune’s large icy satellite (which is slightly larger than Pluto) and to other large icy objects in the Kuiper belt beyond the orbit of Neptune. Pluto has a thin, extended atmosphere, probably methane and nitrogen, which is slowly escaping because of Pluto’s low gravity. This puts it in a somewhat intermediate state between a freely outflowing cometary coma and a bound planetary atmosphere. Spectroscopic evidence shows that methane frost covers much of the surface of Pluto, whereas its largest satellite Charon appears to be covered with water frost. Nitrogen frost has also been detected on Pluto. The density of Pluto is $\sim 2 \text{ g/cm}^3$, suggesting that the rocky component of the dwarf planet accounts for about 70% of its total mass.
There also exist thermally processed asteroids, such as Vesta, whose surface material resembles a basaltic lava flow, and iron–nickel objects, presumably the differentiated cores of larger asteroids that were subsequently disrupted by collisions. The thermal gradient that processed the asteroids appears to be very steep and likely cannot be explained simply by the individual distances of these bodies from the forming proto-Sun. Rather, various special mechanisms such as magnetic induction, short-lived radioisotopes, or massive solar flares have been invoked to explain the heating event that so strongly processed the inner third of the asteroid belt.

The largest asteroid is Ceres, now classified as a dwarf planet, at a mean distance of 2.77 AU from the Sun. Ceres was the first asteroid discovered, by G. Piazzi on January 1, 1801. Ceres is 948 km in diameter, rotates in 9.075 h, and appears to have a surface composition similar to that of carbonaceous chondrite meteorites. The second largest asteroid is Pallas, also a carbonaceous type with a diameter of 532 km. Pallas is also at 2.77 AU, but its orbit has an unusually large inclination of 34.8°. Over 385,000 asteroids have had their orbits accurately determined and have been given official numbers in the asteroid catalog (as of January 2014). Another 247,300 asteroids have been observed well enough to obtain preliminary orbits, 130,200 of them at more than one opposition. Note that these numbers include all objects nominally classified as asteroids: main belt, near-Earth, Trojans, Centaurs, and KBOs (including Pluto, Eris, Haumea, and Makemake).

As a result of the large number of objects in the asteroid belt, impacts and collisions are frequent. Several “families” of asteroids have been identified by their closely grouped orbital elements and are likely fragments of larger asteroids that collided. Spectroscopic studies have shown that the members of these families often have very similar surface compositions, further evidence that they are related. The largest asteroids such as Ceres, Pallas, and Vesta are likely too large to be disrupted by impacts, but most of the smaller asteroids have probably been collisionally processed. Increasing evidence suggests that many asteroids may be “rubble piles”, that is, asteroids that have been broken up but not dispersed by previous collisions and that now form a single but poorly consolidated body.
Beyond the main asteroid belt there exist small groups of asteroids locked in dynamical resonances with Jupiter. These include the Hildas at the 3:2 mean-motion resonance, the Thule group at the 4:3 resonance, and the Trojans, which are in a 1:1 mean-motion resonance with Jupiter. The effect of the resonances is to prevent these asteroids from making close approaches to Jupiter, even though many of the asteroids are in Jupiter-crossing orbits.

The Trojans are particularly interesting. They are essentially in the same orbit as Jupiter, but they librate about points 60° ahead and 60° behind the planet in its orbit, known as the Lagrange L4 and L5 points. These are pseudostable points in the three-body problem (Sun—Jupiter—asteroid) where bodies can remain dynamically stable for extended periods of time. Some estimates have placed the total number of objects in the Jupiter L4 and L5 Trojan swarms as equivalent to the population of the main asteroid belt. Trojan-type 1:1 librators have also been found for the Earth (one), Mars (three), Uranus (one), and Neptune (nine). Interestingly, the Saturnian satellites Dione and Tethys also have small satellites locked in Trojan-type librations in their respective orbits.

Much of what we know about the asteroid belt and about the early history of the solar system comes from meteorites recovered on the Earth. It appears that the asteroid belt is the source of almost all recovered meteorites. A modest number of meteorites that are from the Moon and from Mars, presumably blasted off of those bodies by asteroid and/or comet impacts, have been found. Cometary meteoroids are thought to be too fragile to survive atmospheric entry. In addition, cometary meteoroids typically encounter the Earth at higher velocities than asteroidal debris and thus are more likely to fragment and burn up during atmospheric entry. However, we may have cometary meteorites in our sample collections and simply not yet be knowledgeable enough to recognize them.

Recovered meteorites are roughly equally split between silicate and carbonaceous types, with a few percent being iron—nickel meteorites. The most primitive meteorites (i.e. the meteorites which appear to show the least processing in the solar nebula) are the volatile-rich carbonaceous chondrites. However, even these meteorites show evidence of some thermal processing and aqueous alteration (i.e. processing in the presence of liquid water). Study of carbonaceous and ordinary (silicate) chondrites provides significant information on the composition of the original solar nebula, on the physical and chemical processes operating in the solar nebula, and on the chronology of the early solar system.

The other major group of primitive bodies in the solar system is the comets. Because comets formed farther from the Sun than the asteroids, in colder environments, they contain a significant fraction of volatile ices. Water ice is the dominant and most stable volatile. Typical comets also contain modest amounts of CO, CO2, CH4, NH3, H2CO, and CH3OH, most likely in the form of ices, but possibly also contained within complex organic molecules and/or in clathrate hydrates. Organics make up a significant fraction of the cometary nucleus, as well as silicate grains. F. Whipple described this icy conglomerate mix as “a dirty snowball”, although the term “frozen mudball” may be more appropriate since the comets are more than 60% organics and silicates. It appears that the composition of comets is very similar to the condensed (solid) grains and ices observed in dense interstellar cloud cores where new stars are forming, with little or no evidence of processing in the solar nebula. Thus, comets appear to be the most primitive bodies in the solar system. As a result, the study of comets is extremely valuable for learning about the origin of the planetary system and the conditions in the solar nebula 4.567 billion years ago.

Five cometary nuclei—periodic comets Halley, Borrelly, Wild 2, Tempel 1, and Hartley 2—have been encountered by interplanetary spacecraft and imaged (Figure 1.6). These irregular nuclei range from about 2 to 12 km in mean diameter and have low albedos, only 3—4%. The nuclei exhibit a variety of complex surface morphologies unlike any other bodies in the solar system. It has been suggested that cometary nuclei are weakly bound conglomerations of smaller dirty snowballs, assembled at low velocity and low temperature in the giant planets region (and beyond in the Kuiper belt) of the solar nebula. Thus, comets may be “primordial rubble piles”, in some ways similar to the asteroids. Recent studies have suggested that cometary nuclei, like the asteroids, may have undergone intense collisional evolution, either while resident in the Kuiper belt or in the giant planets region prior to their dynamical ejection to the Oort cloud.

Subtle and not-so-subtle differences in cometary compositions have been observed. However, it is not entirely clear if these differences are intrinsic or due to the physical evolution of cometary surfaces over many close approaches to the Sun. Because the comets that originated among the giant planets have all been ejected to the Oort cloud or to interstellar space, the compositional spectrum resulting from the heliocentric thermal profile is not spatially preserved as it has been in the asteroid belt. Although comets in the classical Kuiper belt are likely located close to their formation distances, physical studies of these distant objects are still in an early stage. There is an observed compositional trend, but it is associated with orbital eccentricity and inclination, rather than semimajor axis.

3.3. Satellites, Rings, and Things

The natural satellites of the planets, listed in the appendix to this volume, show as much diversity as the planets they orbit (see Figure 1.7). Among the terrestrial planets, the
only known satellites are the Earth’s Moon and the two small moons of Mars, Phobos and Deimos. The Earth’s Moon is unusual in that it is so large relative to its primary. The Moon has a silicate composition similar to the Earth’s mantle and a small iron core.

It is now widely believed that the Moon formed as a result of a collision between the proto-Earth and another protoplanet about the size of Mars, late in the accretion of the terrestrial planets. Such “giant impacts” are now recognized as being capable of explaining many of the features of the solar system, such as the unusually high density of Mercury and the large obliquities of several of the planetary rotation axes. In the case of the Earth, the collision with another protoplanet resulted in the cores of the two planets merging, while a fraction of the mantles of both bodies was thrown into orbit around the Earth where some of the material accreted to form the Moon. The tidal interaction between the Earth and the Moon then slowly evolved the orbit of the Moon outward to its present position, at the same time slowing the rotation of both the Earth and the Moon. The giant impact hypothesis is capable of explaining many of the features of the Earth–Moon system, including the similarity in composition between the Moon and the Earth’s mantle, the lack of a significant iron core within the Moon, and the high angular momentum of the Earth–Moon system.

Like most large natural satellites, the Moon has tidally evolved to where its rotation period matches its revolution period in its orbit. This is known as synchronous rotation. It results in the Moon showing the same face to the Earth at all times, although there are small departures from this because of the eccentricity and inclination of the Moon’s orbit.

The Moon’s surface displays a record of the intense bombardment all the planets have undergone over the history of the solar system. Returned lunar samples have been age-dated based on decay of long-lived radioisotopes. This has allowed the determination of a chronology of lunar bombardment by comparing the sample ages with the crater counts on the lunar plains where the samples were collected. The lunar plains, or maria, are the result of
massive eruptions of lava during the first billion years of the Moon’s history. The revealed chronology shows that the Moon experienced a massive bombardment between 4.2 and 3.8 billion years ago, known as the Late Heavy Bombardment. This time period is relatively late as compared with the 100–200 million years required to form the terrestrial planets and to clear the orbital zones of most interplanetary debris. Similarities in crater size distributions on the Moon, Mercury, and Mars suggest that the Late Heavy Bombardment swept over all the terrestrial planets. Recent explanations for the Late Heavy Bombardment have focused on the possibility that it came from the clearing of the outer planets’ zones of their cometary debris. However, the detailed dynamical calculations of the timescales for that process are still being determined.

Like almost all other satellites in the solar system, the Moon has no substantial atmosphere. There is a transient atmosphere due to helium atoms in the solar wind striking the lunar surface and being captured. Argon has been detected escaping from surface rocks and being temporarily cold-trapped during the lunar night. Also, sodium and potassium have been detected, likely the result of sputtering of surface materials due to solar wind particles, as on Mercury. Water ice has been detected in craters at the Moon’s south pole, but in very limited quantities.

Unlike the Earth’s Moon, the two natural satellites of Mars are both small, irregular bodies, and in orbits relatively close to the planet. In fact, Phobos, the larger and closer satellite, orbits Mars faster than the planet rotates. Both of the Martian satellites have surface compositions that appear to be similar to carbonaceous chondrites. This has resulted in speculation that the satellites are captured asteroids. A problem with this hypothesis is that Mars is located close to the inner edge of the asteroid belt, where thermally processed silicate asteroids dominate the asteroid population, and where carbonaceous asteroids are relatively rare. Also, both satellites are located very close to the planet and in near-circular orbits, which is unusual for captured objects.

In contrast to the satellites of the terrestrial planets, the satellites of the giant planets are numerous and are arranged in complex systems. Jupiter has four major satellites, easily
visible in small telescopes from Earth, and 63 known lesser satellites. The discovery of the four major satellites by Galileo in 1610, now known as the Galilean satellites, was one of the early confirmations of the Copernican theory of a heliocentric solar system. The innermost Galilean satellite, Io, is about the same size as the Earth’s Moon and has active vulcanism on its surface as a result of Jupiter’s tidal perturbation and the gravitational interaction with Europa and Ganymede (see Section 3.1). The next satellite outward is Europa, somewhat smaller than Io, which appears to have a thin ice crust overlying a possible liquid water ocean, also the result of tidal heating by Jupiter and the satellite—satellite gravitational interactions. Estimates of the age of the surface of Europa, based on counting impact craters, are very young, suggesting that the thin ice crust may repeatedly break up and reform. The next satellite outward from Jupiter is Ganymede, the largest satellite in the solar system, even larger than the planet Mercury. Ganymede is another icy satellite and shows evidence of tectonic activity and of being partially resurfaced at some time(s) in its past. The final Galilean satellite is Callisto, another icy satellite that appears to preserve an impact record of comets and asteroids dating back to the origin of the solar system. As previously noted, the orbits of the inner three Galilean satellites are locked into a 4:2:1 mean-motion resonance.

The lesser satellites of Jupiter include four within the orbit of Io, and 59 at very large distance from the planet. The latter are mostly in retrograde orbits, which suggests that they are likely captured comets and asteroids. The orbital parameters of many of these satellites fall into several tightly associated groups. This suggests that each group consists of fragments of a larger object that was disrupted, most likely by a collision with another asteroid or comet. Possibly, the collision occurred within the gravitational sphere of Jupiter, which then could have led to the dynamical capture of some of the fragments.

All the close-orbiting Jovian satellites (out to the orbit of Callisto) appear to be in synchronous rotation with Jupiter. However, rotation periods have been determined for two of the outer satellites, Himalia and Elara, and these are approximately 8 and 12 h, respectively, much shorter than their ~250-day periods of revolution about the planet.

Saturn’s satellite system is very different from Jupiter’s in that it contains only one large satellite, Titan, comparable in size to the Galilean satellites, seven intermediate-sized satellites, and 54 smaller satellites. Titan is the only satellite in the solar system with a substantial atmosphere. Clouds of organic compounds in its atmosphere prevent easy viewing of the surface of that moon, although the Cassini spacecraft has had success in viewing the surface at infrared and radar wavelengths. The atmosphere is primarily nitrogen and also contains methane and possibly argon. The surface temperature on Titan has been measured at 94 K, and the surface pressure is 1.5 bar. Cassini radar imaging has revealed a complex surface morphology on Titan that includes rivers, lakes, and possible cryovolcanism.

The intermediate satellites of Saturn all appear to have icy compositions and have undergone substantial processing, possibly as a result of tidal heating and also due to collisions. Orbital resonances exist between several pairs of satellites, and most are in synchronous rotation with Saturn. An interesting exception is Hyperion, which is a highly nonspherical body and which appears to be in chaotic rotation. Another moon, Enceladus, has a ring of material in its orbit that likely has come from geysers discovered at the icy satellite’s south pole. Two other satellites, Dione and Tethys, have two companion satellites each, in the same orbit, which oscillate about the Trojan-libration points for the Saturn–Dione and Saturn–Tethys systems, respectively. Yet another particularly interesting satellite of Saturn is Iapetus, which is dark on one hemisphere and bright on the other and has a narrow ridge circling the satellite at its equator. The dark material appears to be a coating on the satellite’s leading hemisphere that is suspected of coming from Phoebe. The equatorial ridge is believed to be a remnant from a time when the satellite was warmer and larger, but this is by no means certain.

Saturn has one very distant, intermediate-sized satellite, Phoebe, which is in a retrograde orbit and which is suspected of being a captured, early solar system planetesimal, albeit a very large one. Phoebe is not in synchronous rotation, but rather has a rotation period of about 10 h. The 54 known small satellites of Saturn include 11 embedded in or immediately adjacent to the planet’s ring system, 4 Trojan-type librators, and 39 in distant orbits. As with Jupiter, the majority of these distant objects are in retrograde orbits and some are in groups, which suggests that they too are collisional fragments.

The Uranian system consists of five intermediate-sized satellites and 22 smaller ones. Again, these are all icy bodies. These satellites also exhibit evidence of past heating and possible tectonic activity. The satellite Miranda is particularly unusual in that it exhibits a wide variety of complex terrains. It has been suggested that Miranda, and possibly many other icy satellites, were collisionally disrupted at some time in their history, and the debris then reaccreted in orbit to form the currently observed satellites, but preserved some of the older surface morphology. Such disruption/reaccretion phases may have even reoccurred on several occasions for some of the satellites over the history of the solar system. Of the smaller Uranian satellites, 13 are embedded in the ring system and nine are in distant, mostly retrograde orbits. Again, these are likely captured objects.

Neptune’s satellite system consists of one large icy satellite and 13 smaller ones. Triton is somewhat larger than Pluto and is unusual in that it is in a retrograde orbit. As a result, the tidal interaction with Neptune is causing the
satellite’s orbit to decay, and eventually Triton will be torn apart by the planet’s gravity when it passes within the Roche limit. The retrograde orbit is often cited as evidence that Triton must have been captured from interplanetary space and did not actually form in orbit around the planet. Despite its tremendous distance from the Sun, Triton’s icy surface displays a number of unusual terrain types that strongly suggest thermal processing and possibly even current activity. The Voyager spacecraft photographed what appeared to be plumes from “ice volcanoes” on Triton.

Neptune has one intermediate-sized satellite, Nereid, in a distant and eccentric orbit. The lesser satellites of Neptune include six that are either in or adjacent to the ring system and five in distant orbits, three of which are retrograde.

In addition to their satellite systems, all the giant planets have ring systems (Figure 1.8). As with the satellite systems, each ring system is distinctly different from its neighbors. Jupiter has a single ring at 1.72–1.81 planetary radii, discovered by the Voyager spacecraft. The ring has several components, related to the four small satellites in or close to the ring. The micron-sized ring particles appear to be material sputtered off the embedded satellites.

Saturn has an immense, broad ring system extending between 1.11 and 2.27 planetary radii, easily seen in a small telescope from the Earth. The ring system consists of three major rings, known as A, B, and C ordered from the outside in toward the planet, a diffuse ring labeled D inside the C ring and extending down almost to the top of the Saturnian atmosphere, and several other narrow, individual rings.

Closer examination by the Voyager spacecraft revealed that the A, B, and C rings were each composed of thousands of individual ringlets. This complex structure is the result of mean-motion resonances with the many Saturnian satellites, as well as with small satellites embedded within the rings themselves. Some of the small satellites act as gravitational “shepherds”, focusing the ring particles into narrow ringlets. Additional narrow and diffuse rings are located outside the main ring system.

The Uranian ring system was discovered accidentally in 1977 during observation of a stellar occultation by Uranus. A symmetric pattern of five narrow dips in the stellar signal was seen on both sides of the planet. Later observations of other stellar occultations found an additional five narrow rings. Voyager 2 detected several more, fainter, diffuse
rings and provided detailed imaging of the entire ring system.

The success with finding Uranus’ rings led to similar searches for a ring system around Neptune using stellar occultations. Rings were detected but were not always symmetric about the planet, suggesting gaps in the rings. Subsequent Voyager 2 imaging revealed large azimuthal concentrations of material in one of the six detected rings.

All of the ring systems are within the Roche limits of their respective planets, at distances where tidal forces from the planet will disrupt any solid body, unless it is small enough and strong enough to be held together by its own material strength. This has led to the general belief that the rings are disrupted satellites, or possibly material that could never successfully form into satellites. Ring particles have typical sizes ranging from micron-sized dust to meter-sized objects and appear to be made primarily of icy materials, although in some cases contaminated with carbonaceous materials. Jupiter’s ring is an exception because it appears to be composed of carbonaceous and silicate materials, with no ice.

Another component of the solar system is the zodiacal dust cloud, a huge, continuous cloud of fine dust extending throughout the planetary region and generally concentrated toward the ecliptic plane. The cloud consists of dust grains liberated from comets as the nucleus ices sublimate and from collisions between asteroids. Comets are estimated to account for about two-thirds of the total material in the zodiacal cloud, with asteroid collisions providing the rest. Dynamical processes tend to spread the dust uniformly around the Sun, although some structure is visible as a result of the most recent asteroid collisions. These structures, or bands as they are also known, are each associated with specific asteroid collisional families.

Dust particles will typically burn up due to friction with the atmosphere when they encounter the Earth, appearing as visible meteors. However, particles less than about 50 \( \mu \)m in radius have sufficiently large area-to-mass ratios that they can be decelerated high in the atmosphere and can radiate away the energy generated by friction without vaporizing the particles. These particles then settle slowly through the atmosphere and are eventually incorporated into terrestrial sediments. In the 1970s, NASA began experimenting with collecting interplanetary dust particles (IDPs, also known as Brownlee particles because of the pioneering work of D. Brownlee) using high-altitude U2 reconnaissance aircraft. Terrestrial sources of particulates in the stratosphere are rare and consist largely of volcanic aerosols and aluminum oxide particles from solid rocket fuel exhausts, each of which are readily distinguishable from extraterrestrial materials.

The composition of the IDPs reflects the range of source bodies that produce them and include ordinary and carbonaceous chondritic material and suspected cometary particles. Because the degree of heating during atmospheric deceleration is a function of the encounter velocity, recovered IDPs are strongly biased toward asteroidal particles from the main belt, which approach the Earth in lower eccentricity orbits. Nevertheless, suspected cometary particles are included in the IDPs. The cometary IDPs show a random, “botryoidal” (cluster-of-grapes) arrangement of submicron silicate grains similar in size to interstellar dust grains, intimately mixed in a carbonaceous matrix. Voids in cometary IDPs may have once been filled by cometary ices. In 2006, the Stardust spacecraft returned samples of cometary dust collected during a flyby of comet Wild 2; these are providing an important comparison with the IDPs collected by high-flying aircraft. An example of a suspected cometary IDP is shown in Figure 1.9.

Extraterrestrial particulates are also collected on the Earth in Antarctic ice cores, in melt ponds in Greenland, and as millimeter-sized silicate and nickel—iron melt products in ocean sediments. The IDP component in terrestrial sediments can be determined by measuring the abundance of \(^{3}\)He. \(^{3}\)He has normal abundances in terrestrial materials of \(10^{-6}\) or less. The \(^{3}\)He is implanted in the IDP grains during their exposure to the solar wind. Using this technique, one can look for variations in the infall rate of extraterrestrial particulates over time, and such variations are seen, sometimes correlated with impact events on the Earth.

### 3.4. The Solar Wind and the Heliosphere

A largely unseen part of the solar system is the solar wind, an ionized plasma that streams continuously into space...
from the Sun. The solar wind is composed primarily of protons (hydrogen nuclei) and electrons with some alpha particles (helium nuclei) and trace amounts of heavier ions. It is accelerated to supersonic speed in the solar corona and streams outward at a typical velocity of 400 km/s. The solar wind is highly variable, changing with both the solar rotation period of \(\sim 25\) days and with the 22-year solar cycle, as well as on much more rapid timescales. As the solar wind expands outward, it carries the solar magnetic field with it in a spiral pattern caused by the rotation of the Sun. The solar wind was first inferred in the early 1950s by L. Biermann based on observations of cometary Type I plasma tails. The theory of the supersonic solar wind was first described by E. N. Parker in 1958, and the solar wind itself was detected in 1962 by the Explorer 10 spacecraft in Earth orbit, and the Mariner 2 spacecraft while en route to a flyby of Venus.

The solar wind interaction with the planets and the other bodies in the solar system is also highly variable, depending primarily on whether or not the body has its own intrinsic magnetic field. For bodies without a magnetic field, such as Venus and the Moon, the solar wind impinges directly on the top of the atmosphere or on the solid surface, respectively. For bodies like the Earth or Jupiter, which do have magnetic fields, the field acts as a barrier and deflects the solar wind around it. Because the solar wind is expanding at supersonic speeds, a shock wave, or bow shock, develops at the interface between the interplanetary solar wind and the planetary magnetosphere. The planetary magnetospheres can be quite large, extending out \(\sim 12\) planetary radii upstream (sunward) of the Earth, and \(50-100\) radii sunward of Jupiter. Solar wind ions can leak into the planetary magnetospheres near the poles, and these can result in visible aurora, which have been observed on the Earth, Jupiter (Figure 1.10), and Saturn. As it flows past the planet, the interaction of the solar wind with the planetary magnetospheres results in huge magnetotail structures that often extend over interplanetary distances.

All the giant planets, as well as the Earth, have substantial magnetic fields and thus planetary magnetospheres. Mercury has a weak magnetic field, but Venus has no detectable field. Mars has a patchy field, indicative of a past magnetic field at some point in the planet’s history, but it has no organized magnetic field at this time. The Galileo spacecraft detected a magnetic field associated with Ganymede, the largest of the Galilean satellites. However, no magnetic field was detected for Europa or Callisto. The Earth’s Moon has no magnetic field.

The most visible manifestation of the solar wind is cometary plasma tails, which result when the evolving gases in the cometary comae are ionized by sunlight and by charge exchange with the solar wind and then accelerated by the solar magnetic field. The ions stream away from the cometary comae at high velocity in the antisunward direction. Structures in the tail are visible as a result of fluorescence by \(\text{CO}^+\) and other ions, although the most abundant ion in the plasma tails is \(\text{H}_2\text{O}^+\).

At some distance from the Sun, far beyond the orbits of the planets, the solar wind reaches a point where the ram pressure from the wind is equal to the external pressure from the local interstellar wind flowing past the solar system. A termination shock develops upstream of that point, and the solar wind will be decelerated from supersonic to subsonic. Voyager 1 detected the termination shock at 94 AU in 2004 and Voyager 2 detected it at 84 AU in 2007. Beyond this distance is a region called the heliosheath, still dominated by the subsonic solar plasma and extending out another 15–25 AU. The outer boundary of this region is known as the heliopause and defines the limit between solar system-dominated plasma and the interstellar wind. It is not currently known if the flow of interstellar medium past the solar system is supersonic or subsonic. If it is supersonic, then there must additionally be a bow shock beyond the heliopause, where the interstellar medium encounters the obstacle presented by the heliosphere. A diagram of the major features of the heliosphere is shown in Figure 1.11.

The Voyager 1 spacecraft crossed the heliopause in August 2012 and is now in interstellar space. It is 126.9 AU from the Sun and continues to move outward at 3.6 AU/year (as of February 2014). The Voyager 2 spacecraft continues to study the outermost region of the heliosphere, known as the “heliosheath” and is expected to cross the heliopause in 2017. Voyager 2 is currently 104.0 AU from the Sun and is moving outward at 3.3 AU/year. The Voyager 1 and 2 spacecraft are expected to continue to send measurements at least until the year 2020, when they will be at about 148 and 129 AU from the Sun, respectively. To many planetary scientists, the heliopause defines the boundary of the solar system because it marks the changeover from the solar wind to an interstellar medium-dominated space. However, as already noted, the Sun’s gravitational sphere of influence extends out much farther,
picture assumes that the Sun is a typical star and that it
and processes in the formation of the solar system. That
some of the observed processes or events may not be appli-
cable to the formation of our own Sun and planetary system.
We must learn which qualities reflect that often violent evo-
condensed from the natal interstellar cloud. We must learn
itself and study of star formation in nearby giant molecular
clouds. The two sources are radically different. In the case
of the solar system, we have an abundance of detailed in-
formation on the planets, their satellites, and numerous
small bodies. But the solar system we see today is highly
evolved and has undergone massive changes since it first
condensed from the natal interstellar cloud. We must learn
to recognize which qualities reflect that often violent evo-
lution and which truly record conditions at the time of solar
system formation.

In contrast, when studying even the closest star-forming
regions (which are about 140 pc from the Sun), we are
handicapped by a lack of adequate resolution and detail. In
addition, we are forced to take a “snapshot” view of many
young stars at different stages in their formation, and from
that attempt to generate a time-ordered sequence of those
different stages and processes involved. When we observe
the formation of other stars, we also need to recognize that
some of the observed processes or events may not be applic-
cable to the formation of our own Sun and planetary system.

Still, a coherent picture has emerged of the major events
and processes in the formation of the solar system. That
picture assumes that the Sun is a typical star and that it
formed in a similar way to many of the low-mass protostars
we see today.

The birthplace of stars is giant molecular clouds in the
galaxy. These huge clouds of molecular hydrogen have
masses of $10^5-10^6 M_\odot$. Within these clouds are denser
regions or cores where star formation actually takes place.
Some process, perhaps the shock wave from a nearby su-
pernova, triggers the gravitational collapse of a cloud core.
Material falls toward the center of the core under its own
self-gravity and a massive object begins to grow at the
center of the cloud. Heated by the gravitational potential
energy of the infalling matter, the object becomes self-
luminous and is then described as a protostar. Although
central pressures and temperatures are not yet high enough
to ignite nuclear fusion, the protostar begins to heat the
growing nebula around it. The timescale of the infall of the
cloud material for a solar mass cloud is about $10^6$ years.

The infalling cloud material consists of both gas and
dust. The gas is mostly hydrogen (75% by mass) and he-
lium (22%). The dust (2%) is a mix of interstellar grains,
including silicates, organics, and condensed ices. A popular
model suggests that the silicate grains are coated with icy
organic mantles. As the dust grains fall inward, they
experience a pressure from the increasing density of gas
toward the center of the nebula. This slows and even halts
the inward radial component of their motion. However, the
dust grains can still move vertically with respect to the
central plane of the nebula, as defined by the rotational
angular momentum vector of the original cloud core. As a
result, the grains settle toward the central plane.

As the grains settle, they begin to collide with one
another. The grains stick and quickly grow from micro-
scopic to macroscopic objects, perhaps meters in size
(initial agglomerations of grains may look very much like
the suspected cometary IDP in Figure 1.9). This process
continues and even increases as the grains reach the denser
environment at the central plane of the nebula. The meter-
sized bodies grow to kilometer-sized bodies and the
kilometer-sized bodies grow to 100 km-sized bodies. These
bodies are known as planetesimals. As a planetesimal be-
gins to acquire significant mass, its cross-section for ac-
cretion grows beyond its physical cross-section because it
is now capable of gravitationally deflecting smaller plane-
tesimals toward it. These larger planetesimals then “run
away” from the others, growing at an ever increasing rate.

The actual process is far more complex than described
here, and there are many details of this scenario that still
need to be worked out. For example, the role of turbulence
in the nebula is not well quantified. Turbulence would tend
to slow or even prevent the accretion of grains into larger
objects. Also, the role of electrostatic and magnetic effects
in the nebula is not understood.

Nevertheless, it appears that accretion in the central
plane of the solar nebula can account for the growth of
planets from interstellar grains. An artist’s concept of the accretion disk in the solar nebula is shown in Figure 1.12. In the inner region of the solar nebula, close to the forming Sun, the higher temperatures would vaporize icy and organic grains, leaving only silicate grains to form the planetesimals, which eventually merged to form the terrestrial planets. At larger distances where the nebula was cooler, organic and icy grains would condense, and these would combine with the silicates to form the cores of the giant planets. Because the total mass of ice and organics may have been several times the mass of silicates, the cores of the giant planets may actually have grown faster than the terrestrial planets interior to them.

At some point, the growing cores of the giant planets became sufficiently massive to begin capturing hydrogen and helium directly from the nebula gas. Because of the lower temperatures in the outer planets zone, the giant planets were able to retain the gas and continue to grow even larger. The terrestrial planets close to the Sun may have acquired some nebula gas, but probably they could not hold on to it at their higher temperatures.

Observations of protostars in nearby molecular clouds have found substantial evidence for accretionary disks and gas nebulae surrounding these stars. The relative ages of these protostars can be estimated by comparing their luminosity and color with theoretical predictions of their location in the Hertzsprung–Russell diagram. One of the more interesting observations is that the nebula dust and gas around solar mass protostars seem to dissipate after about $10^5$ years. It appears that the nebula and dust may be swept away by mass outflows, essentially superpowerful solar winds, from the protostars. If the Sun formed similarly to the protostars we see today, then these observations set strong limits on the likely formation times of Jupiter and Saturn.

An interesting process that must have occurred during the late stages of planetary accretion is “giant impacts”, i.e. collisions between very large protoplanetary objects. As noted in Section 3.3, a giant impact between a Mars-size protoplanet and the proto-Earth is now the accepted explanation for the origin of the Earth’s Moon. Although it was previously thought that such giant impacts were low-probability events, they are now recognized to be a natural consequence of the final stages of planetary accretion.

Another interesting process late in the accretion of the planets is the clearing of debris from the planetary zones. At some point in the growth of the planets, their gravitational spheres of influence grew sufficiently large that an encounter with a planetesimal would more likely lead to the planetesimal being gravitationally scattered into a different orbit, rather than an actual collision. This would be particularly true for the massive giant planets, both because of their stronger gravitational fields and because of their larger distances from the Sun.

Because it is just as likely that a planet will scatter objects inward as outward, the clearing of the planetary zones resulted in planetesimals being flung throughout the solar system and in a massive bombardment of all planets and satellites. Many planetesimals were also flung out of the planetary system to interstellar space or to distant orbits in the Oort cloud. Although the terrestrial planets are too small to eject objects out of the solar system, they can scatter objects to Jupiter-crossing orbits where Jupiter will quickly dispose of them in about $10^6$ years or less.
The clearing of the planetary zones has several interesting consequences. The dynamical interaction between the planets and the remaining planetesimals results in an exchange of angular momentum. Computer-based dynamical simulations have shown that this causes the semimajor axes of the planets to migrate. In general, Saturn, Uranus, and Neptune are expected to first move inward and then later outward as the ejection of material progresses. Jupiter, which ejects the most material because of its huge mass, migrates inward but by only a few tenths of an astronomical unit.

This migration of the giant planets has significant consequences for the populations of small bodies in the planetary region. As the planets move, the locations of their mean-motion and secular resonances will move with them. This will result in some small bodies being captured into resonances while others will be thrown into chaotic orbits, leading to their eventual ejection from the system or possibly to impacts on the planets and the Sun. The radial migration of the giant planets has been invoked both in the clearing of the outer regions of the main asteroid belt and the inner regions of the Kuiper belt.

Another consequence of the clearing of the planetary zones is that rocky planetesimals formed in the terrestrial planets zone will be scattered throughout the giant planets region, and vice versa, for icy planetesimals formed in the outer planets zone. The bombardment of the terrestrial planets by icy planetesimals is of particular interest, both as an explanation for the Late Heavy Bombardment and as a means of delivering the volatile reservoirs of the terrestrial planets. Isotopic studies suggest that some fraction of the water in the Earth’s oceans may have come from comets and/or volatile-rich asteroids, although not all of it. Also, the discovery of an asteroidal-appearing object, 1996 PW, on a long-period comet orbit has provided evidence that asteroids may indeed have been ejected to the Oort cloud, where they may make up 1—3% of the population there.

5. THE SOLAR SYSTEM’S PLACE IN THE GALAXY

The Milky Way galaxy is classified as a barred spiral with loosely wound arms, Sbc in the Hubble catalog of galaxies. It consists of four major structures: the galactic disk, the central bar, the halo, and the corona (Figure 1.13 and 1.14). As the name implies, the disk is a highly flattened, rotating structure about 15—25 kpc in radius and about 0.5—1.3 kpc thick, depending on which population of stars is used to trace the disk. Note that galactic distances are measured in parsecs and kiloparsecs (1000 pc), where a parsec is defined as the distance where a star would have a parallax of 1 arc-second as viewed from the Earth’s orbit. A parsec is equivalent to 206,264.8 AU, or 3.26 light years.

The galactic disk contains 100—400 million relatively young stars and interstellar clouds, arranged in a multiarm spiral structure. At the center of the disk is the bar, a prolate spheroid about 3 kpc in radius in the plane of the disk, and with a radius of about 1.5 kpc perpendicular to the disk. The bar rotates more slowly than the disk and consists largely of densely packed older stars and interstellar clouds. It does not display spiral structure. At the center of the bar is the galactic nucleus, a complex region only 4—5 pc across (see Figure 1.15), which appears to have a super massive black hole at its center. The mass of the central black hole has been estimated at ~4 million \( M_\odot \).

The halo surrounds both of these structures, extending ~30 kpc from the galactic center. The halo has an oblate spheroid shape and contains older stars and globular clusters of stars. The corona appears to be a yet more distant halo extending 60—100 kpc and consists of ionized gas and dark matter, unobservable except for the effect it has on the dynamics of observable bodies in the galaxy. The corona may be several times more massive than the other three galactic components combined.
The galactic disk is visible in the night sky as the Milky Way, a bright band of light extending across the celestial sphere. When examined with a small telescope, the Milky Way is resolved into thousands or even millions of individual stars and numerous nebulae and star clusters. The direction to the center of the galaxy is in the constellation Sagittarius (best seen from the southern hemisphere in June), and the disk appears visibly wider in that direction, which is the view of the central bulge and bar.

The disk is not perfectly flat; there is evidence for warping in the outer reaches of the disk, between 15 and 25 kpc. The warp may be the result of gravitational perturbations due to encounters with other galaxies and/or with the Magellanic Clouds, two nearby, irregular dwarf galaxies that appear to be in orbit around the Milky Way. In addition, the Milky Way’s central bar appears to be tilted relative to the plane of the galactic disk. The nonspherical shape of the bar and the tilt have important implications for understanding stellar dynamics and the long-term evolution of the galaxy.

Stars in the galactic disk have different characteristic velocities as a function of their stellar classification, and hence age. Low-mass older stars, like the Sun, have relatively high random velocities and, as a result, can move farther out of the galactic plane. Younger, more massive stars have lower mean velocities and thus smaller scale heights above and below the plane. Giant molecular clouds, the birthplace of stars, also have low mean velocities and thus are confined to regions relatively close to the galactic plane. The galactic disk rotates clockwise as viewed from “galactic north”, at a relatively constant velocity of \( \sim 220 \text{ km/s} \). This motion is distinctly non-Keplerian, the result of the nonspherical mass distribution in the disk. The rotation velocity for a circular galactic orbit in the galactic plane defines the Local Standard of Rest (LSR). The LSR is then used as the reference frame for describing local stellar dynamics.

The Sun and the solar system are located approximately \( 8.5 \text{ kpc} \) from the galactic center (although some estimates put it closer at \( \sim 7 \text{ kpc} \) or farther at \( 8.7 \text{ kpc} \)), and \( 5-30 \text{ pc} \) above the central plane of the galactic disk. The Sun and the...
The solar system are moving at approximately 17–22 km/s relative to the LSR. The Sun’s velocity vector is currently directed toward a point in the constellation of Hercules, approximately at right ascension 18 h 0 m, and declination +30°, known as the solar apex. Because of this motion relative to the LSR, the solar system’s galactic orbit is not circular. The Sun and planets move in a quasielliptical orbit between about 8.4 and 9.7 kpc from the galactic center, with a period of revolution of about 225–250 million years. The solar system is currently close to and moving inward toward “perigalacticon”, the point in the orbit closest to the galactic center. In addition, the solar system moves perpendicular to the galactic plane in a harmonic fashion, with an estimated period of 52–74 million years, and an amplitude of ±49–93 pc out of the galactic plane. The uncertainties in the estimates of the period and amplitude of the motion are caused by the uncertainty in the amount of dark matter in the galactic disk. The Sun and planets passed through the galactic plane about 2 to 3 million years ago, moving “northward.”

The Sun and solar system are located at the inner edge of one of the spiral arms of the galaxy, known as the Orion or local arm, although also called the “Orion spur”. Nearby spiral structures can be traced by constructing a three-dimensional map of stars, star clusters, and interstellar clouds in the solar neighborhood. Two well-defined neighboring structures are the Perseus arm, farther from the galactic center than the local arm, and the Sagittarius arm, toward the galactic center. The arms are about 0.5 kpc wide, and the spacing between the spiral arms is ~1.2–1.6 kpc.

The Sun’s velocity relative to the LSR is low as compared with other G-type stars, which have typical velocities of 40–45 km/s relative to the LSR. Stars are accelerated by encounters with giant molecular clouds in the galactic disk. Thus, older stars can be accelerated to higher mean velocities, as noted earlier. The reason(s) for the Sun’s low velocity is not known. Velocity-altering encounters with giant molecular clouds occur with a typical frequency of once every 300–500 million years.

The local density of stars in the solar neighborhood is about 0.11/pc³, although many of the stars are in binary or multiple star systems. The local density of binary and multiple star systems is 0.086/pc³. Most of these are low-mass stars, less massive and less luminous than the Sun. The star nearest to the solar system is Proxima Centauri, which is a low-mass (M ≈ 0.1 M☉), distant companion to Alpha Centauri, which itself is a double star system of two closely orbiting solar-type stars. Proxima Centauri is currently about 1.3 pc from the Sun and about 0.06 pc (1.35 × 10⁴ AU) from the Alpha Centauri pair it is orbiting. The second nearest star is Barnard’s star, a fast-moving red dwarf at a distance of 1.83 pc. The brightest star within 5 pc of the Sun is Sirius, an A1 star (M ≈ 2 M☉) about 2.6 pc away. Sirius is also a double star, with a faint, white dwarf companion. The stars in the solar neighborhood are shown in Figure 1.16.

The Sun’s motion relative to the LSR, as well as the random velocities of the stars in the solar neighborhood, will
occasionally result in close encounters between the Sun and other stars. Using the value above for the density of stars in the solar neighborhood, one can predict that ~12 star systems (single or multiple stars) will pass within 1 pc of the Sun per million years. The total number of stellar encounters scales as the square of the encounter distance. This rate has been confirmed in part by data from the Hipparcos astrometry satellite, which measured the distances and proper motions of ~118,000 stars, and which was used to reconstruct the trajectories of stars in the solar neighborhood.

Based on this rate, the closest stellar approach over the lifetime of the solar system would be expected to be at ~900 AU. Such an encounter would result in a major perturbation of the Oort cloud and would eject many comets into interstellar space. It would also send a shower of comets into the planetary region, raising the impact rate on the planets for a period of about 2–3 million years, and having other effects that may be detectable in the stratigraphic record on the Earth or on other planets. A stellar encounter at 900 AU could also have a substantial perturbative effect on the orbits of comets in the Kuiper belt and scattered disk and would likely disrupt the outer regions of those populations. Obviously, the effect that any such stellar passage will have is a strong function of the mass and velocity of the passing star.

Because the Sun likely formed in a star cluster, and because the Sun will move through denser regions of the galactic disk (in particular, the spiral arms), the encounter rate mentioned above is likely a lower limit and was higher at times in the past. That also means that the closest stellar encounters may have been even closer than 900 to the planetary system.

The advent of space-based astronomy, primarily through Earth-orbiting ultraviolet and X-ray telescopes, has made it possible to study the local interstellar medium surrounding the solar system. The structure of the local interstellar medium has turned out to be quite complex. The solar system appears to be on the edge of an expanding bubble of hot plasma about 120 pc in radius, which appears to have originated from multiple supernovae explosions in the Scorpius–Centaurus OB association. The Sco-Cen association is a nearby star-forming region that contains many young, high-mass O- and B-type stars. Such stars have relatively short lifetimes and end their lives in massive supernova explosions, before collapsing into black holes. The expanding shells of hot gas blown off the stars in the supernova explosions are able to “sweep” material before them, leaving a low-density “bubble” of hot plasma.

Within this bubble, known as the Local Bubble, the solar system is at this time within a small interstellar cloud, perhaps 2–5 pc across, known as the Local Interstellar Cloud. That cloud is apparently a fragment of the expanding shells of gas from the supernova explosions, and there appear to be a number of such clouds within the local solar neighborhood.

6. THE FATE OF THE SOLAR SYSTEM

Stars like the Sun are expected to have lifetimes on the main sequence of about $10^{10}$ years. The main sequence lifetime refers to the time period during which the star produces energy through hydrogen fusion in its core. As the hydrogen fuel in the core is slowly depleted over time, the core contracts to maintain the internal pressure. This raises the central temperature and as a result, the rate of nuclear fusion also increases and the star slowly brightens. Thus, temperatures throughout the solar system will slowly increase over time. Presumably, this slow brightening has already been going on since the formation of the Sun and solar system.

A 1-$M_\odot$ star like the Sun is expected to run out of hydrogen at its core in about $10^{10}$ years. As the production of energy declines, the core again contracts. The rising internal temperature and pressure are then able to ignite hydrogen burning in a shell surrounding the depleted core. The hydrogen burning in the shell heats the surrounding mass of the star and causes it to expand. The radius of the star increases and the surface temperature drops. The luminosity of the star increases dramatically, and it becomes a red giant. Eventually the star reaches a brightness about $10^3$ times more luminous than the present-day Sun, a surface temperature of 3000 K, and a radius of 100–200 solar radii. One hundred solar radii is equal to 0.46 AU, larger than the orbit of Mercury. Two hundred radii is just within the orbit of the Earth. Thus, Mercury and likely Venus will be incorporated into the outer shell of the red giant Sun and will be vaporized.

The increased solar luminosity during the red giant phase will result in a fivefold rise in temperatures throughout the solar system. At the Earth’s orbit this temperature increase will vaporize the oceans and roast the planet at a temperature on the order of ~1400 K or more. At Jupiter’s orbit it will melt the icy Galilean satellites and cook them at a more modest temperature of about 600 K, about the same as current noontime temperatures on the surface of Mercury. Typical temperatures at the orbit of Neptune will be about the same as they are today at the orbit of the Earth. Comets in the inner portion of the Kuiper belt will be warmed sufficiently to produce visible comae.

The lowered gravity at the surface of the greatly expanded Sun will result in a substantially increased solar wind, and the Sun will slowly lose mass from its outer envelope. Meanwhile, the core of the Sun will continue to contract until the central temperature and pressure are great enough to ignite helium burning in the core. During this time, hydrogen burning continues in a shell around the core. Helium burning continues during the red giant phase until the helium in the core is also exhausted. The star again
contracts, and this permits helium burning to ignite in a shell around the core. This is an unstable situation, and the star can undergo successive contractions and reignition pulses, during which it will blow off part or all of its outer envelope into space. These huge mass ejections produce an expanding nebula around the star, known as a planetary nebula (because it looks somewhat like the disk of a giant planet through a telescope). For a star with the mass of the Sun, the entire red giant phase lasts about $7 \times 10^8$ years.

As the Sun loses mass in this fashion, the orbits of the surviving planets will slowly spiral outward. This will also be true for comets in the Kuiper belt and Oort cloud. The gravitational sphere of influence of the Sun will shrink as a result of the Sun’s decreasing mass, so comets will be lost to interstellar space at the outer limits of the Oort cloud.

As a red giant star loses mass, its core continues to contract. However, for an initially $1-M_\odot$ star like the Sun, the central pressure and temperature cannot rise sufficiently to ignite carbon burning in the core, the next phase in nuclear fusion. With no way of producing additional energy other than gravitational contraction, the luminosity of the star plunges. The star continues to contract and cool, until the contraction is halted by degenerate electron pressure in the superdense core. At this point, the mass of the star has been reduced to about 70% of its original mass and the diameter is about the same as the present-day Earth. Such a star is known as a white dwarf. The remnants of the previously roasted planets will be plunged into a deep freeze as the luminosity of the white dwarf slowly declines.

The white dwarf star will continue to cool over a period of about $10^9$ years, to the point where its luminosity drops below detectable levels. Such a star is referred to as a black dwarf. A nonluminous star is obviously very difficult to detect. There is some suggestion that they may have been found through an observing technique known as microlensing events. Dark stars provide one of the possible explanations for the dark matter in the galaxy.

7. CONCLUDING REMARKS

This chapter has provided an introduction to the solar system and its varied members, viewing them as components of a large and complex system. Each of them (the Sun, the planets, their satellites, the comets and asteroids, etc.) is also a fascinating world in its own right. The ensuing chapters provide more detailed descriptions of each of these members of the solar system, as well as descriptions of important physical and dynamical processes, discussions of some of the more advanced ways we study the solar system, the search for life elsewhere in the solar system, and finally, the search for planetary systems around other stars.

BIBLIOGRAPHY


The Origin of the Solar System

John E. Chambers
Department of Terrestrial Magnetism, Carnegie Institution of Washington, Washington, DC, USA

Alex N. Halliday
Department of Earth Sciences, University of Oxford, Oxford, UK

1. INTRODUCTION

The origin of the solar system has long been a fascinating subject posing difficult questions of deep significance. It takes one to the heart of the question of our origins, of how we came to be here and why our surroundings look the way they do. Unfortunately, we currently lack a self-consistent model for the origin of the solar system and other planetary systems. The early stages of planet formation are obscure and we have only a modest understanding of how much the orbits of planets change during and after their formation. At present, we cannot say whether terrestrial planets similar to the Earth are commonplace or highly unusual. Nor do we understand where the water came from that makes our planet habitable.

In the face of such uncertainty, one might ask whether we will ever understand how planetary systems form. In fact, the last 10 years have seen rapid progress in almost every area of planetary science, and our understanding of the origin of the solar system and other planetary systems has improved greatly as a result. Planetary science today is as exciting as it has been at any time since the Apollo landings on the Moon, and the coming decade looks set to continue this trend.

Some key recent developments are

1. Two decades ago, the first planet orbiting another Sun-like star was discovered. Since then, hundreds of new planets have been discovered using ground-based telescopes, and several thousand planetary candidates have been identified by the space-based Kepler mission. Most of the first planets to be found appear to be gas giants similar to Jupiter and Saturn. Recently, many smaller planets have been found, and at least some of these may be akin to terrestrial planets like Earth.

2. In the last 10 years there have been a number of highly successful space missions to other bodies in the solar system, including Mercury, Mars, and several asteroids and comets, as well as the ongoing Cassini mission to Saturn. Information and images returned from these missions have transformed our view of these objects, while spacecraft have recently obtained samples of an asteroid, a comet, and particles from the solar wind. All this information is greatly enhancing our understanding of the origin and evolution of the solar system.

3. The discovery that one can physically separate and analyze stardust—presolar grains that can be extracted from meteorites and that formed in the envelopes of other stars, has meant that scientists can for the first time...
time test decades of theory on how stars work. The parallel development of methods for extracting isotopic information at the submicron scale has opened up a new window to the information stored in such grains.

4. The development of multiple collector inductively coupled plasma mass spectrometry has made it possible to use new isotopic systems for determining the mechanisms and timescales for the growth of bodies early in the solar system.

5. Our theoretical understanding of planet formation has advanced substantially in several areas, including new models for the rapid growth of giant planets, a better understanding of the physical and chemical evolution of protoplanetary disks, and the growing realization of the ways in which planets can migrate substantially during and after their formation.

6. Powerful new computer codes and equations of state have been developed recently, which make it possible to make realistic, high-resolution simulations of collisions between planet-sized bodies. These developments are greatly improving the realism of models for planetary growth, and may offer the solution to some long-standing puzzles about the origin of Mercury, the Moon, and asteroids.

Today, the formation of the solar system is being studied using three complementary approaches.

- Astronomical observations of protoplanetary disks around young stars are providing valuable information about probable conditions during the early history of the solar system and the timescales involved in planet formation. The discovery of new planets orbiting other stars is adding to the astonishing diversity of possible planetary systems, and providing additional tests for theories of how planetary systems form.

- Physical, chemical, and isotopic analysis of meteorites and samples returned by space missions is generating important information about the formation and evolution of objects in the solar system and their constituent materials. This field of cosmochemistry has taken off in several important new directions in recent years, including the determination of timescales involved in the formation of the terrestrial planets and asteroids, and constraints on the origin of the materials that make up the Solar System.

- Theoretical calculations and numerical simulations are being used to examine every stage in the formation of the solar system. These provide valuable insights into the complex interplay of physical and chemical processes involved, and help to fill in some of the gaps when astronomical and cosmochemical data are unavailable.

In this chapter we will describe what we currently know about how the solar system formed, and highlight some of the main areas of uncertainty that await future discoveries.

2. STAR FORMATION AND PROTOPLANETARY DISKS

The solar system formed 4.5–4.6 billion years ago by collapse of a portion of a molecular cloud composed of gas and dust, rather like the Eagle or Orion Nebulae. Some of the stardust from that ancient nebula has now been isolated from primitive meteorites. Their isotopic compositions are vastly different from those of our own solar system and provide fingerprints of nearby stars that preceded our Sun. These include red giants, asymptotic giant branch (AGB) stars, supernovae and novae. From studying modern molecular clouds it has also become clear that stars like our Sun can form in significant numbers in close proximity to each other. Such observation also provide clues as to how our own solar system formed because they have provided us with images of circumstellar disks—the environments in which planetary objects are born.

Observations from infrared telescopes such as the Spitzer Space Telescope have shown that many young stars give off more infrared radiation than would be expected for blackbodies of the same size. This infrared excess comes from micron-sized grains of dust orbiting the star in an optically thick (opaque) disk. Dark, dusty disks can be seen with the Hubble Space Telescope surrounding some young stars in the Orion Nebula (Figure 2.1). These disks have been dubbed proplyds, short for protoplanetary disks. It is thought that protoplanetary disks are mostly composed of gas, especially hydrogen and helium, and in a few cases this gas has been detected, although gas is generally much harder to see than dust. The fraction of stars having a massive disk declines with stellar age, and large infrared excesses are rarely seen in stars older than 10 Myrs. In some cases, such as the disk surrounding the

FIGURE 2.1 Proplyds are young stellar objects embedded in an optically dense envelope of gas and dust. The objects shown here are from the Orion Nebula.
star HR 4796A, there are signs that the inner portion of a disk has been cleared of dust (Figure 2.2), perhaps due to the presence of one or more planets.

Roughly half of stars up to a few hundred million years old have low-mass, optically thin (nearly transparent) disks containing some dust but apparently little or no gas. In a few cases, such as the star Beta Pictoris, the disk can be seen at visible wavelengths if the glare from the star itself is blocked. Dust grains in these disks will be quickly accelerated out of the system by the pressure of radiation from the central star, or destroyed by high-speed collisions with other grains. Any primordial dust should have been removed on a timescale that is short compared to the age of the star. For this reason, the dust in these disks is thought to be second-generation material formed by collisions between asteroids or sublimation from comets orbiting these stars in more massive analogues of the Kuiper belt in our own solar system. These are often referred to as debris disks since asteroids and comets are presumed to be debris left over from planet formation. In a few cases, such as Beta Pictoris, a planet has been discovered orbiting the same star, reinforcing the link between disks and planetary systems.

In the solar system, the planets all orbit the Sun in the same direction, and their orbits are very roughly coplanar. This suggests the solar system originated from a disk-shaped region of material referred to as the solar nebula, an idea going back more than two centuries to Kant and later Laplace. The discovery of disks of gas and dust around many young stars provides strong support for this idea, and implies that planet formation is associated with the formation of stars themselves. Stars typically form in clusters of a few hundred to a few thousand objects in dense regions of the interstellar medium called molecular clouds (see Figure 2.3). The gas in molecular clouds is cold (roughly 10 K) and dense compared to that in other regions of space (roughly $10^4$ atoms per cubic centimeter) but still much more tenuous than the gas in a typical laboratory “vacuum”. Stars in these clusters are typically separated by

---

**FIGURE 2.2** The circumstellar disk surrounding HR 4796A as revealed by interferometry measurements of the infrared excess. Note the area close into the star swept clear of dust, which has presumably been incorporated into planetary objects.

**FIGURE 2.3** This Hubble Space Telescope image of the Orion Nebula shows molecular clouds of gas and dust illuminated by radiation from young stars. Some early stars appear shrouded in dusty disks (see Figure 2.1). Scientists think that our solar system formed by collapse of a portion of a similar kind of molecular cloud leading to formation of a new star embedded in a dusty disk. How that collapse occurred is unclear. It may have been triggered by a shock wave carrying material being shed from another star such as an AGB star or supernova.
about 0.1 parsecs (0.3 light-years), much less than the distance between stars in the Sun’s neighborhood.

It is unclear precisely what causes the densest portions of a molecular cloud (called molecular cloud cores) to collapse to form stars. It may be that contraction of a cloud core is inevitable sooner or later due to the gravitational attraction of material in the core, or an external event may cause the triggered collapse of a core. The original triggered collapse theory was based on the sequencing found in the ages of stars in close proximity to one another in molecular clouds. This suggests that the formation and evolution of some stars triggered the formation of additional stars in neighboring regions of the cloud. However, several other triggering mechanisms are possible, such as the impact of energetic radiation and gas ejected from other newly formed stars, the effects of a nearby, pulsating AGB star, or a shock wave from the supernova explosion of a massive star.

Gas in molecular cloud cores is typically moving. When a core collapses, the gas has too much angular momentum for all the material to form a single, isolated star. In many cases a binary star system forms instead. In others cases, a single protostar forms (called a T Tauri star or pre-main sequence star), while a significant fraction of the gas goes into orbit about the star forming a disk that is typically 100 astronomical units (AU) in diameter. Temperatures in T Tauri stars are initially too low for nuclear reactions to take place. However, T Tauri stars are much brighter than older stars like the Sun due to the release of gravitational energy as the star contracts. The initial collapse of a molecular cloud core takes roughly 10^5 years, and material continues to fall onto both the star and its disk until the surrounding molecular cloud core is depleted.

The spectra of T Tauri stars contain strong ultraviolet and visible emission lines caused by hot gas falling onto the star. This provides evidence that disks lose mass over time as material moves inward through the disk and onto the star, a process called viscous accretion. This process provides one reason why older stars do not have disks, the other reason being planet formation itself. Estimated disk accretion rates range from 10^{-6} to 10^{-9} solar masses per year. The mechanism responsible for viscous accretion is unclear. A promising candidate is magnetorotational instability (MRI), in which partially ionized gas in the disk becomes coupled to the local magnetic field. Because stars rotate, the magnetic field sweeps around rapidly, increasing the orbital velocity of material that couples strongly to it and moving it outward. Friction causes the remaining material to move inward. As a result, a disk loses mass to its star and spreads outward over time. This kind of disk evolution explains why the planets currently contain only 0.1% of the mass in the solar system but have retained more than 99% of its angular momentum. MRI requires a certain fraction of the gas to be ionized, and it may not be effective in all portions of a disk, creating so-called dead zones where material flows inward more slowly and the gas becomes denser. Disks are also eroded over time by photoevaporation. In this process, gas is accelerated when atoms absorb ultraviolet photons from the central star or nearby, energetic stars, until the gas is moving fast enough to escape into interstellar space.

T Tauri stars often have jets of material moving rapidly away from the star perpendicular to the plane of the disk. These jets are powered by the inward accretion of material through the disk coupled with the rotating magnetic field. Outward flowing winds also arise from the inner portions of a disk. T Tauri stars are strong emitters of X-rays, generating fluxes up to 10^4 times greater than that of the Sun during the strongest solar flares. Careful sampling of large populations of young solar mass stars in the Orion Nebula shows that this is normal behavior in young stars. This energetic flare activity is strongest in the first million years and declines at later times, persisting for up to 10^8 years. From this it has been concluded that the young Sun generated 10^5 times as many energetic protons as today. It is thought that reactions between these protons and material in the disk may have provided some of the short-lived isotopes whose daughter products are seen today in meteorites, although the formation of most of these isotopes predate that of the solar system (see Section 4).

The minimum mass of material that passed through the solar nebula can be estimated from the total mass of the planets, asteroids, and comets in the solar system. However, all these objects are depleted in hydrogen and helium relative to the Sun. Ninety percent of the mass of the terrestrial planets is made up of oxygen, magnesium, silicon, and iron (Figure 2.4), and while Jupiter and Saturn are mostly composed of hydrogen and helium, they are enriched in the heavier elements compared to the Sun. When the missing hydrogen and helium is added, the minimum-mass solar nebula (MMSN) turns out to be 1–2% of the Sun’s mass. The major uncertainties in this

---

**FIGURE 2.4** Pie chart showing the bulk composition of the Earth. Most of the iron (Fe), nickel (Ni), and sulfur (S) are in Earth’s core, while the silicate Earth mostly contains magnesium (Mg), silicon (Si), and oxygen (O) together with some iron.
number come from the fact that the interior compositions of the giant planets and the initial mass of the Kuiper belt are poorly known. Not all this mass necessarily existed in the nebula at the same time, but it must have been present at some point. Current theoretical models predict that planet formation is an inefficient process, with some mass falling into the Sun or being ejected into interstellar space, so the solar nebula was probably more massive than the MMSN.

Gas in the solar nebula became hotter as it viscously accreted toward the Sun, releasing gravitational energy and absorbing sunlight. The presence of large amounts of dust meant the inner portions of the nebula were optically thick to infrared radiation so these regions held on to much of this heat. Numerical disk models show that temperatures probably exceeded 1500 K in the terrestrial-planet-forming region early in the disk’s history. Viscous heating mainly took place at the disk midplane where most of the mass was concentrated. The surfaces of the disk would have been much cooler. The amount of energy generated by viscous accretion declined rapidly with distance from the Sun. In the outer nebula, solar irradiation was the more important effect. Protoplanetary disks are thought to be flared, so that their vertical thickness grows more rapidly than their radius, As a result, the surface layers are always irradiated by the central star. For this reason, the surface layers of the outer solar nebula may have been warmer than the midplane.

The nebula cooled over time as the viscous accretion rate declined and dust was swept up by larger bodies, reducing the optical depth. In the inner nebula, cooling was probably rapid. Models show that at the midplane at 1 AU, the temperature probably fell to about 300 K after 10^5 years. Because the energy generated by viscous accretion and solar irradiation declined with distance from the Sun, disk temperatures also declined with heliocentric distance. At some distance from the Sun, temperatures became low enough for water ice to form, a location referred to as the ice line. Initially, the ice line may have been 5–6 AU from the Sun, but it moved inward over time as the nebula cooled. Some asteroids contain hydrated minerals formed by reactions between water ice and dry rock. This suggests water ice was present when these asteroids formed, in which case the ice line would have been no more than 2–3 AU from the Sun at the time.

Meter-sized icy bodies drifted rapidly inward through the solar nebula due to gas drag (see Section 5). When these objects crossed the ice line they would have evaporated, depositing water vapor in the nebular gas. As a result, the inner nebula probably became more oxidizing over time as the level of oxygen from water increased. When the flux of drifting particles dwindled, the inner nebula may have become chemically reducing again, as water vapor diffused outward across the ice line, froze to form ice, and became incorporated into growing planets.

3. METEORITES AND THE ORIGIN OF THE SOLAR SYSTEM

Much of the above is based on theory and observations of other stars. To find out how our own solar system formed it is necessary to study meteorites and interplanetary dust particles (IDPs). These are fragments of rock and metal from other bodies in the solar system that have fallen to Earth and survived passage through its atmosphere. Meteorites and IDPs tend to have broadly similar compositions, and the difference is mainly one of size. IDPs are much smaller of the two, typically 10–100 μm in diameter, while meteorites can range up to several meters in size. Most such objects are quite unlike any objects formed on Earth. Therefore, we cannot readily link them to natural present-day processes as earth scientists do when unraveling past geological history. Yet the approaches that are used are in some respects very similar. The research that is conducted on meteorites and IDPs is dominated by two fields: petrography and geochemistry. Petrography is the detailed examination of mineralogical and textural features. Geochemistry uses the isotopic and chemical compositions. This combined approach to these fascinating archives has provided a vast amount of information on our Sun and solar system and how they formed. We know about the stars and events that predated formation of the Sun, the nature of the material from which the planets were built, the solar nebula, the timescales for planetary accretion, and the interior workings and geological histories of other planets. Not only these, meteorites provide an essential frame of reference for understanding how our own planet Earth formed and differentiated.

The geochemistry of meteorites and IDPs provides evidence that the Sun’s protoplanetary disk as well as the planets it seeded had a composition that was similar in some respects to that of the Sun itself (Figure 2.5). In other respects, however, it is clear the disk was a highly modified residuum that generated a vast range of planetary compositions. The composition of the Sun can be estimated from the depths of lines associated with each element in the Sun’s spectra (although this is problematic for the lightest elements and the noble gases). Today, the Sun contains almost 99.9% of the total mass of the solar system. A sizable fraction of this material passed through the solar nebula at some point, which tells us that the composition of the original nebula would have been similar to that of the Sun today. The challenge is therefore to explain how it is possible that a disk that formed gas giant objects like Jupiter and Saturn with compositions like the Sun, also generated rocky terrestrial planets like the Earth (Figure 2.4).

Most meteorites are thought to come from parent bodies in the main asteroid belt that formed during the first few million years of the solar system. As a result, these objects
carry a record of processes that occurred in the solar nebula during the formation of the planets. In a few cases, the trajectories of falling meteorites have been used to establish that they arrived on orbits coming from the asteroid belt. Most other meteorites are deduced to come from asteroids based on their age and composition. IDPs are thought to come from both asteroids and comets. A few meteorites did not originate in the asteroid belt. The young ages and abundances of the noble gases trapped inside the Shergottite-Nakhlite-Chassignite (SNC) meteorites suggest they come from Mars. Roughly a 100 SNC meteorites have been found to date, and a comparable number of lunar meteorites from the Moon are also known.

The Earth is currently accumulating meteoritic material at the rate of about $5 \times 10^7$ kg per year. At this rate it would take more than $10^{17}$ years to obtain the Earth’s current mass of $5.97 \times 10^{24}$ kg, which is much longer than the age of the universe. While it is thought that the Earth did form as the result of the accumulation of smaller bodies, it is clear that the rate of impacts was much higher while the planets were forming than it is today.

Broadly speaking meteorites can be divided into three types: chondrites, achondrites, and irons, which can be distinguished as follows:

1. **Chondrites** are mixtures of grains from submicron-sized dust to millimeter- to centimeter-sized particles of rock and metal, apparently assembled in the solar nebula. Most elements in chondrites are present in broadly similar ratios to those in the Sun, with the exception of carbon, nitrogen, hydrogen, and the noble gases, which are all highly depleted. For this reason chondrites have long been viewed as representative of the dust and debris in the circumstellar disk from which the planets formed. So, for example, **refractory** elements that would have resided in solid phases in the circumstellar disk have chondritic (and therefore solar) relative proportions in the Earth, even though the volatile elements are vastly depleted. The nonmetallic components of chondrites are mostly silicates such as olivine and pyroxene. **Chondrules** are a major component of most chondrites (see Figure 2.6). These are roughly millimeter-sized rounded beads of rock that formed by melting, either partially or completely. Their mineral-grain textures suggest they cooled over a period of a few hours, presumably in the nebula, with the heating possibly caused by passage through shock waves in the nebular gas. Some chondrules are thought to have formed later in collisions between planetary objects. Most chondrites also contain **calcium—aluminum-rich inclusions** (CAIs, see Figure 2.7), which have chemical compositions similar to those predicted for objects that condensed from a gas of roughly solar composition at very high temperatures. It is possible that CAIs formed in the very innermost regions of the solar nebula close to the Sun. Dating based on radioactive isotopes suggest that CAIs are the oldest surviving materials to have formed in the solar system. CAIs in the Efremovka chondrite are $4.5673 \pm 0.0002$ Ga old based on the $^{235/238}\text{U} - ^{207/206}\text{Pb}$ system, and this date is often used to define the canonical start to the solar system. The oldest chondrules appear to have

![FIGURE 2.5](image_url) The abundances of elements in our Sun and solar system is estimated from the spectroscopic determination of the composition of the Sun and the laboratory analysis of primitive meteorites called carbonaceous chondrites—thought to represent unprocessed dust and other solid debris from the circumstellar disk. To compare the abundances of different elements it is customary to scale the elements relative to 1 million atoms of silicon. The pattern provides powerful clues to how the various elements were created. See text for details. Based on a figure in Broecker, W. S. How to build a habitable planet with kind permission.

![FIGURE 2.6](image_url) Chondrules are spherical objects, sometimes partly flattened and composed of mafic silicate minerals, metals, and oxides. They are thought to form by sudden (flash) heating in the solar nebula. Some formed as much as 2 million years after the start of the solar system. Photograph courtesy of Drs M. Grady and S. Russell and the Natural History Museum, London.
formed at about the same time, but most chondrules are 1–3 million years younger than this (Figure 2.8). The space between the chondrules and CAIs in chondrites is filled with fine-grained dust called matrix. Most chondrites are variably depleted in moderately volatile elements like potassium (K) and rubidium (Rb) (Figure 2.9). This depletion is more a feature of the chondrules and CAIs than the matrix. Chondrites are subdivided into groups of like objects thought to come originally from the same parent body. Currently, about 15 groups are firmly established, half of which are collectively referred to as carbonaceous chondrites.

Chondrite groups

- Eucrite parent body
- Anglite parent body
- Moon
- Earth
- Mars
- Differentiated planetary objects

**FIGURE 2.7** Calcium–aluminum refractory inclusions are found in chondrite meteorites and are thought to be the earliest objects that formed within our solar system. They have a chemical composition consistent with condensation from a hot gas of solar composition. How they formed exactly is unclear but some have suggested they were produced close in to the Sun and then scattered across the disk. Photograph courtesy of Drs M. Grady and S. Russell and the Natural History Museum, London.

**FIGURE 2.8** The current best estimates for the timescales over which very early inner solar system objects and the terrestrial planets formed. The approximated mean life of accretion is the time taken to achieve 63% growth at exponentially decreasing rates of growth. Based on a figure that first appeared in Halliday, A. N., & Kleine, T. (2006).

**FIGURE 2.9** Comparison between the K/U and Rb/Sr ratios of the Earth and other differentiated objects compared with chondrites. The alkali elements K and Rb are both relatively volatile compared with U and Sr, which are refractory. Therefore, these trace element ratios provide an indication of the degree of volatile element depletion in inner solar system differentiated planets relative to chondrites, which are relatively primitive. It can be seen that the differentiated objects are more depleted in moderately volatile elements than are chondrites. Based on a figure that first appeared in Halliday, A. N., & Porcelli, D. (2001).
These tend to be richer in highly volatile elements such as carbon and nitrogen compared to other chondrites, although as with all meteorites these elements are less abundant than they are in the Sun. Ordinary chondrites are more depleted in certain volatile elements than carbonaceous chondrites, and are largely made of silicates and metal grains. Enstatite chondrites are similar but highly reduced. Chondrules are absent from the most primitive, volatile-rich group of carbonaceous chondrites (the CI group), either because their parent body formed entirely from matrixlike material or because chondrule structures have been erased by subsequent reactions with water in the parent body. Chondrites also contain presolar grains that are sub-micron grains that are highly anomalous isotopically and have compositions that match those predicted to form by condensation in the outer envelopes of various stars. These represent a remarkable source of information on stellar nucleosynthesis and can be used to test theoretical models.

2. **Achondrites** are silicate-rich mafic and ultramafic igneous rocks not too dissimilar from those forming on Earth but with slightly different chemistry and isotopic compositions. They clearly represent the near-surface rocks of planets and asteroids that have melted and differentiated. A few achondrites come from asteroids that appear to have undergone only partial differentiation. In principle, it is possible to group achondrites and distinguish which planet or asteroid they came from. The oxygen isotopic composition of a meteorite is particularly useful in this respect. Isotopically, oxygen is extremely heterogeneous in the solar system, and planets that formed in different parts of the nebula seem to have specific oxygen isotope compositions. This makes it possible to link all the Martian meteorites together, for example (Figure 2.10). These meteorites are specifically linked to Mars because nearly all of them are too young to have formed on any asteroid; they had to come from an object that was large enough to be geologically active in the recent past. This was confirmed by a very close match between the composition of the atmosphere measured with the Viking lander and that measured in fluids trapped in alteration products in Martian meteorites. In fact, Martian meteorites provide an astonishing archive of information into how Mars formed and evolved as discussed in Section 6. To date, only one asteroidal source has been positively identified: Vesta, whose spectrum and orbital location strongly suggest it is the source of the howardite, eucrite, and diogenite (HED) meteorites.

3. **Irons** (see Figure 2.11) are largely composed of iron, nickel (about 10% by mass), and sulfides, together with

![FIGURE 2.10](image1.png)

**FIGURE 2.10** The oxygen isotopic composition of the components in chondrites, in particular CAIs, is highly heterogeneous for reasons that are unclear. (The figure shows deviations in parts per thousand relative to Earth’s oceans or SMOW - standard mean ocean water.) The net result of this variability is that different planets possess distinct oxygen isotopic compositions that define an individual mass fractionation lines as shown here for eucrites, howardites, and diogenites, which come from Vesta and Martian (SNC) meteorites, thought to come from Mars. The Moon is thought to have formed from the debris produced in a giant impact between the proto-Earth when 90% formed and an impacting Mars-sized planet sometimes named “Theia”. The fact that the data for lunar samples are collinear with the terrestrial fractionation line could mean that the Moon formed from the Earth, or the planet from which it was created was formed at the same heliocentric distance, or it could mean that the silicate reservoirs of the two planets homogenized during the impact process, for example, by mixing in a vapor cloud from which lunar material condensed. From Halliday, A. N. (2003).

![FIGURE 2.11](image2.png)

**FIGURE 2.11** Iron meteorites are the most abundant kind of meteorite found because they are distinctive and survive long after other kinds of meteorite are destroyed by weathering. In contrast chondrites are the most abundant class of meteorite observed to fall. Some iron meteorites are thought to represent disrupted fragments of planetesimal cores. Others appear to have formed at low pressures, probably as metal-rich pools formed from impacts on asteroids. The Henbury meteorite shown here is a type IIIAB magmatic iron that fell near Alice Springs, Australia, about 5000 years ago. The texture shown on the sawn face are Windmanstatten patterns formed by slow cooling, consistent with an origin from a core located deep within a meteorite parent body. Photograph courtesy of Drs M. Grady and S. Russell and the Natural History Museum, London.
other elements that have a chemical affinity for iron, called siderophile elements. Like chondrites, irons can be grouped according to their likely parent body, and several dozen groups or unique irons have been found. The textures of mineral grains in iron meteorites have been used to estimate how quickly their parent bodies cooled, and thus the depth at which they formed. It appears that most irons are samples of metallic cores of small asteroidal parent bodies, 10–100 km in radius. These appear to have formed very early, probably within a million years of CAIs, when there was considerable heat available from decay of short-lived radioactive nuclides (see Section 6). Others appear to have formed by impact melting at the surface of asteroids and these are later. A rare class of stony-iron meteorites (amounting to about 5% of all nonchondritic meteorites) called pallasites contain an intricate mixture of metal and silicate (Figure 2.12). It is thought these come from the core–mantle boundary regions of differentiated asteroids that broke up during a collision.

Note that there are no clear examples of mantle material within meteorite collections. The isotopic compositions of some elements in irons reveal that they have been exposed to cosmic rays for long periods—up to hundreds of millions of years. This means their parent bodies broke up a long time ago. Because they are extremely hard they survived the collisions that destroyed their parent body as well as any subsequent impacts. In contrast fragments of mantle material (as with samples excavated by volcanoes on the Earth) are extremely friable and are more easily disrupted by collisions.

Survivability is also an issue for meteorites entering Earth's atmosphere and being recovered in recognizable form. Chondrites and achondrites are mainly composed of silicates that undergo physical and chemical alterations on the surface of Earth more rapidly than the material in iron meteorites. Furthermore, iron meteorites are highly distinctive, so they are easier to recognize than stony meteorites. For this reason, most meteorites found on the ground are irons, whereas most meteorites that are seen to fall from the sky (referred to as falls) are actually chondrites. Most falls are ordinary chondrites, which probably reflects the fact that they survive passage through the atmosphere better than the weaker carbonaceous chondrites. The parent bodies of ordinary chondrites may also have orbits in the asteroid belt that favor their delivery to Earth. IDPs are less prone to destruction during passage through the atmosphere than meteorites so they probably provide a less biased sample of the true population of interplanetary material. Most IDPs are compositionally similar to carbonaceous rather than ordinary chondrites and this suggests that the asteroid belt is dominated by carbonaceous-chondrite like material.

Mass spectrometric measurements on meteorites and lunar samples provide evidence that the isotopes of most elements are present in similar proportions in the Earth, Moon, Mars, and the asteroids. The isotopes of elements heavier than hydrogen and helium were made by nucleosynthesis in stars which generate extremely anomalous isotopic compositions compared to the solar system. Since the solar nebula probably formed from material from a variety of sources, the observed isotopic homogeneity was originally interpreted as indicative that the inner solar nebula was very hot and planetary material condensed from a ~2000 K gas of solar composition. However, a variety of observations including the preservation of presolar grains in chondrites suggest that the starting point of planet formation was cold dust and gas. This homogeneity is therefore nowadays interpreted as indicating that the inner nebula was initially turbulent, allowing dust to become thoroughly mixed. CAIs sometimes contain nucleosynthetic isotopic anomalies. This suggests that CAIs sampled varied proportions of the isotopes of the elements before they became homogenized in the swirling disk. With improved mass spectrometric measurements evidence has been accumulating for small differences in isotopic composition in some elements between certain meteorites and those of the Earth and Moon. This area of study that searches for nucleosynthetic isotopic heterogeneity in the solar system is ongoing and is now providing a method for tracking the provenance of different portions of the disk.

However, oxygen and the noble gases are very different in this respect. Extreme isotopic variations have been found for these elements. The different oxygen and noble gas isotope ratios provide evidence of mixing between compositions of dust and those of volatile (gaseous) components. Some of this mixing may have arisen later when
the nebula cooled, possibly because large amounts of isotopically distinct material are thought to have arrived from the outer nebula in the form of water ice. There are also possibilities for generating some of the heterogeneity in oxygen by irradiation within the solar nebula itself. Samples of the solar wind obtained by the Genesis space mission suggest that the oxygen isotope composition of the solar nebula changed over time as the first stages of planet formation took place.

The terrestrial planets and asteroids are not just depleted in nebular gas relative to the Sun. They are also very depleted in moderately volatile elements (elements such as lead, potassium, and rubidium that condense at temperatures in the range 700–1350 K) (Figure 2.9 and 2.13). In chondritic meteorites, the degree of depletion becomes larger as an element’s condensation temperature decreases. It was long assumed that this is the result of the loss of gas from a hot nebula before it cooled. For example, by the time temperatures became cool enough for lead to condense, much of the lead had already accreted onto the Sun as a gas. However, it is clear that moderately volatile elements are depleted in chondrites at least in part because they contain CAIs and chondrules that lost volatiles by evaporation during heating events. The least depleted chondrites (CI carbonaceous chondrites) contain no CAIs or chondrules. Another mechanism for losing moderately volatile elements is planetary collisions. Energetic collisions between large bodies would have generated high temperatures and could have caused further loss of moderately volatile elements. For this reason, the terrestrial planets have compositions that differ from one another and also from chondritic meteorites. The Moon is highly depleted in moderately volatile elements (Figure 2.9) and is thought to be the product of such an energetic planetary collision.

4. NUCLEOSYNTHESIS AND SHORT-LIVED ISOTOPES

With the exception of hydrogen and helium, the elements we see in the solar system were mainly made by nuclear reactions in the interiors of other stars, a process called stellar nucleosynthesis. If one examines Figure 2.5, seven rather striking features stand out.

- The estimated abundances of the elements in the Sun and the solar nebula span a huge range of 13 orders of magnitude. For this reason they are most easily compared by plotting on a log scale of relative abundance such that the number of atoms of Si is $10^6$.
- Hydrogen and helium are by far the most abundant elements in the Sun, as they are elsewhere in the
universe. These two elements were made from subatomic particles shortly after the Big Bang.

- The abundances of the heavier elements generally decrease with increasing atomic number. This is because most of the elements are themselves formed from lighter elements by stellar nucleosynthesis.
- Iron is about 1000 times more abundant than its neighbors in the periodic table because the binding energy of an atomic nucleus is highest for iron. This provided enhanced stability for iron nuclei during nucleosynthesis.
- Lithium, beryllium, and boron are all relatively underabundant compared to other light elements because they are unstable in stellar interiors.
- A saw-toothed variability is superimposed on the overall trend reflecting the relatively high stability of even-numbered isotopes compared to odd-numbered ones.
- All the elements in the periodic table are present in the solar system except those with no long-lived or stable isotopes, viz. technetium (Tc), promethium (Pm), and the transuranic elements.

Elements lighter than iron can be made by nuclear fusion because the process of combining two nuclei to make a heavier nucleus releases energy for elements up to and including iron. Fusion provides the main source of energy in stars, and is activated when the central pressure exceeds a critical threshold, i.e. when a star reaches a certain mass. Larger stars exert more pressure on their cores such that fusion reactions proceed more quickly. Massive stars shine more brightly than small stars, and have shorter lifetimes as a result. When a star has converted all the hydrogen in its core to helium, nuclear reactions will cease if the star is small, or proceed to the next fusion cycle such as the conversion of helium to carbon if the star is sufficiently massive to drive this reaction. Lithium, beryllium, and boron are unstable at the temperatures and pressures of stellar interiors, and they are rapidly consumed. Small amounts of these elements are made by spallation reactions from heavier elements by irradiation in the outer portions of stars.

Nearly all nuclides heavier than iron have to be made by neutron irradiation because their synthesis via fusion would consume energy. Nuclear reactions in stars generate large numbers of neutrons, and these neutrons are readily absorbed by atoms since they are not repelled by the nuclei’s electrical charge. Neutron addition continues until an unstable isotope is made that decays to an isotope of another element, which then receives more neutrons until another unstable nuclide is made and so forth. These are \textit{s-process} isotopes (produced by the \textit{slow} but continuous production of neutrons in stars). However, some of these isotopes cannot be made simply by adding a neutron to a stable nuclide because there is no stable isotope with a suitable mass. Such nuclides are instead created with a very high flux of neutrons such that unstable nuclides produced by neutron irradiation receive additional neutrons before they have time to decay, jumping the gap to very heavy nuclides. These are \textit{r-process} isotopes (produced by a \textit{rapid} burst of neutrons). Such extremely high fluxes of neutrons are generated in supernova explosions and in particular in the cores of very large stars (e.g. 25 solar masses).

The composition of the Sun and solar system represents the cumulative \(~8~\text{billion year previous history of such stellar processes in this portion of the galaxy prior to collapse of the solar nebula (Figure 2.14). It is unknown how constant these processes were. However, the isotopes of some elements in meteorites provide evidence that stellar nucleosynthesis was still going on just prior to the formation of the solar nebula. In fact the formation of the solar system may have been triggered by material being ejected from a massive star as it was exploding, seeding the solar nebula with freshly synthesized nuclides.}

Chondrites show evidence that they once contained short-lived radioactive isotopes probably produced in massive stars shortly before the solar system formed. As already pointed out, most stable isotopes are present in the same ratios in the Earth, the Moon, Mars, and different groups of meteorites, which argues that material in the solar nebula was thoroughly mixed at an early stage. However, a few isotopes such as $^{26}\text{Mg}$ are heterogeneously distributed in chondrites. In most cases, these isotopes are

![Isotopes and time-scales](image-url)
the daughter products of short-lived isotopes. In other words, the excess $^{26}\text{Mg}$ comes from the radioactive decay of $^{26}\text{Al}$. Every atom of $^{26}\text{Al}$ decays to a daughter atom of $^{26}\text{Mg}$; therefore

$$
\frac{(26\text{Mg})_{\text{today}}}{(24\text{Mg})_{\text{original}}} = \frac{(26\text{Mg})_{\text{original}}}{(24\text{Mg})_{\text{original}}} + \frac{(26\text{Al})_{\text{original}}}{(24\text{Mg})_{\text{original}}} \quad (2.1)
$$

Because it is easier to measure these effects using isotopic ratios rather than absolute numbers of atoms we divide by another isotope of Mg:

$$
\frac{(26\text{Mg})}{(24\text{Mg})}_{\text{today}} = \frac{(26\text{Mg})_{\text{original}}}{(24\text{Mg})_{\text{original}}} + \frac{(26\text{Al})_{\text{original}}}{(24\text{Mg})_{\text{original}}} \quad (2.2)
$$

However, the $^{26}\text{Al}$ is no longer extant and so cannot be measured. For this reason we convert Eqn (2.2) to a form that includes a monitor of the amount of $^{26}\text{Al}$ that would have been present determined from the amount of Al today. Aluminum has only one stable nuclide $^{27}\text{Al}$. Hence, Eqn (2.2) becomes

$$
\frac{(26\text{Mg})}{(24\text{Mg})}_{\text{today}} = \frac{(26\text{Mg})_{\text{original}}}{(24\text{Mg})_{\text{original}}} + \left\{ \frac{(26\text{Al})_{\text{original}}}{(27\text{Al})_{\text{original}}} \times \frac{(27\text{Al})_{\text{today}}}{(24\text{Mg})_{\text{today}}} \right\} \quad (2.3)
$$

which represents the equation for a straight line (Figure 2.15). A plot of $^{26}\text{Mg}/^{24}\text{Mg}$ against $^{27}\text{Al}/^{24}\text{Mg}$ for a suite of cogenetic samples or minerals (such as different minerals in the same rock) will define a straight line the slope of which gives the $^{26}\text{Al}/^{27}\text{Al}$ at the time the object formed. This can be related in time to the start of the solar system with Soddy and Rutherford’s equation for radioactive decay:

$$
\frac{(26\text{Al})_{\text{original}}}{(27\text{Al})_{\text{BSSI}}} = \frac{(26\text{Al})}{(27\text{Al})}_{\text{BSSI}} \times e^{-\lambda t} \quad (2.4)
$$

in which BSSI is the bulk solar system initial ratio, $\lambda$ is the decay constant (or probability of decay in unit time), and $t$ is the time that elapsed since the start of the solar system. Using this method and the assumed initial ratio for $(26\text{Al}/27\text{Al})_{\text{BSSI}}$ of $\sim 5 \times 10^{-5}$ (Table 2.1) it has been possible to demonstrate that many chondrules formed 1–3 million years after CAIs.

Over the past 40 years scientists have found evidence that more than 15 short-lived isotopes existed early in the solar system. The main ones are listed in Table 2.1. However, evidence even exists for $^7\text{Be}$ with a half-life of just 53 days.

These short-lived isotopes can be broken down into three types on the basis of their origin in the solar nebula:

1. The Sun and the other stars in its cluster inherited a mixture of isotopes from their parent molecular cloud that built up over time from a range of stellar sources.
2. Some short-lived isotopes were probably injected into the Sun’s molecular cloud core or the solar nebula itself from at least one nearby star, possibly a supernova.
3. It is likely that a limited number of short-lived isotopes were also generated in the innermost regions of the solar nebula when material was bombarded with energetic particles from the Sun.

Determining the origin of a particular isotope and the timing of its production is often difficult. Isotopes with half-lives of less than $10^6$ years must have come from a source close to the solar nebula in order to have survived, while isotopes with longer half-lives may have come from further away as well. Irradiation in the solar nebula could have produced a variety of light isotopes but the relative importance of local production versus external sources is still unclear. Formation in the nebula appears to be a promising source for Be isotopes. However, if all the $^{26}\text{Al}$ had formed this way it seems likely that some of the other isotopes, especially $^{41}\text{Ca}$, would have been more abundant than they actually were. In fact, there is mounting evidence that many of the short-lived isotopes were uniformly distributed in the solar system, which is hard to explain if they formed in a localized region close to the Sun.

Some of the heavier short-lived isotopes that existed in the early solar system (e.g. $^{107}\text{Pd}$ and $^{129}\text{I}$) can only be produced in large amounts in a massive star. For example, a large flux of neutrons is required to produce $^{129}\text{I}$ and this is achievable during the enormously energetic death throws of a massive star undergoing a type II supernova explosion. Many of the isotopic ratios in Table 2.1 are similar, lying in
the range $10^{-6}$ to $10^{-4}$ for isotopes with half-lives of $0.7 \times 10^{6}$ to $30 \times 10^{6}$ years. This is as expected if all these isotopes were synthesized in roughly similar proportions just prior to the start of the solar system. Many of these isotopes have initial abundances similar to those that would be formed by an AGB star. However, models for AGB stars do not predict the amounts of $^{53}$Mn and $^{182}$Hf that once existed. In fact, $^{182}$Hf (half-life $= 8.9$ Myrs) requires a large flux of neutrons of the kind produced in the supernova explosion of a much larger star. It is possible that more than one kind of nucleosynthetic process gave rise to the short-lived isotopes in the early solar system. At present is seems likely that a nearby supernova was involved because the abundance of $^{60}$Fe, which has a fairly short half-life, is too high to be explained by alternative sources. Some isotopes that may have been present have yet to be found, including $^{126}$Sn with a half-life of 0.3 Myrs. This is an $r$-process isotope that should have been present in the early solar system if a supernova occurred nearby. While modeling these processes is complex, it appears that the supernova explosion of a 25-solar-mass star may explain the correct relative abundances of many of the short-lived isotopes, including $^{182}$Hf, provided that roughly 5 solar masses of material was left behind in the form of a supernova remnant or a black hole.

Supernovas are sufficiently energetic that they could tear apart a molecular cloud core rather than cause it to collapse. Shocks waves with a velocity of at least $20-45$ km/s are capable of triggering collapse but if the velocity exceeds $\sim 100$ km/s, a molecular cloud core will be shredded instead. If the supernova was sufficiently far away the shock wave would have slowed by the time it reached the molecular cloud core. However, the supernova cannot have been more than a few tens of parsecs away otherwise $^{41}$Ca (with a half-life of only 0.104 Myrs) would have decayed before it reached the solar nebula. The former presence of $^{41}$Ca in CAIs may provide the best constraint on the time between nucleosynthesis of the short-lived isotopes and their incorporation into the solar system. To do this, it will be necessary to ascertain the particular stellar source(s) that gave rise to these isotopes, so that the initial amount of $^{41}$Ca can be calculated.

5. EARLY STAGES OF PLANETARY GROWTH

Dust grains are a relatively minor constituent of protoplanetary disks, but they represent the starting point for the formation of rocky planets like Earth, and probably also

---

**TABLE 2.1** A List of Short-Lived Isotopes Thought to have Existed in the Early Solar System Based on the Distribution of their Decay Products in Meteorites.

<table>
<thead>
<tr>
<th>Radio-Nuclide</th>
<th>Half-Life (Myrs)</th>
<th>Ratio</th>
<th>Initial Ratio</th>
<th>Daughter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$^{10}$Be</td>
<td>1.5</td>
<td>$^{10}$Be/$^{9}$Be</td>
<td>$1 \times 10^{-3}$</td>
<td>$^{10}$B</td>
</tr>
<tr>
<td>$^{26}$Al</td>
<td>0.71</td>
<td>$^{26}$Al/$^{27}$Al</td>
<td>$5 \times 10^{-5}$</td>
<td>$^{26}$Mg</td>
</tr>
<tr>
<td>$^{36}$Cl</td>
<td>0.30</td>
<td>$^{36}$Cl/$^{35}$Cl</td>
<td>$2 \times 10^{-5}$</td>
<td>$^{36}$S, $^{36}$Ar</td>
</tr>
<tr>
<td>$^{41}$Ca</td>
<td>0.10</td>
<td>$^{41}$Ca/$^{40}$Ca</td>
<td>$1 \times 10^{-8}$</td>
<td>$^{41}$K</td>
</tr>
<tr>
<td>$^{53}$Mn</td>
<td>3.7</td>
<td>$^{53}$Mn/$^{55}$Mn</td>
<td>$6 \times 10^{-6}$</td>
<td>$^{53}$Cr</td>
</tr>
<tr>
<td>$^{60}$Fe</td>
<td>2.6</td>
<td>$^{60}$Fe/$^{54}$Fe</td>
<td>$1 \times 10^{-6}$</td>
<td>$^{60}$Ni</td>
</tr>
<tr>
<td>$^{92}$Nb</td>
<td>35</td>
<td>$^{92}$Nb/$^{93}$Nb</td>
<td>$3 \times 10^{-5}$</td>
<td>$^{92}$Zr</td>
</tr>
<tr>
<td>$^{107}$Pd</td>
<td>6.5</td>
<td>$^{107}$Pd/$^{110}$Pd</td>
<td>$9 \times 10^{-5}$</td>
<td>$^{107}$Ag</td>
</tr>
<tr>
<td>$^{129}$I</td>
<td>15.7</td>
<td>$^{129}$I/$^{127}$I</td>
<td>$1 \times 10^{-4}$</td>
<td>$^{129}$Xe</td>
</tr>
<tr>
<td>$^{135}$Cs</td>
<td>2.3</td>
<td>$^{135}$Cs/$^{133}$Cs</td>
<td>$5 \times 10^{-4}$</td>
<td>$^{133}$Ba</td>
</tr>
<tr>
<td>$^{146}$Sm</td>
<td>68</td>
<td>$^{146}$Sm/$^{144}$Sm</td>
<td>$0.008$</td>
<td>$^{142}$Nd</td>
</tr>
<tr>
<td>$^{182}$Hf</td>
<td>8.9</td>
<td>$^{182}$Hf/$^{180}$Hf</td>
<td>$1 \times 10^{-4}$</td>
<td>$^{180}$W</td>
</tr>
<tr>
<td>$^{205}$Pb</td>
<td>17</td>
<td>$^{205}$Pb/$^{204}$Pb</td>
<td>$1 \times 10^{-3}$</td>
<td>$^{205}$Tl</td>
</tr>
<tr>
<td>$^{244}$Pu</td>
<td>80</td>
<td>$^{244}$Pu/$^{238}$U</td>
<td>$7 \times 10^{-3}$</td>
<td>$^{131, 132, 134, 136}$Xe</td>
</tr>
<tr>
<td>$^{247}$Cm</td>
<td>16</td>
<td>$^{247}$Cm/$^{235}$U</td>
<td>$2 \times 10^{-3}$</td>
<td>$^{235}$U</td>
</tr>
</tbody>
</table>
gas-rich planets like Jupiter. These grains are small, typically 1 \( \mu \text{m} \) in diameter or less. In a microgravity environment, electrostatic forces dominate interactions between such grains, and these forces help grains stick together when they collide. Charge transfer during grain collisions can lead to the formation of grain dipoles that align with one another, forming aggregates up to several centimeters in size. Freshly deposited frost surfaces also make grains stickier, and increase the ability of grain aggregates to hold together during subsequent collisions.

Laboratory experiments show that low-velocity collisions between grains tend to result in sticking, while faster collisions often cause grains to rebound. Irregularly shaped micron-sized grains often stick to one another at collision speeds of up to tens of meters per second. Fluffy aggregates may stick more readily than compact solids as some of the energy of impact goes into compaction. However, the primary components of chondritic meteorites are compact chondrules so further compaction cannot have played a big role in the formation of their parent bodies. In general, sticking forces scale with the surface area of an object, while collisional energy scales with mass and hence volume. As a result, growth becomes more difficult, and break up becomes more likely, as aggregates become larger. It is possible that early growth in the solar nebula took place mainly as the result of large objects sweeping up smaller ones. This idea is supported by recent experiments that have found that small dust aggregates tend to embed themselves in larger ones if they collide at speeds above about 10 m/s.

Dust grains, grain aggregates, and chondrules would have been closely coupled to the motion of gas in the solar nebula. The smallest particles were mainly affected by Brownian motion—collisions with individual gas molecules, which caused the particles to move with respect to one another, leading to collisions. Particles also settled slowly toward the disk’s midplane due to the vertical component of the Sun’s gravitational field. Settling was opposed by gas drag so that each particle fell at its terminal velocity:

\[
v_z = -\left(\frac{\rho}{\rho_{\text{gas}}}\right) \left(\frac{v_{\text{kep}}}{c_s}\right) \left(\frac{r}{a}\right) v_{\text{kep}}
\]

(2.5)

where \( r \) and \( \rho \) are the radius and density of the particle, \( \rho_{\text{gas}} \) is the gas density, \( a \) is the orbital distance from the Sun, \( z \) is the height above the disk midplane, and \( c_s \) is the sound speed in the gas. Here \( v_{\text{kep}} \) is the speed of a solid body moving on a circular orbit, called the \textit{Keplerian velocity}:

\[
v_{\text{kep}} = \sqrt{\frac{GM_{\text{sun}}}{a}}
\]

(2.6)

where \( M_{\text{sun}} \) is the mass of the Sun. Large particles fell faster than small ones, sweeping up material as they went, increasing their vertical speed further. Calculations show that micron-sized particles would grow and reach the midplane in about \( 10^3 - 10^4 \) orbital periods if these were the only processes operating.

If the gas was turbulent, particles would have become coupled to turbulent eddies due to gas drag. Particles of a given size were coupled most strongly to eddies whose turnover (rotation) time was similar to the particle’s stopping time, given by

\[
t_s = \frac{\rho r}{\rho_{\text{gas}} c_s}
\]

(2.7)

Meter-sized particles would have coupled to the largest eddies, with turnover times comparable to the orbital period \( P \). In a strongly turbulent nebula, meter-sized particles would have collided with one another and with smaller particles at high speeds, typically tens of meters per second.

Gas pressure in the nebula generally decreased with distance from the Sun. This means gas orbited the Sun more slowly than solid bodies, which moved at the Keplerian velocity. Large solid bodies thus experienced a headwind of up to 100 m/s. The resulting gas drag removed angular momentum from solid bodies, causing them to undergo \textit{radial drift} toward the Sun. Small particles with \( t_s \ll P \) drifted slowly at terminal velocity. Very large objects with \( t_s \gg P \) were only weakly affected by gas drag and also drifted slowly. Drift rates were highest for meter-sized bodies with \( t_s = P \) (see Figure 2.16), and these drifted inward at rates of 1 AU every few hundred years. Rapid inward drift meant that these bodies collided with smaller particles at high speeds. Rapid drift also meant that meter-sized objects had very short lifetimes, and many were probably lost when they reached the hot innermost regions of the nebula and vaporized.

The short drift lifetimes and high collision speeds experienced by meter-sized particles have led some researchers to conclude that particle growth stalled at this
size because particles were destroyed as fast as they formed. This is often referred to as the *meter-sized barrier*. This remains an open question, however, due to a shortage of experimental data regarding the physics of collisions in a microgravity environment, and uncertainty about the level of turbulence in the solar nebula. It is possible that a small population of lucky objects successfully passed through the meter-sized barrier because they never experienced a destructive, high-speed collision.

Bodies larger than 1 km generally took a long time to drift inward due to gas drag. These objects were also large enough to have appreciable gravitational fields, making them better able to hold on to fragments generated in collisions. For these reasons, growth became easier once bodies became this large. Much effort has been devoted to seeing whether kilometer-sized bodies could have formed directly, avoiding the difficulties associated with the meter-sized barrier. *Gravitational instability* (GI) offers a possible way to do this. If the level of turbulence in the nebula was very low, solid particles would have settled close to the nebula midplane increasing their local concentration. If enough particles became concentrated in one place, their combined gravitational attraction would render the configuration unstable, allowing the region to become gravitationally bound and collapse. If the particles were then able to contract enough to form a single solid body, the resulting object would be roughly 1–10 km in radius. Such an object is called a *planetesimal*.

GI faces severe obstacles, however. As solid particles accumulated near the nebula midplane, they would have begun to drag gas around the Sun at Keplerian speeds, while gas above and below the midplane continued to travel at sub-Keplerian speeds. The velocity difference between the layers generated turbulence, puffing up the particle layer until a balance between vertical sedimentation and turbulence was reached. This balance may have prevented particle concentrations from becoming high enough for GI to occur. Calculations suggest that the solid-to-gas ratio in a vertical column of nebula material had to become roughly unity before GI would take place. This means that the concentration of solid material had to become enhanced by one to two orders of magnitude compared to the nebula as a whole. If a region of the disk did start to undergo GI, it would only contract to form a planetesimal if the relative velocities of the particles in that region became low enough. Turbulence and radial drift both lead to large relative velocities between particles and may have rendered GI ineffective.

The presence of turbulence may not have been entirely detrimental to growth. Numerical simulations show that chondrule-size particles would be strongly concentrated in stagnant regions in a turbulent nebula, a process called *turbulent concentration*. Larger, roughly meter-sized particles would have been concentrated by a second process called the streaming instability. These highly mobile particles tended to accumulate at temporary, high-pressure zones in the turbulent gas. As particles accumulated they began to shield one another from the headwind, slowing their inward drift and allowing more particles to accumulate at the same location. Each of these mechanisms provides a possible route to planetesimal formation in turbulent protoplanetary disks.

The difficulties associated with both the meter-sized barrier and GI mean that the question of how planetesimals formed remains open for now. However, the fact that roughly half of young stars have debris disks of dust thought to come from asteroids and comets implies that growth of large solid bodies occurs in many protoplanetary disks, even if the mechanism remains obscure.

### 6. FORMATION OF TERRESTRIAL PLANETS

The growth of bodies beyond 1 km in size is better understood than planetesimal formation itself. Gravitational interactions and collisions between pairs of planetesimals dominate the evolution from this point onward. A key factor in determining the rate of growth is *gravitational focusing*. The probability that two planetesimals will collide during a close approach depends on their cross-sectional area multiplied by a *gravitational focusing factor* $F_g$:

$$F_g = 1 + \frac{v_{rel}^2}{v_{esc}^2}$$

where $v_{rel}$ is the planetesimals’ relative velocity, and $v_{esc}$ is the escape velocity from a planetesimal, given by

$$v_{esc} = \sqrt{\frac{2GM}{r}}$$

where $M$ and $r$ are the planetesimal’s mass and radius, respectively. When planetesimals pass each other slowly, there is time for their mutual gravitational attraction to focus their trajectories toward each other, so $F_g$ is large, and the chance of a collision is high. Fast-moving bodies typically do not collide unless they are traveling directly toward each other because $F_g \approx 1$ in this case. The relative velocities of planetesimals depend on their orbits about the Sun. Objects with similar orbits are the most likely to collide with each other. In particular, planetesimals moving on nearly circular, coplanar orbits have high collision probabilities while ones with highly *inclined*, *eccentric* (elliptical) orbits do not.

Most close encounters between planetesimals did not lead to a collision, but bodies often passed close enough for their mutual gravitational tug to change their orbits. Statistical studies show that after many such close encounters, high-mass bodies tend to acquire circular, coplanar orbits, while low-mass bodies are perturbed onto eccentric, inclined orbits.
This is called dynamical friction, and is analogous to the equipartition of kinetic energy between molecules in a gas. Dynamical friction means that on average, the largest bodies in a particular region experience the strongest gravitational focusing and therefore they grow the fastest (Figure 2.17). This state of affairs is called runaway growth for obvious reasons. Most planetesimals remained small, while a few objects, called planetary embryos, grew much larger.

Runaway growth continued as long as interactions between planetesimals determined their orbital distribution. However, once embryos became large enough, gravitational perturbations from these objects came to dominate the motion of the smaller planetesimals. This transition took place when

$$M_{\text{emb}} \Sigma_{\text{emb}} > M_{\text{plan}} \Sigma_{\text{plan}}$$

where $M_{\text{emb}}$ and $M_{\text{plan}}$ are the mass of a typical embryo and planetesimal, respectively, and $\Sigma_{\text{emb}}$ and $\Sigma_{\text{plan}}$ are the surface densities of the embryos and planetesimals.

The evolution now entered a new phase called oligarchic growth. The relative velocities of planetesimals were determined by a balance between perturbations from nearby embryos and damping due to gas drag. Embryos continued to grow faster than planetesimals, but growth was no longer unrestrained. Large embryos stirred up nearby planetesimals more than small embryos did, weakening gravitational focusing and slowing growth. As a result, neighboring embryos tended to grow at similar rates. Embryos spaced themselves apart at regular radial intervals, with each one staking out an annular region of influence in the nebula called a feeding zone.

As embryos became larger, they perturbed planetesimals onto highly inclined and eccentric orbits. The planetesimals began to collide with one another at high speeds, causing fragmentation and breakup. A huge number of subkilometer-sized collision fragments were generated, together with a second generation of fine dust particles. Gas drag operates efficiently on small fragments, so their orbits rapidly became almost circular and coplanar. As a result, many fragments were quickly swept up by embryos, increasing the embryos’ growth rates still further.

Numerical calculations show that embryo feeding zones were typically about 10 Hill radii in width, where the Hill radius of an embryo with mass $M$ and orbital radius $a$ is given by

$$r_h = a \left( \frac{M}{3M_{\text{sun}}} \right)^{1/3}$$

(2.10)

If an embryo were to accrete all the solid material in its feeding zone it would stop growing when its mass reached a value called the isolation mass, given by

$$M_{\text{iso}} \approx \left( \frac{8\pi^3}{3M_{\text{sun}}} \right) \frac{\Sigma_{\text{gas}} a^2}{\Sigma_{\text{solid}} b^2} v_{\text{kep}}^2$$

(2.11)

where $\Sigma$ is the surface (column) density of solid material in that region of the disk, and $b \approx 10$ is the width of a feeding zone in Hill radii. The surface density in the Sun’s protoplanetary nebula is not known precisely, but for plausible values, the isolation masses would have been about 0.1 Earth masses at 1 AU, and around 10 Earth masses in the outer solar system. Calculations suggest that bodies approached their isolation mass in the inner solar system roughly $10^5$ years after planetesimals first appeared in large numbers. Growth was slower in the outer solar system, but bodies were probably nearing their isolation mass at 5 AU after $10^6$ years.

Large embryos significantly perturbed nearby gas in the nebula forming spiral waves. Gas passing through these waves had a higher density than that in the surrounding region. Gravitational interactions between an embryo and its spiral waves transferred angular momentum between them. For conditions likely to exist in the solar nebula, the net result was that each embryo lost angular momentum and migrated inward toward the Sun. This is called type-I migration. In an isothermal (uniform temperature) disk, the migration rate is proportional to an embryo’s mass $M$ and the local surface density of gas $\Sigma_{\text{gas}}$ and is given by

$$\frac{d a}{d t} \approx -4 \left( \frac{M}{M_{\text{sun}}} \right) \left( \frac{\Sigma_{\text{gas}} a^2}{M_{\text{sun}}} \right) \left( \frac{v_{\text{kep}}}{c_s} \right)^2 v_{\text{kep}}$$

(2.12)

where $c_s$ is the sound speed in the gas and $v_{\text{kep}}$ is the orbital velocity of a body moving on a circular, Keplerian orbit. Type-I migration became important once embryos grew to

---

**FIGURE 2.17** Runaway growth of a few large planetesimals takes place due to a combination of dynamical friction (which gives large planetesimals circular and coplanar orbits), and gravitational focusing (which increases the chance of a collision between bodies moving on similar orbits).
about 0.1 Earth masses. Migration rates can be uncomfortably fast, with a 10 Earth mass body at 5 AU migrating into the Sun in $10^5$ years in a minimum-mass nebula. It is possible that many objects migrated all the way into the Sun and were lost in this way, and the question of how other bodies survived migration is one of the great unresolved questions of planet formation at present.

Type-I migration rates are modified when radiative transfer within the disk is taken into account. Migration is especially likely to change in regions where there is a discontinuity in the disk such as the ice line or the edge of a dead zone. At these locations, inward migration can slow substantially or even change direction. As a result, there may be particular locations in a disk that are preferred for planet formation since embryos at these points do not migrate, or survive for long enough to outlive the disk.

Oligarchic growth in the inner solar system ended when embryos had swept up roughly half of the solid material. However, these embryos were still an order of magnitude less massive than Earth. Further collisions were necessary to form planets the size of Earth and Venus. With the removal of most of the planetesimals, dynamical friction weakened. As a result, interactions between embryos caused their orbits to become more inclined and eccentric. The embryos’ gravitational focusing factors became small and this greatly reduced the collision rate. As a result, the last stage of planet formation was prolonged, and the Earth may have taken 100 million years to finish growing.

Embryos underwent numerous close encounters with one another before colliding. Each encounter changed an embryo’s orbit, with the result that embryos moved considerable distances radially in the nebula. Numerical calculations show that the orbital evolution must have been highly chaotic (Figure 2.18). As a result, it is impossible to predict the precise characteristics of a planetary system based on observations of typical protoplanetary disks. Other stars with nebulae similar to the Sun may have formed terrestrial planets that are very different from those in the solar system.

The radial motions of embryos partially erased any chemical gradients that existed in the nebula during the early stages of planet formation. Mixing cannot have been complete, however, since Mars and Earth have distinct compositions. Mars is richer in the more volatile rock-forming elements, and the two planets have distinct oxygen isotope mixtures. Unfortunately, we have no confirmed samples of Mercury and Venus, so we know little about their composition. Mercury is known to have an unexpectedly high density, suggesting it has a large iron-rich core and a small mantle. This probably does not reflect compositional differences in the solar nebula since there is no known reason why iron-rich materials would preferentially form closer to the Sun than silicate materials. A more likely explanation is that Mercury suffered a near-catastrophic impact after it had differentiated, and this stripped away much of the silicate mantle. Mercury’s location close to the Sun made it especially vulnerable in this respect since orbital velocities and hence impact speeds are highest close to the Sun.

Earth and Venus are probably composites of 10 or more embryos so their chemical and isotopic compositions represent averages over a fairly large region of the inner solar system. Mars and Mercury are sufficiently small that they may be individual embryos that did not grow much beyond the oligarchic growth stage. It is currently a mystery why Earth and Venus continued to grow while Mars did not. It may be that Mars formed in a low-density region of the nebula or that all other embryos were removed from that region without colliding with Mars.

The Grand Tack model offers one possible explanation for why there was relatively little solid material in the region that gave rise to Mars, and even less in the asteroid belt. This model is based on numerical simulations that show that Jupiter would have migrated inward through the solar nebula until Saturn formed, at which point Jupiter would have migrated outward. If Jupiter migrated inward to about 1.5 AU from the Sun, and then moved outward again to its current location, the planet’s gravity would have pulled most planetesimals and planetary embryos out of the region that now contains Mars and the asteroid belt, scattering these objects into the Sun or out of the solar system. Computer simulations show the Grand Tack model can explain many observed features of the solar system as a result, but the model remains unconfirmed at the present time.
As embryos grew larger, their temperatures increased due to kinetic energy released during impacts and the decay of radioactive isotopes in their interiors. Short-lived isotopes such as $^{26}$Al and $^{60}$Fe, with half-lives of 0.7 and 2.6 Myrs, respectively (Table 2.1), were particularly powerful heat sources early in the solar system. Bodies more than a few kilometers in radius would have melted if they formed within the first 2 million years, when the short-lived isotopes were still abundant. Embryos that melted also differentiated, with iron and siderophile elements sinking to the center to form a core, while lighter silicates formed a mantle closer to the surface.

The abundances of the most highly siderophile elements (such as platinum and osmium) in Earth’s mantle are higher than one would expect to find after the planet differentiated since most siderophile material should have been extracted into the core. The amount of the platinum in the core is sufficient to cover Earth’s surface to a depth of about a meter. However, even that which is residual in Earth’s mantle and which provides our platinum and gold jewelry, is much more than expected unless it was added after core formation had ceased. The most likely explanation for these high abundances therefore is that Earth continued to acquire some material after its core and mantle had finished separating. This late veneer may amount to almost 1% of the total mass of the planet.

The degree to which this late veneer also provided Earth with some of its inventory of volatiles, namely, water, carbon, nitrogen, sulfur, and the noble gases, is uncertain; some argue that the major portion was accreted earlier. Even the amount of water that Earth contains is debated. Earth’s oceans contain about 0.03% of the planet’s total mass. At least as much water exists in the mantle and some think there could be an order of magnitude more. The present amounts of water and other volatiles, even if they were better quantified, may not reflect the original situation. Earth may have also suffered removal of volatiles as it grew from energetic collisions. There would also have been some dissipation of dissociated hydrogen to space. Finally, reactions with iron could have led to segregation to the core of hydrogen, carbon, nitrogen, and sulfur, just like the platinum group elements.

Temperatures at 1 AU are currently too high for water ice to condense, and this was probably also true for most of the history of the solar nebula (pressures were always too low for liquid water to condense). As a result, Earth probably received most of its water as the result of collisions with other embryos or planetesimals that contained water ice or hydrated minerals in their interiors. Planetesimals similar to modern comets almost certainly delivered some water to Earth. However, a typical comet has a probability of only about one in a million of colliding with Earth, so it is unlikely that comets provided the bulk of the planet’s water. The deuterium to hydrogen (D/H) ratio seen in most comets is twice that of Earth’s oceans, which suggests these comets supplied at most about 10% of Earth’s water. To date, one comet has been observed with a D/H ratio similar to Earth, so it is possible that Earth acquired a substantial amount of water from a subpopulation of comets.

Planetesimals from the asteroid belt are another possible source of water. Carbonaceous chondrites are especially promising since they contain up to 10% water by mass in the form of hydrated silicates, and this water would be released upon impact with the Earth. Calculations suggest that if the early asteroid belt was several orders of magnitude more massive than today, it could have supplied the bulk of Earth’s water. From the current meteorite record it seems unlikely that this water could have been delivered after core formation like the late veneer of highly siderophile elements. There are a number of reasons for this but a compelling case comes from the fact that carbonaceous chondrites and Earth’s mantle have different osmium isotope ratios. As a result, the delivery of water to Earth and its acquisition of a late veneer were separate processes that occurred at different times in its history.

The origin of Earth’s atmospheric constituents is also somewhat uncertain. When the solar nebula was still present, planetary embryos probably had thick atmospheres mostly composed of hydrogen and helium captured from the nebula. Most of this atmosphere was lost subsequently by hydrodynamic escape as hydrogen atoms were accelerated to escape velocity by ultraviolet radiation from the Sun, dragging other gases along with them. Much of Earth’s current atmosphere was probably outgassed from the mantle at a later stage. Some noble gases currently escaping from Earth’s interior are similar to those found in the Sun, which suggests they may have been captured into Earth’s mantle from the nebula or were trapped in bodies that later collided with Earth. Most of the xenon produced by radioactive decay of $^{244}$Pu and $^{129}$I (Table 2.1) has been lost, which implies that Earth’s atmosphere was still being eroded 100 Myr after the start of the solar system, possibly by impacts.

Radioactive isotopes can be used to place constraints on the timing of planet formation. The hafnium–tungsten system is particularly useful in this respect since the parent nuclide $^{182}$Hf is lithophile (tending to reside in silicate mantles) while the daughter nuclide $^{182}$W is siderophile (tending to combine with iron during core formation) (Figure 2.19). Isotopic data can be used in a variety of ways to define a timescale for planetary accretion. The simplest method uses a model age calculation, which corresponds to the calculated time when an object or sample would have needed to form from a simple average solar system reservoir, as represented by chondrites, in order generate its...
isotopic composition. For the \(^{182}\text{Hf}^{182}\text{W}\) system this time is given as

\[
    t_{\mathrm{CHUR}} = \frac{1}{\lambda} \ln \left[ \frac{\left(\frac{^{182}\text{Hf}}{^{180}\text{Hf}}\right)_{\mathrm{BSSI}}}{\left(\frac{^{182}\text{W}}{^{184}\text{W}}\right)_{\mathrm{SAMPLE}} - \left(\frac{^{182}\text{W}}{^{184}\text{W}}\right)_{\mathrm{CHONDRITES}}} \right] 
\]

where \(t_{\mathrm{CHUR}}\) is the time of separation from a CHondritic Uniform Reservoir, \(\lambda = (\ln 2/\text{half-life})\) is the decay constant for \(^{182}\text{Hf}\) (0.078 per million years), and \(\left(\frac{^{182}\text{Hf}}{^{180}\text{Hf}}\right)_{\mathrm{BSSI}}\) is the BSSI ratio of \(^{182}\text{Hf}\) to \(^{180}\text{Hf}\). Tungsten-182 excesses have been found in Earth, Mars, and the HED meteorites, which are thought to come from asteroid Vesta, indicating that all these bodies differentiated while some \(^{182}\text{Hf}\) was still present. Iron meteorites, which come from the cores of differentiated planetesimals, have low Hf/W ratios and are deficient in \(^{182}\text{W}\). This means these planetesimals must have formed at a very early stage before most of the \(^{182}\text{Hf}\) had decayed. New, very precise \(^{182}\text{Hf}^{182}\text{W}\) chronometry has shown that some of these objects formed within the first 2 million years of the solar system (Figure 2.8).

New modeling of the latest \(^{182}\text{Hf}^{182}\text{W}\) data for Martian meteorites also provides evidence that Mars grew and started differentiating within about 1 million years of the start of the solar system. This short timescale is consistent with runaway growth described above. So far, isotopic data for other silicate objects has not been so readily explicable in terms of very rapid growth. However, asteroid Vesta has been determined to have certain formed within about 3 million years of the start of the solar system (Figure 2.8).

The existence of meteorites from differentiated asteroids suggests that core formation began early and this is confirmed by \(^{182}\text{Hf}^{182}\text{W}\) chronometry. Therefore, most planetary embryos would have been differentiated when they collided with one another. Although Mars grew extremely rapidly, Earth does not appear to have reached its current size until the giant impact that was associated with the formation of the Moon (see Section 8). \(^{182}\text{Hf}^{182}\text{W}\) chronometry for lunar samples shows that this took place more than 30 Myrs after the start of the solar system. There is other evidence that this could have been as late as 100 Myrs and it has long been recognized that the formation of the Moon probably happened near the end of Earth’s accretion, and this is consistent with the results of Moon-forming impact simulations. This is also consistent with the W isotopic composition of the silicate Earth itself (Figure 2.20). This shows that the Earth accreted at least half of its mass within the first 30 Myrs of the solar system. However, the data are fully consistent with the final stage of accretion being around the time of the Moon-forming impact. Because the Earth accreted over a protracted period rather than in a single event it is the simplest to model the W isotope data in terms of an exponentially decreasing rate of growth (Figure 2.20).

\[
    F = 1 - e^{-\left(1/\tau\right) \times t} \quad (2.14)
\]

where \(F\) is the mass fraction of the Earth that has accumulated, \(\tau\) is the mean life for accretion in Myrs (Figure 2.20), and \(t\) is time in Myrs. This is consistent with the kinds of curves produced by the late George Wetherill who modeled the growth of the terrestrial planets using...
has a basaltic crust. The HED meteorites, which probably come from Vesta, show this crust formed only a few million years after the solar system, according to several isotopic systems. The survival of Vesta’s crust suggests the impact rate in the belt has never been much higher than today since the crust formed. For these reasons, it is thought that most of the asteroid belt’s original mass was removed at a very early stage by a dynamical process rather than by collisional erosion.

The asteroid belt currently contains a number of orbital resonances associated with the giant planets. Resonances occur when either the orbital period or precession period of an asteroid has a simple ratio with the corresponding period for one of the planets. Many resonances induce large changes in orbital eccentricity, causing asteroids to fall into the Sun, or to come close to Jupiter, leading to close encounters and ejection from the solar system. For this reason, there are very few asteroids that orbit the Sun twice every time Jupiter orbits the Sun once, for example. When the nebular gas was still present, small asteroids moving on eccentric orbits would have drifted inward rapidly due to gas drag. After the giant planets had formed, a combination of resonances and gas drag may have transferred most objects smaller than a few hundred kilometers from the asteroid belt into the terrestrial-planet region. Larger planetary embryos would not have drifted very far. However, once oligarchic growth ceased, embryos began to gravitationally scatter one another across the belt. Numerical simulations show that most or all of these bodies would eventually enter a resonance and be removed, leaving an asteroid belt greatly depleted in mass and containing no objects bigger than Ceres. The timescale for the depletion of the belt depends sensitively on the orbital eccentricities of the giant planets at the time, which are poorly known. The belt may have been cleared in only a few million years, but it may have required as much as several hundred million years if the giant planets had nearly circular orbits. Finally, as we saw earlier, the asteroid belt may have lost much of its original mass due to the migration of Jupiter while the solar nebula was still present.

The albedos and spectral features of asteroids vary widely from one body to another, but clear trends are apparent as one moves across the asteroid belt. S-type asteroids, which generally lie in the inner asteroid belt, appear to be more thermally processed than the C-type asteroids that dominate the middle belt. These may include the parent bodies of ordinary and carbonaceous chondrites, respectively. C-types in turn seem more processed than the P-type asteroids that mostly lie in the outer belt. These differences may reflect differences in the composition of solid materials in different parts of the nebula, or differences in the time at which asteroids formed. Ordinary and enstatite chondrites, which probably come from the inner asteroid belt, tend to be dry, while carbonaceous chondrites from the middle and outer belt

7. THE ASTEROID BELT

The asteroid belt currently contains only enough material to make a planet 2000 times less massive than Earth, even though the spatial extent of the belt is huge. It seems likely that this region once contained much more mass than it does today. A smooth interpolation of the amount of solid material needed to form the inner planets and the gas giants would place about 2 Earth masses in the asteroid belt. Even if most of this mass was lost at an early stage, the surface density of solid material must have been at least 100 times higher than it is today in order to grow bodies the size of Ceres and Vesta, which probably formed from different mechanisms (Figure 2.8).

**FIGURE 2.20** The mean life of accretion of the Earth ($\tau$) is the inverse of the time constant for exponentially decreasing oligarchic growth from stochastic collisions between planetary embryos and planets. The growth curves corresponding to several such mean lives are shown including the one that most closely matches the calculation made by the late George Wetherill based on Monte Carlo simulations. The mean life determined from tungsten isotopes (Figure 2.8) is in excellent agreement with Wetherill’s predictions.

Monte Carlo simulations. The W isotope data are consistent with a mean life of between 10 and 15 Myrs depending on the exact parameters used. This is fully consistent with the timescales proposed by Wetherill. From these protracted timescales it is clear that Earth took much longer to approach its current size than Mars or Vesta, which probably formed from different mechanisms (Figure 2.8).

Several regions of the asteroid belt contain clusters of asteroids with similar orbits and similar spectral features, suggesting they are made of the same material. These clusters are fragments from the collisional breakup of larger asteroids. There are relatively few of these asteroid families, which implies that catastrophic collisions are quite rare. This suggests the asteroid belt has contained relatively little mass for most of its history. The spectrum of asteroid Vesta, located 2.4 AU from the Sun, shows that it has a basaltic crust. The HED meteorites, which probably
contain up to 10% water by mass in the form of hydrated minerals. This suggests that temperatures were cold enough in the outer asteroid belt for water ice to form and become incorporated into asteroids where it reacted with dry rock. Temperatures were apparently too high for water ice to condense in the inner asteroid belt. It is possible that some of the objects currently in the asteroid belt formed elsewhere. For example, it has been proposed that many of the parent bodies of the iron meteorites, and possibly Vesta, formed in the terrestrial-planet region and were later gravitationally scattered outward to their current orbits.

Iron meteorites from the cores of melted asteroids are common, whereas meteorites from the mantles of these asteroids are rarely seen. This suggests that a substantial amount of collisional erosion took place at an early stage, with only the strong, iron-rich cores of many bodies surviving. A number of other meteorites also show signs that their parent asteroids experienced violent collisions early in their history. Chondrites presumably formed somewhat later than the differentiated asteroids, when the main radioactive heat sources had mostly decayed. Chondrites are mostly composed of chondrules, which typically formed 1—3 Myr after CAIs. Chondrite parent bodies cannot be older than the youngest chondrules they contain, so they must have formed several Myr after the start of the solar system. For this reason, it appears that the early stages of planet formation were prolonged in the asteroid belt. While chondrites have experienced some degree of thermal processing, their late formation meant that their parent bodies never grew hot enough to melt, which has allowed chondrules, CAIs, and matrix grains to survive.

8. GROWTH OF GAS AND ICE GIANT PLANETS

Jupiter and Saturn are mostly composed of hydrogen and helium. These elements do not form solids or liquids at temperatures and pressures found in protoplanetary disks, so they must have been gravitationally captured from the gaseous component of the solar nebula. Observations of young stars indicate that protoplanetary disks survive for only a few million years, and this sets an upper limit for the amount of time required to form giant planets. Uranus and Neptune also contain significant amounts of hydrogen and helium (somewhere in the range 3—25%), and so they probably also formed quickly, before the solar nebula dispersed.

Jupiter and Saturn also contain elements heavier than helium and they are enriched in these elements compared to the Sun. The gravitational field of Saturn strongly suggests it has a core of dense material at its center, containing roughly 1/5 of the planet’s total mass. Jupiter may also have a dense core containing a few Earth masses of material. The interior structure of Jupiter remains quite uncertain since we lack adequate equations of state for the behavior of hydrogen at the very high pressures found in the planet’s interior. The upper atmospheres of both planets are enriched in elements such as carbon, nitrogen, sulfur, and argon, compared to the Sun. It is thought likely that these enrichments extend deep into the planets’ interiors but this remains uncertain.

Giant planets may form directly by the contraction and collapse of gravitationally unstable regions of a protoplanetary disk. This disk instability is analogous to the gravitational instabilities that may have formed planetesimals, but instead the instability takes place in nebula gas rather than the solid component of the disk. Instabilities will occur if the Toomre stability criterion $Q$ becomes close to or lower than 1, where

$$Q = \frac{M_{\text{sup}} c_s}{2 \pi G \Sigma \nu_{\text{kep}}}$$

(2.15)

where $\nu_{\text{kep}}$ is the Keplerian velocity, $c_s$ is the sound speed, and $\Sigma$ is the local surface density of gas in the disk. Gas in an unstable region quickly becomes much denser than the surrounding material. Disk instability requires high surface densities and low sound speeds (cold gas), so it is most likely to occur in the outer regions of a massive protoplanetary disk. Numerical calculations suggest instabilities will occur beyond about 5 AU in a nebula a few times more massive than the MMSN. What happens to an unstable region depends on how quickly the gas cools as it contracts, and this is the subject of much debate. If the gas remains hot, the dense regions will quickly become sheared out and destroyed by the differential rotation of the disk. If cooling is efficient, simulations show that gravitationally bound clumps will form in a few hundred years, and these may ultimately contract to form giant planets. Initially, such planets would be homogeneous and have the same composition as the nebula. Their structure and composition may change subsequently due to gravitational settling of heavier elements to the center and capture of rocky or icy bodies such as comets.

The evidence for dense cores at the centers of Jupiter and Saturn suggests to many scientists that giant planets form by core accretion rather than disk instability. In this model, the early stages of giant-planet formation mirror the growth of rocky planets, beginning with the formation of planetesimals, followed by runaway and oligarchic growth. However, planetary embryos would have grown larger in the outer solar system for two reasons. First, feeding zones here are larger because the Sun’s gravity is weaker, so each embryo gravitationally holds sway over a larger region of the nebula. Second, temperatures here were cold enough for volatile materials such as tars, water ice, and other ices to condense, so more solid material was available to build large embryos.
In the outer solar system, bodies roughly 10 times more massive than Earth would have formed via oligarchic growth in a million years, provided the disk was a few times more massive than the MMSN. Bodies that grew larger than Mars would have captured substantial atmospheres of gas from the nebula. Such atmospheres remain in equilibrium due to a balance between an embryo’s gravity and an outward pressure gradient. However, there is a critical core mass above which an embryo can no longer support a static atmosphere. Above this limit, the atmosphere begins to collapse onto the planet forming a massive gas envelope that increases in mass over time as more gas is captured from the nebula. As gas falls toward the planet it heats up as gravitational potential energy is released. The rate at which a planet grows depends on how fast this heat can be radiated away. The critical core mass depends on the opacity of the envelope and the rate at which planetesimals collide with the core, but calculations suggest it is in the range 3–20 Earth masses, possibly less if the envelope is enriched in heavy elements. The growth of the envelope is slow at first, but speeds up rapidly once an embryo reaches 20–30 Earth masses. Numerical simulations show that Jupiter-mass planets can form this way in 1–5 million years. Such planets are mostly composed of hydrogen-rich nebular gas, and are enriched in heavier elements due to the presence of a solid core. As with the disk instability, the planet’s envelope may be further enriched in heavy elements by collisions with comets.

Measurements by the Galileo spacecraft showed that Jupiter’s upper atmosphere is enriched in carbon, nitrogen, sulfur, and the noble gases argon, krypton, and xenon by factors of two to three compared to the Sun. If these enrichments are typical of Jupiter’s envelope as a whole, it suggests the planet captured a huge number of comets. Argon can be trapped in cometary ices but only if these ices form at temperatures below about 30 K. Temperatures at Jupiter’s current distance from the Sun were probably quite a lot higher than this. This suggests either that the comets came from colder regions of the nebula or that Jupiter itself migrated inward over a large distance. However, the fact that relatively refractory elements such as sulfur are present in the same enrichment as the noble gases suggests these elements may all have been captured as gases from the nebula along with hydrogen and helium. If so, Jupiter’s envelope must be nonhomogeneous, with the lower layers depleted in heavy elements, perhaps due to exclusion from high-pressure phases of hydrogen, while the upper layers are enriched.

It is unclear why Jupiter and Saturn stopped growing when they reached their current masses. These planets are sufficiently massive that they would continue to grow very rapidly if a supply of gas was available nearby. It is possible, but unlikely, that they stopped growing because the nebula happened to disperse at this point. A more likely explanation is that the growth of these planets slowed because they each became massive enough to clear an annular gap in the nebula around their orbit. Gap clearing happens when a planet’s Hill radius becomes comparable to the vertical thickness of the gas disk, which would have been the case for Jupiter. Gas orbiting a little further from the Sun than Jupiter would have been sped up by the planet’s gravitational pull, moving the gas away from the Sun. Gas orbiting closer to the Sun than Jupiter was slowed down, causing it to move inward. These forces open up a gap in the disk around Jupiter’s orbit, balancing viscous forces that would cause gas to flow back into the gap. Numerical simulations show that generally gaps are not cleared completely, and some gas continues to cross a gap and accrete onto a planet. However, the accretion rate declines as a planet becomes more massive. Saturn’s growth may have been truncated because the combined gravity of Jupiter and Saturn cleared a gap in the disk around both planets. Uranus and Neptune are referred to as ice-giant planets since they contain large amounts of materials such as water and methane that form ices at low temperatures. They contain some hydrogen and helium, but they did not acquire the huge gaseous envelopes that Jupiter and Saturn possess. This suggests the nebula gas had largely dispersed in the region where Uranus and Neptune were forming before they became massive enough to undergo rapid gas accretion. This may be because they formed in the outer regions of the protoplanetary disk, where embryo growth rates were slowest. It is also possible that the nebula dispersed more quickly in some regions than others. In particular, the outer regions of the nebula may have disappeared at an early stage as the gas escaped the solar system due to photoevaporation by ultraviolet radiation.

The presence of a gap modifies planetary migration. Planets massive enough to open a gap still generate spiral density waves in the gas beyond the gap, but these waves are located further away from the planet as a result, so migration is slower. As a planet with a gap migrates inward, gas tends to pile up at the inner edge of the gap, and become rarified at the outer edge, slowing migration as a result. The migration of the planet now becomes tied to the inward viscous accretion of the gas toward the star. The planet, its gap, and the nebular gas all move inward at the same rate, given by

$$\frac{da}{dt} = -1.5\alpha \left(\frac{c_s}{v_{\text{kep}}}\right)^2 v_{\text{kep}}$$

(2.16)

where $\alpha = \nu v_{\text{kep}}/(ac^2)$ and $\nu$ is the viscosity of the nebular gas. This is called type-II migration. Type-II migration slows when a planet’s mass becomes comparable to that of the nebula, and migration ceases as the nebular gas disperses. Migration can also be modified by the presence of
additional giant planets as we saw earlier when discussing the Grand Tack model.

Giant planets in the solar system experienced another kind of migration as they interacted gravitationally with planetesimals moving on orbits between the giant planets and in the primordial Kuiper belt. One consequence of this process was the formation of the Oort cloud of comets. Once Jupiter approached its current mass, many planetesimals that came close to the planet would have been flung far beyond the outer edge of the protoplanetary disk. Some were ejected from the solar system altogether, but others remained weakly bound to the Sun. Over time, gravitational interactions with molecular clouds, other nearby stars, and the galactic disk circularized the orbits of these objects so they no longer passed through the planetary system. Many of these objects are still present orbiting far from the Sun in the Oort cloud. The ultimate source of angular momentum for these objects came at the expense of Jupiter’s orbit, which shrank accordingly. Saturn, Uranus, and Neptune ejected some planetesimals, but they also perturbed many objects inward, which were then ejected by Jupiter. As a result, Saturn, Uranus, and Neptune probably moved outward rather than inward.

As Neptune migrated outward it interacted dynamically with the primordial Kuiper belt of comets orbiting in the very outer region of the nebula. Some of these comets were ejected from the solar system or perturbed inward toward Jupiter. Others were perturbed onto highly eccentric orbits with periods of hundreds or thousands of years, and now form the Scattered Disk, a region that extends out beyond the Kuiper belt but whose objects are gradually being removed by close encounters with Neptune. A sizable fraction of the objects in the region beyond Neptune were trapped in external mean-motion resonances and migrated outward with the planet. Pluto, currently located in the 3:2 mean-motion resonance with Neptune, probably represents one of these objects.

As the giant planets migrated, it is possible that they passed through orbital resonances with one another. Such a resonance crossing may have had a profound impact on every part of the solar system, the combined effects of which have come to be known as the Nice model after the French city in which it was developed. In particular, if Jupiter and Saturn passed through a mean-motion resonance, the orbital eccentricities of all four giant planets would have increased substantially. The orbits of the ice giants would have penetrated deeply into the primordial Kuiper belt, gravitationally scattering large numbers of objects onto unstable orbits crossing those of the other planets. Many comets would have been perturbed into the inner solar system as a result. In addition, the changing orbits of the giant planets would have perturbed many main-belt asteroids into unstable orbits also leading to a flux of asteroids into orbits crossing the inner planets. Currently, it is unclear whether Jupiter and Saturn passed through a resonance, or when this may have happened. It has been proposed that passage through a resonance was responsible for the late heavy bombardment of the inner planets, which occurred 600–700 million years after the start of the solar system and left a clear record of impacts on the Moon, Mars, and Mercury.

9. PLANETARY SATELLITES

Earth’s moon possesses a number of unusual features. It has a low density compared to the inner planets and it has only a very small core. The Moon is depleted in volatile materials such as water. In addition, the Earth–Moon system has a large amount of angular momentum per unit mass. If they were combined into a single body the object would rotate once every 4 h! All these features can be understood if the Moon formed as the result of an oblique impact between Earth and another large, differentiated body, sometimes referred to as Theia, late in Earth’s formation. Theia is the Greek Titaness goddess who was the mother of Selene, the goddess of the Moon.

Numerical simulations of this giant impact show that much of Theia’s core would have sunk through Earth’s mantle to coalesce with Earth’s core. Molten and vaporized mantle material from both bodies was ejected outward. Gravitational torques from the highly nonspherical distribution of matter during the collision gave some of this mantle material enough angular momentum to go into orbit about Earth. This material quickly formed into a disk, from which the Moon accreted. Certain features of the Moon’s composition are very similar to those of the Earth, which means that (1) Theia was formed from similar material, (2) the resulting vapor and debris that condensed to form the Moon totally equilibrated with the outer portions of the Earth, or (3) the Moon is mostly composed of material from Earth rather than Theia, although most numerical simulations tend to find that the opposite is true in this case.

The impact released huge amounts of energy, heating the disk sufficiently that many volatile materials escaped. As a result, the Moon formed mostly from volatile-depleted mantle materials, explaining its current composition. The simulations suggest Theia probably had a mass similar to Mars, which has roughly 1/10 the mass of Earth. We know little about Theia’s composition except that, like Mars, it seems to have been rich in geochemical volatile elements such as rubidium compared to Earth (Figure 2.9). The Earth and the Moon have identical oxygen isotope characteristics (Figure 2.10). It was once thought that this meant Earth and Theia had a similar isotopic composition, but this similarity now appears to be the result of exchange of material between the Earth and protolunar disk while the Moon was forming. The similarity may also mean that the Moon was mostly formed from material ejected from Earth’s mantle rather than the impactor.
The satellites of the giant planets are much smaller relative to their parent planet than the Moon is compared to the Earth. The Moon is roughly 1/80 of the mass of the Earth, whereas the satellite systems of Jupiter, Saturn, and Uranus each contain about 1/10,000 of the mass of their respective planet. The satellites of the giant planets can be divided into two classes with different properties. Those close to their parent planet tend to have nearly circular orbits in the same plane as the planet’s equator and orbiting in the same direction as the planet spins. These are referred to as regular satellites. Satellites orbiting further from the planet tend to have highly inclined and eccentric orbits and are called irregular satellites as a result. The regular satellites tend to be larger and include the Galilean satellites of Jupiter and Saturn’s largest satellite Titan.

The orbits of the regular satellites suggest they formed from circumplanetary disks orbiting each planet like miniature versions of the solar nebula, while the irregular satellites are thought to have been captured later. Large satellites would have moved rapidly inward through a massive circumplanetary disk due to type-I migration, on a timescale that was short compared to the lifetime of the solar nebula. For this reason, it is possible that multiple generations of satellites formed, with the satellites we see today being the last to form. It is also possible that the circumplanetary disks had very low masses containing much less gas than the solar nebula itself. Solid material would have slowly accumulated in these gas-starved disks, while the gas quickly passed through the disk and accreted onto the planet. Large satellites would have formed slowly as a result, limiting the degree to which they were heated by impacts. This idea is consistent with the fact that three of the Galilean satellites of Jupiter have retained volatile materials such as water ice, while Callisto never grew hot enough to differentiate. Orbital resonances involving two or more satellites are common. For example, the inner three Galilean satellites Io, Europa, and Ganymede have orbital periods in the ratio 1:2:4, and resonances are common in the Saturnian satellite system. This contrasts with the absence of resonances between the major planets. The ubiquity of satellite resonances suggests many of the satellites migrated considerable distances during or after their formation, becoming captured in a resonance en route. Some resonances may have arisen as the growing satellites migrated inward through their planet’s accretion disk. Others could have arisen later as tidal interactions between a planet and its satellite caused the satellites to move outward at different rates.

The Neptunian satellite system is different from those of the other giant planets, having relatively few moons with most mass contained in a single large satellite Triton, which is larger than Pluto. Triton is unusual in that its orbit is retrograde, unlike all the other large satellites in the solar system. This suggests it was captured rather than forming in situ. Several capture mechanisms have been proposed, but most are low-probability events, which makes them unlikely to explain the origin of Triton. A more plausible idea is that Triton was once part of a binary planet like the Pluto–Charon system, orbiting around the Sun. During a close encounter with Neptune, the binary components were parted. Triton’s companion remained in orbit about the Sun, taking with it enough kinetic energy to leave Triton in a bound orbit about Neptune. Triton’s orbit would have been highly eccentric initially, but tidal interactions with Neptune caused its orbit to shrink and become more circular over time. As Triton’s orbit shrank it would have disturbed the orbits of smaller satellites orbiting Neptune, leading to their destruction by mutual collisions. This is presumably the reason for the paucity of regular satellites orbiting Neptune today.

10. EXTRASOLAR PLANETS

At the time of writing about 800 confirmed planets are known orbiting stars other than the Sun, and several thousand additional candidates await verification. These are referred to as extrasolar planets or exoplanets. Many of these objects have been found using the Doppler radial velocity technique. This makes use of the fact that the gravitational pull of a planet causes its star to move in an ellipse with the same period as the orbital period of the planet. As the star moves toward and away from the observer, lines in its spectra are alternately blue- and red-shifted by the Doppler effect, indicating the planet’s presence. Current levels of precision allow the detection of gas-giant planets and ice giants, as well as Earth-mass planets that orbit close to their star. The planet’s orbital period \( P \) can be readily identified from the radial velocity variation. The mean radius of the planet’s orbit \( a \) can then be found using Kepler’s third law if the star’s mass \( M_* \) is known:

\[
\frac{a^3}{P^2} = \frac{GM_*}{4\pi^2}
\]

Unfortunately, the Doppler method determines only one component of the star’s velocity, so the orientation of the orbital plane is not known in general. This means one can obtain only a lower limit on the planet’s mass. For randomly oriented orbits, however, the true mass of the planet is most likely to lie within 30% of its minimum value.

Many other extrasolar planets have been detected when they transit across the face of their star, typically causing the star to dim by a small amount for a few hours. Only a small fraction of extrasolar planets generate a transit since their orbital plane must be almost edge on as seen from the Earth. However, the space-based Kepler mission has surveyed more than one hundred thousand stars, with the result
that several thousand possible planets have been found. These await confirmation by ground-based observers using other techniques.

When a planet is observed using both the Doppler and transit methods, its true mass can be obtained since the orientation of the orbital plane is known. If the stellar radius is also known, the degree of dimming yields the planet’s radius and hence its density. The densities of large extrasolar planets observed this way are generally comparable to that of Jupiter and substantially lower than that of Earth. This suggests these planets are composed mainly of gas rather than rock or ice. In one case, hydrogen has been detected escaping from an extrasolar planet. Recently, a number of objects have been found with masses below 10 Earth masses, and it is plausible that these are more akin to terrestrial planets or water-rich worlds with no analogue in the solar system.

Stars with known giant planets tend to have high metallicities, that is, they are enriched in elements heavier than helium compared to most stars in the Sun’s neighborhood (Figure 2.21). (The Sun also has a high metallicity.) The meaning of this correlation is hotly debated, but it is consistent with the formation of giant planets via core accretion (see Section 8). When a star has a high metallicity, its disk will contain large amounts of the elements needed to form a solid core, promoting rapid growth and increasing the likelihood that a gas giant can form before the gas disk disperses.

Both the Doppler velocity and transit techniques are biased toward finding massive planets since these generate a stronger signal. Both are also biased toward detecting planets lying close to their star. In the case of transits, the probability of suitable orbital alignment declines with increasing orbital distance, while for the Doppler velocity method, one generally needs to observe a planet for at least a full orbital period to obtain a firm detection. Despite these biases, it is clear that at least 20% of Sun-like stars have planets and this fraction may be much higher. The fraction of planets with a given mass increases as the planetary mass grows smaller, despite the strong observational bias working in the opposite direction. Roughly 10% of known extrasolar planets have orbital periods of only a few days, which implies their orbits are several times smaller than Mercury’s orbit about the Sun. These planets are often referred to as hot Jupiters or hot Neptunes due to their likely high temperatures. Theoretical models of planet formation suggest it is unlikely that planets will form this close to a star. Instead, it is thought that these planets formed at larger distances and moved inward due to type-I and/or type-II migration. Alternatively, they may have been scattered onto highly eccentric orbits following close encounters with other planets in the same system. In this case, subsequent tidal interactions with the star will circularize a planet’s orbit and cause the orbit to shrink.

More than 100 stars are known to have two or more planets. In a sizable fraction of these cases, the planets are involved in orbital resonances where either the ratio of the orbital periods or precession periods of two planets is close to the ratio of two integers, such as 2:1. This state of affairs has a low probability of occurring by chance, which suggests these planets have been captured into a resonance when the orbits of one or both planets migrated inward. Several of the planetary systems found by the Kepler mission consist of multiple low-mass planets lying close to their star. It seems likely that these planets or their building blocks must have migrated to their current location from more outlying regions of their protoplanetary disk.

11. SUMMARY AND FUTURE PROSPECTS

Thanks to improvements in isotopic chronology we now know the timescales over which the Earth, Moon, Mars, and some asteroids formed. Terrestrial-planet accretion started soon after the solar system formed, leading to the growth of some Mars-sized and smaller objects within the first million years or so. This early accretionary phase was accompanied by widespread melting due to heat generated by short-lived isotopes and the formation of planetary cores. The Moon formed relatively late, at least 30 Myrs after the start of the solar system. This was the last major event in Earth’s formation. These isotopic timescales are consistent with theoretical models that predict rapid runaway and oligarchic growth at early times, to form asteroid- to Mars-sized bodies within a million years, while predicting that Earth took tens of millions of years to grow to its final size.

The presence in Earth’s mantle of nonnegligible amounts of siderophile elements such as platinum and osmium argues that roughly 1% of Earth’s mass arrived after its core had finished forming. For some time it has been postulated that Earth formed in a very dry...
environment and that its water was delivered along with these siderophile elements in a late veneer. This now appears unlikely given the composition of Earth’s mantle. Instead, Earth probably acquired its water earlier, perhaps from carbonaceous-chondrite-like asteroids, before core formation was complete. This implies that the planet held onto much of its water during the giant impact that led to the formation of the Moon.

It now seems that chondrites, the most primitive meteorites in our collection both physically and chemically, actually formed at a rather late stage, after the parent bodies of the iron meteorites had formed. Chondrites escaped melting because the potent heat sources \( ^{26}\text{Al} \) and \( ^{60}\text{Fe} \) had largely decayed by that point. For a long time it has been thought that chondrites, or something similar, provided the basic building blocks of Earth and the other terrestrial planets, but it now seems that the parent bodies of the iron meteorites provide a better analogue in this respect. Currently, we do not have good dynamical or cosmochemical models for how chondrites and their constituents formed. Chondrules, CAIs, matrix grains, and presolar grains all survived in the nebula for several million years, enduring different degrees of thermal processing, and then were collected together into large bodies. The refractory CAIs may have formed close to the Sun prior to being scattered across the disk, perhaps by turbulent motions within the gas. Supporting evidence for this hypothesis comes from the recent discovery of high-temperature condensates in samples from comet Wild 2 returned by the Stardust mission. Where chondrules formed remains unclear, but these objects would have been highly mobile as long as nebular gas was present, and they may have drifted radially over large distances.

The origin of giant planets remains a subject of debate, but the observed correlation between stellar metallicity and the presence of giant planets, and the recent discovery of a Saturn-mass extrasolar planet that appears to have a very massive core, lend weight to the core accretion model. Recent simulations using plausible envelope opacities have discovered that giant planets can form within the typical lifetime of a protoplanetary disk, overcoming a long-standing obstacle for core accretion. It is becoming apparent that planetary migration is an important feature in the formation and early evolution of planetary systems. This presumably explains the fact that extrasolar planets are seen to orbit their stars at a wide range of distances. Planets also migrate when they clear away residual planetesimals. This may have led to a dramatic episode early in the history of the solar system associated with the late heavy bombardment of comets and asteroids onto the Moon and inner planets.

It is impressive to look back on the past two decades of discovery in planetary science partly because the breakthroughs have involved so many diverse areas of research. Technology has been a key driver, be it in the form of more powerful computers, mass spectrometers, instrumentation for planetary missions, or new telescopes and detectors. The near future looks equally exciting. The Atacama Large Millimeter Array promises to transform our knowledge of protoplanetary disks with very high spatial resolution able to observe features as small as 1 AU in size, and sufficient sensitivity to detect many new molecules including organic materials. Space missions will continue to expand our survey of the solar system, such as the New Horizons and Juno probes en route to Pluto and Jupiter, respectively, and the Rosetta spacecraft heading for comet Churyumov-Gerasimenko. In addition to National Aeronautics and Space Administration and European Space Agency, space agencies in Japan, China, and India are also becoming active players in space exploration. The Doppler radial velocity and transit techniques continue to be refined and are set to expand the catalogue of known extrasolar planets. The relatively new microlensing technique is opening up the possibility of finding Earth-mass planets. With luck, the ongoing Kepler mission should finally answer the question of whether Earth-sized planets are common or relatively rare. Here on Earth, continuing analysis of dust samples from comet Wild 2 returned by the Stardust mission, and solar wind samples from the Genesis mission, will enhance our understanding of the cosmochemical evolution of the solar system. New isotopic measurement techniques and a new generation of dynamic secondary ion spectrometers or ion probes are sure to generate exciting discoveries at a rapid pace. All in all, we have much to look forward to.

**BIBLIOGRAPHY**


1. INTRODUCTION: KEPLERIAN MOTION

The study of the motion of celestial bodies within our solar system has played a key role in the broader development of classical mechanics. In 1687, Isaac Newton published his *Principia*, in which he presented a unified theory of the motion of bodies in the heavens and on the Earth. Newtonian physics has been proved to provide a remarkably good description of a multitude of phenomena on a wide range of length scales. Many of the mathematical tools developed over the centuries to analyze planetary motions in the Newtonian framework have found applications for terrestrial phenomena. Deviations of the orbit of Uranus from that predicted by Newton’s Laws led to the discovery of the planet Neptune. However, Newtonian gravity is only an approximation to Einstein’s general theory of relativity.
This was used to explain deviations of Mercury’s orbit that could not be accounted for by Newtonian physics. But general relativistic corrections to planetary motions are quite small, so this chapter concentrates on the rich and varied effects of Newtonian gravitation, together with briefer descriptions of nongravitational forces that affect the motions of some objects in the solar system.

Newton showed that the motion of two spherically symmetric bodies resulting from their mutual gravitational attraction is described by simple conic sections (see Section 2.4). However, the introduction of additional gravitating bodies produces a rich variety of dynamical phenomena, even though the basic interactions between pairs of objects can be straightforwardly described. Even few-body systems governed by apparently simple nonlinear interactions can display remarkably complex behavior, which has come to be known collectively as chaos. The concept of deterministic chaos, now known to play a major role in weather patterns on the Earth, was first conceived in connection with planetary motions (by Poincaré, in the late nineteenth century). On sufficiently long timescales, the apparently regular orbital motion of many bodies in the solar system can exhibit symptoms of this chaotic behavior.

An object in the solar system exhibits chaotic behavior in its orbit or rotation if the motion is sensitively dependent on the starting conditions, such that small changes in its initial state produce different final states. Examples of chaotic motion in the solar system include the rotation of the Saturnian satellite Hyperion, the orbital evolution of numerous asteroids and comets, and the orbit of Pluto. Numerical investigations suggest that the motion of the planetary system as a whole is chaotic, although there are currently no signs of any gross instability in the orbits of the planets. Chaotic motion has probably played an important role in determining the dynamical structure of the solar system, particularly in its early history.

In this chapter, the basic orbital properties of solar system objects (planets, moons, minor bodies, and dust) and their mutual interactions are described. Several examples of important dynamical processes that occur in the solar system are provided and groundwork is laid for describing some of the phenomena that are discussed in more detail in other chapters of this book.

1.1. Kepler’s Laws of Planetary Motion

By analyzing Tycho Brahe’s careful observations of the orbits of the planets, Johannes Kepler deduced the following three laws of planetary motion:

1. All planets move along elliptical paths with the Sun at one focus. The heliocentric distance \( r \) (i.e. the planet’s distance from the Sun) can be expressed as

\[
r = \frac{a(1 - e^2)}{1 + e \cos f},
\]

where \( a \) is the semimajor axis (average of the minimum and maximum heliocentric distances) and \( e \) (the eccentricity of the orbit) = \((1 - b^2/a^2)^{1/2} \), where \( 2b \) is the minor axis of an ellipse. The true anomaly, \( f \), is the angle between the planet’s perihelion (closest heliocentric distance) and its instantaneous position (Figure 3.1).

2. A line connecting a planet and the Sun sweeps out equal areas \( \Delta A \) in equal periods of time \( \Delta t \):

\[
\frac{\Delta A}{\Delta t} = \text{constant}.
\]

Note that the value of this constant differs from one planet to the next.

3. The square of a planet’s orbital period \( P \) about the Sun (in years) is equal to the cube of its semimajor axis \( a \) (in AU):

\[
P^2 = a^3.
\]

1.2. Elliptical Motion, Orbital Elements, and the Orbit in Space

The Sun contains more than 99.8% of the mass of the known solar system. The gravitational force exerted by a body is proportional to its mass (Eqn (3.5)), so to an excellent first approximation, the motion of the planets and many other bodies can be regarded as being solely due to the influence of a fixed central pointlike mass. For bound objects like the planets, which cannot go arbitrarily far from the Sun, the general solution for the orbit is the ellipse described by Eqn (3.1). The orbital plane, although fixed in
space, can be arbitrarily oriented with respect to whatever reference plane is chosen (such as Earth’s orbital plane about the Sun, which is called the ecliptic, or the equator of the primary). The inclination, \(i\), of the orbital plane is the angle between the reference plane and the orbital plane and can range from 0° to 180°. Conventionally, if the orbital angular momentum of the body is aligned with the rotational angular momentum of the primary \(^1\) (or, for heliocentric orbits, with the orbital angular momentum of the Earth), then the inclination is defined to be in the 0°—90° range and the orbit is said to be prograde. Bodies traveling in the opposite direction are defined to have inclinations from 90° to 180° and are said to be on retrograde orbits. The two planes intersect in a line called the line of nodes and the orbit pierces the reference plane at two locations— one as the body passes upward through the plane (the ascending node) and one as it descends (the descending node). A fixed direction in the reference plane is chosen and the angle to the direction of the orbit’s ascending node is called the argument of periapse, \(\omega\). An additional angle, the longitude of periapse, \(\Omega\) is sometimes used in place of \(\omega\). The six orbital elements \(a\), \(e\), \(i\), \(\Omega\), \(\omega\), and \(f\) uniquely specify the location of the object in space (Figure 3.2). The first three quantities \((a\), \(e\), and \(i\)) are often referred to as the principal orbital elements, as they describe the orbit’s size, shape, and tilt, respectively.

2. THE TWO-BODY PROBLEM

In this section, the general solution to the problem of the motion of two objects under the effects of their mutual gravitational interaction is discussed.

2.1. Newton’s Laws of Motion and the Universal Law of Gravitation

Although Kepler’s laws were originally found from careful observation of planetary motion, they were subsequently shown to be derivable from Newton’s laws of motion together with his universal law of gravity. Consider a body of mass \(m_1\) at instantaneous location \(\mathbf{r}_1\) with instantaneous velocity \(\mathbf{v}_1 = d\mathbf{r}_1/dt\) and hence momentum \(\mathbf{p}_1 = m_1\mathbf{v}_1\). The acceleration \(d\mathbf{v}_1/dt\) produced by a net force \(\mathbf{F}_1\) is given by Newton’s second law of motion:

\[
\mathbf{F}_1 = \frac{d(m_1\mathbf{v}_1)}{dt}. \tag{3.4}
\]

Newton’s universal law of gravity states that a second body of mass \(m_2\) at position \(\mathbf{r}_2\) exerts an attractive force on the first body given by

\[
\mathbf{F}_1 = -\frac{Gm_1m_2}{r_{12}^3} \mathbf{r}_{12} = -\frac{Gm_1m_2}{r_{12}^2} \mathbf{r}_{12}, \tag{3.5}
\]

where \(\mathbf{r}_{12} = \mathbf{r}_1 - \mathbf{r}_2\) is the location of particle 1 with respect to particle 2, \(\mathbf{r}_{12}\) is the unit vector in the direction of \(\mathbf{r}_{12}\), and \(G\) is the gravitational constant. Newton’s third law states that for every action there is an equal and opposite reaction; thus, the force on each object of a pair is equal in magnitude but opposite in direction. These facts are used to reduce the two-body problem to an equivalent one-body case in the next subsection.

2.2. Reduction to the One-Body Case

From the foregoing discussion of Newton’s laws, and the two-body problem, the force exerted by body 1 on body 2 is

\[
\frac{d(m_2\mathbf{v}_2)}{dt} = \mathbf{F}_2 = -\mathbf{F}_1 = \frac{Gm_1m_2}{r_{12}^3} \mathbf{r}_{12} = \frac{Gm_1m_2}{r_{12}^2} \mathbf{r}_{12}. \tag{3.6}
\]

Thus, from Eqns (3.4) and (3.6)

\[
\frac{d(m_1\mathbf{v}_1 + m_2\mathbf{v}_2)}{dt} = \mathbf{F}_1 + \mathbf{F}_2 = 0. \tag{3.7}
\]

This is of course a statement that the total linear momentum of the system is conserved, which means that the center of mass of the system moves with constant velocity.

---

1. That is, the dot product of the two angular momenta is nonnegative.
Multiplying Eqn (3.6) by $m_1$ and Eqn (3.5) by $m_2$ and subtracting, the equation for the relative motion of the bodies can be cast in the form

$$\frac{d^2\mathbf{r}_{12}}{dt^2} = \frac{d^2(\mathbf{r}_1 - \mathbf{r}_2)}{dt^2} = -\frac{G\mu_t M}{r_{12}^3} \mathbf{r}_{12}, \quad (3.8)$$

where $\mu_t \equiv m_1 m_2/(m_1 + m_2)$ is called the reduced mass and $M \equiv m_1 + m_2$ is the total mass. Thus, the relative motion is completely equivalent to that of a particle of reduced mass $\mu_t$ orbiting a fixed central mass $M$. For known masses, specifying the elements of the relative orbit and the positions and velocities of the center of mass is completely equivalent to specifying the positions and velocities of both bodies. A detailed solution of the equation of motion (Eqn 3.8) is discussed in any elementary text on orbital mechanics and in most general classical mechanics books. In the remainder of Section 2, a few key results are given.

2.3. Energy, Circular Velocity, and Escape Velocity

The centripetal force necessary to keep an object of mass $\mu_t$ in a circular orbit of radius $r$ with speed $v_c$ is $\mu_t v_c^2/r$. Equating this to the gravitational force exerted by the central body of mass $M$, the circular velocity is

$$v_c = \sqrt{\frac{GM}{r}}. \quad (3.9)$$

Thus the orbital period (the time to move once around the circle) is

$$P = 2\pi r/v_c = 2\pi \sqrt{\frac{r^3}{GM}}. \quad (3.10)$$

The total (kinetic plus potential) energy $E$ of the system is a conserved quantity:

$$E = T + V = \frac{1}{2} \mu_t v^2 - \frac{GM\mu_t}{r}, \quad (3.11)$$

where the first term on the right is the kinetic energy of the system, $T$, and the second term is the potential energy of the system, $V$. If $E < 0$, the absolute value of the potential energy of the system is larger than its kinetic energy, and the system is bound. The body will orbit the central mass on an elliptical path. If $E > 0$, the kinetic energy is larger than the absolute value of the potential energy, and the system is unbound. The relative orbit is then described mathematically as a hyperbola. If $E = 0$, the kinetic and potential energies are equal in magnitude, and the relative orbit is a parabola. By setting the total energy equal to zero, the escape velocity at any separation can be calculated:

$$v_e = \sqrt{\frac{2GM}{r}} = \sqrt{2v_c}. \quad (3.12)$$

For circular orbits it is easy to show (using Eqns (3.9) and (3.11)) that both the kinetic energy and the total energy of the system are equal in magnitude to half the potential energy:

$$T = -\frac{1}{2} V, \quad (3.13)$$

$$E = -\frac{GM\mu_t}{2r}. \quad (3.14)$$

For an elliptical orbit, Eqn (3.14) holds if the radius $r$ is replaced by the semimajor axis $a$:

$$E = -\frac{GM\mu_t}{2a}. \quad (3.15)$$

Similarly, for an elliptical orbit, Eqn (3.10) becomes Newton’s generalization of Kepler’s third law:

$$p^2 = \frac{4\pi^2 a^3}{G(m_1 + m_2)}. \quad (3.16)$$

It can be shown that Kepler’s second law follows immediately from the conservation of angular momentum, $\mathbf{L}$:

$$\frac{d\mathbf{L}}{dt} = (\mu_r \mathbf{r} \times \mathbf{v}) = 0. \quad (3.17)$$

2.4. Orbital Elements: Elliptical, Parabolic, and Hyperbolic Orbits

As noted earlier, the relative orbit in the two-body problem is either an ellipse, a parabola, or a hyperbola depending on whether the energy is negative, zero, or positive, respectively. These curves are known collectively as conic sections and the generalization of Eqn (3.1) is

$$r = \frac{p}{1 + e \cos \phi}, \quad (3.18)$$

where $r$ and $f$ have the same meaning as in Eqn (3.1), $e$ is the generalized eccentricity, and $p$, the semilatus rectum, is a conserved quantity that depends on the initial conditions. For an ellipse, $p = a(1 - e^2)$, as in Eqn (3.1). For a parabola, $e = 1$ and $p = 2q$, where $q$ is the pericentric separation (distance of closest approach). For a hyperbola, $e > 1$ and $p = q(1 + e)$, where $q$ is again the pericentric separation. For all orbits, the three orientation angles $i$, $\Omega$, and $\omega$ are defined as in the elliptical case.

3. PLANETARY PERTURBATIONS AND THE ORBITS OF SMALL BODIES

Gravity is not restricted to interactions between the Sun and the planets or individual planets and their satellites, but rather all bodies feel the gravitational force of one another.
Within the solar system, one body typically produces the dominant force on any given body, and the resultant motion can be thought of as a Keplerian orbit about a primary, subject to small perturbations by other bodies. In this section, some important examples of the effects of these perturbations on the orbital motion are considered.

Classically, much of the discussion of the evolution of orbits in the solar system used perturbation theory as its foundation. Essentially, the method involves writing the equations of motion as the sum of a part that describes the independent Keplerian motion of the bodies about the Sun plus a part (called the disturbing function) that contains terms due to the pairwise interactions among the planets and minor bodies and the indirect terms associated with the backreaction of the planets on the Sun. In general, one can then expand the disturbing function in terms of the small parameters of the problem (such as the ratio of the planetary masses to the solar mass, the eccentricities and inclinations, etc.), as well as the other orbital elements of the bodies, including the mean longitudes (i.e. the location of the bodies in their orbits), and attempt to solve the resulting equations for the time dependence of the orbital elements.

### 3.1. Perturbed Keplerian Motion and Resonances

Although perturbations on a body’s orbit are often small, they cannot always be ignored. They must be included in short-term calculations if high accuracy is required, for example, for predicting when an object passes in front of a star (stellar occultation) or targeting spacecraft. Most long-term perturbations are periodic in nature, their directions oscillating with the relative longitudes of the bodies or with some more complicated function of the bodies’ orbital elements.

Small perturbations can produce large effects if the forcing frequency is commensurate with the natural frequency of oscillation of the responding elements. Under such circumstances, perturbations add coherently, and the effects of many small tugs can build up over time to create a large-amplitude, long-period response. This is an example of resonance forcing, which occurs in a wide range of physical systems.

An elementary example of resonance forcing is given by the simple one-dimensional harmonic oscillator, for which the equation of motion is

\[ m \frac{d^2 x}{dt^2} + m \Gamma^2 x = F_0 \cos \phi t. \tag{3.19} \]

In Eqn (3.19), \( m \) is the mass of the oscillating particle, \( F_0 \) is the amplitude of the driving force, \( \Gamma \) is the natural frequency of the oscillator, and \( \phi \) is the forcing or resonance frequency. The solution to Eqn (3.19) is

\[ x = x_0 \cos \phi t + A \cos \Gamma t + B \sin \Gamma t, \tag{3.20a} \]

where

\[ x_0 \equiv \frac{F_0}{m(\Gamma^2 - \phi^2)}, \tag{3.20b} \]

and \( A \) and \( B \) are constants determined by the initial conditions. Note that if \( \phi \approx \Gamma \), a large-amplitude, long-period response can occur even if \( F_0 \) is small. Moreover, if \( \phi = \Gamma \), this solution to Eqn (3.19) is invalid. In this case, the solution is given by

\[ x = \frac{F_0}{2 m \Gamma} t \sin \Gamma t + A \cos \Gamma t + B \sin \Gamma t. \tag{3.21} \]

The \( t \) in front of the first term on the right-hand side of Eqn (3.21) leads to secular growth. Often this linear growth is moderated by the effects of nonlinear terms that are not included in the simple example provided here. However, some perturbations have a secular component.

Nearly exact orbital commensurabilities exist at many places in the solar system. Io orbits Jupiter twice as frequently as Europa does, which in turn orbits Jupiter twice as frequently as Ganymede does. Conjunctions (at which the bodies have the same longitude) always occur at the same position of Io’s orbit (its perijove). How can such commensurabilities exist? After all, the probability of randomly picking a rational from the real number line is 0, and the number of small integer ratios is infinitely smaller still! The answer lies in the fact that orbital resonances may be held in place as stable locks, which result from nonlinear effects not represented in the foregoing simple mathematical example. For example, differential tidal recession (see Section 7.5) brings moons into resonance, and nonlinear interactions among the moons can keep them there. Other examples of resonance locks include the Hilda asteroids, the Trojan asteroids, Neptune—Pluto, and the pairs of moons about Saturn, Mimas—Tethys and Enceladus—Dione.

Resonant perturbation can also force material into highly eccentric orbits that may lead to collisions with other bodies; this is thought to be the dominant mechanism for clearing the Kirkwood gaps in the asteroid belt (see Section 5.1). Spiral density waves can propagate away from resonant locations in a self-gravitating particle disk perturbed by an orbiting satellite. Density waves are seen at many resonances in Saturn’s rings; they explain most of the structure seen in Saturn’s A ring. The vertical analogs of density waves, bending waves, are caused by resonant perturbations perpendicular to the ring plane due to a satellite in an orbit that is inclined to the ring. Spiral bending waves excited by the moons Mimas and Titan have been seen in Saturn’s rings. In the next few subsections, these manifestations of resonance effects that do not explicitly involve chaos are discussed. Chaotic motion produced by resonant forcing is discussed later in the chapter.
3.2. Examples of Resonances: Lagrangian Points and Tadpole and Horseshoe Orbits

Many features of the orbits considered in this section can be understood by examining an idealized system in which two massive (but typically of unequal mass) bodies move in circular orbits about their common center of mass. If a third body is introduced that is much less massive than either of the first two, its motion can be followed by assuming that its gravitational force has no effect on the orbits of the other bodies. By considering the motion in a frame corotating with the massive pair (so that the pair remain fixed on a line that can be taken to be the \( x \)-axis), Lagrange found that there are five points where particles placed at rest would feel no net force in the rotating frame. Three of the so-called Lagrange points \((L_1, L_2, \text{ and } L_3)\) lie along a line joining the two masses \( m_1 \) and \( m_2 \). The other two Lagrange points \((L_4 \text{ and } L_5)\) form equilateral triangles with the two massive bodies.

Particles displaced slightly from the first three Lagrangian points will continue to move away and hence these locations are unstable. The triangular Lagrangian points are potential energy maxima, which are stable for sufficiently large primary to secondary mass ratio due to the Coriolis force. Provided that the most massive body has at least 25 times the mass of the secondary (which is the case for all known examples in the solar system larger than the Pluto–Charon system), the Lagrangian points \( L_4 \) and \( L_5 \) are stable points. Thus, a particle at \( L_4 \) or \( L_5 \) that is perturbed slightly will start to “orbit” these points in the rotating coordinate system. Lagrangian points \( L_4 \) and \( L_5 \) are important in the solar system. For example, the Trojan asteroids in Jupiter’s Lagrangian points and both Neptune and Mars confine their own Trojans. There are also small moons in the triangular Lagrangian points of Tethys and Dione, in the Saturnian system. The \( L_4 \) and \( L_5 \) points in the Earth–Moon system have been suggested as possible locations for space stations.

3.2.1. Horseshoe and Tadpole Orbits

Consider a moon on a circular orbit about a planet. Figure 3.3 shows some important dynamical features in the frame corotating with the moon. All five Lagrangian points are indicated in the picture. A particle just interior to the moon’s orbit has a higher angular velocity than the moon in the stationary frame and thus moves with respect to the moon in the direction of corotation. A particle just outside the moon’s orbit has a smaller angular velocity and moves away from the moon in the opposite direction. When the outer particle approaches the moon, the particle is slowed down (loses angular momentum) and, provided the initial difference in semimajor axis is not too large, the particle drops to an orbit lower than that of the moon. The particle then recedes in the forward direction. Similarly, the particle at the lower orbit is accelerated as it catches up with the moon, resulting in an outward motion toward the higher, slower orbit. Orbits like these encircle the \( L_3, L_4, \text{ and } L_5 \) points and are called horseshoe orbits. Saturn’s small moons Janus and Epimetheus execute just such a dance, changing orbits every 4 years.

Since the Lagrangian points \( L_4 \) and \( L_5 \) are stable, material can librate about these points individually; such orbits are called tadpole orbits. The tadpole libration width at \( L_4 \) and \( L_5 \) is roughly equal to \((m/M)^{1/2}r\), and the horseshoe width is \((m/M)^{1/3}r\), where \( M \) is the mass of the planet, \( m \) the mass of the satellite, and \( r \) the distance between the two objects. For a planet of Saturn’s mass, \( M = 5.7 \times 10^{29} \text{ g} \), and a typical small moon of mass \( m = 10^{20} \text{ g} \) (e.g. an object with a 30-km radius, with density of \( \sim 1 \text{ g/cm}^3 \)), at a distance of 2.5 Saturnian radii, the tadpole libration half-width is about 3 km and the horseshoe half-width is about 60 km.

3.2.2. Hill Sphere

The approximate limit to a planet’s gravitational dominance is given by the extent of its Hill sphere,

\[
R_H = \left[ \frac{m}{3(M+m)} \right]^{1/3} a, \tag{3.22}
\]
where \( m \) is the mass of the planet and \( M \) is the Sun’s mass. A test body located at the boundary of a planet’s Hill sphere is subjected to a gravitational force from the planet that is comparable in magnitude to the tidal difference between the force of the Sun on the planet and that on the test body. The Hill sphere essentially stretches out to the \( L_1 \) point and is roughly the limit of the Roche lobe (maximum extent of an object held together by gravity alone) of a body with \( m \ll M \). Planetocentric orbits that are stable over long periods are those well within the boundary of a planet’s Hill sphere; the overwhelming majority of natural satellites lie in this region. The trajectories of the outermost planetary satellites, which lie closest to the boundary of the Hill sphere, show large variations in planetocentric orbital paths (Figure 3.4). Stable heliocentric orbits are those that are always well outside the Hill sphere of any planet.

### 3.3. Examples of Resonances: Ring Particles and Shepherding

In the discussions in Section 2, the gravitational force produced by a spherically symmetric body was described. In this section, the effects of deviations from spherical symmetry must be included when computing the force. This is most conveniently done by introducing the gravitational potential \( \Phi(r) \), which is defined such that the acceleration \( \frac{d^2r}{dt^2} \) of a particle in the gravitational field is

\[
\frac{d^2r}{dt^2} = \nabla \Phi. \tag{3.23}
\]

In empty space, the Newtonian gravitational potential \( \Phi(r) \) always satisfies Laplace’s equation

\[
\nabla^2 \Phi = 0. \tag{3.24}
\]

Most planets are very nearly axisymmetric, with the major departure from sphericity being due to a rotationally induced equatorial bulge. Thus, the gravitational potential can be expanded in terms of Legendre polynomials instead of the complete spherical harmonic expansion, which would be required for the potential of a body of arbitrary shape:

\[
\Phi(r, \phi, \theta) = -\frac{Gm}{r} \left[ 1 - \sum_{n=2}^{\infty} J_n P_n(\cos \theta) \left(\frac{R}{r}\right)^n \right]. \tag{3.25}
\]

This equation uses standard spherical coordinates, so that \( \theta \) is the angle between the planet’s symmetry axis and the vector to the particle. The terms \( P_n(\cos \theta) \) are the Legendre polynomials and \( J_n \) are the gravitational moments determined by the planet’s mass distribution. If the planet’s mass distribution is symmetrical about the planet’s equator,
To express the radial epicyclic frequency in terms of the satellite’s effect on the ring particle, we decompose the gravitational potential into its Fourier components. For a particle orbit that is nearly equatorial, the oblateness of a planet causes the line of periapsis to precess and the line of nodes to regress.

Resonances occur where the radial (or vertical) frequency of the ring particles is equal to the frequency of a component of a satellite’s horizontal (or vertical) forcing, as experienced in the rotating frame of the particle. In this case, the resonating particle is always near the same phase in its radial (or vertical) oscillation when it experiences a particular phase of the satellite’s forcing. This situation enables continued coherent “kicks” from the satellite to build up the particle’s radial (or vertical) motion, and significant forced oscillations may thus result. The location and strengths of resonances with any given satellite can be determined by decomposing the gravitational potential of the satellite’s effect on the ring particle into its Fourier components. The disturbance frequency, $\bar{\omega}$, can be written as the sum of integer multiples of the satellite’s angular, vertical, and radial frequencies:

$$\bar{\omega} = j\omega_{ss} + k\mu_{ss} + \ell\kappa_{ss},$$

where the azimuthal symmetry number, $j$, is a nonnegative integer, and $k$ and $\ell$ are integers, with $k$ being even for horizontal forcing and odd for vertical forcing. The subscript $s$ refers to the satellite. A particle placed at distance $r = r_L$ will undergo horizontal (Lindblad) resonance if $r_L$ satisfies

$$\bar{\omega} - j\omega_{ss} = \pm \kappa_{ss}(r_L).$$

It will undergo vertical resonance if its radial position, $r_v$, satisfies

$$\bar{\omega} - j\omega_{ss} = \pm \mu_{ss}(r_v).$$

When Eqn (3.34) is valid for the lower (upper) sign, $r_L$ is referred to as the inner (outer) Lindblad or horizontal resonance. The distance $r_v$ is called an inner (outer) vertical resonance if Eqn (3.35) is valid for the lower (upper) sign. Since all of Saturn’s large satellites orbit the planet well outside the main ring system, the satellite’s angular frequency $\omega_{ss}$ is less than the angular frequency of the particle and inner resonances are more important than the outer ones. When $j \neq 1$, the approximation $\mu = n \approx k$ may be used to obtain the ratio

$$\frac{n(r_{Lss})}{\omega_{ss}} = \frac{j + k + \ell}{j - 1}.$$
The location and strengths of such orbital resonances can be calculated from known satellite masses and orbital parameters and Saturn’s gravity field. Most strong resonances in the Saturnian system lie in the outer A ring, near the orbits of the moons responsible for them. If \( n = \mu = k \), the locations of the horizontal and vertical resonances would be identical: \( r_L = r_v \). However, owing to Saturn’s oblateness, \( \mu > n > k \) and the positions \( r_L \) and \( r_v \) do not coincide, i.e. \( r_v < r_L \). A detailed discussion of spiral density waves, spiral bending waves, and gaps at resonances produced by moons is presented elsewhere in this encyclopedia. (See Planetary Rings.)

4. CHAOTIC MOTION

4.1. Concepts of Chaos

In the nineteenth century, Henri Poincaré studied the mathematics of the circular restricted three-body problem. In this problem, one mass (the secondary) moves in a fixed, circular orbit about a central mass (the primary), while a massless (test) particle moves under the gravitational effect of both masses but does not perturb their orbits. From this work, Poincaré realized that despite the simplicity of the equations of motion, some solutions to the problem exhibit complicated behavior.

Poincaré’s work in celestial mechanics provided the framework for the modern theory of nonlinear dynamics and ultimately led to a deeper understanding of the phenomenon of chaos, whereby dynamical systems described by simple equations can give rise to unpredictable behavior. The whole question of whether or not a given system is stable to sufficiently small perturbations is the basis of the Kolmogorov-Arnol’d-Moser theory, which has its origins in the work of Poincaré.

One characteristic of chaotic motion is that small changes in the starting conditions can produce vastly different final outcomes. Since all measurements of positions and velocities of objects in the solar system have finite accuracy, relatively small uncertainties in the initial state of the system can lead to large errors in the final state for initial conditions that lie in chaotic regions in phase space.

This is an example of what has become known as the “butterfly effect”, first mentioned in the context of chaotic weather systems. It has been suggested that under the right conditions, a small atmospheric disturbance (such as the flapping of a butterfly’s wings) in one part of the world could ultimately lead to a hurricane in another part of the world.

The changes in an orbit that reveal it to be chaotic may occur very rapidly, for example, during a close approach to the planet, or very slowly as perturbations accumulate over millions or even billions of years. Although there have been a number of significant mathematical advances in the study of nonlinear dynamics since Poincaré’s time, the digital computer has been proved to be the most important tool in investigating chaotic motion in the solar system. This is particularly true in studies of the gravitational interaction of all the planets, where there are few analytical results.

4.2. The Three-Body Problem as a Paradigm

The characteristics of chaotic motion are common to a wide variety of dynamical systems. In the context of the solar system, the general properties are best described by considering the planar circular restricted three-body problem. This idealization consists of a (massless) test particle and two bodies of masses \( m_1 \) and \( m_2 \) moving in circular orbits about their common center of mass at constant separation, with all bodies moving in the same plane. The test particle is attracted to each mass under the influence of the inverse square law of force given in Eqn (3.5). In Eqn (3.16), \( r \) is the constant separation of the two masses and \( n = 2 \pi / P \) is their constant angular velocity about the center of mass. Using \( x \) and \( y \) as components of the position vector of the test particle referred to the center of mass of the system (Figure 3.5), the equations of motion of the particle in a reference frame rotating at angular velocity \( n \) are

\[
\dot{x} - 2ny - n^2x = -G \left( \frac{m_1 x + \mu_2}{r_1^2} - \frac{m_2 x - \mu_1}{r_2^2} \right),
\]

\[
\dot{y} + 2nx - n^2y = -G \left( \frac{m_1 r_1^2 + m_2 r_2^2}{r_1^2 r_2^2} \right),
\]

where \( \mu_1 \equiv m_1 a / (m_1 + m_2) \) and \( \mu_2 \equiv m_2 a / (m_1 + m_2) \) are constants and

\[
r_1^2 = (x + \mu_2)^2 + y^2,
\]

\[
r_2^2 = (x - \mu_1)^2 + y^2,
\]

![FIGURE 3.5](Image) The rotating coordinate system used in the circular restricted three-body problem. The masses are at a fixed distance from one another and this is taken to be the unit of length. The position and velocity vectors of the test particle (at point P) are referred to the center of mass of the system at \( O \).
where \( r_1 \) and \( r_2 \) are the distances of the test particle from the masses \( m_1 \) and \( m_2 \), respectively.

These two second-order, coupled, nonlinear differential equations can be solved numerically provided the initial position \((x_0, y_0)\) and velocity \((\dot{x}_0, \dot{y}_0)\) of the particle are known. Therefore the system is deterministic, and at any given time, the orbital elements of the particle (such as its semimajor axis and eccentricity) can be calculated from its initial position and velocity.

The region of space open to the test particle may be constrained by the existence of a constant of the motion called the Jacobi constant, \( C \), given by

\[
C = r_1^2(x^2 + y^2) + 2G\left(\frac{m_1}{r_1} + \frac{m_2}{r_2}\right) - \dot{x}^2 - \dot{y}^2.
\]  
(3.41)

The values of \((x_0, y_0)\) and \((\dot{x}_0, \dot{y}_0)\) fix the value of \( C \) for the system, and this value is preserved for all subsequent motion. At any instant, the particle is at some position on the two-dimensional \((x, y)\) plane. However, since the actual orbit is also determined by the components of the velocity \((\dot{x}, \dot{y})\), the particle can also be thought of as being at a particular position in a four-dimensional \((4D)\) \((x, y, \dot{x}, \dot{y})\) phase space. Note that the use of four dimensions rather than the customary two is simply a means of representing the position and the velocity of the particle at a particular instant in time; the particle’s motion is always restricted to the \(x-y\) plane. The existence of the Jacobi constant implies that the particle is not free to wander over the entire \(4D\) phase space, but rather that its motion is restricted to the three-dimensional “surface” defined by Eqn (3.41). This has an important consequence for studying the evolution of orbits in the problem.

The usual method is to solve the equations of motion; convert \(x, y, \dot{x}, \dot{y}\) into orbital elements such as semimajor axis, eccentricity, longitude of periapse, and mean longitude; and then plot the variation of these quantities as a function of time. However, another method is to produce a surface of section, also called a Poincaré map. This makes use of the fact that the orbit is always subject to Eqn (3.41), where \( C \) is determined by the initial position and velocity. Therefore, if any three of the four quantities \(x, y, \dot{x}, \dot{y}\) are known, the fourth can always be determined by solving Eqn (3.41). One common surface of section that can be obtained for the planar circular restricted three-body problem is a plot of values of \(x\) and \(\dot{x}\) whenever \(y = 0\) and \(\dot{y}\) is positive. The actual value of \(\dot{y}\) can always be determined uniquely from Eqn (3.41), and so the two-dimensional \((x, \dot{x})\) plot implicitly contains all the information about the particle’s location in the \(4D\) phase space. Although surfaces of section make it more difficult to study the evolution of the orbital elements, they have the advantage of revealing the characteristic motion of the particle (regular or chaotic) and a number of orbits can be displayed on the same diagram.

As an illustration of the different types of orbits that can arise, the results of integrating a number of orbits using a mass \(m_1/(m_1 + m_2) = 10^{-3}\) and Jacobi constant \(C = 3.07\) are described next. In each case, the particle was started with the initial longitude of periapse \(\varpi_0 = 0\) and initial mean longitude \(\lambda_0 = 0\). This corresponds to \(x = 0\) and \(y = 0.55, \quad y_0 = 0\), with \(\dot{y} = 0.9290\) determined from the solution of the three-dimensional “surface” defined by Eqn (3.41). Here a set of dimensionless coordinates are used in which \(n = 1\), \(G = 1\), and \(m_1 + m_2 = 1\). In these units, the orbit of \(m_2\) is a circle at distance \(a = 1\) with uniform speed \(v = 1\). The corresponding initial values of the heliocentric semimajor axis and eccentricity are \(a_0 = 0.6944\) and \(e_0 = 0.2065\). Since the semimajor axis of Jupiter’s orbit is 5.202 AU, this value of \(a_0\) would correspond to an asteroid at 3.612 AU.

Figure 3.6 shows the evolution of \(e\) as a function of time. The plot shows regular behavior with the eccentricity varying from 0.206 to 0.248 over the course of the integration. In fact, an asteroid at this location would be close to an orbit—orbit resonance with Jupiter, where the ratio of the orbital period of the asteroid, \(T\), to Jupiter’s period, \(T_J\), is close to a rational number. From Kepler’s third law of planetary motion, \(T^2 \propto a^3\). In this case, \(T/T_J = (a/a_J)^{3/2} = 0.564 \approx 4/7\) and the asteroid orbit is close to a 7:4 resonance with Jupiter

![Figure 3.6](image-url)
resonance with Jupiter. Figure 3.7 shows the variation of the semimajor axis of the asteroid, \(a\) over the same time interval as shown in Figure 3.6. Although the changes in \(a\) are correlated with those in \(e\), they are smaller in amplitude and \(a\) appears to oscillate about the location of the exact resonance at \(a = (4/7)^{2/3} \approx 0.689\). An asteroid in resonance experiences enhanced gravitational perturbations from Jupiter, which can cause regular variations in its orbital elements. The extent of these variations depends on the asteroid’s location within the resonance, which is, in turn, determined by the starting conditions.

The equations of motion can be integrated with the same starting conditions to generate a surface of section by plotting the values of \(x\) and \(\dot{x}\) whenever \(y = 0\) with positive \(\dot{y}\) (Figure 3.8). The pattern of three distorted curves or “islands” that emerges is a characteristic of resonant motion when displayed in such plots. If a resonance is of the form \((p + q)/p\), where \(p\) and \(q\) are integers, then \(q\) is said to be the order of the resonance. The number of islands seen in a surface-of-section plot of a given resonant trajectory is equal to \(q\). In this case, \(p = 4\), \(q = 3\), and three islands are visible.

The center of each island would correspond to a starting condition that placed the asteroid at exact resonance where the variation in \(e\) and \(a\) would be minimal. Such points are said to be fixed points of the Poincaré map. If the starting location was moved farther away from the center, the subsequent variations in \(e\) and \(a\) would get larger, until eventually some starting values would lead to trajectories that were not in resonant motion.

### 4.2.2. Chaotic Orbits

Figures 3.9 and 3.10 show the plots of \(e\) and \(a\) as a function of time for an asteroid orbit with starting values \(x_0 = 0.56\),

---

**FIGURE 3.7** The semimajor axis as a function of time for an object using the same starting conditions as in Figure 3.6. The units of the semimajor axis are such that Jupiter’s semimajor axis (5.202 AU) is taken to be unity.

**FIGURE 3.8** A surface-of-section plot for the same (regular) orbit shown in Figures 3.6 and 3.7. The 2000 points were generated by plotting the values of \(x\) and \(\dot{x}\) whenever \(y = 0\) with positive \(\dot{y}\). The three “islands” in the plot are due to the third-order 7:4 resonance.

**FIGURE 3.9** The eccentricity as a function of time for an object moving in a chaotic orbit starting just outside the 7:4 resonance with Jupiter. The plot was obtained by solving the circular restricted three-body problem numerically using initial values of 0.6984 and 0.1967 for the semimajor axis and eccentricity, respectively. The corresponding position and velocity in the rotating frame were \(x_0 = 0.56\), \(y_0 = 0\), \(x_0 = 0\), and \(\dot{y} = 0.8998\).

**FIGURE 3.10** The semimajor axis as a function of time for an object using the same starting conditions as in Figure 3.9. The units of the semimajor axis are such that Jupiter’s semimajor axis (5.202 AU) is taken to be unity.
$y_0 = 0$, and $\dot{x}_0 = 0$, and $\dot{y}$ determined from Eqn (3.41) with $C = 3.07$. The corresponding orbital elements are $a_0 = 0.6984$ and $e_0 = 0.1967$. These values are only slightly different from those used earlier, indeed the initial behavior of the plots is quite similar to that seen in Figures 3.6 and 3.7. However, subsequent variations in $e$ and $a$ are strikingly different. The eccentricity varies from 0.188 to 0.328 in an irregular manner, and the value of $a$ is not always close to the value associated with exact resonance. This is an example of a chaotic trajectory where the variations in the orbital elements have no obvious periodic or quasi-periodic structure. The anticorrelation of $a$ and $e$ can be explained in terms of the Jacobi constant.

The identification of this orbit as chaotic becomes apparent from a study of its surface of section (Figure 3.11). Clearly, this orbit covers a much larger region of phase space than the previous example. Furthermore, the orbit does not lie on a smooth curve, but is beginning to fill an area of the phase space. The points also help to define a number of empty regions, three of which are clearly associated with the 7:4 resonance seen in the regular trajectory. There is also a tendency for the points to “stick” near the edges of the islands; this gives the impression of regular motion for short periods.

Chaotic orbits have the additional characteristic that they are sensitively dependent on initial conditions. This is illustrated in Figure 3.12, where the variation in $e$ as a function of time is shown for two trajectories; the first corresponds to Figure 3.9 (where $x_0 = 0.56$) and the second has $x_0 = 0.56001$. The initial value of $\dot{y}$ was chosen so that the same value of $C$ was obtained. Although both trajectories show comparable initial variations in $e$, after 60 Jupiter periods it is clear that the orbits have drifted apart. Such a divergence would not occur for nearby orbits in a regular part of the phase space.

The rate of divergence of nearby trajectories in such numerical experiments can be quantified by monitoring the evolution of two orbits that started close together. In a dynamical system such as the three-body problem, there are a number of quantities called the Lyapunov characteristic exponents. Measurement of the local divergence of nearby trajectories leads to an estimate of the largest of these exponents, and this can be used to determine whether or not the system is chaotic. If two orbits are separated in phase space by a distance $d_0$ at time $t_0$, and $d$ is their separation at time $t$, then the orbit is chaotic if

$$d = d_0 \exp \gamma (t - t_0),$$

(3.42)

where $\gamma$ is a positive quantity equal to the maximum Lyapunov characteristic exponent. However, in practice the Lyapunov characteristic exponents can only be derived analytically for a few idealized systems. For practical problems in the solar system, $\gamma$ can be estimated from the results of a numerical integration by writing

$$\gamma = \lim_{t \to \infty} \frac{\ln(d/d_0)}{t - t_0}$$

(3.43)

and monitoring the behavior of $\gamma$ with time. A plot of $\gamma$ as a function of time on a log–log scale reveals a striking difference between regular and chaotic trajectories. For regular orbits, $d \approx d_0$ and a log–log plot has a slope of $-1$. However, if the orbit is chaotic, then $\gamma$ tends to a constant nonzero value. This method may not always work because $\gamma$ is defined only in the limit as $t \to \infty$ and sometimes chaotic orbits may give the appearance of being regular orbits for long periods by sticking close to the edges of the islands, such as those visible in Figure 3.8.
If the nearby trajectory drifts too far from the original one, then $\gamma$ is no longer a measure of the local divergence of the orbits. To overcome this problem, it helps to rescale the separation of the nearby trajectory at fixed intervals. Figure 3.13 shows $\log \gamma$ as a function of $\log t$ calculated using this method for the regular and chaotic orbits described here. This leads to an estimate of $\gamma = 10^{-0.77}$ (Jupiter periods)$^{-1}$ for the maximum Lyapunov characteristic exponent of the chaotic orbit. The corresponding Lyapunov time is given by $1/\gamma$, or in this case ~6 Jupiter periods. This indicates that for this starting condition the chaotic nature of the orbit quickly becomes apparent.

It is important to realize that a chaotic orbit is not necessarily unbounded. The maximum Lyapunov characteristic exponent concerns local divergence and provides no information about the global stability of the trajectory. The phrase “wandering on a leash” is an apt description of objects on bounded chaotic orbits—the motion is contained but yet chaotic at the same time. Another consideration is that numerical explorations of chaotic systems have many pitfalls both in how the physical system is modeled and whether or not the model provides an accurate portrayal of the real system.

### 4.2.3. Location of Regular and Chaotic Regions

The extent of the chaotic regions of the phase space of a dynamical system can depend on a number of factors. In the case of the circular restricted three-body problem, the critical quantities are the values of the Jacobi constant and the mass ratio $\mu_2$. In Figures 3.14 and 3.15, 10 trajectories are shown for each of two different values of the Jacobi constant. In the first case (Figure 3.14), the value is $C = 3.07$ (the same as the value used in Figures 3.8 and 3.11), whereas in Figure 3.15 it is $C = 3.13$. It is clear that the extent of the chaos is reduced in Figure 3.15. The value of $C$ in the circular restricted problem determines how close the asteroid can get to Jupiter. Larger values of $C$ correspond to orbits with greater minimum distances from Jupiter. For the case $\mu_2 = 0.001$ and $C > 3.04$, it is impossible for their orbits to intersect, although the perturbations can still be significant.

Close inspection of the separatrices in Figures 3.14 and 3.15 reveals that they consist of chaotic regions with regular regions on either side. As the value of the Jacobi constant decreases, the extent of the chaotic separatrices
increases until the regular curves separating adjacent resonances are broken down and neighboring chaotic regions begin to merge. This can be thought of as the overlap of adjacent resonances giving rise to chaotic motion. It is this process that permits chaotic orbits to explore regions of the phase space that are inaccessible to the regular orbits. In the context of the Sun–Jupiter–asteroid problem, this observation implies that asteroids in certain orbits are capable of large excursions in their orbital elements.

5. ORBITAL EVOLUTION OF MINOR BODIES

5.1. Asteroids

With more than 380,000 accurately determined orbits and one major perturber (the planet Jupiter), the asteroids provide a natural laboratory in which to study the consequences of regular and chaotic motion. Using suitable approximations, asteroid motion can be studied analytically in some special cases. However, it is frequently necessary to resort to numerical integration. (See Main-Belt Asteroids.)

Investigations have shown that a number of asteroids have orbits that result in close approaches to planets. Of particular interest are asteroids such as 433 Eros, 1038 Ganymed, and 4179 Toutatis, because they are on orbits that bring them close to the Earth. One of the most striking examples of the butterfly effect (see Section 4.1) in the context of orbital evolution is the orbit of asteroid 2060 Chiron, which has a perihelion close to Saturn’s orbit and an aphelion close to Uranus’s orbit. Numerical integrations based on the best available orbital elements show that it is impossible to determine Chiron’s past or future orbit with any degree of certainty since it frequently suffers close approaches to Saturn and Uranus. In such circumstances, the outcome strongly depends on the initial conditions as well as the accuracy of the numerical method. These are the characteristic signs of a chaotic orbit. By integrating several orbits with initial conditions close to the nominal values, it is possible to carry out a statistical analysis of the orbital evolution. Studies suggest that there is a one in eight chance that Saturn will eject Chiron from the solar system on a hyperbolic orbit, while there is a seven in eight chance that it will evolve toward the inner solar system and come under strong perturbations from Jupiter. Telescopic observations of a faint coma surrounding Chiron imply that it is a comet rather than an asteroid; perhaps its future orbit will resemble that of a short-period comet of the Jupiter family.

Numerical studies of the orbital evolution of planet-crossing asteroids under the effects of perturbations from all the planets have shown a remarkable complexity of motion for some objects. For example, the Earth-crossing asteroid 1620 Geographos gets trapped temporarily in a number of resonances with the Earth in the course of its chaotic evolution (Figure 3.16).

A histogram of the number distribution of asteroid orbits in semimajor axis (Figure 3.17) shows that apart from a clustering of asteroids near Jupiter’s semimajor axis at 5.2 AU, there is an absence of objects within 0.75 AU of the orbit of Jupiter. The objects with the same orbital distance (semimajor axis) as Jupiter are the Trojan asteroids (Section 3.2), which librate about the $L_4$ and $L_5$ triangular

\[ \text{FIGURE 3.16} \quad \text{A plot of the semimajor axis of the near-Earth asteroid 1620 Geographos over a backward and forward integration of 100,000 years starting in 1986. Under perturbations from the planets, Geographos moves in a chaotic orbit and gets temporarily trapped in a number of high-order, orbit–orbit resonances (indicated in the diagram) with the Earth. The data are taken from a numerical study of planet-crossing asteroids undertaken by A. Milani and coworkers. Courtesy of Academic Press.} \]

\[ \text{FIGURE 3.17} \quad \text{A histogram of the distribution of the numbered asteroids as of August 2012 with semimajor axis together with the locations of the major Jovian resonances. Most objects lie in the main belt between 2.0 and 3.3 AU, where the outer edge is defined by the location of the 2:1 resonance with Jupiter. The width of each bin is 0.02 AU. Apart from gaps (the Kirkwood gaps) at the 3:1, 5:2, 2:1, and other resonances in the main belt, there are small concentrations of asteroids at the 3:2 and 1:1 resonances (the Hilda and Trojan groups, respectively). Note that observational biases result in overrepresentation of asteroids orbiting near the inner edge of the asteroid belt and underrepresentation of distant asteroids.} \]
Lagrangian points located \( \sim 60^\circ \) ahead of and behind Jupiter.

The cleared region near Jupiter’s orbit can be understood in terms of chaotic motion due to the overlap of adjacent resonances. In the context of the Sun–Jupiter–asteroid restricted three-body problem, the perturber (Jupiter) has an infinite sequence of first-order resonances that lie closer together as its semimajor axis is approached. For example, the 2:1, 3:2, 4:3, and 5:4 resonances with Jupiter lie at 3.3, 4.0, 4.3, and 4.5 AU, respectively. Since each \((p + 1):p\) resonance (where \(p\) is a positive integer) has a finite width in semimajor axis that is almost independent of \(p\), adjacent resonances will always overlap for some value of \(p\) greater than a critical value, \(p_{\text{crit}}\). This value is given by

\[
p_{\text{crit}} \approx 0.51 \left( \frac{m}{m + M} \right)^{-2/7}
\]  

(3.44)

where, in this case, \(m\) is the mass of Jupiter and \(M\) is the mass of the Sun. This equation can be used to predict that resonance overlap and chaotic motion should occur for \(p\) values greater than 4; this corresponds to a semimajor axis near 4.5 AU. Therefore, chaos may have played a significant role in the depletion of the outer asteroid belt.

The histogram in Figure 3.17 also shows a number of regions in the main belt where there are few asteroids. The gaps at 2.5 and 3.3 AU were first detected in 1867 by Daniel Kirkwood using a total sample of fewer than 100 asteroids; these are now known as the Kirkwood gaps. Their locations coincide with prominent Jovian resonances (indicated in Figure 3.17), and this led to the hypothesis that they were created by the gravitational effect of Jupiter on asteroids that had orbited at these semimajor axes. The exact removal mechanism was unclear until the 1980s, when several numerical and analytical studies showed that the central regions of these resonances contained large chaotic zones.

The Kirkwood gaps cannot be understood using the model of the circular restricted three-body problem described in Section 4.2. The eccentricity of Jupiter’s orbit, although small (0.048), plays a crucial role in producing the large chaotic zones that help to determine the orbital evolution of asteroids. On timescales of several hundreds of thousands of years, the mutual perturbations of the planets act to change their orbital elements and Jupiter’s eccentricity can vary from 0.025 to 0.061. An asteroid in the chaotic zone at the 3:1 resonance would undergo large, essentially unpredictable changes in its orbital elements. In particular, the eccentricity of the asteroid could become large enough for it to cross the orbit of Mars. This is illustrated in Figure 3.18 for a fictitious asteroid with an initial eccentricity of 0.15 moving in a chaotic region of the phase space at the 3:1 resonance. Although the asteroid can have periods of relatively low eccentricity, there are large deviations and \(e\) can reach values in excess of 0.3. Taking the eccentricity of Mars’s orbit to be its maximum value of 0.14, this implies that there will be times when the orbits could intersect (Figure 3.19). In this case, the asteroid orbit would be unstable, since it is likely to either impact the surface of Mars or suffer a close approach that would drastically alter its semimajor axis. Although Jupiter provides the perturbations, it is Mars, Earth, or Venus that ultimately removes the asteroids from the 3:1 resonance.
Figure 3.20 shows the excellent correspondence between the distribution of asteroids close to the 3:1 resonance and the maximum extent of the chaotic region determined from numerical experiments.

The situation is less clear for other resonances, although there is good evidence for large chaotic zones at the 2:1 and 5:2 resonances. In the outer part of the main belt, large changes in eccentricity will cause the asteroid to cross the orbit of Jupiter before it gets close to Mars. There may also be perturbing effects from other planets. In fact, it is now known that **secular resonances** have an important role to play in the clearing of the Kirkwood gaps, including the one at the 3:1 resonance. Once again, chaos is involved. Studies of asteroid motion at the 3:2 Jovian resonance indicate that the motion is regular, at least for low values of the eccentricity. This may help to explain why there is a local concentration of asteroids (the Hilda group) at this resonance, whereas others are associated with an absence of material.

Since the dynamical structure of the asteroid belt has been determined by the perturbative effects of nearby planets, it seems likely that the original population was much larger and more widely dispersed. Therefore, the current distribution of asteroids may represent objects that are either recent collision products or that have survived in relatively stable orbits over the age of the solar system. Asteroids can also undergo orbital evolution due to nongravitational forces such as the **Yarkovsky effect**.

### 5.2. Meteorites

Most meteorites are thought to be the fragments of material produced from collisions in the asteroid belt, and the reflectance properties of certain meteorites are known to be similar to those of common types of asteroids. Since most collisions take place in the asteroid belt, the fragments have to evolve into Earth-crossing orbits before they can hit the Earth and be collected as samples.

An estimate of the time taken for a given meteorite to reach the Earth after the collisional event that produced it can be obtained from a measure of its cosmic ray exposure age. Prior to the collisions, the fragment may have been well below the surface of a much larger body, and as such it would have been shielded from all but the most energetic cosmic rays. However, after a collision, the exposed fragment would be subjected to cosmic ray bombardment in interplanetary space. A detailed analysis of meteorite samples allows these exposure ages to be measured.

In the case of one common class of meteorites called the ordinary chondrites, the cosmic ray exposure ages are typically less than 20 million years and the samples show little evidence of having been exposed to high pressure, or “shocking”. Prior to the application of chaos theory to the origin of the Kirkwood gaps, there was no plausible mechanism that could explain delivery to the Earth within the exposure age constraints and without shocking. However, small increments in the velocity of the fragments as a result of the initial collision or orbital changes due to the **Yarkovsky effect** could easily cause them to enter a chaotic zone near a given resonance. (See Meteorites.)

Numerical integrations of such orbits near the 3:1 resonance showed that it was possible for them to achieve eccentricities large enough for them to cross the orbit of the Earth. This result complemented previous research that had established that this part of the asteroid belt was a source region for the ordinary chondrites. In order to obtain agreement between theory and observations, other perturbations such as the **Yarkovsky effect** need to be included.

### 5.3. Comets

Typical cometary orbits have large eccentricities and therefore planet-crossing trajectories are commonplace. Many comets are thought to originate in the Oort cloud at several tens of thousands of arbitrary units from the Sun; another reservoir of comets, known as the Kuiper belt, exists just beyond the orbit of Neptune. Those that have been detected from the Earth are classified as long period (most of which have made single apparitions and have periods >200 y), Halley-type (with orbital periods of 20–200 y) or Jupiter family (with orbital periods <20 y). All comets with orbital periods of less than ~10^3 y have experienced a close approach to Jupiter or one of the other...
giant planets. By their very nature, the orbits of comets are chaotic, since the outcome of any planetary encounter will be sensitively dependent on the initial conditions.

Studies of the orbital evolution of the short-period comet D/Lexell highlight the possible effects of close approaches. A numerical integration has shown that prior to 1767 it was a short-period comet with a semimajor axis of 4.4 AU and an eccentricity of 0.35. In 1767 and 1779, it suffered close approaches to Jupiter. The first encounter placed it on a trajectory which brought it into the inner solar system and close (0.0146 AU) to the Earth, leading to its discovery and its only apparition in 1770, whereas the second was at a distance of ~ 3 Jovian radii. This changed its semimajor axis to 45 AU with an eccentricity of 0.88.

A more recent example is the orbital history of comet D/Shoemaker–Levy 9 prior to its spectacular collision with Jupiter in 1994. Orbit computations suggest that the comet was captured by Jupiter at some time during a 9-year interval centered on 1929. Prior to its capture, it is likely that it was orbiting in the outer part of the asteroid belt close to the 3:2 resonance with Jupiter or between Jupiter and Saturn close to the 2:3 resonance with Jupiter. However, the chaotic nature of its orbit means that it is impossible to derive a more accurate history unless prediscovery images of the comet are obtained. (See Physics And Chemistry Of Comets; Cometary Dynamics.)

### 5.4. Small Satellites and Rings

Chaos is also involved in the dynamics of a satellite embedded in a planetary ring system. The processes differ from those discussed in Section 3.1, because there is a near-continuous supply of ring material and direct scattering by the perturber is now important. In this case, the key quantity is the Hill’s sphere of the satellite. Ring particles on near-circular orbits passing close to the satellite exhibit chaotic behavior due to the significant perturbations they receive at close approach. This causes them to collide with surrounding ring material, thereby forming a gap. Studies have shown that for small satellites, the expression for the width of the cleared gap is

\[
W \approx 0.44 \left( \frac{m_2}{m_1} \right)^{2/7} a
\]  

(3.45)

where \( m_2 \) and \( a \) are the mass and semimajor axis of the satellite, respectively, and \( m_1 \) is the mass of the planet. Thus, an icy satellite with a radius of 10 km and a density of 1 g/cm³ orbiting in Saturn’s A ring at a radial distance of 135,000 km would create a gap approximately 140 km wide.

Since such a gap is wider than the satellite that creates it, this provides an indirect method for the detection of small satellites in ring systems. There are two prominent gaps in Saturn’s A ring: the ~35-km-wide Keeler gap at 136,530 km and the 325-km-wide Encke gap at 133,570 km. The predicted radii of the icy satellites required to produce these gaps are ~2.5 and ~24 km, respectively. In 1991, an analysis of Voyager images by M. Showalter revealed a small satellite, Pan, with a radius of ~10 km orbiting in the Encke gap. In 2005, the moon Daphnis of radius ~3–4 km was discovered in the Keeler gap by the Cassini spacecraft. Voyager 2 images of the dust rings of Uranus show pronounced gaps at certain locations. Although most of the proposed shepherding satellites needed to maintain the narrow rings have yet to be discovered, these gaps may provide indirect evidence of their orbital locations.

### 6. LONG-TERM STABILITY OF PLANETARY ORBITS

#### 6.1. The N-Body Problem

The entire solar system can be approximated by a system of eight planets orbiting the Sun. In a center-of-mass frame, the vector equation of motion for planet \( i \) moving under the Newtonian gravitational effect of the Sun and the eight planets is given by

\[
\dot{\mathbf{r}} = G \sum_{j=0}^{8} \frac{m_j \mathbf{r}_{ij} - \mathbf{r}_i}{r_{ij}^3} (j \neq i) \tag{3.46}
\]

where \( \mathbf{r}_i \) and \( m_i \) are the position vector and mass of planet \( i \) (\( i = 1, 2, \ldots, 8 \)), respectively; \( r_{ij} = r_j - r_i \); and the subscript 0 refers to the Sun. These are the equations of the \( N \)-body problem for the case where \( N = 9 \), and although they have a surprisingly simple form, they have no general, analytical solution. However, as in the case of the three-body problem, it is possible to tackle this problem mathematically by making some simplifying assumptions.

Provided the eccentricities and inclinations of the \( N \) bodies are small and there are no resonant interactions between the planets, it is possible to derive an analytical solution that describes the evolution of all the eccentricities, inclinations, perihelia, and nodes of the planets. This solution, called Laplace—Lagrange secular perturbation theory, gives no information about the longitudinal location of the planets, yet it demonstrates that there are long-period variations in the planetary orbital elements that arise from mutual perturbations. The secular periods involved are typically tens or hundreds of thousands of years, and the evolving system always exhibits regular behavior. In the case of Earth’s orbit, such periods may be correlated with climatic change, and large variations in the eccentricity of Mars are thought to have had important consequences for its climate.

In the early nineteenth century, Pierre-Simon de Laplace claimed that he had demonstrated the long-term
stability of the solar system using the results of his secular perturbation theory. Although the actual planetary system violates some of the assumed conditions (e.g., Jupiter and Saturn are close to a 5:2 resonance), the Laplace—Lagrange theory can be modified to account for some of these effects. However, such analytical approaches always involve the neglect of potentially important interactions between planets. The problem becomes even more difficult when the possibility of near-resonances between some of the secular periods of the system is considered. However, nowadays it is always possible to carry out numerical investigations of long-term stability.

6.2. Stability of the Solar System

Numerical integrations show that the orbits of the planets are chaotic with a timescale for exponential divergence of 4 or 5 million years. The effect is most apparent in the orbits of the inner planets. Despite this chaos, gross changes in planetary orbits are unlikely on astrophysically important timescales. But in 1% of systems integrated forward for 5 billion years, Mercury and Venus suffer a close approach, which can lead to Mercury colliding with another planet or the Sun or being ejected from the solar system.

The chaotic divergence seen in all long-term integrations implies that the accuracy of the deterministic equations of celestial mechanics to predict the future positions of the planets will always be limited by the accuracy with which their orbits can be measured. For example, if the position of Earth along its orbit is known to within 1 cm today, then the exponential propagation of errors that is characteristic of chaotic motion implies that we have no knowledge of Earth’s orbital position 200 million years in the future.

The situation is even less predictable when the gravitational influence of smaller bodies is accounted for. Asteroids exert small perturbations on the orbits of the major planets. These perturbations can be accounted for and do not adversely affect the precision to which planetary orbits can be simulated on timescales of tens of millions of years. However, unlike the major planets, asteroids suffer close approaches to one another. Close approaches between the two largest and most massive asteroids, Ceres and Vesta, led to exponential growth in uncertainty for backward integrations of planetary orbits with doubling times of <10^6 years more than 50—60 million years ago.

Pluto’s orbit is chaotic, partly as a result of its 3:2 resonance with the planet Neptune, although the perturbing effects of other planets are also important. Despite the fact that the timescale for exponential divergence of nearby trajectories (the inverse of the Lyapunov exponent) is about 20 million years, no study has shown evidence for Pluto leaving the resonance.

7. DISSIPATIVE FORCES AND THE ORBITS OF SMALL BODIES

The previous sections describe the gravitational interactions between the Sun, planets, and moons. Solar radiation has been ignored, but this is an important force for small particles in the solar system. Three effects can be distinguished: (1) the radiation pressure, which pushes particles primarily outward from the Sun (micron-sized dust); (2) the Poynting—Robertson drag, which causes centimeter-sized particles to spiral inward toward the Sun; and (3) the Yarkovsky effect, which changes the orbits of meter- to kilometer-sized objects owing to uneven temperature distributions at their surfaces. These effects can be significant as they can lead to secular changes in orbital angular momentum and energy. Each of these effects is discussed in the next three subsections and then the effect of gas drag is examined. In the final subsection, the influence of tidal interactions is discussed; this effect (in contrast to the other dissipative effects described in this section) is most important for larger bodies such as moons and planets. (See Solar System Dust.)

7.1. Radiation Force (Micron-Sized Particles)

The Sun’s radiation exerts a force, \( F_r \), on all other bodies of the solar system. The magnitude of this force is

\[
F_r = \frac{LA}{4\pi cr^2} Q_{pr},
\]

(3.47)

where \( A \) is the particle’s geometric cross-section, \( L \) is the solar luminosity, \( c \) is the speed of light, \( r \) is the heliocentric distance, and \( Q_{pr} \) is the radiation pressure coefficient, which is equal to unity for a perfectly absorbing particle and is of order unity unless the particle is small compared to the wavelength of the radiation. The parameter \( \beta \) is defined as the ratio between the forces due to the radiation pressure and the Sun’s gravity:

\[
\beta \equiv \frac{F_r}{F_g} = 5.7 \times 10^{-5} \frac{Q_{pr}}{\rho R},
\]

(3.48)

where the radius, \( R \), and the density, \( \rho \), of the particle are in c.g.s. units. Note that \( \beta \) is independent of heliocentric distance and that the solar radiation force is important only for micron- and submicron-sized particles. Using the parameter \( \beta \), a more general expression for the effective gravitational attraction can be written:

\[
F_{geff} = \frac{-GM}{r^2} (1 - \beta),
\]

(3.49)

that is, the small particles “see” a Sun of mass \((1 - \beta)M\). It is clear that small particles with \( \beta > 1 \) are repelled by the Sun and thus quickly escape the solar system, unless they
are gravitationally bound to one of the planets. Dust that is released from bodies traveling on circular orbits at the Keplerian velocity is ejected from the solar system if $\beta > 0.5$.

The importance of solar radiation pressure can be seen, for example, in comets. Cometary dust is pushed in the antisolar direction by the Sun’s radiation pressure. The dust tails are curved because the particles’ velocity decreases as they move farther from the Sun, due to conservation of angular momentum. (See Cometary Dynamics; Physics And Chemistry Of Comets.)

### 7.2. Poynting—Robertson Drag
( centimeter-sized grains)

A small particle in orbit around the Sun absorbs solar radiation and reradiates the energy isotropically in its own frame. The particle thereby preferentially radiates (and loses momentum) in the forward direction in the inertial frame of the Sun. This leads to a decrease in the particle’s energy and angular momentum and causes dust in bound orbits to spiral sunward. This effect is called the Poynting—Robertson drag.

The net force on a rapidly rotating dust grain is given by

$$F_{\text{rad}} = \frac{1}{4\pi cr^2} \left[ \frac{L Q_{\text{gr}} A}{c} \right] \left( 1 - \frac{2u v}{c} - \frac{v \theta}{c} \right). \tag{3.50}$$

The first term in Eqn (3.50) is that due to radiation pressure and the second and third terms (those involving the velocity of the particle) represent the Poynting—Robertson drag.

From this discussion, it is clear that small-sized dust grains in the interplanetary medium are removed: (sub) micron-sized grains are blown out of the solar system, whereas larger particles spiral inward toward the Sun. Typical decay times (in years) for circular orbits are given by

$$\tau_{\text{P-R}} \approx 400 \frac{r^2}{\beta}, \tag{3.51}$$

with the distance $r$ in AU.

Particles that produce the bulk of the zodiacal light (at infrared and visible wavelengths) are between 20 and 200 $\mu$m, so their lifetimes in the Earth orbit are on the order of $10^5$ y, which is much less than the age of the solar system. Sources for the dust grains are comets as well as the asteroid belt, where numerous collisions occur between countless small asteroids.

### 7.3. Yarkovsky Effect (meter-sized objects)

Consider a rotating body heated by the Sun. Because of thermal inertia, the afternoon hemisphere is typically warmer than the morning hemisphere, by an amount $\Delta T \ll T$. Let us assume that the temperature of the morning hemisphere is $T - \Delta T/2$, and that of the evening hemisphere is $T + \Delta T/2$. The radiation reaction on a surface element $dA$, normal to its surface, is $dF = 2\pi T^4 dA/3c$. For a spherical particle of radius $R$, the Yarkovsky force in the orbit plane due to the excess emission on the evening side is

$$F_Y = \frac{8}{3} \pi R^2 \frac{\sigma T^4}{c^2} \Delta T \cos \psi, \tag{3.52}$$

where $\sigma$ is the Stefan–Boltzmann constant and $\psi$ is the particle’s obliquity, that is, the angle between its rotation axis and orbit pole. The reaction force is positive for an object that rotates in the prograde direction, $0 < \psi < 90^\circ$, and negative for an object with retrograde rotation, $90^\circ < \psi < 180^\circ$. In the latter case, the force enhances the Poynting–Robertson drag.

The Yarkovsky force is important for bodies ranging in size from meters to several kilometers. Asymmetric outgassing from comets produces a nongravitational force similar in form to the Yarkovsky force. (See Cometary Dynamics.)

### 7.4. Gas Drag

Although interplanetary space generally can be considered an excellent vacuum, there are certain situations in planetary dynamics where interactions with gas can significantly alter the motion of solid particles. Two prominent examples of this process are planetesimal interactions with the gaseous component of the protoplanetary disk during the formation of the solar system and orbital decay of ring particles as a result of drag caused by extended planetary atmospheres.

In the laboratory, gas drag slows solid objects down until their positions remain fixed relative to the gas. In the planetary dynamics case, the situation is more complicated. For example, a body on a circular orbit about a planet loses mechanical energy as a result of drag with a static atmosphere, but this energy loss leads to a decrease in the semimajor axis of the orbit, which implies that the body actually speeds up! Other, more intuitive effects of gas drag are the damping of eccentricities and, in the case where there is a preferred plane in which the gas density is the greatest, the damping of inclinations relative to this plane.

Objects whose dimensions are larger than the mean free path of the gas molecules experience Stokes drag,

$$F_D = -C_D \rho v^2, \tag{3.53}$$

where $v$ is the relative velocity of the gas and the body, $\rho$ is the gas density, $A$ is the projected surface area of the body, and $C_D$ is a dimensionless drag coefficient, which is of order unity unless the Reynolds number is very small. Smaller bodies are subject to Epstein drag,

$$F_D = -A \rho v v', \tag{3.54}$$
where \( \nu' \) is the mean thermal velocity of the gas. Note that as the drag force is proportional to surface area and the gravitational force is proportional to volume (for constant particle density), gas drag is usually most important for the dynamics of small bodies.

The gaseous component of the protoplanetary disk in the early solar system is thought to have been partially supported against the gravity of the Sun by a negative pressure gradient in the radial direction. Thus, less centripetal force was required to complete the balance, and consequently the gas orbited less rapidly than the Keplerian velocity. The “effective gravity” felt by the gas is

\[
\text{g}_{\text{eff}} = -\frac{G M_S}{r^2} - \frac{1}{\rho} \frac{\text{d}p}{\text{d}r}. \tag{3.55}
\]

To maintain a circular orbit, the effective gravity must be balanced by centripetal acceleration, \( m^2 \). For estimated protoplanetary disk parameters, the gas rotated \( \sim 0.5\% \) slower than the Keplerian speed.

Large particles moving at (nearly) the Keplerian speed thus encountered a headwind, which removed part of their angular momentum and caused them to spiral inward toward the Sun. Inward drift was greatest for midsized particles, which have large ratios of surface area to mass yet still orbit with nearly Keplerian velocities. The effect diminishes for very small particles, which are so strongly coupled to the gas that the headwind they encounter is very slow. Peak rates of inward drift occur for particles that collide with roughly their own mass of gas in one orbital period. Meter-sized bodies in the inner solar nebula drift inward at a rate of up to \( 10^6 \) km/y! Thus, the material that survives to form the planets must complete the transition from centimeter to kilometer size rather quickly, unless it is confined to a thin dust-dominated subdisk in which the gas is dragged along at essentially the Keplerian velocity.

Drag induced by a planetary atmosphere is even more effective for a given density, as atmospheres are almost entirely pressure supported, so the relative velocity between the gas and particles is high. As atmospheric densities drop rapidly with height, particles decay slowly at first, but as they reach lower altitudes, their decay can become very rapid. Gas drag is the principal cause of orbital decay of artificial satellites in low Earth orbit.

7.5. Tidal Interactions and Planetary Satellites

Tidal forces are important to many aspects of the structure and evolution of planetary bodies:

1. On short timescales, temporal variations in tides (as seen in the frame rotating with the body under consideration) cause stresses that can move fluids with respect to more rigid parts of the planet (e.g. the familiar ocean tides) and even cause seismic disturbances (although the evidence that the Moon causes some earthquakes is weak and disputable, it is clear that the tides raised by the Earth are a major cause of moonquakes).

2. On long timescales, tides cause changes in the orbital and spin properties of planets and moons. Tides also determine the equilibrium shape of a body located near any massive body; note that many materials that behave as solids on human timescales are effectively fluids on very long geological timescales (e.g. Earth’s mantle).

The gravitational attraction of the Moon and Earth on each other causes tidal bulges that rise in a direction close to the line joining the centers of the two bodies. Particles on the nearside of the body experience gravitational forces from the other body that exceed the centrifugal force of the mutual orbit, whereas particles on the far side experience gravitational forces that are less than the centripetal forces needed for motion in a circle. It is the gradient of the gravitational force across the body that gives rise to the double tidal bulge.

The Moon spins once per orbit, so the same face of the Moon always points toward the Earth and the Moon is always elongated in that direction. Earth, however, rotates much faster than the Earth–Moon orbital period. Thus, different parts of the Earth point toward the Moon and are tidally stretched. If the Earth was perfectly fluid, the tidal bulges would respond immediately to the varying force, but the finite response time of Earth’s figure causes the tidal bulge to lag behind, at the point on the Earth where the Moon was overhead slightly earlier. Since the Earth rotates faster than the Moon orbits, this “tidal lag” on the Earth leads the position of the Moon in inertial space. As a result, the tidal bulge of the Earth accelerates the Moon in its orbit. This causes the Moon to slowly spiral outward. The Moon slows down Earth’s rotation by pulling back on the tidal bulge, so the angular momentum in the system is conserved. This same phenomenon has caused most, if not all, major moons to be in synchronous rotation: the rotation and orbital periods of these bodies are equal. In the case of the Pluto–Charon system, the entire system is locked in a synchronous rotation and revolution of 6.4 days. Satellites in retrograde orbits (e.g. Triton) or satellites whose orbital periods are less than the planet’s rotation period (e.g. Phobos) spiral inward toward the planet as a result of tidal forces.

Mercury orbits the Sun in 88 days and rotates around its axis in 59 days, a 3:2 spin-orbit resonance. Hence, at every perihelion, one of two locations is pointed at the Sun: the subsolar longitude is either 0° or 180°. This configuration is stable because Mercury has both a large orbital eccentricity and a significant permanent deformation that is aligned with the solar direction at perihelion. Indeed, at 0° longitude,
there is a large impact crater, Caloris Planitia, which may be the cause of the permanent deformation.

3. Under special circumstances, strong tides can have significant effects on the physical structure of bodies. Generally, the strongest tidal forces felt by solar system bodies (other than Sun-grazing or planet-grazing comets) are those caused by planets on their closest satellites. Near a planet, tides are so strong that they rip a fluid (or weakly aggregated solid) body apart. In such a region, large moons are unstable, and even small moons, which could be held together by material strength, are unable to accrete because of tides. The boundary of this region is known as Roche's limit. Inside Roche's limit, solid material remains in the form of small bodies and rings are found instead of large moons.

The closer a moon is to a planet, the stronger is the tidal force to which it is subjected. Let us consider Roche's limit for a spherical satellite in synchronous rotation at a distance $r$ from a planet. This is the distance at which a loose particle on an equatorial subplanet point just remains gravitationally bound to the satellite. At the center of the satellite of mass $m$ and radius $R_s$, a particle would be in equilibrium and so

$$\frac{GM}{r^2} = n^2 r, \quad (3.56)$$

where $M(\gg m)$ is the mass of the planet. However, at the equator, the particle will experience (1) an excess gravitational or centrifugal force due to the planet, (2) a centrifugal force due to rotation, and (3) a gravitational force due to the satellite. If the equatorial particle is just in equilibrium, these forces will balance and

$$-\frac{d}{dr}\left(\frac{GM}{r^2}\right)R_s + n^2 r = \frac{Gm}{R^2}. \quad (3.57)$$

In this case, Roche's limit $r_{Roche}$ is given by

$$r_{Roche} = 3^{1/3}\left(\frac{\rho_{planet}}{\rho_s}\right)^{1/3} R_{planet}. \quad (3.58)$$

where $\rho_{planet}$ and $\rho_s$ are the densities of the planet and satellite, respectively, and $R_{planet}$ is the planetary radius. When a fluid moon is considered and flattening of the object due to the tidal distortion is taken into account, the correct result for a liquid moon (no internal strength) is

$$r_{Roche} = 2.456\left(\frac{\rho_{planet}}{\rho_s}\right)^{1/3} R_{planet}. \quad (3.59)$$

Most bodies have some internal strength, which allows bodies with sizes $\leq 100$ km to be stable somewhat inside Roche's limit. Mars's satellite Phobos is well inside Roche's limit; it is subjected to a tidal force equivalent to that in Saturn’s B ring.

4. Internal stresses caused by variations in tides on a body in an eccentric orbit or not rotating synchronously with its orbital period can result in significant tidal heating of some bodies, most notably in Jupiter’s moon Io. If no other forces were present, this would lead to a decay of Io’s orbital eccentricity. In analogy to the Earth–Moon system, the tide raised on Jupiter by Io will cause Io to spiral outward and its orbital eccentricity to decrease. However, there exists a 2:1 mean-motion resonant lock between Io and Europa. Io passes on some of the orbital energy and angular momentum that it receives from Jupiter to Europa, and Io’s eccentricity is increased as a result of this transfer. This forced eccentricity maintains a high tidal dissipation rate and large internal heating in Io, which displays itself in the form of active volcanism. [See Io]

7.6. Tidal Evolution and Resonances

Objects in prograde orbits that lie outside the synchro-nous orbit can evolve outward at different rates, so there may have been occasions in the past when pairs of satellites evolved toward an orbit—orbit resonance. The outcome of such a resonant encounter depends on the direction from which the resonance is approached. For example, capture into resonance is possible only if the satellites are approaching one another. If the satellites are receding, then capture is not possible but the resonance passage can lead to an increase in the eccentricity and inclination. In certain circumstances, it is possible to study the process using a simple mathematical model. However, this model breaks down near the chaotic separatrices of resonances and in regions of resonance overlap.

It is likely that the major satellites of Jupiter, Saturn, and Uranus have undergone significant tidal evolution and that the numerous resonances in the Jovian and Saturnian systems are a result of resonant capture. The absence of orbit—orbit resonances among the major moons in the Uranian system is thought to be related to the fact that the oblateness of Uranus is significantly less than that of Jupiter or Saturn. In these circumstances, there can be large chaotic regions associated with resonances and stable capture may be impossible. However, temporary capture into some resonances can produce large changes in eccentricity or inclination. For example, the Uranian satellite Miranda has an anomalously large inclination of 4°, which is thought to be the result of a chaotic passage through the 3:1 resonance with Umbriel at some time in its orbital history. Under tidal forces, a satellite’s eccentricity is reduced on a shorter timescale than its inclination, and Miranda’s current inclination agrees with estimates derived from a chaotic evol-u-
8. Chaotic Rotation

8.1. Spin—Orbit Resonance

One of the dissipative effects of the tide raised on a natural satellite by a planet is to cause the satellite to evolve toward a state of synchronous rotation, where the rotational period of the satellite is approximately equal to its orbital period. Such a state is one example of a spin—orbit resonance, where the ratio of the spin period to the orbital period is close to a rational number. The time needed for a near-spherical satellite to achieve this state depends on its mass and orbital distance from the planet. Small, distant satellites take a longer time to evolve into the synchronous state than large satellites that orbit close to the planet. Observations by spacecraft and ground-based instruments suggest that most regular satellites are in the synchronous spin state, in agreement with theoretical predictions.

The lowest energy state of a satellite in synchronous rotation has the moon’s longest axis pointing in the approximate direction of the planet—satellite line. Let $\theta$ denote the angle between the long axis and the planet—satellite line in the planar case of a rotating satellite (Figure 3.21). The variation of $\theta$ with time can be described by equating the time variation of the rotational angular momentum with the restoring torque. The resulting differential equation is

$$\ddot{\theta} + \frac{\omega_0^2}{2r^3} \sin 2(\theta - f) = 0,$$

(3.60)

where $\omega_0$ is a function of the principal moments of inertia of the satellite, $r$ is the radial distance of the satellite from the planet, and $f$ is the true anomaly (or angular position) of the satellite in its orbit. The radius is an implicit function of time and is related to the true anomaly by the equation

$$r = \frac{a(1 - e^2)}{1 + e \cos f},$$

(3.61)

where $a$ and $e$ are the constant semimajor axis and the eccentricity of the satellite’s orbit, respectively, and the orbit is taken to be fixed in space.

Equation (3.60) defines a deterministic system where the initial values of $\theta$ and $\dot{\theta}$ determine the subsequent rotation of the satellite. Since $\theta$ and $\dot{\theta}$ define a unique spin position of the satellite, a surface-of-section plot of $(\theta, \dot{\theta})$ once every orbital period, say at every periapse passage, produces a picture of the phase space. Figure 3.22 shows the resulting surface-of-section plots for a number of starting conditions using $e = 0.1$ and $\omega_0 = 0.2$. The chosen values of $\omega_0$ and $e$ are larger than those that are typical for natural satellites, but they serve to illustrate the structure of the surface of section; large values of $e$ are unusual since tidal forces also act to damp eccentricity. The surface of section shows large, regular regions surrounding narrow islands associated with the 1:2, 1:1, 3:2, 2:1, and 5:2 spin—orbit resonances at $\dot{\theta} = 0.5$, 1, 1.5, 2, and 2.5, respectively. The largest island is associated with the strong 1:1 resonance and, although other spin states are possible, most regular satellites, including the Earth’s Moon, are observed to be in this state. Note the presence of diffuse collections of points associated with small chaotic regions at the separatrices of the resonances. These are particularly obvious at the 1:1 spin—orbit state at $\theta = \pi/2$, $\dot{\theta} = 1$. Although this is a completely different dynamical system compared to the circular restricted three-body problem,
there are distinct similarities in the types of behavior visible in Figure 3.22 and parts of Figures 3.14 and 3.15.

In the case of near-spherical objects, it is possible to investigate the dynamics of spin—orbit coupling using analytical techniques. The sizes of the islands shown in Figure 3.22 can be estimated by expanding the second term in Eqn (3.60) and isolating the terms that will dominate at each resonance. Using such a method, each resonance can be treated in isolation and the gravitational effects of nearby resonances can be neglected. However, if a satellite is distinctly nonspherical, \( \omega_0 \) can be large and this approximation is no longer valid. In such cases, it is necessary to investigate the motion of the satellite using numerical techniques.

### 8.2. Hyperion

Hyperion is a satellite of Saturn that has an unusual shape (Figure 3.23). It has a mean radius of 135 km, an orbital eccentricity of 0.1, a semimajor axis of 24.55 Saturn radii, and a corresponding orbital period of 21.3 days. Such a small object at this distance from Saturn has a large tidal despinning timescale, but the unusual shape implies an estimated value of \( \omega_0 = 0.89 \).

The surface of section for a single trajectory is shown in Figure 3.24 using the same scale as Figure 3.22. It is clear that there is a large chaotic zone that encompasses most of the spin—orbit resonances. The islands associated with the synchronous and other resonances survive but in a much reduced form. Although this calculation assumes that Hyperion’s spin axis remains perpendicular to its orbital plane, studies have shown that the satellite should also be undergoing a tumbling motion, such that its axis of rotation is not fixed in space.

*Voyager* observations of Hyperion indicated a spin period of 13 days, which suggested that the satellite was not in synchronous rotation. However, the standard techniques that are used to determine the period are not applicable if it varies on a timescale that is short compared with the timespan of the observations. In principle, the rotational period can be deduced from ground-based observations by looking for periodicities in plots of the brightness of the object as a function of time (the light curve of the object). The results of one such study for Hyperion are shown in Figure 3.25. Since there is no recognizable periodicity, the light curve is consistent with that of an object undergoing chaotic rotation. Hyperion is the first natural satellite that has been observed to have a chaotic spin state, and results from *Cassini* images confirm this result. Observations and numerical studies of Hyperion’s rotation in three dimensions have shown that its spin axis does not point in a fixed direction. Therefore, the satellite also undergoes a tumbling motion in addition to its chaotic rotation.

The dynamics of Hyperion’s motion is complicated by the fact that it is in a 4:3 orbit—orbit resonance with the larger Saturnian satellite Titan. Although tides act to decrease the eccentricities of satellite orbits, Hyperion’s eccentricity is maintained at 0.104 by means of the resonance. Titan effectively forces Hyperion to have this large value of \( e \) and so the apparently regular orbital motion inside the resonance results, in part, in the extent of the chaos in its rotational motion. (See Planetary Satellites.)

### 8.3. Other Satellites

Although there is no evidence that other natural satellites are undergoing chaotic rotation at the present time, it is
possible that several irregularly shaped regular satellites did experience chaotic rotation at some time in their histories. In particular, since satellites have to cross chaotic separatrices before capture into synchronous rotation can occur, they must have experienced some episode of chaotic rotation. This may also have occurred if the satellite suffered a large impact that affected its rotation. Such episodes could have induced significant internal heating and resurfacing events in some satellites. The Martian moon Phobos and the Uranian moon Miranda have been mentioned as possible candidates for this process. If this happened early in the history of the solar system, then the evidence may well have been obliterated by subsequent cratering events. (See Planetary Satellites.)

8.4. Chaotic Obliquity

The fact that a planet is not a perfect sphere means that it experiences additional perturbing effects due to the gravitational forces exerted by its satellites and the Sun, and these can cause long-term evolution in its obliquity (the angle between the planet’s equator and its orbit plane). Numerical investigations have shown that chaotic changes in obliquity are particularly common in the inner solar system. For example, it is now known that the stabilizing effect of the Moon results in a variation of $\pm 1.3^\circ$ in Earth’s obliquity around a mean value of $23.3^\circ$. Without the Moon, Earth’s obliquity would undergo large, chaotic variations. In the case of Mars, there is no stabilizing factor and the obliquity varies chaotically from $0^\circ$ to $60^\circ$ on a timescale of 50 million years. Therefore, an understanding of the long-term changes in a planet’s climate can be achieved only by an appreciation of the role of chaos in its dynamical evolution.

9. EPILOG

It is clear that nonlinear dynamics has provided us with a deeper understanding of the dynamical processes that have helped to shape the solar system. Chaotic motion is a natural consequence of even the simplest systems of three or more interacting bodies. The realization that chaos has played a fundamental role in the dynamical evolution of the solar system came about because of contemporary and complementary advances in mathematical techniques and digital computers. This coincided with an explosion in our knowledge of the solar system and its major and minor members. Understanding how a random system of planets, satellites, ring and dust particles, asteroids, and comets interacts and evolves under a variety of chaotic processes and timescales ultimately means that this knowledge can be used to trace the history and predict the fate of other planetary systems.

BIBLIOGRAPHY


Chapter 4

Planetary Impacts

Richard A.F. Grieve, Gordon R. Osinski and Livio L. Tornabene
Centre for Planetary Science and Exploration, University of Western Ontario, London, ON, Canada

1. IMPACT CRATERS

1.1. Crater Shape

On bodies that have no atmosphere, such as the Moon, even the smallest pieces of interplanetary material can produce impact craters down to micrometer-sized cavities on individual mineral grains. On larger bodies, atmospheric passage results in aerodynamic resistant forces, which decelerate incoming bodies and break up weaker ones. On Earth, for example, impacting bodies with masses below $10^4$ g can lose up to 90% of their velocity during atmospheric penetration and the resultant impact pit is only slightly larger than the projectile itself. Atmospheric effects on larger incoming masses, however, are less severe, and the body impacts with relatively undiminished velocity, producing a crater that is considerably larger than the impacting body.

The processes accompanying such events are rooted in the physics of impact, with the differences in response among the various planets largely being due to differences in the properties of the planetary bodies (e.g., surface gravity, atmospheric density, and target composition and strength). The basic shape of virtually all impact craters is a depression with an upraised rim. With increasing diameter, impact craters become proportionately shallower, with respect to their diameter, i.e. the depth to diameter ratio decreases. They also develop more complicated rims and floors, including the appearance of central topographic peaks and interior rings. It should be noted that not all impact craters are circular in plan view. For example, the rims of some terrestrial, Martian and Venusian impact craters have straight line segments, reflecting preimpact inhomogeneities and structural features in the target rocks.

There are three major subdivisions in shape: simple craters, complex craters, and impact basins. Simple impact structures have the form of a bowl-shaped depression with an upraised rim (Figure 4.1(A)). An overturned flap of ejected target materials exists on the rim, and the exposed rim, walls, and floor define the apparent crater. Observations at terrestrial impact craters reveal that a lens of brecciated target material, roughly parabolic in cross section, exists beneath the floor of this apparent crater (Figure 4.2). This breccia lens is a mixture of different materials in heterogeneous targets, with fractured blocks set in a finer-grained matrix. These are allochthonous materials, having been moved into their present position by the cratering process. Beneath the breccia lens, relatively in-place, or parautochthonous, fractured target materials define the walls and floor of what is known as the true crater (Figure 4.2). In the case of terrestrial simple craters, the depth to the base of the
breccia lens (i.e. the base of the true crater) is roughly twice the depth to the top of the breccia lens (i.e. the floor of the apparent crater). With increasing diameter, simple craters display signs of wall and rim collapse (Figure 4.1(B)), as they evolve into complex craters. The diameter at which this transition takes place varies between planetary bodies and is, to a first approximation, an inverse function of planetary gravity. Other variables, such as target strength, and possibly projectile type, and impact angle and velocity, play a role and the transition actually occurs over a small range in diameter. For example, the transition between simple and complex craters occurs in the 15–25 km diameter range on the Moon. The effect of target strength is most readily apparent on Earth, where complex craters can occur at diameters as small as 2 km in sedimentary target rocks, but do not occur until diameters of 4 km, or greater, in stronger, crystalline target rocks.

Complex craters are highly modified structures with respect to their final form, compared to simple craters. A typical complex crater is characterized by a central topographic feature (e.g. a peak, pit or some combination thereof), a broad, flat floor, and a terraced, inwardly slumped rim area (Figure 4.3(A)). Observations at terrestrial complex craters show that the flat floor consists of a sheet of impact melt rock and/or polymict breccia (Figure 4.4). The central region is structurally complex and, is most commonly occupied by the central peak, which is the topographic manifestation of a much broader and extensive volume of uplifted fractured and faulted parautochthonous rocks that originate from the target beneath the crater (Figure 4.4).

With increasing diameter, a fragmentary ring of interior peaks appears (Figure 4.3(B)), marking the beginning of the morphologic transition from craters to basins. While a single interior ring is required to define a basin, they can be subdivided further into central-peak basins, with both a peak and ring; peak ring basins (Figure 4.5), with a single ring; and multiring basins, with two or more interior rings (Figure 4.6). The transition from central-peak basins to peak-ring basins to multiring basins also represents a sequence with increasing diameter. As with the simple to complex crater transition, there is a small amount of overlap in basin shape near transition diameters.

**FIGURE 4.1** (a) Winslow Crater on Mars is an example of a fresh simple crater. It is ~1 km in diameter and large blocks ejected late in the cratering process can be seen around the rim area. The ejecta can be differentiated into continuous and discontinuous ejecta, which appear as separate “fingers” and “braids” (herringbone pattern) ~1.5–2 crater radii from the rim. (b) Noord Crater is a fresh transitional crater ~8 km in diameter. It has a relatively flat floor and shows extensive slumping in the form of blocks, where the crater wall meets the crater floor. Note that there is no sign of a central structure, typical of bona fide complex craters (Mars Reconnaissance Orbiter).

**FIGURE 4.2** Schematic cross-section of a simple crater, based on terrestrial observations, D is rim diameter and dₐ and dₜ are apparent and true depth, respectively. See text for details.
Ejected target material surrounds impact craters and can be subdivided into continuous and discontinuous ejecta facies (Figures 4.1(A) and 4.3(B)). The continuous deposits are those closest to the crater, being thickest at the rim crest. In the case of simple craters, the net effect of the ejection process is to invert the stratigraphy at the rim, which may continue into the continuous ejecta depending on target heterogeneity (i.e. the deepest materials are deposited near the rim, and the shallowest are most distal). However, as the distance from the crater rim increases, the ejecta are emplaced at higher velocities and, therefore, land with higher kinetic energies, resulting in the mixing of ejecta with local surface material. Thus, at increasing distance from the crater, the final ejecta blanket on the ground includes increasing amounts of local materials. Secondary crater fields, resulting from the impact of larger, coherent blocks and clods of ejecta, surround fresh craters and are particularly evident on bodies that lack or have thin atmospheres, such as the Moon, Mercury (Figure 4.3(B)), and Mars. On the Moon and Mercury, they are often associated with typically bright or high-albedo “rays” that define an overall radial pattern to the primary crater (See: Mercury, The Moon). Two principal processes have been suggested to explain the rays. The first is a compositional effect, where the ejecta are chemically different from the material on which it is deposited. While this most often results in rays that are brighter than the surrounding material, the reverse can also occur. The second effect is a consequence of “maturity” due to prolonged exposure to “space weathering” agents like radiation and micrometeoroid bombardment on surface materials (See Main-Belt Asteroids). Fresher material excavated by an impact and deposited in the rays is generally brighter than the more mature material of the deposition surface: however, this contrasts with the rays on Mars, which are most recognizable in thermal images. More recently, numerical simulations have suggested the rays (at least, on airless bodies) could be the result of the interaction of impact-induced shock waves and preexisting surface depressions.

Many Martian craters display examples of apparently fluidized ejecta (Figure 4.7). They have been called...
“fluidized—ejecta”, “rampart”, or “pedestal” craters, where their ejecta deposits indicate emplacement as a ground-hugging flow. Most hypotheses on the origin of these features invoke the presence of ground ice (or water), which, upon heating by impact, is incorporated into the ejecta in either liquid or vapor form. This, then, provides lubrication for the mobilized material (See: Mars: Surface and Interior).

On Venus, radar data indicate that impact craters more than 15–20 km in diameter exhibit central peaks and/or peak rings (Figure 4.8) and appear, for the most part, to be similar to complex craters and basins on the other terrestrial planets. Many of the craters smaller than 15 km, however, have rugged, multiple floors or occur as crater clusters (See: Venus: Surface and Interior). This is attributed to the effects of the dense atmosphere of Venus (surface pressure of ~90 bar), which effectively crushes and breaks up smaller impacting bodies, so that they result in clusters of relatively shallow craters. Also due to atmospheric effects, there is a deficit in the number of expected craters with diameters up to 35 km, and there are no craters smaller than 3 km in diameter on Venus. In principle, this atmospheric effect on small impacting bodies occurs on Earth. Due to its less dense atmosphere, however, the fragments remain relatively close together in the terrestrial case and the net effect is similar to the impact of a coherent impacting body.

In many cases, craters on Venus have ejecta deposits out to greater distances than expected from simple ballistic emplacement and the distal deposits are clearly lobate (Figure 4.8). These deposits likely owe their origin to entrainment by the dense atmosphere and/or the high portion of impact melt that would be produced on a high gravity and high surface temperature planet, such as Venus. Another unusual feature on Venus is radar-dark zones surrounding some craters that can extend three to four

FIGURE 4.5 (a) The 200 km diameter peak ring basin Lowell on Mars (Mars Odyssey). An extensive deposit interior to the inner ring is likely dunes or related peri-glacial (near-surface rock—ice interactions) that occurred well after the formation of the basin. (b) The approximately 290 km peak ring basin Rachmaninoff on Mercury. The dark, smooth fractured deposits interior to the inner ring of this basin may represent extensive deposits rich in impact melt that were emplaced during crater formation (Messenger).

FIGURE 4.6 At a diameter of ~1000 km, as defined by the outer ring, the Cordillera mountains, Orientale is the youngest and best-preserved lunar multiring basin (Lunar Orbiter Images).
crater diameters from the crater center (Figure 4.8). They are believed to be due to the modification of the surface roughness by the atmospheric shock wave produced by the impacting body. Small crater clusters have dark haloes and dark circular areas with no central crater form have been observed. In these latter cases, the impacting body did not survive atmospheric passage, but the accompanying atmospheric shock wave had sufficient energy to interact with the surface to create a dark, radar-smooth area (See Venus: Surface and Interior). The situation is somewhat analogous to the 1908 Tunguska event, when a relatively small body exploded over Siberia at an altitude of ~ 10 km, and the resultant atmospheric pressure wave leveled some 2000 km² of forest. Most recently, on 15 February, 2013, a meteoritic body, with an estimated original mass of 10,000 t, entered the Earth’s atmosphere on a shallow trajectory over Russia. It exploded in an air burst at a height of 23.3 km. Damage to infrastructure in the nearby city of Chelyabinsk due to the atmospheric shock wave resulted in the injury of some 1500 people but no fatalities. It is believed to have been the largest meteoritic object to enter the Earth’s atmosphere since the 1908 Tunguska event.

Remarkable ring structures occur on the Galilean satellites of Jupiter, Callisto, and Ganymede (See: Ganymede and Callisto). The largest is the 4000-km feature Valhalla on Callisto (Figure 4.9), which consists of a bright central area up to 800 km in diameter, surrounded by a darker terrain with bright ridges 20–30 km apart. This zone is about 300 km wide and gives way to an outer zone with graben or rift-like features 50–100 km apart. These (very) multiring basins are generally considered to be of impact origin, but with the actual impact crater confined to the central area. In one working hypothesis, the exterior rings are formed as a result of the original crater puncturing the outer, strong shell, or lithosphere, of these bodies. This permitted the weaker, underlying layer, the asthenosphere, to flow toward the crater, setting up stresses that led to fracturing and the formation of circumscribing scarps and graben.

On Callisto and Ganymede, there is also a unique class of impact craters that no longer have an obvious crater form but appear as bright, or high-albedo, spots on the surfaces
of these bodies. These are known as palimpsests and are believed to have begun as complex craters but have had their topography relaxed by the slow, viscous creep of the target’s icy crust over time. Palimpsests are old impact features and may have been formed when the icy satellites were young and relatively warm, with a thin crust possibly incapable of retaining significant topography.

Impact craters on icy satellites display a wide range of morphologies, some of which have no counterpart on rocky bodies. On these icy satellites, most craters larger than 25 km have a central pit or central dome, rather than a central peak. Pit and dome craters are shallower than other craters of comparable size, and it has been suggested that the pits are due to the formation of slushy or fluid material by impact melting and the domes are due to uplift of the centers of the craters as a result of layers in the crust with different mechanical properties. The fact that some craters on these icy bodies are anomalous has been ascribed to a velocity effect, as higher impact velocities result in greater melting of the target, or to changes in the mechanical behavior of the crust and its response to impact with time. Interpretations of the origin of the various anomalous crater forms on the icy satellites, however, are generally not well constrained.

The 2011–2012 Dawn mission to the asteroid Vesta imaged the largest complex central peak crater in the solar system (See: Main Belt Asteroids). It is the 505 km diameter Rheasilvia structure, which is centered on the south pole. The central peak rises some 23 km above the crater floor, making it the tallest mountain in the solar system. The fact that such a large impact crater has a complex crater form, as opposed to ringed basin form as observed on major planetary bodies, is due to the low gravity of Vesta and serves to dramatically illustrate the effect of the variation in planetary gravity on final crater form with crater size. The Rheasilvia impact is believed to have excavated some 1% of Vesta’s mass and is the most likely source of the Howardite–Eucrite–Diogenite (HED) group of differentiated meteorites (See: Meteorites).

### 1.2. Crater Dimensions

The depth-diameter relations for craters on the terrestrial or silicate planets are given in Table 4.1. Relations are in the form \( d = aD^b \) where \( d \) is apparent depth, \( D \) is rim crest diameter, and units are in kilometers. Other relations involving parameters such as rim height, rim width, central peak diameter, and peak height can be found in the literature. Due to low rate of crater-modifying process, such as erosion, and the abundance of high-resolution data from the Apollo missions, morphometric relations for fresh impact craters were well defined for the Moon. Recent planetary missions have produced laser altimetry data and have, for example, resulted in digital terrain models of all lunar craters with \( D > 30 \) km.

Simple craters have similar apparent depth–diameter relationships on all the terrestrial planets (Table 4.1). At first glance, terrestrial craters appear to be shallower than their planetary counterparts. Compared to the other terrestrial planets, erosion is most severe on Earth, and crater rims and floors are rapidly affected by erosion and subsequent deposition, respectively. Few terrestrial craters have well-preserved rims, and it is common to measure terrestrial crater depths with respect to the ground surface, which is known and is assumed to erode more slowly. In the case of other planetary bodies, depths have been measured most often by the shadow that the rim casts on the crater floor, although some recent planetary missions have produced laser or stereo-derived altimetry data. That is, the topographic measure is a relative one between the rim crests and the floor. Thus, the measurements of depth for Earth and for other planetary bodies differ. For the very few cases in which the rim is well preserved in terrestrial craters, depths from the top of the rim to the crater floor are comparable to those of similar-sized simple craters on the other terrestrial planets.

Unlike simple craters, the depths of complex craters with respect to their diameters do vary between the terrestrial planets (Table 4.1). While the sense of variation is that increasing planetary gravity shallows final crater depths, this is not a strict relationship. For example, Martian complex craters are shallower than equivalent-sized Mercurian craters (Table 4.1), even though the

<table>
<thead>
<tr>
<th>Planetary body</th>
<th>Exponent ( b )</th>
<th>Coefficient ( a )</th>
<th>Gravity (cm (^{-2} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Simple Craters</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moon</td>
<td>1.010</td>
<td>0.196</td>
<td>162</td>
</tr>
<tr>
<td>Mars</td>
<td>1.019</td>
<td>0.204</td>
<td>372</td>
</tr>
<tr>
<td>Mercury</td>
<td>0.98</td>
<td>0.18</td>
<td>378</td>
</tr>
<tr>
<td>Earth</td>
<td>1.06</td>
<td>0.13</td>
<td>981</td>
</tr>
<tr>
<td><strong>Complex Central Peak Craters</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moon</td>
<td>0.301</td>
<td>1.044</td>
<td>162</td>
</tr>
<tr>
<td>Mars</td>
<td>0.25</td>
<td>0.53</td>
<td>372</td>
</tr>
<tr>
<td>Mercury</td>
<td>0.415</td>
<td>0.492</td>
<td>378</td>
</tr>
<tr>
<td>Venus</td>
<td>0.30</td>
<td>0.40</td>
<td>891</td>
</tr>
<tr>
<td>Sedimentary</td>
<td>0.12</td>
<td>0.30</td>
<td>981</td>
</tr>
<tr>
<td>Crystalline</td>
<td>0.15</td>
<td>0.43</td>
<td>981</td>
</tr>
</tbody>
</table>
surface gravities of the two planets are very similar. This is probably a function of differences between target materials, with the trapped volatiles and relatively abundant sedimentary deposits making Mars’ surface, in general, a weaker target. Mars has also evidence of wind and water processes, which will reduce crater-related topography by erosion and sedimentary infilling. The secondary effect of target strength is also well illustrated by the observation that terrestrial complex craters in sedimentary targets are shallower than those in crystalline targets (Table 4.1). Target effects are also apparent in the recent Lunar Orbiter Laser Altimeter (LOLA) topographic data, with “young” complex craters in the lunar highlands being generally deeper and having higher central peaks than equivalent sized complex craters on the lunar mare.

Data from the Galileo mission indicates that depth—diameter relationships for craters on the icy satellites Callisto, Europa, and Ganymede have the same general trends as those on the rocky terrestrial planets. Interestingly, the depth—diameter relationship for simple craters is equivalent to that on the terrestrial planets. Although the surface gravities of these icy satellites is only 13–14% of that of the Earth, the transition diameter to complex crater forms occurs at ~3 km, similar to that on the Earth. This may be a reflection of the extreme differences in material properties between icy and rocky worlds. There are also inflections and changes in the slopes of the depth—diameter relationships for the complex craters, with a progressive reduction in absolute depth at diameters larger than the inflection diameter. These anomalous characteristics of the depth—diameter relationship have been attributed to changes in the physical behavior of the crust with depth and the presence of subsurface oceans (See Europa; Ganymede and Callisto).

2. IMPACT PROCESSES

The extremely brief timescales and extremely high energies, velocities, pressures, and temperatures that accompany impact are not encountered, as a group, in other geologic processes and make studying impact processes inherently difficult. Small-scale impacts can be produced in the laboratory by firing projectiles at high velocity (generally below about 8 km/s) at various targets. Some insights can also be gained from observations of high-energy, including nuclear, explosions. “Hydrocode” numerical models have been used to simulate impact crater formation. The planetary impact record also provides constraints on the process. The terrestrial record is an important source of ground-truth data, especially with regard to the subsurface nature and spatial relations at impact craters, and the effects of impact on rocks.

When an interplanetary body impacts a planetary surface, it transfers most of its kinetic energy to the target. The energy released in the impact of a 1 kg body with a velocity of approximately 2 km/s is equivalent to that in 1 kg of high explosives. The energy density of impacting interplanetary bodies is even higher, however, as the mean impact velocity on the terrestrial planets for asteroidal bodies ranges from ~12 km/s for Mars to over ~25 km/s for Mercury. The impact velocity of comets is even higher. Long-period comets (those with orbital periods greater than 200 years) have an average impact velocity with Earth of ~55 km/s; whereas, short-period comets have a somewhat lower average impact velocity (See Comet Populations and Cometary Dynamics).

2.1. Crater Formation

On impact, a shock wave propagates back into the impacting body and also into the target. The latter shock wave compresses and heats the target, while accelerating the target material (Figure 4.10). The direction of this acceleration is perpendicular to the shock front, which is roughly hemispherical, so material is accelerated downward and outward. As a state of stress cannot be maintained at a free surface, such as the original ground surface or the edges and rear of the impacting body, a series of secondary release or “rarefaction” waves are generated, which bring the shock-compressed materials back to ambient pressure. As the rarefaction wave interacts with the target material, it alters the direction of the material set in motion by the shock wave, changing some of the outward and downward motions in the relatively near-surface materials to outward and upward, leading to the ejection of material and the growth of a cavity. Directly below the impacting body, however, the two wave fronts are more nearly parallel, and material is still driven downward (Figure 4.10).

These motions define the cratering flow-field and a cavity grows by a combination of upward ejection and downward displacement of target materials. This “transient cavity” reaches its maximum depth before its maximum radial dimensions, but it is usually depicted in illustrations at its maximum growth in all directions (Figure 4.10). At this point, it is parabolic in cross section and, at least for the terrestrial case, has a depth-to-diameter ratio of about 1:3. As simple craters throughout the solar system appear to have similar depth-diameter ratios, the 1:3 ratio for the transient cavity can probably be treated as universal.

An asteroidal body of density 3 g/cm³ impacting crystalline target rocks at 25 km/s will generate initial shock velocities in the target faster than 20 km/s, with corresponding velocities over 10 km/s for the materials set in motion by the shock wave. The shock wave pressure decays with propagated distance and there is a decay in the strength of the cratering flow-field with distance, until it finally ceases to be able to displace target materials and the
**FIGURE 4.10** Schematic model of the formation of simple (left) and complex (right) craters for a typical impact. In the modification stage of complex craters, the arrows labeled “a” to “c” to represent a time sequence. See text for details.
formation of the transient cavity stops. Transient-cavity growth is an extremely rapid event. For example, the formation of a 2.5 km diameter transient cavity will take only about 10 s on Earth.

The cratering process is sometimes divided into stages: initial contact and compression, excavation, and modification. In reality, however, it is a continuum with different volumes of the target undergoing different stages of the cratering process at the same time (Figure 4.10). As the excavation stage draws to a close, the direction of movement of target material changes from outward to inward, as the unstable transient cavity collapses to a final topographic form more in equilibrium with gravity. This is the modification stage, with collapse ranging from landslides on the cavity walls of the smaller simple craters to complete collapse and modification of the transient cavity, involving the uplift of the center and collapse of the rim area to form central peaks and terraced, structural rims in larger complex craters (Figure 4.10).

The interior breccia lens of a typical simple crater is the result of this collapse. As the cratering flow comes to an end, the fractured and oversteepened cavity walls become unstable and collapse inward, carrying with them a lining of shocked and melted debris (Figure 4.10). The inward-collapsing walls undergo more fracturing and mixing, eventually coming to rest as the bowl-shaped breccia lens of mixed unshocked and shocked target materials that partially fill simple craters (Figure 4.10). The collapse of the walls increases the rim diameter, such that the final crater diameter in the terrestrial environment is about 20% larger than that of the transient cavity. This is offset by the shallowing of the cavity accompanying production of the breccia lens, with the final apparent crater being about half the depth of the original transient cavity (Figure 4.10). The collapse process is rapid and probably takes place on timescales comparable to those of transient-cavity formation.

Much of our understanding of complex-crater formation comes from observations at terrestrial craters, where it has been possible to trace the movement of beds to show that central peaks are the result of the uplift of rocks from depth (Figure 4.4). Shocked target rocks, analogous to those found in the floors of terrestrial simple craters, constitute the central peak at the centers of complex structures, with the central structure representing the uplifted floor of the original transient cavity. The amount of uplift determined from terrestrial data corresponds to a value of approximately one tenth of the final rim—crest diameter. Further observations at terrestrial complex craters indicate excavation is also limited to the central area and that the transient cavity diameter was about 50–65% of the diameter of the final crater. Radially beyond this, original near-surface units are preserved in the down-dropped annular floor. The rim area is a series of fault terraces, progressively stepping down to the floor (Figure 4.3(A)).

Although models for the formation of complex craters are less constrained than those of simple craters, there is a general consensus that, in their initial stages, complex craters were not unlike simple craters. At complex craters, however, the downward displacements in the transient cavity floor observed in simple craters are not locked in and the cavity floor rebounds upward (Figure 4.10). As the maximum depth of the transient cavity is reached before the cavity’s maximum diameter, it is likely that this rebound and reversal of the flow-field in the center of a complex crater occurs while the diameter of the transient cavity is still growing by excavation (Figure 4.10). With the upward movement of material in the transient cavity’s floor, the entire rim area of the transient cavity collapses downward and inward (Figure 4.10), greatly enlarging the crater’s diameter compared to that of the transient cavity. There have been a number of reconstructions of large lunar craters, in which the terraces are restored to their original, preimpact positions, resulting in estimated transient cavity diameters of about 60% of the final rim—crest diameter. It is clear that uplift and collapse, during the modification stage at complex craters, is extremely rapid and also takes place on time scales comparable to those of transient cavity formation. During the modification stage, the target materials behave as if they were temporally very weak, with the mechanical properties of “normal” fractured rock being restored on final crater formation. A number of mechanisms, including “thermal softening” and “acoustic fluidization”, by which strong vibrations cause the rock debris to behave as a fluid, have been suggested as mechanisms to produce the required weakening of the target materials.

There is less of a consensus on the formation of rings within impact basins. The most popular hypothesis for central peak basins is based on the results of modeling; namely, that the rings represent uplifted material in excess of what can be accommodated in a central peak (Figure 4.10). This may explain the occurrence of both peaks and rings in central peak basins but offers little explanation for the absence of peaks and the occurrence of only rings in peak ring and multi-ring basins. A number of analogies have been drawn with the formation of “craters” in liquids and semiconsolidated materials such as muds, where the initial uplifted peak of material has no strength and collapses completely, sometimes oscillating up and down several times. At some time in the formation of ringed basins, however, the target rocks must regain their strength, so as to preserve the interior rings. An alternative explanation is that the uplift process proceeds, as in central peak craters, but the uplifted material in the very center is essentially fluid due to impact melting. In large impact events, the depth of impact melting may reach and even exceed the depth of the transient cavity floor. When the transient cavity is uplifted in such events, the central,
melted part has no strength and, therefore, cannot form a positive topographic feature, such as a central peak. Only rings from the unmelted portion of the uplifted transient cavity floor can form some distance out from the center. This is one working hypothesis for the formation of peak ring basins on Mercury (Figure 4.3(B)), which has the largest population and population per area of peak ring structures amongst all the terrestrial planets.

One of the principal characteristics of impact events is the formation and emplacement of ejecta deposits. Recent observations suggest that ejecta may be emplaced as a multistage process. The generation of the continuous ejecta blanket occurs during the excavation stage of cratering, via conventional ballistic ejection, followed by more minor radial surface flow. Most recently, it has been hypothesized that this is followed by the late-stage emplacement of more melt-rich, ground-hugging flows, during the terminal stages of excavation and the modification stage of crater formation (Figure 4.10). Ejecta deposited in this latter is relatively minor in terms of total volume and is influenced by several factors, most importantly planetary gravity, surface temperature and the physical properties of the target rocks.

2.2. Changes in the Target Rocks

The target rocks are initially highly compressed by the passage of the shock wave, transformed into high-density phases, and then rapidly decompressed by the rarefaction wave. As a result, they do not recover fully to their pre-shock state but are of slightly lower density, with the nature of their constituent minerals changed. The collective term for these shock-induced changes in minerals and rocks is **shock metamorphism**. Shock metamorphic effects are found naturally in many lunar samples and meteorites and at terrestrial impact craters. They have also been produced in nuclear explosions and in the laboratory, through shock effects are microscopic in character. The most obvious are solid-state effects. Apart from shatter cones, all other diagnostic shock effects are microscopically indistinguishable. Such features are generally not readily distinguished from those produced by endogenic geologic processes, such as tectonism. There is, however, a unique, brittle, shock-metamorphic effect, which results in the development of unusual, striated, and horse-tailed conical fractures, known as shatter cones (Figure 4.11). Shatter cones are best developed at relatively low shock pressures (5–10 GPa) and in fine-grained, structurally homogeneous rocks, such as carbonates, quartzites, and basalts.

Apart from shatter cones, all other diagnostic shock effects are microscopic in character. The most obvious are **planar deformation features** and **diaplectic glasses**. Planar deformation features are intensely deformed, are a few micrometers wide, and are arranged in parallel sets (Figure 4.11). They are best known from the common silicate minerals, quartz and feldspar, for which shock-recovery experiments have calibrated the onset shock pressures for particular crystal orientations. They develop initially at ~10 GPa and continue to 20–30 GPa in crystalline rocks. The increasing effects of shock pressure are mirrored by changes in X-ray characteristics, indicative of the increasing breakdown of the internal crystal structure of individual minerals to smaller and smaller domains.

By shock pressures of ~30–40 GPa, quartz and feldspar are converted to diaplectic (from the Greek, “to strike”) glass in crystalline rocks. These are solid-state glasses, with no evidence of flow, that exhibit the same outline as the original crystal. For this reason, they are higher for complex geologic materials with fractures, etc. This is the pressure–volume point beyond which the shocked material no longer deforms elastically and permanent changes are recorded on recovery from shock compression.

The peak pressures generated on impact control the upper limit of shock metamorphism. These vary with the type of impacting body and target material but are, principally, a function of impact velocity, reaching into the hundreds to thousands of GPa. For example, the peak pressure generated when a stony asteroidal body impacts crystalline rock at 15 km/s is over 300 GPa, not much less than the pressure at the center of the Earth (~390 GPa). Shock metamorphism is also characterized by strain rates that are orders of magnitude higher than those produced by internal geologic processes. For example, the duration of regional metamorphism associated with tectonism on Earth is generally considered to be in the millions of years. In contrast, the peak strains associated with the formation of a crater 20 km in diameter are attained in less than a second.
sometimes referred to as thomomorphic (from the Greek, “same shape”) glasses. The variety produced from plagioclase is known as maskelynite and was originally discovered in the Shergotty meteorite in 1872. The thermodynamics of shock processes are highly irreversible, so the pressure-volume work that is done during shock compression is not fully recovered upon decompression. This residual work is manifested as waste heat and, as a result, shock pressures of $40 - 50 \text{ GPa}$ are sufficient to initiate melting in some minerals (Figure 4.11). For example, feldspar grains show incipient melting and flow at shock pressures of $\sim 45 \text{ GPa}$. It is important to note that in porous and potentially volatile-rich sedimentary rocks, the pressures required for the formation of shock features are substantially less than for dense nonporous crystalline rocks. For example, diaplectic quartz glass in sandstones begins to form at pressures as low as $\sim 5.5 \text{ GPa}$ and, between $10$ and $20 \text{ GPa}$, almost complete conversion of quartz to diaplectic glass has been observed.

Regardless of the target, melting tends initially to be mineral specific, favoring mineral phases with the highest compressibilities and to be concentrated at grain boundaries, where pressures and temperatures are enhanced by reflections and refractions of the shock wave. In detail, as the shock wave travels through multicomponent systems, such as rocks, it becomes a complex system of multiple reflected and refracted local shock fronts, which may result in the localization of particular shock metamorphic phenomena. The effects of the complex interactions of shock reflections and refractions on melting are most obvious when comparing the pressures required to melt particulate materials, such as those that make up the lunar regolith (see The Moon), and solid rock of similar composition. Shock recovery experiments indicate that intergranular melts can occur at pressures as low as $30 \text{ GPa}$ in particulate basaltic material, compared to $45 \text{ GPa}$ necessary for the onset of melting of solid basalt.

Most minerals undergo transitions to dense, high-pressure phases during shock compression. Little is known, however, about the mineralogy of the high-pressure phases, as they generally revert to their low-pressure forms during decompression. Nevertheless, metastable high-pressure phases are sometime preserved, as either high-pressure polymorphs of preexisting low-pressure phases or high-pressure assemblages due to mineral breakdown. Some known high-pressure phases, such as diamond
from carbon or stishovite from quartz (SiO₂), form during shock compression. Others, such as coesite (SiO₂), form by reversion of such minerals during pressure release. Several high-pressure phases that have been noted in shocked meteorites, however, are relatively rare at terrestrial craters. This may be due to postshock thermal effects, which are sufficiently prolonged at a large impact crater to inhibit preservation of metastable phases.

2.2.2. Melting

The waste heat trapped in shocked rocks is sufficient to result in whole-rock melting above shock pressures of ∼60 GPa for crystalline rocks and ∼30–35 GPa for sandstones. Thus, relatively close to the impact point, a volume of the target rocks is melted and can even be vaporized (Figure 4.10). Ultimately, these liquids cool to form impact melt rocks. These occur as glassy bodies in ejecta and breccias, as dikes in the crater floor, as pools and lenses within the breccia lenses of simple craters (Figures 4.2 and 4.10), and as annular sheets surrounding the central structures and lining the floors of complex craters and basins (Figures 4.4, 4.5(B), and 4.11). Some terrestrial impact melt rocks were initially misidentified as having a volcanic origin. In general, however, impact melt rocks are compositionally distinct from volcanic rocks. They have compositions determined by a mixture of the compositions of the target rocks, in contrast to volcanic rocks that have compositions determined by internal partial melting of more mafic and refractory progenitors, within the planetary body’s mantle or crust.

Impact melt rocks can also contain shocked and unshocked fragments of rocks and minerals. During the cratering event, as the melt is driven down into the expanding transient cavity (Figure 4.10), it overtakes and incorporates less-shocked materials such as clasts, ranging in size from small grains to large blocks. Impact melt rocks that cool quickly generally contain large fractions of clasts, while those that cool more slowly show evidence of melting and resorption of the clastic debris, which is possible because impact melts are initially a superheated mixture of liquid melt and vapor. This is another characteristic that sets impact melt rocks apart from volcanic rocks, which are generally erupted at their melting temperature and no higher.

3. IMPACTS AND PLANETARY EVOLUTION

Impacts are fundamental to the origin and evolution of the Solar System. The current working hypothesis for early Solar System history is that, at an early stage, solid material from the preplanetary disk formed a large number of kilometer-sized planetesimals through collisions or impacts (See: The Origin of The Solar System). This planetesimal assemblage evolved through additional impacts to form a small number of larger planetary embryos. Whether a planetesimal grows or erodes through impacts depends on the impact velocity of the collision. Retention of colliding planetesimal material requires the impact velocity to be less than three to five times the escape velocity of the target planetesimal.

As the impact flux has varied through geologic time, so has the potential for impact to act as an evolutionary agent. The ancient highland crust of the Moon records almost the complete record of cratering since its formation. Crater counts combined with isotopic ages on returned lunar samples have established an estimate of the cratering rate on the Moon and its variation with time. Terrestrial data have been used to extend knowledge of the cratering rate, at least in the Earth–Moon system, to more recent geologic time. The lunar data are generally interpreted as indicating an exponential decrease in the rate until ∼4.0 billion years (Ga) ago, a slower decline for an additional billion years, and a relatively constant rate, within a factor of two, since ∼3.0 Ga ago. The actual rate before ∼4.0 Ga ago is imprecisely known, as there is the question of whether the ancient lunar highlands reflect all of the craters that were produced (i.e. a production population) or only those that have not been obliterated by subsequent impacts (i.e. an equilibrium population). Thus, it is possible that the oldest lunar surfaces give only a minimum estimate of the ancient cratering rate. Similarly, there is some question as to whether the largest recorded events, represented by the major multiring basins on the Moon, occurred over the relatively short time period of 4.2–3.8 Ga ago (the “called lunar cataclysm”) or were spread more evenly with time (See: The Moon).

3.1. Impact Origin of Earth’s Moon

The impacts of the greatest magnitude dominate the cumulative effects of the much more abundant smaller impacts in terms of affecting planetary evolution. In the case of Earth, this would be the massive impact that likely produced the Moon. Earth is unique among the terrestrial planets in having a large satellite and the origin of the Moon has always presented a problem. The suggestion that the Moon formed from a massive impact with Earth was originally proposed some 35 years ago, but, with the development of complex numerical calculations and more efficient computers, it has been possible more recently to model such an event. Most models involve the oblique impact of a Mars-sized object with the proto-Earth, which produces an Earth-orbiting disk of impact-produced vapor and debris, consisting mostly of mantle material from Earth and the impacting body. This disk, depleted in volatiles and enriched in refractory elements, would cool, condense, and accrete to form the Moon.
Planetary Impacts

<text>

less than Mt. Everest at the least topographic relaxation, is over 8 km, somewhat multiring basin at associated with the Orientale Basin (Figure 4.6), the youngest features of the Moon. For example, the topography asso-

<text>

materials ejected from major craters in the lunar highlands is few pristine, igneous rocks from the early lunar crust, with impact products, are in the range of 3.8 \( \times \) 4.0 Ga old. Only a few pristine, igneous rocks from the early lunar crust, with ages >3.9 Ga, occur in the Apollo collection. Computer simulations indicate that the cumulative thickness of materials ejected from major craters in the lunar highlands is 2–10 km. Beneath this, the crust is believed to be brecciated and fractured by impacts to a depth of 20–25 km.

The large multiring basins define the major topographic features of the Moon. For example, the topography associated with the Orientale Basin (Figure 4.6), the youngest multiring basin at \( \sim 3.8 \) Ga and, therefore, the basin with the least topographic relaxation, is over 8 km, somewhat less than Mt. Everest at \( \sim 9 \) km. The impact energies released in the formation of impact basins in the 1000 km size range are on the order of \( 10^{27} \)–\( 10^{28} \) J, one to 10 million times the present annual output of internal energy of Earth. The volume of crust melted in a basin-forming event of this size is on the order of a \( 1 \times 10^6 \) km\(^3\). Although the majority of crater ejecta is generally confined to within \( \sim 2.5 \) diameters of the source crater, this still represents essentially hemispheric redistribution of materials in the case of an Orientale-sized impact on the Moon.

Following formation, these impact basins localized subsequent endogenic geologic activity in the form of tectonism and volcanism. A consequence of such a large impact is the uplift of originally deep-seated isotherms and the subsequent tectonic evolution of the basin, and its immediate environs, is then a function of the gradual loss of this thermal anomaly, which could take as long as a billion years to dissipate completely. Cooling leads to stresses, crustal fracturing, and basin subsidence. In addition to thermal subsidence, the basins may be loaded by later mare volcanism, leading to further subsidence and stress.

All the terrestrial planets experienced the formation of large impact basins early in their histories. Neither Earth nor Venus, however, retains any record of this massive bombardment, so the cumulative effect of such a bombardment on the Earth is unknown. Basin-sized impacts will have also affected any existing atmosphere, hydrosphere, and potential biosphere. For example, the impact on the early Earth of a body in the 500 km size range, similar to the present day asteroids Pallas and Vesta would be sufficient to evaporate the world’s present oceans, if only 25% of the impact energy were used in vaporizing the water. Such an event would have effectively sterilized the surface of Earth. The planet would have been enveloped by an atmosphere of hot rock and water vapor that would radiate heat downward onto the surface, with an effective temperature of a few 1000 degrees. It would take thousands of years for the water-saturated atmosphere to rain out and reform the oceans. Models of impact’s potential to frustrate early development of life on Earth indicate that life could have survived in a deep marine setting at \( 4.2 \)–\( 4.0 \) Ga, but smaller impacts would continue to make the surface inhospitable until \( \sim 4.0 \)–\( 3.8 \) Ga.

3.2. Early Crustal Evolution

Following planetary formation, the subsequent high rate of bombardment by the remaining “tail” of accretionary debris is recorded on the Moon and the other terrestrial planets and the icy satellites of the outer solar system that have preserved some portion of their earliest crust. Due to the age of its early crust, the relatively large number of space missions, and the availability of samples, the Moon is the source of most interpretations of the effects of such an early, high flux. In the case of the Moon, a minimum of 6000 craters with diameters greater than 20 km are believed to have been formed during this early period. In addition, \( \sim 45 \) impacts produced basins, ranging in diameter from Bailly at 300 km, through the South Pole–Aitken Basin at 2600 km, to the putative Procellarum Basin at 3500 km, the existence of which is still debated. The results of the Apollo missions demonstrate clearly the dominance of impact in the nature of the samples from the lunar highlands. Over 90% of the returned samples from the highlands are impact rock units, with 30–50% of the hand-sized samples being impact melt rocks. The dominance of impact as a process for change is also reflected in the age of the lunar highland samples. The bulk of the near-surface rocks, which are impact products, are in the range of 3.8–4.0 Ga old. Although the majority of crater ejecta is generally confined to within \( \sim 2.5 \) diameters of the source crater, this still represents essentially hemispheric redistribution of materials in the case of an Orientale-sized impact on the Moon.

Following formation, these impact basins localized subsequent endogenic geologic activity in the form of tectonism and volcanism. A consequence of such a large impact is the uplift of originally deep-seated isotherms and the subsequent tectonic evolution of the basin, and its immediate environs, is then a function of the gradual loss of this thermal anomaly, which could take as long as a billion years to dissipate completely. Cooling leads to stresses, crustal fracturing, and basin subsidence. In addition to thermal subsidence, the basins may be loaded by later mare volcanism, leading to further subsidence and stress.

All the terrestrial planets experienced the formation of large impact basins early in their histories. Neither Earth nor Venus, however, retains any record of this massive bombardment, so the cumulative effect of such a bombardment on the Earth is unknown. Basin-sized impacts will have also affected any existing atmosphere, hydrosphere, and potential biosphere. For example, the impact on the early Earth of a body in the 500 km size range, similar to the present day asteroids Pallas and Vesta would be sufficient to evaporate the world’s present oceans, if only 25% of the impact energy were used in vaporizing the water. Such an event would have effectively sterilized the surface of Earth. The planet would have been enveloped by an atmosphere of hot rock and water vapor that would radiate heat downward onto the surface, with an effective temperature of a few 1000 degrees. It would take thousands of years for the water-saturated atmosphere to rain out and reform the oceans. Models of impact’s potential to frustrate early development of life on Earth indicate that life could have survived in a deep marine setting at \( 4.2 \)–\( 4.0 \) Ga, but smaller impacts would continue to make the surface inhospitable until \( \sim 4.0 \)–\( 3.8 \) Ga.

3.3. Biosphere Evolution

Evidence from the Earth–Moon system suggests that the cratering rate had essentially stabilized to something approaching a constant value by 3.0 Ga. Although major basin-forming impacts were no longer occurring, there were still occasional impacts resulting in craters in the size range of a few 100 km. The terrestrial record contains remnants of the Sudbury, Canada, and Vredefort, South Africa, structures, which have estimated original crater diameters of \( \sim 250 \) km and \( \sim 300 \) km, respectively, and
ages of $\sim 2 \text{ Ga}$. Events of this size are unlikely to have caused significant long-term changes in the solid geosphere, but they likely affected the biosphere of Earth. In addition to these actual Precambrian impact craters, a number of anomalous spherule beds with ages ranging from $\sim 2.0$ to $3.5 \text{ Ga}$ have been discovered relatively recently in Australia and South Africa. Geochemical and physical evidence (shocked quartz) indicate an impact origin for some of these beds; at present, however, their source craters are unknown. If, as indicated, one of these spherule beds in Australia is temporally correlated to one in South Africa, its spatial extent would be in excess of 32,000 km$^2$.

At present, the only case of a direct physical and chemical link between a large impact event and changes in the biostratigraphic record is at the “Cretaceous—Paleogene boundary”, which occurred $\sim 65$ million years (Ma) ago. The worldwide physical evidence for impact includes: shock-produced, microscopic planar deformation features in quartz and other minerals; the occurrence of stishovite (a high-pressure polymorph of quartz) and impact diamonds; high-temperature minerals believed to be vapor condensates; and various, generally altered, impact-melt spherules. The chemical evidence consists primarily of a geochemical anomaly, indicative of an admixture of meteoritic material. In undisturbed North American sections, which were laid down in swamps and pools on land, the boundary consists of two units: a lower one, linked to ballistic ejecta, and an upper one, linked to atmospheric dispersal in the impact fireball and subsequent fallout over a period of time. This fireball layer occurs worldwide, but the ejecta horizon is known only in North America.

The Cretaceous—Paleogene boundary marks a mass extinction in the biostratigraphic record of the Earth. Originally, it was thought that dust in the atmosphere from the impact led to global darkening, the cessation of photosynthesis, and cooling. Other potential killing mechanisms have been suggested. Soot, for example, has also been identified in boundary deposits, and its origin has been ascribed to globally dispersed wildfires. Soot in the atmosphere may have enhanced or even overwhelmed the effects produced by global dust clouds. Recently, increasing emphasis has been placed on understanding the effects of vaporized and melted ejecta on the atmosphere. Models of the thermal radiation produced by the ballistic re-entry of ejecta condensed from the vapor and melt plume of the impact indicate the occurrence of a thermal-radiation pulse on Earth’s surface. The pattern of survival of land animals 65 Ma ago is in general agreement with the concept that this intense thermal pulse was the first global blow to the biosphere.

Although the record in the Cretaceous—Paleogene boundary deposits is consistent with the occurrence of a major impact, it is clear that many of the details of the potential killing mechanism(s) and the associated mass extinction are not fully known. The “killer crater” has been identified as the $\sim 180 \text{ km}$ diameter structure, known as Chicxulub, buried under $\sim 1 \text{ km}$ of sediments on the Yucatan peninsula, Mexico. Variations in the concentration and size of shocked quartz grains and the thickness of the boundary deposits, particularly the ejecta layer, pointed toward a source crater in Central America. Shocked minerals have been found in deposits both interior and exterior to the Chicxulub structure, as have impact melt rocks, with an isotopic age of 65 Ma.

Chicxulub may hold the clue to potential extinction mechanisms. The target rocks include beds of anhydrite (CaSO$_4$), and model calculations for the Chicxulub impact indicate that the SO$_2$ released would have sent anywhere between 30 billion and 300 billion tons of sulfuric acid into the atmosphere, depending on the exact impact conditions. Studies have shown that the lowering of temperatures following large volcanic eruptions is mainly due to sulfuric acid aerosols. Models, using both the upper and lower estimates of the mass of sulfuric acid created by the Chicxulub impact, lead to a calculated drop in global temperature of several degrees Celsius. The sulfuric acid would eventually return to Earth as acid rain, which would cause the acidification of the upper ocean and potentially lead to marine extinctions. In addition, impact heating of nitrogen and oxygen in the atmosphere would produce NO$_x$ gases that would affect the ozone layer and, thus, the amount of ultraviolet radiation reaching the Earth’s surface. Like the sulfur-bearing aerosols, these gases would react with water in the atmosphere to form nitric acid, which would result in additional acid rains.

The frequency of Chicxulub-size events on Earth is on the order of one every $\sim 100 \text{ Ma}$. Smaller, but still significant, impacts occur on shorter timescales and could affect the terrestrial climate and biosphere to varying degrees. Some model calculations suggest that dust injected into the atmosphere from the formation of impact craters as small as 20 km could produce global light reductions and temperature disruptions. Such impacts occur on Earth with a frequency of approximately two or three every million years but are not likely to have a serious affect upon the biosphere. The most fragile component of the present environment, however, is human civilization, which is highly dependent on an organized and technologically complex infrastructure for its survival. Though we seldom think of civilization in terms of millions of years, there is little doubt that if civilization lasts long enough, it could suffer severely or even be destroyed by an impact event.

Impacts can occur on historical timescales. For example, the Tunguska event in Russia in 1908 was due to the atmospheric explosion of a relatively small body at an altitude of $\sim 10 \text{ km}$. The energy released, based on that required to produce the observed seismic disturbances, has
been estimated as being equivalent to the explosion of ~10 megatons of Trinitrotoluene (TNT). Although the air blast resulted in the devastation of ~2000 km² of Siberian forest, there was no loss of human life. Events such as Tunguska occur on timescales of a 1000 of years. Fortunately, 70% of the Earth’s surface is ocean and most of the land surface is not densely populated. Such oceanic impacts, however, could result in devastating tsunami waves in coastal areas.

In addition to the aforementioned deleterious effects of meteorite impacts, it has become apparent over the past decade that impact events produce several beneficial effects with respect to microbial life. Most importantly, impact events are now known to produce several habitats that are highly conducive to life and that were not present before the impact event. Major habitats include (1) impact-generated hydrothermal systems, which could provide habitats for thermophilic and hyperthermophilic microorganisms, (2) impact-processed crystalline rocks, which have increased porosity and translucence compared to unshocked materials, improving microbial colonization, (3) impact glasses, which, similar to volcanic glasses, provide an excellent readily available source of bioessential elements, and (4) impact crater lakes, which form protected sedimentary basins with various niches and that increase the preservation potential of fossils and organic material. Thus, impact craters, once formed on Early Earth, and, by analogy on Mars and other planets, may have represented prime sites that served as protected niches, where life could have survived and evolved and, more speculatively, perhaps originated.

Of these habitats, the one that has received most attention is the impact-generated hydrothermal system. This derives from the longstanding suggestion that hydrothermal systems might have provided habitats or “cradles” for the origin and evolution of early life on Earth and possibly other planets, such as Mars. This is consistent with the most ancient organisms in the terrestrial tree of life being thermophilic (optimum growth temperatures >50 °C) or hyperthermophilic (optimum growth temperatures >80 °C) in nature. Studies of a number of impact structures on Earth suggest that most impact events that result in the formation of complex impact craters (i.e. >2–4 and >5–10 km diameter on Earth and Mars, respectively) are potentially capable of generating a hydrothermal system. Studies of the Haughton impact structure in the Canadian Arctic suggest that there are six main locations within and around impact craters where impact-generated hydrothermal deposits can form: (1) crater-fill impact melt rocks and melt-bearing breccias; (2) interior of central uplifts; (3) outer margin of central uplifts; (4) impact ejecta deposits; (5) crater rim region; and (6) post-impact crater lake sediments. The question of whether impact-generated hydrothermal systems form in craters elsewhere in the Solar System remains open, however, in 2010, such evidence was presented from the Toro Crater on Mars.

It has become apparent over the past couple of decades that impact events have profoundly affected the evolution of life on Earth and may have also influenced life’s origins. There is also the outstanding question of the potential transfer of life from another planet to Earth through impact events. Experiments have shown that certain organisms can survive the impact process and the harsh conditions of space, at least for the time span that these experiments were conducted, which is obviously limited by the human life-span and research careers. Whether life could have been ejected, survived the potential several millennia journey through Space, survive the impact on Earth and then have the ability to colonize this planet, remains conjecture at this time.

4. IMPACTS AS PLANETARY PROBES

Impacts serve to probe the nature of the subsurface of planetary bodies through the processes of excavation, exposure and uplift. For example, in terrestrial complex structures lithologies exposed in the central uplift originally from depths approximately one tenth of the final rim diameter.

4.1. Water and Ices

As noted earlier, it has been known since the late 1970s that many Martian impact craters possess, “lobate”, “rampart”, “layered”, or “fluidized” ejecta (Figure 4.7). According to the most recent Martian crater database, which contains 79,723 craters ≥3 km in diameter, approximately 50% of all Martian craters possess clearly discernable ejecta blankets and, of these, over 40% possess layered ejecta morphologies. Although the formation of layered ejecta morphologies continues to be debated, it is generally accepted that subsurface volatiles (water-ice) played a role in the formation of these unique ejecta morphologies. In addition to layered ejecta, evidence of deeper subsurface volatiles has been recently bolstered by the discovery of a globally wide-spread crater related pitted deposit (Figure 4.12) observed in 205 craters ranging from 1 to 150 km in diameter. Interestingly, crater-related pitted deposits have also been observed associated with the freshet craters on the asteroid Vesta, and are consistent with the former Martian work that suggests that there may be volatile-rich materials beneath the scale of this large differentiated asteroid. There is also direct evidence of shallow subsurface ice on Mars has been brought to light by small impact craters (10s of meters scale; Figure 4.13). High-resolution images indicate that meter-to-decameter scale impact craters occur frequently on Mars (~ 20 craters over a 7 year period). Subsequent observations from High Resolution Imaging Science Experiment
(HiRISE) with coordinated Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) of five newly formed craters within the mid-latitude regions of Mars revealed water ice excavated from just meters below the surface. Historically, rocks returned from the Moon were composed entirely of anhydrous minerals and, thus, believed to be completely “bone-dry”. This was typically considered a testament to the energetic and violent planet-scale impact origin of the Moon. However, recent spectral data of the surface of the Moon indicates the presence of adsorbed water molecules and/or hydroxyl (OH\(^+\)) associated with lunar surface materials. Based on a few assumptions, the hydrated proportion likely amounts to \(\sim 0.1\)% of these surface materials. Interestingly, the hydration spectral signature has been noted to correlate with the ejecta blankets of several fresh lunar craters, suggesting that some of these relatively water-rich materials can retain their water just beneath the surface, where they are protected by interactions with cosmic rays and the solar wind.

The concept of an impact as a natural subsurface probe was utilized in a recent lunar mission designed to create an artificial impact and observe in real-time the ejected plume of impact-liberated subsurface materials. The Lunar Crater Observation and Sensing Satellite detected an estimated \(\sim 6\)% by mass of water in the top few meters of the lunar regolith that was ejected by a crater formed from the impact of a spent Centaur rocket stage into a crater near the lunar south pole.

4.2. Spectral Composition

Several recent studies of craters that take advantage of crater as probes emphasize distinct compositions between the exposed subsurface materials and the surrounding surface around impact craters. From a survey of over 100 lunar central peaks in complex craters with Clementine multispectral images, some general conclusions can be drawn regarding the upper and lower lunar crust. These include: (1) that the lunar crust is extremely anorthositic, consistent with the “magma ocean” model (See The Moon), (2) crustal composition gradually increases in mafic content with depth, although mafic compositions are generally rare
in central peaks, and (3) the lower crust is more compositionally diverse than the highlands. The strong anorthositic character of the lunar crust, as observed through central peaks, was confirmed by the Selene multiband imager, which identified some peaks that approached 100% pure anorthosite. Recent results are generally consistent with the Clementine spectral images of central peaks, with the exception of the discovery of a rare occurrence of Mg-spinel-rich deposits associated with the central peak of Theophilus Crater. As Theophilus lies on the rim of the Nectaris Basin, and with the only other known occurrence of Mg-spinel-rich also being associated with another large impact basin, it has been suggested that these materials may originate from deep within the lunar interior.

The Thermal Emission Imaging System (THEMIS) was the first subkilometer-scale spectral instrument flown in Mars orbit that could effectively be used to detect crater-scale spectral contrasts. Although a great deal of the surface of Mars is spectrally obscured by surface dust, including many crater central peaks, a few THEMIS observations of crater-related deposits or outcrops show higher concentrations of mafic minerals (pyroxene and olivine), with respect to their surrounding surfaces (Figures). A THEMIS-based study also discovered an undetermined high-silica spectral component associated with some crater deposits in the northwestern part of Syrtis Major. Once interpreted to represent exposed granitoids, these crater-related deposits have since been observed to contain altered silicate phases, including smectites, hydrated glasses and zeolites. As such, they are now interpreted to phases altered prior to impact or via multiple impact alteration pathways, including hydrothermalism. Additional craters associated with hydrated silicates are now being observed by the Observatoire pour la Minéralogie, l’Eau, les Glaces et l’Activité (OMEGA) and CRISM across Mars. Two independent global-scale classifications of alteration phases using OMEGA and CRISM data indicate that ~70% of the total occurrences of known alteration phases on Mars are associated with impact craters. With the exception of small simple craters, the origin of altered phases within complex craters, especially associated with crater central peaks, is difficult to constrain as large impact events generate a hydrothermal system that typically dominates the central uplift region. Detailed studies of multiple data sets are required to distinguish between these different origins.

4.3. Morphologic and Geologic

High-resolution visible images from recent planetary missions have been documenting exquisite geologic and structural details of craters. For example, images from HiRISE and the Context Camera on the Mars Reconnaissance Orbiter have revealed exposed bedrock in central uplifts of complex craters. The exposed parautochthonous bedrock can be divided into three outcrop-scale textural classes: (1) layered, (2) massive and fractured and (3) brecciated. Of these three classes, the layered bedrock class is the most strongly correlated with specific surface geologic units on Mars. Central peaks with layers occur predominately in units interpreted as extensive and voluminous lava deposits or plains (e.g., Tharsis and the surrounding regions). These uplifts have layered mega-blocks that are consistent with sampled stratigraphic sections of cyclic volcanism (i.e. alternating lava flows and pyroclastics). Although not surprising, the application of resolving, identifying and mapping bedrock characteristics in crater central uplifts on planetary bodies, other than Earth, is a new and novel approach for aiding planetary scientists to determine the geologic history of the surface and subsurface and to elucidate various aspects of central uplift formation. How different target rocks are exposed and deformed are particularly informative with respect to the impact process. For example, it has been hypothesized that the final orientation of layers that are faulted and rotated during central uplifts show a relationship to the impactor trajectory. Breccia injection dikes and impact melt deposits associated with crater central uplifts, which are now resolvable in HiRISE images of Mars, are also providing important clues with respect to the impact process.

BIBLIOGRAPHY


1. SUMMARY OF PLANETARY VOLCANIC FEATURES

1.1. Earth

The ~70% of Earth’s surface represented by the crust forming the floors of the oceans consists of volcanic rocks generally erupted within the last 300 Ma (million years) from long lines of volcanoes located along ridges near the centers of ocean basins. This geologically youthful age stimulated the development of the theory of plate tectonics, which explained the locations and distributions of volcanoes over Earth’s surface. Midocean ridge volcanoes erupt magmas called basalts that are relatively metal-rich and silica- and volatile-poor, and these volcanoes mark the constructional margins of Earth’s rigid crustal plates. Basalts are the products of the partial melting of mantle rocks due to decreasing pressure at the tops of convection cells in which temperature variations cause the solid mantle to deform and flow on very long timescales. Basalt compositions are closely related to the bulk composition of the mantle, which makes up most of Earth’s volume outside the iron-dominated core. The volcanic edifices produced by ocean-floor volcanism consist mainly of relatively fluid (low-viscosity) lava flows with lengths from a few kilometers to a few tens of kilometers. Lava flows erupted along the midocean ridges simply add to the topography of the edges of the growing plates as the plates move slowly (~10 mm/year) away from the ridge crest (See Earth as a Planet: Atmosphere and Oceans; Earth as a Planet: Surface and Interior).

Lavas erupted from vents located some distance away from the ridge crest build roughly symmetrical edifices with convex-upward shapes described as shield volcanoes (having relatively shallow flank slopes) or domes (having relatively steeper flanks). Some of these vent systems are not related to the spreading ridges at all, but instead mark the locations of “hot spots” in the underlying mantle, unusually vigorous plumes of mantle material feeding magmas through the overlying plate. Because the plate moves over the hot spot, a chain of shield volcanoes can be built up in this way, marking the trace of the relative motion. The largest shield volcanoes on Earth form such a line of volcanoes, the Hawaiian Islands, and the two largest of these edifices, Mauna Loa and Mauna Kea, rise ~10 km above the ocean floor and have basal diameters of about 200 km.

Eruptive activity on shield volcanoes tends to be concentrated either at the summit or along linear or arcuate zones radiating away from the summit, called rift zones. The low viscosity of the basaltic magmas released in...
Hawaiian-style eruptions on these volcanoes (Figure 5.1) allows the lava flows produced to travel relatively great distances (a few tens of km), and is what gives shield volcanoes their characteristic wide, low profiles. It is very common for a long-lived reservoir of magma, a magma chamber, to exist at a depth of a few to several kilometers below the summit. This reservoir, which is roughly equant in shape and may be up to 1–3 km in diameter, intermittently feeds surface eruptions, either when magma ascends vertically in the volcano summit region or when magma flows laterally in a subsurface fracture called a dike, which most commonly follows an established rift zone, to erupt at some distance from the summit. In many cases, magma fails to reach the surface and instead freezes within its fracture, forming an intrusion. The summit reservoir is fed, probably episodically, from partial melt zones in the mantle beneath. Rare but important events in which a large volume of magma leaves such a reservoir lead to the collapse of the rocks overlying it, and a characteristically steep-sided crater called a caldera is formed, with a width similar to that of the underlying reservoir.

Volcanoes erupting relatively silica-rich and volatile-rich magma (andesite or, less commonly, rhyolite) mark the destructive margins of plates, where the plates bend downward to be subducted into the interior and at least partly remelted. These volcanoes tend to form an arcuate pattern (called an island arc when the volcanoes rise from the sea floor) marking the trace on the surface of the zone where the melting is taking place, at depths on the order of 100–150 km. The andesitic magmas thus produced are the products of melting of a mixture of subducted ocean floor basalt, sedimentary material that had been washed onto the ocean floor from the continents (which are themselves an older, silica-rich product of the chemical differentiation of Earth), seawater trapped in the sediments, and the primary mantle materials into which the plates are subducted. Thus, andesites are much less representative of the current composition of the mantle. Andesite magmas are rich in volatiles (mainly water, carbon dioxide, and sulfur compounds), and their high silica contents give them high viscosities, making it hard for gas bubbles to escape. As a result, andesitic volcanoes often erupt explosively in Vulcanian-style eruptions, producing localized pyroclastic deposits with a wide range of grain sizes; alternatively, they produce relatively viscous lava flows that travel only short distances (a few kilometers) from the vent. The combination of short flows and localized ash deposits tends to produce steep-sided, roughly conical volcanic edifices.

When large bodies of very silica-and volatile-rich magma (rhyolite) accumulate—in subduction zones or, in some cases, where hot spots exist under continental areas, leading to extensive melting of the continental crustal rocks—the potential exists for very large-scale explosive eruptions to occur, in which finely fragmented magma is blasted at high speed from the vent to form a convecting eruption cloud, called a Plinian cloud, in the atmosphere. These clouds may reach heights up to 50 km, and pyroclastic fragments fall from them to create a characteristic deposit spreading downwind from the vent (Figure 5.2). Under certain circumstances, the cloud cannot convect in a stable fashion and collapses to form a fountain-like structure over the vent, which feeds a series of pyroclastic flows—mixtures of incandescent pyroclastic fragments, volcanic gas, and entrained air—that can travel for at least tens of kilometers from the vent at speeds in excess of...
100 m/s, eventually coming to rest to form a rock body called ignimbrite. These fall and flow deposits may be so widespread around the vent that no appreciable volcanic edifice is recognizable; however, there may still be a caldera, or at least a depression, at the vent site due to the collapse of the surface rocks to replace the large volume of material erupted from depth.

It should be clear from the foregoing descriptions that the distribution of the various types of volcano and characteristic volcanic activity seen on Earth are intimately linked with the processes of plate tectonics. A major finding to emerge from the exploration of the Solar System over the last 30 years is that this type of large-scale tectonism is currently confined to the Earth and may never have been active on any of the other bodies. Virtually all of the major volcanic features that we see elsewhere can be related to the eruption of mantle melts similar to those associated with the mid-ocean ridges and oceanic hotspots on Earth. However, differences between the physical environments (acceleration due to gravity, atmospheric conditions) of the other planets and Earth lead to significant differences in the details of the eruption processes and the deposits and volcanic edifices formed.

1.2. The Moon

Analyses of the samples collected from the Moon by the Apollo missions in the 1970s showed that there were two major rock types on the lunar surface. The relatively bright rocks forming the old, heavily cratered highlands of the Moon are a primitive crust that formed about 4.5 Ga (billion years) ago by the accumulation of solid minerals at the cooling top of a possibly 300 km thick melted layer referred to as a magma ocean. This early crust was extensively modified, mainly prior to about 3.9 Ga ago, by the impacts of comets and asteroids with a wide range of sizes to form impact craters and basins. Some of the larger craters and basins (the mare basins) were later flooded episodically by extensive lava flows, many more than 100 km long, to form the darker rocks visible on the lunar surface (See The Moon, Planetary Impacts).

Radiometric dating of samples from lava flow units showed that these mare lavas were mostly erupted between 3 and 4 Ga ago, forming extensive, relatively flat deposits inside large basins. Individual flow units, or at least groups of flows, can commonly be distinguished using multispectral remote-sensing imagery on the basis of their differing chemical compositions, which give them differing reflectivities in the visible and near-infrared parts of the spectrum. In composition these lavas are basaltic, and their detailed mineralogy shows that they are the products of partial melting of the lunar mantle at depths between 150 and more than 400 km, the depth of origin increasing with time as the lunar interior cooled. Melting experiments on samples, supported by theoretical calculations based on their mineralogies, show that these lavas were extremely fluid (i.e. had very low viscosities, at least a factor of 3–10 less than those of typical basalts on Earth). This allowed them to travel for great distances, often more than 100 km (Figure 5.3) from their vents; it also meant that they had a tendency to flow back into, and cover up, their vents at the ends of the eruptions. Even so, it is clear from the flow directions that the vents were mainly near the edges of the interiors of the basins that the flows occupy. Many vents were probably associated with the arcuate rilles found in similar positions. These are curved grabens, trench-like depressions parallel to the edges of the basins formed as parts of the crust sink between pairs of parallel faults caused by tension. This marginal tension, due to the weight of the lava ponded in the middle of the basin, makes it easier for cracks filled with magma to reach the surface in these places.

A second class of lunar volcanic features associated with the edges of large basins is the sinuous rilles. These are meandering depressions, commonly hundreds of meters wide, tens of meters deep, and tens of kilometers long, which occur mainly within the mare basals. Some are discontinuous, giving the impression of an underground tube that has been partly revealed by partial collapse of its roof, and these are almost certainly the equivalent of lava tube systems (lava flows whose top surface has completely solidified) on Earth. Other sinuous rilles are continuous open channels all along their length; these generally have origins in source

![Figure 5.3](image_url)
depressions two or three times wider than the rille itself, and
become narrower and shallower with increasing downslope
distance from the source. These sinuous rilles appear to have
been caused by long-duration lava flows that were very
turbulent, i.e. the hot interior was being constantly mixed
with the cooler top and bottom of the flow. As a result the
flows were able to heat up the preexisting surface until some
of its minerals melted, allowing material to be carried away
and an eroded channel to form.

In contrast to the lava flows and lava channels, two
types of pyroclastic deposit are recognized on the Moon.
There are numerous deposits called dark mantles, often
roughly circular and up to at least 200 km in diameter,
where the fragmental lunar surface regolith is less reflective
than usual, and spectroscopic evidence shows that it con-
tains a component of small volcanic particles in addition to
the locally derived rock fragments. The centers of these
regions are commonly near the edges of mare basins,
suggesting that the dark mantle deposits were produced by
the same source vents as the lava flows. Chemical analyses
of the Apollo lava samples show that the Moon’s mantle
is severely depleted in common volatiles like water and
carbon dioxide due to its hot origin (See The Moon) and
suggest that the main gas released from mare lava vents
was carbon monoxide, produced in amounts up to a few
100 parts per million by weight as a result of a chemical
reaction between free carbon and metal oxides, mainly iron
oxide, in the magma as it neared the surface.

Several small, dark, fragmental deposits occur on the
floor of the old, 90-km-diameter impact crater Alphonsus.
These patches, called dark haloes, extend for a few kilo-
meters from the rims of subdued craters that are centered
on, and elongated along, linear fault-bounded depressions
(called linear rilles) on the crater floor. It is inferred that
these are the sites of less energetic volcanic explosions.

Localized volcanic constructs such as shield volcanoes
and domes are rare on the Moon, though more than 200 low,
shield-like features with diameters mainly in the range
3–10 km are found in the Marius region within Oceanus
Procellarum, in northeast Mare Tranquilitatis, and in the
region between the craters Kepler and Copernicus.
Conspicuously absent are edifices with substantial summit
calderas. This implies that large, shallow magma reservoirs
are very rare, almost certainly a consequence of the diffi-
culty with which dense magmas rising from the mantle
penetrate the low-density lunar crust. However, a few
collapse pits with diameters up to 3 km do occur, located
near the tops of domes or aligned along linear rilles.

1.3. Mars

About half of the surface of Mars consists of an ancient
crust containing impact craters and basins. Spectroscopic
evidence from orbiting spacecraft suggests that it is
composed mainly of volcanic rocks. The other half of the
planet consists of relatively young, flat, lower-lying, plains-
forming units that are a mixture of wind-blown sediments,
lava flows, and rock debris washed into the lowlands by
episodes of water release from beneath the surface.
Combining orbital observations with analyses made by
the several probes that have landed successfully on the
surface suggests that most of the magmas erupted on Mars
are basalts or basaltic andesites (See Mars: Surface and
Interior, Mars Site Geology and Geochemistry).

The most obvious volcanic features on Mars are five
extremely large (~600 km diameter, heights up to
~20 km) shield volcanoes (Olympus Mons, Ascraeus
Mons, Pavonis Mons, Arsia Mons and Alba Mons) with the
same general morphology as basaltic shield volcanoes
found on Earth (Figure 5.4). These volcanoes are sur-
rounded by overlapping lava aprons that collectively form a
huge volcanic rise. A second but smaller rise contains the
volcanoes Elysium Mons, Hecates Tholus and Albor
Tholus. There are also about 20 smaller shield volcanoes on
Mars. Counts of small impact craters seen in high-
resolution images from orbiting spacecraft show that the
ages of the lava flow units on the volcanoes range from
more than 3 Ga to less than ~50 Ma. Complex systems of
nested and intersecting calderas are found on the larger
shields, implying protracted evolution of the internal
plumbing of each volcano, typified by cycles of activity in
which a volcano is sporadically active for ~1 Ma and then
dormant for ~50 Ma. Individual caldera depressions are up
to at least 30 km in diameter, much larger in size than any
found on Earth, and imply the presence of very large

![FIGURE 5.4 The Olympus Mons shield volcano on Mars with the Hawaiian islands superimposed for scale. NASA image with overlay by P.J. Mouginis-Mark.](image-url)
shallow magma reservoirs during the active parts of the volcanic cycles. The large size of these reservoirs, like that of the volcanoes themselves, is partly due to the low acceleration due to gravity on Mars and partly due to the absence of plate tectonics, which means that a mantle hot spot builds a single large volcano, rather than a chain of small volcanoes as on Earth. The availability of large volumes of melt in the mantle beneath some of the largest shield volcanoes has apparently led to the production of giant swarms of dikes, propagating radially away from the volcanic centers for more than 2000 km in some cases. The locations of these dikes are indicated by the presence of narrow (1–5 km wide) grabens where the intrusions have exerted extensional stresses on the crust.

Most shields appear to have flanks dominated by lava flows, many more than 100 km long. However, there are examples of sinuous channels like the sinuous rilles on the Moon, presumably caused by hot, turbulent, high-speed lavas melting the ground over which they flow. Some of the older and more eroded edifices, like Tyrrenhus Mons and Hadriacus Mons, appear to contain high proportions of relatively weak, presumably pyroclastic, rocks. There is a hint, from the relative ages of the volcanoes and the stratigraphic positions of the mechanically weaker layers within them, that pyroclastic eruptions were commoner in the early part of Mars’s history. It is also possible that some of the plains-forming units, generally interpreted as weathered lava flows, in fact consist of pyroclastic fall or flow deposits. Not all of the sources of these deposits have been identified with certainty — there are a few dozen massifs, each tens of km in size, in the Martian highlands that have been tentatively interpreted as ancient degraded volcanoes.

1.4. Venus

Because of its dense, optically opaque atmosphere, the only detailed synoptic imaging of the Venus surface comes from orbiting satellite-based radar systems. Despite the differences between optical and radar images (radar is sensitive to both the dielectric constant and the roughness of the surface on a scale similar to the radar wavelength), numerous kinds of volcanic features have been unambiguously detected on Venus. Large parts of the planet are covered with plains-forming lava flows, having well-defined lobate edges and showing the clear control of topography on their direction of movement (Figure 5.5). The lengths (which can be up to several hundred kilometers) and thicknesses (generally significantly less than 30 m, since they are not resolvable in the radar altimetry data) of these flows suggest that they are basaltic in composition. This interpretation is supported by the (admittedly small) amounts of major-element chemical data obtained from six Soviet probes that soft-landed on the Venus surface. Some areas show concentrations of particularly long flows called fluctüs (Latin for floods) and incised channels called canali. Most of the lava plains, judging by the numbers of superimposed impact craters, were emplaced within the last ~700 Ma (See Venus: Surface and Interior).

Many areas within the plains and within other geological units contain groupings (dozens to hundreds) of small volcanic shields or domes, from less than one to several kilometers in diameter. At least 500 such shield fields have been identified. Some of the individual volcanoes have small summit depressions, apparently due to magma withdrawal and collapse, and others are seen to feed lava flows. Quite distinct from these presumably basaltic shields and domes is a class of larger, steep-sided domes with diameters of a few tens of kilometers and heights up to ~1 km. The surface morphologies of these domes suggest that most were emplaced in a single episode, and current theoretical modeling shows that their height-to-width ratio is similar to that expected for highly viscous silicic (perhaps rhyolitic) lavas on Earth.

Many much larger volcanic constructs occur on Venus. About 300 of these are classed as intermediate volcanoes and have a variety of morphologies, not all including extensive lava flows. A further 150, with diameters between 100 and about 600 km, are classed as large volcanoes. These are generally broad shield volcanoes, with extensive systems of lava flows and heights above the surrounding plains of up to 3 km.
Summit calderas are quite common on the volcanoes, ranging in size from a few kilometers to a few tens of kilometers. There are two particularly large volcano-related depressions, called Sacajawea and Colette, located on the upland plateau Lakshmi Planum. With diameters of ~200 km and depths of ~2 km, these features appear to represent the downward sagging of the crust over some unusually deep-seated site of magma withdrawal.

Finally, there are a series of large, roughly circular features on Venus, which, though intimately linked with the large-scale tectonic stresses acting on the crust (they range from a few hundred to a few thousand kilometers in diameter), also have very strong volcanic associations. These are the coronae, novae, and arachnoids. Though defined in terms of the morphology of circumferential, moat-like depressions and radial fracture systems, these features commonly contain small volcanic edifices (shields or domes), small calderas, or lava flows, the latter often apparently fed from elongate vents coincident with the distal parts of radial fractures. In such cases, it seems extremely likely that the main feature is underlain by some kind of magma reservoir, which feeds the more distant eruption sites via lateral dike systems.

1.5. Mercury

Much of the surface of Mercury is a heavily cratered ancient terrain like that of the Moon and parts of Mars. However, data from the MESSENGER spacecraft orbiting Mercury have confirmed suspicions from the earlier Mariner 10 flyby mission that the relatively smooth plains-forming units dispersed among the craters (the inter-crater plains) are vast areas of basaltic lava flows (Figure 5.6). As in the case of the Moon, these lavas appear to be characterized by high eruption rates, great travel distances, and the ability to commonly drown their own vents at the end of the eruption, making the identification of source areas difficult.

Localized explosive volcanic activity is indicated by the presence of more than 30 rimless depressions up to ~40 km in diameter surrounded by spectrally distinctive deposits that grade progressively into the surrounding terrain. The lateral extents of these deposits, up to ~30 km, imply up to ~1 wt% of volatiles in the erupting magmas. The nature of the volatiles is uncertain, but a high abundance of sulfur is present in crustal rocks. Additional evidence for volcanic activity includes several irregularly shaped, rimless, steep-sided pits up to ~30 km in diameter that commonly occur on the floors of impact craters. These may indicate collapse after withdrawal of shallow bodies of intruded magma or the decomposition of magma-linked metal sulfide deposits due to the high daytime temperatures (See Mercury).

1.6. Io

The bulk density of Io suggests that it has a silicate composition, similar to that of the inner, Earth-like planets. Io and the Earth’s Moon also have similar sizes and masses, and it might therefore be expected by analogy with the Moon that any volcanic activity on Io would have been confined to the first 1 or 2 billion years of its life. However, as the innermost satellite of the gas-giant Jupiter, Io is subjected to strong tidal forces. An orbital period resonance driven by the mutual gravitational interactions of Io, Europa, and Ganymede causes the orbit of Io to be slightly elliptical. This, coupled with the fact that Io rotates synchronously (i.e. the orbital period is the same as that of the axial rotation), means that the interior is subjected to periodic tidal flexing. The inelastic part of this deformation generates heat on a scale that far outweighs any remaining heat of formation or heat from the decay of naturally radioactive elements. As a result, Io is currently the most volcanically active body in the solar system. At any one time there may be up to a dozen erupting vents. Roughly half of these produce lava flows from fissure vents (Figure 5.7) associated with calderas located at the centers of very low shield-like features, and half produce umbrella-shaped eruption clouds into which gases and small pyroclasts are being ejected at speeds of up to 1000 m/s to reach heights up to 300 km (Figure 5.8) (See Io).

The main gases detected in the eruption clouds are sulfur and sulfur dioxide, and much of the surface is coated with highly colored deposits of sulfur and sulfur compounds that have been degassed from the interior over solar...
system history and are now concentrated in the near-surface layers. However, it seems very likely, based on the presence of hot lava flows and the fluid dynamic and thermodynamic analysis of the eruption clouds, that the underlying cause of the activity is the ascent of very hot basaltic magmas from the mantle. The average eruption rate on Io is so great that the materials forming the surface layers at any one time are buried to depths of order 30 km in only a few million years. This rapid subsidence makes the geothermal gradient very nonlinear, and volatiles like sulfur and sulfur dioxide do not melt until depths of ~20 km area reached. At these depths, these volatiles can be entrained into magma in dikes propagating up from the deeper mantle. The magma may be inherently quite volatile poor, but the addition of the buried volatiles makes it anomalously volatile rich and drives the extremely explosive eruptions. Most of the volatiles condense as they expand and cool and, along with the silicate pyroclasts, eventually fall back to the surface, where their subsequent burial provides the materials to drive future explosive eruptions.

1.7. The icy Satellites: Cryo-Volcanism

Many of the satellites of the gas-giant planets have bulk densities indicating that their interiors are mixtures of silicate rocks and the ice of common volatiles, mainly water. On some of these bodies (e.g. Jupiter’s satellites Ganymede and Europa, Uranus’s satellite Ariel, Neptune’s satellite Triton, and Saturn’s large satellite Titan), flow-like features are seen, reminiscent of very viscous lava flows (See: Icy Satellites).

However, there is no spectroscopic evidence for silicate magmas having been erupted on these bodies, and the flow-like features have forced us to recognize that there is a more general definition of volcanism than that employed so far. Volcanism is the generation of partial melts from the internal materials of a body and the transport out onto the surface of some fraction of those melts. In the ice-rich bodies, it is the generation of liquid water from solid ice that mimics the partial melting of rocks, in the process called cryo-volcanism.

The ability of this water to erupt at the surface is influenced by its content of volatiles like ammonia and methane. Since the surface temperatures of most of these satellites are very much less than the freezing temperature of water (even when the freezing point is lowered by the presence of compounds like ammonia), and since they do not have appreciable atmospheres (except Titan), the fate of
any liquid water erupting at the surface is complex. Cooling will produce ice crystals at all boundaries of the flow and, being less dense than liquid water, these crystals will rise toward the flow surface. Because of the negligible external pressure, evaporation (boiling) will take place within the upper few hundred millimeters of the flow. The vapor produced will freeze as it expands, to settle out as a frost or snow on the surrounding surface. The boiling process extracts heat from the liquid and adds to the rate of ice crystal formation. If enough ice crystals collect at the surface of a flow, they will impede the boiling process, and if a stable ice raft several 100-mm-thick forms, it will suppress further boiling. Thus, if it is thick enough, a liquid water flow may be able to travel a significant distance from its eruption site. It is even possible that solid ice may form flow-like features, in essentially the same way that glaciers flow on Earth, though the very low temperatures will make the timescales much longer.

If liquid water produced below the surface of an icy satellite contains a large enough amount of volatiles it will erupt explosively at high speed in what, near the vent, is the equivalent of a Plinian eruption. The expanding volatiles could cause the eruption cloud to spread sideways (like the umbrella-shaped plumes on Io) and disperse the water droplets, rapidly freezing to snow, over a wide area. If the eruption speed is high enough and the parent body small enough some of the snow may be ejected with escape velocity. Data from the Cassini spacecraft provided graphic evidence for this process occurring near the south pole of Saturn’s small satellite Enceladus. The orbit of Enceladus is very close to the brightest of Saturn’s many rings, the E ring, which appears to be composed of particles of ice. It now seems clear that these are derived directly from Enceladus, having been ejected fast enough to escape from the satellite but not from Saturn itself (See: Planetary Rings). The volatiles driving the water release include nitrogen, methane and carbon dioxide with traces of propane, ethane, acetylene and ammonia.

1.8. The Differentiated Asteroids

The meteorites that fall to the Earth’s surface are fragments ejected from the surfaces of asteroids during mutual collisions. Most meteorites are pieces of silicate rock and, whereas many contain minerals consistent with them never having been strongly heated, the mineralogy of others can only be explained if they are either solidified samples of what was once magma or pieces of what was once a mantle that partially melted and then cooled again after magma was removed from it. Additionally, some meteorites are pieces of a nickel-iron-sulfur alloy that was once molten but subsequently cooled slowly. Taken together these observations imply that some asteroids went through a process of extensive chemical differentiation by melting to form a crust, mantle and core. The trace element composition of the meteorites from these differentiated asteroids shows that they were heated by the radioactive decay of a group of short-half-life isotopes that were present at the time the Solar System formed, the most important of which was $^{26}$Al which has a half life of $\sim 0.75$ Ma. Thus all of the heating, melting and differentiation must have taken place within an interval of only a few million years. Yet during this brief period, asteroids as small as 100 km in diameter were undergoing the same patterns of mantle melting, melt rise to the surface, and explosive and effusive eruptions that would only start to occur on Earth, Mars and Venus many tens of millions of years later.

Earth-based spectroscopic evidence, now supported by remote-sensing measurements from the orbiting Dawn spacecraft, very strongly suggests that the asteroid 4 Vesta is the parent body of one group of volcanically-generated surface, crust and upper-mantle rocks, the Howardite–Eucrite–Diogenite group of meteorites. Unfortunately Vesta has been so modified by impact cratering that no obvious volcanic features like lava flows are visible in the Dawn images. We have not yet identified any other differentiated asteroids with such certainty, but know from their compositions that the Aubite and Ureilite meteorites are rocks from the mantles of two different asteroids on which violently explosive eruptions ejected magma that should have become their crustal rocks into space at escape velocity. Acapulcoite and Lodranite meteorites are rocks from the shallow crust or upper mantle of a body that produced rather small amounts of gas during mantle melting, so that in these meteorites we see gas bubbles trapped in what was once magma traveling through fractures toward the surface. Finally the nickel-iron meteorites cluster into many tens of chemically-similar groups implying that at least this number of differentiated asteroids once existed but have since been largely fragmented in mutual collisions. The importance of these meteorites is that they give us copious samples of the very deep interiors of their parent bodies as well as the surfaces; such deep samples will not be available for a very long time for Venus, Mars and Mercury and are rare even for the Earth and Moon (See: Meteorites; Asteroids).

2. CLASSIFICATION OF ERUPTIVE PROCESSES

Volcanic eruption styles on Earth were traditionally classified mainly in terms of the observed composition and dispersal of the eruption products. Over the last 30 years it has been realized that they might be more systematically classified in terms of the physics of the processes involved. This has the advantage that a similar system can be adopted for all planetary bodies, automatically taking account of the ways in which local environmental factors (especially
surface gravity and atmospheric pressure) lead to differences in the morphology of the deposits of the same process occurring on different planets.

Eruptive processes are classified as either explosive or effusive. An effusive eruption is one in which lava spreads steadily away from a vent to form one or more lava flows, whereas in explosive eruptions magma emerging through the vent is torn apart, as a result of the coalescence of expanding gas bubbles, into clots of liquid that are widely dispersed. The clots cool while in flight above the ground and may be partly or completely solid by the time they land to form a layer of pyroclasts. There is some ambiguity concerning this basic distinction between effusive and explosive activity, because many lava flows form from the coalescence, near the vent, of large clots of liquid that have been disrupted by gas expansion but that have not been thrown high enough or far enough to cool appreciably. Thus, some, especially Hawaiian-style, eruptions have both an explosive and an effusive component at the same time.

There is also ambiguity about the use of the word explosive in a volcanic context. Conventionally, an explosion involves the sudden release of a quantity of material that has been confined in some way at a high pressure. Most often the expansion of trapped gas drives the explosion process. In volcanology, the term explosive is used not only for this kind of abrupt release of pressurized material, but also for any eruption in which magma is torn apart into pyroclasts that are accelerated by gas expansion, even if the magma is being erupted in a steady stream over a long time period. In eruption styles falling into the first category include Strombolian, Vulcanian, and phreato-magmatic activity, whereas those falling into the second include Hawaiian and Plinian activity. All of these styles are discussed in detail below.

### 3. EFFUSIVE ERUPTIONS AND LAVA FLOWS

Whatever the complications associated with prior gas loss, an effusive eruption is regarded as taking place once lava leaves the vicinity of a vent as a continuous liquid flow. The morphology of a lava flow, both while it is moving and after it has come to rest as a solid rock body, is an important source of information about the rheology (the deformation properties) of the lava, which is determined largely by its chemical composition, and about the rate at which the lava is being delivered to the surface through the vent. Because lava flows basically similar to those seen on Earth are so well exposed on Mars, Venus, Mercury, the Moon and Io, a great deal of effort has been made to understand lava emplacement mechanisms.

In general, lava contains some proportion of solid crystals of various minerals and also gas bubbles. Above a certain temperature called the liquidus temperature, all the crystals will have melted, and the lava will be completely liquid. Under these circumstances, lavas containing less than about 20% by volume of gas bubbles will have almost perfectly Newtonian rheologies, which means that the rate at which the lava deforms, the strain rate, is directly proportional to the stress applied to it under all conditions. This constant ratio of the stress to the strain rate is called the Newtonian viscosity of the lava. At temperatures below the liquidus but above the solidus (the temperature at which all the components of the lava form completely solid minerals), the lava in general contains both gas bubbles and crystals and has a non-Newtonian rheology. The ratio of stress to strain rate is now a function of the stress, and is called the apparent viscosity. At high crystal or bubble contents, the lava may develop a nonzero strength, called the yield strength, which must be exceeded by the stress before any flowage of the lava can occur. The simplest kind of non-Newtonian rheology is one in which the increase in stress, after the yield strength is exceeded, is proportional to the increase in strain rate: the ratio of the two is then called the Bingham viscosity and the lava is described as a Bingham plastic.

The earliest theoretical models of lava flows treated them as Newtonian fluids. Such a fluid released on an inclined plane will spread both downslope and sideways indefinitely (unless surface tension stops it, a negligible factor on the scale of lava flows). Some lavas are channeled by preexisting topography, and so it is understandable that they have not spread sideways. However, others clearly stop spreading sideways even when there are no topographic obstacles, and quickly establish a pattern in which lava moves downhill in a central channel between a pair of stationary banks called levées. Also, lavas do not flow downhill indefinitely once the magma supply from the vent ceases: they commonly stop moving quite soon afterward, often while the front of the flow is on ground with an appreciable slope and almost all of the cooling lava is still at least partly liquid. Also, liquid lava present in a channel at the end of an eruption does not drain completely out of the channel: a significant thickness of lava is left in the channel floor. These observations led to the suggestion that no lavas are Newtonian, and attempts were made to model flows as the simplest non-Newtonian fluids, Bingham plastics.

The basis of these models is the idea that the finite thickness of the levées or flow front can be used to determine the yield strength of the lava and that the flow speed in the central channel can be used to give its apparent, and hence Bingham, viscosity. Multiplying the central channel width by its depth and the mean lava flow speed gives the volume flux (the volume per second) being erupted from the vent. Laboratory experiments were used to develop these ideas, and they have been applied by numerous
workers to field observation of moving flows on Earth and to images of ancient flows on other planets. For flows on Earth it is possible to deduce all of the parameters just listed; for ancient flow deposits one can obtain the yield strength unambiguously, but only the product of the viscosity and volume flux can be determined.

There is a possible alternative way to estimate the volume flux if it can be assumed that the flow unit being examined has come to rest because of cooling. An empirical relationship has been established for cooling-limited flows on Earth between the effusion rate from the vent and the length of a flow unit, its thickness, and the width of its active channel. If a flow is treated as cooling-limited when in fact it was not (the alternative being that it was volume-limited, meaning that it came to rest because the magma supply from the vent ceased at the end of the eruption), the effusion rate will inevitably be an underestimate by an unknown amount. Cooling-limited flows can sometimes be recognized because they have breakouts from their sides where lava was forced to form a new flow unit when the original flow front came to rest.

Lava rheologies and effusion rates have been estimated in this way for lava flows on Mars, the Moon, and Venus. It should be born in mind, when assessing these published estimates, that a major failing of simple models like the Bingham model is that they assign the same rheological properties to all of the material in a flow. However, lava that has resided in a stationary levée near the vent for a long period will have suffered vastly more cooling than the fresh lava emerging from the vent and will have very different properties. More elaborate models have been evolved since the earliest work, including some that apply to broadly spreading lava lobes that do not have a well-defined levée-channel structure, and others that treat the levées and central channel as separate materials with differing rheologies, but no model yet accounts for all of the factors controlling lava flow emplacement. With this caution, the rheological properties found suggest that essentially all of the lavas studied so far on the other planets have properties similar to those of basaltic to intermediate (basaltic–andesite) lavas on Earth. Many of these lavas have lengths up to several hundred kilometers, to be compared with basaltic flow lengths up to a few tens of kilometers on Earth in geologically recent times, and this implies that they were erupted at much higher volume fluxes than is now common on Earth. There is a possibility, however, that some of these flow lengths have been overestimated. If a flow comes to rest so that its surface cools, but the eruption that fed it continues and forms other flow units alongside it, a breakout may eventually occur at the front of the original flow. A new flow unit is fed through the interior of the old flow, and the cooled top of the old flow, which has now become a lava tube, acts as an excellent insulator. As a result, the breakout flow can form a new unit almost as long as the original flow, and a large, complex compound flow field may eventually form in this way. Unless spacecraft images of the area have sufficiently high resolution for the compound nature of the flows to be recognized, the total length of the group of flows will be interpreted as the length of a single flow, and the effusion rate will be greatly overestimated.

There are, however, certain volcanic features on the Moon and Mars that may be less ambiguous indicators of high effusion rates: the sinuous rilles. The geometric properties of these meandering channels—widths and depths that decrease away from the source, lengths of tens to a few hundred kilometers—are consistent with the channels being produced by very fluid lava erupted at a very high volume flux for a long time. The turbulent motion of the initial flow, meandering downhill away from the vent, led to efficient heating of the ground on which it flowed, and it can be shown theoretically that both mechanical and thermal erosion of the ground surface are expected to have occurred on a timescale from weeks to months. The flow, typically ~10 m deep and moving at ~10 m/s, slowly subsided into the much deeper channel that it was excavating. Beyond a certain distance, the lava cooled to the point where it could no longer erode the ground, and it continued as an ordinary surface lava flow. The volume eruption rates deduced from the longer sinuous rille channel lengths are very similar to those found for the longest conventional lava flow units. Modeling studies show that the turbulence leading to efficient thermal erosion was probably encouraged by a combination of unusually steep slope and unusually low lava viscosity. A few sinuous channels associated with lava plains are visible on Venus, but the lengths of some of the Venus channels are several to 10 times as great as those seen on the Moon and Mars. It is not yet clear if the thermal erosion process is capable of explaining these channels by the eruption of low-viscosity basalts, or whether some more exotic volcanic fluid (or some other process) was involved.

There are numerous uncertainties in using the foregoing relationships to estimate lava eruption conditions. Thus, there have been many studies of the way heat is transported out of lava flows, taking account of the porosity of the lava generated by gas bubbles, the effects of deep cracks extending inward from the lava surface, and the external environmental conditions—the ability of the planetary atmosphere to remove heat lost by the flow by conduction and convection, and by radiation (whether or not an atmosphere is present). However, none of these has yet dealt in sufficient detail with turbulent flows, or with the fact that cooling must make the rheological properties of a lava flow a function of distance inward from its outer surface, so that any bulk properties estimated in the ways described earlier can only be approximations to the detailed behavior of the interior of the lava flow. There is clearly
some feedback between the way a flow advances and its internal pattern of shear stresses. For example, lava flows on Earth have two basic surface textures. Basaltic flows erupted at low effusion rates or while still hot near their vents have smooth, folded surfaces with a texture called pahoehoe (a Hawaiian word), the result of plastic stretching of the outer skin as the lava advances; at higher effusion rates, or at lower temperatures farther from the vent, the surface fractures in a more brittle fashion to produce a very rough texture called ‘a‘ā. A similar but coarser, rough, blocky texture is seen on the surfaces of more andesitic flows. Because there is a possibility of relating effusion rate and composition to the surface roughness of a flow in this way, there is a growing interest in obtaining relatively high resolution radar images of planetary surfaces (and Earth’s surface) in which, as in the Magellan images of Venus, the returned signal intensity is a function of the small-scale roughness.

4. EXPLOSIVE ERUPTIONS

4.1. Basic Considerations

Magmas ascending from the mantle on Earth commonly contain volatiles, mainly water and carbon dioxide together with sulfur compounds and halogens. All of these have solubilities in the melt that are both pressure and temperature dependent. The temperature of a melt does not change greatly if it ascends rapidly enough toward the surface, but the pressure to which it is subjected changes enormously. As a result, the magma generally becomes saturated in one or more of the volatile compounds before it reaches the surface. Only a small degree of supersaturation is needed before the magma begins to exsolve the appropriate volatile mixture into gas bubbles, especially if the magma contains unmelted crystals on which bubbles can nucleate. As a magma ascends to shallower levels, existing bubbles grow by decompression and new ones form. It is found empirically that once the volume fraction of the magma occupied by the bubbles exceeds some value in the range 65—80%, the foam-like fluid can no longer deform fast enough in response to the shear stresses applied to it and as a result disintegrates into a mixture of released gas and entrained clots and droplets that form the pyroclasts. The eruption is then, by definition, explosive. The pyroclasts have a range of sizes dictated by the viscosity of the magmatic liquid, in turn a function of its composition and temperature, the rate at which the decompression is taking place, essentially proportional to the rise speed of the magma, and the rate at which the magma is being sheared, a function of its rise speed and the conduit width.

It is not a trivial matter for the volume fraction of gas in a magma to become large enough to cause disruption into pyroclasts. The lowest pressure to which a magma is ever exposed is the planetary surface atmospheric pressure. On Venus this ranges from about 10 MPa in lowland plains to about 4 MPa at the tops of the highest volcanoes; on Earth it is about 0.1 MPa at sea-level (and 30% less on high volcanoes) but much higher, up to 60 MPa, on the deep ocean floor; on Mars it ranges from about 500 Pa at the mean planetary radius to about 50 Pa at the tops of the highest volcanoes; and it is essentially zero on the Moon, Mercury and Io. If the magma volatile content is small enough, then even at atmospheric pressure too little gas will be exsolved to cause magma fragmentation. Using the solubilities of common volatiles in magmas, calculations show that explosive eruptions can occur on Earth as long as the water content exceeds 0.07 wt% in a basalt. On Mars the critical level is 0.01 wt%. On Venus, however, a basalt would have to contain about 2 wt% water before explosive activity could occur, even at highland sites; this is greater than is common in basalts on Earth, and leads to the suggestion that explosive activity may never happen on Venus, at least at lowland sites, or may happen only when some process leads to the local concentration of volatiles within a magma. Examples of this are discussed later. Finally, the negligible atmospheric pressures on the Moon, Mercury and Io mean that miniscule amounts of magmatic volatiles can in principle cause some kind of explosive activity there.

The above discussion assumes that released magmatic volatiles are the only source of explosive activity. However, many Vulcanian and phreato-magmatic explosive eruptions involve interaction of erupting magma with solid or liquid volatiles already present at the surface (almost always water or ice on Earth and Mars; mainly sulfur compounds on Io). The total weight fraction of gas in the eruption products in such cases will depend on the detailed nature of the interaction as well as the composition and inherent volatile content of the magma; this is a critical factor in understanding the very explosive activity on Io.

4.2. Strombolian Activity

Strombolian eruptions, named for the style of activity common on the Italian volcanic island Stromboli, are an excellent example of how the rise speed, gas content, and viscosity of a magma are critical in determining the style of explosive activity that occurs. While the magma as a whole is ascending through a fracture in the planetary crust, bubbles of exsolved gas are rising through the liquid at a finite speed determined by the liquid viscosity and the bubble sizes. If the magma rise speed is negligible, for example, when magma is trapped in a shallow reservoir or a shallow intrusion, and if its viscosity is low, as in the case of a basalt, there may be enough time for gas bubbles to rise completely through the magma and escape into overlying fractures that convey the gas to the surface, where it escapes or is added to the atmosphere if there is one. Subsequent
eruption of the residual liquid will be essentially perfectly effusive. If a low-viscosity magma is rising to the surface at a slow enough speed, most of the gas will still escape as bubbles rise to the liquid surface and burst. Because relatively large bubbles (those that nucleated first and have decompressed most) will rise through the liquid faster than very small bubbles, it is common in basalts for large bubbles to overtake and coalesce with small ones. The even larger bubbles produced in this way rise even faster and overtake additional smaller bubbles. A runaway situation can develop in which a single large bubble completely fills the diameter of the vent system apart from a thin film of magma lining the walls of the fracture. In extreme cases the bubble may have a much greater vertical extent than its width, in which case it is called a slug of gas. As this body of gas emerges at the surface of the slowly rising liquid magma column, it bursts, and a discrete layer of magma forming the upper “skin” of the bubble or slug disintegrates into clots and droplets up to tens of cm in size. These are blown outward by the expanding gas (Figure 5.9; see also color insert). The pyroclasts produced accumulate around the vent to form a cinder cone that can be up to several tens of meters in size. The time interval between the emergence of successive bubbles or slugs from a vent may range from seconds to at least minutes, making this a distinctly intermittent type of explosive activity. If the largest rising gas bubble does not completely fill the vent, continuous overflow of a lava lake in the vent may take place to form one or more lava flows at the same time that intermittent explosive activity is occurring, resulting in a simultaneously effusive and explosive eruption.

A second method of producing gas slugs has been suggested for some Strombolian eruptions on Earth, in which gas bubbles form during convection in a body of magma beneath the surface and drift upward to accumulate into a layer of foam at the top of the magma body. When the vertical extent of the foam layer exceeds a critical value it collapses. Liquid magma drains from between the bubbles, which coalesce into a large gas pocket that can now rise through any available fracture to the surface. The argument is that if a fracture had been already present, the high effective viscosity of the foam would have inhibited its rise into the fracture, whereas the viscosity of the pure gas is low enough to allow this to occur. If a fracture was not already present, the changing stresses due to the foam collapse may be able to create one.

As long as any volatiles are exsolved from a low-viscosity magma rising sufficiently slowly to the surface, some kind of Strombolian explosive activity, however feeble, should occur at the vent on any planet, even at the high pressures on Venus or on Earth’s ocean floors (where there is now evidence for such activity from submersible vehicles). Strombolian eruptions commonly involve excess pressures in the bursting bubbles of only a few tenths of a MPa, so that the amount of gas expansion that drives the dispersal of pyroclasts is small. Pyroclast ranges in air on Earth can be several tens to at most a few hundred meters, and ranges would be much smaller in submarine Strombolian events on the Earth’s ocean floor or on Venus because of the higher ambient pressure. Strombolian eruptions on Mars would eject pyroclasts to distances about five times greater than on Earth because of the lower gravity and atmospheric pressure; as a result the deposits formed would have a 25-fold lower relief than on Earth, and perhaps as a result no examples have yet been unambiguously identified in spacecraft images.

4.3. Vulcanian Activity

In a slowly rising viscous magma, it is relatively difficult for gas bubbles to escape from the melt. Particularly if the magma stalls as a shallow intrusion, slow diffusion of gas and rise of bubbles in the liquid concentrate gas in the upper part of the intrusion, and the gas pressure in this region rises. The pressure rise is greatly enhanced if any volatiles existing near the surface (groundwater on Earth; ground ice on Mars; sulfur or sulfur dioxide on Io) are evaporated. Eventually the rocks overlying the zone of high pressure break under the stress and the rapid expansion of the trapped gas drives a sudden, discrete explosion in which fragments of the overlying rock and disrupted magma are scattered around the vent: this is called Vulcanian activity.
named for the Italian volcanic island Vulcano. Again, as long as any volatiles are released from the magma or are present in the near-surface layers of the planet, activity of this kind can occur. Several Vulcanian events on Earth involving fairly viscous magmas have been analyzed in enough detail to provide estimates of typical pressures and gas concentrations. Bombs approaching a meter in size thrown as far as 5 km imply pressures as high as a few MPa in regions that are tens of meters in size. The gas mass fractions in the explosion products can be up to 10%.

On Mars, with the same initial conditions, the lower atmospheric pressure would cause much more gas expansion to accelerate the ejected fragments, and the lower atmospheric density would exert much less drag on them; also the lower gravity would allow them to travel farther for a given initial velocity. The result is that the largest clasts could travel up to 50 km. This means that the roughly circular deposit from a localized, point-source explosion would be spread over an area 100 times greater than on Earth, being on average 100 times thinner. Apart from the possibility that the pattern of small craters produced by the impact of the largest boulders on the surface might be recognized, such a deposit, with almost no vertical relief and having very little influence on the preexisting surface, would almost certainly go unnoticed in even the highest-resolution spacecraft images, and indeed no such features have yet been identified. However, if the explosion involves a larger, more complex, and especially elongate vent structure, there would not be such large differences. In the Elysium region of Mars a large, water-carved channel, Hrad Vallis, has a complex elongate source depression that appears to have been excavated by a Vulcanian explosion when a dike injected a sill into the ice-rich permafrost of the cryosphere—the outer few kilometers of the crust which is so cold that any H₂O must be present as ice. As heat from the sill magma melted the ice and boiled the resulting water in the cryosphere, violent expansion of the vapor forced intimate mixing of magma and lumps of cryosphere, encouraging ever more vapor production. Soon all of the cryosphere above the sill was thrown out in what is called a fuel—coolant interaction (here the fuel is the magma and the coolant is the ice) to produce a deposit extending about 35 km on either side of the 150 km-long depression. Residual heat from the magma melted the remaining ice in the shattered cryosphere rocks so that for a while, until it froze again, there was liquid water present to form a characteristic “muddy” appearance in the deposit (Figure 5.11).

A Vulcanian explosion on Venus would also be very different from its equivalent on Earth. In this case, the high atmospheric pressure would tend to suppress gas dispersion, allowing the deposit to be more confined.
expansion, producing a low initial velocity in the ejecta, and the atmospheric drag would also be high. Pyroclasts that would have reached a range of 5 km on Earth would travel less than 200 m on Venus. On the one hand, this should concentrate the eruption products around the vent and make the deposit more obvious; however, the resolution of the best radar images from Magellan is only ~75 m, and so such a deposit would represent only three or four adjacent pixels, which again would probably not be recognized.

On the Moon a number of Vulcanian explosion products are seen. The dark halo craters on the floor of the impact crater Alphonsus have ejecta deposits with ranges up to 5 km. Since the Moon has no atmosphere, the preceding arguments suggest that lunar Vulcanian explosions should eject material to very great ranges. However, the Alphonsus event seems to have involved the intrusion of basaltic magma into the ~10 m thick layer of fragmental material forming the regolith in this area, and the strength of the resulting mixture of partly welded regolith and chilled basalt was quite low. Thus only a small amount of pressure buildup occurred before the retaining rock layer fractured. As a result, the initial speeds of ejected pyroclasts were low and their ranges were unusually small.

4.4. Hawaiian Activity

In some cases, especially where low viscosity basaltic magma travels laterally in dikes at shallow depth, enough gas bubble coalescence and bubble rise occurs for much of the gas to be lost into cracks in the rocks above the dike. Magma then emerges from the vent as a lava flow. However, when basaltic magmas rise mainly vertically at appreciable rates (more than about 1 m/s), some gas bubble coalescence occurs but little gas is lost, and the magma is released at the vent in a nearly continuously explosive manner. A lava fountain, more commonly called a fire fountain, forms over the vent, consisting of pyroclastic clots and droplets of liquid entrained in a magmatic gas stream that fluctuates in its upward velocity on a timescale of a few seconds. The largest clots of liquid, up to tens of cm in size, rise some way up the fountain and fall back around the vent to coalesce into a lava pond that overflows to feed lava flows—the effusive part of the eruption—whereas smaller clasts travel to greater heights in the fountain. Some of the intermediate-sized pyroclasts cool as they fall from the outer parts of the fountain and collect around the lava pond in the vent to build up a roughly conical edifice called an ash cone, cinder cone, or scoria cone, the term used depending on the sizes of the pyroclasts involved, ash being smallest. Such pyroclastic cones are commonly asymmetric owing to the influence of the prevailing wind.

Atmospheric gases are entrained into the edge of the fire fountain and heated by contact with the hot pyroclasts and mixing with the hot magmatic gas. In this way, a convecting gas cloud is formed over the upper part of the fountain, entraining the smallest pyroclasts so that they take part fully in the convective motion. The whole cloud spreads downwind and cools, and eventually the pyroclasts are released again to form a layer on the ground, the smallest particles being deposited at the greatest distances from the vent. This whole process, involving formation of lava flows and pyroclastic deposits at the same time, is called Hawaiian eruptive activity (Figure 5.12). This style of activity should certainly have occurred on Mars, but may be suppressed in basaltic magmas on Venus by the high atmospheric pressure, especially in lowland areas, unless, as noted earlier, magma volatile contents are several times higher than is common on Earth.

Figure 5.13 shows qualitatively how the combination of erupting mass flux and magma gas content in a Hawaiian eruption on Earth determines the nature and size of the possible products: a liquid lava pond at the vent that directly feeds lava flows; a pile of slightly cooled pyroclasts accumulating fast enough to weld together and form a “rootless” lava flow; a cone in which almost all of the pyroclasts are welded together; or a cone formed from pyroclasts that have had time to cool while in flight so that none, or only a few, weld on landing. Theoretical analyses based on the trends seen in Figure 5.13 confirm that hot lava ponds around vents on Earth are expected to be no more than a few tens of meters wide even at very high mass eruption rates. On the Moon, the greater gas expansion due to the lack of an atmosphere causes very thorough disruption of the magma (even at the low gas contents implied by analysis of the Apollo samples) and
gives the released volcanic gas a high speed. This, together with the lower gravity, allows greater dispersal of pyroclasts of all sizes, and provides an explanation of the 100—300 km wide dark mantle deposits as the products of extreme dispersal of the smallest, 30—100 μm sized particles.

Nevertheless, it appears that hot lava ponds up to ~5 km in diameter could have formed around basaltic vents on the Moon if the eruption rates were high enough—as high as those postulated to explain the long lava flows and sinuous rilles. The motion of the lava in such ponds would have been thoroughly turbulent, thus encouraging thermal erosion of the base of the pond, and this explains why the circular to oval depressions seen surrounding the sources of many sinuous rilles have just these sizes. Similar calculations for the Mars environment show that, as long as eruption rates are high enough, the atmospheric pressure and gravity are low enough on Mars to allow similar hot lava source ponds to have formed there, again in agreement with the observed sizes of depressions of this type that are seen.

Some noticeable differences occur when Hawaiian eruptions take place from very elongate fissure vents. Instead of a roughly circular pyroclastic cone containing a lava pond feeding one main lava flow, a pair of roughly parallel ridges forms, one on either side of the fissure, called spatter ramparts. Along the parts of the fissure where the eruption rate is highest, pyroclasts may coalesce as they land to form lava flows, so that there are gaps in the ramparts from which the flows spread out. A striking example of this has been found on Mars (Figure 5.14).

4.5. Plinian Activity

In the case of a basaltic magma very rich in volatiles, or (much more commonly on Earth) in the case of a volatile-rich andesitic or rhyolitic magma, fragmentation in a steadily erupting magma is very efficient, and most of the pyroclasts formed are small enough to be entrained by the gas stream. Furthermore, the speed of the mixture emerging from the vent, which is proportional to the square root of the amount of gas exsolved from the magma, will be much higher (perhaps up to 500 m/s) than in the case of a basaltic Hawaiian eruption (where speeds are commonly less than 100 m/s). The fire fountain in the vent now entrains so much atmospheric gas that it develops into a very strongly convecting eruption cloud in which the heat content of the pyroclasts is converted in the buoyancy of the entrained gas. The resulting cloud rises to a height that is proportional to the fourth root of the magma eruption rate (and hence the heat supply rate). Such clouds may reach heights of several tens of kilometers on Earth. Only the very coarsest pyroclasts fall out near the vent, and almost all of the erupted material is dispersed over a wide area from the higher parts of the eruption cloud (Figure 5.15). This activity is termed Plinian, after Pliny’s description of the A.D. 79 eruption of Vesuvius.

Not all eruptions of this type produce stable convection clouds. If the vent is too wide or the eruption speed of the magma is too low, insufficient atmospheric gas is entrained to provide the necessary buoyancy for convection, and a
collapsed fountain forms over the vent, feeding large pyroclastic density currents or smaller, more episodic pyroclastic surges.

Mars is the obvious place other than Earth to look for explosive eruption products: the low atmospheric pressure encourages explosive eruptions to occur and the atmospheric density is high enough to allow convecting eruption clouds to form, at least up to ~20 km. Stable eruption clouds much higher than this cannot form on Mars because the atmosphere becomes too thin to provide the required amount of entrained gas. Nevertheless, the smaller sizes expected for pyroclasts on Mars than Earth mean that winds can transport particles for great distances. Very extensive friable layered deposits are seen in the Arabia Terra and Terra Meridiani areas and in the Medusae Fossae formation. Computations combining eruption cloud formation models with global atmospheric circulation models have shown that large explosive eruptions from the major volcanoes are readily able to explain these deposits.

Although the large magma gas contents needed suggest that large-scale, steady (Plinian) explosive eruptions are rare on Venus, it is possible to calculate the heights to which their eruption clouds would rise. The high density and temperature of the atmosphere lead to rise heights about a factor of two smaller than on Earth for the same eruption rate, and very large (at least a few tens of meters) clasts may be transported into near-vent deposits. At distances greater than a few kilometers from the vent, pyroclastic fall deposits will not be very different from those on Earth. A few examples of elongate markings on the Venus surface have been proposed as fall deposits, but no detailed analysis of them has yet been carried out.

The conditions that cause a steady explosive eruption to generate pyroclastic density currents instead of feeding a stable, convecting eruption cloud are fairly well understood. If the eruption rate exceeds a critical value (which increases with increasing gas content of the mixture emerging through the vent and decreases with increasing vent diameter), stable convection is not possible whatever the nature of the atmosphere. Since pyroclastic density current formation is linked to high eruption rate and, in general, to high eruption speed, which will encourage a great travel distance, it would not be surprising if such large-scale pyroclastic deposits distributed radially around a vent were the products of high discharge rate eruptions of gas-rich magmas. Many of the flanking deposits of some Martian volcanoes, especially Tyrrenus Mons and Hadriacus Mons, may have been produced in this way.

Short-lived or intermittent explosive eruptions (e.g. Vulcanian explosions, phreato-magmatic explosions, or events in which a gas-rich, high-viscosity lava flow or dome disintegrates into released gas and pyroclasts as a result of excessive gas pressure) can also produce small-scale pyroclastic density currents. Because these are shorter-lived and have characteristically different grain size distributions, they are called surges. The least well understood aspect of these phenomena is the way in which the magmatic material interacts with the overlying atmosphere. As a result, it is currently almost impossible to predict in detail what the results of this kind of activity on Mars or Venus would look like. Such deposits, by the nature of the way they are generated, would not be very voluminous, however, and so would be spread very thinly, and might not be recognized if they were able to travel far from the vent.

4.6. Phreato-Magmatic Activity

Some types of eruption on Earth are controlled by the vigorous interaction of magma with surface or shallow subsurface water. If an intrusion into water-rich ground causes steam explosions, these are called phreatic events (from the Greek word for a well). If some magma also reaches the surface, the term used is phreato-magmatic, as distinct from normal, purely magmatic eruptions. When the equivalents of Strombolian or Hawaiian explosive events take place from eruption sites located in shallow water, they lead to much greater fragmentation of the magma than usual because of the thermal stresses induced as pyroclasts are chilled by contact with the water. This activity is usually called Surtseyan, named after an eruption that formed the island of Surtsey off the south coast of Iceland. A much more vigorous and long-lived eruption under similar circumstances leads to a pyroclastic fall deposit similar to that
of a Plinian event, but again involving greater fragmentation of magma; the result is called phreatic-Plinian activity. Since the word phreatic does not specifically refer to water as the nonmagmatic volatile involved in these kinds of explosive eruption, it seems safe to apply these terms, as appropriate, to the various kinds of interactions between magma and liquid sulfur or sulfur dioxide forming the plumes currently seen on Io. These eruptions appear to involve about 30% by weight volatiles mixed with the magma; these proportions are close to the optimum for converting the heat of the magma to kinetic energy of the explosion products. Phreatic and phreato-magmatic eruptions should also have occurred on Mars in the distant past if, as many suspect, the atmospheric pressure was high enough to allow liquid water to exist on the surface.

4.7. Dispersal of Pyroclasts into a Vacuum

The conditions in the region above the vent in an explosive eruption on a planet with an appreciable atmosphere (e.g. Venus, Earth, and Mars) are very different from those when the atmospheric pressure is very small (much less than about 1 Pa), as on the Moon, Mercury and Io (and differentiated asteroids in the distant past). If the mass of atmospheric gas displaced from the region occupied by the eruption products after the magmatic gas has decompressed to the local pressure is much less than the mass of the magmatic gas, convecting eruption clouds cannot form in eruptions that would have been classed as Hawaiian or Plinian on Earth. In the region immediately above the vent, the gas expansion involves a series of shock waves. Relatively large pyroclasts will pass through these shocks with only minor deviations in their trajectories, but intermediate-sized particles may follow very complex paths, and few studies have yet been made of these conditions. The magmatic gas eventually expands radially into space, accelerating to reach a limiting velocity that depends on its initial temperature. As the density of the gas decreases, its ability to exert a drag force on pyroclasts also decreases. On bodies the size of the Moon, even the smallest particles eventually decoupled from the gas and fell back to the planetary surface, though in gas-rich eruptions on asteroids these particles were commonly ejected into space.

These are the conditions that led to the formation of the ancient dark mantle deposits on the Moon, with ultimate gas speeds on the order of 500 m/s, leading to ranges up to 150 km for small pyroclasts 30–100 μm in size. They are also the conditions that exist now in the eruption plumes on Io, though with an added complication. The driving volatiles in the Io plumes appear to be mainly sulfur and sulfur dioxide, evaporated from the solid or liquid state by intimate mixing with rising basaltic magma in what are effectively phreato-magmatic eruptions. The Io plume heights imply gas speeds just above the vent of ~1000 m/s, and these speeds are consistent with the plume materials being roughly equal mixtures of basaltic pyroclasts and evaporated surface volatiles. As the gas phase expands to very low pressures, both sulfur and sulfur dioxide will condense, forming small solid particles that rain back onto the surface along with the silicate pyroclasts to be recycled again in future eruptions.

A final point concerns pyroclastic eruptions on the smallest atmosphereless bodies, the asteroids. Basaltic partial melts formed within these bodies were erupted at the surface at speeds that depended on the released volatile content. This is estimated to have been as much as 0.2–0.3 wt%, leading to speeds up to 150 m/s. These speeds are greater than the escape velocities from asteroids with diameters less than about 200 km, and so instead of falling back to the surface, pyroclasts would have been expelled into space, eventually to spiral into the Sun. This process explains the otherwise puzzling fact that we have many meteorites (e.g. the aubrites and ureilites) representing samples of the residual material left in the mantle of at least two asteroids after partial melting events, but have only a tiny number of meteorites from these asteroids containing grains with the expected partial melt composition.

5. INFERENCES ABOUT PLANETARY INTERIORS

The presence of the collapse depressions called calderas at or near the summits of many volcanoes on Earth, Mars, Venus, and Io suggests that it is common on all of these bodies for large volumes of magma to accumulate in reservoirs at relatively shallow depths. Theories of magma accumulation suggest that the magma in these reservoirs must have an internal pressure greater than the stress produced in the surrounding rocks by the weight of the overlying crust. This excess pressure may be due to the formation of bubbles by gas exsolution, or to the fact that heat loss from the magma to its cooler surroundings causes the growth of crystals that are less dense than the magmatic liquid and so occupy a larger volume. Most commonly, a pressure increase leads to fracturing of the wall of the reservoir and to the propagation of a magma-filled crack, called a dike, as an intrusion into the surrounding rocks. If the dike reaches the surface, an eruption occurs, and removal of magma from the reservoir allows the wall rocks to relax inward elastically as the pressure decreases. If magma does not reach the surface, the dike propagates underground until either the magma within it chills and comes to rest, or the pressure within the reservoir falls to the point where there is no longer a great enough stress at the dike tip for rock fracturing to continue.

Under certain circumstances, an unusually large volume of magma may be removed from a shallow reservoir,
reducing the internal pressure beyond the point where the reservoir walls behave elastically. Collapse of the overlying rocks may then occur to fill the potential void left by the magma, and a caldera (or, on a smaller scale, a pit crater) will form. The circumstances causing large-volume eruptions on Earth include the rapid eruption to the surface immediately above the reservoir of large volumes of low-density, gas-rich silicic (rhyolitic) magma, and the drainage of magma through extensive lateral dike systems extending along rift zones to distant flank eruption sites on basaltic volcanoes. This latter process appears to have been associated with caldera formation on Kilauea volcano in Hawaii, and it is tempting to speculate that the very large calderas on some of the Martian basaltic shield volcanoes (especially Pavonis Mons and Arsia Mons) are directly associated with the large-volume eruptions seen on their flanks.

The size of a caldera must be related to the volume of the underlying magma reservoir, or more exactly to the volume of magma removed from it in the caldera-forming event. If the reservoir is shallow enough, the diameter of the caldera is probably similar to that of the reservoir. Diameters from 1 to 3 km are common on basaltic volcanoes on Earth and on Venus, with depths up to a few hundred meters implying magma volumes less than about 10 km³. In contrast, caldera diameters up to at least 30 km occur on several volcanoes on Mars and, coupled with caldera depths up to 3 km, imply volumes ranging up to as much as 10,000 km³. The stresses implied by the patterns of fractures on the floors and near the edges of some of these Martian calderas suggest that the reservoirs beneath them are centered on depths of the order of 10–15 km, about three to four times greater than the depths of basaltic reservoirs on Earth. The simplest models of the internal structures of volcanoes suggest that, due to the progressive closing of cavities in rocks as the pressure increases, the density of the rocks forming a volcanic edifice should increase, at first quickly and then more slowly, with depth. Rising magma from deep partial melt zones may stall when its density is similar to that of the rocks around it, so that it is neither positively nor negatively buoyant, and a reservoir may develop in this way. Since the pressure at a given depth inside a volcano is proportional to the acceleration due to gravity, and since on Mars this is about three times less than that on Earth or Venus, the finding that Martian magma reservoirs are centered three to four times deeper than on Earth is not surprising. However, these simple models do not address the reason for the Martian calderas being very much wider than any of those on Earth or many of those on Venus. On Io we see some caldera-like structures, not necessarily associated with obvious volcanic edifices, that are even wider (but not deeper) than those on Mars, though we have too little information about the internal structure of Io’s crust to interpret this observation unambiguously. Clearly, much is still not understood about the formation and stability of shallow magma bodies.

Evidence for significant shallow magma storage is very rare on the Moon. The large volumes observed for the great majority of eruptions in the later part of lunar volcanic history, and the high effusion rates inferred for them, imply that almost all of the eruptions took place directly from large bodies of magma stored at very great depth—at least at the base of the crust and possibly in partial melting zones in the lunar mantle. Not all the dikes propagating up from these depths will have reached the surface, however, and some shallow dike intrusions almost certainly exist. Recent work suggests that many of the linear rilles on the Moon represent the surface deformation resulting from the emplacement of such dikes, having thicknesses of at least 100 m, horizontal and vertical extents of ~100 km, and tops extending to within 1 or 2 km of the surface. Minor volcanic activity associated with some of these features, as in the case of Rima Hyginus, would then be the result of gas loss and small-scale magma redistribution as the main body of the dike cooled.

The emplacement of very large dike systems extending most or all of the way from mantle magma source zones to the surface is not confined to the Moon. It has long been assumed that such structures must have existed to feed the high-volume basaltic lava flow sequences called flood basalts that occur on Earth every few tens of millions of years. These kinds of feature are probably closely related to the systems of giant dikes, tens to hundreds of meters wide and traceable laterally for many hundreds to more than 1000 km, that are found exposed in very ancient rocks on the Earth. The radial patterns of these ancient dike swarms suggest that they are associated with major areas of mantle upwelling and partial melting, with magma migrating vertically above the mantle plume to depths of a few tens of kilometers and then traveling laterally to form the longest dikes. Some of the radial surface fracture patterns associated with the novae and coronae on Venus are almost certainly similar features that have been formed more recently in that planet’s geologic history. On Mars the systems of linear grabens, some of which show evidence of localized eruptive vents, extending radially from large shield volcanoes such as Arsia Mons, also bear witness to the presence of long-lived mantle upwellings generating giant dike swarms. It seems that there may be a great deal of similarity between the processes taking place in the mantles of all of the Earth-like planets; it is the near-surface conditions, probably strongly influenced by the current presence of its oceans, that drive the plate tectonic processes distinguishing the Earth from its neighbors.
BIBLIOGRAPHY


Chapter 6

Magnetic Field Generation in Planets

Sabine Stanley
Department of Physics, University of Toronto, Toronto, ON, Canada

Chapter Outline

1. Planetary Magnetic Field Observations 121
   1.1. Sources of Observed Magnetic Fields 121
   1.2. Spatial Characteristics of Dynamo-Generated Fields 122
   1.3. Temporal Characteristics of Observed Magnetic Fields 125
2. The Dynamo Mechanism 125
   2.1. What is a Dynamo? 125
   2.2. Necessary Conditions for a Dynamo 126
   2.3. Dynamo Generation Regions in Planets 126
3. The Standard Planetary Dynamo 127
   3.1. Driving Forces 127
   3.2. Fluid Motions in Dynamo Regions 127
   3.3. Generation Mechanisms 127
   3.4. Beyond the Standard Dynamo 129
4. Simulations and Experiments 129
   4.1. Numerical Dynamo Simulations 129
   4.2. Dynamo Experiments 131
5. Planetary Dynamos 131
   5.1. Earth 131
   5.2. Mercury 133
   5.3. Jupiter 133
   5.4. Saturn 133
   5.5. Uranus and Neptune 134
   5.6. Ganymede 134
   5.7. Ancient Moon 134
   5.8. Ancient Mars 134
   5.9. Small Bodies 135
   5.10. Extrasolar Planets 135
   5.11. Planetary Bodies Lacking Dynamos 135
6. Conclusions and Future Prospects 135
Bibliography 135

1. PLANETARY MAGNETIC FIELD OBSERVATIONS

The Earth’s magnetic field has been used for navigation since at least the eleventh century AD, but it was not until the seventeenth century that the source of this magnetic field was attributed to the Earth’s interior. Namely, in 1600, William Gilbert published “De Magnete” in which he described his experiments involving magnetic measurements of a lodestone sphere. He concluded that “Globus terrae sid magneticus & magnes” which can be loosely translated as “The Earth is a great magnet”. Further work by Gellibrand, Halley, Gauss, and others established that the Earth’s field was predominantly axially dipolar, but varied in time (see Kono, 2007 for a nice overview of the history of geomagnetism).

It was not until the mid-twentieth century that magnetic fields of other planets were observed. The first observation was indirectly of Jupiter’s magnetic field in the 1950s through its radio emissions. These emissions result from interactions of the solar wind with the planet’s magnetosphere (see chapter on Planetary Magnetospheres). The radio emissions were intense enough and in an appropriate bandwidth such that they could be detected from the Earth. Discovery of other planetary magnetic fields awaited visits by planetary spacecraft missions with magnetometers. Table 6.1 provides highlights of magnetic missions to planetary bodies and their main magnetic discoveries.

1.1. Sources of Observed Magnetic Fields

Observed planetary magnetic fields can result from a variety of processes. For example, external sources that generate observed planetary magnetic fields include Venus’ ionospheric currents, Earth’s magnetospheric currents, and electromagnetic induction in the saltwater oceans of Europa, Ganymede, and Callisto due to
Jupiter’s time varying field. In contrast, internal sources include remanent magnetization in crustal rocks on Mars, Earth, and the Moon or self-sustaining dynamos in the deep interiors of planets. This latter process will be the main focus of this chapter.

1.2. Spatial Characteristics of Dynamo-Generated Fields

Dynamo-generated magnetic fields are typically distinguished from other sources by their spatial and temporal
characteristics. Dynamo-generated fields are global in structure and vary on timescales related to the fluid motions in the planetary interior. Figure 6.1 shows the radial component of the magnetic field at the surface of the planets with actively generated dynamos. Jupiter’s moon Ganymede also has a dynamo-generated field, but the data can only constrain the dipole moment. Hence a map of its surface magnetic field is not included in the figure. The planetary fields likely contain much smaller scale structure that has not yet been resolvable by available data; however, based on the large-scale fields, major similarities and differences between planetary magnetic fields are obvious.

The surface magnetic fields of Mercury, Earth, Jupiter, Saturn, and possibly Ganymede are dominated by their axial dipolar components. In contrast, fields of Uranus and Neptune do not show this dominance and instead, higher order multipoles and nonaxisymmetric components are as prominent as the axial dipole component. In terms of secondary features, one also notices that Saturn’s observed field is purely axisymmetric (i.e. there is no variation in the zonal direction) and Mercury’s field has a fairly large northward offset between its geographic equator and magnetic equator compared to the other planets.

In order to analyze the spectral components of the field more quantitatively, the surface magnetic field can be represented using surface spherical harmonics. By assuming that observations are made in a current-free region (i.e. that current $\mathbf{j} = 0$), Ampere’s law implies

![Figure 6.1](image)

**Figure 6.1** Surface radial magnetic field of planets with active dynamos. Units are $\mu T$. 
that the magnetic field $\mathbf{B}$ is solenoidal (i.e. curl free) and hence that it can be written as the gradient of a scalar potential $V$:

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} = 0 \Rightarrow \mathbf{B} = -\nabla V$$

(where $\mu_0$ is the magnetic permeability of free space). Combining this with Gauss’ law for magnetism yields the result that the magnetic scalar potential is the solution to Laplace’s equation:

$$\nabla \cdot \mathbf{B} = 0 \Rightarrow \nabla^2 V = 0.$$

Ignoring any external field sources, the potential can be written in spherical coordinates as:

$$V(r, \theta, \phi) = a \sum_{l=1}^{\infty} \frac{(a/r)^{l+1}}{r} \sum_{m=-l}^{l} \left[g^m_l \cos(m\phi) + \frac{1}{l} h^m_l \sin(m\phi)\right] P_l^m(\cos \theta)$$

where $l$ and $m$ are the spherical harmonic degree and order, respectively; $r$ is the radius; $\phi$ is the longitude; $\theta$ is the colatitude; $a$ is the planetary radius; and $P_l^m$ are the associated Legendre polynomials. In this expansion, axisymmetric terms are given by $m = 0$ terms, and the successive multipoles are determined by $l$. For example, $l = 1$, 2, and 3 represent the dipole, quadrupole, and octupole components respectively. The amplitudes of each harmonic are given by their respective Gauss coefficients $g^m_l$ and $h^m_l$.

Perhaps the easiest way to visualize the similarity in spectral content of the different planetary magnetic fields is through power spectra. Defining the power in each degree and order using the mean-square field intensity:

$$p(l, m, r) = (l + 1)(a/r)^{2l+4} \left[(g^m_l)^2 + (h^m_l)^2\right]$$

the power in each degree is found by summing over all orders and the power in each order can be found by summing over all degrees. Figure 6.2 plots the power as a function of degree and order for the planets in Figure 6.1. Since the purpose of this plot is to compare between the planets, only the lowest degrees (i.e. largest length scales) are plotted (maximum degree up to three). Observational data for Earth provide spectra to much higher degree, a recent model can be found in Finlay et al. (2010).

The equations above for the magnetic scalar potential and power depend on the distance from the source region. It is common practice to calculate the Gauss coefficients and power spectra at the respective planetary surface radius because we are limited to making observations outside the planet. However, if our goal is to compare and contrast planetary magnetic fields, then it is more appropriate to choose the dynamo source region radius since this removes the arbitrary differences in distance between the surface and the dynamo source regions for the planets.

There is an inherent danger in extrapolating the field deeper in the planet due to four factors. First, the smaller scale fields will increase in power much faster than the larger scale fields. Since the smaller scale fields are the least resolved at the surface, this can result in significant errors in the extrapolation. Second, to extrapolate using the potential field expansion in Gauss coefficients given above, the region between the surface and the top of the dynamo source region must be an insulator. Any significant electrical conductivity will introduce errors into the field extrapolation. Third, any sources of magnetism between the surface and dynamo source region (e.g. crustal magnetism in terrestrial planets) needs to be accounted for if we are only interested in the dynamo-generated field. Fourth, the radius of the top of the
dynamo source region is not well known for all the planets. This is especially a problem for the giant planets since they experience a gradual increase in conductivity with depth without a significant compositional change.

Being aware of these limitations, we tentatively plot the power spectra at the top of the dynamo source regions in Figure 6.3. The radii of the dynamo source regions used for the figure are given in Table 6.2.

Figure 6.3 demonstrates that the fields at the top of the dynamo source region are not as dipolar dominated as the surface fields, but Uranus and Neptune are the only planets for which the dipole ($l = 1$) and axisymmetric ($m = 0$) components are not the largest contributors to the spectra. The planet whose magnetic field spectrum most resembles that of Earth is Jupiter. Of the planets with a dominant axial dipole, Mercury’s field seems to have a relatively large quadrupole, whereas Saturn appears to have a relatively large octupole. Based on current data, both Saturn and Mercury have little to no spectral contributions from nonaxial terms.

1.3. Temporal Characteristics of Observed Magnetic Fields

Dynamo-generated fields are expected to display a myriad of temporal behavior reflecting the temporal nature of the fluid motions generating the fields. Variations in Earth’s observed field are discussed in Section 5.1. There is little information about the temporal behavior of other planetary magnetic fields due to a lack of magnetic data resolution, both temporally and spatially. No undisputed magnetic secular variation has been observed for Jupiter, Saturn, or Mercury, the three planets that have been visited by multiple magnetic missions with a sufficient time interval to carry out a secular variation study. However, we do expect these fields to exhibit secular variation based on our observations of Earth’s magnetic field as well as the solar magnetic field, also generated by a dynamo.

2. THE DYNAMO MECHANISM

2.1. What is a Dynamo?

A dynamo is the process by which mechanical energy is converted to electromagnetic energy through induction. In
planets, the mechanical energy is due to fluid motions and the resulting electromagnetic energy produces the observed planetary magnetic fields. The main equation governing dynamo action is the magnetic induction equation:

$$\frac{\partial \vec{B}}{\partial t} = \nabla \times (\vec{v} \times \vec{B}) + \eta \nabla^2 \vec{B}$$

which can be derived from Maxwell’s equations and Ohm’s law in the magnetohydrodynamic limit.

In this equation, the time variation of magnetic field \( \vec{B} \) is the result of: (1) the interaction of velocity fields \( \vec{v} \) and magnetic fields (represented in the first term on the right-hand side of the equation which we will call the “induction” term) and (2) the diffusion of the field through Ohmic dissipation (represented by the second term on the right-hand side of the equation). The magnetic diffusivity \( \eta = (\sigma \mu)^{-1} \) is inversely proportional to the electrical conductivity \( \sigma \) and the magnetic permeability \( \mu \).

### 2.2. Necessary Conditions for a Dynamo

In order for a planet to have a dynamo-generated magnetic field, it must contain an electrically conducting fluid region undergoing motions to generate induction. These conditions are easily discernible by examining the magnetic induction equation, i.e. nonzero \( \sigma \) and \( \nabla \times \vec{v} \) are required for a nonzero induction term.

However, there are further necessary conditions for the vigor and morphology of the motions that result from the fact that the magnetic field must not decay away due to Ohmic dissipation. A common measure of the required vigor of motions comes from the critical magnetic Reynolds number condition. In order for dynamo action to be sustainable, the induction term must be larger than the diffusion term in the magnetic induction equation. By using some characteristic velocity \( V \) and length scale \( L \) to represent the magnitude of terms on the right-hand side of the magnetic induction equation, the ratio of the induction to diffusion terms is given by the magnetic Reynolds number, \( \text{Re}_M \):

$$\text{Re}_M = \frac{|\nabla \times (\vec{v} \times \vec{B})|}{|\eta \nabla^2 \vec{B}|} \approx \frac{VL}{\eta}$$

This number must be larger than some critical value \( \text{Re}_c \) in order for dynamo action to occur. Lower bounds can be placed on this critical value using analytic techniques (see Jones (2008) for some common ones). Depending on the choice of characteristic length and velocity scales, the bounds are typically around \( \pi \) to \( \pi^2 \); however, these bounds do not take into account the required complexity of the fluid motions (see below). Investigations of \( \text{Re}_M \) in numerical simulations of dynamos give values around 20–50.

It is believed that active planetary dynamos have \( \text{Re}_M \) much greater than the critical value. For example, using the secular variation of the field as a characteristic velocity and the core radius as a characteristic length scale results in \( \text{Re}_M = O(10^3) \) for the Earth.

In addition to the vigor of convection, there are also necessary conditions on the morphology of the velocity field. Antidynamo theorems demonstrate that the flow in a spherical geometry must have a radial component. This rules out some standard fluid motions from being dynamo-capable. For example, in spherical coordinates \((r, \theta, \phi)\), differential rotation of the form \( \vec{v}(r, \theta, \phi) = v(s) \phi \) (where \( s = r \sin \theta \) is the cylindrical radial coordinate and \( \phi \) is the longitudinal direction) cannot produce a dynamo alone. Similarly, solid body rotation due to the rotation of the planet alone cannot generate a dynamo, although it has an important influence on the flow morphology. Analytic expressions for flows capable of generating a dynamo have been found (see Jones (2008) for a review); however the minimum sufficient conditions for a dynamo are not currently known.

### 2.3. Dynamo Generation Regions in Planets

The basic necessity for a planetary dynamo is a fluid electrically conducting region in the planet. In terrestrial planets, this region is the liquid layer of the iron-rich core and hence the conductivity is metallic. Although the giant planets likely also contain deep rocky layers with iron-rich cores, these are not the source of the observed magnetic fields for these planets. This is because other materials with good conductivity are undergoing motions in a much larger fraction of these planetary interiors closer to their surfaces.

Jupiter and Saturn, composed predominantly of hydrogen, possess extreme temperatures and pressures in their interiors. Hydrogen under these extreme conditions can metallize and hence produce a good electrical conductor. Even at pressures somewhat lower than the transition to a metallic state, hydrogen can be an effective semiconductor, and if velocities are large enough in this region, the conductivity may be sufficient to generate a dynamo. The approximate radii at which these transitions occur are given in Table 6.2 along with other properties of the dynamo source regions in planets.

In Uranus and Neptune, the hydrogen-rich layer does not extend to high enough pressures to metallize. Instead, the large water-rich portion of these planets reaches pressures and temperatures allowing for the dissociation of molecules and hence a significant ionic conductivity. Although ionic conductivities are not as large as metallic conductivities, the length scales and likely velocities in these regions still result in highly supercritical magnetic Reynolds numbers for these bodies.
3. THE STANDARD PLANETARY DYNAMO

In this section, we discuss the “canonical” planetary dynamo. Perhaps not coincidentally, this canonical dynamo is considered a good representation of Earth’s dynamo. As spacecraft missions have provided details of other planetary magnetic fields, it has become obvious that this standard picture is not applicable to all planets and that the differences between planetary magnetic fields must be explained by considering each planet’s dynamo region properties more carefully. In Section 3.4, we consider some of the details beyond this standard picture.

In the canonical planetary dynamo, fluid motions are generated in an electrically conducting spherical shell surrounding a solid inner core. The physical properties of the dynamo region (e.g. thermal and electrical conductivities and viscosity) are assumed to be constants.

3.1. Driving Forces

The fluid motions required for dynamo action must have a power source. The most commonly accepted source is gravitational potential energy release due to cooling of the planet. Planetary formation results in significant amounts of heat trapped in planetary interiors. The planets then slowly cool over time. In the simplest picture, this results in an unstable thermal stratification in the planetary dynamo region where hotter (and hence less dense) material lies below colder (and hence more dense) material. Above a critical temperature difference across the dynamo region, convection will occur in which the hotter (i.e. more buoyant) material is transported outward.

In addition to heat of formation, other sources of buoyancy in the dynamo region may include:

1. Radiogenic heat sources: If significant concentrations of radiogenic elements are in the dynamo region, then heat from radioactive decay can contribute to the thermal energy in the core and drive convection.

2. Compositional convection: Planetary cooling can result in the generation of compositional variations that can also lead to buoyancy differences and hence convection. For example, as the Earth’s core cools, the solid inner core freezes out, releasing a light element-rich fluid at the base of the liquid core. Since this fluid contains less iron than the surrounding fluid, it is less dense and hence buoyant. In Mercury and Ganymede, details of the iron–sulfur system suggest that an iron-rich solid (called “iron snow”) may condense out at midlayers or at the top of the cores resulting in more dense material at outer radii and hence a negative buoyancy. This can also drive convection.

The rate at which a planet can cool is ultimately determined by how much heat can be removed from the outer layers. In the terrestrial planets, the metallic cores are surrounded by rocky mantles with very different material properties than the cores. Heat transfer in the mantle layer is at a very different pace than that in the core and the amount of heat that can be removed from the core is ultimately determined by how much heat can be transferred through the mantle. Details of the mantle structure are therefore important in determining the cooling properties of the cores of terrestrial planets.

Convection is believed to be the most likely source of motions generating dynamos in planetary cores for two reasons. First, the abundance of heat in most planetary dynamo regions makes this power source capable of generating motions for long times. Second, the form of convection in rapidly rotating fluids results in flow morphologies that are very conducive to dynamo action. These morphologies will be explored in Section 3.2.

3.2. Fluid Motions in Dynamo Regions

Dynamo generation regions are spherical shells undergoing rapid rotation. This results in specific flow morphologies:

1. Convective flows: Due to the Taylor–Proudman theorem, flows in rapidly rotating, low-viscosity fluid show much smaller variation in the axial direction compared to the cylindrical radial and azimuthal directions. Since convection aims to move buoyant parcels outward, the combination of outward motions and rapid rotation results in columnar motions where fluid in entire vertical spans move outward. Rapid rotation is therefore very good at organizing convective fluid motions on a large scale (see Figure 6.4).

2. Meridional flows: These flows are approximately parallel to the rotation axis inside the convection columns caused, for example, by geometric effects due to the boundary curvature.

3. Zonal flows: Flows in the azimuthal (i.e. longitudinal or zonal) direction are typically generated by (a) thermal winds due to latitudinal variations in buoyancy, (b) magnetic winds due to latitudinal variations in magnetic fields, or (c) Reynolds stresses due to correlations in small-scale velocity fields. These flows result in differential rotation in the dynamo generation region.

3.3. Generation Mechanisms

Fluid motions must possess a certain amount of complexity in order to generate a dynamo. As discussed in Section 2.2, the flows must be vigorous enough, but must also meet
morphology requirements. There are two categories of motions which are likely to occur in planetary cores that work very well in combination to generate a dynamo: (1) helical flows and (2) differential rotation.

To discuss magnetic field generation mechanisms, it is common to use a poloidal/toroidal decomposition of the field. This is possible because magnetic fields obey Gauss’ law for magnetism and so they can be fully represented with two scalar functions:

$$B_T = \nabla \times (T \vec{r}) + \nabla \times (P \vec{r})$$

where $B_T$ and $B_P$ are orthogonal “toroidal field” and “poloidal field” components and $T$ and $P$ are the toroidal and poloidal scalar functions respectively. Note from the equation above that toroidal field has no radial component, whereas poloidal field generally has components in all three directions. In simplest terms, magnetic field generation can be envisioned as the result of helical flows and differential rotation acting upon some initial toroidal and poloidal fields to generate new fields. If this can be done in such a way as to reproduce the initial fields before their Ohmic decay, then the dynamo is self-sustaining. Heuristic depictions of these generation mechanisms are shown in Figure 6.5.

Helical flows are a natural result of convection in planetary dynamo regions due to their spherical geometry and rapid rotation. Figure 6.5(a) depicts helical flow stretching and twisting a magnetic field line to generate field with an orthogonal component to the original field. The amount of helicity is measured through the helicity parameter $h = \vec{v} \cdot \vec{w}$, where $\vec{w} = \nabla \times \vec{v}$ is the vorticity in the fluid. If this flow acts upon a toroidal magnetic field, then it can generate poloidal field from it. Similarly, if this flow acts upon a poloidal magnetic field, then it can generate toroidal field from it.

The process of generating magnetic fields through helical motions is sometimes called the “\( \alpha \)-effect”. This name is borrowed from mean field dynamo theory, popular in studies of astrophysical dynamos, but has a slightly different connotation in planetary dynamo theory. Specifically, the \( \alpha \)-effect in mean field theory represents the combined effect of turbulent microscopic motions, whereas in planetary dynamo theory, the helical motions described above can be macroscopic in scale and laminar (as opposed to turbulent). In general terms, one can consider the \( \alpha \)-effect as being the result of motions with helicity.

Differential rotation refers to the shearing motion of zonal flows. As discussed in Section 2.2, these motions alone cannot generate a dynamo; however, they can be used in combination with other motions to very effectively generate magnetic field. Figure 6.5(b) depicts cylindrical differential rotation acting on a poloidal magnetic field. The differential rotation shears the magnetic field line generating a magnetic field component orthogonal to the original magnetic field. The strength of the generated toroidal magnetic field will depend on the shear and the amplitude of the poloidal field where the shear is strong.
The process of magnetic field generation through differential rotation is sometimes called the “ω-effect”. Here the connotation is similar to that in mean field theory as it describes the effect of large-scale zonal flows on stretching magnetic fields.

The dynamo generation cycle can then be envisioned in the following way: If we initially have a poloidal magnetic field, then either differential rotation or helical flows can act to generate toroidal magnetic field from it. Once this toroidal field is created, then helical flows can act to generate poloidal magnetic field from it. In this way, we are able to regenerate the field we started with and the cycle of field generation can continue in a self-sustained manner.

There is debate as to whether the ω-effect is an important contributor to magnetic field generation in planets. It is possible to generate a dynamo solely through the ω-effect and hence the ω-effect is not needed in principle. However, there are mechanisms to generate strong differential rotation in planets, for example, through thermal winds or Reynolds stresses. Since we cannot observe fluid motions deep in planets directly, investigations of the importance of these processes are mainly carried out through computational and laboratory experiments (see Section 4).

3.4. Beyond the Standard Dynamo

Additions to the standard dynamo appear to be necessary to explain unique features of specific planetary dynamos. These include:

1. *Alternative driving mechanisms:* In addition to convection, other driving mechanisms for fluid motions have been proposed and may be relevant for specific planets. Motions due to precession, tides, and boundary driving have all been invoked as possibilities. Although the basic flows due to these forcings are purely toroidal and laminar, and hence, not good at generating dynamos, instabilities of the motions due to these forcings have been shown to be capable of dynamo action. These mechanisms may be important at certain times in planetary evolution when convection is not capable of driving a dynamo. For example, the relatively late lunar dynamo may have been the result of precession or boundary forcings.

2. *Stably stratified layers:* The entire electrically conducting fluid regions may not be convecting either because of compositional or thermal stratification. For example, data and models suggest that an outer thin layer of Earth’s core may be stably stratified. Thermal evolution models for Uranus and Neptune also suggest that the deepest water-rich regions of these planets are stably stratified. The presence of helium rain in Jupiter and Saturn may also result in stably stratified layers.

3. *Dynamo region geometry:* The dynamo source region is a spherical shell as opposed to a full sphere in most planets. The thickness of the spherical shell has important implications for the convective motions. For example, in Earth, the solid inner core is relatively small and may not affect the location of convection columns, but in a planetary body with a larger inner core (and hence a thinner convecting shell), the fluid motions must allow for the inner core boundary.

4. *Radially-varying physical properties:* The material properties of the fluid, such as the electrical and thermal conductivities, kinematic viscosity, and density are pressure and temperature dependent. The variation of these properties as a function of depth in the dynamo source region may have important implications for dynamo generation. This is likely to be more important for the giant planets than the terrestrial planets since the more massive giant planets experience larger ranges of pressure and temperature in their dynamo source regions.

5. *Laterally-varying boundary conditions:* Conditions at dynamo region boundaries may not be homogeneous. For example, in the Earth, laterally varying heat flux at the core-mantle boundary may have significant influence on the dynamo.

6. *Influence of external fields:* If a source external to the planetary body generates significant magnetic fields in the dynamo source region, these can affect the dynamo. For example, Jupiter’s magnetic field is relatively strong at Ganymede and hence may influence Ganymede’s dynamo generation. Similarly, magnetospheric currents at Mercury may generate magnetic fields that can be appreciable in Mercury’s core.

4. SIMULATIONS AND EXPERIMENTS

In combination with planetary magnetic field observations, properties of planetary dynamos are investigated through numerical simulations and laboratory experiments. It is currently not feasible to accurately represent the parameter regime of planetary interiors with simulations and experiments; however, they are useful tools in investigating mechanisms and force balances in planetary dynamos.

4.1. Numerical Dynamo Simulations

Dynamo simulations computationally solve the governing equations for planetary dynamos. Current models include
some of the additions discussed in Section 3.4, but here we will describe the equations solved for the simplest standard planetary dynamo model: that of a Boussinesq, electrically conducting fluid driven by thermal convection in a spherical rotating shell with a linear gravity profile. Equations solved numerically are then:

1. Momentum equation:

\[
\frac{\partial \vec{v}}{\partial t} + \vec{v} \cdot \nabla \vec{v} + 2 \vec{\Omega} \times \vec{v} = -\nabla \tilde{P} - \alpha T \frac{g_0}{r_0} \hat{r} + \frac{1}{\rho_o} \hat{r} \times \vec{B} + \nu \nabla^2 \vec{v}
\]

2. Magnetic induction equation:

\[
\frac{\partial \vec{B}}{\partial t} = \nabla \times (\vec{v} \times \vec{B}) + \eta \nabla^2 \vec{B}
\]

3. Energy equation:

\[
\frac{\partial T}{\partial t} + \vec{v} \cdot \nabla T = \kappa \nabla^2 T + Q
\]

where \( \vec{\Omega} \) is the angular velocity of the body, \( \tilde{P} \) is the modified pressure, \( \rho_o \) is the background constant density, \( \alpha \) is the thermal expansion coefficient, \( T \) is the temperature, \( g_0 \) is the gravitational acceleration at the top of the dynamo source region (i.e. at \( r_0 \)), \( \nu \) is the kinematic viscosity, \( \kappa \) is the thermal diffusivity, and \( Q \) is the volumetric heat source.

Typically, these equations are nondimensionalized using characteristic scales for the dimensional variables. There are a variety of ways to carry out the nondimensionalization and we offer one example here. Choosing the dynamo region shell thickness \( D \) as a length scale, the magnetic diffusion time \( \tau = D^2/\eta \) as the timescale, a magnetostrophic balance estimate \( B = (2 \Omega \mu_0 \rho \eta)^{1/2} \) as the magnetic field scale, the superadiabatic temperature difference \( \Delta T \) across \( D \) as the temperature scale, and using Ampère’s law to represent the current density in terms of the magnetic field, the nondimensional equations can be written in the form:

\[
\frac{E}{P_m} \left( \frac{\partial \vec{v}}{\partial t} + \vec{v} \cdot \nabla \vec{v} \right) + 2 \vec{\Omega} \times \vec{v} = -\nabla \tilde{P} - Ra_{th} T \hat{r} + \hat{r} \times \vec{B} + \nu \nabla^2 \vec{v}
\]

\[
\frac{\partial \vec{B}}{\partial t} = \nabla \times (\vec{v} \times \vec{B}) + \nabla^2 \vec{B}
\]

Due to numerical constraints, simulations are not able to work in the appropriate parameter regime for planetary dynamo regions. Specifically, the Ekman and magnetic Prandtl numbers are much larger in simulations than in planets, and the Rayleigh number is probably much smaller in simulations than in planets (although it is not well constrained). However, insights into mechanisms and force balances are used to extrapolate results from simulations to planetary conditions using scaling laws, but it is unclear whether these derived scaling laws hold over the many orders of magnitude of extrapolation needed to bridge the gap between simulations and planets.

Although dynamo simulations work in parameter regimes far from that of planets, they are capable of reproducing many salient features of observed planetary magnetic fields. For example, models for Earth’s dynamo can produce axially dipolar-dominated fields, smaller scale spectral features, reversals, and other secular variation features that have been observed. This may be coincidental but may also be the result of the fact that dynamo models
are producing accurate force balances even if the quantitative values for the parameters are not correct. For example, although the Ekman number is $\sim 10$ orders of magnitude too large in simulations, the value is still very small ($\sim O(10^{-7})$) indicating that viscous forces are much weaker than Coriolis forces in the models, as we would expect them to be in planetary dynamo regions.

Numerical simulations for other planets are generated by including some of the additions discussed in Section 3.4, where applicable. These are discussed in detail in Section 5 for the respective planets.

The ultimate goal of numerical simulations is to use them to understand the processes occurring in planetary dynamos. Numerical modelers are using the most advanced computational resources available to push parameters as close as possible to planetary values, and the models will only improve in the future. Other efforts include covering wide ranges of parameter space to develop scaling laws for various observables (see Section 5). These types of models provide valuable insights into fluid motions at more extreme parameter values than possible in fully three-dimensional self-consistent models.

Other numerical methods aim to simplify the models to work in more challenging parameter regimes. For example, “quasi-geostrophic” models solve the equations governing fluid motions solely in the equatorial plane and then use constraints from rapid rotation to infer the motions outside of this plane. These types of models provide valuable insight into fluid motions at more extreme parameter values than possible in fully three-dimensional self-consistent models.

### 4.2. Dynamo Experiments

Building a laboratory experiment that can generate a self-sustaining dynamo is incredibly challenging. The main reason being that the small length scales of experiments result in small magnetic Reynolds numbers (i.e. below critical values) unless velocities are made extremely large. Typical experiments use liquid metals such as sodium or gallium. Early experiments, such as the Karlsruhe and Riga dynamos, used a series of pipes to create flow morphologies that are known analytically to be conducive to dynamo action. They demonstrated the growth of magnetic field intensity; and hence dynamo action; however, they are somewhat nonplanetary-like in geometry.

Present day experiments aim to generate dynamos in more homogenous geometries (e.g. in cylindrical or spherical tanks). In these experiments, the flows are not as constrained as the pipe flows and instead, fluid motions are generated through propellers or boundary differential rotation (see Figure 6.6 for some examples). Generating a laboratory dynamo using convective motions (i.e. through buoyancy) is currently not feasible. In addition, experiments face difficulties in mimicking the radial form of the gravity force in planets.

These experiments have provided important insights into the role of turbulence in helping and hindering dynamos and also provide a means to investigate different regions of parameter space from numerical simulations. For a nice review of dynamo experiments, see Lathrop and Forest (2011).

## 5. PLANETARY DYNAMOS

Here we outline the major features of planetary dynamos in a comparative fashion. Information comes from magnetic field and other spacecraft observations, as well as theoretical and experimental studies of planetary interior properties. We briefly discuss results from numerical dynamo simulations, but for a deeper review of planetary dynamo simulations, see Stanley and Glatzmaier (2010).

### 5.1. Earth

The Earth’s magnetic field (also known as the geomagnetic field) is dominated by its axial dipole component. The nondipolar component of the field includes two strong normal polarity flux spot pairs, particularly evident in the northern hemisphere over Canada and Russia. Based on paleomagnetic records, these spots appear to be long-lived and it has been suggested that they are features associated with convection columns in the core. These convection columns might be relatively stationary due to thermal influences from the mantle convection morphology at the core-mantle boundary. In addition to these flux spots, the Earth also has intermittent normal and reverse flux patches in equatorial regions, some of which drift westward in time. Figure 6.7 shows the radial component of the Earth’s magnetic field at the core-mantle boundary.

Records from satellites such as Magsat and Oersted, ground observatories, and ship logs have provided fairly detailed records of the geomagnetic field over the past 400 years. In addition, paleomagnetic records from crustal rocks on the sea floor provide data on magnetic field reversals over the past $\sim 180$ Myrs (i.e. up to the age of the oldest seafloor). Prior to this, we rely on paleomagnetic fields in continental rocks.

These data demonstrate the time variability of the geomagnetic field. Reversals occur sporadically, on average, every half a million years. In recent times, equatorial flux spots drift with a speed of approximately 0.2°/year and the north geomagnetic pole meanders about the geographic pole at about 10 km/year.

The geomagnetic field is generated in the Earth’s iron-rich core. The fluid outer core surrounds a solid inner
core that is enriched in iron compared to the outer core. The inner core grows in time as the Earth cools and thermal evolution models suggest that it is ~1 billion years old. As it grows, it expels light elements resulting in a source of compositional buoyancy at the base of the outer core, in addition to the thermal buoyancy available from core cooling, latent heat, and possibly, radiogenic elements. Paleomagnetic records suggest that the Earth’s field is at least 3.4 billion years old, implying that the geodynamo was active before the solid inner core began to grow.

Earth is the only planet with evidence of a solid inner core. Although the Earth’s inner core is relatively small

**FIGURE 6.7** Radial component of the Earth’s magnetic field at the core-mantle boundary. *Figure from Jones (2011).*
likely that they will appear in the near future. Models with the combination of a large dipole offset and fields. There has not been much investigation of producing terrestrial fields and the core field produces weaker observed models that appeal to feedback between external magnetic elements such as sulfur or silicon in addition to iron in order to depress the melting temperature and hence reduce the speed of solid inner core growth. Radar measurements from Earth, as well as MESSENGER data, have independently confirmed that the Mercury core contains a liquid layer. The size of the inner core is unknown and depends strongly on the fraction of light elements in the core. Prior to the Mariner 10 mission, it was considered unlikely that Mercury would have a dynamo based on thermal evolution models for such a small body. Essentially, since smaller planets cool faster, it was suggested that Mercury’s iron core should have fully solidified by the present day making dynamo action impossible. To explain the observed global magnetic field, most likely the result of a dynamo, researchers suggested that the core must also contain light elements such as sulfur or silicon in addition to iron in order to depress the melting temperature and hence reduce the speed of solid inner core growth. Radar measurements from Earth, as well as MESSENGER data, have independently confirmed that the Mercury core contains a liquid layer. The size of the inner core is unknown and depends strongly on the fraction of light elements in the core. Specific features of Mercury’s magnetic field are difficult to explain with a “standard” dynamo model. First, the dipole field is about two to three orders of magnitude weaker than that expected by scaling laws that estimate a dynamo-generated field’s strength. Second, Mercury’s axial dipole offset (a measure of its axial quadrupole component) is fairly large, whereas its dipole tilt (a measure of its nonaxial dipole component) is fairly small. This combination is difficult to produce with a standard dynamo. Several numerical dynamo models have attempted to explain the weak dipole intensity by appealing to additions to the standard dynamo model (like those in Section 3.4). For example, numerical models with stably stratified layers either in the outer region of the core, or at mid-depth, can produce weak dipole fields, as can models with a relatively large or relatively small solid inner core. In addition, models that appeal to feedback between external magnetospheric fields and the core field produce weaker observed fields. There has not been much investigation of producing models with the combination of a large dipole offset and small dipole tilt due to the freshness of this data, but it is likely that they will appear in the near future.

5.2. Mercury

Mercury’s magnetic field was first observed by the Mariner 10 mission during flybys of the planet in the mid-1970s. Data from two of the flybys found a weak dipole moment. More recent data from the MESSENGER mission has confirmed the Mariner 10 results and has also provided more detailed constraints on the low-degree spectral components of the field. Prior to the Mariner 10 mission, it was considered unlikely that Mercury would have a dynamo based on thermal evolution models for such a small body. Essentially, since smaller planets cool faster, it was suggested that Mercury’s iron core should have fully solidified by the present day making dynamo action impossible. To explain the observed global magnetic field, most likely the result of a dynamo, researchers suggested that the core must also contain light elements such as sulfur or silicon in addition to iron in order to depress the melting temperature and hence reduce the speed of solid inner core growth. Radar measurements from Earth, as well as MESSENGER data, have independently confirmed that the Mercury core contains a liquid layer. The size of the inner core is unknown and depends strongly on the fraction of light elements in the core. Specific features of Mercury’s magnetic field are difficult to explain with a “standard” dynamo model. First, the dipole field is about two to three orders of magnitude weaker than that expected by scaling laws that estimate a dynamo-generated field’s strength. Second, Mercury’s axial dipole offset (a measure of its axial quadrupole component) is fairly large, whereas its dipole tilt (a measure of its nonaxial dipole component) is fairly small. This combination is difficult to produce with a standard dynamo. Several numerical dynamo models have attempted to explain the weak dipole intensity by appealing to additions to the standard dynamo model (like those in Section 3.4). For example, numerical models with stably stratified layers either in the outer region of the core, or at mid-depth, can produce weak dipole fields, as can models with a relatively large or relatively small solid inner core. In addition, models that appeal to feedback between external magnetospheric fields and the core field produce weaker observed fields. There has not been much investigation of producing models with the combination of a large dipole offset and small dipole tilt due to the freshness of this data, but it is likely that they will appear in the near future.

5.3. Jupiter

Magnetic field data for Jupiter has come primarily from the Voyager I and II and Galileo missions. Jupiter’s field is similar to Earth’s field in morphology, being axially dipolar dominated with a dipole tilt of approximately 10° and a large-scale spectral structure similar to Earth’s field (see Figure 6.3). The field is generated in the electrically conducting hydrogen region of the planet which extends out to about 0.8–0.9 Jupiter radii. Dynamo simulations for Jupiter generally include radially varying physical properties such as density and electrical conductivity since the dynamo generation region extends through many pressure-scale heights. Although the electrical conductivity increases with depth, the deepest layers are also the most dense and hence, experience the slowest fluid velocities. It is therefore possible that dynamo generation is limited to the outermost layers of the dynamo region where the combination of fluid velocities and electrical conductivity produce the most appreciable magnetic Reynolds numbers. Data from the upcoming Juno mission are expected to provide a significant improvement to the resolution of the field, which should allow testing of dynamo region geometry.

5.4. Saturn

Saturn’s magnetic field data has come from Voyager I and II in the 1970s and the 1980s and more recently from Cassini since 2004. Like Jupiter, the field is dominated by its axial dipole, but Saturn is unique in the lack of any observed nonaxisymmetric field components. In addition, the octupole component is larger than the quadrupole component suggesting a preference for odd harmonics in the field. Because the axial octupole Gauss coefficient (g3) has the same sign as the axial dipole Gauss coefficient (g1), the field is concentrated in the polar regions compared to equatorial regions. This is the opposite of what is observed in Earth’s or Jupiter’s field today, although standard numerical dynamo simulations suggest that nondipolar components can vary significantly in time. Hence, this may just be the result of capturing the magnetic field in an untypical configuration or point toward a different dynamo mechanism.

The observation of a purely axisymmetric field is problematic for a standard dynamo explanation. First, Cowling’s theorem demonstrates that a perfectly axisymmetric magnetic field cannot be generated by a dynamo. Second, no other planetary magnetic field demonstrates the same amount of axisymmetry as Saturn (although future data from Mercury and Ganymede may alter this statement). Third, standard dynamo models cannot reproduce this level of axisymmetry in the observed field.

Like Jupiter, Saturn’s dynamo region is its hydrogen-rich metallic layer. Saturn may differ from Jupiter by the
presence of a helium rain layer at pressures and temperatures where hydrogen becomes metallic. Quantum mechanical simulations have demonstrated that at these characteristic pressures and temperature in Saturn, helium may become immiscible in hydrogen and as it separates from the mixture it will be negatively buoyant compared to the hydrogen and hence “rain out”. At deeper pressure, helium may then become miscible again (for a review, see McMahon et al., 2012). It is possible that such a layer also exists in Jupiter, but based on thermal evolution calculations, this layer should be much thicker in Saturn than Jupiter and hence, this may be a decent mechanism to explain the difference between these planetary magnetic fields.

The leading theory explaining Saturn’s axisymmetric field involves this layer: If Saturn’s dynamo is surrounded by this stably stratified electrically conducting layer, and if differential rotation exists in this layer, then the non-axisymmetric field may be preferentially attenuated through the electromagnetic skin effect resulting in a surface field with much more axisymmetry than would occur without such a layer. Dynamo simulations for Saturn explore the effects of stably stratified layers on axisymmetrizing the field.

5.5. Uranus and Neptune

The magnetic field data of Uranus and Neptune come from single flybys of the planets by the Voyager II mission in the 1980s. The data revealed that, unlike the other planets, these ice giants’ magnetic fields were not dominated by their axial dipoles and instead, contained roughly equal contributions from different field harmonics.

Standard numerical dynamo models typically do not produce nonaxially dipolar-dominated fields, except in isolated regions of parameter space or for very large buoyancy forcing relative to Coriolis forcing. Although these may be viable explanations for these planets’ field morphologies, these models do not explain another observation of the ice giants: in addition to the anomalous magnetic fields, data also demonstrate that these planets have low intrinsic heat flows.

The dynamo source regions in the ice giants are the water-rich layers at depths such that a significant ionic conductivity results (~0.7–0.8 planetary radii). Thermal evolution simulations suggest that the low heat flows are explained if the ice giant interiors are not fully convective, and hence that the inner regions of these ice layers are stably stratified to convection. Numerical dynamo simulations that incorporate a stably stratified interior fluid region below a relatively thin convective shell (where the dynamo generation occurs) can reproduce the magnetic field observations for the ice giants.

5.6. Ganymede

Ganymede’s magnetic field was discovered by the Galileo mission in the mid-1990s. Although magnetic induction signatures were found for Europa, Ganymede, and Callisto, resulting from currents generated in salt-water oceans in these bodies, Ganymede was the sole Galilean satellite to demonstrate a self-sustained dynamo-generated field. There is little data available for the spectrum of the field aside from a dipole moment.

Like Mercury, Ganymede’s small size suggests that there must be a significant fraction of light elements such as sulfur in its core to keep it liquid at present day. It is unknown whether Ganymede has a solid inner core, but the dynamo may have a compositional driving source if the liquid core is in a regime where it freezes at the outer boundary, releasing negatively buoyant iron-rich fluid, rather than freezing at the inner boundary like in the Earth’s core. Dynamo studies including different buoyancy source distributions intended to mimic these solidification processes have been carried out.

5.7. Ancient Moon

The lunar magnetic field was mapped by the Lunar Prospector mission, but this field is due to remanent magnetization in the lunar crust rather than an active dynamo. Paleomagnetic data from lunar samples indicate that the field was most likely due to a dynamo that was active from at least 4.2 to 3.5 billion years ago.

The driving source for the lunar dynamo is unclear since the small size of the core suggests that thermal convection would not provide enough energy to drive a dynamo for such a long time after formation. Therefore, alternative mechanisms such as precession or boundary forcing due to oblique impacts have been suggested.

5.8. Ancient Mars

The Martian magnetic field was studied by the Mars Global Surveyor mission in the 1990s. Similar to the Moon, the Martian magnetic field is due to remanent magnetization in the crustal rocks. The magnetizing field was most likely due to a dynamo active in early Martian history, before ~3.9 billion years. The crustal field displays a correlation with the hemispheric crustal dichotomy (see Chapter on Mars: Surface and Interior) with the southern hemisphere containing more intense fields than the northern hemisphere.

The magnetic dichotomy may be due to postdynamo crustal reworking that preferentially removed magnetism from crust in the northern hemisphere, or it may be due to the morphology of the magnetizing field while the rocks were forming. For example, mechanisms suggested to explain the hemispheric crustal dichotomy include degree-
one mantle circulation or a large glancing impact in the northern hemisphere. Both these mechanisms could result in hemispheric thermal variations at Mars’ core-mantle boundary. Numerical dynamo simulations have demonstrated that these thermal variations can result in hemispheric dynamos, where the field is much stronger in the southern hemisphere than in the northern hemisphere. It is therefore possible that the difference in crustal magnetization between the hemispheres is the result of a difference in the intensity of the magnetizing field between the hemispheres.

5.9. Small Bodies

Paleomagnetic studies of classes of meteorites such as the Angrites, some carbonaceous chondrites and the HED (Howardite-Eucrite-Diogenite) meteorites demonstrate that small bodies such as planetesimals and asteroids may have possessed dynamos in the early solar system. Although it is difficult to sustain a convective driving force for a long time in such small bodies, the presence of appreciable heat sources in the early solar system, such as radiogenic elements $^{126}\text{Al}$ and $^{60}\text{Fe}$, provide enough driving power for a short time (around 10 million years) to allow iron core formation and then to generate a dynamo. Scaling studies demonstrate that the surface fields are strong enough to magnetize the crusts and explain the observed meteorite magnetism (Weiss et al., 2010).

5.10. Extrasolar Planets

Although one would expect some extrasolar planets to have active dynamo-generated magnetic fields, no extrasolar planetary magnetic field has been unambiguously discovered. However there are potential detection methods that could be feasible in the near future. One possibility of discovery is through radio emissions from the interaction of stellar winds with planetary magnetic fields (similar to the radio emissions observable from Jupiter in our own solar system). New advanced radio antennae such as the Low-Frequency Array might be capable of such detection.

5.11. Planetary Bodies Lacking Dynamos

The only planet in our solar system with no evidence of past or present dynamo action is Venus. This is somewhat unexpected considering the similarity between Venus’ and Earth’s interior structures. The difference cannot be explained by Venus’ slower rotation. The most likely explanation for a lack of dynamo action today in Venus is that the core is not cooling vigorously enough to generate strong convection. This may be due to the fact that Venus experiences a different mode of mantle convection than the Earth, possibly due to a lack of water in Venus’ mantle. Venus appears to experience sluggish, rigid lid convection with episodic large-scale overturn. During the sluggish stage, the mantle may not remove enough heat from the core to generate core convection and hence no dynamo generation occurs.

An interesting question remains as to whether a dynamo action today in Venus onsets during episodic overturning events? Unfortunately, the surface temperatures are above the Curie temperature for most crustal rocks and therefore, it is unlikely that the surface rocks contain appreciable remanent magnetization from the last overturning episode, which occurred approximately 700 million years ago.

Aside from the Moon and Ganymede, no other satellites or small planetary bodies have observed magnetic fields (expect for those inferred from meteorites discussed in the previous section).

6. CONCLUSIONS AND FUTURE PROSPECTS

The past four decades have provided a wealth of new data from satellite missions on planetary magnetic fields. In combination with numerical simulations and laboratory experiments over the past two decades, new insights on planetary magnetic field generation have resulted. The near future promises some exciting advances. First, new data from planned and active magnetic missions such as Cassini (at Saturn until 2017), JUNO (to Jupiter), Juice (the Jupiter system), BepiColombo (to Mercury), and Swarm (to Earth) will provide great improvements in data resolution and possibly allow study of secular variation of planetary magnetic fields other than Earth. Second, ongoing improvements to computational models and hardware, in addition to laboratory dynamo experiments, will improve our knowledge of the fluid dynamics and magnetohydrodynamics of planetary dynamo regions. Third, new paleomagnetic techniques and data sets will add to our knowledge of past planetary magnetic fields. Finally, planetary magnetic field generation is not an isolated process and so improvements in our knowledge of planetary interior composition, structure, and dynamics will also be used to provide greater constraints on planetary magnetic fields.

BIBLIOGRAPHY


1. WHAT IS A MAGNETOSPHERE?

The term magnetosphere was coined by T. Gold in 1959 to describe the region above the ionosphere in which the magnetic field of the Earth controls the motions of charged particles. The magnetic field traps low-energy charged particles and forms the Van Allen belts, torus-shaped regions in which high-energy ions and electrons (tens of keV and higher) drift around the Earth. The control of charged particles by the planetary magnetic field extends many Earth radii into space but finally terminates near 10 Earth radii in the direction toward the Sun. At this distance, the magnetosphere is confined by a low-density magnetized plasma called the solar wind that flows radially outward from the Sun at supersonic speeds. (Plasmas are highly ionized gases composed of electrically charged particles in equal proportions of positive charge on ions and negative charge on electrons whose properties are dominated by their electromagnetic interactions.) Qualitatively, a planetary magnetosphere is the volume of space from which the solar wind is excluded by a planet’s magnetic field. (A schematic illustration of the terrestrial magnetosphere is given in Figure 7.1, which shows how the solar wind is diverted around the magnetopause, a surface that surrounds the volume containing the Earth, its distorted magnetic field, and the plasma trapped within that field.) This qualitative definition is far from precise. Most of the time, solar wind plasma is not totally excluded from the region that we call the magnetosphere. Some solar wind plasma finds its way in and indeed many important dynamical phenomena give clear evidence of intermittent direct links between the solar wind and the plasmas governed by a planet’s magnetic field. Moreover, unmagnetized planets in the flowing solar wind carve out cavities whose properties are sufficiently similar to those of true magnetospheres to allow us to include them in this discussion. Moons embedded in the flowing plasma of a planetary magnetosphere create interaction regions resembling those that surround unmagnetized planets. If a moon is sufficiently strongly magnetized, it may carve out a true magnetosphere completely contained within the magnetosphere of the planet.

Magnetospheric phenomena are of both theoretical and phenomenological interest. Theory has benefited from the data collected in the vast plasma laboratory of space in which different planetary environments provide the analogue of different laboratory conditions. Furthermore, magnetospheric plasma interactions are important to diverse elements of planetary science. For example, plasma trapped in a planetary magnetic field can interact...
strongly with the planet’s atmosphere, heating the upper layers, generating neutral winds, ionizing the neutral gases and affecting the ionospheric flow. Energetic ions and electrons that precipitate into the atmosphere can modify atmospheric chemistry. Interaction with plasma particles can contribute to the isotopic fractionation of a planetary atmosphere over the lifetime of a planet. Impacts of energetic charged particles on the surfaces of planets and moons can modify surface properties, changing their albedos and spectral properties. The motions of charged dust grains in a planet’s environment are subject to both electrodynamic and gravitational forces; recent studies of dusty plasmas show that the former have been critical in determining the role and behavior of dust in the solar nebula as well as being significant in parts of the present-day solar system.

In Section 2, the different types of magnetospheres and related interaction regions are introduced. Section 3 presents the properties of observed planetary magnetic fields and discusses the mechanisms that produce such fields. Section 4 reviews the properties of plasmas contained within magnetospheres, describing their distribution, their sources, and some of the currents that they carry. Section 5 covers magnetospheric dynamics, both steady and “stormy”. Section 6 addresses the interactions of moons with planetary plasmas. Section 7 concludes the chapter with remarks on plans for future space exploration.

2. TYPES OF MAGNETOSPHERES

2.1. The Heliosphere

The solar system is dominated by the Sun, which forms its own magnetosphere referred to as the heliosphere. (See The Sun.) The size and structure of the heliosphere are governed by the motion of the Sun relative to the local interstellar medium, the density of the interstellar plasma, and the pressure exerted on its surroundings by the outflowing solar wind that originates in the solar corona. (See The Solar Wind.) The corona is a highly ionized gas, so hot that it can
escape the Sun’s immense gravitational field and flow outward at supersonic speeds. Through much of the heliosphere, the solar wind speed is not only supersonic but also much greater than the Alfvén speed \( v_A = B / (\mu_0 \rho)^{1/2} \), the speed at which rotational perturbations of the magnetic field propagate along the magnetic field in a magnetized plasma. (Here \( B \) is the magnetic field magnitude, \( \mu_0 \) is the magnetic permeability of vacuum, and \( \rho \) is the mass density of the plasma.) The solar wind is threaded by magnetic field lines that map back to the Sun. A useful and picturesque description of the field contained within a plasma relies on the idea that if the conductivity of a plasma is sufficiently large, the magnetic field is frozen into the plasma and field lines can be traced from their source by following the motion of the plasma to which it is frozen. Because the roots of the field lines remain linked to the rotating Sun (the Sun rotates about its axis with a period of approximately 25 days), the field lines twist in the form of an Archimedean spiral as illustrated in Figure 7.2. In the direction of the Sun’s motion relative to the interstellar plasma, the outflow is terminated by the forces exerted by the interstellar plasma. Elsewhere the flow is diverted within the boundary of the heliosphere. Thus, the Sun and the solar wind are (largely) confined within the heliospheric cavity; the heliosphere is the biggest of the solar system magnetospheres.

Our knowledge of the heliosphere beyond the orbits of the giant planets was for decades principally theoretical, but data acquired by Voyager 1 and 2 since their last planetary encounters in 1989 have provided important evidence in situ of the structure of the outer heliosphere. The solar wind density continues to decrease as the inverse square of the distance from the Sun; as the plasma becomes sufficiently tenuous, the pressure of the interstellar plasma impedes its further expansion. The solar wind slows down abruptly across a shock (referred to as the termination shock) before reaching the heliopause, the boundary that separates the solar wind from the interstellar plasma. (The different plasma regimes are schematically illustrated in Figure 7.3.) Voyager 1 encountered the termination shock on December 16, 2004, at a distance of 94 AU (AU is an

---

**FIGURE 7.2** The magnetic field of the Sun is carried by the solar wind away from the Sun and winds into a spiral. The heliospheric current sheet (colored magenta in the inset three-dimensional diagram) separates magnetic fields of opposite polarities and is warped into a “ballerina skirt” by combined effects of the Sun’s spin and the tilt of the magnetic field. The main diagram (two-dimensional projection) shows a cut through the inner heliosphere in the ecliptic plane (the plane of Earth’s orbit); the radial flow of the solar wind and the rotation of the Sun combine to twist the solar magnetic field (yellow lines) into a spiral. A parcel of solar wind plasma (traveling radially at an average speed of 400 km/s) takes about 4 days to travel from the Sun to Earth’s orbit at 1 AU. The dots and magnetic field lines labeled 1, 2, and 3 represent snapshots during this journey. Energetic particles emitted from the Sun travel much faster than the bulk solar wind and reach the Earth in minutes to hours. Traveling at the speed of light, solar photons reach the Earth in 8 min. Credit: Van Allen and Bagenal (1999).
astronomical unit, equal to the mean radius of Earth’s orbit or about $1.5 \times 10^8$ km) from the Sun and entered the heliosheath, the boundary layer between the termination shock and the heliopause. The encounter with the termination shock had long been anticipated as an opportunity to identify the processes that accelerate a distinct class of cosmic rays, referred to as anomalous cosmic rays (ACRs). ACRs are extremely energetic singly charged ions (energies of the order of $10$ MeV/nucleon) produced by ionization of interstellar neutrals. The mechanism that accelerates them to high energy is not established. Some models propose that these particles are ionized and accelerated near the termination shock. Although the Voyager data show no sign of a change in the energy spectrum or the intensity of the flux across the termination shock the connection of ACRs to the shock itself may be nonlocal, which could reconcile the observations with the theory. However, at this time there is not full understanding of the mechanism that produces ACRs.

Various sorts of electromagnetic waves and plasma waves have been interpreted as coming from the termination shock or the heliopause. Bursts of radio emissions that do not weaken with distance from known sources within the solar systems were observed intermittently by Voyager between 1983 and 2004. They are thought to be emissions generated when an interplanetary shock propagating outward from the Sun reaches the heliopause. Plasma waves driven by electron beams generated at the termination shock and propagating inward along the spiral field lines of the solar wind were also identified. As Voyager continues its journey out of the solar system, it will encounter the heliopause and enter the interstellar plasma beyond. Although the schematic heliosphere of Figure 7.3 suggests that beyond the heliopause, there is a region of interstellar wind, there is increasing evidence that the upstream flow may be submagnetosonic, in which case no shock develops. With the Voyager spacecraft continuing to provide data, direct evidence of the properties of the local interstellar medium will be beamed back to earth.

2.2. Magnetospheres of the Unmagnetized Planets

Earth has a planetary magnetic field that has long been used as a guide by such travelers as scouts and sea voyagers. However, not all of the planets are magnetized. Table 7.1 summarizes some key properties of some of the planets including their surface magnetic field strengths. The planetary magnetic field of Mars is extremely small, and the

FIGURE 7.3 Schematic illustration of the heliosphere. The direction of plasma flow in the local interstellar medium relative to the Sun is indicated, and the boundary between solar wind plasma and interstellar plasma is identified as the heliopause. A broad internal shock, referred to as the termination shock, is shown within the heliopause. Such a shock, needed to slow the outflow of the supersonic solar wind inside of the heliopause, is a new feature in this type of magnetosphere. Beyond the heliopause, the interstellar flow is diverted around the heliosphere and a shock that slows and diverts flow may or may not exist. Credit: Fisk, 2005.
planetary magnetic field of Venus is nonexistent. (See Mars and Venus: Surface and Interior.) The nature of the interaction between an unmagnetized planet and the supersonic solar wind is determined principally by the electrical conductivity of the body. If conducting paths exist across the planet’s interior or ionosphere, then electric currents flow through the body and into the solar wind where they create forces that slow and divert the incident flow. The diverted solar wind flows around a region that is similar to a planetary magnetosphere. Mars and Venus have ionospheres that provide the required conducting paths. The barrier that separates planetary plasma at these planets from solar wind plasma is referred to as an ionopause in analogy to the magnetopause of a magnetized planet. Earth’s Moon, with no ionosphere and a very low-conductivity surface, does not deflect the bulk of the solar wind incident on it. Instead, the solar wind runs directly into the surface, where it is absorbed. (See The Moon.) The absorption leaves the region immediately downstream of the Moon in the flowing plasma (the wake) devoid of plasma, but the void fills in as solar wind plasma flows toward the center of the wake. The different types of interaction are illustrated in Figure 7.4.

The magnetic structure surrounding Mars and Venus has features much like those found in a true magnetosphere surrounding a strongly magnetized planet. This is because the interaction causes the magnetic field of the solar wind to drape around the planet. The draped field stretches out downstream (away from the Sun), forming a magnetotail. The symmetry of the magnetic configuration within such a tail is governed by the orientation of the magnetic field in the incident solar wind, and that orientation changes with time. For example, if the interplanetary magnetic field (IMF) is oriented northward, the east—west direction lies in the symmetry plane of the tail and the northern lobe field (see Figure 7.1 for the definition of lobe) points away from

---

**TABLE 7.1** Properties of the Solar Wind and Scales of Planetary Magnetospheres

<table>
<thead>
<tr>
<th></th>
<th>Mercury</th>
<th>Venus</th>
<th>Earth</th>
<th>Mars</th>
<th>Jupiter</th>
<th>Saturn</th>
<th>Uranus</th>
<th>Neptune</th>
<th>Pluto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance, $a_{\text{planet}}$ (AU)$^1$</td>
<td>0.31–0.47</td>
<td>0.723</td>
<td>1$^2$</td>
<td>1.524</td>
<td>5.2</td>
<td>9.5</td>
<td>19</td>
<td>30</td>
<td>30–50</td>
</tr>
<tr>
<td>Solar wind density (amu/cm$^3$)$^2$</td>
<td>35–80</td>
<td>16</td>
<td>8</td>
<td>3.5</td>
<td>0.3</td>
<td>0.1</td>
<td>0.02</td>
<td>0.008</td>
<td>0.008–0.003</td>
</tr>
<tr>
<td>Radius, $R_p$ (km)</td>
<td>2439</td>
<td>6051</td>
<td>6373</td>
<td>3390</td>
<td>71,398</td>
<td>60,330</td>
<td>25,559</td>
<td>24,764</td>
<td>1153</td>
</tr>
<tr>
<td>Surface magnetic field, $B_0$ (nT)</td>
<td>195</td>
<td>–</td>
<td>30,600</td>
<td>–</td>
<td>430,000</td>
<td>21,400</td>
<td>22,800</td>
<td>13,200</td>
<td>Unknown</td>
</tr>
<tr>
<td>$R_{mp}$ ($R_{\text{planet}}$)$^3$</td>
<td>1.4–1.6 $R_M$</td>
<td>–</td>
<td>10 $R_E$</td>
<td>–</td>
<td>46 $R_I$</td>
<td>20 $R$</td>
<td>25 $R_U$</td>
<td>24 $R_N$</td>
<td></td>
</tr>
<tr>
<td>Observed size of magnetosphere ($R_{\text{planet}}$)</td>
<td>1.5 $R_M$</td>
<td>–</td>
<td>8–12 $R_E$</td>
<td>–</td>
<td>63–93 $R_I$</td>
<td>22–27 $R_S$</td>
<td>18 $R_U$</td>
<td>23–26 $R_N$</td>
<td>Unknown</td>
</tr>
<tr>
<td>Observed size of magnetosphere (km)</td>
<td>$3.6 \times 10^3$</td>
<td>–</td>
<td>$7 \times 10^4$</td>
<td>–</td>
<td>$7 \times 10^6$</td>
<td>$1 \times 10^6$</td>
<td>$5 \times 10^5$</td>
<td>$6 \times 10^3$</td>
<td></td>
</tr>
</tbody>
</table>

$^1$1 AU = 1.5 × 10$^8$ km.

$^2$The density of the solar wind fluctuates by about a factor of 5 about typical values of $\rho_{sw} \sim (8 \cdot \text{amu/cm}^3)/a_{\text{planet}}^2$.

$^3$Magnetopause nose distance, $R_{mp}$, is calculated using $R_{mp} = (\frac{B_0^2}{2 \mu_0 u^2})^{1/3}$ for typical solar wind conditions of $\rho_{sw}$ given above and $u \sim$ 400 km/s. For outer planet magnetospheres, this is usually an underestimate of the actual distance (Kivelson & Russell, 1995).

---

FIGURE 7.4 Schematic illustrations of the interaction regions surrounding, top, a planet like Mars or Venus, which is sufficiently conducting that currents close through the planet or its ionosphere (solar magnetic field lines are shown in yellow to red and are draped around and behind the planet) and, bottom, a body like the Moon, which has no ionosphere and low surface and interior conductivity. Credit: Steve Bartlett.
the Sun, while the southern lobe field points toward the Sun. A southward-oriented IMF would reverse these polarities, and other orientations would produce rotations of the symmetry axis.

Much attention has been paid to magnetic structures that form in and around the ionospheres of unmagnetized planets. Magnetic flux tubes of solar wind origin pile up at high altitudes at the dayside ionopause where, depending on the solar wind dynamic pressure, they may either remain for extended times, thus producing a magnetic barrier that diverts the incident solar wind, or penetrate to low altitudes in localized bundles. Such localized bundles of magnetic flux are often highly twisted structures stretched out along the direction of the magnetic field. Such structures, referred to as flux ropes, are illustrated in Figure 7.5.

Although, in the present epoch, Mars has only a small global scale magnetic field and interacts with the solar wind principally through currents that link to the ionosphere, there are portions of the surface over which local magnetic fields block the access of the solar wind to low altitudes. “Mini-magnetospheres” extending up to 1000 km form above the regions of intense crustal magnetization in the southern hemisphere; these mini-magnetospheres protect portions of the atmosphere from direct interaction with the solar wind. As a result, the crustal magnetization may have modified the evolution of the atmosphere and may still contribute to the energetics of the upper atmosphere.

2.3. Interactions of the Solar Wind with Asteroids, Comets, and Pluto

Asteroids are small bodies (<1000 km radius and more often only tens of kilometers) whose signatures in the solar wind were first observed by the Galileo spacecraft in the early 1990s. (See Main-Belt Asteroids.) Asteroid-related disturbances are closely confined to the regions near to and downstream of the magnetic field lines that pass through the body, and thus the interaction region is fan shaped as illustrated in Figure 7.6 rather than bullet shaped like Earth’s magnetosphere and there is no shock standing upstream of the disturbance in the solar wind. The signature found by Galileo in the vicinity of the asteroid Gaspra suggested that the asteroid is magnetized at a level similar to the magnetization of meteorites. Because the measurement locations were remote from the body, its field was not measured directly, and it is possible that the putative magnetic signature was a fortuitous rotation of the IMF. Data from other asteroids do not establish unambiguously the strength of their magnetic fields. A negligibly small magnetic field was measured by the

![Interior structure of flux rope](image)

![Distribution of flux ropes](image)

**Figure 7.5** Schematic illustration of a flux rope, a magnetic structure that has been identified in the ionosphere of Venus (shown as black dots within the ionosphere) and extensively investigated (a low-altitude pass of the Pioneer Venus Orbiter is indicated by the dashed curve). The rope (see earlier) has an axis aligned with the direction of the central field. Radially away from the center, the field wraps around the axis, its helicity increasing with radial distance from the axis of the rope. Structures of this sort are also found in the solar corona, near the magnetopause, and in the magnetotails of magnetized planets. Credit: Steve Bartlett.
Comets are also small bodies. The spectacular appearance of an active comet, which can produce a glow over a large visual field extending millions of kilometers in space on its approach to the Sun, is somewhat misleading because comet nuclei are no more than tens of kilometers in diameter. It is the gas and dust released from these small bodies by solar heating that we see spread out across the sky. Some of the gas released by the comet remains electrically neutral, with its motion governed by purely mechanical laws, but some of the neutral matter becomes ionized either by photoionization or by exchanging charge with ions of the solar wind. The newly ionized cometary material is organized in interesting ways that have been revealed by spacecraft measurements in the near neighborhood of comets Halley, Giacobini-Zinner, Borrelly, and others. Figure 7.7 shows schematically the types of regions that have been identified, illustrating clearly that the different gaseous regions fill volumes of space many orders of magnitude larger than the actual solid comet. The solar wind approaching the comet first encounters the expanding neutral gases blown off the comet. As the neutrals are ionized by solar photons, they extract momentum from the solar wind, and the flow slows a bit. Passing through a shock that further decelerates the flow, the solar wind encounters ever-increasing densities of newly ionized gas of cometary origin, referred to as pickup ions. Energy is extracted from
the solar wind as the pickup ions are swept up, and the flow slows further. Still closer to the comet, in a region referred to as the cometopause, a transition in composition occurs as the pickup ions of cometary origin begin to dominate the plasma composition. Close to the comet, at the contact surface, ions flowing away from the comet carry enough momentum to stop the flow of the incident solar wind. Significant asymmetry of the plasma distribution in the vicinity of a comet may arise if strong collimated jets of gas are emitted by the cometary nucleus. Such jets have been observed at Halley’s comet and at comet Borrelly.

Pluto is also a small body that was classified as a planet until 2006 and more recently reclassified as a dwarf planet. Pluto’s interaction with the solar wind has not yet been observed, but it is worth speculating about what that interaction will be like to test our understanding of comparative planetology. (See Pluto.) The solar wind becomes tenuous and easily perturbed at large distances from the Sun (near 30 AU), and either escaping gases or a weak internal magnetic field could produce an interaction region many times Pluto’s size. At some phases of its 248-year orbital period, Pluto moves close enough to the Sun for its surface ice to sublimate, producing an atmosphere and possibly an ionosphere. Models of Pluto’s atmosphere suggest that the gases would then escape and flow away from the planet. If the escape flux is high, the solar wind interaction would then appear more like that at a comet than like that at Venus or Mars. Simulations show a very asymmetric shock surrounding the interaction region for a small but possible neutral escape rate. Pluto’s moon, Charon, may serve as a plasma source within the magnetosphere, and this could have interesting consequences of the type addressed in Section 6 in relation to the moons of Jupiter and Saturn. As is the case for small asteroids and comets, ions picked up in the solar wind at Pluto have gyroradii and ion inertial lengths that are large compared with the size of the obstacle, a situation that adds asymmetry and additional complexity to the interaction. For most of its orbital period, Pluto is so far from the Sun that its atmosphere disappears and its interaction with the solar wind is more likely to resemble that of the Moon, with absorption occurring at the sunward surface and a void developing in its wake. It seems unlikely that a small icy body will have an internal magnetic field large enough to produce a magnetospheric interaction region, but one must recognize that actual observations of the magnetic fields of small bodies have repeatedly challenged our ideas about magnetic field generation.

### 2.4. Magnetospheres of Magnetized Planets

In a true magnetosphere, the scale size is set by the distance, \( R_{\text{MP}} \), along the planet–Sun line at which the sum of the pressure of the planetary magnetic field and the pressure exerted by plasma confined within that field balance the dynamic pressure of the solar wind. (The dynamic pressure is \( \rho u^2 \) where \( \rho \) is the mass density and \( u \) is its flow velocity in the rest frame of the planet. The thermal and magnetic pressures of the solar wind are small compared with its dynamic pressure.) Assuming that the planetary magnetic field is dominated by its dipole moment and that the plasma pressure within the magnetosphere is small, one can estimate \( R_{\text{MP}} \) as

\[
R_{\text{MP}} \approx R_P \left( \frac{B_0^2}{2 \mu_0 \rho u^2} \right)^{1/6}
\]

where \( u \) is the surface equatorial field of the planet and \( R_P \) is its radius. Table 7.1 gives the size of the magnetosphere, \( R_{\text{MP}} \), calculated from its internal field and observed for the different planets and shows the vast range of scale sizes both in terms of the planetary radii and of absolute distance.

Within a magnetosphere, the magnetic field differs greatly from what it would be if the planet were placed in a vacuum. The field is distorted, as illustrated in Figure 7.1, by currents carried on the magnetopause and in the plasma trapped within the magnetosphere. Properties of the trapped plasma and its sources are discussed in Section 4. An important source of magnetospheric plasma is the solar wind. Figure 7.1 makes it clear that, along most of the boundary, solar wind plasma would have to move across magnetic field lines to enter the magnetosphere. The Lorentz force of the magnetic field opposes such motion. However, in the polar cusp shocked solar wind plasma of the magnetosheath easily penetrates the boundary by moving along the field. Other processes that enable solar wind plasma to penetrate the boundary are discussed in Section 5.

### 3. PLANETARY MAGNETIC FIELDS

Because the characteristic timescale for thermal diffusion is greater than the age of the solar system, the planets tend to have retained their heat of formation. At the same time, the characteristic timescale for diffusive decay of a magnetic field in a planetary interior is much less than the age of the planets. Consequently, primordial fields and permanent magnetism on a planetary scale are small and the only means of providing a substantial planetary magnetic field is an internal dynamo (see Magnetic Field Generation).

For a planet to have a magnetic dynamo, it must have a large region that is fluid, electrically conducting, and undergoing convective motion. The deep interiors of the planets and many larger satellites are expected to contain electrically conducting fluids: terrestrial planets and the larger satellites have differentiated cores of liquid iron alloys; at high pressures in the interiors of the giant planets Jupiter and Saturn, hydrogen behaves like a liquid metal; for Uranus and Neptune, a water—ammonia—methane mixture forms a deep conducting “ocean”. (See Interiors of the Giant Planets.) The fact that some planets and satellites do not have dynamos tells us that their
interiors are stably stratified and do not convect or that the interiors have solidified. Models of the thermal evolution of terrestrial planets show that as the object cools, the liquid core ceases to convect, and further heat is lost by conduction alone. In some cases, such as the Earth, convection continues because the nearly pure iron solidifies out of the alloy in the outer core, producing an inner solid core and creating compositional gradients that drive convection in the liquid outer core. The more gradual cooling of the giant planets also allows convective motions to persist.

Of the eight planets, six are known to generate magnetic fields in their interiors. Exploration of Venus has provided an upper limit to the degree of magnetization comparable with the crustal magnetization of the Earth, suggesting that its core is stably stratified and that it does not have an active dynamo. The question of whether Mars does or does not have a weak internal magnetic field was disputed for many years because spacecraft magnetometers had measured the field only far above the planet’s surface. The first low-altitude magnetic field measurements were made by Mars Global Surveyor in 1997. It is now known that the surface magnetic field of Mars is very small (|B| < 10 nT or 1/3000 of Earth’s equatorial surface field) over most of the northern hemisphere but that in the southern hemisphere there are extensive regions of intense crustal magnetization as already noted. Pluto has yet to be explored. Models of Pluto’s interior suggest that it is probably differentiated, but its small size makes one doubt that its core is convecting and any magnetization is likely to be remnant. Earth’s moon has a negligibly small planetscale magnetic field, although localized regions of the surface are highly magnetized. Jupiter’s large moons are discussed in Section 6.

The characteristics of the six known planetary fields are listed in Table 7.2. Assuming that each planet’s magnetic field has the simplest structure, a dipole, we can characterize the magnetic properties by noting the equatorial field strength (B_0) and the tilt of the axis with respect to the planet’s spin axis. For all the magnetized planets other than Mercury, the surface fields are on the order of a Gauss = 10^{-4} T, meaning that their dipole moments are of order R_p^3 10^{-4} T, where R_p is the planetary radius (i.e. the dipole moments scale with planetary size).

The degree to which the dipole model is an oversimplification of more complex structure is indicated by the ratio of maximum to minimum values of the surface field. This ratio has a value of 2 for a dipole. The larger values, particularly for Uranus and Neptune, are indications of strong nondipolar contributions to the planets’ magnetic fields. Similarly, the fact that the magnetic axes of these two planets are strongly tilted (see Figure 7.8) also suggests that the dynamos in the icy giant planets may be significantly different than those of the planets with aligned, dipolar planetary magnetic fields.

The size of a planet’s magnetosphere (R_MP) depends not only on the planet’s radius and magnetic field but also on the ambient solar wind density, which decreases as the inverse square of the distance from the Sun. (The solar wind speed is approximately constant with distance from the Sun.) Thus, it is not only planets with strong magnetic fields that have large magnetospheres but also the planets Uranus and Neptune whose weak magnetic fields create moderately large magnetospheres because of the weak force exerted by the tenuous solar wind far from the Sun. Table 7.1 shows that the measured sizes of planetary magnetospheres generally agree quite well with the theoretical R_MP values. Jupiter, for which the plasma pressure inside the magnetosphere is sufficient to further “inflate” the magnetosphere, is the only notable exception. Jupiter’s strong internal field combines with the relatively low solar wind

<table>
<thead>
<tr>
<th>TABLE 7.2 Planetary Magnetic Fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnetic moment (M_earth)</td>
</tr>
<tr>
<td>-----------------------------------</td>
</tr>
<tr>
<td>Surface magnetic field at dipole equator (nT)</td>
</tr>
<tr>
<td>Maximum/minimum</td>
</tr>
<tr>
<td>Dipole tilt and sense</td>
</tr>
<tr>
<td>Oliquidity</td>
</tr>
<tr>
<td>Solar wind angle</td>
</tr>
</tbody>
</table>

1M_earth = 7.906 x 10^23 Gauss cm^3 = 7.906 x 10^3 T m^3.
2Ratio of maximum surface field to minimum (equal to 2 for a centered dipole field).
3Angle between the magnetic axis and rotation pole (S or N).
4The inclination of the equator to the orbit.
5Range of angle between the radial direction from the Sun and the planet’s rotation axis over an orbital period.
density at 5 AU to make the magnetosphere of Jupiter a huge object—about 1000 times the volume of the Sun, with a tail that extends at least 6 A.U. in the antisunward direction, beyond the orbit of Saturn. If the jovian magnetosphere were visible from Earth, its angular size would be much larger than the size of the Sun, even though it is at least four times farther away. The magnetospheres of the other giant planets are smaller (although large compared with the Earth’s magnetosphere), having scales of about 20 times the planetary radius, comparable with the size of the Sun. Mercury’s magnetosphere is extremely small because the planet’s magnetic field is weak and the solar wind close to the Sun is very dense. Figure 7.9 compares the sizes of several planetary magnetospheres.

Although the size of a planetary magnetosphere depends on the strength of a planet’s magnetic field, the configuration and internal dynamics depend on the field orientation (illustrated in Figure 7.8). At a fixed phase of planetary rotation, such as when the dipole tilts toward the Sun, the orientation of a planet’s magnetic field is described by two angles (tabulated in Table 7.2): the tilt of the magnetic field with respect to the planet’s spin axis and the angle between the planet’s spin axis and the solar wind direction, which is generally within a few degrees of being radially outward from the Sun. Because the direction of the spin axis with respect to the solar wind direction varies only over a planetary year (many Earth years for the outer planets), and the planet’s magnetic field is assumed to vary only on geological timescales, these two angles are constant for the purposes of describing the magnetospheric configuration at a particular epoch. Earth, Jupiter and Saturn have small dipole tilts and relatively small obliquities. This means that changes of the orientation of the magnetic field with respect to the solar wind over a planetary rotation period and seasonal variations, though detectable, produce only subtle magnetospheric effects. Thus, Mercury, Earth, Jupiter, and Saturn have reasonably symmetric, quasi-stationary magnetospheres, with the first three exhibiting a wobble at the planetary rotation period owing to their \( \sim 10^9 \) dipole tilts. In contrast, the large dipole tilt angles of Uranus and Neptune imply that the orientation of their magnetic fields with respect to the interplanetary flow direction varies greatly over a planetary rotation period, resulting in highly asymmetric magnetospheres that vary at the period of planetary rotation. Furthermore, Uranus’ large obliquity means that its magnetosphere undergoes strong seasonal changes of its global configuration over its 84-year orbital period.

4. MAGNETOSPHERIC PLASMAS

4.1. Sources of Magnetospheric Plasmas

Magnetospheres contain considerable amounts of plasma whose sources are both internal and external (see Table 7.3). The main source of plasma in the solar system is the Sun. The solar corona, the upper atmosphere of the Sun (which has been heated to temperatures of 1–2 million Kelvin), streams away from the Sun at a more or less steady rate of \( 10^9 \) kg/s carrying approximately equal numbers \( (8 \times 10^{35} \text{ s}^{-1}) \) of electrons and ions. The boundary between the solar wind and a planet’s magnetosphere, the magnetopause, is not entirely plasma tight. Wherever the IMF
has a component antiparallel to the planetary magnetic field near the magnetopause boundary, magnetic reconnection (discussed in Section 5) is likely to occur, and solar wind plasma can enter the magnetosphere across the magnetopause. Solar wind material is identified in the magnetosphere by its energy and characteristic composition of protons (H⁺) with ~4% alpha particles (He²⁺) and trace heavy ions, many of which are highly ionized.

A secondary source of plasma is the ionosphere. Although ionospheric plasma is generally cold and gravitationally bound to the planet, a small fraction of particles can acquire sufficient energy to escape up magnetic field lines and into the magnetosphere. In some cases, field-aligned potential drops accelerate ionospheric ions and increase the escape rate. Ionospheric plasma has a composition that reflects the composition of the planet’s atmosphere (e.g., abundant O⁺ for the Earth and H⁺ for the outer planets).

The sulfur and oxygen ions that dominate Jupiter’s magnetospheric plasma are formed by breakdown and ionization of sulfur and sulfur dioxide whose source is the volcanoes of Io. Energetic ions impacting Io’s surface or atmosphere can sputter off ions of lower energy through a direct interaction and can also create an extensive cloud of neutral atoms that are subsequently ionized, possibly far from the satellite. Although the sputtering process, which removes at most a few microns of surface ice per thousand years, is probably insignificant in geological terms, sputtering has important consequences for the optical properties of satellite or ring surfaces.

Water-product ions are ubiquitous in Saturn’s magnetosphere. Initially it was thought that these ions were formed from neutrals sputtered from ring particles and from the surfaces of icy satellites, but it was hard to account for the observed densities. The dominant source of the water group ions was established only after measurements and images from the Cassini orbiter revealed that plumes or geysers spouting from the surface of the tiny moon Enceladus are the dominant source of the water group ions at Saturn.

In the hot tenuous plasmas of planetary magnetospheres, collisions between particles are very rare. By contrast, in the cold dense plasmas of a planet’s ionosphere, collision rates are high enough to allow ionospheric plasmas to conduct currents. Cold, dense, collision-dominated plasmas are expected to be in thermal equilibrium, but such equilibrium was not originally expected for the hot, tenuous collisionless plasmas of the magnetosphere. Surprisingly, even hot tenuous plasmas in space are generally found not far from equilibrium (i.e., their particle distribution functions are observed to be approximately Maxwellian, although the ion and electron populations often have different temperatures). This fact is remarkable because some of the source mechanisms tend to produce particles whose initial energies fall in a very narrow range and timescales for equilibration by means of Coulomb collisions are usually much longer than transport timescales. A distribution close to Maxwellian is achieved by interaction with waves in the plasma. Space plasmas support many different types of plasma waves that can grow in amplitude when free energy is present in the form of non-Maxwellian energy distributions, unstable spatial
distributions, or anisotropic velocity—space distributions of newly created ions. Interactions between plasma waves and particle populations not only bring the bulk of the plasma toward thermal equilibrium but also accelerate or scatter suprathermal particles.

Plasma detectors mounted on spacecraft can provide detailed information about the particles’ velocity distribution, from which bulk parameters such as density, temperature, and flow velocity are derived, but plasma properties are determined only in the vicinity of the spacecraft. Data from planetary magnetospheres other than Earth’s are limited in duration and spatial coverage, so there are considerable gaps in our knowledge of the changing properties of the many different plasmas in the solar system. Some of the most interesting space plasmas, however, can be remotely monitored by observing emissions of electromagnetic radiation. Dense plasmas, such as Jupiter’s plasma torus, comet tails, Venus’s ionosphere, and the solar corona, can radiate collisionally excited line emissions at optical or UV wavelengths. Radiative processes, particularly at UV wavelengths, can be significant sinks of plasma energy. Figure 7.10 shows an image of optical emission from the plasma that forms a ring or torus deep within Jupiter’s magnetosphere near the orbit of its moon, Io (see Section 6). Observations of these emissions give compelling evidence of the temporal and spatial variability of the Io plasma torus. Similarly, when magnetospheric particles bombard the planets’ polar atmospheres, various auroral emissions are generated from radio to X-ray wavelengths and these emissions can also be used for remote monitoring of the system. (See Atmospheres of the Giant Planets.) Thus, our knowledge of space plasmas is based on combining the remote sensing of plasma phenomena with available spacecraft measurements that provide “ground truth” details of the particles’ velocity distribution and of the local electric and magnetic fields that interact with the plasma.

4.2. Energetic Particles

Significant populations of particles at keV–MeV energies, well above the energy of the thermal population, are found in all magnetospheres. The energetic particles are largely trapped in long-lived radiation belts (summarized in Table 7.4) by the strong planetary magnetic field. Where do these energetic particles come from? Since the interplanetary medium contains energetic particles of solar and galactic origins an obvious possibility is that these energetic particles are “captured” from the external medium. In most cases, the observed high fluxes are hard to explain without identifying additional internal sources. Compositional evidence supports the view that some fraction of the thermal plasma is accelerated to high energies, either by tapping the rotational energy of the planet, in the cases of Jupiter and Saturn, or by acceleration in the distorted and dynamic magnetic field in the magnetotails of Earth, Uranus, and Neptune. If the energy density of the energetic particle populations is comparable with the magnetic field energy density, currents develop that significantly modify the planetary magnetic field. Table 7.4 shows that this occurs at Jupiter and Saturn, where the high particle pressures inflate and stretch out the magnetic field and generate a strong azimuthal current in the magnetodisc. Even though

<table>
<thead>
<tr>
<th>TABLE 7.3 Plasma Characteristics of Planetary Magnetospheres</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
</tr>
<tr>
<td>Maximum density (cm⁻³)</td>
</tr>
<tr>
<td>Composition</td>
</tr>
<tr>
<td>Dominant source</td>
</tr>
<tr>
<td>Strength (ions/s)</td>
</tr>
<tr>
<td>(kg/s)</td>
</tr>
<tr>
<td>Lifetime</td>
</tr>
<tr>
<td>Plasma motion</td>
</tr>
</tbody>
</table>

¹Inside plasmasphere.
²Outside plasmasphere.

Based on Bagenal (2009) and Kivelson (2006)
Uranus and Neptune have significant radiation belts, the energy density of particles remains small compared with the magnetic energy density and the azimuthal current is very weak. In Earth’s magnetosphere, the azimuthal current, referred to as the ring current, is extremely variable, as discussed in Section 5. Relating the magnetic field produced by the azimuthal current to the kinetic energy of the trapped particle population (scaled to the dipole magnetic energy external to the planet), we find that even though the total energy content of magnetospheres varies by many orders of magnitude and the sources are very different, the net particle energy builds up to only 1/1000 of the magnetic field energy in each magnetosphere. Earth, Jupiter, and Saturn all have energetic particle populations close to this limit. The energy in the radiation belts of Uranus and Neptune is much below this limit, perhaps because it is harder to trap particles in nondipolar magnetic fields.

We have commented on ways in which particles gain energy. Where do these energetic particles go? Most appear to diffuse inward toward the planet. Loss processes for energetic particles in the inner magnetospheres of the terrestrial planets arise largely from inward diffusion to low

| TABLE 7.4 Energetic Particle Characteristics in Planetary Magnetospheres |
|-------------------------|---------|---------|---------|---------|---------|
|                        | Earth   | Jupiter | Saturn  | Uranus  | Neptune |
| Phase space density1   | 20,000  | 200,000 | 60,000  | 800     | 800     |
| Plasma beta2           | <1      | >1      | >1      | ~0.1    | ~0.2    |
| Ring current, ΔB (nT)3 | 10−200  | 200     | 10      | <1      | <0.1    |
| Auroral power (W)      | 10^{10} | 10^{12} | 10^{11} | 5 × 10^9| 2−8 × 10^7 |

1The phase space density of energetic particles (in this case 100 MeV/Gauss ions) is measured in units of (cm^{-2} sr MeV)^{-1} and is listed near its maximum value.
2The ratio of the particle pressure to the magnetic pressure of a plasma, nkT/(\mathbf{B}^2/2\mu_0). These values are typical for the body of the magnetosphere. Higher values are often found in the tail plasma sheet and, in the case of the Earth, at times of enhanced ring current.
3The magnetic field produced at the surface of the planet due to the ring current of energetic particles in the planet’s magnetosphere.
altitudes or charge exchange with neutral clouds, and scattering by waves so that the particles stream into the upper atmospheres of the planets. At Jupiter and Saturn, energetic particles can be lost through absorption by rings and satellites. Some energetic particles precipitate into the atmosphere at high latitudes where they excite auroral emission and deposit large amounts of energy, at times exceeding the local energy input from the Sun. The presence of high fluxes of energetic ions and electrons of the radiation belts must be taken into account in designing and operating spacecraft. At Earth, relativistic electron fluxes build to extremely high levels during magnetically active times referred to as storm times. High fluxes of relativistic electrons affect sensitive electronic systems and have caused anomalies in the operation of spacecraft. At Earth damaging levels of relativistic electrons occur intermittently but in the inner part of Jupiter’s magnetosphere, such high fluxes are always present. Proposed missions to Jupiter’s moon Europa must be designed with attention to the fact that the energetic particle radiation near Europa’s orbit is punishingly intense.

5. DYNAMICS

Magnetospheres are ever-changing systems. Changes in the solar wind, in plasma source rates, and in energetic cosmic ray fluxes can couple energy, momentum, and additional particle mass into the magnetosphere and thus drive magnetospheric dynamics. Sometimes the magnetospheric response is direct and immediate. For example, an increase of the solar wind dynamic pressure compresses the magnetosphere. Both the energy and the pressure of field and particles then increase even if no particles have entered the system. Sometimes the change in both field and plasma properties is gradual, similar to a spring being slowly stretched. Sometimes, as for a spring stretched beyond its breaking point, the magnetosphere responds in a very nonlinear manner, with both field and plasma experiencing large-scale abrupt changes. These changes can be identified readily in records of magnetometers (a magnetometer is an instrument that measures the magnitude and direction of the magnetic field), in scattering of radio waves by the ionosphere or emissions of such waves from the ionosphere, and in the magnetic field configuration, plasma conditions and flows, and energetic particle fluxes measured by a spacecraft moving through the magnetosphere itself.

Auroral activity is the most dramatic signature of magnetospheric dynamics and it is observed on distant planets as well as on Earth. Accounts of the terrestrial aurora (the lights flickering in the night sky that inspired fear and awe) date to ancient days, but the oldest scientific records of magnetospheric dynamics are measurements of fluctuating magnetic fields at the surface of the Earth. Consequently, the term geomagnetic activity is used to refer to magnetospheric dynamics of all sorts. Fluctuating magnetic signatures with timescales from seconds to days are typical. For example, periodic fluctuations at frequencies between ~1 mHz and ~1 Hz are called magnetic pulsations. In addition, impulsive decreases in the horizontal north—south component of the surface magnetic field (referred to as the H-component) with timescales of tens of minutes occur intermittently at latitudes between 65° and 75° often several times a day. The field returns to its previous value typically in a few hours. These events are referred to as substorms. A signature of a substorm at a ~70° latitude magnetic observatory is shown in Figure 7.11. The H-component decreases by hundreds to 1000 nT (the Earth’s surface field is 31,000 nT near the equator). Weaker signatures can be identified at lower and higher latitudes. Associated with the magnetic signatures and the current systems that produce them are other manifestations of magnetospheric activity including particle precipitation and auroral activation in the polar region and changes within the magnetosphere previously noted.

The auroral activity associated with a substorm can be monitored from above by imagers on spacecraft at high polar altitudes. The dramatic intensification of the brightness of the aurora as well as its changing spatial extent can thereby be accurately determined. Figure 7.12 shows an image of the aurora taken by the Far Ultraviolet Imaging System on the IMAGE spacecraft on July 15, 2000. Note that the intense brightness is localized in a high-latitude band surrounding the polar regions. This region of auroral activity is referred to as the auroral oval. Only during very
intense substorms does the auroral region move far enough equatorward to be visible over most of the United States. The intensity of substorms and other geomagnetic activity is governed to some extent by the speed of the solar wind but of critical importance is the orientation of the magnetic field embedded in the solar wind incident on a magnetized planet. The fundamental role of the magnetic field in the solar wind may seem puzzling. It is the orientation of the IMF that is critical, and at Earth substorm activity normally occurs following an interval during which the interplanetary field has been tilted southward. The issue is subtle. Magnetized plasma flowing through space is frozen to the magnetic field. The high conductivity of the plasma prevents the magnetic field from diffusing through the plasma, and, in turn, the plasma particles are bound to the magnetic field by a \( \mathbf{v} \times \mathbf{B} \) Lorentz force that causes the particles to spiral around a field line. How, then, can a plasma ion or electron move from a solar wind magnetic field line to a magnetospheric field line?

The coupling arises through a process called reconnection, which occurs when plasmas bound on flux tubes with oppositely directed fields approach each other sufficiently closely. The weak net field at the interface may be too small to keep the plasma bound on its original flux tube and the field connectivity can change. Newly linked field lines will be bent at the reconnection location. The curvature force at the bend accelerates plasma away from the reconnection site. At the dayside magnetopause, for example, solar wind magnetic flux tubes and magnetospheric flux tubes can reconnect in a way that extracts energy from the solar wind and allows solar wind plasma to penetrate the magnetopause. A diagram first drawn in a French café by J. W. Dungey in 1961 (and reproduced frequently thereafter) provides the framework for understanding the role of magnetic reconnection in magnetospheric dynamics (Figure 7.13). Shown in the diagram on the top are southward-oriented solar wind field lines approaching the dayside magnetopause. Just at the nose of the magnetosphere, the northern ends of the solar wind field lines break their connection with the southern ends, linking instead with magnetospheric fields. Accelerated flows develop near the reconnection site. The reconnected field lines are dragged tailward by their ends within the solar wind, thus forming the tail lobes. When the magnetic field of the solar wind points strongly northward at Jupiter or Saturn, reconnection is also thought to occur at the low-latitude dayside magnetopause, but the full process has not yet been documented by observations, although there is some evidence that auroral displays intensify at Jupiter as at Earth when magnetopause reconnection is occurring. At Earth, if the reconnection is persistent, disturbances intensify. Energetic particle fluxes increase and significant fluxes of energetic particles may appear even at low latitudes and the ring current (see Section 4.3) intensifies. If dayside reconnection occurs at Earth, the solar wind transports magnetic flux from the dayside to the nightside. The path of the foot of the flux tube crosses the center of the polar cap, starting at the polar edge of the dayside auroral zone and moving to the polar edge of the nightside auroral zone as shown schematically in Figure 7.13(a). Ultimately that flux must return, and the process is also shown, both in the magnetotail where reconnection is shown closing a flux tube that had earlier been opened on the dayside and in the polar cap (Figure 7.13(b)) where the path of the foot of the flux tube appears at latitudes below the auroral zone, carrying the flux back to the dayside. In the early stage of a substorm (between A and B in Figure 7.11), the rate at which magnetic flux is transported to the nightside is greater than the rate at
which it is returned to the dayside. This builds up stress in the tail, reducing the size of the region within the tail where the magnetic configuration is dipole like and compressing the plasma in the plasma sheet (see Figure 7.1). Only after reconnection starts on the nightside (at (B) in Figure 7.11) does flux return to the dayside. Complex magnetic structures form in the tail as plasma jets both earthward and tailward from the reconnection site. In some cases, the magnetic field appears to enclose a bubble of tailward-moving plasma called a plasmoid. At other times, the magnetic field appears to twist around the earthward- or tailward-moving plasma in a flux rope (see Figure 7.5). Even on the dayside magnetopause, twisted field configurations seem to develop as a consequence of reconnection, and, because these structures are carrying flux tailward, they are called flux transfer events or FTEs.

The diversity of the processes associated with geomagnetic activity, their complexity and the limited data on which studies of the immense volume of the magnetosphere must be based have constrained our ability to understand details of substorm dynamics. However, both new research tools and anticipated practical applications of improved understanding have accelerated progress toward the objective of being able to predict the behavior of the magnetosphere during a substorm. The new tools available in this century include a fleet of spacecraft in orbit around and near the Earth (ACE, Wind, Polar, Geotail, Cluster, Double Star Themis, and several associated spacecraft) that make coordinated measurements of the solar wind and of different regions within the magnetosphere, better instruments that make high time resolution measurements of particles and fields, spacecraft imagers covering a broad spectral range, ground radar systems, and networks of magnetometers. The anticipated applications relate to the concept of forecasting space weather (See Space Weather) much as we forecast weather on the ground. An ability to
anticipate an imminent storm and take precautions to protect spacecraft in orbit, astronauts on space stations, and electrical systems on the surface (which can experience power surges during big storms) has been adopted as an important goal by the space science community, and improvements in our understanding of the dynamics of the magnetosphere will ultimately translate into a successful forecasting capability.

Dynamical changes long studied at Earth are also expected in the magnetospheres of the other planets. In passes through Mercury’s magnetosphere, the Mariner spacecraft observed substorms that lasted for minutes and at the present epoch the Messenger spacecraft is compiling measurements that further characterize the dynamics of the system. FTEs can occur on Mercury’s magnetopause with such frequency that the observations are described as showers of FTEs. The occurrence of FTEs at Jupiter’s magnetopause is infrequent, suggesting that the formation process is controlled by the plasma parameters of the solar wind.

Substorms or related processes should also occur at the outer planets, but the timescale for global changes in a system is expected to increase as its size increases. For a magnetosphere as large as Jupiter’s, the equivalent of a substorm is not likely to occur more often than every few days or longer, as contrasted with several each day for Earth. Until December 1995 when Galileo began to orbit Jupiter, no spacecraft had remained within a planetary magnetosphere long enough to monitor its dynamical changes. Data from Galileo’s 8-year orbital reconnaissance of Jupiter’s equatorial magnetosphere demonstrate unambiguously that this magnetosphere like that of Earth experiences intermittent injections of energetic particles and, in the magnetotail, unstable flows correlated with magnetic perturbations of the sort that characterize terrestrial substorms. Yet the source of the disturbances is not clear. The large energy density associated with the rotating plasma suggests that centrifugally driven instabilities must themselves contribute to producing these dynamic events. Plasma loaded into the magnetosphere near Io may ultimately be flung out down the magnetotail, and this process may be intermittent, possibly governed both by the strength of internal plasma sources and by the magnitude of the solar wind dynamic pressure that determines the location of the magnetopause. Various models have been developed to describe the pattern of plasma flow in the magnetotail as heavily loaded magnetic flux tubes dump plasma on the nightside, but it remains ambiguous what aspects of the jovian dynamics are internally driven and what aspects are controlled by the solar wind.

Whether or not the solar wind plays a role in the dynamics of the jovian magnetosphere, it is clear that a considerable amount of solar wind plasma enters Jupiter’s magnetosphere. One way to evaluate the relative importance of the solar wind and Io as plasma sources is to estimate the rate at which plasma enters the magnetosphere when dayside reconnection is active and compare that estimate with the few $10^{28}$ ions/s whose source is Io. If the solar wind near Jupiter flows at 400 km/s with a density of 0.5 particles/cm$^3$, it carries $\sim 10^{31}$ particles/s onto the circular cross-section of a magnetosphere with $>50 R_J$ radius. If reconnection is approximately as efficient as it is at Earth, where a 10% efficiency is often suggested, and if a significant fraction of the solar wind ions on reconnected flux tubes enter the magnetosphere, the solar wind source could be important, and, as at Earth, the solar wind may contribute to the variability of Jupiter’s magnetosphere. Galileo data are still being analyzed in the expectation that answers to the question of how magnetospheric dynamics are controlled are contained in the archives of the mission and the Juno mission that will reach Jupiter in October, 2016, will provide additional useful data.

Cassini has been monitoring Saturn’s magnetosphere since mid-2004. Earlier passes through the magnetosphere (Voyager 1 and 2 and Pioneer 10) were too rapid to provide insight into the dynamics of Saturn’s magnetosphere or even to identify clearly the dominant sources of plasma. Cassini, in orbit at Saturn since 2004, has characterized such dynamical features as boundary oscillations, current sheet reconfigurations, and other aspects of magnetotail dynamics (Figure 7.14).

Saturn’s magnetic moment is closely aligned with its spin axis, and there is no evident longitudinal magnetic asymmetry. Nonetheless, both the intensity of radio emissions and magnetic perturbations are modulated at approximately the planetary rotation period. The period changes slowly with Saturn season, which makes it likely that the source of periodicity is not linked to the deep interior of the planet, but there is not yet consensus on the source of the periodicity.

The energetic particle detector on Cassini is capable of “taking pictures” of particle fluxes over large regions of the magnetosphere. The technique relies on the fact that if an energetic ion exchanges charge with a slow-moving neutral, a fast-moving neutral particle results. The energetic particle detector then acts like a telescope, collecting energetic neutrals instead of light and measuring their intensity as a function of the look direction. The images show that periodic intensifications and substorm-like acceleration are present at Saturn.

It is still uncertain just how particle transport operates at Saturn and how the effects of rotation compare in importance with convective processes imposed by interaction with the solar wind. It is already clear that with a major source of plasma localized close to the planet (see discussion of the plume of Enceladus in Section 6), Saturn’s magnetospheric dynamics resemble those of Jupiter more closely than those of Earth. Observations scheduled
in the coming years will provide greater insight into this matter.

6. INTERACTIONS WITH MOONS

Embedded deeply within the magnetosphere of Jupiter, the four Galilean moons (Io, Europa, Ganymede, and Callisto whose properties are summarized in Table 7.5) are immersed in magnetospheric plasma that corotates with Jupiter (i.e. flows once around Jupiter in each planetary spin period). At Saturn, Titan, shrouded by a dense atmosphere, Enceladus, and the other icy moons are also embedded within the flowing plasma of a planetary magnetosphere. (See Titan.) In the vicinity of these moons, interaction regions with characteristics of induced or true magnetospheres develop. The scale of each interaction region is linked to the size of the moon and to its electromagnetic properties. Ganymede, Callisto, and

TABLE 7.5 Properties of Selected Moons of Jupiter and Saturn

<table>
<thead>
<tr>
<th>Moon</th>
<th>Orbit Distance ($R_p$)</th>
<th>Rotation Period (Earth days)</th>
<th>Radius (km)</th>
<th>Radius of Core (moon radii)</th>
<th>Mean Density (kg/m$^3$)</th>
<th>Surface $B$ at Dipole Equator (nT)</th>
<th>Approx. Average $B_{avg}$ (nT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Io</td>
<td>5.9</td>
<td>1.77</td>
<td>1821</td>
<td>0.25–0.5</td>
<td>3550</td>
<td>&lt;200</td>
<td>−1900</td>
</tr>
<tr>
<td>Europa</td>
<td>9.4</td>
<td>3.55</td>
<td>1570</td>
<td></td>
<td>2940</td>
<td>0 or small</td>
<td>−420</td>
</tr>
<tr>
<td>Ganymede</td>
<td>15</td>
<td>7.15</td>
<td>2631</td>
<td>0.25–0.5</td>
<td>1936</td>
<td>750</td>
<td>−90</td>
</tr>
<tr>
<td>Callisto</td>
<td>26</td>
<td>16.7</td>
<td>2400</td>
<td></td>
<td>1850</td>
<td>0 or small</td>
<td>−30</td>
</tr>
<tr>
<td>Enceladus</td>
<td>4</td>
<td>1.37</td>
<td>252</td>
<td></td>
<td>1609</td>
<td>0</td>
<td>−325</td>
</tr>
<tr>
<td>Titan</td>
<td>20</td>
<td>15.9</td>
<td>2575</td>
<td></td>
<td>1900</td>
<td>0 or small</td>
<td>−5.1</td>
</tr>
</tbody>
</table>

$^1$Jupiter’s rotation period is 9 h 55 min, so corotating plasma moves faster than any of the moons.
$^2$Core densities are assumed in the range from 5150 to 8000 kg/m$^3$. This corresponds to maximum and minimum core radii, respectively.
$^3$The magnetic field of Jupiter at the orbits of the moons oscillates in both magnitude and direction at Jupiter’s rotation period of 9 h 55 min. The average field over a planetary rotation period is southward oriented (i.e. antiparallel to Jupiter’s axis of rotation). Neither the orbits nor the spin axes of the moons are significantly inclined to Jupiter’s equatorial plane, so we use averages around the moon’s orbit from the model of Khurana (1997).
Titan are similar in size to Mercury; Io and Europa are closer in size to Earth’s Moon; Enceladus is tiny (mean radius 252 km) but its interaction region is greatly expanded by a cloud of vapor that it introduces into the local magnetosphere. This means that Enceladus as well as Io is a major source of the plasma in which it is embedded. A plume at Enceladus injects $\sim 10^{28}$ water molecules/s or 0.3 tons/s into Saturn’s inner magnetosphere. Dust, too, is ejected in the plume and has been found to modify plasma dynamics in critical ways. Approximately 1 ton per second of ions is introduced into Jupiter’s magnetosphere by the source at Io, thus creating the Io plasma torus alluded to in Section 4. The other moons, particularly Europa and Titan, are weaker plasma sources.

The magnetospheric plasma sweeps by the moons in the direction of their orbital motion because the Keplerian orbital speeds are slow compared with the speed of local plasma flow. Plasma interaction regions develop around the moons, with details depending on the properties of the moon. Only Ganymede, which has a significant internal magnetic moment, produces a true moon magnetosphere.

The interaction regions at the moons differ in form from the model planetary magnetosphere illustrated in Figure 7.1. An important difference is that no bow shock forms upstream of the moon. This difference can be understood by recognizing that the speed of plasma flow relative to the moons is smaller than either the sound speed or the Alfvén speed, so that instead of experiencing a sudden decrease of flow speed across a shock surface, the plasma flow can be gradually deflected by distributed pressure perturbations upstream of a moon. The ratio of the thermal pressure to the magnetic pressure is typically small in the surrounding plasma, and this minimizes the changes of field geometry associated with the interaction. Except for Ganymede, the magnitude of the magnetic field changes only very near the moon. Near each of the unmagnetized moons the magnetic field rotates because the plasma tied to the external field slows near the body but continues to flow at its unperturbed speed both above and below, producing a draped magnetic field. The effect is that expected if the field lines were “plucked” by the moon. The regions containing rotated field lines are referred to as Alfvén wings. Within the Alfvén wings, the field connects to the moon and its surrounding ionosphere. Plasma on these flux tubes is greatly affected by the presence of the moon. Energetic particles may be depleted as a result of direct absorption, but low-energy plasma densities may increase locally because the moon’s atmosphere serves as a plasma source. In many cases, strong plasma waves, a signature of anisotropic or non-Maxwellian particle distributions, are observed near the moons.

In the immediate vicinity of Io, both the magnetic field and the plasma properties are substantially different from those in the surrounding torus because currents associated with ionization of pickup ions greatly affect the plasma properties in Io’s immediate vicinity. When large perturbations were first observed near Io it seemed possible that they were signatures of an internal magnetic field, but multiple passes established that the signatures near Io can be interpreted purely in terms of currents flowing in the plasma.

Near Titan, the presence of an extremely dense atmosphere and ionosphere also results in a particularly strong interaction whose effects on the field and the flow were observed initially by Voyager 1 and have been extensively explored by the Cassini orbiter. Saturn’s magnetospheric field drapes around the Titan’s ionosphere much as the solar wind field drapes to produce the magnetosphere of Venus, a body that like Titan has an exceptionally dense atmosphere.

It was field draping that provided the first hint of the presence of an ion source at Enceladus, deep within the magnetosphere at 4 $R_S$. The anomalous bending of the magnetic field alerted investigators to the likelihood that high-density ionized matter was present below the south pole of the moon. Cassini’s trajectory was modified on some orbits to enable imaging instruments to survey the region. A plume of vapor, largely water, was observed to rise far above the surface. This geyser is a major source of Saturn’s magnetospheric plasma and is the reason that Enceladus plays a role at Saturn much like that of Io at Jupiter.

One of the great surprises of the Galileo mission was the discovery that Ganymede’s internal magnetic field not only exists but is strong enough to stand off the flowing plasma of Jupiter’s magnetosphere and to carve out a bubble-like magnetospheric cavity around the moon. A schematic of the cross-section of the magnetosphere in the plane of the background field and the upstream flow is illustrated in Figure 7.15. Near Ganymede, both the magnetic field and the plasma properties depart dramatically from their values in the surroundings. A true magnetosphere forms with a distinct magnetopause separating the flowing jovian plasma from the relatively stagnant plasma tied to the moon. Within the magnetosphere, there are two types of field lines. Those from low latitudes have both ends linked to Ganymede and are called closed field lines. Little plasma from sources external to the magnetosphere is present on those field lines. The field lines in the polar regions are linked at one end to Jupiter. The latter are the equivalent of field lines linked to the solar wind in Earth’s magnetosphere and are referred to as open field lines. On the open field lines, the external plasma and energetic charged particles have direct access to the
interior of the magnetosphere. The particle distributions measured in the polar regions are extremely anisotropic because the moon absorbs a large fraction of the flux directed toward its surface. Where the energetic particles hit the surface, they change the reflectance of the ice, so the regions of open field lines can be identified in images of Ganymede's surface and compared with the regions inferred from magnetic field models. The two approaches are in good agreement. As expected, the angular distribution of the reflected particles has also been found to be modified by Ganymede's internal dipole field.

Ganymede’s dipole moment is roughly antiparallel to Jupiter’s, implying that the field direction reverses across the near equatorial magnetopause. This means that magnetic reconnection is favored. There is some evidence that such reconnection occurs intermittently. Should future missions allow a systematic study of this system, it will be of interest to establish directly how the properties of the upstream plasma control the reconnection process and whether, with steady upstream conditions, it occurs relatively steadily or whether it occurs with some periodic or aperiodic modulation.

7. CONCLUSIONS

We have described interactions between flowing plasmas and diverse bodies of the solar system. The interaction regions all manifest some of the properties of magnetospheres. Among magnetospheres of magnetized planets, one can distinguish (a) the large, symmetric, and rotation-dominated magnetospheres of Jupiter and Saturn; (b) the small magnetosphere of Mercury where the solar wind drives rapid circulation of material through the magnetosphere (see Mercury); and (c) the moderate-sized and highly asymmetric magnetospheres of Uranus and Neptune, whose constantly changing configuration does not allow substantial densities of plasma to build up. The Earth’s magnetosphere is an interesting hybrid of the first two types, with a dense corotating plasmasphere close to the planet and tenuous plasma, circulated by the solar wind driven convection, in the outer region. All of these magnetospheres set up bow shocks in the solar wind. The nature of the interaction of the solar wind with nonmagnetized objects depends on the presence of an atmosphere that becomes electrically conducting when ionized. Venus and Mars have tightly bound atmospheres so that the region of interaction with the solar wind is close to the planet on the sunward-facing side, with the IMF draped...
back behind the planet to form a magnetotail. Bow shocks form in front of both these magnetospheres. The regions on the surface of Mars where strong magnetization is present produce mini-magnetospheres whose properties are being explored. Comets cause the solar wind field to drape much as at Venus and Mars; they produce clouds extended over millions of kilometers. The interaction of the solar wind with the cometary neutrals weakens or eliminates a bow shock. Small bodies like asteroids disturb the solar wind without setting up shocks. Within the magnetospheres of Saturn and Jupiter, the large moons interact with the subsonic magnetospheric flow, producing unique signatures of interaction with fields that resist draping. No shocks have been observed in these cases.

The complex role of plasmas trapped in the magnetosphere of a planetary body must be understood as we attempt to improve our knowledge of the planet’s internal structure, and this means that the study of magnetospheres links closely to the study of intrinsic properties of planetary systems. Although our understanding of the dynamo process is still rather limited, the presence of a planetary magnetic field has become a useful indicator of properties of a planet’s interior. As dynamo theory advances, extensive data on the magnetic field may provide a powerful tool from which to learn about the interiors of planets and large satellites. For example, physical and chemical models of interiors need to explain why Ganymede has a magnetic field while its neighbor of similar size, Callisto, does not, and why Uranus and Neptune’s magnetic fields are highly nondipolar and tilted while Jupiter’s and Saturn’s fields are nearly dipolar and aligned.

Continued exploration of the plasma and fields in the vicinity of planets and moons is needed to reveal features of the interactions that we do not yet understand. We do not know how effective reconnection is in the presence of the strong planetary fields in which the large moons of Jupiter are embedded. We have not learned all we need to know about moons as sources of new ions in the flow. We need many more passes to define the magnetic fields and plasma distributions of some of the planets and all of the moons because single passes do not provide constraints sufficient to determine more than the lowest order properties of the internal fields. Temporal variability of magnetospheres over a wide range of timescales makes them inherently difficult to measure, especially with a single spacecraft. Spurred by the desire to understand how the solar wind controls geomagnetic activity, space scientists combine data from multiple spacecraft and from ground-based instruments to make simultaneous measurements of different aspects of the Earth’s magnetosphere or turn to multiple spacecraft missions like Cluster and Themis and the much anticipated Magnetospheric Multiscale Mission. As it orbited Jupiter, the Galileo spacecraft mapped out different parts of the jovian magnetosphere, monitoring changes and measuring the interactions of magnetospheric plasma with the Galilean satellites. Cassini in orbit around Saturn for many years provides ever more complete coverage of the properties of another magnetosphere and its interaction with Enceladus, Titan, and the other moons. The properties of the magnetic and plasma environment of Mars are still being clarified by spacecraft measurements. Messenger at Mercury will fully characterize the mysterious magnetic field of this planet and its magnetosphere. And finally, Pluto beckons as the prototype of an important new group of solar system bodies that the New Horizons spacecraft with explore in 2016. As new technologies lead to small lightweight instruments, we look forward to missions that will determine which of the many small bodies of the solar system have magnetic fields and help us understand the complexities of magnetospheres large and small throughout the solar system.

BIBLIOGRAPHY


INTRODUCTION

In this part of the book we describe the rotation of planets and moons of the solar system. Planetary rotation can be divided into the rotation speed around an axis and the orientation of this axis (or another axis of the planet) in space. Here, we summarize the main observed rotational characteristics of the planets of the solar system and explain how the rotation might have evolved during the history of the solar system. On short timescales, planetary rotation is variable and yields information on the interior structure of planets. Most of us know that the rotation of a boiled egg noticeably differs from that of a raw egg. This simple observation shows that information on the inside of an object, here an egg, can be obtained from its rotation. The same idea applies to the rotation of celestial bodies.

The chapter is organized in 10 sections: first we describe the observed rotation states of the planets (Section 1); next we discuss the origin of the spin and its long-term evolution (see Section 2), as well as the long-term evolution of the orientation (Section 3); and Section 4 explains the flattening of planets induced by the rotation. The next five sections (Sections 5, 6, 7, 8, and 9) are dedicated to a description and study of the variable components of rotational motion of planets (precession, nutation, length-of-day (LOD) variations, libration, and polar motion) with a particular emphasis on terrestrial planets. These sections also provide insight into how information on the interior and dynamics of a planet can be obtained from observations of the rotation. Methods for observing the rotation of planets are described in the last section (Section 10); the particular case of the Earth is also treated there.

1. OBSERVED ROTATION STATE OF PLANETS

As Kepler (1571—1630) has shown, all planets of the solar system move on approximately elliptical orbits. The dimensions of these orbits are characterized by the semimajor axis $a$ and the semiminor axis $b$ (see Figure 8.1). The elliptical form of the orbit is uniquely determined by the distance between the center of the ellipse and one of its foci, also referred to as the parameter of the ellipse and denoted by $f$. Alternatively, it is described by the eccentricity $e$, which is defined as the ratio of the distance between the two foci to the length of the major axis: $e = 2f/2a = f/a$. The
The eccentricity of an ellipse is between 0 and 1. When the eccentricity is 0, the foci coincide with the center point and the figure is a circle. The Sun for planets or the central planet for satellites occupies one of the focal points of the ellipse. Since the planets and most satellites of the solar system revolve in an almost circular motion around the Sun or their parent planet, their orbital eccentricity is generally much closer to 0 than to 1. An eccentricity of 1 characterizes a parabola, and \( e > 1 \), a hyperbola. Some comets are on parabolic and hyperbolic orbits (See Comet Populations and Cometary Dynamics).

Besides the orbital motion, the planets and satellites (or moons; we will use those two words designating the same objects of our solar system interchangeably) also rotate around an internal axis. This rotation axis or spin axis differs from the perpendicular to the orbital plane by an angle called the obliquity. The obliquities, rotation periods, and revolution periods are provided in Table 8.1 for the planets of our solar system. The rotation period given is the sidereal rotation period, which is defined as the period between two passes of a given point at the surface of the planet to the same direction in space. It must be mentioned that the rotation of the surface layers of the giant gaseous planets, which can be obtained from direct visual observations, does not necessarily represent the internal rotation rate. It is usually accepted that the rotation axis orientation can be obtained from the magnetic dipole axis and that the rotation speed can be obtained from the rotation of the magnetic field (magnetically linked to the deep interior) or from natural radio emissions varying periodically with the rotation. For a planet such as Venus with a thick atmosphere (or a moon such as Titan) surface details can hardly be seen in the visible and the rotation is determined from radar observations of the surface and from measurements of the gravity field rotating with the planet.

The planets rotate about their axis in about 1 day or less, except for the two planets closest to the Sun, Mercury and Venus (see Table 8.1). Both these planets and Jupiter rotate about an axis almost perpendicular to their orbital plane, in contrast to most other planets. These observations suggest that the innermost planets have significantly changed their rotation in the about 4.6 billion years since the origin of the solar system, as will be further explained in the next sections (see Section 2 and Section 3). The near alignment of Jupiter’s rotation axis with that of the Sun is primordial and a not unexpected result of conservation of angular momentum during the formation of this by far the most massive planet of the solar system. Surprisingly, half of the planets have an obliquity close to 25°, including the Earth. Uranus is peculiar in the sense that its rotation axis is almost in the orbital plane. Most probably, Uranus did not retain its original spin, but has been tilted onto its side by a large impact event. Equally unique is that Venus and Uranus have a retrograde rotation instead of a counterclockwise rotation as seen from above the Sun’s North Pole as for all other planets.

### 2. ORIGIN AND LONG-TERM SPIN EVOLUTION

#### 2.1. Origin

The solar system planets acquired rotational angular momentum during the formation process. As a consequence of conservation of angular momentum, the planets are expected to form with the same sense of rotation as their orbital motion. Gas and ice giant planets mainly exchange angular momentum with the gas disk, whereas terrestrial planets can achieve fast rotation speeds by collisions with planetesimals. The spin state after formation is expected to be mainly
determined by the random giant impacts that terrestrial planets encountered at the final stages of planet formation.

Not all current spin characteristics summarized in Table 8.1 correspond to the initial values after formation since both the obliquities and the rotation rates changed on long timescales. As already mentioned, for Mercury and Venus, in particular, even important changes probably occurred. The long-term changes in the rotation of planets and satellites are mainly due to friction associated with tides raised by the central body (the Sun for planets or the parent planet for satellites). Tidal dissipation also explains why most large natural satellites rotate with a period equal to their orbital period.

2.2. Tidal Dissipation

2.2.1. Tidal Torque

Planets are periodically deformed by the gravitational attraction of the Sun and their natural satellites. These tides are never perfectly elastic responses to the tidal forcing and lead to the dissipation of a small amount of energy in the planets. This input of energy to the planets is so small that it can generally be neglected for the thermal evolution and internal dynamics of the planets of the solar system (although it can be important for exoplanets near their mother star). Nevertheless, it can drastically change the orbital and rotational dynamics of planets. For satellites, the effect of tidal dissipation (the tides are mainly raised by the parent planet) is more pronounced and can also be important for internal processes. For example, Io’s volcanism is due to dissipation of energy associated with the large tides raised by Jupiter on Io (See Io: The Volcanic Moon).

Deviations from perfectly elastic behavior of the materials inside planets delay the response of the planet to an applied tidal force. The tidal bulge therefore does not form instantaneously and, because of the rotation of the planet relative to the Sun, it becomes slightly misaligned with respect to the direction to the body raising the tides (see Figure 8.3). As a result of the asymmetrical orientation of the tidal bulge with respect to the Sun, the Sun exerts a gravitational torque on the planet, which tends to alter the rotation of the planet (see Figure 8.3). The magnitude of the torque can be expressed as \( T = \frac{3}{8} k_P M_S \left( \frac{R_P}{d} \right)^6 \sin 2\varepsilon \), where \( G \) is the universal gravitational constant, \( M_S \) is the mass of the Sun, \( R_P \) is the radius of the planet, \( d \) is the distance between the planet and the Sun, \( \varepsilon \) is the angle between the long symmetry axis of the tidal bulge and the direction to the Sun, and \( k_P \) is the tidal Love number of the planet specifying the response of the planet to a unit tidal potential. In general, this torque is small because of the large distances in the solar system. It rapidly decreases with increasing distance \( d \) between the planet and the Sun according to \( d^{-6} \). As it is a gravitational torque that is acting on a tidal bulge, which itself is created by tidal forces, the torque is also proportional to the square of the mass of the Sun (see Figure 8.2). A gravitational torque can be exerted by any body on a planet, such as the Moon for the Earth (see Figure 8.3). The torque can be large for small distance \( d \) between the planet and the perturbing body and for large mass of the tide-raising body. This mass dependence also explains why Io, which is at about the same distance from Jupiter as the moon is from the Earth, is geologically very active with volcanism on a much larger scale than on Earth, whereas the Earth’s moon is geologically quiet (See The Moon).

2.2.2. Long-Term Spin Evolution

The tides slow down the rotation of a planet or a satellite when its rotation period is shorter than the orbital period. In that case, the tidal bulge closest to the perturbing body will be behind the direction to the tide-raising body (e.g. the Sun...
in the case of Mercury, the Earth in the case of the Moon, and Jupiter in the case of Io) with respect to the direction of orbital motion (see Figures 8.2 and 8.3). The gravitational tidal torque from the central body then acts to slow down the rotation. For a circular orbit \((d = a)\), the despinning timescale \(\tau\) can be obtained directly from the change in the angular momentum of the planet as

\[
\tau = \frac{2}{3} Q_p \frac{Q_r}{k_p GM_p} C_p \left(\frac{a}{R_p}\right)^6.
\]

Here \(Q_p\) is the planet’s or satellites’ rotation frequency, \(C_p\) is its polar moment of inertia, \(R_p\) is its radius, and \(M_p\) is the mass of the primary. \(Q_{fr}\), the quality factor of tidal dissipation, is defined as the inverse of the sine of the lag angle. It is of the order of 100 for terrestrial planets but can be as small as 10 if the interior is partially molten, as is likely the case for Io. The quality factor is several orders of magnitude larger for the giant planets since gases are much less dissipative than liquids or solids. On Mercury, the closest planet, the Sun has the largest tidal effect of all planets in the solar system. Tides can despin Mercury from an initially faster spin with a period of about a day to rotation periods commensurate with the orbital period on a timescale of a few 100 million years, much shorter than the age of the solar system (see Figure 8.2). Although the initial spin of Mercury is unknown, this shows that Mercury must necessarily be spinning slowly. Besides Mercury, tidal friction by the Sun could also significantly despin Venus over the age of the solar system. It also affects the rotation of the Earth, but has only a very small influence on the rotation of the more distant Mars and the gas and ice giants.

In contrast to the decelerating effect of the tidal torque for rapidly rotating planets, the rotation is accelerated when the orbital period is shorter than the rotation period. Taken together, this suggests that planets and satellites will ultimately evolve to a situation in which their rotation period is equal to the orbital period.

In contrast to planets, most large and medium-sized satellites are indeed observed to have rotation periods that are almost exactly equal to their revolution periods. This is the most common example of a spin—orbit resonance in the solar system. The spin and orbital motion are said to be resonant because the ratio of their periods is a ratio of integers. The resonance is referred to as the 1:1 spin—orbit resonance (see Figure 8.4) and is an end state of tidal friction processes. Once captured in such a state, the satellite always shows approximately the same face to the central planet (see Figure 8.4).

The despinning timescale for most satellites is less than 1 million years and often even much shorter. The main tide-raising body here is the central planet. Although the planets are much less massive than the Sun (its mass is about 1000 times larger than that of the largest planet Jupiter), the distance between the planets and their satellites is also much shorter. Typical distances are of the order of 100 thousand kilometers, whereas distances to the Sun are on the order of 100 million kilometers and more. Because of the strong dependence on distance, the planets can have a larger tidal effect on the rotation of their satellites than the Sun has on planets.

In a circular orbit, tidal friction will ultimately drive the rotation of the planet to a state synchronous with the orbital motion (we neglect obliquity here). The final equilibrium rotation rate for a satellite in an eccentric orbit is slightly faster than synchronous due to the larger accelerating effect on rotation near pericenter than the decelerating effect near apocenter. Rotation accelerates near pericenter because the orbital motion there is faster than the rotational motion and the tidal bulge will be ahead of the satellite—planet line. Near apocenter the orbital motion is slower than the rotational motion resulting in a gravitational torque tending to slow down the rotation. The synchronous rotation state is, nevertheless, stable for a slightly eccentric orbit if the gravitational torque of the central planet on a permanent asymmetry of the satellite is larger than the tidal torque at synchronous rotation averaged over an orbital period. For the largest satellites of the solar system for which the ellipsoidal shape has been measured, this condition is indeed fulfilled.

Besides the Sun, the Moon also contributes to the deceleration of the rotation of the Earth by tidal friction. Because the Moon is so much closer to the Earth than is the Sun, its decelerating effect is even about a factor 2 larger than that of the Sun.

Friction at the core—mantle boundary (CMB) associated with differential rotation between a liquid core and a solid mantle can also change the rotation of a planet on long timescales. The friction is associated with either the viscosity of the liquid core or with the electromagnetic interaction between mantle and core (See Earth: Surface And Interior). In the latter case, the rotational motions of...
the fluid core relative to the neighboring solid regions (mantle or inner core) at any of the tidal frequencies results in an oscillatory sweep of the magnetic field through the fluid, producing a Lorentz force which causes the fluid itself to be partially dragged with the field. Since the dissipated energy must come from the rotation, core—mantle friction torques cause the rotation to slow down.

In addition, a dense atmosphere can have an effect on the long-term evolution of planets. This is especially the case for Venus, for which the atmospheric pressure at the surface is 92 times that on Earth (See Venus Atmosphere). The atmosphere and surface of a planet like the Earth and Venus is heated mostly by solar radiation at the subsolar point. This causes pressure changes and atmosphere motions, or thermal tides, away from the subsolar point. This creates an atmospheric “bulge”, which is almost perpendicular to the planet—Sun direction, but not exactly because of the associated dissipation and a finite response time. For a rapidly rotating planet such as the Earth, the torque on the thermal atmospheric tides tends to accelerate the rotation but the effect is more than a factor 10 smaller than the decelerating effect on rotation from gravitational tides. For the slowly rotating planet Venus, however, the atmospheric tides have a large influence on its rotation. The rotation of Venus is thought to be the result of a balance between solid body tidal torques, which drive Venus to synchronous rotation, and atmospheric torques, which drive it away.

Tidal dissipation can occur everywhere in a planet or satellite, but some layers may have a dominant contribution. For Io, a partially molten asthenosphere may be important. On the Earth, dissipation induced by the tides in the oceans and in particular in the shallow seas plays the most important role. The presence of oceans therefore increases the change in the rotation period of a planet. In the Devonian, the LOD on Earth was about 22 h. In 100 years, it will take two more milliseconds for the Earth to complete a full rotation around its axis. At that rate, the day will be an hour longer after more than 100 million years.

Both Pluto and Charon have evolved to a double-synchronous state in which the orbital period and the rotation period of both bodies are equal. Both Pluto and Charon always show the same face to the other body.

2.2.3. Effects on the Orbit

Tidal interactions change not only the rotation of the planet or satellite but also the orbital motion. As long as the system of the planet and the tide-raising body can be considered isolated, the total angular momentum of the system must be conserved. Since the planet loses rotational angular momentum due to friction of tides in the planet, the angular momentum of orbital motion must increase. As a result the semimajor axis increases, making the orbit wider. This effect due to the Sun is very small for planets. Since the angular momentum associated with orbital motion is many orders of magnitude larger than that of rotational motion, tidal braking is much less efficient in changing the orbit of planets than in changing their rotational motion.

Because of the shorter distance between planets and satellites than between the Sun and planets, this process of orbital change is more important for planet—satellite systems. For example, because of dissipation of tidal energy in the Earth associated with tides raised by the Moon, the Moon recedes from the Earth and the Earth slows down its rotation. After one century, the semimajor axis of the Moon increases by 3–4 m (~3.5 cm/year).

The Earth also raises tides on the Moon and as a result, tidal friction in the Moon also changes the semimajor axis of the orbit. This change, however, is much smaller than that induced by tides in the Earth mainly because periodic tides in the Moon only exist for an eccentric orbit. For an exactly circular orbit, any material point in a satellite in a 1:1 resonance would exactly feel the same gravitational force from the central planet and there would be no periodic tides. Since the deviation of a circular orbit, measured by the eccentricity, is small, tidal dissipation in the Moon contributes less than 1% to the change in its semimajor axis. For satellites in the outer solar system orbiting a gas or ice giant, the situation can be different since tidal dissipation in a solid body is more efficient than in a gas. For example, tidal dissipation in Io, the closest of the Galilean satellites to Jupiter, contributes about as much to the change in orbital motion as dissipation in Jupiter.

Besides changing the semimajor axis, tidal friction also changes the orbital eccentricity. Since despining is a faster process than circularization because of the smaller angular momentum involved in rotation than in orbital motion, satellites still have an elliptical orbit upon reaching an equilibrium rotation state. Nevertheless, their eccentricity usually is very small. Mercury, on the other hand, has a large eccentricity of 0.2059. With such a large eccentricity, the equilibrium rotation rate, at which the mean tidal torque averaged over one orbit is equal to zero, is about 1.26 times the mean motion. This, however, is not the rotation state occupied by Mercury. Observations show that Mercury is in a 3:2 spin—orbit resonance, meaning that its orbital period is 1.5 times longer than the rotation period (see Figure 8.5 and also the beginning of Movie 6). As a result, a solar day on Mercury—the time between two consecutive passes of the Sun through the local meridian—is twice as long as a year, a quite remarkable situation compared to our situation on Earth.

The 3:2 spin—orbit resonance is dynamically stable for the shape of Mercury and the current eccentricity, and Mercury has probably been captured into this 3:2 resonant state while spinning down from an earlier faster rotation. Because of large oscillations in the eccentricity of Mercury driven by other planets, the resonance may have been
crossed many times before final capture. This increases the likelihood of capture in the 3:2 spin–orbit resonance. Capture into the \(3:2\) spin–orbit resonance is the most probable capture but Mercury might also have been captured before in another resonance, which subsequently, probably during the period of Late Heavy Bombardment, was destabilized by large impacts. If Mercury initially rotated retrogradely, an initial capture into the 1:1 resonance is even the most likely. The observed asymmetrical distribution of large impact basins may be an indication for such a different primordial resonant state.

3. LONG-TERM EVOLUTION OF THE ORIENTATION

Since dissipation associated with tidal deformation delays the appearance of the tidal bulge with respect to a perfectly elastic body that reacts instantaneously, the tidal bulge is carried outside the orbital plane by rotation during the time lag for obliquities different from \(0^\circ\) to \(180^\circ\). The gravitational torque of the central body on the tidal bulge therefore also has a component in the orbital plane. As a result, besides changing the rotation rate of a planet, the tidal torque can also change the obliquity. The effect can be important for those planets and satellites that also changed their rotation rate as a result of tidal dissipation. For those slowly rotating bodies, meaning that their rotation period is not too different from their orbital period, the tidal torque tends to drive the obliquity to zero in the case of zero eccentricity since the obliquity is the cause of torque. Mercury, Venus, and most satellites are indeed observed to have a small obliquity.

For Mercury and satellites in spin–orbit resonance, the tidal evolution has also led to a stable equilibrium state for the orientation of the body, called the Cassini state. In this state, the rotation axis and the orbit normal remain coplanar with the normal to the Laplace plane, while the rotation axis and the orbit normal rotate (or precess) about the normal to the Laplace plane with the same period. The Laplace plane is defined as the plane about which the orbit rotates (or precesses) and is equal to the mean orbital plane (see Figure 8.6). The obliquity is then constant but nonzero as tides try to bring the spin axis to the orbit normal which is itself moving on long timescales.

Radar observations have shown that Mercury occupies the Cassini state within the observational error. Mercury’s obliquity is equal to \(2.04 \pm 0.08\) arcmin. The observed spin pole position differs by \(2.7\) arcsec from the predicted Cassini position, within the \(5\)-arcsec uncertainty on the orientation. This observation is very useful for constraining the interior structure of Mercury because it allows the polar moment of inertia of Mercury to be calculated. The latter is a measure of the internal mass distribution in the radial direction (See Mercury). The obliquity in the Cassini state is

![Figure 8.5](image-url) Rotation and revolution of Mercury around the Sun in a spin–orbit 3:2 resonance.

![Figure 8.6](image-url) Spin axis (s), normal to the orbital plane (orb), both being very close to one another, and normal to the Laplace plane (k) with numerical values for Mercury’s case.
theoretically connected to the polar moment of inertia by a relation which also involves the degree-two gravitational coefficients $J_2$ and $C_{22}$, and some orbital parameters. For the recently determined values from MErcury Surface, Space ENVironment, Geochemistry and Ranging (MESSENGER) radio tracking ($J_2 = (5.03 \pm 0.02) \times 10^{-5}$ and $C_{22} = (0.809 \pm 0.006) \times 10^{-5}$), we have $C = (0.346 \pm 0.014) M_p R_p^2$, where $M_p$ is the mass of the planet Mercury and $R_p$ is its radius. A further rotational constraint and joint inferences on the core of Mercury are described in Section 8.

For most satellites the obliquity has not yet been accurately determined but is generally small. Based on radar images obtained from the Cassini spacecraft, the obliquity of Titan is estimated to be $0.32 \pm 0.02\degree$. The observations indicate that Titan is close to but does not exactly occupy the Cassini state since the rotation axis deviates by $0.12 \pm 0.02\degree$ from the plane formed by the orbit normal and the normal to the Laplace plane. The obliquity of Titan is a few times larger than expected for a rigid Titan in the Cassini state, suggesting that either Titan has a subsurface ocean or that the obliquity is excited by other causes like its atmosphere.

Since obliquity is defined as the angle between the rotation axis and the normal to the orbital plane, changes in obliquity can have two different causes. Obliquity can change not only due to a change in the orientation of the planet (or spin axis) in inertial space but also due to variations in the orientation of the orbital plane, called orbital precession. The total motion of the spin axis with respect to the orbit normal depends critically on the ratio of the orbital precession period to the period of the precession of the planet. Precession is the slow change in the orientation of a planet due to gravitational torques exerted by other solar system bodies in which the tip of the rotation axis describes a large circle on the celestial sphere around the normal to the orbital plane. The physics of precession will be explained in Section 5. For precession much faster than the precession of the orbit, the spin axis precesses about the slowly moving orbit normal and tracks that motion effectively such that the obliquity remains nearly constant. On the other hand, if the precession of the spin axis is much slower than changes in the orientation of the orbit, the spin axis will precess about the averaged orbit normal. If the ratio is about one, a resonant situation occurs and the obliquity variations can be large and chaotic. Mars is in such a situation and shows large changes in obliquity over a relatively short period of time. For example, over the past million years the obliquity of Mars fluctuated between about 15\degree and 35\degree. The other three terrestrial planets could also have experienced large chaotic variations in obliquity during some period of their history. Therefore, the obliquities of the terrestrial planets cannot be considered as primordial.

The presence of the Moon increases the Earth’s precession rate, or decreases the precession period, by about a factor 3 with respect to an Earth without a Moon. As a result the precession period of the Earth is much shorter than the orbital precession. Therefore the obliquity of the Earth nowadays remains nearly constant. It is $23.44\degree$ at present and oscillates between $22.1$ and $24.5\degree$ with a period of 41,000 years. These changes in the orientation of the Earth create changes in the insolation from the Sun on the Earth’s surface, producing climate change. Milanković (1879–1958) first put forward the idea that variations in eccentricity, axial tilt, and precession of the Earth’s orbit determine climate changes of the Earth through orbital forcing. Such climate changes are particularly important for Mars because of its large obliquity variations. Note that Mars has no large moon (the moons Phobos and Deimos are rocky bodies of only about 20 km and 10–15 km size) to increase the precession rate and Mars’ precession is also slower than that of the Earth since it is further from the Sun (see Section 5). Its period is 171,000 years, close to some planetary perturbations on the orbit.

Besides the gravitational torque on the gravitationally forced tides, the orientation of a planet can also evolve as a result of a gravitational torque on the thermal tides in the atmosphere and of core—mantle friction torques. For the slowly rotating Venus, core—mantle friction torque is thought to dominate the obliquity evolution and to drive Venus to an orientation with its polar axis perpendicular to its orbital plane. If Venus initially had a rapid prograde rotation like Mars and the Earth, it may have slowed down so much that it started developing a retrograde rotation. In an alternative scenario, however, the rotation axis of Venus has flipped direction under the influence of strong atmospheric torques during the rapid rotation phase and slowed down later. Due to planetary perturbations, the obliquity never reaches a zero value.

4. ROTATIONAL FLATTENING OF PLANETS

Let us consider a spherical deformable planet rotating with angular velocity $\Omega$. A point $P$ on the surface ($r = a_0$ where $a_0$ is the mean radius of the planet) or inside the planet at a distance $r$ from the center of the planet and at a colatitude $\theta$ experiences a centrifugal acceleration of magnitude:

$$a_{cf} = \Omega^2 r \sin \theta.$$  

The centrifugal acceleration can be written as the gradient of the centrifugal potential $V_{cf}$ as:

$$\overrightarrow{a_{cf}} = -\nabla V_{cf} \quad \text{with} \quad V_{cf} = \frac{1}{2} \Omega^2 r^2 \sin^2 \theta.$$  

The centrifugal acceleration is largest at the equator and decreases to zero at the poles. Since the centrifugal
acceleration is perpendicular to the rotation axis and oriented outward, rotation will flatten planets and expand them in the equatorial regions into the shape of an oblate spheroid.

The gravitational acceleration resulting from self-gravitation also derives from a potential. The gravitational potential at the point \( P \) is proportional to the mass \( M \) of the planet and inversely proportional to the distance to the center. As we consider the planet to be of an oblate spheroidal form, the potential also depends on latitude. For positions exterior to the planet we have

\[
V = \frac{GM}{r} \left( 1 + J_2 \left( \frac{r_{\text{eq}}}{r} \right)^2 \frac{3 \cos^2 \theta - 1}{2} \right) + \frac{1}{2} \Omega^2 r^2 \sin^2 \theta,
\]

where \( J_2 = \frac{C_{\text{eff}}}{M r_{\text{eq}}^2} \) is the form factor of the planet with \( \bar{A} = \frac{A + B}{2}, A \) and \( B \) being the equatorial moments of inertia, \( C \) being the polar moment of inertia (\( A < B < C \)), and \( r_{\text{eq}} \) being the equatorial radius of the planet; for an oblate spheroidal planet we have \( \bar{A} = \frac{A + B}{2} = A = B \).

The total gravity at the surface of the planet can be written:

\[
\vec{g} = -\nabla V.
\]

A planet is said to be in hydrostatic equilibrium when gravity is balanced by a differential pressure (\( \rho \)) force at each position inside the planet with density \( \rho \):

\[
\vec{\nabla} \rho = -\vec{g}.
\]

The centrifugal potential then forces the constant gravity surfaces or constant potential surfaces to be ellipsoidal. The constant potential surfaces coincide with surfaces of constant pressure and constant density. By using the equation for the total potential and expressing the radial coordinate of the surface as \( r = a_0 (1 - \frac{2}{3} \alpha P_2 (\cos \theta)) \), the flattening \( \alpha \) of the equipotential surface of the planet is seen to be given by \( \alpha = \frac{2}{3} J_2 + \frac{q}{2} \) where \( q = \frac{\Omega r_{\text{eq}}^2}{GM} \) is the ratio of the centrifugal acceleration to the gravity at the equator of radius \( r_{\text{eq}} \). Since \( J_2 \), too, is nonzero due to the rotation, the flattening is proportional to the square of the rotation frequency, as is the centrifugal acceleration. Therefore, fast rotating planets show a more distinct equatorial bulge than the slower rotating planets Mercury and Venus.

For bodies rotating synchronously with their orbital motion like most of the large natural satellites including the Moon, the tidal forces also have an important static component, which further deforms the satellites in addition to the centrifugal acceleration. The static tidal bulge also takes the form of an oblate ellipsoid but with the long axis in the direction to the central planet at pericenter. As a result, synchronous satellites have the shape of a triaxial ellipsoid with three different principal moments of inertia \( A, B, \) and \( C \). These principal moments of inertia are computed from the mass repartition inside the planet:

\[
\begin{align*}
A &= \int \rho(\vec{r})(x^2 + z^2) \, d^3r, \\
B &= \int \rho(\vec{r})(x^2 + z^2) \, d^3r, \\
C &= \int \rho(\vec{r})(x^2 + y^2) \, d^3r.
\end{align*}
\]

When the coordinate axes are chosen to coincide with the principal axes of inertia, the external gravitational potential for a triaxial planet can be expressed as

\[
\frac{GM}{r} \left( 1 - J_2 \left( \frac{r_{\text{eq}}}{r} \right)^2 \frac{3 \cos^2 \theta - 1}{2} \right) + C_{22} \left( \frac{r_{\text{eq}}}{r} \right)^2 (3 \sin^2 \theta \cos 2\lambda),
\]

where \( \theta \) and \( \lambda \) are the colatitude and longitude and \( C_{22} \) is the degree-two sectorial coefficient of the gravitational potential and is related to the equatorial moment of inertia by \( C_{22} = \frac{\theta - A}{4Mr_{\text{eq}}^2} \).

5. PRECESSION

For a planet whose rotation axis is tilted with respect to the orbital plane (Table 8.1), the equatorial bulge is out of the equatorial plane during the orbital motion. As a result, the Sun exerts a gravitational torque on the planet tending to twist the equator toward the orbital plane of the planet. All other objects in the solar system in principle also exert a torque on the planet, but apart from torques of large and nearby satellites (such as the Moon for the Earth) these torques are several orders of magnitude smaller. As the planet is rotating, it acquires additional angular momentum in the direction of the torque and reacts as a spinning top. The main effect is precession, which is the slow motion of the rotation axis and the planet in space around the perpendicular to the orbital plane (see Figure 8.7).
Movie 1 shows first the motion of a top. It then shows the long-term variation of the orientation of the Earth in space, the precession.

Supplementary video related to this chapter can be found at http://dx.doi.org/10.1016/B978-0-12-415845-0.00008-6.

The following is/are the supplementary data related to this chapter: Movie 1: Precession of the Earth in space. (file EarthPrecession.mpg)

The gravitational torque exerted by the Sun on a planet can be computed from the relative position of the celestial bodies. From the point of view of the center of mass of the planet, the Sun describes an apparent motion in the orbital plane with angular revolution velocity \( n \), commonly referred to as the mean motion, around the planet. According to Kepler’s third law, we have

\[
    n^2 = \frac{G(M_{\text{Sun}} + M_{\text{Planet}})}{a^3} \equiv \frac{GM_{\text{Sun}}}{d^3}
\]

where \( M_{\text{Sun}} \) and \( M_{\text{Planet}} \) are the masses of the Sun and the planet, respectively. In order to describe the position of the Sun, we define a reference frame \((\hat{X}, \hat{Y}, \hat{Z})\) with origin in the mass center of the planet and three orthogonal axes.

Two axes \((\hat{X}, \hat{Y})\) are chosen in the orbital plane, and the third axis \(\hat{Z}\) is perpendicular to the orbit. The position of the Sun can then be expressed by the following three coordinates with respect to that frame (see Figure 8.8):

\[
    \xi = d \cos (nt) \\
    \eta = d \sin (nt) \\
    \zeta = 0
\]

Here \( d \) is the distance between the centers of mass of the planet and the Sun at a given position in the elliptical orbit.

The X-axis of the frame \(\hat{X}\) is chosen in the direction of the ascending equinox, the intersection between the equator and the orbital plane where the Sun crosses from below to above the orbital plane. We also consider here that the Sun is at the equinox at time \( t = 0 \).

The gravitational acceleration of a material element at \( \vec{r} \) inside the planet due to the Sun is \(-\nabla W^{\text{Sun}}(\vec{r}, t)\), where \( W^{\text{Sun}} \) is the gravitational potential of the Sun. The force on the element occupying a volume \( dV \) at that position inside the planet is \(-\rho(\vec{r})\nabla W^{\text{Sun}}(\vec{r}, t)dV\), where \( \rho(\vec{r}) \) is the density of matter at \( \vec{r} \). In a reference frame at the planet’s mass center, this acceleration is the centripetal acceleration of the planet as a whole toward the Sun, which is responsible for the relative orbital motion of the planet around the Sun. We neglect deviations from spherical symmetry in the Sun and then have

\[
    W^{\text{Sun}}(\vec{r}, t) = -\frac{GM_{\text{Sun}}}{|\vec{d} - \vec{r}|} \\
    = -\frac{GM_{\text{Sun}}}{\sqrt{d^2 - 2(\vec{d} \cdot \vec{r}) + r^2}} \equiv -\frac{GM_{\text{Sun}}}{d} \\
    \times \left( 1 + \frac{d \cdot r}{2d^2} + \frac{3(d \cdot \vec{r})^2}{2d^4} + \ldots \right)
\]

where \( \vec{d} \) is the position vector of the Sun with respect to the planet and is time dependent because of the motion of the Sun relative to the planet frame, where \( |\vec{d}| = d \) and \( |\vec{r}| = r \).

In the reference frame \((\hat{x}, \hat{y}, \hat{z})\) tied to the planet, the torque \( \vec{T} \) exerted on the planet by the Sun may be calculated by expressing the force (and hence the torque) on a mass element in the planet in terms of the gravitational potential due to the Sun at the location of the mass element, making vectorial multiplication with \( \vec{r} \), and then integrating over the whole planet. The torque \( \vec{T} \) that it produces on an element of matter at \( \vec{r} \) is

\[
    \vec{T} = -\int \rho(\vec{r}) \vec{r} \times \nabla W^{\text{Planet}}(\vec{r}, t)dV \\
    = -\frac{GM_{\text{Sun}}}{d} \int \rho(\vec{r}) \\
    \times \left( \frac{\vec{r} \times \vec{d}}{d^2} + \frac{3}{2d^4} \left( \frac{\vec{d} \cdot \vec{r}}{2d^2} \right) \right) dV
\]
where the integration is over the volume of the planet. The three components of the torque integral acting on the rotating planet can then be expressed as
\[
\begin{pmatrix}
\Gamma_x \\
\Gamma_y \\
\Gamma_z
\end{pmatrix} = \frac{3GM_{\text{Sun}}}{d^5} \int \rho(\vec{r}) \left\{ \mathbf{d}^2 + 3(xd_x + yd_y + zd_z) \right\} \times \begin{pmatrix}
yd_z - zd_y \\
zd_x - xd_z \\
xd_y - yd_x
\end{pmatrix} \, dV
\]

The components \((d_x, d_y, d_z)\) are the coordinates of \(\mathbf{d}\) and \((x, y, z)\) are the coordinates of \(\mathbf{r}\) in the \((\hat{x}, \hat{y}, \hat{z})\) reference frame. We choose the axes \(x, y,\) and \(z\) to coincide with the principal axes of inertia of the static planet, which we refer to as the principal axis coordinate system ("static" means that variations in the density distribution due to tidal deformations and other causes are not considered). The first term of the above integral drops out due to the fact that the reference frame is tied to the center of mass and the equator of the planet. Since we have chosen a principal axis coordinate system, the matrix of inertia reduces to a diagonal matrix; therefore, \(\int \rho(\vec{r}) \mathbf{y} \mathbf{y} dV = \int \rho(\vec{r}) \mathbf{x} \mathbf{z} dV = \int \rho(\vec{r}) \mathbf{z} \mathbf{x} dV = 0\). By using the definitions of the principal moments of inertia, we then have
\[
\begin{pmatrix}
\Gamma_x \\
\Gamma_y \\
\Gamma_z
\end{pmatrix} = \frac{3GM_{\text{Sun}}}{d^5} \begin{pmatrix}
(C - B) & d_y & d_z \\
(A - C) & d_x & d_z \\
(B - A) & d_x & d_y
\end{pmatrix}
\]

For rapidly rotating planets, the difference between the two equatorial moments of inertia is a few orders of magnitude smaller than the difference between the polar moment of inertia and the mean equatorial moment of inertia. Therefore, the planet can be considered as axially symmetric with \(A \approx B\). In this case, \(\Gamma_z = 0\), and so the torque vector lies in the equatorial plane. It, moreover, is perpendicular to the equatorial projection \((d_x, d_y)\) of \(\mathbf{d}\), so it is perpendicular to the direction to the Sun.

The torque is maximal when the Sun is at the highest points above the equator (at summer and winter solstices). At both points, the torque tends to align the equator with the orbit plane. It is zero at the equinoxes. The torque will change the angular momentum, and thus the rotation of the planet according to extension of Newton’s third law to rotational motion is \(\frac{d\mathbf{H}}{dt} = \mathbf{T}\).

For a rigid body rotating around an axis of symmetry (say the polar axis \(\hat{z}\)), the angular momentum can be expressed as the product of the polar moment of inertia, \(C\), and its angular velocity \(\Omega_z : \mathbf{H} = C\Omega_z \hat{z}\), where \(\hat{z}\) is the unit vector in the \(z\) direction. Since the torque \(\mathbf{T}\) is in the equatorial plane and the angular momentum is perpendicular to the equatorial plane, we have
\[
\mathbf{H} \cdot \mathbf{T} = 0,
\]
and hence
\[
\mathbf{H} \cdot \frac{d\mathbf{H}}{dt} = 0 \quad \text{or} \quad \frac{d(\mathbf{H} \cdot \mathbf{H})}{dt} = 0.
\]

Therefore, the magnitude of angular momentum does not change with time. As a result, the torque can only change the orientation of the planet but not its rotation speed.

The average torque is oriented in the direction of the line joining the equinoxes \((\hat{X})\). For a circular orbit, it can easily be calculated from the above expression of the torque by using expressions for the position of the Sun with respect to the planet. Instead of considering a reference frame with rotating axes coinciding with principal axes of inertia as we did above, we will now change to nonrotating equatorial \(x\)- and \(y\)-axes and choose the \(x\)-axis to coincide with the \(X\)-axis of the reference frame tied to the orbital plane \((\hat{X}, \hat{Y})\) (see Figure 8.8). We can do so because we consider a biaxial planet. Any equatorial axis can then be considered as a principal axis of inertia. The coordinates of the Sun \(d_x, d_y,\) and \(d_z\) in the \(\hat{x}, \hat{y},\) and \(\hat{z}\) directions of this frame can then easily be seen to be
\[
\begin{align*}
d_x &= d \cos(nt) \\
d_y &= d \cos \epsilon \sin(nt) \\
d_z &= d \sin \epsilon \sin(nt).
\end{align*}
\]

For the average torque we then have,
\[
\begin{align*}
\Gamma_x &= \frac{3m^2}{2(C - A) \sin \epsilon} \cos \epsilon, \quad \Gamma_y = 0 \quad \text{and} \quad \Gamma_z = 0.
\end{align*}
\]

The fact that there is only an average torque component in the \(X\)-axis can be understood from the geometry of the problem by realizing that the torque is oriented along the line joining the equinoxes when it takes its maximum value at the solstices and diminishes to zero and changes its orientation to the perpendicular direction when the Sun moves to the equinoxes. The average torque is therefore perpendicular to the plane formed by the rotation axis and the orbit normal. As a result, the angular momentum of the planet cannot acquire an additional component in the plane of the orbit normal and rotation axis. The average torque, therefore, cannot change the obliquity of an oblate spheroidal (biaxial) planet and results in a gyroscopic motion in which the rotation axis moves on the surface of a cone with the orbit normal as its axis (precession). In a triaxial case, there will be an additional change in the orientation of the cone with respect to space.
As shown by the above expression, the magnitude of the average solar torque on the planet is determined by the obliquity \( \varepsilon \), the mean motion \( n \), and the parameter \(( C - \overline{A})\). The torque would vanish if the planet were spherically symmetric or if the obliquity \( \varepsilon \) were zero, i.e. if the orbit of the Sun were to lie in the equatorial plane itself. It also decreases for smaller mean motion, or increasing distance to the Sun.

The change in the rotation can be calculated from the angular momentum equation \( \frac{d\mathbf{L}}{dt} = \mathbf{T} \), which is valid only in an inertial reference frame. We therefore need an expression for the torque in an inertial reference frame. We have shown that the average torque has only an \( X \)-component, but the \( X \)-axis considered is not an inertial axis as it changes with the precession we are studying here. Nevertheless, during one orbital evolution of the planet it almost does not change as precession is very slow, which we will see below. Therefore, the \( X \)-component will be to a very good approximation equal to the \( X \)-component of an inertial frame.

Because of precession, the line of the equinoxes and thus the \( \overline{X} \)-axis defined above changes its orientation in space (see Figure 8.9).

Therefore, as seen from Figure 8.9, the rate of change in the \( X \)-direction due to precession is \( \dot{\psi} \sin \varepsilon \). The rate of change in the \( X \)-component of the angular momentum vector, \( \mathbf{H}_X \), is thus equal to the rate of change in the celestial longitude angle \( \psi \sin \varepsilon \) times the angular momentum amplitude \( C\Omega \), as can be seen by projecting the rate of change of the unit angular momentum vector \( \mathbf{H}_X \) on the equator. Therefore we have

\[
\Gamma_x = \mathbf{H}_x = C\Omega\dot{\psi} \sin \varepsilon.
\]

By using this and using the expression of the \( X \)-component of the torque, the precession rate can be expressed as

\[
\dot{\psi} = \frac{3n^2}{2\Omega} \frac{(C - \overline{A})}{C} \cos \varepsilon.
\]

Or equivalently by using \( J_2 = \frac{C - \overline{A}}{Mr^2} \) as

\[
\dot{\psi} = \frac{3n^2}{2\Omega} \frac{Mr^2}{C} J_2 \cos \varepsilon.
\]

For the Earth, the time needed to perform one cycle around the orbit normal is about 25,600 years for a dynamical flattening around \( 1/300 \left( \frac{C - \overline{A}}{C} = 3.274 \times 10^{-3} \right) \). For the other planets, the periods of precession may be very different and increase with increasing dynamical flattening defined in Section 4 and decrease with increasing distance from the Sun. For example for Mars, as seen above, the precession period is about 171,000 years and corresponds to a dynamical flattening of about \( 1/900 \left( \frac{C - \overline{A}}{C} = 1.074 \times 10^{-3} \right) \). For Venus, the precession period is thought to be about 29,000 years, for Mercury about 550 years.

Precession is a very useful observable for the internal geophysics of the Earth and Mars because its rate is inversely proportional to the polar moment of inertia \( C \). Since all other quantities influencing the precession rate are well known (the mean rotation rate, the mean orbital motion, the obliquity, and the difference in moments of inertia from the \( J_2 \) coefficient), observation of the precession yields an estimate of the polar moment of inertia, which is a measure of the radial mass distribution in the planet. For Mars, precession is one of the main and best-determined constraints on the interior structure. Also for the Earth, precession is used in addition to seismology, to determine the internal structure (See Earth: Surface and Interior and Probing the Interiors of Planets with Geophysical Tools).

6. NUTATION

As explained above, the torque changes with time during the orbital motion. Besides having a component perpendicular to the plane formed by the angular momentum and the orbit normal, the torque also has a component perpendicular to the line joining the equinoxes. The former will induce periodic changes to the precession, while the latter will cause the obliquity to change periodically. In addition, due to interaction with other solar system bodies, the orbit changes. All the orbital parameters of the orbit of the planet around the Sun, such as the eccentricity, change with time. As a consequence, the gravitational torque acting on the planet changes with time, and thus also the orientation in space. The changes are periodic with periods equal to harmonics of the orbital motion around the Sun and to the periods with which the orbital elements of the planet around the Sun change, which are themselves related to the orbital periods of the planets perturbing the orbital motion. The situation is similar if one considers moon(s) around a
planet. Figure 8.10 shows these changes in the orientation of the planet in space.

The Movie 2 shows the periodic variations of the orientation of the Earth in space, the nutation.

Supplementary video related to this chapter can be found at http://dx.doi.org/10.1016/B978-0-12-415845-0.00008-6.

The following is/are the supplementary data related to this chapter: Movie 2: Nutation of the Earth in space. (file EarthNutation.mpg)

While precession carries the pole of the axis at a uniform rate in an anticlockwise sense along a circle on the surface of the celestial sphere, centered on the normal to the ecliptic plane, nutation consists of small deviations from this uniform motion, both over the precessional path (called nutation in longitude) and perpendicular to it on the celestial sphere (called nutation in obliquity). The resulting path of the pole appears wiggly, as in Figure 8.10. For a hypothetical planet with a circular orbit, the torque variation has a period equal to half the orbital period. Since the planets of the solar system all have orbits close to circular, the main nutation will also have a semiannual period.

The amplitude of the nutation with a period half the orbital period depends on the shape of the planet, the obliquity, the rotation speed, and the orbital speed. For a circular orbit, this can be shown as follows.

As we have done for the constant part of the torque, starting from its above expression (see Section 5) and substituting the expressions for the position of the Sun with respect to the planet, the time-dependent torque as a function of the angular velocity \( n \) and the obliquity \( \epsilon \) can thus be written as:

\[
\Gamma_x = \frac{3n^2}{2} (C - \bar{A}) \sin \epsilon \cos \epsilon (1 - \cos 2nt),
\]

\[
\Gamma_y = -\frac{3n^2}{2} (C - \bar{A}) \sin \epsilon \sin 2nt, \quad \text{and} \quad \Gamma_z = 0
\]

For \( F = \frac{3\mu^2}{2} (C - \bar{A}) \sin \epsilon \), the torque becomes,

\( \Gamma_x = F \cos \epsilon (1 - \cos 2nt), \quad \Gamma_y = -F \sin 2nt, \quad \text{and} \quad \Gamma_z = 0. \)

As shown above, the constant term leads to precession. The time-dependent part causes nutation. The magnitude of the time-dependent part of the solar torque on the planet is determined by the mean motion \( n \), the obliquity \( \epsilon \), and the parameter \( (C - \bar{A}) \). As previously for precession, the torque would vanish if the planet were spherically symmetric or if the obliquity \( \epsilon \) were zero, i.e. if the orbit of the Sun were to lie in the equatorial plane itself.

We also see that the torque has a period of half a year (frequency \( 2n \)). The time variation of \( (\Gamma_x, \Gamma_y) \) describes an elliptical path in the equatorial plane. The periodic part of the torque components are

\[
\Gamma_x = -F \cos \epsilon \cos 2nt, \quad \Gamma_y = -F \sin 2nt.
\]

The semiannual nutation, which is the largest of the nutations of solar origin, is the result of this periodic term in the torque. Both the torque and the nutation may be resolved, like any periodic elliptical motion, into two counterrotating circular motions:

\[
\begin{pmatrix}
\Gamma_x \\
\Gamma_y
\end{pmatrix} = \frac{F}{2} \begin{pmatrix}
(1 + \cos \epsilon) \cos 2nt \\
(1 - \cos \epsilon) \sin 2nt
\end{pmatrix}.
\]

The first of the two column vectors on the right-hand side of this equation represents a vector in prograde (counterclockwise) motion, while the second one is a retrograde motion. The prograde part of the nutation is due to the prograde part of the torque and the retrograde part of the nutation results from the retrograde part of the torque. Prograde means that the angle increases with time counterclockwise like the rotation of the Earth; retrograde means that the angle increases clockwise with time. The superposition of two motions, one clockwise and the other counterclockwise, gives rise to an elliptical motion. The direction of the elliptical motion depends on which amplitude of the prograde and retrograde terms is largest. For the semiannual nutation, the first (prograde) term is larger than the second (retrograde) term \((1 + \cos \epsilon) > (1 - \cos \epsilon)\), and the elliptical semiannual nutation is prograde.

A lot of nutation terms other than semiannual arise when one takes into account the deviations from the simple model that we considered here: the orbit of the Sun relative to the planet is not circular but elliptical, so that \( d \) is not constant, the simple harmonic time dependence of \( d_t \) and \( d_c \) is a simplification, and there are other planetary perturbations of the orbit of the Earth around the Sun. The present simplified picture serves to highlight the essential aspects of the nutation—precession phenomenon.

Nutation have up to now only been unambiguously observed for the Earth. Very long baseline interferometry (VLBI, see Section 10.3) observations provide the precession and nutation of the Earth with a precision below the tenth of a milliarcsecond (see Section 10). For this semiannual nutation, the amplitudes of the nutations in
longitude and obliquity are about 1300 mas and 570 mas, respectively, which corresponds to a quasicircular motion at the Earth's surface of about 16 m as seen from space (see Table 8.2). The largest nutations are of lunar origin and have a period of about 18.6 years (and 9.3 years to a minor extent), which arises from the precession of the lunar orbit around the ecliptic. The amplitudes of nutation in longitude and obliquity of the 18.6-year nutation are about 17200 mas and 9200 mas, respectively (for the 9.3-year nutation they are about 200 mas and 100 mas, respectively). The nutations of solar origin at 1 year have amplitudes provided in Table 8.2.

Nutation for a rigid planet can easily be computed from the torque as performed above for precession, given that the ephemerides (relative positions of the celestial bodies) are known. Those theoretical values, however, do not all correspond to the observed values since the planets are deformable and thus nonrigid and contain a liquid layer such as the outer core of the Earth. Thanks to the high precision of the observations, information on the interior of the Earth can be obtained. In order to do so, models for the precession and nutation of a nonrigid body have been developed. A short description of such a model is given in Section 9 on the wobbles. That section also discusses results of a comparison between observational results and theoretical modeling for the Earth and explains what theory predicts for the nutations of Mars.

7. LOD VARIATIONS

The rotation of a planet is approximately uniform, meaning that the planet rotates with an almost constant rate. Small variations in the rotation rate occur due to various reasons. For the terrestrial planets with an atmosphere, the largest of these changes are due to the atmosphere dynamics. For Mercury the largest rotation variations are most likely due to the gravitational torque exerted by the Sun. This will be further developed in the next section (see Section 8). The gas and ice planets, which are primarily composed of H, He, and hydrides of the light elements O, C, and N (i.e. water H2O, methane CH4, and ammonia NH3) are mainly fluid. Different large atmospheric patterns such as the great Red Spot on Jupiter and large-scale circulations such as banded structures have been observed. Since the global mean rotation rate of the giant planets is not well known (e.g. within a few minutes for Saturn) here we only consider how exchanges of angular momentum within the planet can affect the rotation of the surface of the terrestrial planets.

If we consider an isolated planet or neglect external torques, the angular momentum of the system is conserved:

$$\frac{d\vec{H}}{dt} = 0.$$ Any change in the angular momentum a layer of the planet must necessarily be balanced by a change in angular momentum of another layer. Therefore, the rotation of the solid outer part of the terrestrial planet (solid) can change due to angular momentum exchange with other layers. For example, the Earth’s rotation rate changes due to interactions and angular momentum exchange of the mantle and crust with the atmosphere, the liquid outer core, and the solid inner core (See Earth: Surface and Interior). As an example, we consider angular exchange with the atmosphere (atm) only. Because of conservation of angular momentum, we have

$$\frac{dH_{\text{solid}}}{dt} = -\frac{dH_{\text{atm}}}{dt}.$$ Changes in the winds and in the mass distribution in the atmosphere both affect the angular momentum of the atmosphere (H_{\text{atm}}) and consequently change the angular momentum and rotation of the solid planet (H_{\text{solid}}). For example, stronger winds from the west will increase the angular momentum of the atmosphere and reduce the angular momentum of the solid planet. Likewise, if mass is transferred to the polar regions, the atmospheric angular momentum decreases and the planet rotates faster. This effect is particularly important for Mars, in particular since during winter about one-fourth of the total mass of the CO2 atmosphere of Mars condenses at the winter pole. If the planet can be considered to be rigid, any change in the angular momentum of the solid planet transfers immediately into changes in the angular velocity with respect to a uniform rotation \(\Omega\) along the z-axis:

$$-\frac{dH_{\text{atm}}}{dt} = \gamma \vec{\Omega}$$
For a rigid body rotating around an axis of symmetry (say the polar axis $\hat{z}$), the angular momentum can be expressed as the product of the polar moment of inertia, $C$, and its angular velocity $\Omega_z$: $H_z = C\Omega_z$. We then have

$$\frac{dH_z}{dt} = C\frac{d\Omega_z}{dt}.$$ 

The change in the rotation of the planet is therefore proportional to the change in the angular momentum of the atmosphere and inversely proportional to the polar moment of inertia, which expresses the rotational inertia. The most important changes in the angular momentum in the atmosphere are due to the seasons. The Earth’s LOD changes at seasonal timescales by a few milliseconds (see Figure 8.12 and Figure 8.13). Variations due to weather systems on much shorter timescales can change the rotation of the Earth by up to a few milliseconds.

As an illustration, Movie 3 shows the uniform rotation of the Earth viewed from space and a comparison with the nonuniform case. Movie 4 shows first the
nonuniform rotation, then the nutation, and next the polar motion (defined in Section 9 as the motion of the rotation axis in a frame tied to the planet). Movie 5 shows first the combination of all motions, then the nonuniform rotation, and separately the nutation, then the polar motion. In these animations we have exaggerated the motions and did not respect the real timing for the Earth. They are only intended to illustrate the physical phenomena.

Supplementary video related to this chapter can be found at http://dx.doi.org/10.1016/B978-0-12-415845-0.00008-6.

The following is/are the supplementary data related to this chapter: Movie 3: Earth uniform rotation. (file EarthUniformRotation.wmv)
Supplementary video related to this chapter can be found at http://dx.doi.org/10.1016/B978-0-12-415845-0.00008-6.

The following is/are the supplementary data related to this chapter: Movie 4: Earth nonuniform rotation nutation. (file EarthNonUniformRotationNutationPM.wmv)
Supplementary video related to this chapter can be found at http://dx.doi.org/10.1016/B978-0-12-415845-0.00008-6.

The following is/are the supplementary data related to this chapter: Movie 5: Earth nonuniform orientation. (file EarthNonUniformOrientation.wmv)

Exchange of angular momentum with the liquid outer core is also an important cause of rotation variations for the Earth and is a likely cause of rotation variations for the other terrestrial planets, in particular Mercury. At decadal timescales, flow in the Earth’s core related to variations in the geodynamo (See Magnetic Field Generation) leads to changes in the angular momentum of the core which are partially transferred to the solid parts of the planet through several coupling mechanisms at the CMB and inner core boundary (ICB). In a first approximation, the motion in the core is often supposed to be geostrophic, such that the flow has constant velocity on cylinders coaxial with the rotation axis. Each cylinder is rotating at its own rotation angular velocity and these rotation speeds are varying with time and linked through the so-called torsional oscillations of which the amplitudes varies with time. From the velocity field in the whole core, one can compute the angular momentum of the core, and hence, the variation of the LOD induced by the time variation of this quantity.

The coupling mechanisms are mainly due to the gravitational torque resulting from the core—mantle gravitational interactions, the electromagnetic torque, and the pressure torque mainly related to the differential motions of the core with respect to the mantle on the bumpy boundary of the core. Figure 8.14 illustrates these coupling mechanisms in the case of the atmosphere. Figure 8.11 shows the LOD variations of the Earth over several decades. Both the decadal variations related to the decadal changes in the core angular momentum and seasonal variations due to interaction with the atmosphere are clearly visible. Figure 8.12 considers Earth rotation variations over 12 years and better shows the seasonal and shorter variations due to interaction with the atmosphere. At that timescale, angular momentum exchange with the oceans contributes about 10% of the rotation variations.

Variations in the angular momentum of a fluid layer can be computed if the velocity field in the fluid is known. For
the atmosphere or the ocean, general circulation models can be used and the angular momentum can be calculated with the help of the following integral over the volume of the fluid layer

\[ H_{\text{fluid}} = \frac{r_{\text{surface}}^4}{g} \int_0^{2\pi} \int_0^{2\pi} P_{\text{surface}} (\theta, \lambda) \sin^2 \theta \begin{pmatrix} \cos \theta \cos \lambda & \cos \theta \sin \lambda \\ \sin \theta & \sin \theta \end{pmatrix} d\theta d\lambda + \frac{r_{\text{surface}}^3}{g} \int_0^{2\pi} \int_0^{2\pi} \begin{pmatrix} u_\theta \sin \lambda + u_\lambda \cos \theta \cos \lambda \\ -u_\theta \cos \lambda + u_\lambda \cos \theta \sin \lambda \\ -u_\lambda \cos \lambda \end{pmatrix} dp d\theta d\lambda \]

\[ = H_{\text{atmosphere}} + H_{\text{wind}} \]

The first term depends on the pressure and thus on the total atmospheric mass rotating with the planet. It is the angular momentum associated with a global rotation of the atmosphere with the planet and is called the mass term or inertia term. The second term involves the relative velocity with respect to the solid planet and is called the wind term. The sum of time derivative of these two terms can also be expressed in terms of torques acting on the surface.

The changes in the rotation of the planet can be computed from Global Circulation models (GCMs) providing maps of the hydrostatic pressure on the surface of the planet and maps of winds at the surface of the planet.

These exchanges of angular momentum exist on Mars as well. Mars’ atmosphere consists of 95% carbon dioxide (CO2) and, although it is much more diluted and has a surface pressure hundred times smaller than that on the Earth, it undergoes large seasonal changes related to the CO2 sublimation and condensation process (See Mars Atmosphere: History and Surface Interactions). About one-fourth of the atmosphere is participating in this process.
forming large ice caps in the winter of each hemisphere (the maximum mass of the north and south seasonal caps that is produced is $3.0 \times 10^{15}$ and $5.5 \times 10^{15}$ kg, respectively; the total mass of the atmosphere is about $2.5 \times 10^{16}$ kg). Using GCMs, it is possible to compute the angular momentum in the Martian atmosphere and therewith to estimate the large changes in the LOD. LOD variations have amplitudes at the same level as that on Earth, at the level of several tenths of millisecond with annual and semiannual periods. Wind is found to induce rotation angle variations with an amplitude of 14 mas for the annual period and 76 mas for the semiannual period. The effect of the ice caps is small because mass at the polar ice caps has only a small contribution to the polar moment of inertia $C$, whereas atmospheric masses can be a Mars radius further from the polar axis and cause a larger change in the polar moment of inertia. The amplitude of the rotation variations is at the level of 500 mas on the equator over 1 year corresponding to annual changes in the LOD at the level of 0.3 ms. For the semiannual amplitude, we have 170 mas on the equator, corresponding to LOD changes at the level of 0.2 ms. A global dust storm as it might happen on Mars may induce variations in the LOD of several percents of the total seasonal effect.

LOD variations on Venus are expected to be small because Venus does not have seasons as the Earth and Mars (See Venus Atmosphere).

8. LIBRATION

No planet or satellite is exactly spherically symmetric. In a lowest order approximation, planets and satellites take the form of an ellipsoid (see Section 4). For rapidly rotating planets, an oblate spheroid (biaxial ellipsoid) flattened at the poles represents well the shape (the equator is a perfect circle but any meridian circle passing through the poles is an ellipsoid flattened at the equator). For a triaxial ellipsoid planet or satellite, the equator is also an ellipse.

The three principal moments of inertia are then different ($A < B < C$), while for the spheroid $A = B$. Due to the rotational and orbital motion the long axis of the planet or satellite in the equatorial plane, coinciding with the axis of smallest principal moment of inertia $A$ (see Section 4), is generally not in the direction to the central body. Therefore, the central body exerts a gravitational torque on the planet or satellite, which therefore accelerates or decelerates its rotation depending on its orientation with respect to the central body. When the planet’s rotation is much faster than the orbital motion as, for example, for the Earth and Mars, the effects on rotation are very small. However, in the case of a spin–orbit resonance, the torque of the central body can cause relatively large variations in the rotation rate. These longitudinal librations are important for Mercury and also for satellites in a 1:1 spin–orbit resonance. Figure 8.15 illustrates the geometry of the problem for satellites in a 1:1 spin–orbit resonance. Although the rotational period is exactly equal to the rotational period, the satellites do not exactly show the same face to the central planet because the speed of orbital motion varies according to Kepler’s second law as a result of the eccentricity of the orbit.

We neglect the very small obliquity of Mercury and the synchronous satellites. The polar component of the gravitational torque exerted by the central body can then be expressed as (see Section 5)

$$I_3 = \frac{3}{2} (B - A) \frac{GM_{\text{central}}}{r^3} \sin 2\xi$$

where $r$ is the distance between the mass centers of the central body and the planet or satellite, and $\xi$ is the angle between the direction of the long axis and the direction to the central body ($\xi = f - \varphi$, where $f$ is the true anomaly and $\varphi$ is the rotation angle between the long axis of the planet or satellite and the major axis of the orbit) (see Figure. 8.16).

For satellites in a 1:1 spin–orbit resonance, the rotation angle is close to the mean anomaly describing the mean angle associated with orbital motion and defined as $M = \frac{2\pi}{T} (t - t_0)$, where $t_0$ is the time when the satellite passes through pericenter. One can then introduce a small libration angle $\gamma = \varphi - M$.

For Mercury, the rotation angle changes faster than the orbital angle by a factor of about 1.5 and the small libration angle is defined as $M = \frac{2\pi}{T} (t - t_0)$. At the pericenter, the libration angle describes the difference between the direction of the long axis and the direction to the central body. It can be determined by solving the equation for the change in angular momentum due to the above torque.

By expressing the true anomaly and the distance between the mass centers in terms of the semimajor axis, the
eccentricity, and the mean anomaly for a Keplerian orbit, the torque can be seen to consist of a term proportional to the libration angle and periodic terms with period equal to the orbital period and its harmonic frequencies. If we restrict ourselves to the largest term with period equal to the orbital period and assume that Mercury behaves rigidly, the equation for angular momentum of Mercury can be expressed as the pendulum equation

\[
\frac{d^2 \gamma}{dM^2} + \frac{3}{2} \frac{B - A}{C} e \left( \gamma - \frac{123}{8} e^2 \right) \gamma = \frac{3}{2} \frac{B - A}{C} \left[ (1 - 11 e^2 + \ldots) \sin M \right].
\]

Also, for synchronously rotating satellites, the governing equation for libration can be expressed as a pendulum equation. The amplitude of the libration angle of the main libration at orbital period of 88 days, \( \gamma_{88\text{days}} \), can then be approximately expressed as

\[
\gamma_{88\text{days}} = \frac{3}{2} \frac{B - A}{C} (1 - 11 e^2 + \ldots).
\]

For rigid synchronous satellites, the amplitude of libration at the orbital period, \( \gamma_{\text{orbital period}} \), can be approximately expressed as

\[
\gamma_{\text{orbital period}} = 6 \frac{B - A}{C} e.
\]

An important difference with respect to the libration amplitude \( \gamma_{88\text{days}} \) of Mercury in 3:2 spin–orbit resonance is that the libration amplitude \( \gamma_{\text{orbital period}} \) for a 1:1 spin–orbit resonance decreases linearly with the eccentricity. The decrease with eccentricity for a satellite in a 1:1 spin–orbit resonance can be understood by considering the limit case of a circular orbit. In that case, there would not be any forced libration in longitude since the orbital speed is then constant and the satellite always shows exactly its same face to the central planet.

As is typical for a pendulum equation, a free libration is also possible. When the long axis does not point toward the Sun at perihelion, the averaged gravitational torque on Mercury tends to restore the alignment, and the long axis will librate around the direction to the Sun at perihelion. The free libration period of a rigid Mercury is given by

\[
P_{\text{free}} = \sqrt{\frac{1}{\frac{3}{2} \frac{B - A}{C} e \left( \gamma - \frac{123}{8} e^2 \right)}} \frac{2\pi}{n}.
\]

Free libration is also due to the gravitational interaction between the Sun and Mercury, as is forced libration, but the distinctive feature of free libration as opposed to forced libration is that the amplitude and phase of free libration cannot be determined without knowing its excitation and dissipation, as is the case for a free harmonic oscillator. Free libration has a much longer period of several years than the main forced libration at 88 days and has not yet been unambiguously observed.

An illustration of the libration of Mercury is shown in Movie 6.

Supplementary video related to this chapter can be found at http://dx.doi.org/10.1016/B978-0-12-415845-0.00008-6.

The following is/are the supplementary data related to this chapter: Movie 6: Libration of Mercury. (file MercuryLibration.wmv)

Like nutation, libration depends on the interior. The main effect, and probably the only effect that is substantially larger than the precision of current observation techniques, is due to the liquid outer core of Mercury. A liquid outer core effectively decouples the libration of the mantle from the core. The total torque on the mantle from the Sun and the core is equal to the total torque on an entirely rigid Mercury but the polar moment of inertia, which represents the resistance to rotational forcing, is smaller than the total planetary polar moment of inertia. Therefore, the libration of Mercury with a liquid outer core is larger than that of an
entirely solid Mercury by about a factor \( C/C_m \). The libration period with a liquid core is also shorter than that without by a factor \( (C_m/C)^{1/2} \). Based on radar observations using two different large radio telescopes (see Section 10), the amplitude of the 88-day libration of Mercury has been determined as \((38.5 \pm 1.6)\) arcsec. This means that a point at the equator of Mercury will be periodically offset from its equilibrium position at constant rotation rate with an amplitude of \((355 \pm 19)\) m. By using \( B - A = (3.235 \pm 0.024) \times 10^{-5} \) MR\(^2\), estimated from radio tracking observations of the National Aeronautics and Space Administration (NASA) MESSENGER mission currently in orbit around Mercury, the polar moment of the mantle can then be determined as \( C_m = (0.148 \pm 0.006) \) MR\(^2\), less than half the value for the total planetary moment of inertia determined from the obliquity value. Therefore, Mercury cannot be entirely solid and a global liquid layer must exist. For an entirely solid Mercury, the libration amplitude at the equator would be about 190 m. The comparison of the rotational constraints on \( C \) and \( C_m \) with theoretical moment of inertia values for models of the interior structure of Mercury shows that the liquid core has a radius of about 2000 km. The core of Mercury is probably larger than the core of the larger planet Mars. It is relatively much larger than the core of the Earth, which has a radius of about 55% of the total radius of the Earth. For Mercury, the relative core radius is about 80%. This clearly shows that Mercury has had a different formation history.

9. WOBBLES AND THE INTERIORS OF TERRESTRIAL planets

Besides precession and nutation which describe the changes in the orientation of a planet in space, the position of the rotation axis can also change with respect to a reference frame tied to the planet. The term wobble is used in a very broad sense for any periodic or quasiperiodic motion of the instantaneous rotation axis of a planet or moon with respect to the figure axis, i.e. conventionally the \( z \)-axis of the frame tied to the planet, irrespective of the frequency or the physical origin of the motion. In general, wobble cannot occur without accompanying nutation, and equivalently, nutation is always associated with wobble, as we explain below. Depending on the period, nutation is either much larger or much smaller than wobble for planets, like the Earth, rotating much faster than the orbital motion. More specifically, at periods much longer than a day in the terrestrial frame, the wobble has a larger amplitude than the corresponding motion of the instantaneous rotation axis in space. The frequency of the latter motion is quasi-diurnal due to the rotation of the planet. Consider for instance a static or zero-frequency phenomenon in a frame tied to the planet. With respect to inertial space, the signal will have a diurnal frequency due to the rotation of the planet. In general, if the wobble frequency in the terrestrial frame is \( \sigma \), the corresponding frequency in the celestial frame is \( \sigma' = \sigma + \Omega \), the nutation frequency. Since the gravitational torque inducing nutation has long periods with respect to an inertial reference frame, the nutation frequencies \( \sigma' \) are much smaller than \( \Omega \) and the associated wobble has retrograde quasi-diurnal frequencies. At these periods, the nutation amplitude is much larger than that of the wobble.

Traditionally, the two motions, wobble and nutation, are represented by two cones, with the smaller one rotating inside the larger one without gliding (see Figure 8.17).

The study of the wobble of a planet is performed by observing the instantaneous position of the rotation axis (i.e. the direction of the instantaneous angular velocity vector \( \Omega \)) from the direction of the mean angular velocity vector \( \Omega \), which is also the direction of the figure axis. The components of \( \Omega \) in the terrestrial frame are usually denoted by

\[
\begin{pmatrix}
0 \\
0 \\
m_1 \\
1 \\
m_2 \\
m_3 
\end{pmatrix}
\]

FIGURE 8.17 On the left, Euler–Poinsot representation of the rotation axis motion in space (blue) and in the terrestrial frame (red) for the Earth; on the right, representation of the nutation and diurnal wobble for the Earth.
where \( m_1, m_2, \) and \( m_3 \) are small quantities; variations of \( m_3 \)
represent the fractional variations in the spin rate, which
manifest themselves in LOD variations as seen in Section 7;
\( m_1 \) and \( m_2 \) express the deviation of the instantaneous rotation
axis with respect to the uniform rotation axis in the
plane perpendicular to the figure axis \( \hat{z} \) and so describes
wobble (or polar motion, see Section 9.1) in a reference
frame \((\hat{x}, \hat{y}, \hat{z})\) tied to the planet and rotating uniformly as
introduced before.

In order to demonstrate the link between nutation and
wobble, we introduce the Euler angles relating the inertial
\((\vec{X}, \vec{Y}, \vec{Z})\) and body-fixed \((\hat{x}, \hat{y}, \hat{z})\) reference frames (see
Figure 8.18).

Three rotations are needed to pass from one frame to the
other:

1. A rotation of angle \( \Psi \) around \( \hat{z} \):
   \[
   R(\Psi) = \begin{pmatrix}
   \cos \Psi & \sin \Psi & 0 \\
   -\sin \Psi & \cos \Psi & 0 \\
   0 & 0 & 1
   \end{pmatrix},
   \]

2. A rotation of angle \( \theta \) around \( \hat{x} \):
   \[
   R(\theta) = \begin{pmatrix}
   1 & 0 & 0 \\
   0 & \cos \theta & -\sin \theta \\
   0 & \sin \theta & \cos \theta
   \end{pmatrix},
   \]

3. A rotation of angle \( \phi \) around \( \hat{Z} \):
   \[
   R(\phi) = \begin{pmatrix}
   \cos \phi & \sin \phi & 0 \\
   -\sin \phi & \cos \phi & 0 \\
   0 & 0 & 1
   \end{pmatrix}.
   \]

The body-fixed reference frame rotates with respect to the
inertial reference frame with the instantaneous rotation
vector \( \vec{\Omega} \). Therefore, the Euler angles change with time.
The rotation vector \( \vec{\Omega} \) can thus be expressed as a function
of the time derivatives of the Euler angles. Over an infin-
itesimal time \( dt \) the change in the relative positions of the
two reference frames can be obtained from the infinitesimal
changes in the Euler angles by performing the sum of three
rotations over (1) the angle \( d\Psi (= \Psi dt) \) around the axis \( \hat{Z} \)
of the frame \((\vec{X}, \vec{Y}, \vec{Z})\) in space, (2) the angle \( d\theta = \theta dt \) around
the axis of intersection between the plane \((\vec{X}, \vec{Y})\) and the plane
\((\hat{x}, \hat{y})\), and (3) the angle \( d\phi = \phi dt \) around the \( \hat{z} \) of
the frame \((\hat{x}, \hat{y}, \hat{z})\). The instantaneous rotation vector
can thus be expressed as
\[
\vec{\Omega} = \vec{\Psi} \vec{Z} + \dot{\theta} \vec{x} + \phi \vec{z}.
\]

For the Earth (and Mars), the equatorial terrestrial frame
\((\vec{x}, \vec{y}, \hat{z})\) is tied to the planet and the celestial frame
\((\vec{X}, \vec{Y}, \vec{Z})\) in space is tied to the ecliptic. The relative
orientation of these frames differs from the representation
ing Figure 8.18 classically used for Euler angles as the
\( \vec{Y} \)-axis in the ecliptic lies in the Northern Hemisphere and
the obliquity \( \epsilon \) is defined as \( \epsilon = -\theta \). We then have
\[
\vec{\Omega} = \vec{\Psi} \vec{Z} + (-\epsilon) \vec{x} + \phi \vec{z}.
\]

In the terrestrial reference frame, the instantaneous rotation
vector is given by
\[
\vec{\Omega} = \begin{pmatrix}
-\Psi \sin \epsilon \sin \phi - \dot{\epsilon} \cos \phi \\
-\Psi \sin \epsilon \cos \phi + \dot{\epsilon} \sin \phi \\
\Psi \cos \epsilon + \phi
\end{pmatrix}.
\]

Identifying this expression with the above expression for the
instantaneous rotation vector involving \( m_1, m_2, \) and \( m_3, \)
we obtain the famous kinematic equation of Euler:
\[
\dot{\vec{\Psi}} = -(\Omega_1 + i\Omega_2)e^{i\phi} = -\Omega(m_1 + im_2)e^{i\Omega t}.
\]

This equation provides the relation between the nutation
angles \( \Psi \) and \( \epsilon \), and the wobble components \( (m_1, m_2) \) and
proves that nutation and wobble are intimately related.

### 9.1. Polar Motion

The wobble of the instantaneous rotation axis has several
components grouped in different frequency bands. One of
them is the small-amplitude retrograde diurnal frequency
component related to nutation. Larger wobble, commonly
referred to as polar motion, occurs at long periods in the
body-fixed reference frame. For the Earth, for example, the
position where the rotation axis intersects the surface
changes seasonally. The separation from the mean position
remains below 20 m (see Figure 8.19).
**Movie 7** shows the polar motion of the Earth viewed from space.

Supplementary video related to this chapter can be found at http://dx.doi.org/10.1016/B978-0-12-415845-0.00008-6.

The following is/are the supplementary data related to this chapter: Movie 7: Polar motion for the Earth. (file EarthPolarMotion.mpg)

As for LOD variations, seasonal polar motion of the Earth is mainly due to the effect of the atmosphere and the other geophysical fluids such as the ocean and the hydrosphere. As the rotation changes, it can be computed from the angular moment exchange between the solid Earth and the fluid layers (conservation of the total angular momentum).

The wobbles of a planet are computed from the Liouville equations describing the angular momentum conservation equation for a nonrigid planet. The angular momentum of a rotating body $\vec{H}$ is the sum of two terms, one related to the global rotation (as seen previously in Section 0) and defined as the product of the inertia matrix $\vec{I}$ and the rotation vector $\vec{\Omega}$, and the other related to the relative moment of inertia $\vec{h}$, due to the relative motion with respect to the corotating frame in the fluid parts of the planet:

$$\vec{H} = \vec{I} \vec{\Omega} + \vec{h}$$

Angular momentum conservation can be expressed in a reference frame corotating with the mantle as

$$\frac{\partial \vec{H}}{\partial t} + \vec{\Omega} \times \vec{H} = \vec{T}$$

where $\vec{T}$ is the external torque, and $\partial/\partial t$ expresses the time derivative in the corotating reference frame. The inertia matrix in a reference frame tied to the moment of inertia can be written as:

$$\vec{I} = \begin{pmatrix} A & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & C \end{pmatrix} + \vec{\epsilon}$$

where $\vec{\epsilon}$ is the incremental inertia matrix due to the deformation of the planet and $A$, $B$, and $C$ are the equatorial and polar moments of inertia as seen before (see Section 4). For a biaxial ellipsoidal planet (typical for fast rotating planets), $A = B$.

Using the expressions:

$$\begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix} \quad \vec{T} = \begin{pmatrix} \Gamma_1 \\ \Gamma_2 \\ \Gamma_3 \end{pmatrix} \quad \vec{\epsilon} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{pmatrix}$$

**FIGURE 8.19** Polar motion as computed from observation on the IERS Website of the Observatoire de Paris (http://hpiers.obspm.fr/eop-pc/).
and limiting ourselves to terms of the first order in the small quantities, one obtains the so-called Liouville equations:

\[
\begin{align*}
\Omega \dot{m}_1 + \Omega^2 (C - A) m_2 + \Omega c_{13} - \Omega^2 c_{23} + \dot{h}_1 - \Omega h_2 &= \Gamma_1 \\
\Omega \dot{m}_2 - \Omega^2 (C - A) m_1 + \Omega c_{23} + \Omega^2 c_{13} + \dot{h}_2 + \Omega h_1 &= \Gamma_2 \\
\Omega \dot{m}_3 + \Omega c_{33} + \dot{h}_3 &= \Gamma_3
\end{align*}
\]

For a rigid planet (meaning that \( c_{13} = c_{23} = 0 \), and \( h_1 = h_2 = 0 \) in the above equation), the equations simplify to:

\[
\begin{align*}
\Omega \dot{m}_1 + \Omega^2 (C - A) m_2 &= \Gamma_1 \\
\Omega \dot{m}_2 - \Omega^2 (C - A) m_1 &= \Gamma_2 \\
\Omega \dot{m}_3 &= \Gamma_3
\end{align*}
\]

The Liouville equations can be brought to a simpler form by using complex notation. For one wobble component, we write the time dependence of the wobble as \( \exp(i\sigma t) \), where \( \sigma \) is the wobble frequency. The time derivative of the variables can then be expressed as \( i\sigma \) times the variable. On taking the complex sum of the first two equations and introducing \( m = m_1 + im_2 \) and \( \Gamma = \Gamma_1 + i\Gamma_2 \), one obtains:

\[
\Omega \dot{m} - \Omega^2 (C - A) m = \Gamma
\]

Without forcing (\( \Gamma = 0 \)), polar motion (if excited) is possible for a particular frequency, the free Euler frequency

\[
\sigma_{\text{Euler}} = \frac{(C - A)}{A} \Omega.
\]

The Euler frequency depends only on the dynamical flattening (and thus on the moments of inertia) and the rotation rate of the planet. If the free polar motion is excited, the rotation axis has a circular motion in the \((\vec{x}, \vec{y})\)-plane at the Euler period. The amplitude depends on the excitation and damping.

In reality, terrestrial planets contain a liquid core and some also a solid inner core. For those internal layers, additional small deviations from the mean rotation must be considered since they will not, in general, rotate precisely with the mantle. Considering in a first approximation that the planet can still be considered rigid (deformations are neglected) and neglecting a possible solid inner core, two complex equations can be derived for the complex polar motion of the mantle \( m \) and the complex polar motion of the core \( m_f \). Those Liouville equations can be expressed as

\[
\begin{align*}
i\sigma \Omega \dot{m} + i\sigma \Omega A_f m_f - i\Omega^2 e_A m + i\Omega^2 A_f m_f &= \Gamma \\
&= -ie\Omega^2 A\phi \\
i\sigma \Omega A_f m + i\sigma \Omega A_f m_f + i\Omega^2 (1 + e_1) A_f m_f &= 0
\end{align*}
\]

These two equations are two linear equations in \( m \) and \( m_f \). They can be expressed in matrix form as

\[
\begin{pmatrix}
\sigma - e\Omega & (\sigma + \Omega) \frac{A_f}{A} \\
\sigma & \sigma + \Omega (1 + e_1)
\end{pmatrix}
\begin{pmatrix}
m \\
m_f
\end{pmatrix}
= \begin{pmatrix}
-e\Omega \phi \\
0
\end{pmatrix}.
\]

In the absence of any forcing \( \Gamma \) or \( \phi \), polar motion is possible again for a particular frequency, the frequency of the free wobble. The frequencies providing a solution of the above equations for no external forcing are computed from equating to zero the determinant of the matrix. Since polar motion occurs mainly at seasonal periods, the frequency of each polar motion contribution is small with respect to the rotation frequency for rapidly rotating planets. Therefore, the matrix simplifies further and the frequency of the free wobble \( \sigma_{\text{free}} \) can be expressed as

\[
\sigma_{\text{free}} = \frac{A}{A_m} e\Omega = \frac{(C - A)}{A_m} \Omega.
\]

The existence of a liquid core decouples, to first order, the mantle wobble from that of the core. Therefore, the period of the free wobble is proportional to the moment of inertia of the mantle and not that of the whole planet, thereby decreasing the period with respect to an entirely solid planet. The free frequency is, therefore, different from the Euler frequency. Moreover, deformation changes the frequency, which for terrestrial planets is called the Chandler frequency. The associated polar motion is known as the Chandler wobble. For the Earth, the Euler period is 305 days and the Chandler period is 433 days, showing that non-rigid effects are very important. The different non-rigid contributions to the period are shown in Figure 8.20 for the Earth. Besides elasticity, the existence of the ocean and the mantle inelasticity also increase the period.
Besides the free polar motion, the atmosphere and other fluid layers of the Earth also excite polar motion at seasonal periods. Both the annual component and the Chandler Wobble contribution have amplitudes of about a few meters. The seasonal and Chandler components sometimes add up constructively providing a large amplitude, or sometimes subtract destructively providing a small amplitude in polar motion. At the beginning of 2013, polar motion was very small (see Figure 8.19).

In principle, as Mars has an atmosphere and seasons, the polar motion of Mars will also consist of a wobble with a period of about 200 days. It has, therefore, not yet been observed. Mars is also expected to have a Chandler (see Figure 8.19 for the Earth). The seasonal and Chandler components sometimes add up constructively providing a large amplitude in polar motion. At the beginning of 2013, polar motion was very small (see Figure 8.19). The FCN period can be computed from

\[
\begin{pmatrix}
-\Omega & \sigma' \left( \frac{A_t}{A} \right) \\
-\Omega & \sigma' + e_f \Omega
\end{pmatrix}
\begin{pmatrix}
m \\
m_f
\end{pmatrix}
= \begin{pmatrix}
-e \Omega \varphi \\
0
\end{pmatrix}
\]

The solutions of this system show that both \( m \) and \( m_f \) are resonantly amplified by the FCN frequency. As the nutation can be computed from \( m \), as shown above, one immediately sees that the FCN amplifies the planet orientation changes induced by the tidal gravitational torque. Observing the nutation will thus provide information on whether the core is liquid or entirely solid and even, when the FCN can be exactly determined, on the core moment of inertia \( I_t \) and the core flattening \( e_c \), and hence on the core radius and density, as the core hydrostatic dynamical flattening and core moment of inertia depend on those quantities.

For the Earth, the FCN has a period \( \frac{1}{\sigma_{\text{FCN}}} \) of about \( -430 \) days in the celestial frame. For Mars, the FCN period has not yet been observed but theoretical computations based on plausible interior models for Mars provide a range of values between \(-220\) and \(-280\) days. The FCN period could thus be very close to the ter-annual Martian nutation at 229 days and thus largely amplify it, making it better observable and usable for deducing information on the Martian core.

### 10. Observation of the Rotation of Terrestrial Planets

#### 10.1. Ground-Based Observations

Rotations of solid planets can be observed from Earth-based telescopes. The principle uses the transmission by Earth-based radar telescopes of a circularly polarized monochromatic X-band radio signal to the planet. Radar echoes from the planet surface are reflected back to the Earth and received at several large telescopes. The signal exhibits spatial irregularities in the wave front caused by the constructive and destructive interference of waves scattered by the irregular surface. Because of the rotation of the planet, the irregularities in the wave front, also called speckles, sweep over the different receiving stations with time (see Figure 8.21). As the radar speckle patterns are tied to the surface rotation, the observations determine the rotation of the top surface layer.
10.2. Spacecraft Observation

Rotation observations can be performed by using spacecraft orbiting a planet and even by spacecraft during flybys. Several methods can be used to determine the rotation. For example, as the orbit of a spacecraft is sensitive to the gravity of the planet and as the planet with its mass anomalies rotates and the spacecraft moves, the spacecraft is sensitive to the planet rotation as well. In order to determine the gravity field and the planet rotation, one uses the radio link between the spacecraft and the Earth. A radio signal is sent from Earth to the spacecraft and sent back by a transponder on the spacecraft to Earth after a coherent turnaround. One measures on Earth the Doppler shift on the radio signal in order to determine the relative velocity between the Earth and the spacecraft. Knowing the Earth rotation and orientation almost perfectly in space (at centimeter level from VLBI measurements), one determines the spacecraft velocity, position, and orbit.

Successive images taken from the spacecraft of the surface of the planet can also be used to determine the relative changes between the images with respect to space. This obviously requires good knowledge of the position of the spacecraft, which can be determined from radiosonde. The ESA BepiColombo mission to Mercury, in particular, will be able to determine the libration of Mercury in this way. This rotation experiment is represented in Movie 8.

Supplementary video related to this chapter can be found at http://dx.doi.org/10.1016/B978-0-12-415845-0.00008-6

A very valuable way of observing the rotation, the rotation variations, and the orientation changes is to use landers (or rovers when they are fixed for a long period). A two-way direct radio link from the Earth to the lander provides the relative velocity of the lander with respect to the ground stations on Earth and therewith, provides, again knowing were the Earth is in space, the orientation and rotation of the Moon or planet as a function of time. The geometry coverage of the experiment and the precision that can be reached with such observations are such that this is the best method to get precession and nutation; in particular, it is and will be used for Mars in the future (see Figure 8.22).

The radio links are normally in S-band, X-band or Ka-band. The X-band is mostly used and preferred with respect to the S-band as it is less sensitive to perturbations from the ionosphere and plasma. The ground stations on the Earth are quite big; they have a diameter of up to 70 m (see Figure 8.23). NASA has installed DSN antennas (Deep Space Network), and European Space Agency, ESTRACK stations (ESA TRACKing station) to communicate with spacecraft.

For measurements of the rotation of the Moon, one also uses ranging by lasers (Lunar Laser Ranging, LLR). In this method, a laser station on the Earth (see Figure 8.25) sends...
a laser beam to the Moon. The beam is reflected back to the Earth with retroreflectors (corner cubes reflecting the laser pulses back in the same direction as the received direction, see Figure 8.24). The Apollo missions have deposited several retroreflectors (Figure 8.24) on the surface of the Moon (locations in Figure 8.26), which are still used for determining the Moon–Earth distance and the librations of the Moon.

### 10.3. The Particular Case of the Earth

The Earth is a particular case as we are able to build relatively large equipment such as VLBI antennas, LLR stations, and satellite laser ranging (SLR) instruments. LLR is a technique explained in the previous paragraph used mainly for obtaining the distance between the Earth and the Moon, and SLR is essentially used for computing the gravity field of the Earth. VLBI is a technique based on multiple radio astronomy telescopes on Earth, at which the signals from very distant astronomical radio sources, such as quasars, are collected simultaneously and processed. The distances and changes of distance between the radio telescopes are then computed using the time differences between the arrivals of the radio signal at the different telescopes. This allows determining the rotation of the Earth and the Earth’s orientation in space, since the radio sources are essentially fixed in space on account of their large distance (or with a very small and well-determined proper motion). This technique is very precise as it also uses very precise frequency references (hydrogen maser), thereby providing the orientation of the Earth at the sub-centimeter level.
In parallel, there is a well-known geodetic technique used on the Earth: the Global Navigation Satellite System (GNSS). It provides positions of receivers with respect to satellites orbiting around the Earth with a very high precision (below the centimeter for geodesic receivers). As the satellites are orbiting in space and the receivers are on the Earth’s surface, this in turn can provide additional information on the Earth’s rotation and orientation. And, as polar motion expresses the variation of the rotation axis with respect to the figure axis of the Earth, GNSS also provides information on polar motion. Moreover, we have high spatial coverage of measuring GNSS devices, which ensures a precise determination of polar motion, also at the subcentimeter level.

**BIBLIOGRAPHY**


1. INTRODUCTION

The evolution of the interior of terrestrial bodies is the result of the balance between the heat accumulated during planetary formation and generated by slowly decaying radioactive isotopes present in mantle rocks and the heat that is lost and radiated to space over billions of years. Understanding this balance requires the characterization of the processes that control the way heat is transported from the core and mantle to the surface.

At the extreme temperature and stress levels of deep interiors, the thermally activated migration of crystalline defects allows rocky materials to flow over geological timescales like extremely viscous liquids. Powered by internal heat sources and heat from the core, terrestrial mantles undergo convection in response to buoyancy forces due to temperature variations, which cause thermal expansion and contraction. Convection transports thermal energy very efficiently and ultimately controls planetary cooling.

While the Earth’s surface is fragmented into plates whose slow movement away from midocean ridges toward convergent margins is a direct expression of the underlying convection, all other terrestrial bodies of the solar system are characterized by a single immobile plate. Convection occurs beneath a thick stagnant lid, which isolates the mantle retarding its cooling, and experiences deformation to a much lower extent than the Earth’s surface. The thermal history of planetary bodies crucially depends on whether the cold upper layers are mobile and participate in convection, as for the Earth with plate tectonics, or stagnant and separated from the interior, as for Mercury, Venus, Mars, and the Moon. The consequences of mantle convection are multiple and profound: volcanism, the generation of magnetic fields via dynamo action, or the formation of the tectonic landforms that shape planetary surfaces and crusts all involve physical processes intimately related to convective heat transport at depth.

Apart from the notable exception of seismic tomography (see Chapter 21), which provides at present a unique window on the structure of the Earth’s mantle, the depths of planets and moons remain largely inaccessible to direct inspection. The study of mantle rocks at high-pressure and high-temperature conditions combined with observational constraints on the composition, geology, magnetization, and thermal state of planetary surfaces provide a framework upon which theoretical models of mantle convection can be built. These models are a fundamental tool to characterize the numerous processes that govern the dynamics of planetary interiors and influence their evolution.

Whether or not life has influenced the evolution of the interior of the Earth to an extent even remotely rivaling its effect on the evolution of the atmosphere and oceans (see Chapter 20) is presently a matter of debate. With plate...
tectonics cycling matter between the surface reservoirs and the interior, there is a venue for cycling biologically altered rock and volatiles with the interior where these may physically and chemically interact with mantle rock. As the concentration of volatiles, water in particular, may have profound effects on the mantle rheology, and thus on the vigor of the convective engine, there is a potential venue for life having a profound effect on tectonics.

2. FORMATION AND EARLY EVOLUTION OF TERRESTRIAL BODIES

The terrestrial bodies of the solar system are the end products of a complex process of accretion (see Chapter 2). The gravitational collapse that led to the formation of the Sun around 4.5 billion years ago was accompanied by the generation of a disk of cold gas and dust—the solar nebula. This so-called protoplanetary disk, which orbited the newly born Sun owing to the conservation of angular momentum, was the source of the materials from which all planetary bodies formed. Although the existence of the solar nebula can only be inferred indirectly from the physical and chemical evidence left in the solar system, protoplanetary disks orbiting young stars are now routinely detected via infrared- and radioastronomy, which strongly substantiate the hypothesis that the solar system itself developed from similar conditions.

The accretion of planetary bodies consists of three main phases. First, the collective gravity of localized swarms of centimeter-sized dust particles leads to the formation of planetesimals—solid bodies ranging in size from about 1 to 100 km. Once initiated, this process is expected to take place rapidly (less than about $10^5$ years). However, the occurrence of sizable particle concentrations in the disk can be sporadic both in space and time, with the consequence that planetesimal formation can proceed over a relatively extended period of a few million years. Afterward, within $10^3$–$10^6$ years, planetesimals undergo a runaway growth (also called oligarchic growth), with large bodies growing faster than small ones until few tens of planetary embryos of lunar to Martian mass are formed. Finally, over a timescale of tens of millions of years, relatively rare collisions between embryos of comparable size, and between embryos and leftover planetesimals, give rise to a small number of terrestrial planets.

Intimately related to the formation of terrestrial bodies, from small asteroids to satellites and planets, is their differentiation into a metallic core enriched in iron and a rocky mantle consisting of silicate compounds (Figure 9.1). Core—mantle differentiation, which occurred contemporary with or shortly following the accretion, is a simple consequence of the larger density of iron and its alloys with respect to silicates. The most stable configuration of a rotating mass subject to its own gravity is in fact an oblate spheroid with the material of highest density located at its center. Albeit conceptually simple, this picture is complicated by problems related to the availability of the materials involved in the differentiation process and to the conditions and time at which these were added to the accreting mass.

The separation of solid metal from solid silicate due to gravitational instability is too slow to have played an important role during accretion. Therefore, core—mantle differentiation requires some form of melting. Core materials, having a lower melting temperature than silicates, can coexist as a liquid phase with a solid silicate matrix and sink through it via grain-scale percolation or in the form of large molten diapirs descending through fractures and dykes. Alternatively, metal—silicate separation can take place if both phases are in a liquid state. The feasibility of core formation requires thus an understanding of whether and how, during accretion, it was possible to reach temperatures large enough to cause melting of metallic and silicate materials.

Three main processes have been recognized as capable to provide a sufficient amount of energy to produce widespread melting during the early history of terrestrial bodies. First, the radioactive isotopes $^{26}$Al and $^{60}$Fe, known to be abundant in the solar nebula from geochemical analyses of
primitive meteorites, have short half-lives (see Table 9.1). Because of the energy released by the decay of these radionuclides, planetesimals that accreted within the first few million years of the solar system could have experienced the degree of melting required for core–mantle differentiation. As a consequence, at least part of the planetesimals that contributed to the formation of embryos first, and planets later, were likely already differentiated, and their cores built up the larger cores of the bodies to which they accreted. Second, it is important to recognize that the late stages of accretion involving impacts between large bodies resulted in strongly increased temperatures due to the conversion of kinetic and gravitational energy into heat. The upper bound to the total heat of accretion can be calculated based on the gravitational binding energy $E$, the energy required to pull apart to infinity all the accreted material, or, from the opposite point of view that is relevant here, the amount of energy liberated upon accretion from material pulled together from infinity:

$$E = \frac{3}{5} GM^2 \frac{1}{R},$$

(9.1)

where $G$ is the universal gravitational constant and $M$ and $R$ are the total mass and radius of the accreted body, respectively. The temperature change $\Delta T$ associated with the energy $E$ is

$$\Delta T = \frac{E}{M c_p},$$

(9.2)

where $c_p$ is the heat capacity ($\approx 10^3$ J/kg/K for silicate minerals and $\approx 800$ J/kg/K for metallic core materials). According to Eqn (9.2), if all accretionary energy were delivered instantaneously, the resulting temperature rise would be $\approx 10^3 - 10^4$ K for bodies ranging in size and mass from Mercury to Earth. From Figure 9.2, which shows the solidus and liquidus curves for the Earth’s mantle, it can be quickly realized that a temperature increase of a few thousand degrees would be sufficient to cause large-scale melting, possibly even partial vaporization, of the silicate materials. Indeed, it is now widely accepted that, because of violent impacts with embryos such as the one of Martian size from which the Moon probably originated (so-called giant impacts, Figure 9.3), terrestrial planets likely experienced one or multiple episodes of widespread melting that led to the formation of deep magma oceans. However, it should be also noted that accretionary heat is neither delivered through a single episode nor homogeneously. Also, depending on the size and velocity of the impactor, while part of its energy can be actually buried at depth and retained by the target body, part of it is also quickly radiated back to space. Nevertheless, even though it only gives a crude maximum heating estimate, the simple analysis above illustrates that the energy involved in the formation of planetary bodies was prodigious.

The third main source of energy that contributed to the early melting of planetary interiors is the process of core–mantle differentiation itself. No matter whether iron sinks toward the center of the planet through a fully or partially molten magma ocean, the associated redistribution of mass results in the release of gravitational potential energy in the form of additional heat. Let us consider, for example, a hypothetical Moon-forming impact involving

<table>
<thead>
<tr>
<th>Isotope</th>
<th>Specific Heat Production (W/kg)</th>
<th>Half-life (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$^{26}$Al</td>
<td>0.355</td>
<td>$7.2 \times 10^5$</td>
</tr>
<tr>
<td>$^{60}$Fe</td>
<td>0.068</td>
<td>$2.6 \times 10^6$</td>
</tr>
<tr>
<td>$^{238}$U</td>
<td>$9.46 \times 10^{-5}$</td>
<td>$4.47 \times 10^9$</td>
</tr>
<tr>
<td>$^{235}$U</td>
<td>$5.69 \times 10^{-4}$</td>
<td>$7.04 \times 10^8$</td>
</tr>
<tr>
<td>$^{232}$Th</td>
<td>$2.64 \times 10^{-5}$</td>
<td>$1.40 \times 10^{10}$</td>
</tr>
<tr>
<td>$^{40}$K</td>
<td>$2.92 \times 10^{-5}$</td>
<td>$1.25 \times 10^9$</td>
</tr>
</tbody>
</table>

Isotopes of Al and Fe have very short half-lives and were only important for internal heating of planetesimals during the early stages of the solar system. Isotopes of U, Th, and K have half-lives comparable with the age of the solar system and are relevant for the long-term evolution of planetary interiors.
the proto-Earth and an embryo having, respectively, a mass 80–90% and 10–20% of that of the present-day Earth. Assuming that both bodies already possess a core that occupies half of their radius and that the two cores will completely merge, the mean temperature change in the final core can be estimated to be ~1000–2000 K, sufficient to induce substantial melting of the lower mantle and to generate a strongly superheated core.

**FIGURE 9.3** Hydrodynamic simulation of the collision between two bodies of similar masses as an example of the Moon-forming impact. Time is shown in hours, distance in units of $10^3$ km, and temperature in Kelvin. The final planet is surrounded by a disk with very similar composition, from which the Moon would later originate. This simulation illustrates the extreme heating caused by energy transfer associated with large-scale impacts. *After Canup (2012).*
In synthesis, thanks to the heat due to the decay of short-lived radioactive isotopes as well as to the conversion of kinetic energy of impactors and of gravitational potential energy associated with core formation, near the end of the accretion period, terrestrial bodies were likely hot and chemically differentiated into a fully or partially molten silicate mantle and a liquid metallic core possibly superheated with respect to the overlying mantle. The time for reaching this stage is relatively short compared with the age of the solar system. It varies from a few million years for small asteroids and planets like Mars, to tens of millions of years for the Earth. This condensed timescale predicted by theoretical models of planetary accretion and differentiation is supported by the study of isotopic systems. In particular, during the past decade, the application of the Hf–W decay system to the available inventory of meteorites has repeatedly proved to be a very effective tool for timing various stages of the early evolution of rocky bodies (see Chapters 2 and 28).

The idea that planets were extremely hot shortly after their formation is not new. The famously wrong calculation of the age of the Earth made by William Thompson (better known as Lord Kelvin) in 1862 was based on the assumption that our planet was initially completely molten and that, after solidifying, reached its present temperature by cooling via conduction. After all, temperature measurements in mines suggested a geothermal gradient of 20–30 K/km, which clearly indicated that the Earth was cooling. Kelvin’s estimates that ranged from few ten to few hundred million years, however, were far too short because the continental crust, where the above temperature measurements were taken, is in a nearly steady-state balance as a consequence of internal heat production due to the decay of radiogenic isotopes and mantle heat flow from below (Section 3), both of which were ignored by Kelvin. The fact that the history of planetary bodies started from a hot state, possibly associated with the existence of vast regions of molten silicates, can have important consequences. The process of metal–silicate separation is certainly the most dramatic differentiation event in the history of a planet, but the solidification of a magma ocean can also represent a fundamental step in the development of the interior structure and composition. Although the existence of a terrestrial magma ocean is considered as highly probable, its evidence is not compelling, particularly because plate tectonics and mantle convection have contributed to erase all the traces that this event may have left at the surface. Nevertheless, a basal magma ocean, which crystallized slowly enough to allow localized patches of dense melt to survive until present, has been invoked to explain the existence of a deep primordial reservoir of incompatible species, which could account for the ultralow seismic shear wave velocities detected in some regions of the lowermost Earth’s mantle (see Chapter 21). Geological and geochemical arguments suggest that smaller bodies such as Mars or the Moon experienced a high degree of silicate melting early in their history. In this respect, the Moon provides the most notable and widely accepted example. The widespread presence of plagioclase minerals all over the Moon’s highlands (see Chapters 23 and 24) is best explained in the framework of a solidifying magma ocean. In fact, assuming its fractional crystallization, i.e., the formation upon cooling of specific rock compositions due to the segregation of certain elements in the liquid magma out of the solid phase, plagioclase, having a relatively low density, is positively buoyant and tends to rise to the surface where it solidifies and forms a primordial crust (see Section 6.1).

Completion of large-scale mantle solidification ideally defines the end of the cataclysmic events associated with the formation and early differentiation of terrestrial bodies. The subsequent evolution of their interiors is characterized by secular cooling, the slow and continuous loss of the primordial heat accumulated during accretion and core formation and of the heat generated by the decay of long-lived radiogenic isotopes present in the mantle.

3. SUBSOLIDUS CONVECTION

Shortly after the discovery of natural radioactivity in 1896 by Henri Becquerel, Pierre Curie noticed that a speck of radium continuously and spontaneously emits a vast amount of heat, much higher than that released in any known chemical reaction involving the same quantity of matter. Geologists quickly realized that the decay of radioactive elements could be responsible for heating the interior of the planet and in turn affect Kelvin’s estimate of the age of the Earth. Accurate dating techniques based on radioactive decay were rapidly developed and in 1911 the British geologist Arthur Holmes demonstrated that the most ancient rock sample from the large collection he examined was 1.6 billion years old. Holmes laid in this way the foundations of modern geochronology, although it was not until 1955 that the first meteorite was dated, making it finally possible to place the age of the Earth at 4.5 billion years.

Kelvin’s idea according to which the Earth was simply undergoing secular cooling was initially replaced by the concept of steady-state conductive balance. The heat flowing out of the Earth was still believed to be lost only by conduction, but also balanced by the heat generated by radioactive sources, which were thought to be concentrated exclusively in the crust overlying a solid interior depleted of heat producing elements (see Section 6.1). Using heat flow measurements from continental regions, this model could actually well predict typical concentrations of heat producing elements in continental rocks. The assumption that the heat sources were concentrated solely in the crust
also implied that the heat flow in oceanic regions, where the crust was known to be thinner than under continents (Figure 9.4(a)), would be much lower. However, the first measurements carried out in the 1950s in the Pacific and the Atlantic showed that the oceanic heat flow away from midocean ridges is very similar to that measured in continental regions (Figure 9.4(b)). This surprising equivalence was then attributed to the existence of convection currents in the mantle. The high thermal gradients observed near the Earth’s surface are the expression of the upper thermal boundary layer associated with mantle convection (Figure 9.7). Across this layer, heat is transported by thermal diffusion (i.e. by conduction), while transport by convection dominates beneath it, with the bulk temperature, well below its melting value, increasing with depth along a nearly adiabatic gradient (see Figure 9.2).

The process of thermal (or natural) convection refers to the transport of heat due to bulk fluid motion. Although the

FIGURE 9.4 Global maps of the Earth’s crustal thickness from a compilation of seismic data (a) and surface heat flow combining heat flux measurements and a cooling model for the oceanic lithosphere (b). *After Jaupart and Mareschal (2011).*
mantle is by all means solid in that it propagates elastic waves on short timescales, on geological timescales of the order of millions of years it deforms like a highly viscous fluid in a way similar to the more familiar one with which the ice in glaciers flows on timescales of years. One way to study the fluid behavior of the Earth, which led to the first empirical estimates of rock viscosity in the 1930s, is to investigate the viscous response of the mantle to the last deglaciation. The vast ice sheets that covered large portions of the northern hemisphere around 20,000 years ago depressed the Earth’s surface below sea level. As a consequence of their quite abrupt melting, the mantle has been responding by flowing to compensate the mass deficit caused by the disappearance of the ice load (Figure 9.5(a)).

The characteristic time \( \tau \) for the \textbf{viscous relaxation} of the surface topography associated with this process can be calculated as

\[
\tau = \frac{4\pi \eta}{\rho g \lambda}
\]  

(9.3)

where \( \eta \) is the viscosity, \( \rho \) is the density, \( g \) is the gravity acceleration, and \( \lambda \) is a characteristic spatial length of the surface load (i.e. the ice sheet). From observations of elevated beach terraces, it is possible to determine \( \tau \) and \( \lambda \). Assuming \( \rho = 3300 \text{ kg/m}^3 \), \( g = 9.8 \text{ m/s}^2 \), and \( \tau = 4400 \text{ year} \) and \( \lambda = 3000 \text{ km} \) as based on the data of Figure 9.5(b), we obtain \( \eta = 1.1 \times 10^{21} \text{ Pa s} \) for the viscosity of the upper mantle. This value is enormous when compared with the viscosity of more ordinary materials. For example, at ambient conditions, water has a viscosity of \( \sim 10^{-3} \text{ Pa s} \), while the typical value for honey is \( \sim 10 \text{ Pa s} \) and that for ice is \( \sim 10^{13} \text{ Pa s} \).

Mantle rocks, similarly as they deform in response to surface loading, also flow at subsolidus temperatures as a consequence of temperature differences, thereby advecting heat. An increase in the temperature of a fluid parcel produces a reduction in density due to volumetric thermal expansion. If the fluid is subject to a gravitational field, buoyancy forces arise that tend to lift the heated fluid (the opposite argument clearly applies in the presence of a decrease in temperature). Let us consider a homogeneous plane layer of thickness \( D \) whose upper and lower surfaces are maintained at constant temperatures \( T_0 \) and \( T_1 \), respectively, with \( T_1 > T_0 \). Fluid near the hotter lower boundary will become lighter than the overlying fluid and tend to rise, while fluid near the colder upper boundary will become heavier and tend to sink. By performing an analysis of \textbf{linear stability} of this system, it can be established under which conditions small disturbances will decay and lead to a stable configuration, or grow by overcoming the viscous resistance of the fluid and lead instead to a motion of finite amplitude. Only one nondimensional parameter is sufficient to characterize the stability of a fluid layer heated from below and cooled from above. This parameter is the \textbf{Rayleigh number}:

\[
Ra = \frac{\rho^2 c_p g (T_1 - T_0) \alpha D^3}{k \eta},
\]  

(9.4)

where \( \rho \), \( \eta \), and \( g \) have been defined above; \( c_p \) is the isobaric heat capacity; \( \alpha \) is the coefficient of thermal expansion; and \( k \) is the thermal conductivity. If \( Ra \) is greater than a certain critical threshold \( Ra_{ct} \), thermal convection will set in. Figure 9.6 shows this critical Rayleigh number as a function of the nondimensional wave number (i.e. the inverse of the wavelength) of the applied perturbation. At a wave number of \( \pi \sqrt{2}/2 \), the curve attains its minimum

\FIGURE 9.5 (a) Surface subsidence due to loading of an ice sheet of characteristic wavelength \( \lambda \) and subsequent viscous uplift following rapid ice melting. (b) Data points for the uplift \( h \) of the mouth of the Angerman River, Sweden, as a function of time before present compared with an exponential relaxation model of the kind \( h \sim \exp(-t/\tau) \), with \( \tau \) defined as in Eqn (9.3). After Turcotte and Schubert (2002).
convection in this case is 867.8. Also note that although the above estimates are only valid for a fluid enclosed in a horizontal layer, in the case of spherical shells, the critical Rayleigh number for both bottom and internally heated convection is not much different, being of the order of $\sim 10^7$. Essentially on the grounds of these analyses, Arthur Holmes proposed as early as in 1931, well before the acceptance of the plate tectonics paradigm in the 1960s, that the Earth’s interior experiences solid-state convection. Indeed, by setting in Eqn (9.5) $D = 2890$ km (the thickness of the Earth’s mantle), $g = 9.8$ m/s$^2$, $\rho = 4000$ kg/m$^3$, $c_p = 1200$ J/kg, $\alpha = 3 \times 10^{-7}$ K$^{-1}$, $k = 3$ W/m/K, $H = 10^{-11}$ W/kg (see Eqn (9.6) and Table 9.1), and $\eta = 10^{21}$ Pa s as derived from the analysis of postglacial uplift, we obtain $Ra_H \sim 10^7$. This value is much greater than the critical threshold predicted by the theory and thus indicates that the mantle of the Earth is in a highly supercritical regime and undergoes vigorous thermal convection.

A similar argument applies to the other terrestrial bodies of the solar system (Mercury, Mars, Venus, and the Moon) whose mantle is also thought to undergo thermal convection. Internal heating is provided by the decay of the long-lived radioactive isotopes of uranium, thorium, and potassium whose specific heat production rates and half-lives are reported in Table 9.1. Their relative concentration in the mantle can vary significantly among different planets. Nevertheless, an order-of-magnitude estimate can be made considering the relative abundance of heat producing elements as measured for chondrites, i.e. stony meteorites that have not been modified by melting and differentiation and that are thought to form the building blocks of terrestrial planets. In one of the
most primitive type, the carbonaceous Ivuna (CI), radiogenic isotopes have the following proportions: \(7.4 \times 10^{-9}\) kg/kg uranium, \(29 \times 10^{-9}\) kg/kg thorium, and \(550 \times 10^{-6}\) kg/kg potassium. Radioactive substances decay exponentially according to their time constant and the heat production as a function of time is given by

\[
H(t) = \sum_i C_i H_i \exp\left(\frac{-\ln(2)t}{\tau_i}\right), \tag{9.6}
\]

where \(C_i\) is the concentration of the \(i\)th element and \(H_i\) and \(\tau_i\) are the corresponding heat production rate and half-life, respectively. Using Eqn (9.6), we find that the heat production of a CI chondritic body would vary between \(2.4 \times 10^{-11}\) W/kg 4.5 Ga ago and \(0.34 \times 10^{-11}\) W/kg today.

All terrestrial planets, with the possible exception of Mercury, have a mantle characterized by a supercritical Rayleigh number, which is indicative of a convecting interior. In Table 9.2, we report ranges of possible Rayleigh numbers for the terrestrial planets and the Moon. On the one hand, differences in the Rayleigh number among different bodies are primarily due to the thickness of their mantles, which varies from \(\sim 400\) km for Mercury to \(2890\) km for the Earth, and scales with the third power for different bodies are primarily due to the thickness of their interior. In Table 9.2, we report ranges of possible Rayleigh numbers for the terrestrial planets and the Moon. On the one hand, differences in the Rayleigh number among different bodies are primarily due to the thickness of their mantles, which varies from \(\sim 400\) km for Mercury to \(2890\) km for the Earth, and scales with the third power for bottom-heated convection (Eqn (9.4)) or with the fifth power for internally heated convection (Eqn (9.5)). On the other hand, for a given body, the ranges reported in the table are the consequence of a rather poor knowledge of the viscosity (see Section 4).

### Table 9.2 Order-of-Magnitude Estimate of the Reference Rayleigh Number for the Terrestrial Planets and the Moon

<table>
<thead>
<tr>
<th>Planetary Body</th>
<th>Radius (km)</th>
<th>Outer Core Radius (km)</th>
<th>Rayleigh Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
<td>2440</td>
<td>1940–2140</td>
<td>Subcritical–10^5</td>
</tr>
<tr>
<td>Venus</td>
<td>6052</td>
<td>3089</td>
<td>10^7–10^9</td>
</tr>
<tr>
<td>Earth</td>
<td>6371</td>
<td>3480</td>
<td>10^7–10^9</td>
</tr>
<tr>
<td>Moon</td>
<td>1737</td>
<td>150–400</td>
<td>10^5–10^8</td>
</tr>
<tr>
<td>Mars</td>
<td>3390</td>
<td>1400–1900</td>
<td>10^6–10^8</td>
</tr>
</tbody>
</table>

Rayleigh numbers are calculated from Eqn (9.5) using the following parameters: \(\rho \sim 10^3\) kg/m\(^3\), \(c_p \sim 10^3\) J/kg K, \(\gamma \sim 1\) m/s\(^2\), \(a \sim 10^{-5}\) K, \(H \sim 10^{-11}\) W/kg, \(k \sim 1\) W/m K, \(\eta \sim 10^{19}–10^{21}\) Pa s (see Table 9.3), and mantle thickness \(D\) calculated as the difference between the radius of a given body and the radius of its core. Note that, apart from the Earth, the lack of seismic observations and uncertainties in the gravity field measurements from which interior structure models can be derived, the size of the outer core of the other terrestrial bodies is not well constrained. This is particularly relevant for Venus whose core size was simply obtained by rescaling Earth’s parameters.

Core radii are from Breuer, et al. (2010). Thermal evolution and magnetic field generation in terrestrial planets and satellites, Space Science Reviews, 152, 449–500.

### 4. Rock Rheology and Modes of Convection

The way rocks deform in response to applied stresses (i.e. their rheology) depends on several factors such as rock composition, temperature, pressure, and level of deviatoric stress (the portion of the stress field responsible for actual distortions in a continuum as opposed to isotropic volume changes caused by the mean stress, i.e. the pressure). In the crust and the shallow mantle, rocks undergo primarily elastic and brittle deformation, while at the higher pressures and temperatures characteristic of the deep interior, ductile behavior becomes dominant.

At low stresses, the relation between stress and strain is linear and solids behave elastically (the strain is a measure of deformation that accounts for the relative displacement between particles in a stressed body). Upon compressing or placing under tension a crystalline material, interatomic forces tend to resist compression or expansion and to keep atoms in their lattice position. This elastic deformation is completely reversible. As the stress level increases, a critical value known as yield stress can be reached. Above this value, deformation becomes irreversible and is accommodated through brittle failure or ductile flow (Figure 9.8). The former manifests itself either through the creation of new cracks and faults, or through frictional sliding along existing fractures. The latter consists instead in a fluidlike motion characterized by an effective viscosity that controls how temporal variations of the strain (strain rates) vary as a function of applied stresses.

Viscous flow is the principal mode of subsolidus deformation in planetary mantles and takes place via two main mechanisms: diffusion and dislocation creep. These mechanisms control the thermally activated migration of crystalline defects. In diffusion creep, which dominates at relatively low stresses, atomic vacancies, i.e. empty sites in the crystal lattice, can move in response to applied stresses within grains (Herring—Nabarro creep), or along grain boundaries (Coble creep). In dislocation creep, which becomes relevant at higher stresses, deformation is due to the motion of one-dimensional linear imperfections of the lattice (so-called dislocations). Both theory and laboratory experiments indicate that the dependence of the strain rate \(\dot{\varepsilon}\) on the applied stress \(\sigma\) can be expressed in the following form, valid for both diffusion and dislocation creep:

\[
\dot{\varepsilon} = A \sigma^n \exp \left(-\frac{E_a + V_a P}{RT}\right), \tag{9.7}
\]

where \(A\) is a material prefactor, \(n\) is the stress exponent, \(P\) is the hydrostatic pressure, \(T\) is the absolute temperature, \(R\) is the universal gas constant, \(E_a\) is the activation energy, and \(V_a\) is the activation volume. In diffusion creep the relationship between stress and strain rate is linear (Newtonian).
and \( n = 1 \). Dislocation creep is instead a nonlinear (non-Newtonian) deformation mechanism with \( n > 1 \) (typically \( n = 3.5 \)). In the exponential term of \( \text{Eqn (9.7)} \) (Arrhenius law), the activation enthalpy \( H_a \equiv E_a + V_a P \) accounts, through the term \( E_a \), for the energy necessary to form vacancies and for the energy barrier that atoms must overcome to migrate into a vacancy site, and, through the term \( V_a P \), for the fact that pressure tends to render these two processes more difficult. Introducing the viscosity \( \eta \) as the proportionality factor relating stress and strain rate, i.e. \( \tau = 2\eta \dot{\varepsilon} \), we have from \( \text{Eqn (9.7)} \):

\[
\eta = \frac{1}{2} A^{1/n} e^{(1-n)/n} \exp\left(\frac{E_a + V_a P}{nRT}\right). \tag{9.8}
\]

Although \( \text{Eqn (9.8)} \) depends explicitly only on strain rate, pressure, and temperature, the prefactor \( A \) is itself a function of several parameters that can change the viscosity by several orders of magnitude. The most significant are grain size (relevant for diffusion creep only), melt fraction, and concentration of water. Large grains tend to increase the viscosity, while partial melt and water content tend to decrease it. \( \text{Table 9.3} \) lists numerical values of the parameters that appear in \( \text{Eqn (9.8)} \) and the corresponding viscosity calculated at a reference temperature of 1600 K and at a pressure of 3 GPa for water-free (dry) and water-saturated (wet) olivine, the most abundant mineral of the Earth’s upper mantle. The effect of water is large: its presence can reduce the effective viscosity almost by two orders of magnitude. It should be also noted that the uncertainties in the parameters that enter the rheological relation (\( \text{Eqn (9.8)} \)) are also large. Laboratory experiments in which these parameters can be determined are particularly difficult. In order for deformations to be accurately measured, the strain rates considered in the laboratory are typically orders of magnitude greater than those expected in planetary interiors, with the result that significant extrapolations must be applied.

Diffusion and dislocation creep occur simultaneously. However, at given conditions of temperature, pressure, and strain rate, the effective viscosity of a stressed material is determined by the weakest mechanism, i.e. the one that delivers the smallest viscosity. In \( \text{Figure 9.9} \) we show a map of viscosity as a function of temperature for dry olivine obtained using parameters from \( \text{Table 9.3} \) and different strain rates, from \( 10^{-10} \) to \( 10^{-14} \), as appropriate for the slow creeping flow of mantle rocks. Diffusion creep tends to prevail at low strain rates and relatively high temperatures, while, as the strain rate is raised, the effective viscosity associated with dislocation creep dominates over the whole temperature range considered. The diagram also illustrates the large sensitivity of viscosity to temperature changes. For example, for diffusion creep, a decrease in temperature from 1200 to 1000 K causes an increase in viscosity of more than two orders of magnitude.

At the relatively low temperatures characteristic of planetary surfaces and shallow interiors, the exponential dependence in \( \text{Eqn (9.8)} \) causes the viscosity to become so large that ductile flow is no longer possible. At these conditions, terrestrial bodies develop a so-called stagnant lid, an immobile upper layer that cannot participate in mantle convection. This situation is the most common in the solar

<table>
<thead>
<tr>
<th>Diffusion Creep</th>
<th>Dislocation Creep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dry ( (\text{Pa} \cdot \text{s}) )</td>
<td>1.9 \times 10^{-11}</td>
</tr>
<tr>
<td>Wet ( (\text{Pa} \cdot \text{s}) )</td>
<td>2.4 \times 10^{-16}</td>
</tr>
<tr>
<td>( n )</td>
<td>1</td>
</tr>
<tr>
<td>( E_a ) (kJ/mol)</td>
<td>3 \times 10^5</td>
</tr>
<tr>
<td>( V_a ) (m^3/mol)</td>
<td>6 \times 10^{-6}</td>
</tr>
<tr>
<td>( \eta ) (Pa s)</td>
<td>6.3 \times 10^{20}</td>
</tr>
</tbody>
</table>

The prefactor \( A \) was calculated assuming a shear modulus of 80 GPa, grain size of \( 10^{-3} \) m, Burgers vector’s length of \( 5 \times 10^{-19} \) m. To calculate the viscosity from \( \text{Eqn (9.8)} \), we also assumed reference temperature and pressure of 1600 K and 3 GPa, respectively, and a strain rate of \( 10^{-14} \). Parameters after Kohlstedt et al. (1995), Rheology of the upper mantle: A synthesis. Science 260, 771–778.
system. Little doubts exist that Mercury, Mars, and the Moon have been one-plate bodies throughout all or most of their history. Although Venus, because of its young surface (Chapter 15), is generally believed to have experienced an episodic regime in which short events of surface mobilization are interspersed between long phases of quiescence associated with a stagnant lid, at present it also possesses an immobile surface. The Earth is the only planetary body in a mobile lid regime with a surface split in tectonic plates that are an active part of the mantle convection system. The consequences of this difference in tectonic modes on the dynamics and evolution of the interior are profound. A qualitative glimpse of the impact that a stagnant lid regime has on mantle convection can be obtained from Figure 9.10, which shows the temperature field from numerical simulations of an Earth-like planet in stagnant (Figure 9.10(a)) and mobile lid regimes (Figure 9.10(b)) (see also Section 5). In both cases, we assumed the rheology to be only governed by diffusion creep of dry olivine with temperature- and pressure-dependent viscosity. In the first case, the upper layers are immobile because of their high viscosity. Convection is driven primarily by bottom boundary layer instabilities that result in upwelling plumes, and, to a lesser extent, by sinking downwellings originating from the bottom part of the lid, which are only slightly colder than the surrounding mantle. The stagnant lid clearly acts as a thermal insulator that tends to keep the interior of the planet warm.

In order for surface mobilization to be possible, other deformation mechanisms must be taken into account. A convective behavior like the one portrayed in Figure 9.10(b) results from considering, in addition to diffusion and/or dislocation creep, a mechanism of brittle failure. Recall from the discussion at the beginning of this

**FIGURE 9.9** Viscosity as a function of temperature for dry olivine calculated from the parameters of Table 9.3. The solid line refers to the viscosity of diffusion creep, while the dashed lines to the viscosity of dislocation creep computed for different values of the strain rate $\dot{\varepsilon}$. At a given temperature, the effective viscosity is that of the weakest mechanism. For example, with $\dot{\varepsilon} = 10^{-18}/s$, the material will deform via dislocation creep for $T < 1200$ K and via diffusion creep for $T > 1200$ K.

**FIGURE 9.10** Temperature distribution from numerical simulations of thermal convection in a two-dimensional cylindrical geometry. (a) Stagnant lid and (b) mobile lid regime. In both, cases the system is cooled from above and heated from both below and within. In the first, the uppermost cold layers are immobile because of their high viscosity and exert a blanketing effect that tends to keep the interior warm. In the second, the introduction of an yielding mechanism allows for surface mobilization and subduction, which in turn causes a strong cooling of the interior. Directions of flow velocity are indicated by white arrows whose length is proportional to the magnitude of the velocity vector. The central region is occupied by the core, which is not modeled. *Simulations courtesy of Hättig C., DLR Berlin.*
section that deformation becomes irreversible when the yield stress is exceeded. The brittle nature of rocks at relatively low pressures at which fracture or frictional sliding can occur can be described with the Byerlee’s law:

$$\sigma_y = C + \mu P; \quad (9.9)$$

where $\sigma_y$ is the yield stress, $C$ is the cohesive strength, $\mu$ is the friction coefficient, and $P$ is the hydrostatic pressure. Remembering now that the effective viscosity $\eta_{\text{eff}}$ is controlled by the weakest mechanism, we can write

$$\eta_{\text{eff}} = \left( \frac{1}{\eta_{\text{diff}}} + \frac{1}{\eta_{\text{disl}}} + \frac{1}{\eta_y} \right)^{-1}, \quad (9.10)$$

where $\eta_{\text{diff}}$ and $\eta_{\text{disl}}$ are diffusion and dislocation creep viscosities as calculated from Eqn (9.8) and $\eta_y \equiv \sigma_y/2\dot{\varepsilon}$ is a plastic viscosity associated with the yielding mechanism (Eqn (9.9)). When stresses generated by mantle convection are large enough, the stagnant lid can locally fail. The effective viscosity (Eqn (9.10)) drops, making it possible to strongly localize deformation in regions that serve as nucleation points for the initiation of subduction.

Figure 9.10(b) was obtained using a rheological model accounting for diffusion creep and brittle deformation (see also Figure 9.11(a)). Two important features emerge from the resulting mode of mobile lid convection that render it remarkably different from the stagnant lid mode. First, the mobilization and accompanying foundering of the cold thermal boundary layer is a significantly stronger driver of convection than the sublid instabilities arising in the presence of a stagnant lid. Indeed, it is widely recognized that Earth’s mantle convection is driven primarily by buoyancy forces associated with the subduction of oceanic plates. Second, the fact that the cold upper thermal boundary layer actively participates in convection yields a continuous cooling of the interior as opposed to the stagnant lid case in which the mantle tends to remain homogeneously hot and to cool largely because of the decay of radioactive heat sources. As we shall see in the following section, the ability of a planet to cool more effectively has important consequences for the evolution of the interior and its geophysical implications.

![FIGURE 9.11](image-url) (a) Numerical simulations of mantle convection with plate tectonics in a three-dimensional spherical shell. The top line shows two views of the viscosity field at the surface. Rigid plates moving with uniform velocity are separated by both divergent and convergent narrow boundaries with low viscosity at which plates form and are recycled into the mantle. The bottom line shows temperature isosurfaces representing the subduction of cold downwellings. (After van Heck & Tackley (2008).) (b) Numerical simulation of the formation of Mars’ crustal dichotomy by mantle convection. The top panel shows a hot upwelling plume (yellow) rising from the core—mantle boundary (red) able to generate a large region of partial melt (light blue). Migration of the melt toward the surface leads to the formation of a crust (Section 6.1) whose thickness in the southern hemisphere is much greater than in the northern one (bottom panel). After Sramek & Zhong (2012).
5. MODELING INTERIOR DYNAMICS AND EVOLUTION

Heat transport by conduction and subsolidus convection are the main physical processes underlying the dynamics and thermal evolution of terrestrial planetary interiors. Plate tectonics, which has been operating on the Earth for billions of years; magma production and volcanism, which have shaped the surfaces of all terrestrial bodies; and the generation of a magnetic field, currently operating on the Earth and Mercury only but likely active in a remote past also on Mars and the Moon, are all ultimately related to the way heat is transported in the interior.

Since solid mantles exhibit a fluidlike behavior over geological timescales, their dynamics is governed by the same set of fundamental equations employed for describing the motion of more ordinary fluids like, for example, water. A crucial difference is due to the large viscosity of rocks. Because of it, mantle materials respond instantaneously to applied stresses, causing viscous forces, which tend to resist motion, to be constantly in equilibrium with the buoyancy forces that drive it. As a consequence, inertial forces are negligible and convective flows in the mantle are always laminar, as opposed to the turbulent flows that frequently characterize convection in fluids with low viscosity, such as flows in the ocean and atmosphere. Note however, that at sufficiently high Rayleigh numbers (see Figure 9.7 and Table 9.2), mantle flow, albeit laminar, is highly time dependent, so that small variations in initial conditions are able to induce large differences in the flow at subsequent times, which renders mantle convection an intrinsically chaotic process.

In general, modeling mantle convection requires to solve the conservation equations of mass, linear momentum, thermal energy, and chemical composition, appropriate for viscous media with negligible inertia. These build a set of coupled partial differential equations whose analytical solution is only available under a handful of highly simplifying assumptions. Solutions to these equations at conditions relevant for planetary interiors are usually based on numerical models, which represent today the primary tool for investigating mantle dynamics and evolution. The conservation equations must be supplemented by a suitable equation of state relating density to temperature and chemical composition, by a rheological law describing the viscosity field (see Eqns (9.8) and (9.10)), and by a series of material parameters (e.g. coefficients of thermal expansion and conduction, heat capacity, etc.) that in first approximation can be considered constant but, in general, like the viscosity, are themselves functions of pressure, temperature, and composition. Since the atmosphere and the (liquid) outer core do not exert any significant traction on the solid mantle, it is usually assumed that shear stresses vanish at the planetary surface and at the core—mantle boundary. The average surface temperature of a planetary body is controlled by the atmosphere, or the lack thereof, and remains approximately constant throughout the evolution governed by solid-state convection, although Venus may represent an exception in this sense (see Chapter 15). The bottom mantle temperature at the interface with the core, instead, can change considerably over long timescales. In first approximation, in fact, a liquid core underlying the mantle can be treated as an isothermal heat bath of given density and heat capacity from which as much heat can be extracted as required by the heat flowing through the bottom thermal boundary layer associated with the convecting mantle (see Figure 9.7(d)). In addition, heat production decreases with time because of the decay of long-lived radioactive isotopes (Eqn (9.6)), ultimately causing the mantle to cool over billions of years of evolution.

In recent years, numerical simulations have witnessed dramatic progress thanks to the steadily increasing availability of large computational resources. Dynamic approaches based on the solution of the full set of conservation equations of mantle convection allow now for the treatment of complex problems in three-dimensional spherical shell geometry. For example, self-consistent models of Earth’s plate tectonics on the sphere have started to appear (Figure 9.11(a)). Or, in the framework of Mars’ interior dynamics, theories based on numerical simulations have been put forth supporting the idea that its hemispheric dichotomy, the strong difference in the thickness of the crust beneath the smooth northern lowlands and the rugged southern highlands (see Chapter 17), is the result of an endogenic process associated with a peculiar plume of mantle convection (Figure 9.11(b)).

Nevertheless, despite the relative easiness with which computing resources can be now accessed, the growing complexity of modern three-dimensional simulations presents significant challenges related both to technical difficulties associated with the development of adequate numerical methodologies and to the long calculation times that such simulations typically demand. For these reasons, a simpler approach based on so-called parametrized convection models is often adopted as a valuable alternative to fully dynamic simulations. These are one-dimensional models based on the solution of energy balance equations for the mantle and core. Instead of resolving the flow responsible for convective heat transport, appropriate scaling laws are employed to parameterize it. Being derived from experiments and numerical models, these scaling laws can describe the energetics of stagnant or mobile lid convection. Boundary layer theory is then used to determine the thickness of the thermal boundary layers across which heat is transported by conduction. In this way, the interior evolution can be obtained by constructing a radial thermal profile for the entire planet assuming that the temperature increases linearly in the boundary layers and adiabatically in the mantle and core.
Figure 9.12 illustrates a simple thermal evolution model calculated from parameters typical of Mars’ mantle and considering parameterizations of heat transport appropriate for stagnant lid convection, as expected for Mars, and for mobile lid or plate tectonics-like convection, for illustrative purposes. In the first case, the lid grows rapidly (Figure 9.12(b)) and the mantle temperature remains confined within a relatively narrow range (solid line in Figure 9.12(a)). On the contrary, in the plate tectonics case in which the lid can not grow but is continuously recycled into the mantle, cooling of the interior is much more efficient (see also Figure 9.10) and the mantle temperature exhibits a large decrease over time (solid line in Figure 9.12(a)). At the end of the evolution, the two temperatures differ by as much as \( \sim 600 \, \text{K} \), corresponding to a difference of \( 3 \times 10^{29} \, \text{J} \) in total energy loss between the two models (Figure 9.12(d)). The absence of a stable insulating layer in the plate tectonics case also leads to an initially higher, but steadily decreasing, surface heat flow with respect to the stagnant lid case. After about 2.5 Ga, however, heat flows become comparable and decrease with similar rates until today (Figure 9.12(c)).

An important characteristic of the evolution of vigorously convecting mantles with strongly temperature-dependent viscosity is that the thermal state and surface heat flow at present are only slightly dependent on the initial temperature distribution (Figure 9.13). In fact, any increase of the mantle temperature is accompanied by a reduction in mantle viscosity (see Eqn (9.8)) that increases convective vigor and leads to a more rapid heat loss. On the contrary, upon mantle cooling, the viscosity tends to increase, reducing convective vigor and rendering heat transfer less efficient. As a result of this feedback mechanism, small temperature changes can produce large variations in the heat flux, with the consequence that, after a sufficiently long time, the temperature is buffered at nearly constant values.

6. CONSTRAINTS ON AND MODELS OF THE EVOLUTION OF PLANETARY INTERIORS

The variety and complexity of physical and chemical processes, as well as the number of poorly known parameters that potentially affect the evolution of the interior over a wide range of spatial and temporal scales is extremely large. In order to understand the processes that govern the dynamics of the mantle and make inferences on the actual history or present-day status of planetary interiors, it is thus...
essential to integrate a number of observational constraints into numerical simulations. To this end, geophysical, geological, and geochemical evidence collected remotely or in situ by artificial satellites and lander spacecrafts, and, as far as the Earth is concerned, in the field or in the laboratory, needs to be considered in theoretical evolution models. In the following subsections, we discuss selected examples in which key aspects and processes of the evolution of the Earth and the other terrestrial bodies can be elucidated with the methods described in the previous section combined with different types of constraints.

6.1. Crust Formation and Volcanic History

The crust is the brittle layer that occupies the outermost part of terrestrial bodies (see Figure 9.1). When the mantle temperature rises above the solidus (see Figure 9.2), partial melt forms. At relatively low pressures (less than about 10 GPa), silicate melts are less dense than solids and can buoyantly percolate via porous flow or migrate upward through dykes and fractures over significantly shorter timescales (up to about $10^5$ years) than those that characterize subsolidus deformation ($\sim 10^6$ years). Subsequent cooling and solidification of the extracted magma, either at depth or following eruption at the surface, lead to the formation of the crust. The composition of the crust is representative of the degree of melting experienced by the parent material, and hence it contains hints on the thermal state of the mantle from which the melt originated. In addition, the thickness of the crust and its time of emplacement are further constraints that can be employed to characterize the evolution of the interior.

The crust is essentially the result of a differentiation process that alters the chemical composition of the rocks involved. Upon melting, so-called incompatible elements tend to partition in the liquid phase. The term compatibility refers to the tendency of a trace element (i.e. an element present in concentrations much smaller than 1%) to remain in the solid phase or to move preferentially into the liquid one. This depends on the ability of a given element’s size and charge to fit the crystal structure in which it finds itself. Incompatible elements are, for example, volatile substances, such as water and carbon dioxide, which are also greenhouse gases. Volatiles-enriched magmas can eventually feed volcanoes whose eruptions release the volatiles in the atmosphere with important consequences for planetary climates. Heat producing elements are also incompatible, with the consequence that planetary crusts tend to be enriched in these elements, while the residual mantle is depleted in them. Because of partial melting, the depleted residuum can acquire a lower density than the original mantle, as for harzburgite formation upon extraction of basaltic melt, and also a higher viscosity if hydrated rocks are involved from which water is removed by the melt (see Section 4 and Table 9.3). Furthermore, the upward transport of melt represents an additional mechanism that quickly removes heat from the interior. The dynamics and evolution of the mantle are thus affected in several fundamental ways by the processes related to the generation of crustal material. On the one hand, the redistribution of heat sources and the extraction of melt influence the heat budget of the interior and, in turn, secular cooling. On the other hand, the creation of compositional heterogeneity can affect mantle rheology and generate additional density differences, which can enhance or retard thermally driven mantle flow and induce complex effects such as the emergence of convection in separate layers. As a consequence, mantle dynamics cannot be simply treated as a problem of thermal convection, but rather becomes one of thermo-compositional (or thermochemical) convection.
The fractional crystallization of a cooling magma ocean (see Section 2) may lead to the formation of the first primordial crust. For this, however, observational evidence among terrestrial bodies is scarce. As noted before, an exception is represented by the Moon with its ancient anorthositic crust that likely formed because of the flotation of buoyant plagioclase magma over denser iron-rich melts (see Section 2 and Chapters 23 and 24). The bulk of planetary crusts is thought to form later as a consequence of partial melting in the convecting solid mantle. On Earth, the most common way to build this secondary crust is through decompression melting below spreading centers, where hot material melts while rising adiabatically (despite being cooled) because of the steeper slope of the mantle adiabat with respect to the solidus (see Figure 9.2). This mechanism is responsible for the formation of oceanic crust (typically basaltic). Thanks to plate tectonics, the crust is continuously recycled into the mantle at convergent margins via subduction, whose associated melting processes, including remelting of basalt, lead to the formation of continental crust. This last process is absent in one-plate planets, which lack a mechanism of surface recycling. Decompression melting is also more difficult because a thick stagnant lid obstructs the rise of material near the surface where the solidus temperature is lowest. Partial melting is more likely induced by mantle heating, which in turn can occur in two forms. On the one hand, because of the combined effects of inefficient heat transport through the stagnant lid and of internal heat generation, the mantle temperature below the lid can rise above the solidus, thereby creating a global layer of partial melt. Its surface extrusion can be made difficult by the thick stagnant lid through which it has to percolate. Indeed, it is estimated that only a small part of the melt can reach the surface, while the rest leads to intrusive volcanism forming crust at depth. On the other hand, the mantle temperature can increase locally above the solidus because of ascending hot mantle plumes (see Figures 9.7 and 9.10), which are also thought to be responsible for intraplate volcanism on Earth.

A direct and accurate determination of the crustal thickness is only possible using seismological methods. With the exception of a limited amount of data collected in the framework of the Apollo program of lunar exploration, seismic measurements are available at present only for the Earth. Alternatively, the thickness of the crust can be obtained from the joint analysis of gravity and topography data, which, however, suffers from problems of nonuniqueness. While the study of the Earth’s crust obviously benefits from the easiness of collecting data directly from the field, remote sensing methods need to be employed to investigate the surface of the other terrestrial bodies in order to derive information about their crusts. Image-based analyses of lava flows as well as visible-, infrared-, and nuclear spectroscopy (see Chapters 50 and 54) can all be used to characterize the composition, mineralogy, and chemistry of planetary surfaces and shed light on the magmatic processes that may have led to the observations.

The way theoretical models are employed to gain insight into the thermochemical history of terrestrial planets can be well exemplified by their application to the evolution of the mantle and crust of Mars. Although it is widely accepted that the planet has been volcanically active throughout most of its history—indeed the detection of lava flows as young as a few tens of millions of years is indicative of volcanic activity that lasted until a recent past, it is also believed that the bulk of the Martian crust, including the prominent dichotomy (see Chapter 17 and Figure 9.11(b)), formed approximately within the first 500 million years of evolution. Figure 9.14 shows results from simulations of Mars’ thermochemical evolution in which suitable parameterizations have been employed to describe

![Figure 9.14](image-url)
not only the heat transport by stagnant lid convection but also partial melting and accompanying crust formation. Two families of models distinguished by low (blue) and high (red) initial temperatures are shown for three different reference viscosities. As already discussed above, a lower viscosity not only promotes more efficient convection and cooling but also earlier formation of the crust. An initially hot mantle leads to a rapid generation of large partial melt zones and to the growth of the crustal thickness, which reaches values in excess of 120 km or more depending on the viscosity. The insulating effect of the crust due to its low thermal conductivity and its enrichment in heat producing elements let heat accumulate at the base of the stagnant lid, which tends to thin, causing erosion of the crust from below and recycling in the mantle (this is marked by the abrupt change in the slope of red curves in Figure 9.14(b)). Isotopic characteristics of the Martian meteorites (see Chapters 2 and 28) indicate an early mantle differentiation event about 4.5 Ga ago accompanied by the formation of distinct geochemical reservoirs with little mixing occurring afterward. Therefore, crustal recycling predicted by these models is generally believed to be incompatible with this observation. Crustal erosion is not observed in models started from low temperatures that predict thicknesses between 50 and 75 km, which are within the range of estimates derived from analyses of gravity and topography data. Finally, timing of crust formation is a function of the viscosity. Low values, indicative of a wet rheology and hence of a hydrated mantle, tend to be preferred as they lead to the emplacement of the bulk crust at earlier times.

6.2. Surface Heat Flow and Mantle Heat Budget

Understanding the global energy budget of a planet requires assessing to what extent the loss of primordial heat (secular cooling) is balanced by the heat production due to the decay of long-lived radiogenic elements. In this context, the Earth offers probably the best example for testing models and ideas on planetary dynamics and evolution because of the availability of relatively tight observational constraints. The most recent estimates place the global present-day surface heat flow at 46 TW, corresponding to 90 mW/m². This value is the sum of the contribution from ocean basins (32 TW, corresponding to 106 mW/m²) and continental regions (14 TW, corresponding to 66 mW/m²). The former is estimated by combining a cooling model of the oceanic lithosphere with the ages of the sea floor deduced from marine magnetic anomalies. The latter is based on actual heat flow measurements from continents and their margins. The total heat production of the mantle and continental crust is placed by geochemical models at about 20 TW, of which continents are responsible for approximately 7 TW.

Continental heat sources, however, are stored in the continental lithosphere, which, being highly buoyant and stable, does not participate in convective heat transfer. Therefore, subtracting the continental heat production from the global heat flow yields \( 46 - 7 = 39 \) TW for the heat flowing out of the mantle, and \( 20 - 7 = 13 \) TW for its heat production. The ratio of mantle heat production to heat loss is termed Urey ratio \((Ur)\). For the Earth, we thus have \( Ur = \frac{13}{39} = 0.33 \). However, this is only a preferred value; consideration of various uncertainties in the above estimates, the largest of which are related to the actual mantle composition and hence its heat sources content, yields Urey ratios between about 1/5 and 1/2.

Present-day surface heat flow and interior heat production can be employed to constrain models of the thermochemical history of the Earth’s interior. Figure 9.15 shows results from one such model. The authors of this study conducted dynamic calculations in a two-dimensional spherical annulus. In their simulations, they considered two end-member initial mantle temperatures (1600 and 2500 K) representative of a “cold” and “hot” initial state, accounted for a viscoplastic rheology able to simulate a surface with platelike behavior (see Figure 9.10(b)), and modeled the effects associated with magmatism and crustal production along with the accompanying recycling into the mantle via plate tectonics and subduction. Because of the self-regulation effect induced by the strong temperature dependence of the viscosity, after about 1.5 Ga, average mantle temperatures (Figure 9.15(a)) and surface heat flows (Figure 9.15(b)) tend to converge despite the large differences in their initial values (see Figure 9.13). At the end of the evolution, surface heat flows lie in a narrow range close to the present-day estimate of 32 TW for ocean basins, which is the appropriate value in this context as the model does not account for the presence of continents. For the case with an initial mantle temperature of 2500 K, Figure 9.15(c) and (d) illustrate how the heat flow is partitioned between different mechanisms and the resulting Urey ratios, respectively. The total surface heat flow is the result of a complex heat budget that involves several contributors (Figure 9.15(c)): heat escaping the core, radioactive heat production, heat flux due to magma cooling and solidification, and secular cooling. Variations in surface plate motion and in the formation of hot plumes able to generate significant amounts of partial melt are responsible for the largest oscillations in the evolution of the heat flow associated with secular cooling and magmatism. This strong variability clearly illustrates the highly time-dependent behavior of the mantle caused by the chaotic nature of convection. Figure 9.15(d) shows that the total Urey ratio oscillates around an approximately constant value that is compatible with the range predicted on the basis of observations and geochemical models. Without considering the contribution of magma transport (convective heat flow in
Figure 9.15(d)), the Urey ratio lies systematically above the expected range, indicating that, particularly at early time, magmatism is an essential contributor to the global heat budget of a planet and its cooling.

### 6.3. Core–Mantle Boundary Heat Flow and Magnetic Field History

The generation of magnetic fields in terrestrial bodies is the result of a so-called dynamo action. A necessary condition for a magnetic dynamo to be feasible is that the electrically conductive, iron-rich fluids, which make up the liquid part of metallic cores, undergo vigorous convection. The existence of a liquid core is certain for the Earth, and Mercury, which, incidentally, are also the only terrestrial bodies possessing at present an active magnetic field (note, however, that Jupiter’s icy moon Ganymede also generates a self-sustained field). Nonetheless, several lines of evidence suggest that the cores of Mars, Venus, and the Moon should also be, at least partly, liquid. Although the possibility that the cores of these bodies have completely solidified by now cannot be excluded, the fact that they do not possess a global magnetic field is most likely explained by the absence of an adequate source of buoyancy able to drive intense convection. For a dynamo-generated field to be feasible, it is necessary that the core is cooled by the mantle at a sufficiently high rate. On the one hand, if the heat flowing from the core into the mantle exceeds the heat conducted along the core adiabatic temperature profile, thermal buoyancy can be large enough to overcome ohmic losses (the dissipation of energy into heat caused by electrical resistance). In this case, a thermal dynamo can be realized. On the other hand, a large heat flow at the core–mantle boundary can also cool the liquid core below its liquidus temperature, leading to the onset of the solidification of the inner core. Upon solidification, lighter elements such as sulfur and oxygen, which are usually thought to be alloyed with iron because of their cosmochemical abundance and siderophile nature (i.e. their tendency to dissolve preferentially in iron), tend to be expelled from the newly forming solid phase and be enriched in the remaining liquid. Owing to their lower density with respect to the alloyed composition, light elements provide a source of chemical buoyancy able to drive convection. In this case,
we speak of chemical or compositional dynamos. Therefore, the evolution of the mantle and, in particular, of the core—mantle boundary temperature play a central role in the generation of planetary magnetic fields; if the mantle cools efficiently, e.g., via plate tectonics and subduction, a dynamo is possible; if instead heat escape is difficult, e.g., because of the presence of a thick stagnant lid, a dynamo is still possible but likely less efficient.

Despite its similarity with the Earth in terms of size and, probably, of interior structure, it is known from orbital measurements that Venus does not possess at present an Earth-like magnetic field. This fact does not rule out the possibility that a dynamo operated in the past. Indeed, different scenarios have been proposed such as the existence of an early dynamo or of an intermittent one. The latter would be due to episodic events of plate tectonics, spaced out by long periods of stagnant lid convection, which may have cooled the mantle sufficiently to make a dynamo action possible. However, any reconstruction of the magnetic field history of Venus remains rather speculative. In fact, the high temperature of the surface (∼740 K), and hence also of the crust, are close to the specific temperature (Curie temperature) above which a permanent magnetization can no longer be recorded from magnetic minerals such as magnetite and hematite. Therefore, information on Venus’ magnetic activity from analyses of the magnetization of its surface cannot be expected to be retrieved.

The situation is different for Mars and the Moon. In the former case, a strong remnant magnetization of the crust induced by a global magnetic field has been observed in ancient terrains. The absence of magnetic imprint in younger regions corresponding to large impact basins has been then interpreted as an indication that a Martian dynamo ceased to operate around 4 billion years ago. As far as the Moon is concerned, although it also has no internally generated magnetic field at present, paleomagnetic data combined with radiometric ages of lunar samples retrieved by the Apollo missions indicate that a global field did operate approximately between 4.2 and 3.2 billion years ago, suggesting that an internal field started about 300 million years after the formation of the lunar core. Nevertheless, an even earlier start cannot be ruled out since the very first crust, which might have recorded a magnetic activity, may now be lost.

Timing of magnetic field activity is an important constraint for thermal evolution models as these can be used to predict whether the heat flow from the core into the mantle is compatible with the generation of a (thermally driven) dynamo and its onset and cessation. Figures 9.16 and 9.17 show two models that have been proposed to explain the magnetic field histories of Mars and the Moon, respectively. In the first, which is based on a parametrized description of Mars’ thermochemical evolution, a supercritical heat flow that exceeds the heat conducted along the core adiabat can be obtained in two different ways: either by invoking a plate tectonics mode of convection for the early evolution of Mars (solid lines) or by considering stagnant lid convection for the entire evolution (dashed-dotted lines), but assuming that the core, right after its formation, is superheated with respect to the mantle (by 250 K in this case). On the one hand, plate tectonics is such an efficient mechanism for cooling the interior that it allows for a supercritical heat flow even when assuming that the mantle and core have the same initial temperature. On the other hand, an initially superheated core is a rather natural consequence of the process of core formation. Given the absence of unambiguous evidence for early plate tectonics on Mars, the last simpler scenario appears thus more plausible also in light of its ability to provide a history of crust formation and evolution more consistent with geological and geophysical observations (not shown in the figure).

The case of the lunar magnetic field presented in Figure 9.17 well illustrates the usefulness of two- and three-dimensional models to study processes in which the dynamics of the interior with its complex time dependence plays a central role. Beside the formation of plagioclase, which gave rise to the anorthositic crust (see Sections 2 and 6.1), models of crystallization of the lunar magma ocean also predict the formation of a dense layer of ilmenite (titanium-rich) cumulates at relatively shallow depths.
Because of gravitational instability, ilmenites may have sunk to the core−mantle boundary carrying along an overlying layer strongly enriched in heat producing elements. Accumulation of this mixed and dense layer at the bottom of the mantle creates a thermal blanket that initially prevents core cooling and magnetic field generation. Subsequent heating due to radioactive heat sources progressively increases the thermal buoyancy of the layer, which can ultimately become unstable and rise buoyantly, generating a sufficiently high heat flux to overcome the critical value for the onset of a dynamo. Depending on the actual density of the sunken layer, this model can actually predict the correct timing for the onset and disappearance of the lunar magnetic field and also for the eruption of mare basalts early in the lunar history.

6.4. Lithosphere and Surface Deformation, and Thermal Evolution

Theoretical inferences and observations on the state of deformation of the surface and interior of planetary bodies can also serve to constrain their evolution. In particular, the determination of the elastic response of the lithosphere to surface loads due to geological structures (ice fields, volcanoes, rift systems, etc.) can be linked to the thermal state of the mantle, thereby helping to reconstruct the history of the interior. The lithosphere represents the mechanical boundary layer of convective terrestrial bodies whose viscosity is a strong function of temperature. It comprises the crust and the upper layers of the mantle that undergo elastic and plastic−brittle deformation, and lies above the ideal depth below which fully ductile behavior becomes dominant (see Figure 9.8). The thickness of the elastic part of the lithosphere can be calculated by combining flexural models based on given rheological assumptions with gravity and topography data. The base of the elastic lithosphere, which marks a transition in the mechanical behavior of the mantle, can also be identified with a specific isotherm (∼1050 K assuming a dry olivine rheology). Knowledge of the mantle temperature at a given depth allows for an estimate of the surface heat flow at the loading site. Since, in addition, the elastic lithosphere can support loading stresses over geologically long time intervals (∼10⁶ years), such heat flow is not representative of the current thermal state but rather of the conditions at the time of formation of the load responsible for the deformation, which remained “frozen” in the lithosphere. This information can be employed to constrain the thermal history of a planet.

Figure 9.18 shows results of combined analyses of lithospheric flexure and gravity/topography data for Mars (Figure 9.18(a)) and Venus (Figure 9.18(b)). In the former
case, the elastic thickness correlates well with the age of the surface. While during the Noachian period elastic thicknesses are around 20 km, they quickly increase to values between 50 and 100 km in the Hesperian, and further rise in the Amazonian, with present-day values reaching 300 km. Small elastic thicknesses during the earliest evolution, followed by larger ones at later times are generally consistent with models of Mars’ thermal history that predict the heat flow to decrease rapidly at the beginning of the evolution and more slowly later on (see Figure 9.13). On the contrary, for Venus, which possesses a surface of nearly uniform age (see Chapter 15), no correlation can be recognized between elastic thickness and age of the surface (Figure 9.18(b)), which renders this kind of analyses useful to characterize the recent state of the planet but not its long-term history.

On stagnant lid bodies, whose surface has been preserved for billions of years, the deformation caused by certain tectonic phenomena can also help to understand processes associated with the evolution of the interior. An interesting example in this sense is offered by Mercury, whose surface is characterized by a widespread system of tectonic landforms termed lobate scarps (Chapter 13). These are crust-breaking thrust faults that are interpreted as the result of periods of global planetary contraction. Since the vertical displacement on a fault scales with the length of the fault itself, measurements of the length of lobate scarps from photogeological mapping can be used to retrieve the total amount of contractional strain experienced by the planet and in turn the corresponding decrease in planetary radius. The latest estimates based on the analysis of the images delivered by the camera onboard of the MESSENGER (MERCURY Surface Space Environment Geochemistry and Ranging) spacecraft indicate that Mercury experienced a global contraction of ~2−4 km throughout its history, depending on whether certain secondary tectonic features are also an expression of global contraction or not. As the interior heats up or cools, the corresponding increase or decrease in volume due to thermal expansion and contraction causes the planet’s radius to change accordingly. The observed amount of contraction can be thus used as a constraint in evolution models. Figure 9.19 shows results from one such model. Besides variations due to heating and cooling of the interior, the production of crust also affects the calculation of global radius changes because of the variations in density (and hence in volume) associated with the extraction of partial melt (Section 6.1). During the first 1.5−2 billion years of evolution, mantle heating, accompanied by the production of partial melt and crust formation, causes the planet to expand. Afterward, cooling takes over causing contraction to continue at a nearly constant rate until present. At the end of the evolution, a decrease in planetary radius of ~4 km following the expansion phase is obtained, in agreement with the observations.

6.5. Life, Plate Tectonics and the Evolution of the Interior

The continuous cycling of rock and volatiles between the surface reservoirs and the mantle and crust through plate tectonics opens pathways of biologically altered rock to the interior where this may react with the mantle (compare Figure 9.20(a)). Norman Sleep and colleagues (see
Dennis Hönning and colleagues (see Figure 9.20) have used the well-established fact that the biosphere enhances the erosion rate on continents to consider the possible effects on the mantle water budget and continental growth rate of a biologically enhanced sedimentation rate. It is well established that the rate of water transfer to the mantle will increase with the thickness of the sedimentary layer on a subducting slab (Figure 9.20(a)). On the one hand, the sedimentary layer by virtue of its own porosity will carry water. But it will also reduce the rate of shallow dewatering of the oceanic crust due to overlying and interbedded sediments of low permeability (e.g. clay-rich deposits) sealing off the oceanic crust. An increased flux of water to depth would increase the rate of melting of the slab and adjacent mantle in the source region of andesitic volcanism (see Figure 9.20(a)) thereby increasing the rate of continental growth. The increased rate of subduction of water would further increase the water content of the mantle. Since the viscosity of the mantle will decrease with increasing water content as we have discussed in Section 4, the convection rate would increase and thus the rate of turnover of the plate tectonic engine and the rate of subduction indicating a self-regulatory system. The process would also factor in the carbon silicate cycle that buffers the surface temperature of the Earth. Taking everything else constant, an increased turnover rate of plate tectonics should result in a cooling of the atmosphere.

Figure 9.20(b) shows how the plate tectonics system will evolve in a simplified model that is assumed to evolve toward stable states. The figure shows which stable states the model can attain in a phase plane defined by the mantle water content and the continental surface coverage. The solid line joins states for which the continental surface area is stable (it will not change with time). The dashed line joins states for which the mantle water content is stable. Where the two lines cross, both continental surface coverage and mantle water content are constant. The two points labeled $F_{\text{wet}}$ and $F_{\text{dry}}$ are attractors. The arrows show how states away from these points will evolve toward one of these. The third intersection point between $F_{\text{wet}}$ and $F_{\text{dry}}$ is not an attractor. Rather, the system would evolve away from this point should it ever be reached. For the model parameter combination shown—representative of the present Earth—almost all initial states will evolve toward $F_{\text{wet}}$. The area of attraction for $F_{\text{dry}}$ is small and is labeled by the Roman letter I. Thus, for the present Earth, $F_{\text{wet}}$ is the stable attractor toward which the Earth has evolved. Reducing the weathering rate will enlarge the zone of attraction for $F_{\text{dry}}$ and may even lead to the disappearance altogether of $F_{\text{wet}}$. Thus it has been concluded that an abiotic Earth would likely have only a small coverage of continents and a comparatively dry mantle.

It should be said that the model described above assumes plate tectonics to be operating. The conditions under which plate tectonics is working are not completely understood but it is largely held that it requires a wet and sufficiently vigorously convecting mantle. It can thus be speculated that an abiotic world would not only have a significantly dryer mantle than a biotic one but also may lack plate tectonics altogether. Plate tectonics could thus be a biosignature.
7. CONCLUDING REMARKS AND PERSPECTIVES

The combination of theoretical models with a variety of observational constraints has been dramatically improving our understanding of the evolution of terrestrial bodies. Ever more sophisticated numerical simulations have been helping to unravel the basic principles governing the workings of planetary interiors, despite their inaccessibility to direct observation. Nevertheless, the number and complexity of the physical and chemical processes at play in the mantle and core over a wide range of spatial and temporal scales requires continuous advancements in the models as well as in the way these make use of observational data, both in terms of inputs and constraints. Several fundamental questions are still open. Research on the early evolution following planetary accretion and magma ocean solidification is in its infancy, although these processes could significantly affect the modes of interior heat transfer and hence the entire evolution. Despite the possibility to simulate the behavior of the Earth as a plate tectonics planet (Figure 9.11), a consistent physical theory of this phenomenon explaining its initiation and the ultimate reason why it occurs on Earth but not on the other known bodies is still lacking. Coupling between interior and atmospheric processes (see Section 6.1) has received little attention so far, but it could be a key aspect to understanding the evolution of planets, like e.g. Venus, on which variations in surface conditions, of temperature in particular, can be large enough to induce significant rheological changes.

Progress from mineral physics in the analysis of rocks at conditions resembling those of deep interiors is fundamental to restrict the number and range of free parameters that enter evolution models. Despite the somewhat simplistic view according to which terrestrial planets consist of ordered shells with near-constant properties (Figure 9.1), the study and imaging of the Earth’s interior demonstrates that important heterogeneities exist at all scales, and there is no reason to believe that this should not also be the case of other terrestrial bodies. The integration of consistent descriptions of mantle mineralogy in simulations of thermochemical convection will be thus an essential aspect of the forthcoming generation of evolution models. The quality and amount of surface observations from spacecraft missions continue to increase and to be a

FIGURE 9.20 (a) Schematic cartoon depicting Earth’s global plate tectonics and water cycle. Water is represented by large and small dots, its path by black arrows, and movement of the oceanic plate by white arrows. Initial water uptake occurs within the submarine oceanic crust and sediments. Water loss first occurs after the subduction trench through dewatering, followed by the formation of the water-rich partial melt. The partial melt drives arc volcanism and continental crust formation. A fraction of the water contained in the subducting plate is regassed into the mantle. The water leaves the convecting mantle at midoceanic ridges (MOR) as free volatiles closing the cycle or becomes part of the newly formed oceanic crust. (b) Phase plan spanned by mantle water concentration and continental surface area for a model of the present biotic Earth. The dashed and the solid isolines indicate a steady state for mantle water concentration and the continental area, respectively. The arrows indicate state vectors for a trajectory. Points labeled $F_{wet}$ and $F_{dry}$ are attractors toward which the system will evolve given enough time. The parameters of $F_{wet}$ are close to the parameters of the present Earth. The zone of attraction for $F_{dry}$ is small and is indicated by the Roman letter $I$, while the Roman letter $II$ indicates the zone of attraction of $F_{wet}$ and comprises basically the rest of the phase plane. The size of zone $I$ will increase with decreasing sedimentation rate. For an abiotic Earth with a considerably smaller erosion rate than for the present Earth, the system will likely evolve toward the dry state depending to some extent on initial conditions. It is possible, if not likely, that such a world would lack plate tectonics altogether. After Höning, et al. (2013).
major driver of research on planetary interiors. The level of realism with which numerical models will be able to reconstruct the evolution of terrestrial bodies will crucially depend on our ability to simultaneously account for the largest possible number of observational constraints.

**BIBLIOGRAPHY**


L19312.
1. INTRODUCTION

Life is widespread on the Earth and appears to have been present on the planet since early in its history. Biochemically all life on Earth is similar and seems to share a common origin. Throughout geological history life has significantly altered the environment of the Earth while at the same time adapting to this environment. It would not be possible to understand the Earth as a planet without the consideration of life. Thus, life is a planetary phenomenon and is arguably the most interesting phenomenon observed on planetary surfaces.

Everything we know about life is based on the example of life on Earth. Generalization to other areas or extended forms of life must proceed with this caveat. Although we remain uncertain of the process or the time for its origin, the advent of life on Earth was established within 1 billion years after the formation of the planet. While life also requires energy and nutrients, liquid water is the defining ecological requirement for life on Earth. Thus, a liquid water environment is currently the best indicator of where to search for extraterrestrial life. Looking out into the Solar System we see evidence for liquid water. Europa appears to have a liquid water ocean underneath a global ice surface—the evidence is indirect but persuasive. Enceladus has geysers erupting from its South Polar area powered by subsurface liquid water. There are several lines of evidence that suggest past liquid water on Mars. Direct images from orbiting spacecraft show fluvial features on the surface of Mars. Orbital infrared spectrometers have found local regions that show minerals formed in liquid water environments. The Mars Exploration Rovers and the Curiosity Rover have also found evidence for past aqueous activity at their landing sites on Mars. Our understanding of life, albeit limited to one example and one planet, would suggest that life is possible on other planets whenever conditions allow for environments like those on Earth—energy, nutrients, and most critically liquid water. This suggests the possibility of early microbial life on Mars and forms the basis for a search for Earth-like planets orbiting other stars. Studies of a second example of life—a second genesis—to which we can compare and contrast terrestrial biochemistry will be the beginning of a more general understanding of life as a process in the universe. This implies a search for more
than just fossils but a search for the biochemical remains of organisms, dead or alive.

2. WHAT IS LIFE?

Our understanding of life as a phenomenon is currently based only on the study of life on Earth. One of the profound results of biology is the realization that all life forms on Earth share a common physical and genetic makeup. The impression of vast diversity that we experience in nature is a result of manifold variations on a single fundamental biochemistry. The biochemistry of life is based on 20 amino acids and five nucleotide bases. Added to this are the few sugars, from which are made the polysaccharides, and the simple alcohols and fatty acids that are the building blocks of lipids. This simple collection of primordial biomolecules (Figure 10.1) represents the set from which the rest of biochemistry derives.

![Figure 10.1: The basic molecules of life.](image-url)
Except for glycine, the **amino acids** in Figure 10.1 can have either left handed (L) or right handed (D) symmetry. Figure 10.2 shows the two versions, known as enantiomers (from the Greek *enantios* meaning opposite), for alanine. Life uses only the L-enantiomer to make proteins, although there are some bacteria that use certain D-forms in their cell walls and many others have **enzymes** that can convert the D-form to the L-form. In addition, L-**amino acids** other than the 20 listed in Figure 10.1 are occasionally used in proteins and are sometimes used directly, for example, as toxins by fungi and plants. We do not yet understand how and why life acquired a preference for the L-amino acids over the D-amino acids; this is one of the key observations that theories for the origins of life seek to explain.

The genetic materials of life—DNA (deoxyribonucleic acid) and RNA (ribonucleic acid)—are both constructed from nucleotide bases that form the alphabet of life’s genetic code. In DNA these are adenine (A), thymine (T), cytosine (C), and guanine (G). In RNA thymine is replaced by uracil (U). The nucleic acids each provide a four-letter alphabet in which the codes for the construction of proteins are based. This information recording system is found in all living systems.

The biochemical unity of life, in particular the genetic unity, strongly suggests that all living things on Earth are descendant from a common ancestor. This is the phylogenetic unity of life as shown in Figure 10.3. These genetic trees are obtained by comparing the ribosomal RNA within each organism. There are sections within the RNA that are remarkably similar within all life forms. These conserved sections show only random point changes and not evolutionary trends. Thus, the similarity between the genetic sequences of any two organisms is a measure of their evolutionary distance, or more precisely the time elapsed since they shared a common ancestor. When viewed in this way, life on Earth is divided into three main groups: the eucarya, the bacteria, and the archaea. The eucarya include the multicellular life forms encompassing all plants and animals. The bacteria are the familiar bacteria including intestinal bacteria, common soil bacteria, and the pathogens. The archaea are a different class of microorganisms that are found in unusual and often harsh environments such as hypersaline ponds and H₂ rich anaerobic sediments. All methane-producing microbes are archaea. Archaea are also found in soils and grow on and in humans, producing methane in the gut. Archaea are not known to be human pathogens or to produce substances that are toxic to humans. Why some bacteria but no archaea are pathogenic is not yet understood.

### 2.1. The Ecology of Life: Liquid Water

In addition to describing the building blocks of life it is instructive to consider what life does. In this regard it is
possible to define a set of ecological or functional requirements for life. There are four fundamental requirements for life on Earth: energy, carbon, liquid water, and a few other elements. These are listed in Table 10.1 along with the occurrence of these environmental factors in the Solar System.

Energy is required for life from basic thermodynamic considerations. Typically on the Earth this energy is provided by sunlight, which is a thermodynamically efficient (low entropy) energy source. Some limited systems on Earth are capable of deriving their energy from chemical reactions (e.g., methanogenesis, \( \text{CO}_2 + 4\text{H}_2 \rightarrow \text{CH}_4 + 2\text{H}_2\text{O} \)) and do not depend on photosynthesis. On Earth these systems are confined to locations where the more typical photosynthetic organisms are not able to grow, and it is not known if an ecosystem that was planetary in scale or survived over billions of years could be based solely on chemical energy. There are no known organisms on Earth that make use of temperature gradients to derive energy. These organisms would be analogous to a Carnot heat engine. Table 10.2 lists some of the most important metabolic reactions by which living systems generate energy. This list includes autotrophs (which derive energy from nonbiological sources) as well as heterotrophs (which derive energy by the consumption of organic material, usually other life forms).

Elemental material is required for life, and on Earth carbon has the dominant role as the backbone molecule of biochemistry. Life almost certainly requires other elements as well. Life on Earth utilizes a vast array of the elements available on the surface. However, this does not prove that these elements are absolute requirements for life. Other than \( \text{H}_2\text{O} \) and C, the elements N, S, and P are probably the leading candidates for the status of required elements. Table 10.3 lists the distribution of elements in the cosmos and on the Earth and compares these with the common elements in life.

As indicated in Table 10.1, sunlight and the elements required for life are common in the Solar System. What appears to be the ecologically limiting factor for life in the Solar System is the stability of liquid water. Liquid water is a necessary requirement for life on Earth. Liquid water is key to biochemistry because it acts as the solvent in which biochemical reactions take place and furthermore it interacts with many biochemistries in ways that influence their properties. For example, water forms hydrogen bonds with some parts of a large molecule, the hydrophilic groups, and repels other parts, the hydrophobic groups, thereby forcing these molecules to curl up with their hydrophobic groups in the interior and the hydrophilic groups on the exterior in contact with the water. Certain organisms, notably lichen and some algae, are able to utilize water in the vapor phase if the relative humidity is high enough. Many organisms can continue to metabolize at temperatures well below the freezing point of pure water because their intracellular material contains salts and other solutes that lower the freezing point of the solution. No microorganism currently known is able to obtain water directly from ice. Many organisms, such as the snow algae Chlamydomonas nivalis, thrive in liquid water associated with ice but in these circumstances the organisms are the beneficiaries of external processes that melt the ice. There is no known occurrence of an organism using metabolic methods to overcome the latent heat of fusion of ice thereby liquefying it.

Because liquid water is universally required for known life and because it appears to be rare in the Solar System the search for life beyond the Earth begins first with the search for liquid water.

### 2.2. Generalized Theories for Life

There have been many attempts at a definition of life and perhaps such a definition would aid in our investigation for life on other planets and help unravel the origins of life on Earth. However, it is probable that there will never be a simple definition of life and it may not be necessary in a search for life on other worlds. Despite the fundamental unity of biochemistry and the universality of the genetic code, no single definition has proven adequate in describing the single example of life on Earth. Many of the attributes that we would associate with life, for example,
### TABLE 10.2 Examples of Metabolic Pathways

**Heterotrophy**
1. Fermentation
   \[ \text{C}_6\text{H}_{12}\text{O}_6 \rightarrow 2\text{CO}_2 + 2\text{C}_2\text{H}_5\text{OH} \]
2. Anaerobic respiration
   \[ \text{C}_6\text{H}_{12}\text{O}_6 + 12\text{NO}_3 \rightarrow 6\text{CO}_2 + 6\text{H}_2\text{O} + 12\text{NO}_2^- \]
3. Aerobic respiration
   \[ \text{C}_6\text{H}_{12}\text{O}_6 + 6\text{O}_2 \rightarrow 6\text{CO}_2 + 6\text{H}_2\text{O} \]

**Photoautotrophy**
1. Anoxic photosynthesis
   \[ 12\text{CO}_2 + 12\text{H}_2\text{S} + \text{hv} \rightarrow 2\text{C}_6\text{H}_{12}\text{O}_6 + 9\text{S} + 3\text{SO}_4^2^- \]
2. Oxygenic photosynthesis
   \[ 6\text{CO}_2 + 6\text{H}_2\text{O} + \text{hv} \rightarrow \text{C}_6\text{H}_{12}\text{O}_6 + 3\text{O}_2 \]

**Chemoautotrophy**

#### Anaerobic
1. Methanogens
   \[ \text{CO}_2 + 4\text{H}_2 \rightarrow \text{CH}_4 + 2\text{H}_2\text{O} \]
   \[ \text{CO} + 3\text{H}_2 \rightarrow \text{CH}_4 + \text{H}_2\text{O} \]
   \[ 4\text{CO} + 2\text{H}_2\text{O} \rightarrow \text{CH}_4 + 3\text{CO}_2 \]
2. Acetogens
   \[ 2\text{CO}_2 + 4\text{H}_2 \rightarrow \text{CH}_3\text{COOH} + 2\text{H}_2\text{O} \]
3. Sulfate reducers
   \[ \text{H}_2\text{SO}_4 + 4\text{H}_2 \rightarrow \text{H}_2\text{S} + 4\text{H}_2\text{O} \]
4. Sulfur reducers
   \[ \text{S} + \text{H}_2 \rightarrow \text{H}_2\text{S} \]
5. Thionic denitrifiers
   \[ \text{H}_2\text{S} + 2\text{NO}_3^- \rightarrow \text{SO}_4^{2-} + \text{H}_2\text{O} + \text{N}_2\text{O} \]
   \[ 3\text{S} + 4\text{NO}_3^- + \text{H}_2 \rightarrow 3\text{SO}_4^{2-} + 2\text{N}_2 + 2\text{H}^+ \]
6. Iron reducers
   \[ 2\text{Fe}^{3+} + \text{H}_2 \rightarrow 2\text{Fe}^{2+} + 2\text{H}^+ \]

#### Aerobic
1. Sulfide oxidizers
   \[ 2\text{H}_2\text{S} + 3\text{O}_2 \rightarrow 2\text{SO}_4^2^- + 2\text{H}_2\text{O} \]
2. Iron oxidizers
   \[ 4\text{FeO} + \text{O}_2 \rightarrow 2\text{Fe}_2\text{O}_3 \]

### TABLE 10.3 Elemental Abundances by Mass

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cosmic (%)</th>
<th>Earth’s Crust (%)</th>
<th>Humans (%)</th>
<th>Bacteria (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H 70.7</td>
<td>O 46.6</td>
<td>O 64</td>
<td>O 68</td>
</tr>
<tr>
<td>2</td>
<td>He 27.4</td>
<td>Si 29.7</td>
<td>C 19</td>
<td>C 15</td>
</tr>
<tr>
<td>3</td>
<td>O 0.958</td>
<td>Al 8.13</td>
<td>H 9</td>
<td>H 10.2</td>
</tr>
<tr>
<td>4</td>
<td>C 0.304</td>
<td>Fe 5.00</td>
<td>N 5</td>
<td>N 4.2</td>
</tr>
<tr>
<td>5</td>
<td>Ne 0.174</td>
<td>Ca 3.63</td>
<td>Ca 1.5</td>
<td>P 0.83</td>
</tr>
<tr>
<td>6</td>
<td>Fe 0.126</td>
<td>Na 2.83</td>
<td>P 0.8</td>
<td>K 0.45</td>
</tr>
<tr>
<td>7</td>
<td>N 0.110</td>
<td>K 2.59</td>
<td>S 0.6</td>
<td>Na 0.40</td>
</tr>
<tr>
<td>8</td>
<td>Si 0.0706</td>
<td>Mg 2.09</td>
<td>K 0.3</td>
<td>S 0.30</td>
</tr>
<tr>
<td>9</td>
<td>Mg 0.0656</td>
<td>Ti 0.44</td>
<td>Na 0.15</td>
<td>Ca 0.25</td>
</tr>
<tr>
<td>10</td>
<td>S 0.0414</td>
<td>H 0.14</td>
<td>Cl 0.15</td>
<td>Cl 0.12</td>
</tr>
</tbody>
</table>
self-replication, self-ordering, response to environmental stimuli, can be found in nonliving systems, fire, crystals, bimetallic thermostats, respectively. Furthermore, there are various and peculiar life forms such as viruses and giant cell-less slime molds that defy even a biological definition of life in terms of the cell or the separation of internal and external environments. In attempting a resolution of this problem the most useful definition of life is that it is a system that develops Darwinian evolution: reproduction, mutation, selection (Table 10.4). This is an answer to the question what does life do?

We are able to answer the questions, what does life need and what does life do, even if we do not have a closed form compact definition of life. Thus, the requirements for life listed in Table 10.1 and the functions of life in Table 10.4 are therefore very general and it is probably unwise to apply more restrictive criteria. For example, for evolution to occur some sort of information storage mechanism is required. However, it is not certain that this information mechanism needs to be a DNA/RNA-based system or even that it be expressed in structures dedicated solely for replication.

While on the present Earth all life uses dedicated DNA and RNA systems for genetic coding there is evidence that at one time genetic and structural coding were combined into one molecule, RNA. In this so-called RNA world there would have been no distinction between genotype (genetic) and phenotype (structural) molecular replicating systems—both of these processes would have been performed by an RNA replicating molecule. In present biology the phenotype is composed of proteins for the most part. This example illustrates the difficulty in determining which aspects of biochemistry are fundamental and which are the result of the peculiarities of life’s history on Earth.

In basing our consideration of life, on the distribution we observe here on Earth as a general phenomenon, we suffer simultaneously from the problem that there is only one kind of life on this planet while the variety of that life is too complex to allow for precise definitions or characterizations. Thus, we can neither extrapolate nor be specific in our theories for life.

Some scientists have suggested that living systems elsewhere in the universe may exhibit vast differences from terrestrial biology, and have proposed a variety of alternative life forms. One postulated alternative life form is based on the substitution of ammonia for water. Certainly ammonia is an excellent solvent—in some respects better than water. The range of temperatures over which ammonia is liquid is prevalent in the universe (melting point: −78°, normal boiling point −33°, liquid at room temperature when mixed with water) and the elements that compose it are abundant in the cosmos. Other scientists have suggested the possibility that silicon may be used as a substitute for carbon in alien life forms. However, silicon does not form polymeric chains either as readily or as long as carbon does and its bonds with oxygen (SiO2) are much stronger than carbon bonds (CO2) rendering its oxide essentially inert.

Although speculations of alien life capable of using silicon in place of carbon or ammonia in place of water are intriguing, no specific experiments directed toward alternative biochemistries have been designed. Thus, we have no strategies for where or how to search for such alternate life or its fossils. More significantly, these speculations have not contributed to our understanding of life. One can only conclude that our unique understanding of terrestrial life is based on Earth systems and wide ranging speculations regarding alternate chemistries are currently too limited to be fruitful. Perhaps some day we will develop general theories for life or, more likely, have many sources of life to compare thereby allowing for complete theories. Basing our theories on Earth-like life should be considered a necessary first approach and not a fundamental limitation.

3. THE HISTORY OF LIFE ON EARTH

There are several sources of information about the origin of life on Earth. These include the physical record, the genetic record, the metabolic record, and laboratory simulations. The physical record includes the collection of sedimentary and fossil evidence of life. This record is augmented by theoretical models of the Earth and the Solar System all of which provide clues to conditions billions of years ago when the origin of life is thought to have occurred. There is also the record stored in the genomes of living systems that comprises the collective gene pool of our planet. Genetic information tells us the path of evolution as shaped by environmental pressures, biological constraints, and random events that connect the earliest genomic organism, through the last universal common ancestor to the present tree of life (Figure 10.3). There is also the record of metabolic pathways in the biochemistry of organisms that have evolved in response to changes in the environment while simultaneously causing changes to that environment. All these records are palimpsests in that they have been overwitten—often repeatedly—over time. Laboratory simulations of prebiotic chemistry—the chemistry assumed present before life—can provide clues to the conditions and chemical solutions leading up to the origin of life. Experiments of DNA/RNA replication sequences can provide

<table>
<thead>
<tr>
<th>TABLE 10.4 Properties of Life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties of Life</td>
</tr>
<tr>
<td>Mutation</td>
</tr>
<tr>
<td>Selection</td>
</tr>
<tr>
<td>Reproduction</td>
</tr>
</tbody>
</table>
clues to the selection process that optimizes mutations as well as provide a basic understanding of reproduction. Perhaps one day the process that initiates life will be studied in the laboratory or discovered on another planet.

The major events in the history of life are shown in Figure 10.4. As the Earth was forming about 4.5 Gyr ago its surface would have been inhospitable to life. The gravitational energy released by the formation of the planet would have kept surface temperatures too high for liquid water to exist. Eventually as the heat flow subsided rain would have fallen for the first time and life could be sustained in liquid water. However, it is possible that subsequent impacts could have been large enough to sterilize the Earth by melting, excavating, and vaporizing the planetary surface, removing all liquid water. Thus, life may have been frustrated in its early starts. Following a sufficiently large impact, the entire upper crust of the Earth would be ejected into outer space and any remnant left as a magma ocean. Barring these catastrophic events, however, sterilizing the Earth is a difficult task since it is not sufficient to merely heat the surface to high temperatures. At the present time, microorganisms survive at the bottom of the ocean and even kilometers below the surface of the planet. An Earth-sterilizing impact must not only completely evaporate the oceans but also must then heat the surface and subsurface of the Earth such that the temperature does not fall anywhere below about 200 °C—which is required for heat sterilization of dry, dormant organisms. This is a difficult requirement since the time it takes heat to diffuse down a given distance scales as the square of the distance. Thus, heat must be applied a million times longer to sterilize to a depth of 1 km compared to a depth of 1 m.

It is not known when the last life-threatening impact occurred. As shown schematically in Figure 10.4 the rate of impact, extrapolated from the record on the moon, rises steeply before 3.8 Gyr ago. Thus, it is likely that the Earth was not continuously suitable for life much before 3.8 Gyr ago. There is persuasive evidence that microbial life was present on the Earth as early as 3.4 Gyr ago. This evidence includes microbial fossils and stromatolites. Stromatolites are large features—often many meters in size—that can be formed by the lithification of laminated microbial mats—although physical processes can result in similar forms (See Figure 10.5). Phototactic microorganisms living on the bottom of a shallow lake or ocean shore may be periodically covered with sediment carried in by spring runoff, for example. To retain access to sunlight the organism must move up through this sediment layer and establish a new microbial zone. After repeated cycles a layered series of mats are formed by lamination of the sediments containing the organic material. One characteristic of these biogenic mats that distinguishes them from nonbiologically caused layering is that the response is phototactic not gravitational. Thus, the layered structure is

**FIGURE 10.4** Major events in the history of the Earth and Mars. The period of moist surface conditions on Mars may have corresponded to the time during which life originated on Earth. The similarities between the two planets at this time raise the possibility of the origin of life on Mars.
Stromatolites are an important form of fossil evidence of life because they form macroscopic structures that could be found on Mars. It is therefore possible that a search for stromatolites near the shores of an ancient Martian lake or bay could be conducted in the near future. Expecting microbial communities to have formed stromatolites on Mars is not entirely misplaced geocentricism. The properties of a microbial mat community that results in stromatolite formation need only be those associated with photosynthetic uptake of CO$_2$. There are broad ecological properties that we expect to hold on Mars even if the details of the structure move more toward the side to reach light. In this way stromatolites can sometimes be distinguished from similar but nonbiological laminae. Often stromatolites are not usually flat but is more often domed-shaped because covered microorganisms in a lower layer on the periphery of the structure move more toward the side to reach light. In this way stromatolites can sometimes be distinguished from similar but nonbiological laminae. Often stromatolites contain microfossils—further testimony to their biological origin.

Microbial life—possibly capable of photosynthesis and mobility—appears to have originated early in the history of the Earth possibly before the end of the late bombardment 3.8 Gyr ago and almost certainly not later than 3.4 Gyr ago. This suggests that the time required for the onset of life was brief. If the Greenland sediments are taken as evidence for life it suggests that, within the resolution of the geological record, life arose on Earth as soon as a suitable habitat was provided. The microbial mats at 3.4 Gyr ago put an upper limit of 400 million years on the length of time it took for life to arise after clement conditions were present.

It is possible, in principle, to determine which organism on the Earth is the most similar to the last universal common ancestor. To do so one must determine which organism has changed the least compared to all other organisms. For example, if some taxon of organism contains a certain mutation but many do not one can trace the mutation to an ancestor common to all organisms in that taxon. Within this related group of organisms the most primitive traits can be established based on how widespread they are. Traits that are found in all or most of the major groupings should be primitive, particularly if these traits are found in groups that diverged early. Traits found in only a few recently related groups are probably younger traits. This line of reasoning applied to the entire phylogenetic tree would indicate which organism extant today has the most primitive set of traits. This organism would therefore be the most similar to the common ancestor. Studies of this type have indicated that the organisms alive today that are most similar, genetically and hence presumably ecologically, to the common ancestor are the thermophilic hydrogen-metabolizing bacteria and perhaps the sulfur-metabolizing bacteria. The arrow in Figure 10.3 represents the suggested position of the last common ancestor.

It is important to note here that the last universal common ancestor is not necessarily representative of the first organisms on Earth but was merely the last organism (or group of organisms) from which all life forms today are known to have descended. The common ancestor may have existed within a world of multiple lineages none of which are in evidence today. If all life on Earth has indeed descended from a sulfur bacterium living in a hot springs environment this could be the result of at least three possibilities. First, it may be the case that hot sulfurous environments are important in the origin of life and the common ancestor may represent this primal cell. Second, the common ancestor may have been a survivor of a catastrophe that destroyed all other life forms. The survival of the common ancestor may have been the result of its ability to live deep within a hydrothermal system. Third, the nature of the common ancestor may be serendipitous with no implications as to origin or evolution of the biosphere.

For over 2 Gyr after the earliest evidence for life, life on the Earth was composed of microorganisms only. There were certainly bacteria and possibly one-celled eukaryotes as well. There seemed to be a major change in the environment of the Earth with the rise of photosynthetically produced oxygen beginning at about 2.5 Gyr ago, reaching significant levels about 1 Gyr ago and culminating about 600 Myr ago. (Figure 10.4 shows a timeline of Earth’s history with these events.) Soon after the development of high levels of oxygen in the atmosphere multicellular life forms appeared. These rapidly radiated into the major phylum known today (as well as many that have no known living representatives). In time organisms adapted to land conditions to continue their evolution.
environments in addition to aqueous environments, and plants and animals appeared.

4. THE ORIGIN OF LIFE

There are numerous and diverse theories for the origin of life currently under serious consideration within the scientific community. A diagram and classification of current theories for the origin of life on Earth are shown in Figure 10.6. At the most fundamental level, theories may be characterized within two broad categories: theories that suggest that life originated on Earth (Terrestrial in Figure 10.6) and those that suggest that the origin took place elsewhere (Extraterrestrial in Figure 10.6). The extraterrestrial or panspermia theories suggest that life existed in outer space and was transported by meteorites, asteroids, or comets to a receptive Earth. In this case the origin of life is not related to environments possible on the early Earth. Along similar lines, life may have been ejected by impacts from another planet in the Solar System and jettisoned to Earth—or vice versa. Furthermore, it has been suggested in the scientific literature that life may have been purposely directed to Earth (Directed Panspermia in Figure 10.6) by an intelligent species from another planet.

The terrestrial theories are further subdivided into organic origins (carbon based) and inorganic origins (mineral based). Mineral-based theories suggest that life’s first components were mineral substrates that organized and synthesized clay organisms. These organisms have evolved via natural selection into the organic-based life forms visible on Earth today. The majority of theories that do not invoke an extraterrestrial origin require an organic origin for life on Earth. Theories postulating an organic origin suggest that the initial life forms were composed of the same basic building blocks present in biochemistry today—organic material. If life arose in organic form then there must have been a prebiological source of organics. The Miller–Urey experiments and their successors have demonstrated how organic material may have been produced naturally in the primordial environment of Earth (endogenous production in Figure 10.6). An alternative to the endogenous production of organics on early Earth is the importation of organic material by celestial impacts and debris—comets, meteorites, interstellar dust particles, and comet dust particles. A comparison of these sources is shown in Table 10.5. Table 10.6 lists the organics found in the Murchison meteorite and compares these with the organics produced in a Miller–Urey abiogenic synthesis. Organic origins differ mainly in the type of primal energy sources: photosynthetic, chemosynthetic, or heterotrophic. The phototrophs and chemotrophs (collectively called autotrophs) use energy sources that are inorganic (sunlight and chemical energy, respectively), whereas heterotrophs acquire their energy by consuming organics (Table 10.2).

Hydrothermal vent environments have been suggested for the subsurface origin of chemotrophic life. In the absence of sunlight these organisms must utilize chemical energy (e.g. \( \text{CO}_2 + 4\text{H}_2 \rightarrow \text{CH}_4 + 2\text{H}_2\text{O} + \text{energy} \)). Alternatively phototrophic life utilizes solar radiation from the surface for prebiotic synthesis. These organisms with the ability to chemosynthesize and photosynthesize can assimilate their own energy from materials in their

![Diagram of theories for the origin of life](image-url)

**FIGURE 10.6** Diagrammatic representation and classification of current theories for the origin of life.
**TABLE 10.5** Sources of Prebiotic Organics on Early Earth

<table>
<thead>
<tr>
<th>Source</th>
<th>Energy Dissipation (J/year)</th>
<th>Organic Production (in a Reducing Atmosphere (kg/year))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lightning</td>
<td>$1 \times 10^{18}$</td>
<td>$3 \times 10^9$</td>
</tr>
<tr>
<td>Coronal discharge</td>
<td>$5 \times 10^{17}$</td>
<td>$2 \times 10^9$</td>
</tr>
<tr>
<td>Ultraviolet light ($\lambda &lt; 270$ nm)</td>
<td>$1 \times 10^{22}$</td>
<td>$2 \times 10^{11}$</td>
</tr>
<tr>
<td>Ultraviolet light ($\lambda &lt; 200$ nm)</td>
<td>$6 \times 10^{20}$</td>
<td>$3 \times 10^{9}$</td>
</tr>
<tr>
<td>Meteor entry shocks</td>
<td>$1 \times 10^{17}$</td>
<td>$1 \times 10^{9}$</td>
</tr>
<tr>
<td>Meteor postimpact plumes</td>
<td>$1 \times 10^{20}$</td>
<td>$2 \times 10^{10}$</td>
</tr>
<tr>
<td>Interplanetary dust</td>
<td>—</td>
<td>$6 \times 10^7$</td>
</tr>
</tbody>
</table>

**TABLE 10.6** Comparison of the Amino Acids in Murchison Meteorite and in an Electric Discharge Synthesis, Normalized to Glycine

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Murchison Meteorite</th>
<th>Electric Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Alanine</td>
<td>&gt;50</td>
<td>&gt;50</td>
</tr>
<tr>
<td>α-Amino-n-valeric acid</td>
<td>&gt;50</td>
<td>&gt;50</td>
</tr>
<tr>
<td>α-Aminoisovaleric acid</td>
<td>10</td>
<td>&gt;50</td>
</tr>
<tr>
<td>Saline</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Norvaline</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Isovaline</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Proline</td>
<td>10</td>
<td>0.1</td>
</tr>
<tr>
<td>Pimelic acid</td>
<td>0.1</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>β-Alanine</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>β-Amino-n-valeric acid</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>δ-Aminoisovaleric acid</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>γ-Aminovaleric acid</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>Sarcosine</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>N-Ethyl glycine</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>N-Methyl alanine</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
environment. One feature that the various theories for the origin of life have in common is the requirement for liquid water. This is because the chemistry of even the earliest life requires a liquid water medium. This is true if the primal organism appears fully developed (panspermia), if it engages in organic chemistry, as well as for the clay inorganic theories.

For many years the standard theory for the origin of life posited a terrestrial organic origin requiring endogenous production of organics leading to the development of heterotrophic organisms. This was generally known as the primordial “soup” theory. Recently, there has been serious consideration for the chemotrophic origin of life and at the present time the scientific community is split between these two views.

5. LIMITS TO LIFE

In considering life beyond the Earth it is useful to quantitatively determine the limits that life has been able to reach on this planet with respect to environmental conditions. Life is not everywhere. There are environments on Earth in which life has not been able to effectively colonize even though these environments could be suitable for life. Perhaps the largest life-free zone on Earth is the polar ice sheets. Here there is abundant energy, carbon, and nutrients (from atmospheric deposition) to support life. However, water is available only in the solid form. No organism on Earth has adapted to use metabolic energy to liberate water from ice even though the energy required per molecule is only ~1% of the energy produced by photosynthesis per molecule. Table 10.7 lists the limits to life as we currently know them. The lower temperature limit clearly ties to the presence of liquid water while the higher temperature limits seem to be determined by the stability of proteins, also in liquid water. Life can survive at extremely low light levels—corresponding to 100 AU, roughly three times the distance between Pluto and the Sun. Salinity and pH also allow for a wide range. Water activity, effectively a measure of the relative humidity of a solution or vapor, can support life only for values above 0.6 for yeasts, lichens, and molds. Bacteria require levels above 0.8. Radiation-resistant organisms such as Deinococcus radiodurans can easily survive radiation doses of 1–2 Mrad and higher when in a dehydrated or frozen state.

6. LIFE IN THE SOLAR SYSTEM

Because the knowledge of life is restricted to the unique but varied case found here on Earth, the most practical approach to the search for life on the other planets has been to proceed by way of analogy with life on Earth. The argument for the origin of life on another world is then based on the similarity of other planetary environments with the postulated environments on early Earth. Whatever process led to the establishment of life in one of these environments on Earth could then be logically expected to have led to the origin of life on this comparable world. The more exact the comparison between the early Earth and another planet the more compelling is the argument by analogy. This comparative process should be valid for all the theories for the origin of life listed in Figure 10.6—ranging from panspermia to the standard theory.

Following this line of reasoning further we can conclude that if similar environments existed on two worlds and life arose in both of them then these life forms should be comparable in their broad ecological characteristics. If sunlight was the available energy source, CO2 the available carbon source, and liquid water the solvent then one could expect phototrophic autotrophs using sunlight to fix carbon dioxide with water as the medium for chemical reactions. Our knowledge of the Solar System suggests such an

<table>
<thead>
<tr>
<th>TABLE 10.7 Limits to Life</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameter</strong></td>
</tr>
<tr>
<td>Lower temperature</td>
</tr>
<tr>
<td>Upper temperature</td>
</tr>
<tr>
<td>Low light</td>
</tr>
<tr>
<td>pH</td>
</tr>
<tr>
<td>Salinity</td>
</tr>
<tr>
<td>Water activity</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Radiation</td>
</tr>
</tbody>
</table>
environment could have existed on Mars early in its history as well as Earth early in its history. Thus, while life forms independently originating on these two planets would have different biochemical details they would be recognizably similar in many fundamental attributes. This approach—by analogy to Earth life and the early Earth—provides a specific search strategy for life elsewhere in the Solar System. The key element of that strategy is the search for liquid water habitats.

Spacecrafts have now visited or flown past comets, asteroids, and most of the large worlds in the Solar System except Pluto—and one is en route to Pluto at the time of this writing. Observatory missions have studied all the major objects in the Solar System as well. We can do a preliminary assessment of the occurrence of liquid water habitats—and indirectly life—in the Solar System.

6.1. Mercury and the Moon

Mercury and the Moon appear to have few prospects for liquid water, now or anytime in the past. These virtually airless worlds have negligible amounts of the volatiles (such as water and carbon dioxide) essential for life. There are no geomorphological features that indicate fluid flow. There is evidence that permanently shaded regions of the polar areas on the Moon and Mercury act as traps for water ice. However, there is no indication that the pressure and temperature were ever high enough for liquid water to exist at the surface on either body. (See Mercury, Moon.)

6.2. Venus

Venus currently has a surface that is clearly inhospitable to life. There is no liquid water on the surface and the temperature is over 450 °C at an atmospheric pressure of 92 times the Earth’s. There is water on Venus but only in the form of vapor and clouds in the atmosphere. The most habitable zone on Venus is at the level in the atmosphere where the pressure is about half of the sea level on Earth. At that location, there are clouds composed of about 25% water and 75% sulfuric acid at a temperature of about 25 °C; these might be reasonable conditions for life. It is possible therefore to speculate that life can be found, or survive if implanted, in the clouds of Venus. What argues against this possibility is the fact that clouds on Earth—at similar pressures and temperatures—do not harbor life. We do not know of any life forms that thrive in cloud environments. Perhaps the essential elements are there but a stable environment is required. (See Venus, Atmosphere.)

Theoretical considerations suggest that Venus and Earth may have initially had comparable levels of water. In this case Venus may have had a liquid water surface early in its history when it was cooler—4 billion years ago, due to the reduced brightness of the fainter early sun. Unfortunately, all record of this early epoch has been erased on Venus and the question of the origin of life during such a liquid water period remains untestable. (See Venus, Surface and Interior.)

6.3. Mars

Of all the extraterrestrial planets and smaller objects in the Solar System, Mars is the one that has held the most fascination in terms of life. Early telescopic observations revealed Earth-like seasonal patterns on Mars. Large white polar caps that grew in the winter and shrunk in the summer were clearly visible. Regions of the planet’s surface near the polar caps appeared to darken beginning at the start of each polar cap’s respective spring season and then spreading toward the equator. It was natural that these changes, similar to patterns on the Earth, would be attributed to like causes. Hence, the polar caps were thought to be water ice and the wave of darkening was believed to have been caused by the growth of vegetation. The nineteenth century arguments for the existence of life, and even intelligent life, on Mars culminated in the book Mars as the Abode of Life by Percival Lowell in 1908 and the investigations of the celebrated canals. The Mars revealed by spacecraft exploration is decidedly less alive than Lowell’s anticipation but its standing as the most interesting object for biology outside Earth still remains.

6.3.1. The Viking Results

In 1976, the Viking landers successfully reached the Martian surface while the two orbiters circled the planet repeatedly photographing and monitoring the surface. The primary objective of the Viking mission was the search for microbial life. Previous reconnaissance of Mars by the Mariner flyby spacecraft and the photographs returned from the Mariner 9 orbiter had already indicated that Mars was a cold dry world with a thin atmosphere. There were intriguing features indicative of past fluvial erosion but there was no evidence for current liquid water on the surface. It was thought that any life to be found on Mars would be microbial. The Viking biology package consisted of three experiments shown schematically in Figure 10.7.

The pyrolytic release (PR) experiment searched for evidence of photosynthesis as a sign of life. The PR was designed to see if Martian microorganisms could incorporate CO₂ under illumination. The experiment could be performed under dry conditions—similar to the Martian surface—or it could be run in a humidified mode. The CO₂ in the chamber was labeled with radioactive carbon, which could then be detected in any organic material synthesized during the experiment. The very first run of the PR experiment produced a significant response. It was well below the typical response observed when biotic soils from Earth had
been tested in the experiment but it was much larger than the noise level. Subsequent trials did not reproduce this high result and this initial response was attributed to a start-up anomaly, possibly some small prelaunch contamination.

The gas exchange (GEx) experiment searched for heterotrophs, which are microorganisms capable of consuming organic material. The GEx was designed to detect any gases that the organisms released as a by-product of their metabolism—bacterial flatulence. After a sample was placed in the chamber the soil was first equilibrated with water vapor and then combined with a nutrient solution. At prescribed intervals, a sample of the gas above the sample was removed and analyzed by a gas chromatograph.

The GEx results were startling. When the Martian soil was merely exposed to water vapor it released oxygen gas at levels of 70–770 nmol per gram soil, much larger than could be explained by the release of ambient atmospheric oxygen that had been absorbed onto the soil grains. The GEx results are summarized in Table 10.8. It was clear that some chemical or biological reaction was responsible for the oxygen release. A biological explanation was deemed unlikely since the reactivity of the soil persisted even after it had been heat sterilized to temperatures of over 160 °C. Furthermore, adding the nutrient solution did not change the result that some chemical in the soil was highly reactive with water.

**FIGURE 10.7** Schematic diagram of the Viking biology experiments.

---

**TABLE 10.8 A Comparison of GEx O2 and LR 14C Results**

<table>
<thead>
<tr>
<th>Sample</th>
<th>GEx O₂ (nmol/cm³)</th>
<th>Source² (Trapped O₂)</th>
<th>LR CO₂ (nmol/cm³)</th>
<th>Oxidant² (Hypochlorite³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viking 1 (surface)</td>
<td>770</td>
<td>16 ppm/m</td>
<td>~30</td>
<td>1 ppm/m</td>
</tr>
<tr>
<td>Viking 2 (surface)</td>
<td>194</td>
<td>4</td>
<td>~30</td>
<td>1</td>
</tr>
<tr>
<td>Viking 2 (subrock)</td>
<td>70</td>
<td>1.5</td>
<td>~30</td>
<td>1</td>
</tr>
</tbody>
</table>

¹After Klein, 1979.
²Assuming a bulk soil density of 1.5 g/cm³.
³Based on Quinn et al., 2013.
The labeled release (LR) experiment also searched for evidence of heterotrophic microorganisms. In the LR experiment, a solution of water containing seven organic compounds was added to the soil. The carbon atoms in each organic compound were radioactive. A radiation detector in the headspace detected the presence of radioactive CO\(_2\) released during the experiment. Any carbon metabolism in the soil would be detected as organisms consumed the organics and released radioactive CO\(_2\).

When the LR experiment was performed on Mars there was a steady release of radioactive CO\(_2\); the results are summarized in Table 10.8. When the soil sample was heat sterilized before exposure to the nutrient solution no radioactive CO\(_2\) was detected. The results of the LR experiment were precisely those expected if there were microorganisms in the soil sample. Taken alone the LR results would have been a strong positive indication for life on Mars.

In addition to the three biology experiments there was another instrument that gave information pivotal to the interpretation of the biological results. This was a combination of a gas chromatograph and a mass spectrometer (GCMS). This instrument received Martian soil samples from the same sampling arm that provided soil to the biology experiments. The sample was then heated to release any organics. The decomposed organics were carried through the gas chromatograph and identified by the mass spectrometer. No Martian organics were reported and all signals were attributed to cleaning agents used on the spacecraft before launch. The limit on the concentration of organics that would remain undetectable by the GCMS was one part per billion. A part per billion of organic material in a soil sample represents over a million individual bacteria, each the size of a typical *Escherichia coli*. This may not seem to rule out a biological explanation for the LR results. However, all life is composed of organic material and it is constantly exuded and processed in the biosphere. On Earth, it is difficult to imagine life without a concomitant matrix of organic material. This apparent absence of organic material is the main argument against a biological interpretation of the positive LR results.

The initial explanation for the reactivity of the Martian soil and the apparent absence of organics focused on possible atmospheric oxidants such as hydrogen peroxide produced by ultraviolet light in the atmosphere and deposited onto the soil surface. However, the finding by the Phoenix mission to Mars that the dominant form of chlorine on Mars is as perchlorate has provided the key to understanding the Viking results. Perchlorates in the Viking samples when heated to 500°C would have decomposed into reactive O and Cl oxidizing any organics present and producing the trace chlorinated organic compounds detected. The presence of perchlorates has been confirmed by results from the Curiosity Rover at the Martian equator.

Furthermore, ionizing radiation can decompose perchlorate in the soil on Mars and result in the formation of hypochlorite, other lower oxidation state oxychlorine species, with a concomitant production of O\(_2\) gas that remains trapped in the salt crystal. The hypochlorite provides an explanation for the LR results and the trapped O\(_2\) gas provides an explanation for the GEx results. These reactive forms of chlorine would have broken down any naturally occurring organic material or any material carried in by meteorites on the Martian surface. Table 10.8 also lists the concentration of the sources necessary to explain the Viking results for perchlorate-based models of the chemistry of the LR oxidant.

Amplifying the apparently negative results of the Viking biology experiments, the environment of Mars appears to be inhospitable to life. Although the atmosphere contains many of the elements necessary for life—it is composed of 95% CO\(_2\) with a few percent N\(_2\) and Argon and trace levels of water—the mean surface pressure is less than 1% of sea level pressure on the Earth, and the mean temperature is −60°C. The mean surface pressure is close to the triple point pressure of water. This is the minimum pressure at which a liquid state of water can exist. The low pressures and low temperatures make it unlikely that water will exist as a liquid on Mars. Due to seasonal transport, the available surface water on Mars is trapped as ice in the polar regions. In the locations at low elevation where the pressures and temperatures are sufficient to support liquid water, the surface is desiccated. Even saturated brine solutions cannot exist in equilibrium with the atmosphere near the equator. The absence of liquid water on the surface of Mars is probably the most serious argument against the presence of life anywhere at the surface of the planet. A second significant hazard to life on the Martian surface is the presence of solar ultraviolet light in the wavelengths between 190 and 300 nm. This radiation, which is largely shielded from Earth’s surface by atmospheric oxygen and the ozone layer, is highly effective at destroying terrestrial organisms. Wavelengths below 190 nm are absorbed even by the present thin Martian CO\(_2\) atmosphere. Compounding the effects of UV irradiation, and perhaps caused by it, are possible chemical oxidants that are thought to exist in the Martian soil. Such strong oxidants have been suggested as the causative agent for the chemical reactivity observed at the Viking sites. (See Mars: Atmosphere and Volatile History.)

### 6.3.2. Early Mars

There is considerable evidence that early in its history Mars did have liquid water on its surface. Images from the many orbiters show complex dendritic valley networks that are believed to have been carved by liquid water. These valleys are predominantly found in the heavily cratered, hence ancient, terrains in the southern hemisphere. This would
suggest that the period of liquid water on Mars occurred contemporaneously with the end of the last stages of heavy cratering, about 3.8 Gyr ago. This epoch is the same at which life is thought to have originated on Earth (Figure 10.4). (See Mars, Surface and Interior.)

Figure 10.8 shows part of Nanedi Vallis on Mars. The canyon snakes back and forth—characteristic of liquid flow. On the floor of the canyon there appears a small channel. Presumably this channel was the flow of the river that carved the canyon. It would have taken considerable flow, although not necessarily continuous flow, for this river to have carved the much larger canyon. This image provides what is perhaps the best evidence from orbit that liquid water that flowed on the surface of Mars is stable flow for long periods of time.

The presence of liquid water habitats on early Mars at approximately the time that life is first evident on Earth suggests that life may have originated on Mars during the same time period. Liquid water is the most critical environmental requirement for life on Earth and the general

similarity between Earth and Mars leads us to assume that life on Mars would be similar in this basic environmental requirement. More exotic approaches to life on Mars cannot be ruled out nor are they supported by any available evidence.

It is interesting therefore to consider how evolution may have progressed on Mars by comparison with the Earth. The histories of Earth and Mars are compared in Figure 10.4. In this figure it is seen that the period between 4.0 and 3.5 Gyr ago is the time when life is most likely to have evolved on both planets. On Earth life persists and remains essentially unchanged for several billion years until the cumulative effects of O2 production induces profound changes on the atmosphere of that planet. On Mars, conditions become unsuitable for life (no liquid water on the surface) in a billion years or less. Thus, it is likely that if there were any life on early Mars it remained microbial.

The evidence of liquid water on early Mars, particularly that provided by the valley networks, suggests that the climate on early Mars may have been quite different than the present. It is generally thought that the surface temperature must have been close to freezing, much warmer than the present −60°C. These warmer temperatures are thought to have occurred as a result of a greatly enhanced greenhouse due to a thick (1−5 atm) CO2 atmosphere. However, CO2 condensation may have limited the efficacy of the CO2 greenhouse but theoretical models indicate that CO2 clouds or CH4 could enhance the greenhouse and maintain warmer temperatures. The detection of clays in ancient sediments on Mars is another indication that liquid water was present.

If Mars did have a thick CO2 atmosphere this strengthens the comparison to the Earth, which is thought to have also had a thick CO2 atmosphere early in its history. The duration of a thick atmosphere on Mars and the concomitant warm, wet surface conditions are unknown but simple climate models suggest that significant liquid water habitats could have existed on Mars for ∼0.5 Gyr after the mean surface temperature reached freezing. This model is based on the presence of deep ice-covered lakes (over 30 m) such as those in the dry valleys of the Antarctic where mean annual temperatures are −20°C.

If we divide the possible scenario for the history of water on the surface of Mars into four epochs, the first epoch would have warm surface conditions and liquid water—this is the epoch of clay formation. As Mars gradually loses its thick CO2 atmosphere the second and third epochs would be characterized by low temperatures but still relatively high atmospheric pressures. This is because the temperature would drop rapidly as the pressure decreased. During the second epoch temperatures would rise above freezing during some of the year and liquid water habitats would require a perennial ice cover. However, by epoch

**FIGURE 10.8** Liquid water on another world. Mars Global Surveyor image showing Nanedi Vallis in the Xanthe Terra region of Mars. Image covers an area 9.8 km by 18.5 km; the canyon is about 2.5 km wide. This image is the best evidence we have of liquid water anywhere outside the Earth. *Photo from NASA/Malin Space Sciences.*
three the temperature would never rise above freezing; the only liquid water would be found in porous rocks with favorable exposures to sunlight. In epoch four the pressure would fall too low for the presence of liquid water. These epochs might possibly be associated with the production of sulfates as shown in the timeline in Figure 10.4.

A point worth emphasizing here is that the biological requirement is for liquid water per se. Current difficulties in understanding the composition and pressure of the atmosphere need not lessen the biological importance of the direct evidence for the presence of liquid water. In fact, as we observe in the Antarctic dry valleys, ecosystems can exist when the mean temperatures are well below freezing. Mars need not have ever been above freezing for life to persist.

The particular environment on the early Earth in which life originated is not known. However, this does not pose as serious a problem to the question of the origin of life on Mars as might be expected. The reason is that all the environments found on the early Earth would be expected to be found on Mars; these include hydrothermal sites, hot springs, lakes, oceans (that is planetary scale water reservoirs), volcanoes, tidal pools (solar tides only), wetlands, salt flats, and others. Thus, whatever environment or combination of environments that was needed for life to get started on Earth should have been present on Mars as well—and at the same time.

Since the rationale for life on Mars early in its history is based on analogs with fossil evidence for life on the early Earth it is natural to look to the fossil record on Earth as a guide to how relics of early Martian life might be found. The most persuasive evidence for microbial life on the early Earth comes from stromatolites as discussed before. The resulting structures can be quite large—they are macroscopic fossils generated by microorganisms.

6.3.3. Subsurface Life on Mars

Although there is currently no direct evidence to support speculations about extant life on Mars, there are several interesting possibilities that cannot be ruled out at this time. Protected subsurface niches associated with hydrothermal activity could have continued to support life even after surface conditions became inhospitable. Liquid water could be provided by the heat of geothermal or volcanic activity melting permafrost or other subsurface water sources. Gases from volcanic activity deep in the planet could provide reducing power (as CH₄, H₂, or H₂S) percolating up from below and enabling the development of a microbial community based upon chemolithoautotrophy (see chemolithoautotrophy). An example is a methanogen (or acetogen) that uses H₂ and CO₂ in the production of CH₄. Such ecosystems have been found deep underground on the Earth consuming H₂ produced by the reaction of water with basaltic rock—a plausible reaction for subsurface Mars. However, their existence is neither supported, nor excluded, by current observations of Mars. Tests for such a subsurface system involve locating active geothermal areas associated with ground ice or detecting trace quantities of reduced atmospheric gases that would leak from such a system. It is interesting to consider the recent reports of CH₄ in the atmosphere of Mars at the tens of parts per billion level, and highly variable in space and time. If these reports are confirmed, it may be that this CH₄ may be related to subsurface biological activity. However, nonbiological sources of CH₄ are also possible. The reports are unlikely to be correct as CH₄ is expected to be a long-lived, and hence well-mixed gas in the Martian atmosphere.

While it certainly seems clear that volcanic activity on Mars has diminished over geological time, intriguing evidence for recent (on the geological timescale) volcanic activity comes from the young crystallization ages (all less than 1 Gyr) of the Shergotty meteorite (and other similar meteorites thought to have come from Mars). Volcanic activity by itself does not provide a suitable habitat for life—liquid water that may be derived from the melting of ground ice is also required. Presumably, the volcanic source in the equatorial region would have depleted any initial reservoir of ground ice and there would be no mechanism for renewal—although there are indications of geologically recent volcano/ground ice interactions at equatorial regions. Closer to the poles, ground ice is stable. It is conceivable that a geothermal heat source could result in cycling of water through the frozen ice-rich surface layers. The heat source would be melting and drawing in water from any underlying reservoir of groundwater or ice that might exist. (See Meteorites.)

Another line of reasoning also supports the possibility of subsurface liquid water. There are outflow channels on Mars that appear to be the result of the catastrophic discharge of subsurface aquifers of enormous sizes. There is evidence based on craters and stratigraphic relations that these have occurred throughout Martian history. If this is the case then it is possible that intact aquifers remain. This would have profound implications for exobiology (as well as human exploration). Furthermore, it suggests that the debris field and outwash regions associated with the outflow channel may hold direct evidence of life that existed within the subsurface aquifer just prior to its catastrophic release.

The collection of available water on Mars in the polar regions naturally suggests that summer warming at the edges of the permanent water ice cap may be a source of meltwater, even if short lived. In the polar regions of Earth there are complex microbial ecosystems that survive in transient summer meltwater. However, on Mars the temperature and pressures remain too low for liquid water to
form. Any energy available is lost due to sublimation of the ice before any liquid is produced. It is unlikely that there are even seasonal habitats at the edge of the polar caps. This situation may be different over longer timescales. Changes in the obliquity axis of Mars can significantly increase the amount of insolation reaching the polar caps in summer. If the obliquity increases to over about 50° then the increased temperatures, atmospheric pressures, and polar insolation that result may cause summer liquid water melt streams and ponds at the edge of the polar cap.

In addition to the discovery of perchlorate discussed above, the Phoenix mission, at 68°N, confirmed the presence of ice-cemented ground at a depth below the surface that appeared consistent with vapor-deposited ice. The depth was 4–6 cm. However, in addition to ice-cemented soil there was relatively pure light-toned ice (Figure 10.9). This ice was unexpected, and it appears consistent with the formation of excess ice by soil ice accretion, such as would occur by vapor deposition during times of thermal expansion and contraction. Figure 10.9 shows the light-toned ice at the Phoenix landing site. The change, due presumably to evaporation over the four-sol period, indicates that the light-toned material is indeed ice and not salt or carbonate.

The ice-cemented soils in the northern plains of Mars, such as the Phoenix site, are possibly the best location on Mars for recent habitability. The presence of ice near the surface provides a source of H₂O. The atmospheric surface pressure over the northern plains is well above the triple point of water, so the liquid phase even of pure water would be stable against boiling. This situation is in contrast with the ice-rich southern polar regions, which are at high elevation. Note that the pressure at the Viking 2 lander site located at 49°N never fell below 750 Pa; the triple point of water is 610 Pa. Thus, all that would be needed to provide liquid water activity capable of supporting life is sufficient energy to melt the subsurface ice. This may have occurred as recently as 5 Myr ago, which is when calculations indicate that Mars had an orbital tilt of 45°, compared to the present value of 25°. The summer insolation in the polar regions of Mars at summer solstice for an obliquity of 45° is about twice that for an obliquity of 25°. When Mars had an obliquity of 45°, the polar regions (especially 68°N) received roughly the same level of summer sunlight as Earth’s polar regions do at the present time.

The polar regions may harbor remnants of life in another way. Tens of meters beneath the surface the temperature is well below freezing (−70 °C) and does not change from summer to winter. It is likely that these permafrost zones have remained frozen—particularly in the southern hemisphere—since the end of the intense crater formation period. In this case there may be microorganisms frozen within the permafrost that date back to the time when liquid water was common on Mars, over 3.5 Gyr ago. On Earth permafrost of such age does not exist, but there are sediments in the polar regions that have been frozen for many millions of years. When these sediments are exhumed and samples extracted using sterile techniques viable bacteria are recovered. The sediments on Mars have been frozen much longer (1000 times) but the temperatures are also much colder. Thus, it may be possible that intact microorganism could be

![FIGURE 10.9](image.png) (a) Light-toned ice at the Phoenix landing site. (b) The change, due presumably to evaporation over the four-sol period, indicates that the light-toned material is indeed ice and not salt or carbonate. Image credit NASA/University of Arizona.
recovered from the Martian permafrost. Natural radiation from U, Th, and K in the soil would be expected to have killed any organism but their biochemical remains would be available for study. The southern polar region seems like the best site for searching for evidence of ancient microorganisms since the terrain there can be dated to the earliest period of Martian history as determined by the number of observed craters.

6.3.4. Meteorites from Mars

Of the thousands of meteorites known, there are over 30 that are thought to have come from Mars. It is certain that these meteorites came from a single source because they all have similar ratios of the oxygen isotopes—values distinct from terrestrial, lunar, or asteroidal ratios. These meteorites can be grouped into four classes. Three of these classes contain all but one of the known Mars meteorites and are known by the name of the type specimen; the S (Shergotty), N (Nakhla), and C (Chassigny) class meteorites. The S, N, and C meteorites are relatively young, having crystallized from lava flows between 200 and 1300 million years ago (see Figure 10.4). Gas inclusions in two of the S-type meteorites contain gases similar to the present Martian atmosphere as measured by the Viking Landers—proving that this meteorite, and by inference the others as well, came from Mars. The fourth class of Martian meteorite is represented by the single specimen known as ALH84001. Studies of this meteorite indicate that it formed on Mars about 4.5 Gyr ago in warm, reducing conditions. There are even indications that it contains Martian organic material and appears to have experienced aqueous alteration after formation. This rock formed during the time period when Mars is thought to have had a warm, wet climate capable of supporting life.

It has been suggested that ALH84001 contains evidence for life on Mars based on four observations. (1) Polycyclic aromatic hydrocarbons similar to molecules found in interstellar space are present inside ALH84001. (2) Carbonate globules are found in the meteorite that are enriched in $^{13}$C over $^{12}$C. The isotopic shift is within the range that on Earth, and indicates organic matter derived from biogenic activity. (3) Magnetite and iron sulfide particles are present that are similar to those produced by microbial activity. (4) Features are seen that could be fossils of microbial life, except that they are much smaller than any bacteria on Earth. As a result of this and the study, most scientists currently prefer a nonbiological explanation for all these results. Only the magnetite result is generally considered relevant, although not conclusive, evidence related to life.

ALH84001 does not provide convincing evidence of past life on Mars, when compared to the multiple lines of evidence for life on Earth 3.4 Gyr ago including fossil evidence. However, the ALH84001 results do provide strong support to the suggestion that conditions suitable for life were present on Mars early in its history. When compared to the SNC meteorites, ALH84001 indicates that Mars experienced a transition from a warm reducing environment with organic material present to a cold oxidizing environment in which organic material was unstable.

6.4. The Giant Planets

The “habitable zone” in the inner Solar System provides the temperature conditions that can support liquid water on a planetary surface, but the outer Solar System is richer in the organic material from which life is made. This comparison, which shows the ratio of carbon to heavy elements (all elements other than H and He) for various objects in the Solar System, is shown in Figure 10.10. Earth is in fact depleted in carbon with respect to the average Solar System value by a factor of about $10^4$. It may be interesting then to consider life in the organic-rich outer Solar System.

The giant planets Jupiter, Saturn, Uranus, and Neptune, do not have firm surfaces on which water could accumulate and form a reservoir for life. Here the only clement zone would be that region of the clouds in which temperatures were in the range suitable for life. Cloud droplets would provide the only source of liquid water. Such an environment might provide the key elements needed for life as well as an energy source in the form of sunlight. (See Atmospheres of the Giant Planets.)

There have been speculations that life, including advanced multicellular creatures, could exist in such an environment. However, such speculations are not supported by considerations of the biological state of clouds on Earth.

![Carbon abundance in the solar system](image-url)

**FIGURE 10.10** Ratio of carbon atoms to total heavy atoms (heavier than He) for various solar system objects illustrating the depletion of carbon in the inner solar system. The x-axis is not a true distance scale but the objects are ordered by increasing distance from the sun. Mars is not shown since the size of its carbon reservoir is unknown.
There are no organisms that have adapted themselves to live exclusively in clouds on Earth even in locations where clouds are virtually always present. This niche remains unfilled on Earth and by analogy is probably unfilled elsewhere in the Solar System.

Following this line of thought leads us to search for environments suitable for life on planetary bodies with surfaces. In the outer Solar System this focuses us on the moons of the giant planets. Of particular interest are Europa, Titan, and Enceladus.

### 6.4.1. Europa

Europa, one of the moons of Jupiter, appears to be an airless ice-shrouded world. However, theoretical calculations suggest that under the ice surface of Europa there may be a layer of liquid water sustained by tidal heating as Europa orbits Jupiter. The Galileo Spacecraft imaging showed features in the ice consistent with a subsurface ocean and the magnetometer indicated the presence of a global layer of slightly salty liquid water. The surface of Europa is crisscrossed by streaks that are slightly darker than the rest of the icy surface. If there is an ocean beneath a relatively thin ice layer then these streaks may represent cracks where the water has come to the surface. (See Small Satellites.)

There are many ecosystems on Earth that thrive and grow in water that is continuously covered by ice. These are found in both the Arctic and Antarctic regions. In addition to the polar oceans where sea ice diatoms perform photosynthesis under the ice cover, there are perennially ice-covered lakes in the Antarctic continent in which microbial mats based on photosynthesis are found in the water beneath a 4 m ice cover. The light penetrating these thick ice covers is minimal—about 1% of the incident light. Using these Earth-based systems as a guide it is possible that sunlight penetrating through the cracks (the observed streaks) in the ice of Europa could support a transient photosynthetic community. Alternatively, if there are hydrothermal sites on the bottom of the Europan ocean it may be possible that chemosynthetic life could survive there—by analogy to life at hydrothermal vent sites at the bottom of the Earth’s oceans. The biochemistry of hydrothermal sites on Earth does depend on O₂ produced at the Earth’s surface. On Europa, a chemical scheme like that suggested for subsurface life on Mars would be appropriate (H₂ + CO₂).

The main problem with life on Europa is the question of its origin. Lacking a complete theory for the origin of life, and lacking any laboratory synthesis of life, we have to base our understanding of the origin of life on other planets on analogy with the Earth. It has been suggested that hydrothermal vents may have been the site for the origin of life on Earth and in this case the prospects for life in a putative ocean on Europa are improved. However, the early Earth contained many environments other than hydrothermal vents, such as surface hot springs, volcanoes, lake and ocean shores, tidal pools, and salt flats. If any of these environments were the locale for the origin of the first life on Earth then the case for an origin on Europa is weakened considerably.

### 6.4.2. Titan

Titan, the largest moon of the planet Saturn, has a substantial atmosphere composed primarily of N₂ and CH₄ with many other organic molecules present. The temperature at the surface is close to 94 K and the surface pressure is 1.5 times the pressure of Earth at sea level. The surface does not appear to have expansive oceans as once suggested but numerous small lakes have been discovered in the north polar region. However, the ground beneath the Huygens Probe was wet with liquid CH₄, the result of a slight but constant drizzle. (See Titan.)

The spacecraft data from the Voyager and Cassini/Huygens missions, as well as ground-based studies, indicate that there is an optically thick haze in the upper atmosphere. The haze is composed of organic material and the atmosphere contains many organic molecules heavier than CH₄. Photochemical models suggest that these organics are produced from CH₄ and N₂ through chemical reactions driven by solar photons and by magnetospheric electrons. The observed organic species and even heavier organic molecules are predicted to result from these chemical transformations. Laboratory simulations of organic reactions in Titan-like gas mixtures produce solid refractory organic substances (tholin) and similar processes are expected to occur in Titan’s atmosphere.

Conditions on Titan are much too cold for liquid water to exist, although the pressure is in an acceptable regime. For this reason it is unlikely that Earth-like life could originate or survive there. However, the organic material in Titan’s atmosphere provides a potential source of energy and the liquid methane on the surface could provide a possible liquid medium for an alternate type of life. Life in liquid methane could use active transport and large size to overcome the low solubility of organics in liquid methane and enzymes to catalyze reactions at the low temperatures. If carbon-based life in liquid methane existed on Titan it could be widespread. With or without life, Titan remains interesting because it is a naturally occurring Miller–Urey experiment in which simple compounds are transformed into more complex organics. A detailed study of this process may yield valuable insight into how such a mechanism might have operated on the early Earth.

There is also some speculation that under unusual conditions Titan may have liquid water on or near the surface. This could have occurred early in its formation.
when the gravitational energy released by the formation of Titan would have heated it to high temperatures. More recently, impacts could conceivably melt local regions generating warm subsurface temperatures that could last for thousands of years. Whether such brief episodes of liquid water could have led to water-based life remains to be tested.

6.4.3. Enceladus

The Cassini mission has documented geysers erupting from the south polar region of Enceladus. (See Small Satellites.) Associated with this outflow of water, CH₄, other organics, and NH₃ have been identified. The source of the water has been shown to be a slightly salty subsurface liquid water reservoir heated and pressurized by subsurface heat flow. Such a subsurface habitat could support the sort of anaerobic chemosynthetic life that has been found on Earth. These systems are based on methanogens that consume H₂ produced by geochemical reactions or by radioactive decay. The age or lifetime of any subsurface liquid water on Enceladus is not known, which adds uncertainty to speculation about the origin of life. The theories for the origin of life on Earth, shown in Figure 10.6, which would apply to Enceladus are pan-sperma and a chemosynthetic origin of life. The same theories that would apply to Europa. If there is subsurface life in the liquid water reservoirs on Enceladus then the geysers would be carrying these organisms out into space. Here they would quickly become dormant in the cold vacuum of space and would then be killed by solar ultraviolet radiation. But these dead, frozen microbes would still retain the biochemical and genetic molecules of the living forms. Thus, a Stardust-like mission moving through the plume of Enceladus’ geysers might collect life forms for return to Earth. This might provide the easiest way to get a sample of a second genesis of life.

6.5. Asteroids

Asteroids seem like unlikely locations for life to have originated. Certainly they are too small to support an atmosphere sufficient to allow for the presence of liquid water at the present time. However, asteroids, particularly the so-called carbonaceous type, are thought to contain organic material. Thus, they might have played a role in the delivery of organics to the prebiotic Earth. A more intriguing aspect of some asteroids is the presence of hydrothermally altered materials. This seems to indicate that the asteroids were once part of a larger parent body. Furthermore, conditions on this larger parent body were such that liquid water was present—at least in thin films. Containing both organic material and liquid water, the parent bodies of these asteroids are thus interesting targets in the search for extraterrestrial life forms. However, a thorough assessment of this possibility will require a more detailed study of carbonaceous asteroids in the asteroid belt. Meteorites found on the Earth provide only a glimpse of small fragments of these objects and no signs of extraterrestrial life have been found. But the samples are small and the potential for contamination by Earth life is great.

6.6. Comets

Comets are also known to be rich in organic material. However, unlike asteroids, comets also contain a large fraction of water. In their typical state this water is frozen as ice—unsuitable for life processes. As a comet approaches the sun its surface is warmed considerably, but this leads only to the sublimation of the water ice. Liquid does not form because the pressure at the surface of the comet is much too low.

There has been the suggestion that soon after their formation the interior of large comets would have been heated by short-lived radioactive elements (²⁶⁴Al) to such an extent that the core would have melted. In this case there would have been a subsurface liquid water environment similar to that postulated for the present day Europa. Again the question of the origin of life in such an environment rests on the assumption that life can originate in an isolated deep dark underwater setting.

7. HOW TO SEARCH FOR LIFE ON MARS, EUROPA, OR ENCELADUS

If we were to find organic material in the subsurface of Mars, or in the ice of Europa, or entrained in the geysers of Enceladus, how could we determine if it was the product of a system of biology or merely abiotic organic material from meteorites or photochemistry? If the life is related to Earth life it should be easy to detect. We now have very sensitive methods, such as the amplification of DNA, fluorescent antibody markers, etc., for detecting life from Earth. The case of Earth-like life is the easiest but it is also the least interesting. If the life is not Earth-like then the probes specific to our biology are unlikely to work. We need a general way to determine a biological origin. The question is open and possibly urgent. As we plan missions to Mars and Europa we may have the opportunity to analyze the remains of alien biology.

One practical approach makes use of the distinction between biochemicals and organic matter that is not dependent on a particular organic molecule but results from considering the pattern of the organics in a sample. Abiotic processes will generate a smooth distribution in molecular types without sharp distinctions between similar molecules, isotopes, or chemical chirality. If we
consider a generalized phase space of all possible organic molecules then for an abiotic production mechanism the relative concentration of different types will be a smooth function. In contrast to abiotic mechanisms, biological production will not involve a wide range of possible types. Instead, biology will select a few types of molecules and build biochemistry up from this restricted set. Thus, organic molecules that are chemically very similar may have widely different concentrations in a sample of biological organics. An example of this on Earth is the 20 amino acids used in proteins and the selection of life for the left-handed version of these amino acids. To maximize efficiency life everywhere is likely to evolve this strategy of using a few molecules repeatedly. It may be that other life forms discover the same set of biomolecules that Earth life uses because these are absolutely the most efficient and effective set under any planetary conditions. But it may also be that life elsewhere uses a different set that is optimal given the specific history and conditions of that world. We can search for the repeated use of a set of molecules without knowing in advance what the members of that set will be.

We can apply this approach to the search for biochemistry in the Solar System. Samples of organic material collected from Mars and Europa can be tested for the prevalence of one chirality of amino acid over the other. More generally a complete analysis of the relative concentration of different types of organic molecules might reveal a pattern that is biological even if that pattern does not involve any of the biomolecules familiar from Earth life. Interestingly, if a sample of organics from Mars or Europa shows a preponderance of D-amino acids this will suggest the presence of extant or extinct life and at the same time show that this life is distinct from Earth life. This same conclusion would apply to any clearly biological pattern that is distinct from the pattern of Earth life.

The pattern of biological origin in organic material can potentially persist long after the organisms themselves are dead. Eventually, this distinctive pattern will be destroyed as a result of thermal and radiation effects. Below the surface of Mars, both temperature and radiation are low so this degradation should not be significant. On Europa the intense radiation may destroy the biological signature after several million years at depths to about 1 m below the surface ice.

8. LIFE ABOUT OTHER STARS

In the Solar System we find only our own planet with clear signs of life. Mars, Europa, and Enceladus provides some hopes of finding past or present liquid water but not comparable to the richness of water and life on Earth. Our understanding of life as a planetary phenomenon would clearly benefit from finding another Earth-like planet, around another Sun-like star, that harbored life.

One way of formulating the probability of life, and intelligent life, elsewhere in the galaxy is known as the Drake equation after Frank Drake, a pioneer in the search for extraterrestrial intelligence. The equation and the terms that comprise it are listed in Table 10.9. The most accurately determined variable in the Drake equation at this time is $R^*$, the number of stars forming in the galaxy each year. Since we know that there are about $10^{11}$ stars in our galaxy and that their average lifetime is about $10^{10}$ years, then $R^* \approx 10$ stars per year. All the other terms are uncertain and can be only estimated by extrapolating from what has occurred on Earth. Estimates by different authors for $N$, the number of civilizations in the galaxy capable of communicating by radio waves, range from one to millions. Perhaps the most uncertain term is $L$, the length of time that a technologically advanced civilization can survive.

| TABLE 10.9 The Probability of Life, and Intelligent Life, Elsewhere in the Galaxy |
|---------------------------------------|-----------------------------------------------|
| The Drake Equation $N = R_\ast \times \, f_p \times \, n_e \times \, f_l \times \, f_i \times \, f_c \times \, L$ |
| $N$ | The number of civilizations in the galaxy |
| $R_\ast$ | The number of stars forming each year in the galaxy |
| $f_p$ | The fraction of stars possessing planetary systems |
| $n_e$ | The average number of habitable planets in a planetary system |
| $f_l$ | The fraction of habitable planets in which life originates |
| $f_i$ | The fraction of life forms that develop intelligence |
| $f_c$ | The fraction of intelligent life forms that develop advanced technology |
| $L$ | The length of time, in years, that a civilization survives |
The primary criterion for determining whether a planet can support life is the availability of water in the liquid state. This in turn depends on the surface temperature of the planet that is controlled primarily by the distance to a central star. Life appeared so rapidly on Earth after its formation that it is possible that other planets may only have had to sustain liquid water for a short period of time for life to originate. However, it is important to note that the origin of life is not understood and its probability is completely unknown. Planets orbiting a variety of star types could satisfy this criterion at some time in their evolution. The development of advanced life on Earth, and in particular intelligence, took much longer, almost 4 billion years. Earth maintained habitable conditions for the entire period of time. Locations about stars in which temperatures are conducive to liquid water for such a long period of time have been called continuously habitable zones (CHZs). Calculations of the CHZ about main sequence stars indicate that the mass of the star must be less than 1.5 times the mass of our sun for the CHZ to persist for more than 2 billion years.

An interesting result of these calculations is that the current habitable zone for the sun has an inner limit at about 0.8 AU and extends out to between 1.3 and 1.6 AU, depending on the way clouds are modeled. Thus, while Venus is not in the habitable zone, Earth and Mars both are. This calculation would suggest that Mars is currently habitable. But we see no indication of life. This is owing to the fact that the distance from the sun is not the only determinant for the presence of liquid water on a planet’s surface. The presence of a thick atmosphere and the resultant greenhouse effect is required as well. On Earth the natural greenhouse effect is responsible for warming the Earth by 30°C; without the greenhouse effect the temperature would average −15°C. Mars does not have an appreciable greenhouse effect and hence its temperature averages −60°C. If Earth were at the same distance from the sun that Mars is it would probably be habitable. The reason is the thermostatic effect of the long-term carbon cycle. This cycle is driven by the burial of carbon in seafloor sediments as organic material and carbonates. The formation of carbonates is due to chemical erosion of the surface rocks. Subduction carries this material to depths where the high temperatures release the sedimentary CO₂ gas. These gases escape to the surface in volcanoes that lie on the boundary arc of the subduction zones. The thermostatic action of this cycle results because the erosion rate is strongly dependent on temperature. If the temperature were to drop, erosion would slow down. Meanwhile, the outgassing of CO₂ would result in a buildup of this greenhouse gas and the temperature would rise. Conversely, higher temperatures would result in higher erosion rates and a lowering of CO₂ again stabilizing the temperature.

Mars became uninhabitable because it lacks plate tectonics and hence has no means of recycling the carbon-containing sediments. As a result, the initial thick atmosphere that kept Mars warm has dissipated, presumably into carbonate rocks located on the floor of ancient lake and ocean basins on Mars. Mars lacks plate tectonics because it is too small, 10 times smaller than the Earth, to maintain the active heat flows that drive tectonic activity. The low gravity of Mars and the absence of a magnetic field also contributed to the loss of its atmosphere. Hence, planetary size and its effect on geological activity also play a role in determining the surface temperature and thereby the presence of liquid water and life.

The concept of a galactic habitable zone (GHZ) extends the CHZ to the Milky Way galaxy. The defining feature of the GHZ is the probability of planet formation rather than planet habitability for the CHZ. Planet formation appears to correlate with the concentration of heavy elements and indicates that many galaxies should have a GHZ.

9. CONCLUSION

Life is a planetary phenomenon. We see its profound influences on the surface of one planet—the Earth. Its origin, history, present reach, and global scale interactions remain a mystery primarily because we have only one datum. Many questions about life await the discovery of another life form with which to compare. Mars early in its history is probably the best prospective target in the search for extraterrestrial life forms, although Europa and Enceladus are also promising candidates due to the likely presence of liquid water beneath a surface ice shell and the possibility of associated hydrothermal vent activity. In any case it is likely that our true understanding of life is to be found in the exploration of other worlds—both those with and without life forms. We have only just begun to search.

BIBLIOGRAPHY


Chapter 11
The Sun
Markus J. Aschwanden
Lockheed Martin ATC Solar and Astrophysics Laboratory, Palo Alto, CA, USA

Chapter Outline
1. Introduction 235
2. The Solar Interior 236
   2.1. Standard Models 236
   2.2. Thermonuclear Energy Source 238
   2.3. Neutrinos 238
   2.4. Helioseismology 238
   2.5. Solar Dynamo 239
3. The Photosphere 239
   3.1. Granulation and Convection 239
   3.2. Photospheric Magnetic Field 240
   3.3. Sunspots 240
4. The Chromosphere and Transition Region 241
   4.1. Basic Physical Properties 241
   4.2. Chromospheric Dynamic Phenomena 242
5. The Corona 243
   5.1. Active Regions 243
   5.2. Quiet-Sun Regions 243
   5.3. Coronal Holes 244
   5.4. Hydrostatics of Coronal Loops 244
   5.5. Dynamics of the Solar Corona 244
   5.6. The Coronal Magnetic Field 245
   5.7. MHD Oscillations of Coronal Loops 246
   5.8. MHD Waves in Solar Corona 247
   5.9. Coronal Heating 248
6. Solar Flares and CMEs 248
   6.1. Magnetic Reconnection 249
   6.2. Filaments and Prominences 249
   6.3. Solar Flare Models 249
   6.4. Flare Plasma Dynamics 252
   6.5. Particle Acceleration and Kinematics 253
   6.6. Hard X-Ray Emission 254
   6.7. Gamma-Ray Emission 254
   6.8. Radio Emission 256
   6.9. Coronal Mass Ejections 256
7. Final Comments 258
Bibliography 259

1. INTRODUCTION
The Sun is the central body and energy source of our solar system. The Sun is our nearest star, but otherwise it represents a fairly typical star in our galaxy, classified as G2-V spectral type, with a radius of \( r_\odot \approx 700,000 \text{ km} \), a mass of \( m_\odot \approx 2 \times 10^{33} \text{ g} \), a luminosity of \( L_\odot \approx 3.8 \times 10^{26} \text{ W} \), and an age of \( t_\odot \approx 4.6 \times 10^9 \text{ years} \) (Table 11.1). The distance from the Sun to the Earth is called an Astronomical Unit (AU) and amounts to \( \sim 150 \times 10^6 \text{ km} \). The Sun lies in a spiral arm of our galaxy, the Milky Way, at a distance of 8.5 kiloparsecs from the galactic center. Our galaxy contains \( \sim 10^{12} \) individual stars, many of which are likely to be populated with similar solar systems, according to the rapidly increasing detection of extrasolar planets over the past years; the binary star systems are very unlikely to harbor planets because of their unstable, gravitationally disturbed orbits. The Sun is for us humans of particular significance, first because it provides us with the source of all life and second because it furnishes us with the closest laboratory for astrophysical plasma physics, magnetohydrodynamics (MHD), atomic physics, and particle physics. The Sun still represents the only star from which we can obtain spatial images, in many wavelengths.

The basic structure of the Sun is sketched in Figure 11.1. The Sun and the solar system were formed together from an interstellar cloud of molecular hydrogen some 5 billion years ago. After gravitational contraction and subsequent collapse, the central object became the Sun, with a central temperature hot enough to ignite thermonuclear reactions, the ultimate source of energy for the entire solar system. The chemical composition of the Sun is of 92.1% hydrogen and 7.8% helium by number (or 27.4% He by mass), and 0.1% of heavier elements (or 1.9% by mass,
mostly C, N, O, Ne, Mg, Si, S, Fe). The central core, where hydrogen burns into helium, has a temperature of \( \approx 15 \text{ million K} \) (Figure 11.1). The solar interior further consists of a radiative zone, where energy is transported mainly by radiative diffusion, a process where photons with hard X-ray (kiloelectronvolt) energies get scattered, absorbed, and reemitted. The outer one-third of the solar interior is called the convective zone, where energy is transported mostly by convection. At the solar surface, photons leave the Sun in optical wavelengths, with an energy that is about a factor of \( 10^5 \) lower than the original hard X-ray photons generated in the nuclear core, after a random walk of \( \approx 10^5 - 10^6 \) years.

The irradiance spectrum of the Sun is shown in Figure 11.2, covering all wavelengths from gamma rays, hard X-rays, soft X-rays, extreme ultraviolet (EUV), ultraviolet, white light, infrared, to radio wavelengths. The quiet Sun irradiates most of the energy in visible (white light) wavelengths, to which our human eyes have developed the prime sensitivity during evolution. Emission in EUV is dominant in the solar corona because it is produced by ionized plasma in the coronal temperature range of \( \approx 1 - 2 \) million K. Emissions in shorter wavelengths require higher plasma temperatures and thus occur during flares only. Flares also accelerate particles to nonthermal energies, which cause emission of hard X-rays, gamma rays, and radio wavelengths, but to a highly variable degree.

2. THE SOLAR INTERIOR

The physical structure of the solar interior is mostly based on theoretical models that are constrained (1) by global quantities (age, radius, luminosity, total energy output; see Table 11.1), (2) by the measurement of global oscillations (helioseismology), and (3) by the neutrino flux, which now constrains for the first time elemental abundances in the solar interior, since the neutrino problem has been solved in the year 2001.

2.1. Standard Models

There are two types of models of the solar interior: (1) hydrostatic equilibrium models and (2) time-dependent numerical simulations of the evolution of the Sun.
starting from an initial gas cloud to its present state today, after ~8% of the hydrogen has been burned into helium. The standard hydrostatic model essentially calculates the radial run of temperature, pressure, and density that fulfill the conservation of mass, momentum, and energy in all internal spherical layers of the Sun, constrained by the boundary conditions of radius, temperature, and radiation output (luminosity) at the solar surface, the total mass, and the chemical composition. Furthermore, the ideal gas law and thermal equilibrium is assumed, and thus the radiation is close to that of an ideal blackbody. The solar radius has been measured by triangulation inside the solar system (e.g. during a Venus transit) and by radar echo measurements. The mass of the Sun has been deduced from the orbital motions of the planets (Kepler’s laws) and from precise laboratory measurements of the gravitational constant. The solar luminosity is measured by the heat flux received at the Earth. From these standard models, a central temperature of ~15 million K, a central density of ~150 g/cm³, and a central pressure of $2.3 \times 10^{17}$ dyne/cm² have been inferred. Fine-tuning of the standard model is obtained by including convective transport and by varying the (inaccurately known) helium abundance.

**FIGURE 11.2** The solar irradiance spectrum from gamma rays to radio waves. The spectrum is shifted by 12 orders of magnitude in the vertical axis at $\lambda = 1$ mm to accommodate for the large dynamic range in spectral irradiance. Courtesy of H. Malitson and NASA/NSSDC.
2.2. Thermonuclear Energy Source

The source of solar energy was understood in the 1920s, when Hans Bethe, George Gamow, and Carl Von Weizsäcker identified the relevant nuclear chain reactions that generate solar energy. The main nuclear reaction is the transformation of hydrogen into helium, where 0.7% of the mass is converted into radiation (according to Einstein’s energy equivalence, \( E = mc^2 \)), the so-called \( p-p \) chain, which starts with the fusion of two protons into a nucleus of deuterium \( ^2\text{He} \), and, after chain reactions involving \( ^3\text{He} \), \( ^7\text{Be} \), and \( ^7\text{Li} \) produces helium \( ^4\text{He} \),

\[
p + p \rightarrow ^2\text{He} + e^+ + \nu_e \\
^2\text{He} + p \rightarrow ^3\text{He} + \gamma \\
^3\text{He} + ^3\text{He} \rightarrow ^4\text{He} + p + p
\]

or

\[
^3\text{He} + ^4\text{He} \rightarrow ^7\text{Be} + \gamma \\
^7\text{Be} + e^- \rightarrow ^7\text{Li} + \nu_e \\
^7\text{Li} + p \rightarrow ^8\text{Be} + \gamma \rightarrow ^4\text{He} + ^4\text{He}
\]

One can estimate the Sun’s lifetime by dividing the available mass energy by the luminosity, where we assumed that only about a fraction of 0.1 of the total solar mass is transformed because only the innermost core of the Sun is sufficiently hot to sustain nuclear reactions.

\[
t_\odot = 0.1 \times 0.007 m_\odot c^2 / L_\odot \approx 10^{10} \text{years}
\]

An alternative nuclear chain reaction occurring in the Sun and stars is the carbon–nitrogen–oxygen (CNO) cycle,

\[
^{12}\text{C} + p \rightarrow ^{13}\text{N} + \gamma \\
^{13}\text{N} \rightarrow ^{13}\text{C} + e^+ + \nu_e \\
^{13}\text{C} + p \rightarrow ^{14}\text{N} + \gamma \\
^{14}\text{N} + p \rightarrow ^{15}\text{O} + \gamma \\
^{15}\text{O} \rightarrow ^{15}\text{N} + e^+ + \nu_e \\
^{15}\text{N} + p \rightarrow ^{12}\text{C} + ^4\text{He}
\]

The \( p-p \) chain produces 98.5% of the solar energy, and the CNO cycle produces the remainder, but the CNO cycle is faster in stars that are more massive than the Sun.

2.3. Neutrinos

Neutrinos interact very little with matter, unlike photons, and thus most of the electronic neutrinos \( (\nu_e) \), emitted by the fusion of hydrogen to helium in the central core, escape the Sun without interactions and a very small amount is detected at the Earth. Solar neutrinos have been detected since 1967, pioneered by Raymond Davis, Jr., using a chlorine tank in the Homestake Gold Mine in South Dakota, but the observed count rate was about one-third of the theoretically expected value, causing the puzzling neutrino problem that persisted for the next 35 years. However, Pontecorvo and Gribov predicted already in 1969 that low-energy solar neutrinos undergo a “personality disorder” on their travel to the Earth and oscillate into other atomic flavors of muonic neutrinos \( (\nu_\mu) \) (from a process involving a muon particle) and tauonic neutrinos \( (\nu_\tau) \) (from a process involving a tauon particle), which turned out to be the solution of the missing neutrino problem for detectors that are only sensitive to the highest energy (electronic) neutrinos, such as the Homestake chlorine tank and the gallium detectors Gallium Experiment (GALLEX) in Italy and SAGE in Russia. Only the Kamiokande and Super-Kamiokande-I pure-water experiments and the Sudbury Neutrino Observatory (SNO, Ontario, Canada) heavy-water experiments are somewhat sensitive to the muonic and tauonic neutrinos. It was the SNO that measured in 2001 for the first time all three lepton flavors and, in this way, brilliantly confirmed the theory of neutrino (flavor) oscillations. Today, after the successful solution of the neutrino problem, the measured neutrino fluxes are sufficiently accurate to constrain the helium abundance and heavy element abundances in the solar interior.

2.4. Helioseismology

During 1960–1970, global oscillations were discovered on the solar surface in visible light, which became the field of helioseismology. Velocity oscillations were first measured by R. Leighton and then interpreted in 1970 as standing sound waves in the solar convection zone by R. Ulrich, C. Wolfe, and J. Leibacher. These acoustic oscillations, also called \( p \)-modes (pressure-driven waves), are detectable from fundamental up to harmonic numbers of \( \sim 1000 \) and are most conspicuous in dispersion diagrams, \( \omega(k) \), where each harmonic shows up as a separate ridge, when the oscillation frequency \( (\omega) \) is plotted as function of the wavelength \( \lambda \) (i.e. essentially the solar circumference divided by the harmonic number). Frequencies of the \( p \)-mode correspond to periods of \( \sim 5 \) min. An example of a \( p \)-mode standing wave is shown in Figure 11.3 (left), which appears like a standing wave on a drum skin. Each mode is characterized by the number of radial, longitudinal, and latitudinal nodes, corresponding to the radial quantum number \( n \), the azimuthal number \( m \), and the degree \( l \) of spherical harmonic functions. Since the density and temperature increase monotonically with depth inside the Sun, the sound speed varies as a function of radial
distance from the Sun center. P-mode waves excited at the solar surface propagate downward and are refracted toward the surface. The low harmonics penetrate very deep, whereas high harmonics are confined to the outermost layers of the solar interior. By measuring the frequencies at each harmonic, the sound speed can be inverted as a function of the depth; in this way, the density and temperature profile of the solar interior can be inferred and unknown parameters of theoretical standard models can be constrained, such as the abundance of helium and heavier elements. By exploiting the Doppler effect, frequency shifts of the p-mode oscillations can be used to measure the internal velocity rates as a function of depth and latitude, as shown in Figure 11.3 (right). A layer of rapid change in the internal rotation rate was discovered this way at the bottom of the convection zone, the so-called tachocline (at 0.693 ± 0.002 solar radius, with a thickness of 0.039 ± 0.013 solar radius).

Besides the p-mode waves, gravity waves (g-modes), where buoyancy rather than pressure supplies the restoring force, are suspected in the solar core. These gravity waves are predicted to have long periods (hours) and very small velocity amplitudes, but they have not yet been convincingly detected.

Global helioseismology detects p-modes as a pattern of standing waves that encompass the entire solar surface; however, local deviations of the sound speed can also be detected beneath sunspots and active regions, a diagnostic that is called local helioseismology. Near sunspots, p-modes are found to have oscillation periods in the order of 3 min, compared to 5 min in active region plages and quiet-Sun regions.

2.5. Solar Dynamo

The Sun is governed by a strong magnetic field (much stronger than those on planets), which is generated with a magnetic field strength of $B \approx 10^{5}$ G in the tachocline, the thin shear layer sandwiched between the radiative and the convective zone. Buoyant magnetic flux tubes rise through the convection zone (due to the convective instability obeying the Schwarzschild criterion) and emerge at the solar surface in active regions, where they form sunspots with magnetic field strengths of $B \approx 10^{3}$ G and coronal loops with field strengths of $B \approx 10^{2}$ G at the photospheric footpoints, and $B \approx 10$ G in larger coronal heights. The differential rotation on the solar surface is thought to wind up the surface magnetic field, which then fragments under the magnetic stress, circulates meridionally to the poles, and reorients from the toroidally stressed state (with field lines oriented in the east–west direction) at solar maximum into a poloidal dipole field (connecting the North with the South Pole) in the solar minimum. This process is called the solar dynamo, which flips the magnetic polarity of the Sun every ~11 years (the solar cycle), or returns to the same magnetic configuration every ~22 years (the Hale cycle). The solar cycle controls the occurrence rate of all solar activity phenomena—from sunspot numbers, active regions, to flares, and coronal mass ejections (CMEs).

3. THE PHOTOSPHERE

The photosphere is a thin layer at the solar surface that is observed in white light. The irradiance spectrum in Figure 11.2 shows the maximum at visible wavelengths, which can be fitted with a blackbody spectrum with a temperature of $T \approx 6400$ K at wavelengths of $\lambda \geq 2000$ Å, which is the solar surface temperature. The photosphere is defined as the range of heights from which photons directly escape, which encompasses an optical depth range of $0.1 \leq \tau \leq 3$ and translates into a height range of $h \approx 300$ km for the visible wavelength range.

3.1. Granulation and Convection

The photospheric plasma is only partially ionized; there are fewer than 0.001 electrons per hydrogen atom at the
photospheric temperature of $T = 6400$ K at $\lambda = 5000$ Å. These few ionized electrons come mostly from less abundant elements with a low ionization potential, such as magnesium, while hydrogen and helium are almost completely atomic. The magnetic field is frozen into the gas under these conditions. However, the temperature is rapidly increasing below the photospheric surface, exceeding the hydrogen ionization temperature of $T = 11,000$ K at a depth of 50 km, where the number of ionized electrons increases to 0.1 electrons per hydrogen atom, and the opacity increases by a similar factor. The high opacity of the partially ionized plasma impedes the heat flow. Moreover, a stratification with a temperature gradient steeper than an adiabatic gradient is unstable to convection (Schwarzschild criterion). Thus the partially ionized photosphere of the Sun, as well as of other low-mass stars (with masses $m < 2m_\odot$) are therefore convective.

The observational manifestation of subphotospheric convection is the granulation pattern (Figure 11.4, right), which contains granules with typical sizes of $\sim 1000$ km and lifetimes of $\tau \approx 7$ min. The subphotospheric gas flows up in the bright centers of granulation cells, and then cools by radiating away some heat at the optically thin photospheric surface, and, while cooling, becomes denser and flows down in the intergranular lanes. This convection process can now be reproduced with numerical simulations that include hydrodynamics, radiative transfer, and atomic physics of ionization and radiative processes (Figure 11.4, left). The convection process is also organized on larger scales, exhibiting cellular patterns on scales of $\sim 5000$–10,000 km (mesogranulation) and on scales of $\sim 20,000$ km (supergranulation).

3.2. Photospheric Magnetic Field

Most of what we know about the solar magnetic field is inferred from observations of the photospheric field, from the Zeeman effect of spectral lines in visible wavelengths (e.g. Fe 5250 Å). From two-dimensional (2D) maps of the photospheric magnetic field strength, we extrapolate the coronal three-dimensional (3D) magnetic field, or try to trace the subphotospheric origin from emerging magnetic flux elements. The creation of magnetic flux is thought to occur in the tachocline at the bottom of the convection zone, from where it rises upward in the form of buoyant magnetic flux tubes and emerges at the photospheric surface. The strongest fields emerge in sunspots, amounting to several kilogauss field strengths, and fields with strengths of several 100 G also emerge all over in active regions, often in the form of a leading sunspot trailed by following groups of opposite magnetic polarity. Due to the convective motion, small magnetic flux elements that emerge in the center of granulation cells are then swept to the intergranular lanes, where often unresolved small concentrations are found, with sizes of less than a few 100 km. The flow velocities due to photospheric convection are on the order of $\sim 1$ km/s. In the quiet Sun, away from active regions, the mean photospheric magnetic field amounts to a few Gauss.

3.3. Sunspots

Sunspots are the areas with the strongest magnetic fields, and therefore a good indicator of the solar activity (Figure 11.5, bottom). The butterfly diagram shows that sunspots (or active regions) appear first at higher latitudes
early in the solar cycle and then drift equatorward toward the end of the solar cycle (Figure 11.5, top). Since all solar activity phenomena are controlled by the magnetic field, they have a similar solar cycle dependence as sunspots, such as the flare rate, active region area, global soft X-ray brightness, and radio emission. The appearance of dark sunspots lowers the total luminosity of the Sun only by about 0.15% at sunspot maximum, and thus the variation of the sunlight has a negligible effect on the Earth’s climate. The variation of the EUV emission, which affects the ionization in the Earth’s ionosphere, however, has a more decisive impact on the Earth’s climate.

An individual sunspot consists of a very dark central umbra, surrounded by a brighter, radially striated penumbra. The darkness of sunspots is attributed to the inhibition of convective transport of heat, emitting only about 20% of the average solar heat flux in the umbra and being significantly cooler (≈ 4500 K) than the surroundings (≈ 6000 K). Their diameters range from 3600 to 50,000 km, and their lifetime ranges from a week to several months. The magnetic field in the umbra is mostly vertically oriented, but it is strongly inclined over the penumbral, nearly horizontally. Current theoretical models explain the interlocking comb structure of the filamentary penumbra with outward submerged field lines that are pumped down by turbulent, compressible convection of strong descending plumes.

Sunspots are used to trace the surface rotation since Galileo in 1611. The average sidereal differential rotation rate is

$$\omega = 14.522 - 2.84 \sin^2 \Phi \text{ °/day}$$

where $\Phi$ is the heliographic latitude. The rotation rate of an individual feature, however, can deviate from this average by a few percent because it depends on the anchor depth to which the feature is rooted, since the solar internal differential rate varies radially (Figure 11.3, right).

4. THE CHROMOSPHERE AND TRANSITION REGION

4.1. Basic Physical Properties

The chromosphere (from the Greek word χρῶμος, color) is the lowest part of the solar atmosphere, extending to an average height of ≈ 2000 km above the photosphere. The first theoretical concepts conceived the chromosphere as a spherical layer around the solar surface (in the 1950s; Figure 11.6, left), while later refinements included the diverging magnetic fields (canopies) with height (in the 1980s; Figure 11.6, middle), and finally ended up with a very inhomogeneous mixture of cool gas and hot plasma, as a result of the extremely dynamic nature of chromospheric
phenomena (in the 2000s; Figure 11.6, right). According to hydrostatic standard models assuming local thermodynamic equilibrium (LTE), the temperature first reaches a minimum of \( T = 4300 \) K at a height of \( h \approx 500 \) km above the photosphere and rises then suddenly to \( \sim 10,000 \) K in the upper chromosphere at \( h \approx 2000 \) km, but the hydrogen density drops by about a factor of \( 10^6 \) over the same chromospheric height range. These hydrostatic models have been criticized because they neglect the magnetic field, horizontal inhomogeneities, dynamic processes, waves, and non-LTE conditions.

Beyond the solar limb (without having the photosphere in the background), the chromospheric spectrum is characterized by emission lines; these lines appear dark on the disk as a result of photospheric absorption. The principal lines of the photospheric spectrum are called the Fraunhofer lines, including, for example, hydrogen lines (H I; with the Balmer series H\( \alpha \) (6563 Å), H\( \beta \) (4861 Å), H\( \gamma \) (4341 Å), H\( \delta \) (4102 Å)), calcium lines (Ca II; K 3934 Å, H 3968 Å), and helium lines (He I; D\( _3 \) 5975 Å).

4.2. Chromospheric Dynamic Phenomena

The appearance and fine structure of the chromosphere varies enormously depending on which spectral line, wavelength, and line position (core, red wing, or blue wing) is used because of their sensitivity to different temperatures (and thus altitudes) and Doppler shifts (and thus velocity ranges). In the H and K lines of Ca II, the chromospheric images show a bright network surrounding supergranulation cells, which coincide with the large-scale subphotospheric convection cells. In the Ca II K2 or in ultraviolet continuum lines (1600 Å), the network and internetwork appear grainier. The so-called bright grains have a high contrast in wavelengths that are sensitive to the temperature minimum (4300 K), with an excessive temperature of 30–360 K and with spatial sizes of \( \sim 1000 \) km. The bright points in the network are generally associated with magnetic elements that collide, which then heat the local plasma after magnetic reconnection. In the intranetwork, bright grains result from chromospheric oscillations that produce shock waves. There are also very thin spaghetti-like elongated fine structures visible in H\( \alpha \) spectroheliograms (Figure 11.7, left), which are called fibrils around sunspots. More vertically oriented fine structures are called mottles on the disk or spicules above the limb. Mottles appear as irregular threads, localized in groups around and above supergranules, at altitudes of 700–3000 km above the photosphere, with lifetimes of \( 12–20 \) min, and are apparently signatures of upward and downward motions of plasmas.
with temperatures of $T = 8000-15,000$ K and velocities of $v = 5-10$ km/s. Spicules (Figure 11.7, right) are jetlike structures of plasma with temperatures of $T \approx 10,000$ K that rise to a maximum height of $h \approx 10,000$ km into the lower corona, with velocities of $v \approx 20$ km/s. They carry a maximum flux of 100 times the solar wind into the low corona. Recent numerical simulations by DePontieu and Erdelyi show that global (helioseismic) p-mode oscillations leak sufficient energy from the global resonant cavity into the chromosphere to power shocks that drive upward flows and form spicules. There is also the notion that mottles, fibrils, and spicules could be unified, being different manifestations of the same physical phenomenon at different locations (quiet Sun, active region, above the limb), in analogy to the unification of filaments (on the disk) and prominences (above the limb).

5. THE CORONA

It is customary to subdivide the solar corona into three zones, which all vary their size during the solar cycle: (1) active regions, (2) quiet-Sun regions, and (3) coronal holes.

5.1. Active Regions

Active regions are located in areas of strong magnetic field concentrations, visible as sunspot groups in optical wavelengths or magnetograms. Sunspot groups typically exhibit a strongly concentrated leading magnetic polarity, followed by a more fragmented trailing group of opposite polarity. Because of this bipolar nature, active regions are mainly made up of closed magnetic field lines. Due to the permanent magnetic activity in terms of magnetic flux emergence, flux cancellation, magnetic reconfigurations, and magnetic reconnection processes, a number of dynamic processes such as plasma heating, flares, and CMEs occur in active regions. Consequences of plasma heating in the chromosphere are upflows into coronal loops, which give active regions the familiar appearance of numerous filled loops, which are hotter and denser than the background corona, producing bright emission in soft X-rays and EUV wavelengths. In the EUV image shown in Figure 11.8, active regions appear in white.

5.2. Quiet-Sun Regions

Historically, the remaining areas outside of active regions were dubbed quiet-Sun regions. Today, however, many dynamic processes have been discovered all over the solar surface, so that the term quiet Sun is considered to be a misnomer, only justified in relative terms. Dynamic processes in the quiet Sun range from small-scale phenomena such as network heating events, nanoflares, explosive events, bright points, and soft X-ray jets, to large-scale structures, such as transequatorial loops or coronal arches. The distinction between active regions and quiet-Sun regions becomes more and more blurred because most of the large-scale structures that overarch quiet-Sun
regions are rooted in active regions. A good working definition is that quiet-Sun regions encompass all closed magnetic field regions (excluding active regions), which demarcates the quiet-Sun territory from coronal holes (that encompass open magnetic field regions).

5.3. Coronal Holes

The northern and southern polar zones of the solar globe have generally been found to be darker than the equatorial zones during solar eclipses. Max Waldmeier thus dubbed those zones as coronal holes (i.e. Koronale Löcher in German). Today it is fairly clear that these zones are dominated by open magnetic field lines, which act as efficient conduits for flushing heated plasma from the corona into the solar wind, whenever they are fed by chromospheric upflows at their footpoints. Because of this efficient transport mechanism, coronal holes are empty of plasma most of the time, and thus appear much darker than the quiet Sun, where heated plasma flowing upward from the chromosphere remains trapped, until it cools down and precipitates back to the chromosphere. A coronal hole is visible in Figure 11.8 at the North Pole, where the field structures point radially away from the Sun and show a cooler temperature ($T \leq 1.0$ MK; dark blue in Figure 11.8) than the surrounding quiet-Sun regions.

5.4. Hydrostatics of Coronal Loops

Coronal loops are curvilinear structures aligned with the magnetic field. The cross-section of a loop is essentially defined by the spatial extent of the heating source because the heated plasma distributes along the coronal magnetic field lines without cross-field diffusion, since the thermal pressure is much less than the magnetic pressure in the solar corona. The solar corona consists of many thermally isolated loops, where each one has its own gravitational stratification, depending on its plasma temperature. A useful quantity is the hydrostatic pressure scale height $\lambda_p$, which depends only on the electron temperature $T_e$,

$$\lambda_p(T_e) = \frac{2k_B T_e}{\mu m_H G_{\odot}} = 47,000 \frac{T_e}{1 \text{ MK}} \text{ (km)}$$

where $m_H$ denotes hydrogen mass.

Observing the solar corona in soft X-rays or EUV, which are both optically thin emissions, the line-of-sight integrated brightness intercepts many different scale heights, leading to a hydrostatic weighting bias toward systematically hotter temperatures in larger altitudes above the limb. The observed height dependence of the density needs to be modeled with a statistical ensemble of multi-hydrostatic loops. Measuring a density scale height of a loop requires careful consideration of projection effects, loop plane inclination angles, cross-sectional variations, line-of-sight integration, and the instrumental response functions. Hydrostatic solutions have been computed from the energy balance between the heating rate, the radiative energy loss, and the conductive loss. The major unknown quantity is the spatial heating function, but analysis of loops in high-resolution images indicate that the heating function is concentrated near the footpoints, say at altitudes of $h \leq 20,000$ km. Of course, a large number of coronal loops are found to be not in hydrostatic equilibrium, while nearly hydrostatic loops have been found preferentially in the quiet corona and in older dipolar-active regions. An example of an active region (recorded with the Transition Region and Coronal Explorer (TRACE) about 10 h after a flare) is shown in Figure 11.9, which clearly shows super-hydrostatic loops where the coronal plasma is distributed over up to four times larger heights than expected in hydrostatic equilibrium (Figure 11.9, bottom).

5.5. Dynamics of the Solar Corona

Although the Sun appears lifeless and unchanging to our eyes, except for the monotonic rotation that we can trace
from the sunspot motions, there are actually numerous vibrant dynamic plasma processes continuously happening in the solar corona, which can be detected mainly in EUV and soft X-rays. There is currently a paradigm shift stating that most of the apparently static structures seen in the corona are probably controlled by plasma flows and intermittent heating. It is, however, not easy to measure and track these flows with our remote sensing methods, like the apparently motionless rivers seen from an airplane. For slow flow speeds, the so-called laminar flows, there is no feature to track, while the turbulent flows may be easier to detect because they produce whirls and vortices that can be tracked. A similar situation happens in the solar corona. Occasionally, a moving plasma blob is detected in a coronal loop; it can be used as a tracer. Most of the flows in coronal loops seem to be subsonic (like laminar flows) and thus featureless. Occasionally, we observe turbulent flows, which clearly reveal motion, especially when cool and hot plasma mixes by turbulence and thus yields contrast by emission and absorption in a particular temperature filter. Motion can also be detected with Doppler shift measurements, but this yields only the flow component along the line of sight. There is increasing evidence that flows are ubiquitous in the solar corona.

There are a number of theoretically expected dynamic processes. For instance, loops at coronal temperatures are thermally unstable when the radiative cooling time is shorter than the conductive cooling time, or when the heating scale height falls below one-third of a loop half-length. Recent observations show ample evidence for the presence of flows in coronal loops, as well as evidence for impulsive heating with subsequent cooling, rather than a stationary hydrostatic equilibrium. High-resolution observations of coronal loops reveal that many loops have a superhydrostatic density scale height, far in excess of hydrostatic equilibrium solutions (Figure 11.9, top). Time-dependent hydrodynamic simulations are still in a very exploratory phase, and hydrodynamic modeling of the transition region, coronal holes, and the solar wind remains challenging due to the number of effects that cannot easily be quantified by observations, such as unresolved geometries, inhomogeneities, time-dependent dynamics, and MHD effects.

The coronal plasma is studied with regard to hydrostatic equilibria in terms of fluid mechanics (hydrostatics), with regard to flows in terms of fluid dynamics (hydrodynamics), and including the coronal magnetic field in terms of MHD. The coronal magnetic field has many effects on the hydrodynamics of the plasma. It can play a passive role in the sense that the magnetic geometry does not change (e.g. by channeling particles, plasma flows, heat flows, and waves along its field lines or by maintaining a thermal insulation between the plasmas of neighboring loops or flux tubes). On the other hand, the magnetic field can play an active role (where the magnetic geometry changes), such as exerting a Lorentz force on the plasma, building up and storing nonpotential energy, triggering an instability, changing the topology (by various types of magnetic reconnection), and accelerating plasma structures (filaments, prominences, and CMEs).

5.6. The Coronal Magnetic Field

The solar magnetic field controls the dynamics and topology of all coronal phenomena. Heated plasma flows along magnetic field lines and energetic particles can only propagate along magnetic field lines. Coronal loops are nothing other than conduits filled with heated plasma, shaped by the geometry of the coronal magnetic field, where cross-field diffusion is strongly inhibited. Magnetic field lines take on the same role for coronal phenomena as do highways for street traffic. There are two different magnetic zones in the solar corona that have fundamentally different properties: open-field and closed-field regions. Open-field regions (white zones above the limb in Figure 11.10), which always exist in the polar regions, and sometimes extend toward the equator, connect the solar surface with the interplanetary field and are the source of the fast solar wind (~800 km/s). A consequence of the open-field configuration is efficient plasma transport out into the heliosphere, whenever chromospheric plasma is heated at the footpoints. Closed-field regions (gray zones in Figure 11.10), in contrast, contain mostly closed-field lines in the corona up to heights of about one solar radius, which open up at higher altitudes and connect eventually to the heliosphere, but produce a slow solar wind component of ~400 km/s. It is the closed-field regions that contain all the bright and overdense coronal loops, produced by filling with chromospheric plasma, that stays trapped in these closed-field lines. For loops reaching altitudes higher than about one solar radius, plasma confinement starts to become leaky, because the thermal plasma pressure exceeds the weak magnetic field pressure that decreases with height (plasma-β parameter <1).

The magnetic field on the solar surface is very inhomogeneous. The strongest magnetic field regions are in sunspots, reaching field strengths of $B = 2000–3000$ G. Sunspot groups are dipolar, oriented in an east–west direction (with the leading spot slightly closer to the equator), and with opposite leading polarity in both hemispheres, reversing every 11-year cycle (Hale’s law). Active regions and their plages comprise a larger area around sunspots, with average photospheric fields of $B \approx 100–300$ G, containing small-scale pores with typical fields of $B \approx 1000$ G. The background magnetic field in the quiet Sun and in coronal holes has a net field of $B \approx 0.1–0.5$ G, while the absolute field strengths in
resolved elements amount to \( B = 10-50 \) G. Our knowledge of the solar magnetic field is mainly based on measurements of Zeeman splitting in spectral lines, whereas the coronal magnetic field is reconstructed by extrapolation from magnetograms at the lower boundary, using a potential or force-free field model. The extrapolation through the chromosphere and transition region is, however, uncertain due to unknown currents and non-force-free conditions. The fact that coronal loops exhibit generally much less expansion with height than potential field extrapolations underscores the inadequacy of potential field extrapolations. Direct measurements of the magnetic field in coronal heights are still in their infancy.

### 5.7. MHD Oscillations of Coronal Loops

Much like the discovery of helioseismology four decades ago, it was recently discovered that the solar corona also contains an impressively large ensemble of plasma structures that are capable of producing sound waves and harmonic oscillations. Thanks to the high spatial resolution, image contrast, and time cadence capabilities of the Solar and Heliospheric Observatory (SoHO) and TRACE spacecraft, oscillating loops; prominences, or sunspots; and propagating waves have been identified and localized in the corona and transition region and studied in detail since 1999. These new discoveries established a new discipline that became known as coronal seismology. Even though the theory of MHD oscillations was developed several decades earlier, only the new imaging observations provide diagnostics on length scales, periods, damping times, and densities that allow a quantitative application of the theoretical dispersion relations of MHD waves. The theory of MHD oscillations has been developed for homogeneous media, single interfaces, slender slabs, and cylindrical flux tubes. There are four basic speeds in flux tubes: (1) the Alfvén speed \( v_A = B_0/\sqrt{4\pi \rho_0} \), (2) the sound speed \( c_s = \sqrt{\gamma P_0/\rho_0} \), (3) the cusp or tube speed \( c_T = \left(1/c_s^2 + 1/v_A^2\right)^{-1/2} \), and (4) the kink or mean Alfvén speed \( c_k = \left[(\rho_0 v_A^2 + \rho_e v_A^2)/(\rho_0 + \rho_e)\right]^{1/2} \). For coronal conditions, the dispersion relation reveals a slow-mode branch (with acoustic phase speeds) and a fast-mode branch of solutions (with Alfvén speeds). For the fast-mode branch, a symmetric (sausage) mode and an asymmetric (kink) mode can be distinguished. The fast kink mode produces transverse amplitude oscillations of coronal loops, which have been detected with TRACE (Figure 11.11), having periods in the range of \( P = 2-10 \) min, and can be used to infer the coronal magnetic field strength, thanks to its nondispersive nature. The fast sausage mode is highly dispersive and is subject to a long-wavelength cutoff, so that standing wave oscillations are only possible for thick and high-density (flare and postflare) loops, with periods in the range of \( P \approx 1 \) s to 1 min. Fast sausage-mode oscillations with periods of \( P \approx 10 \) s have recently been imaged for the first time with the Nobeyama radioheliograph, and there are numerous earlier reports on nonimaging detections with periods of \( P \approx 0.5-5 \) s. Finally, slow-mode acoustic oscillations have been detected in flarelike loops with Solar Ultraviolet Measurements of Emitted Radiation having periods in the range of \( P \approx 5-30 \) min. All loop oscillations observed in the solar corona have been found to be subject to strong damping, typically with decay times of only one or two periods. The relevant damping mechanisms are resonant absorption for fast-mode oscillations (or alternatively phase mixing, although requiring an extremely low Reynolds number), and thermal conduction for slow-mode acoustic oscillations. Quantitative modeling of
5.8. MHD Waves in Solar Corona

In contrast to standing modes (with fixed nodes), propagating MHD waves (with moving nodes) have also been discovered in the solar corona recently. Propagating MHD waves result mainly when disturbances are generated impulsively, on timescales faster than the Alfvénic or acoustic travel time across a structure.

Propagating slow-mode MHD waves (with acoustic speed) have been recently detected in coronal loops with TRACE and SoHO/EIT (Figure 11.12); they are usually being launched with 3-min periods near sunspots, or with 5-min periods in plage regions. These acoustic waves propagate upward from a loop footpoint and are quickly damped; they have never been detected in downward direction at the opposite loop side. Propagating fast-mode MHD waves (with Alfvénic speeds) have recently been discovered in a loop in optical (Solar Eclipse Coronal Imaging System eclipse) data, as well as in (Nobeyama) radio images. Ubiquitous low-amplitude waves with Alfvénic phase speeds have also been detected in the corona with the Coronal Multi-Channel Polarimeter at the National Solar Observatory (NSO).

Besides from coronal loops, slow-mode MHD waves have also been detected in plumes in open-field regions in coronal holes, while fast-mode MHD waves have not yet been detected in open-field structures. However, spectroscopic observations of line broadening in coronal holes provide strong support for the detection of Alfvén waves, based on the agreement with the theoretically predicted height-dependent scaling between line broadening and density, $\Delta \nu(h) \propto n_e(h)^{-1/4}$.

The largest manifestations of propagating MHD waves in the solar corona are global waves that spherically propagate after a flare and/or CME over the entire solar surface. These global waves were discovered earlier in Hz, called Moreton waves, and recently in EUV, called EIT
waves (Figure 11.13), usually accompanied with a coronal dimming behind the wave front, suggesting evacuation of coronal plasma by the CME. The speed of Moreton waves is about three times greater than that of EIT waves, which still challenges dynamic MHD models of CMEs.

5.9. Coronal Heating

When Bengt Edlén and Walter Grotian identified Fe IX (nine-times ionized iron) and Ca XIV (14-times ionized calcium) lines in the solar spectrum in 1943, a coronal temperature of $T \approx 1\,\text{MK}$ was first inferred from the formation temperature of these highly ionized atoms. A profound consequence of this measurement is the implication that the corona then consists of a fully ionized hydrogen plasma. Comparing this coronal temperature with the photospheric temperature of 6400 K, we are confronted with the puzzle of how the 200 times hotter coronal temperature can be maintained, the so-called coronal heating problem. Of course, there is also a chromospheric heating problem and a solar wind heating problem. If only thermal conduction were at work, the temperature in the corona should steadily drop down from the chromospheric value with increasing distance, according to the second law of thermodynamics. Moreover, since we have radiative losses by EUV emission, the corona would just cool off in a matter of hours to days, if the plasma temperature could not be maintained continuously by some heating source.

The coronal heating problem has been narrowed down by substantial progress in theoretical modeling with MHD codes; new high-resolution imaging with the Yohkoh Soft X-ray Telescope (SXT), EIT, TRACE, Hinode, and the Atmospheric Imaging Assembly onboard the Solar Dynamics Observatory (SDO); and with more sophisticated data analysis using automated pattern recognition codes. The total energy losses in the solar corona range from $F \approx 3 \times 10^5\,\text{erg/cm}^2\text{s}$ in quiet-Sun regions to $F \approx 10^7\,\text{erg/cm}^2\text{s}$ in active regions. Two main groups of direct current (DC) and alternating current models involve as a primary energy source chromospheric footpoint motion or upward leaking Alfvén waves, which are dissipated in the corona by magnetic reconnection, current cascades, MHD turbulence, Alfvén resonance, resonant absorption, or phase mixing. There is also strong observational evidence for solar wind heating by cyclotron resonance, while velocity filtration seems not to be consistent with EUV data. Progress in theoretical models has mainly been made by abandoning homogeneous flux tubes, but instead including gravitational scale heights and more realistic models of the transition region and taking advantage of numerical simulations with 3D MHD codes (by Boris Gudiksen and Aake Nordlund). From the observational side, we can now unify many coronal small-scale phenomena with flarelike characteristics, subdivided into microflares (in soft X-rays) and nanoflares (in EUV) solely by their energy content. Scaling laws of the physical parameters corroborate their unification. They provide a physical basis to understand the frequency distributions of their parameters and allow estimation of their energy budget for coronal heating. Synthesized data sets of microflares and nanoflares in EUV and soft X-rays have established that these impulsive small-scale phenomena match the radiative loss of the average quiet-Sun corona (Figure 11.14), which points to small-scale magnetic reconnection processes in the transition region and lower corona as primary heating sources.

6. SOLAR FLARES AND CMEs

Rapidly varying processes in the solar corona, which result from a loss of magnetic equilibrium, are called eruptive phenomena, such as flares, CMEs, or eruptive filaments and prominences. The fundamental process that drives all these phenomena is magnetic reconnection.
6.1. Magnetic Reconnection

The solar corona has dynamic boundary conditions: (1) the solar dynamo in the interior of the Sun constantly generates new magnetic flux from the bottom of the convection zone (i.e. the tachocline) which rises by buoyancy and emerges through the photosphere into the corona, (2) the differential rotation as well as convective motion at the solar surface continuously wrap up the coronal field, and (3) the connectivity to the interplanetary field has to constantly break up to avoid excessive magnetic stress. These three dynamic boundary conditions are the essential reasons why the coronal magnetic field is constantly stressed and has to adjust by restructuring the large-scale magnetic field by topological changes, called magnetic reconnection processes. Of course, such magnetic restructuring processes occur wherever magnetic stresses build up (e.g. in filaments, in twisted sigmoid-shaped loops, and along sheared neutral lines). Topological changes in the form of magnetic reconnection always liberate free nonpotential energy, which is converted into heating of plasma, acceleration of particles, and kinematic motion of coronal plasma. Magnetic reconnection processes can occur in a slowly changing quasi-steady way, which may contribute to coronal heating (Section 5.9), but more often happen as sudden violent processes that are manifested as flares and CMEs.

Theory and numerical simulations of magnetic reconnection processes in the solar corona have been developed for steady 2D reconnection (Figure 11.15, top), bursty 2D reconnection, and 3D reconnection. Only steady 2D reconnection models can be formulated analytically; they provide basic relations for inflow speed, outflow speed, and reconnection rate, but represent oversimplifications for most (if not all) observed flares. A more realistic approach seems to be bursty 2D reconnection models (Figure 11.15, bottom), which involve the tearing mode and coalescence instability and can reproduce the sufficiently fast temporal and small spatial scales required by solar flare observations. The sheared magnetic field configurations and the existence of coronal and chromospheric null points, which are now inferred more commonly in solar flares, require ultimately 3D reconnection models, possibly involving null point coalescence, spine reconnection, fan reconnection, and separator reconnection. Magnetic reconnection operates in two quite distinct physical parameter domains: (1) in the chromosphere during magnetic flux emergence, magnetic flux cancellation, and the so-called explosive events and (2) under coronal conditions during microflares, flares, and CMEs.

6.2. Filaments and Prominences

Key elements in triggering flares and/or CMEs are erupting filaments. A filament is a current system above a magnetic neutral line that builds up gradually over days and erupts during a flare or CME process. The horizontal magnetic field lines overlying a neutral line (i.e. the magnetic polarity inversion line) of an active region are filled with cool gas (of chromospheric temperature), embedded in the much hotter tenuous coronal plasma. On the solar disk, these cool dense features appear dark in Hα or EUV images, in absorption against the bright background, and are called filaments, while the same structures appear bright above the limb, in emission against the dark sky background, where they are called prominences. Thus, filaments and prominences are identical structures physically, while their dual name just reflects a different observed location (inside or outside the disk). A further distinction is made regarding their dynamic nature: Quiescent filaments/prominences are long-lived stable structures that can last for several months, while eruptive filaments/prominences are usually associated with flares and CMEs (see example in Figure 11.16).

6.3. Solar Flare Models

A flare process is associated with a rapid energy release in the solar corona, believed to be driven by stored nonpotential magnetic energy and triggered by an instability in the magnetic configuration. Such an energy release process results in acceleration of nonthermal particles and in
FIGURE 11.15  (a) Top: Geometry of the Sweet–Parker (top) and Petschek reconnection model (second panel). The geometry of the diffusion region (gray box) is a long thin sheet ($\Delta \gg d$) in the Sweet–Parker model, but much more compact ($\Delta \approx d$) in the Petschek model (second panel). The Petschek model also considers slow-mode MHD shocks in the outflow region. (b) Numeric MHD simulation of a magnetic reconnection process in a sheared arcade. The grayscale represents the mass density difference ratio, and the dashed lines show the projected magnetic field lines in the vicinity of the reconnection region, at two particular times of the reconnection process. Location a corresponds to a thin compressed region along the slowly rising inner separatrix and location b to a narrow downflow stream outside of the left outer separatrix, and c indicates a broader upflow that follows along the same field lines. Courtesy of Judith Karpen.
heating of coronal/chromospheric plasma. These processes emit radiation in almost all wavelengths: radio, white light, EUV, soft X-rays, hard X-rays, and even gamma rays during large flares. The energy range of flares extends over many orders of magnitude. Small flares that have an energy content of $10^{-6}$ to $10^{-9}$ of the largest flares fall into the categories of microflares and nanoflares (Figure 11.14), which are observed not only in active regions but also in quiet-Sun regions. Some of the microflares and nanoflares have been localized above the photospheric network and are thus also dubbed network flares or network heating events. There are also a number of small-scale phenomena with rapid time variability for which it is not clear whether they represent miniature flare processes (e.g. active region transients, explosive events, blinkers). It is conceivable that some are related to photospheric or chromospheric magnetic reconnection processes, in contrast to flares that always involve coronal magnetic reconnection processes.

The best known flare/CME models entail magnetic reconnection processes that are driven by a rising filament/prominence, flux emergence, converging flows, or shear motion along the neutral line. Flare scenarios with a driver perpendicular to the neutral line (rising prominence, flux emergence, convergence flows) are formulated as 2D reconnection models, while scenarios that involve shear along the neutral line (tearing-mode instability, quadrupolar flux transfer, the magnetic breakout model, and sheared arcade interactions) require 3D descriptions. A 2D reconnection model involving a magnetic X-point is shown in Figure 11.17 (left); a generalized 3D version involving a

---

**FIGURE 11.16** Erupting filament observed with TRACE at 171 Å on July 19, 2000, 23:30 UT, in Active Region 9077. The dark filament mass has temperatures around 20,000 K, while the hot kernels and threads contain plasma with temperatures of 1.0 MK or more. The erupting structure extends over a height of 75,000 km here. Courtesy of TRACE and NASA.

**FIGURE 11.17** Left: A version of the standard 2D X-type reconnection model for two-ribbon flares, pioneered by Carmichael, Sturrock, Hirayama, and Kopp–Pneumann (CSHKP), which also includes the slow and fast shocks in the outflow region, the upward-ejected plasmoid, and the locations of the soft X-ray bright flare loops. (Courtesy of Saku Tsuneta.) Right: 3D version of the two-ribbon flare model, based on the observed evolution during the Bastille Day (July 14, 2000) flare: (a) low-lying, highly sheared loops above the neutral line first become unstable; (b) after loss of magnetic equilibrium, the filament jumps upward and forms a current sheet according to the model by Forbes and Priest. When the current sheet becomes stretched, magnetic islands form and coalescence of islands occurs at locations of enhanced resistivity, initiating particle acceleration and plasma heating; (c) the lowest lying loops relax after reconnection and become filled due to chromospheric evaporation (loops with thick linestyle); (d) reconnection proceeds upward and involves higher lying, less-sheared loops; (e) the arcade gradually fills up with filled loops; (f) the last reconnecting loops have no shear and are oriented perpendicular to the neutral line. At some point, the filament disconnects completely from the flare arcade and escapes into interplanetary space.
highly sheared neutral line is sketched in Figure 11.17 (right). There are more complex versions like the magnetic breakout model, where a second arcade triggers reconnection above a primary arcade. Observational evidence for magnetic reconnection in flares includes the 3D geometry, reconnection inflows, outflows, detection of shocks, jets, ejected plasmoids, and secondary effects like particle acceleration, conduction fronts, and chromospheric evaporation processes. Flare images in soft X-rays often show the cusp-shaped geometry of reconnecting field lines (Figure 11.18, top), while EUV images invariably display the relaxed postreconnection field lines after the flare loops cooled down to EUV temperatures in the postflare phase (Figure 11.18, middle and bottom).

6.4. Flare Plasma Dynamics

The flare plasma dynamics and associated thermal evolution during a flare consists of a number of sequential processes: plasma heating in coronal reconnection sites, chromospheric flare plasma heating (either by precipitating nonthermal particles or by downward propagating heat conduction.
fronts), chromospheric evaporation in the form of upflowing heated plasma, and cooling of postflare loops. The initial heating of the coronal plasma requires anomalous resistivity because Joule heating with classical resistivity is unable to explain the observed densities, temperatures, and rapid timescales in flare plasmas. Other forms of coronal flare plasma heating, such as slow shocks, electron beams, proton beams, or inductive currents, are difficult to constrain with currently available observables. The second stage of chromospheric heating is more thoroughly explored, based on the theory of the thick-target model, with numeric hydrodynamic simulations, and with particle-incell simulations. Important diagnostics on chromospheric heating are also available from $\text{H}_\alpha$, white light, and UV emission, but quantitative modeling is still quite difficult because of the chromospheric opacities and partial ionization. The third stage of chromospheric evaporation has been extensively explored with hydrodynamic simulations, in particular to explain the observed Doppler shifts in soft X-ray lines, while application of spatial models to imaging data is quite sparse. Also, certain types of slow-drifting radio bursts seem to contain information on the motion of chromospheric evaporation fronts. The fourth stage of postflare loop cooling is now understood to be dominated by thermal conduction initially and by radiative cooling later on. However, spatiotemporal temperature modeling of flare plasmas (Figure 11.19) has not yet been fitted to observations in detail.

6.5. Particle Acceleration and Kinematics

Particle acceleration in solar flares is mostly explored by theoretical models because neither macroscopic nor microscopic electric fields are directly measurable by remote sensing methods. The motion of particles can be described in terms of acceleration by parallel electric fields, drift velocities caused by perpendicular forces (i.e. $E \times B$ drifts), and gyromotion caused by the Lorentz force of the magnetic field. Theoretical models of particle acceleration in solar

![FIGURE 11.19](image-url) A 2D numerical MHD simulation of a solar flare with chromospheric evaporation and anisotropic heat conduction in the framework of a 2D magnetic reconnecting geometry. The temporal evolution of the plasma temperature (top row) and density (bottom row) is shown. The temperature and density scale are shown in the bars on the right side. The simulation illustrates the propagation of thermal conduction fronts and the upflows of chromospheric plasma in response. Courtesy of Takaaki Yokoyama and Kazunari Shibata.
flares can be broken down into three groups: (1) DC electric field acceleration, (2) stochastic or second-order Fermi acceleration, and (3) shock acceleration. In the models of the first group, there is a paradigm shift from large-scale DC electric fields (of the size of flare loops) to small-scale electric fields (of the size of magnetic islands produced by the tearing mode instability). The acceleration and trajectories of particles is studied more realistically in the inhomogeneous and time-varying electromagnetic fields around magnetic X-points and O-points of magnetic reconnection sites, rather than in static, homogeneous, large-scale Parker-type current sheets. The second group of models entails stochastic acceleration by gyroresonant wave–particle interactions, which can be driven by a variety of electrostatic and electromagnetic waves, supposed that wave turbulence is present at a sufficiently enhanced level and that the MHD turbulence cascading process is at work. The third group of acceleration models includes a rich variety of shock acceleration models, which is extensively explored in magnetospheric physics and could cross-fertilize solar flare models. Two major groups of models are studied in the context of solar flares (i.e. first-order Fermi acceleration or shockdrift acceleration and diffusive shock acceleration). New aspects are that shock acceleration is now applied to the outflow regions of coronal magnetic reconnection sites, where first-order Fermi acceleration at the standing fast shock is a leading candidate. Traditionally, evidence for shock acceleration in solar flares came mainly from radio type II bursts. New trends in this area are the distinction of different acceleration sites that produce type II emission: flare blast waves, the leading edge of CMEs (bow shock), and shocks in internal and lateral parts of CMEs. In summary, we can say that (1) all three basic acceleration mechanisms seem to play a role to a variable degree in some parts of solar flares and CMEs, (2) the distinctions among the three basic models become more blurred in more realistic (stochastic) models, and (3) the relative importance and efficiency of various acceleration models can only be assessed by including a realistic description of the electromagnetic fields, kinetic particle distributions, and MHD evolution of magnetic reconnection regions pertinent to solar flares.

Particle kinematics, the quantitative analysis of particle trajectories, has been systematically explored in solar flares by performing high-precision energy-dependent time delay measurements with the large-area detectors of the Compton Gamma-Ray Observatory (CGRO). There are essentially five different kinematic processes that play a role in the timing of nonthermal particles energized during flares: (1) acceleration, (2) injection, (3) free-streaming propagation, (4) magnetic trapping, and (5) precipitation and energy loss. The time structures of hard X-ray and radio emission from nonthermal particles indicate that the observed energy-dependent timing is dominated either by free-streaming propagation (obeying the expected electron time-of-flight dispersion) or by magnetic trapping in the weak diffusion limit (where the trapping times are controlled by collisional pitch angle scattering). The measurements of the velocity dispersion from energy-dependent hard X-ray delays allows then to localize the acceleration region, which was invariably found in the cusp of postflare loops (Figure 11.20).

6.6. Hard X-Ray Emission

Hard X-ray emission is produced by energized electrons via collisional bremsstrahlung, most prominently in the form of thick-target bremsstrahlung when precipitating electrons hit the chromosphere. Thin-target bremsstrahlung may be observable in the corona for footpoint-occulted flares. Thermal bremsstrahlung dominates only at energies of $\lesssim 15$ keV. Hard X-ray spectra can generally be fitted with a thermal spectrum at low energies and with a single or double power law nonthermal spectrum at higher energies. Virtually all flares exhibit fast (subsecond) pulses in hard X-rays, which scale proportionally with flare loop size and are most likely spatiotemporal signatures of bursty magnetic reconnection events. The energy-dependent timing of these fast subsecond pulses exhibit electron time-of-flight delays from the propagation between the coronal acceleration site and the chromospheric thick-target site. The inferred acceleration site is located about 50% higher than the soft X-ray flare loop height, most likely near X-points of magnetic reconnection sites (Figure 11.20). The more gradually varying hard X-ray emission exhibits an energy-dependent time delay with opposite sign, which corresponds to the timing of the collisional deflection of trapped electrons. In many flares, the time evolution of soft X-rays roughly follows the integral of the hard X-ray flux profile, which is called the Neupert effect. Spatial structures of hard X-ray sources include: (1) footpoint sources produced by thick-target bremsstrahlung, (2) thermal hard X-rays from flare loop tops, (3) above-the-loop-top (Masuda-type) sources that result from nonthermal bremsstrahlung from electrons that are either trapped in the acceleration region or interact with reconnection shocks, (4) hard X-ray sources associated with upward soft X-ray ejecta, and (5) hard X-ray halo or albedo sources due to backscattering at the photosphere. In spatially extended flares, the footpoint sources assume ribbonlike morphology if mapped with sufficient sensitivity. The monthly hard X-ray flare rate varies about a factor of 20 during the solar cycle, similar to magnetic flux variations implied by the monthly sunspot number, as expected from the magnetic origin of flare energies.

6.7. Gamma-Ray Emission

The energy spectrum of flares (Figure 11.21) in gamma-ray wavelengths (0.5 MeV–1 GeV) is more structured than in
hard X-ray wavelengths (20–500 keV) because it exhibits both continuum emission as well as line emission. There are at least six different physical processes that contribute to gamma-ray emission: (1) electron bremsstrahlung continuum emission, (2) nuclear deexcitation line emission, (3) neutron capture line emission at 2.223 MeV, (4) positron annihilation line emission at 511 keV, (5) pion-decay radiation at $\geq 50$ MeV, and (6) neutron production. The ratio of continuum to line emission varies from flare to flare, and gamma-ray lines can completely be overwhelmed in electron-rich flares or flare phases. When gamma-ray lines are present, they provide a diagnostic of the elemental abundances, densities, and temperatures of the ambient plasma in the chromosphere, as well as of the directivity and pitch angle distribution of the precipitating protons and ions that have been accelerated in coronal flare sites, presumably in magnetic reconnection regions. Critical issues that have been addressed in studies of gamma-ray data are the maximum energies of coronal acceleration mechanisms, the ion/electron ratios (because selective acceleration of ions indicate gyroresonant interactions), the ion/electron timing (to distinguish between simultaneous or second-step acceleration), differences in ion/electron transport (e.g. neutron sources were recently found to be displaced from electron sources), and the first ionization potential effect of chromospheric abundances (indicating enhanced abundances of certain ions that could be preferentially accelerated by gyroresonant interactions). Although detailed modeling of gamma-ray line profiles provides significant constraints on elemental abundances and physical properties of the ambient chromospheric plasma, as well as on the energy and pitch angle distribution of accelerated particles, little information or constraints could be retrieved about the timescales and geometry of the acceleration mechanisms, using gamma-ray data. Nevertheless, the high spectral and imaging resolution of the Ramaty High-Energy Spectroscopic Solar Imager (RHESSI) spacecraft facilitates promising new data for a deeper understanding of ion acceleration in solar flares.
6.8. Radio Emission

Radio emission in the solar corona is produced by thermal, nonthermal, and high-relativistic electrons, and thus provides useful diagnostics complementary to EUV, soft X-rays, hard X-rays, and gamma rays. Thermal or Maxwellian distribution functions produce in radio wavelengths either free-free emission (bremsstrahlung) for low magnetic field strengths or gyroresonant emission in locations of high magnetic field strengths, such as above sunspots, which are both called incoherent emission mechanisms. Since EUV and soft X-ray emission occurs in the optically thin regime, the emissivity adds up linearly along the line of sight. Free-free radio emission is somewhat more complicated because the optical thickness depends on the frequency, which allows direct measurement of the electron temperature in optically thick coronal layers in metric and decimetric frequencies up to \( v \leq 1 \text{ GHz} \) and \( v \geq 2 \text{ GHz} \), free-free emission becomes optically thin in the corona, but gyroresonance emission at harmonics of \( s = 2, 3, 4 \) dominates in strong-field regions. In flares, high-relativistic electrons are produced that emit gyrosynchrotron emission, which allows for detailed modeling of precipitating and trapped electron populations in time profiles recorded at different microwave frequencies.

Unstable non-Maxwellian particle velocity distributions, which have a positive gradient in parallel (beams) or perpendicular (losscones) direction to the magnetic field, drive gyroresonant wave–particle interactions that produce coherent wave growth, detectable in the form of coherent radio emission. Two natural processes that provide these conditions are dispersive electron propagation (producing beams) and magnetic trapping (producing losscones). The wave–particle interactions produce growth of Langmuir waves, upper-hybrid waves, and electron–cyclotron maser emission, leading to a variety of radio burst types (type I, II, III, IV, V, decimetric radio bursts (DCIM), Figure 11.22), which have been mainly explored from (nonimaging) dynamic spectra, while imaging observations have been rarely obtained. Although there is much theoretical understanding of the underlying wave–particle interactions, spatiotemporal modeling of imaging observations is still in its infancy. A solar-dedicated, frequency-agile imager with many frequencies (Frequency-Agile Solar Radio (FASR) telescope) is in the planning stage and might provide more comprehensive observations.

6.9. Coronal Mass Ejections

Every star is losing mass, caused by dynamic and eruptive phenomena in its atmosphere, which accelerate plasma or particles beyond the escape speed. Inspecting the Sun, our nearest star, we observe two forms of mass loss: the steady solar wind outflow and the sporadic ejection of large plasma structures, or CMEs. The solar wind outflow amounts to \( \sim 2 \times 10^{-10} \text{(g/cm}^2 \text{s)} \) in coronal holes, and to \( \leq 4 \times 10^{-11} \text{(g/cm}^2 \text{s)} \) in active regions. The phenomenon of CME occurs with a frequency of about one event per day, carrying a mass in the range of \( m_{\text{CME}} \approx 10^{14}–10^{16} \text{ g} \).
which corresponds to an average mass loss rate of \( m_{\text{CME}} / (\Delta t \cdot 4\pi R_o^2 \approx 2) \times 10^{-14} - 2 \times 10^{-12} \text{g/cm}^2\text{s} \), which is \( \leq 1\% \) of the solar wind mass loss in coronal holes, or \( \leq 10\% \) of the solar wind mass in active regions. The transverse size of CMEs can cover a fraction up to more than a solar radius, and the ejection speed is in the range of \( v_{\text{CME}} \approx 10^2 - 10^3 \text{ (km/s)} \). A CME structure can have the geometric shape of a fluxrope, a semishell, or a bubble (like a light bulb, see Figure 11.24), which is the subject of much debate, because of ambiguities from line-of-sight projection effects and the optical thinness. Recent 3D reconstructions with data from the dual Solar Terrestrial Relationships Observatory (STEREO) spacecraft, however, clarified the 3D geometry of CMEs considerably. There is a general consensus that a CME is associated with a release of magnetic energy in the solar corona, but its relation to the flare phenomenon is controversial. Even big flares (at least GOES M-class) have no associated CMEs in 40% of the cases. A long-standing debate focused on the question of whether a CME is a by-product of the flare process, or vice versa. This question has been settled in the view that flares and CMEs are two aspects of a large-scale magnetic energy release, but the two terms evolved historically from two different observational manifestations (i.e., flares, which mainly denote the emission in hard X-rays, soft X-rays, and radio waves, and CMEs, which refer to the white-light emission of the erupting mass in the outer corona and heliosphere). Recent studies, however, clearly established the coevolution of both processes triggered by a common magnetic instability. A CME is a dynamically evolving plasma structure, propagating outward from the Sun into interplanetary space, carrying a frozen-in magnetic flux and expanding in size. If a CME structure travels toward the Earth, which is mostly the case when launched in the western solar hemisphere, due to the curvature of the Parker spiral interplanetary magnetic field, such an Earth-directed event can engulf the Earth’s magnetosphere and generate significant geomagnetic storms. Obviously such geomagnetic storms can cause disruptions of global communication and navigation networks, can cause failures of satellites and commercial power systems, and thus are the subject of high interest.

Theoretical CME models include at least seven categories: (1) thermal blast models, (2) dynamo models, (3) mass loading models, (4) tether release or straining models, (5) quadrupolar breakout models, and (6) kink or torus instability models. Numerical MHD simulations of CMEs are currently produced by combinations of a fine-scale grid that entails the corona and a connected large-scale grid that encompasses propagation into interplanetary space, which can reproduce CME speeds, densities, and the coarse geometry. The trigger that initiates the origin of a CME seems to be related to previous photospheric shear motion and subsequent kink instability of twisted structures (Figure 11.23). The geometry of CMEs is quite complex, exhibiting a variety of topological shapes from spherical semishells to helical fluxropes (Figure 11.24), and the density and temperature structure of CMEs is currently investigated with multiwavelength imagers. The height-time, velocity, and acceleration profiles of CMEs seem to establish two different CME classes: gradual CMEs associated with propagating interplanetary shocks and impulsive CMEs caused by coronal flares. The total energy of CMEs (i.e., the sum of magnetic, kinetic, and gravitational energy) seems to be conserved in some events, and the total energy of CMEs is comparable to the energy range estimated from flare signatures. A phenomenon closely associated with CMEs is coronal dimming (Figure 11.13),
which is interpreted in terms of an evacuation of coronal mass during the launch of a CME. The propagation of CMEs in interplanetary space provides diagnostic information on the heliospheric magnetic field, the solar wind, interplanetary shocks, solar energetic particle events, and interplanetary radio bursts.

7. FINAL COMMENTS

The study of the Sun, our nearest Star, is systematically moving from morphological observations (sunspots, active regions, filaments, flares, CMEs) to a more physics-based modeling and theoretical understanding, in terms of nuclear physics, magnetoconvection, MHD, magnetic reconnection, and particle physics processes. The major impact of physics-based modeling came from the multiwavelength observations from solar-dedicated space-based (Hinode, Solar Maximum Mission (SMM), Yohkoh, CGRO, SoHO, TRACE, RHESSI, Hinode, (STEREO, SDO, and ground-based instruments (in radio, Hz, and white-light wavelengths). Major achievements over the past decades are the advancement of new disciplines such as helioseismology and coronal seismology and the solution of the neutrino problem; however, there are still unsolved outstanding problems such as the coronal heating problem and particle acceleration mechanisms. We can optimistically expect substantial progress from future solar-dedicated space missions (Solar Orbiter, Solar Probe) and ground-based instruments (The Advanced Technology Solar Telescope (ATST) and the FASR telescope).
**BIBLIOGRAPHY**


The Solar Wind is a plasma, that is an ionized gas, that permeates interplanetary space. It exists as a consequence of the supersonic expansion of the Sun’s hot outer atmosphere, the solar corona. The solar wind consists primarily of electrons and protons, but alpha particles and many other ionic species are also present at low abundance levels. At the orbit of Earth, 1 Astronomical Unit (AU) from the Sun, typical solar wind densities, flow speeds, and temperatures are of the order of 8 protons/cm$^3$, 440 km/s, and $1.2 \times 10^5$ K respectively; however, the solar wind is highly variable in both space and time. A weak magnetic field embedded within the solar wind plasma is effective both in excluding some low-energy cosmic rays from the solar system and in channeling energetic particles from the Sun into the heliosphere. The solar wind plays an essential role in shaping and stimulating planetary magnetospheres and the ionic tails of comets. (See Planetary Magnetospheres.)

1. **DISCOVERY**

1.1. **Early Indirect Observations**

In 1859 R. Carrington made one of the first white light observations of a solar flare. He noted that a major geomagnetic storm began approximately 17 h after the flare and tentatively suggested that a causal relationship might exist between the solar and geomagnetic events. Subsequent observations revealed numerous examples of associations between solar flares and large geomagnetic storms. In the early 1900s F. Lindemann suggested that this could be explained if large geomagnetic storms result from an interaction between the geomagnetic field and plasma clouds ejected into interplanetary space by solar activity. Early studies of geomagnetic activity also noted that some geomagnetic storms tend to recur at the ~27 day rotation period of the Sun as observed from Earth, particularly during declining years of solar activity. This observation...
led to the suggestion that certain regions on the Sun, commonly called M (for magnetic)-regions, occasionally produce long-lived charged particle streams in interplanetary space. Furthermore, because some form of auroral and geomagnetic activity is almost always present at high geomagnetic latitudes, it was inferred that charged particles from the Sun almost continuously impact and perturb the geomagnetic field.

Observations of modulations in galactic cosmic rays in the 1930s also suggested that plasma and magnetic fields are ejected from the Sun during intervals of high solar activity. For example, S. Forbush noted that cosmic ray intensity often decreases suddenly during large geomagnetic storms, and then recovers slowly over a period of several days. Moreover, cosmic ray intensity varies in a cycle of \( \sim 11 \) years, but roughly \( 180^\circ \) out of phase with the solar activity cycle. One possible explanation of these observations was that magnetic fields embedded in plasma clouds from the Sun sweep cosmic rays away from the vicinity of Earth.

In the early 1950s, L. Biermann concluded that there must be a continuous outflow of charged particles from the Sun to explain the fact that ionized tails of comets always point away from the Sun. He estimated that a continuous particle flux of the order of \( 10^{10} \) protons/cm\(^2\)s was needed at 1 AU to explain the comet tail observations. He later revised his estimate downward to a value of \( \sim 10^9 \) protons/cm\(^2\)s, closer to the average observed solar wind proton flux of \( \sim 3.8 \times 10^8 \) protons/cm\(^2\)s at 1 AU.

1.2. Parker’s Solar Wind Model

Apparently inspired by these diverse observations and interpretations, E. Parker, in 1958, formulated a radically new model of the solar corona in which the solar atmosphere is continually expanding outward. Before Parker’s work most theories of the solar atmosphere treated the corona as static and gravitationally bound to the Sun except for sporadic outbursts of material into space at times of high solar activity. S. Chapman had constructed a model of a static solar corona in which heat transport was dominated by electron thermal conduction. For a 10\(^6\) K corona Chapman found that even a static solar corona must extend far out into space. Parker realized, however, that a static model leads to pressures at large distances from the Sun that are seven to eight orders of magnitude larger than estimated pressures in the interstellar plasma. Because of this mismatch at large heliocentric distances, he reasoned that the solar corona could not be in hydrostatic equilibrium and must therefore be expanding. His consideration of the hydrodynamic (i.e. fluid) equations for mass, momentum, and energy conservation for a hot solar corona led him to unique solutions for the coronal expansion that depended on the coronal temperature close to the surface of the Sun. Parker’s model produced low flow speeds close to the Sun, supersonic flow speeds far from the Sun, and vanishingly small pressures at large heliocentric distances. In view of the fluid character of the solutions, Parker called this continuous, supersonic coronal expansion the “solar wind”. The region of space filled by the solar wind is now known as the “heliosphere”.

1.3. First Direct Observations of the Solar Wind

Several Russian and American space probes in the 1959–1961 era penetrated interplanetary space and found tentative evidence for a solar wind. Firm proof of the wind’s existence was provided by C. Snyder and M. Neugebauer, who flew a plasma experiment on Mariner 2 during its epic 3-month journey to Venus in late 1962. Their experiment detected a continual outflow of plasma from the Sun that was highly variable, being structured into alternating streams of high- and low-speed flows that lasted for several days each. Several of the high-speed streams recurred at roughly the rotation period of the Sun. Average solar wind proton densities (normalized for a 1 AU heliocentric distance), flow speeds, and temperatures during this 3-month interval were 5.4 cm\(^{-3}\), 504 km/s, and \( 1.7 \times 10^5 \) K respectively, in essential agreement with Parker’s predictions. The Mariner 2 observations also showed that helium, in the form of alpha particles, is present in the solar wind in variable amounts; the average alpha particle abundance relative to protons of 4.6% being about a factor of 2 lower than estimates of the helium abundance within the Sun. Finally, measurements made by Mariner 2 confirmed that the solar wind carried a magnetic field whose strength and orientation in the ecliptic plane were much as predicted by Parker (see Section 3).

Despite the good agreement of observations with Parker’s model, we still do not fully understand the processes that heat the solar corona and accelerate the solar wind. Parker simply assumed that the corona is heated to a very high temperature, but he did not explain how the heating was accomplished. Moreover, it is now known that electron heat conduction is insufficient to power the coronal expansion. Present models for heating the corona and accelerating the solar wind generally fall into two classes: heating and acceleration by waves generated by convective motions below the photosphere; and bulk acceleration and heating associated with transient events in the solar atmosphere such as magnetic reconnection. Present observations are incapable of distinguishing between these and other alternatives.

2. STATISTICAL PROPERTIES IN THE ECLIPTIC PLANE AT 1 AU

Table 12.1 summarizes a number of statistical solar wind properties derived from spacecraft measurements in the
TABLE 12.1 Statistical Properties of the Solar Wind at 1 AU

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>STD</th>
<th>Most Probable</th>
<th>Median</th>
<th>5–95% Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$ (cm$^{-3}$)</td>
<td>8.7</td>
<td>6.6</td>
<td>5.0</td>
<td>6.9</td>
<td>3.0–20.0</td>
</tr>
<tr>
<td>$V_{sw}$ (km/s)</td>
<td>468</td>
<td>116</td>
<td>375</td>
<td>442</td>
<td>320–710</td>
</tr>
<tr>
<td>B (nT)</td>
<td>6.2</td>
<td>2.9</td>
<td>5.1</td>
<td>5.6</td>
<td>2.2–9.9</td>
</tr>
<tr>
<td>A(He)</td>
<td>0.047</td>
<td>0.019</td>
<td>0.048</td>
<td>0.047</td>
<td>0.017–0.078</td>
</tr>
<tr>
<td>$T_p$ ($\times 10^5$ K)</td>
<td>1.2</td>
<td>0.9</td>
<td>0.5</td>
<td>0.95</td>
<td>0.1–3.0</td>
</tr>
<tr>
<td>$T_e$ ($\times 10^5$ K)</td>
<td>1.4</td>
<td>0.4</td>
<td>1.2</td>
<td>1.33</td>
<td>0.9–2.0</td>
</tr>
<tr>
<td>$T_a$ ($\times 10^5$ K)</td>
<td>5.8</td>
<td>5.0</td>
<td>1.2</td>
<td>4.5</td>
<td>0.6–15.5</td>
</tr>
<tr>
<td>$T_e/T_p$</td>
<td>1.9</td>
<td>1.6</td>
<td>0.7</td>
<td>1.5</td>
<td>0.37–5.0</td>
</tr>
<tr>
<td>$nV_{sw}$ ($\times 10^8$/cm$^2$ s)</td>
<td>3.8</td>
<td>2.4</td>
<td>2.6</td>
<td>3.1</td>
<td>1.5–7.8</td>
</tr>
<tr>
<td>$C_s$ (km/s)</td>
<td>63</td>
<td>15</td>
<td>59</td>
<td>61</td>
<td>41–91</td>
</tr>
<tr>
<td>$C_A$ (km/s)</td>
<td>50</td>
<td>24</td>
<td>50</td>
<td>46</td>
<td>30–100</td>
</tr>
</tbody>
</table>

ecliptic plane at 1 AU. The table includes mean values, standard deviations about the mean values, most probable values, median values, and the 5–95% range limits for the proton number density ($n$), the flow speed ($V_{sw}$), the magnetic field strength ($B$), the alpha particle abundance relative to protons (A(He)), the proton temperature ($T_p$), the electron temperature ($T_e$), the alpha particle temperature ($T_a$), the ratio of the electron and proton temperatures ($T_e/T_p$), the ratio of alpha particle and proton temperatures ($T_a/T_p$), the number flux ($nV_{sw}$), the sound speed ($C_s$), and the Alfvén speed ($C_A$). All solar wind parameters exhibit considerable variability; moreover, variations in solar wind parameters are often coupled to one another. Proton temperatures are considerably more variable than electron temperatures, and alpha particle temperatures are almost always higher than electron and proton temperatures. Alpha particles and the protons tend to have nearly equal thermal speeds and therefore temperatures that differ by a factor of about 4. The solar wind flow is usually both supersonic and super-Alfvénic. Finally, we note that the Sun yearly loses ~6.8 x 10$^{19}$ g to the solar wind, a very small fraction of the total solar mass of ~2 x 10$^{33}$ g.

3. NATURE OF THE HELIOSPHERIC MAGNETIC FIELD

In addition to being a very good thermal conductor, the solar wind plasma is an excellent electrical conductor. The electrical conductivity of the plasma is so high that the solar magnetic field is “frozen” into the solar wind flow as it expands away from the Sun. Because the Sun rotates, magnetic field lines in the equatorial plane of the Sun are bent into Archimedeanspirals (Figure 12.1) whose inclinations relative to the radial direction depend on heliocentric distance and the speed of the wind. At 1 AU the average field line in the equatorial plane is inclined ~45° to the radial direction.

In Parker’s simple model the magnetic field lines out of the equatorial plane take the form of helices wrapped about cones of constant latitude. These helices are evermore elongated at higher solar latitudes and eventually approach radial lines over the solar poles. The equations describing Parker’s model of the magnetic field far from the Sun are

$$B_r(r, \phi, \theta) = B(r_0, \phi_0, \theta)(r_0/r)^2,$$

$$B_\phi(r, \phi, \theta) = -B(r_0, \phi_0, \theta)(\omega r_0^2/V_{sw}r)\sin \theta,$$

$$B_\theta = 0$$

Here $r$, $\phi$, and $\theta$ are radial distance, longitude, and latitude in a Sun-centered spherical coordinate system, respectively, $B_r$, $B_\phi$, and $B_\theta$ are the magnetic field components, $\omega$ is the Sun’s angular velocity (2.9 x 10$^{-6}$ radians/sec), $V_{sw}$ is the flow speed (assumed constant with distance from the Sun), and $\phi_0$ is an initial longitude at a reference distance $r_0$ from Sun center. This model is in reasonably good agreement with suitable averages of the heliospheric magnetic field measured over a wide range of heliocentric distances and latitudes. However, the instantaneous orientation of the
field usually deviates substantially from that of the model field at all distances and latitudes. Moreover, there is evidence that the magnetic field lines wander in latitude as they extend out into the heliosphere. This appears to be a result of field line foot point motions associated with differential solar rotation (the surface of the Sun rotates at different rates at different latitudes) and convective motions in the solar atmosphere.

4. CORONAL AND SOLAR WIND STREAM STRUCTURE

The solar corona is highly nonuniform, being structured by the complex solar magnetic field into arcades, rays, holes (regions relatively devoid of material), and streamers. (See The Sun.) The strength of the Sun’s magnetic field falls off sufficiently rapidly with height above the solar surface that it is incapable of containing the coronal expansion at altitudes above $\sim$0.5–1.0 solar radii. The resulting solar wind outflow produces the “combed-out” appearance of coronal structures above those heights in eclipse photographs.

As first seen in observations of the solar wind by Mariner 2 in the ecliptic plane, the solar wind is also highly nonuniform. In the ecliptic plane it tends to be organized into alternating streams of high- and low-speed flows. Figure 12.2 illustrates certain characteristic aspects of this stream structure. Five high-speed streams are clearly evident in the figure. The fourth and fifth streams were reencounters with the first and second streams, respectively, on the following solar rotation. Each high-speed stream was asymmetric with the speed rising more rapidly than it fell and each stream was essentially unipolar in the sense that $B_r$ was either positive or negative throughout the stream. Reversals in field polarity occurred in the low-speed flows between the streams. Those polarity reversals correspond to crossings of the heliospheric current sheet (discussed in more detail in the following section) that separates solar wind regions of opposite magnetic polarity. The magnetic field strength, proton temperature and density, and total pressure all peaked on the leading edges of the streams and the solar wind flow there was deflected first westward (positive flow azimuth) and then eastward. This pattern of variability is highly repeatable from one stream to the next and is the inevitable consequence of the evolution of the streams as they progress outward from the Sun (see Section 6).

**FIGURE 12.1** Configuration of the heliospheric magnetic field in the solar equatorial plane for a uniform radial solar wind flow.

**FIGURE 12.2** Solar wind stream structure at 1 AU for a 42-day interval in 2005. From top to bottom: 1-hr averages of solar wind proton density, proton temperature, total (plasma + field) pressure, bulk flow speed, flow azimuth angle (positive in the sense of Earth’s motion about the Sun), radial component of the heliospheric magnetic field, and field magnitude. Vertical lines mark crossing of the heliospheric current sheet. Plus and minus signs in the fourth panel indicate magnetic polarities, outward from and inward toward the Sun, respectively. Changes in magnetic field polarity are best determined by reversals in the solar wind suprathermal electron strahl flow polarity (parallel or antiparallel to the magnetic field; see Section 10.3). Adapted from Gosling (2010).
Recurrent high-speed streams originate primarily in coronal holes, which are large nearly unipolar regions in the solar atmosphere having relatively low density. Low-speed flows, on the other hand, tend to originate in the coronal streamer belt that straddles regions of magnetic field polarity reversals in the solar atmosphere. Both coronal and solar wind stream structures evolve considerably from one solar rotation to the next as the solar magnetic field, which controls that structure, continuously evolves. It is now clear that the mysterious M-regions, hypothesized long before the era of satellite X-ray observations of the Sun, are to be identified with coronal holes, and the long-lived particle streams responsible for recurrent geomagnetic activity are to be identified with high-speed solar wind streams. (See Sun–Earth Connection.)

5. THE HELIOSPHERIC CURRENT SHEET AND SOLAR LATITUDE EFFECTS

5.1. The Sun’s Large-Scale Magnetic Field and the Ballerina Skirt Model

During the declining phase of the solar activity cycle and near solar activity minimum the Sun’s large-scale magnetic field well above the photosphere often appears to be approximately that of a dipole. The solar magnetic dipole is tilted with respect to the Sun’s rotation axis; this tilt changes with the advance of the solar cycle. As illustrated in the left portion of Figure 12.3, near solar activity minimum the solar magnetic dipole tends to be aligned nearly with the rotation axis, while during the declining phase of activity it is generally inclined at a considerable angle relative to the rotation axis. Near solar maximum the Sun’s large-scale field is probably not well approximated by a dipole.

When the solar magnetic dipole and the solar rotation axis are closely aligned, the heliospheric current sheet, which is effectively the extension of the solar magnetic equator into the solar wind, coincides roughly with the solar equatorial plane. On the other hand, at times when the dipole is tilted substantially, the heliospheric current sheet is warped and resembles a ballerina’s twirling skirt, as illustrated in the right portion of Figure 12.3. Successive outward ridges in the current sheet (folds in the skirt) correspond to successive solar rotations and are separated radially by about 4.7 AU when the flow speed at the current sheet is 300 km/s. The maximum solar latitude of the current sheet in this simple picture is equal to the tilt angle of the magnetic dipole axis relative to the solar rotation axis.

5.2. Solar Latitude Effects

During the declining phase of the solar activity cycle and near solar activity minimum stream structure and solar wind variability are largely confined to a relatively narrow latitude band centered on the solar equator. This is illustrated in the upper left and right panels of Figure 12.4, which show solar wind speed as a function of solar latitude measured by Ulysses during the declining phases of the last two solar cycles. (Ulysses was in a ~6.2-year solar orbit that took it to solar latitudes of ±80°.) At this phase of the solar cycle the solar wind is dominated by stream structure at low latitudes, but flows at a nearly constant speed of ~850 km/s at high latitudes. This latitude effect is a consequence of the following: (1) solar wind properties change rapidly with distance from the heliospheric current sheet, with flow speed generally being a minimum in the vicinity of the current sheet; and (2) the heliospheric current sheet is commonly tilted relative to the solar equator, but is usually found within about ±30° of it during the declining phase of the solar cycle. The width of the band of solar wind variability changes as the solar magnetic dipole tilt changes. The upper middle panel of Figure 12.4 demonstrates that, in contrast, in the years surrounding solar activity maximum the band of solar wind variability extends up to the highest latitudes sampled by Ulysses.
6. EVOLUTION OF STREAM STRUCTURE WITH HELIOCENTRIC DISTANCE

6.1. Kinematic Stream Steepening and the Dynamic Response

Because the coronal expansion is spatially variable, at low latitudes alternately slow and fast plasma is directed outward along any radial line from the Sun as the Sun rotates (with a period of 27 days as seen from Earth). Faster moving plasma overtakes slower moving plasma ahead while outrunning slower moving plasma behind. Because radially aligned parcels of plasma within a stream originate from different locations on the Sun, they are threaded by different magnetic field lines and thus cannot interpenetrate one another except during relatively infrequent magnetic reconnection events. The result is that the leading edges of high-speed streams steepen with increasing distance from the Sun, producing the asymmetric stream profiles obvious in Figure 12.2. As the streams steepen, plasma and field on the leading edge of a stream is compressed, causing an increase in plasma density, temperature, field strength and pressure there, while plasma and field on the trailing edge becomes increasingly rarefied. The buildup of pressure on the leading edge of a stream produces forces that accelerate the low-speed wind ahead and decelerate the high-speed wind within the stream itself. The net result is a transfer of momentum and energy from the fast-moving wind to the slow-moving wind.

6.2. Shock Formation

As long as the amplitude of a high-speed solar wind stream is sufficiently small, it gradually damps with increasing heliocentric distance in the manner just described. However, when the difference in flow speed between the crest of a stream and the trough ahead is greater than about twice the local fast mode speed, $C_f$ (the fast mode speed is the characteristic speed with which small amplitude pressure signals propagate in a plasma: $C_f = \left( C_s^2 + C_A^2 \right)^{0.5}$), ordinary pressure signals do not propagate sufficiently fast to move the slow wind out of the path of the oncoming high-speed stream. In that case the pressure eventually increases nonlinearly and shock waves form on either side of the high-pressure region (see Figure 12.5). The leading shock, known as a forward shock, propagates into the low-speed wind ahead and the trailing shock, known as a reverse shock, propagates back through the stream. Both shocks are, however, convected away from the Sun by the highly supersonic and super-Alfvénic flow of the wind. The major accelerations and decelerations associated with stream evolution then occur discontinuously at the shocks, giving a stream speed profile the appearance of a double sawtooth wave. The stream amplitude decreases and the compression region expands with increasing heliocentric distance as the shocks propagate. Observations indicate that the shocks often do not form until the streams are well beyond 1 AU. Nevertheless, because $C_f$ generally decreases with
increasing heliocentric distance, virtually all large-amplitude solar wind streams steepen into shock wave structures at heliocentric distances beyond $\sim 3$ AU. At heliocentric distances immediately beyond the orbit of Jupiter ($\sim 5.4$ AU) a large fraction of the mass in the solar wind is found within compression regions bounded by shock waves on the rising portions of damped high-speed streams. The basic structure of the solar wind in the solar equatorial plane in the distant heliosphere thus differs considerably from that observed at 1 AU. Stream amplitudes are severely reduced and short wavelength structure is damped out. The dominant structures at low latitudes in the outer heliosphere are expanding compression regions that interact and merge with one another to form what are commonly called global merged interaction regions.

### 6.3. Stream Evolution in Two and Three Dimensions

When the coronal expansion is spatially variable but time stationary, a steady flow pattern such as sketched in Figure 12.6 develops in the equatorial plane. This entire pattern corotates with the Sun and the compression regions are known as corotating interaction regions, or CIRs; however, only the pattern rotates—each parcel of solar wind plasma moves outward nearly radially as indicated by the black arrows. The region of high pressure associated with a CIR is nearly aligned with the magnetic field line spirals in the equatorial plane and the pressure gradients are thus nearly perpendicular to those spirals. Consequently, at 1 AU the pressure gradients that form on the rising speed portions of high-speed streams have transverse as well as radial components. In particular, not only is the low-speed plasma ahead of a high-speed stream accelerated to a higher speed, but also it is deflected in the direction of solar rotation. In contrast, the high-speed plasma near the crest of the stream is both decelerated and deflected in the direction opposite to solar rotation. These transverse deflections produce the systematic west-east flow direction changes observed near the leading edges of quasi-stationary high-speed streams (see Figure 12.2).

There is an interesting three-dimensional aspect to stream evolution, ultimately associated with the fact that the solar magnetic dipole typically is tilted relative to the solar rotation axis. That tilt causes CIRs in the northern and southern solar hemispheres to have opposed meridional tilts that, particularly beyond about 3 AU, can be discerned in plasma data as systematic north-south deflections of the flow at CIRs. The meridional tilts are such that the forward waves in both hemispheres propagate toward the opposite
hemispheres, whereas the reverse waves in both hemispheres propagate poleward. As a result, forward shocks in the outer heliosphere near solar minimum are generally confined to the low-latitude band of solar wind variability, while the reverse shocks are commonly observed both within the band of variability and poleward of it. However, the reverse waves seldom reach latitudes more than $\sim 15^\circ$ above the low-latitude band of variability.

7. CORONAL MASS EJECTIONS AND TRANSIENT SOLAR WIND DISTURBANCES

7.1. Coronal Mass Ejections

The solar corona evolves on a variety of timescales closely connected with the evolution of the coronal magnetic field. (See The Sun.) The most rapid and dramatic evolution in the corona occurs in events known as coronal mass ejections or CMEs (Figure 12.7(a)). CMEs originate in closed field regions in the corona where the magnetic field normally is sufficiently strong to constrain the coronal plasma from expanding outward. Typically, these closed field regions are found in the coronal streamer belt that encircles the Sun and that underlies the heliospheric current sheet. The outer edges of CMEs often have the optical appearance of closed loops such as the event shown in Figure 12.7(a). Few CMEs ever appear to sever completely their magnetic connection with the Sun. During a typical CME, somewhere between $10^{15}$ and $10^{16}$ g is ejected into the heliosphere. Ejection speeds near the Sun range from less than 50 km/s in some of the slower events to greater than 2500 km/s in the fastest ones. The average CME speed at $\sim 5$ solar radii is close to the median ecliptic solar wind speed of $\sim 440$ km/s. Since observed solar wind speeds near 1 AU are never less than $\sim 280$ km/s, the slowest CMEs are further accelerated enroute to 1 AU.

7.2. Origins, Associations with Other Forms of Solar Activity, and Frequency of Occurrence

The processes that trigger CMEs and that determine their sizes and outward speeds are only poorly understood; there is presently no consensus on the physical processes responsible for initiating or accelerating these events, although it is clear that stressed magnetic fields are the underlying cause of these events and that CMEs play a fundamental role in the long-term evolution of the structure of the solar corona. They appear to be an essential part of the way the corona responds to the evolution of the solar magnetic field associated with the advance of the solar activity cycle. Indeed, the release of a CME is one way that the solar atmosphere reconfigures itself in response to changes in the solar magnetic field. CMEs are commonly, but not always, observed in association with other forms of solar activity such as eruptive prominences and solar flares.

---

**FIGURE 12.7** (a) A CME as imaged by the LASCO/C3 coronagraph on SOHO on 20 April 1998. The Sun, indicated by the white circle, has been occulted within the instrument. The field of view of the image is 30 solar diameters. (The SOHO/LASCO data are produced by a consortium of the Naval Research Laboratory (USA), Max-Planck-Institut fur Sonnensystemforschung (Germany), Laboratoire d’Astrophysique de Marseille (France), and the University of Birmingham (UK). SOHO is a project of international cooperation between ESA and NASA.) (b) A sketch of a solar wind shock disturbance produced by a fast ICME directed toward Earth. Red and magenta arrows indicate the ambient magnetic field and that threading the ICME, respectively. Blue arrows indicate the suprathermal electron strahl flowing away from the Sun along the magnetic field. The ambient magnetic field is compressed by its interaction with the ICME and is forced to drape around the ICME. Adapted from a sketch originally in Crooker (2000) and from Zurbuchen and Richardson (2006).
From a historical perspective one might be led to expect that large solar flares are the prime cause of CMEs; however, it is now clear that flares and CMEs are separate, but closely related, phenomena associated with magnetic disturbances on the Sun. Like other forms of solar activity, CMEs occur with a frequency that varies in a cycle of \( \sim 11 \) years. On average, the Sun emits about 3.5 CMEs/day near the peak of the solar activity cycle, but only about 1 CME every 10 days near solar activity minimum.

### 7.3. Heliospheric Disturbances Driven by Fast CMEs

As illustrated in Figure 12.7(b), fast CMEs produce transient solar wind disturbances that, in turn, often are the cause of large geomagnetic storms and major space weather events in general. (See Sun—Earth Connections.) Figure 12.8 shows calculated radial speed and pressure profiles of a simulated solar wind disturbance driven by a fast CME at the time the disturbance first reaches 1 AU. As indicated by the insert in the top portion of the figure, the disturbance was initiated at the inner boundary of the one-dimensional fluid calculation by abruptly raising the flow speed from 275 to 980 km/s, sustaining it at this level for 6 h, and then returning it to its original value of 275 km/s. The initial disturbance thus mimics a uniformly fast spatially limited CME with an internal pressure equal to that of the surrounding solar wind plasma. A region of high pressure develops on the leading edge of the disturbance as the CME overtakes the slower wind ahead. This region of higher pressure is bounded by a forward shock on its leading edge that propagates into the ambient solar wind ahead and by a reverse shock on its trailing edge that propagates backward into and eventually through the CME. Both shocks are, however, carried away from the Sun by the highly supersonic flow of the solar wind. A rarefaction forms on the trailing edge of the disturbance as the CME outruns slower plasma behind. This rarefaction produces a deceleration of the rear portion of the CME and an acceleration of the trailing wind. The overall interaction with the ambient wind produces an expanding CME whose radial width at \( \sim 0.8 \) AU is greater than its width when introduced into the simulation.

Except for the reverse shock, which observations and more detailed three-dimensional calculations indicate are ordinarily present only near the central portions of the disturbances, the simple calculation shown in Figure 12.8 is consistent with observations of many solar wind disturbances obtained near 1 AU in the ecliptic plane and illustrates to first order the radial and temporal evolution of an interplanetary disturbance driven by a fast CME (now commonly called an interplanetary coronal mass ejection, ICME, when observed in the solar wind). In the example illustrated, the ICME slows from an initial speed of 980 km/s to less than 600 km/s by the time the leading edge of the disturbance reaches 1 AU. This slowing is a result of momentum transfer to the ambient solar wind ahead and behind and proceeds at an ever-slower rate as the disturbance propagates outward. Figure 12.9 displays selected plasma and magnetic field data from a solar wind disturbance driven by an ICME observed near 1 AU. The shock is distinguished in the data by discontinuous increases in flow speed, density, temperature and field strength. The plasma identified as the ICME had a higher flow speed than the ambient solar wind ahead of the shock. In this case it was also distinguished by counterstreaming suprathermal electrons (indicative of a closed magnetic field topology, see Section 10.3), anomalously low proton temperatures, somewhat elevated helium abundance, and a strong smoothly rotating magnetic field that indicates that the field topology was that of a magnetic flux rope (see Figure 12.7(b)).
7.4. Characteristics of ICMEs

The identification of ICMEs in solar wind plasma and field data is still something of an art; however, shocks serve as useful fiducials for identifying fast ICMEs. Table 12.2 provides a summary of plasma and field signatures that qualify as unusual compared with the normal solar wind, but that are commonly observed as a number of hours after shock passage. Most of these anomalous signatures are observed elsewhere in the solar wind as well where, presumably, they serve to identify those numerous relatively low-speed ICMEs that do not drive shock disturbances. Few ICMEs at 1 AU exhibit all of these characteristics, and some of these signatures are more commonly observed than are others.

Most ICMEs expand as they propagate outward through the heliosphere. ICME radial thicknesses are variable; at 1 AU the typical ICME has a radial width of ~0.2 AU whereas at Jupiter’s orbit ICMEs can have radial widths as large as 2.5 AU. Magnetic reconnection occurs relatively rarely at the leading and/or trailing edges of ICMEs, but when it does occur there it erodes away portions of those ICMEs. Approximately one-third of all ICMEs in the ecliptic plane have sufficiently high speeds relative to the ambient solar wind to drive shock disturbances at 1 AU; the slower ICMEs do not drive shock disturbances and simply coast along with the rest of the solar wind. Typically ICMEs cannot be distinguished from the normal solar wind at 1 AU on the basis of either their speed or density (the event in Figure 12.9 is an example). Near solar activity maximum ICMEs account for 15–20% of the solar wind in the ecliptic plane at 1 AU, while near solar activity minimum they account for less than 1%. The Earth intercepts about 72 ICMEs/yr near solar activity maximum and ~8 ICMEs/yr near solar activity minimum. ICMEs are much less common at high heliographic latitudes, particularly near activity minimum when ICMEs are confined largely to the low-latitude band of solar wind variability.

7.5. The Magnetic Field Topology of ICMEs and the Problem of Magnetic Flux Balance

The coronal expansion carries a portion of the solar magnetic field outward to form the heliospheric magnetic field. In the quiescent wind these field lines are usually “open” in
the sense that they connect to field lines of the opposite polarity only in the very distant heliosphere. CMEs, on the other hand, originate in closed field regions in the corona not previously participating directly in the solar wind expansion and inject new closed magnetic flux into the heliosphere. In the absence of magnetic reconnection, such injections would lead to a continual buildup of magnetic flux in the heliosphere, which is not observed: measurements reveal that solar rotation averages of the heliospheric magnetic field strength in the ecliptic plane at 1 AU vary by a factor of about 2 roughly in phase with the 11-year solar activity cycle. That variation is determined by the competition between flux ejection into the heliosphere by CMEs and flux removal by reconnection in the magnetic legs of CMEs and at the heliospheric current sheet. Such reconnection is effective in reducing the magnetic flux in the heliosphere only when it occurs sunward of the Alfvén point where the solar wind flow becomes super-Alfvénic. It remains to be determined which of these reconnection sites is dominant in balancing the new flux ejected into the heliosphere by CMEs.

Figure 12.10 illustrates that reconnection within the magnetic legs of CMEs is inherently three-dimensional in nature and, when occurring between adjacent magnetic loops, produces helical magnetic field lines that are partially disconnected from the Sun as well as new closed field lines relatively low in the solar atmosphere. Sustained three-dimensional magnetic reconnection in the magnetic legs eventually produces a mixture of closed, open, and disconnected field lines threading an ICME as well as additional closed magnetic loops low in the solar atmosphere. All of the types of reconnection illustrated in Figure 12.10 reduce the amount of magnetic flux that a CME adds to the heliosphere and all of the magnetic topologies produced by such reconnection are apparent in suprathermal electron observations of various ICMEs at 1 AU. Recent sequences of coronagraph and heliospheric images indicate that reconnection also commonly occurs at the heliospheric current sheet inside the Alfvén point, producing heliospheric field lines that are disconnected from the Sun.

7.6. Field Line Draping About Fast ICMEs

Because ICME plasma and ambient wind plasma are threaded by different field lines, they cannot, in general, interpenetrate one another. Consequently, the ambient plasma and magnetic field ahead must be deflected away from the path of a fast ICME in much the same manner as the solar wind is deflected around Earth’s magnetosphere. Figure 12.7(b) illustrates that such deflections cause the ambient magnetic field to drape about the ICME. The degree of draping and the resulting orientation of the field ahead of an ICME depend upon the relative speed between the ICME and the ambient plasma, the shape of the ICME and the original orientation of the magnetic field in the ambient plasma. Draping plays an important role in reorienting the magnetic field ahead of a fast ICME. On the other hand, conditions and processes back at the Sun largely determine field orientations within ICMEs. As a final point of interest, Figure 12.7(b) also illustrates that, just as the bow wave in front of a boat moving through water is considerably broader in extent than is the boat that produces it, so too is the shock in front of a fast ICME somewhat broader in extent than is the ICME that drives it. As a result, spacecraft often encounter ICME-driven shocks without also encountering the ICMEs that drive them.

8. VARIATION WITH DISTANCE FROM THE SUN

For a structureless solar wind, the speed remains nearly constant beyond the orbit of Earth, the density falls off with heliocentric distance, \( r \), as \( r^{-2} \), and the magnetic field decreases with distance as described by the equations in Section 2. The temperature also decreases with increasing heliocentric distance due to the spherical expansion of the plasma; however, the precise nature of the decrease depends upon particle species and the relative importance of such things as collisions, heat conduction, turbulence dissipation, and plasma instabilities. Protons and electrons evolve differently with increasing heliocentric distance. For an adiabatic expansion of an isotropic plasma the temperature falls off as \( r^{-4/3} \); for a plasma dominated by heat conduction the temperature falls as \( r^{-27} \). Observations reveal that both proton and electron temperatures inside
5 AU decrease with distance somewhere between the adiabatic and conduction-dominated extremes.

Of course, the solar wind is not structureless. The continual interaction of high- and low-speed flows with increasing heliocentric distance produces a radial variation of speed that differs considerably from that predicted for a structureless wind. High-speed flows decelerate and low-speed flows accelerate with increasing heliocentric distance as a result of momentum transfer (see Sections 6 and 7). Consequently, at low solar latitudes far from the Sun (beyond \(\sim 15\) AU) the solar wind flows at 400–500 km/s most of the time (Figure 12.11). Only rarely are substantial speed perturbations observed at these distances; these relatively rare events usually are associated with disturbances driven by very large and fast ICMEs that require a greater-than-usual distance to share their momentum with lower speed wind.

9. TERMINATION OF THE SOLAR WIND

Interstellar space is filled with a dilute gas of neutral and ionized particles and is threaded by a weak magnetic field. In the absence of the solar wind, the interstellar plasma would penetrate deep into the solar system in the same fashion as do interstellar neutral particles. However, because of the magnetic fields embedded in both, the interstellar and solar wind plasmas cannot easily interpenetrate one another. The result is that the solar wind creates a cavity in the interstellar plasma; however, magnetic reconnection at the boundary between the two plasmas may allow limited interpenetration of the interstellar and solar wind plasmas in thin layers there.

The details of the solar wind’s interaction with the interstellar plasma are still somewhat speculative largely because, until recently, we lacked direct observations of this interaction. Figure 12.12 shows what are believed to be the major elements of the interaction. The Sun and heliosphere move at a speed of \(\sim 23\) km/s relative to the interstellar medium. A bow wave must stand in the interstellar plasma upstream of the heliosphere to initiate the slowing and deflection of the plasma around the heliosphere. Recent remote measurements indicate that this bow wave is probably not a shock, although labeled that way in Figure 12.12. The heliopause is the outermost boundary of the heliosphere. Sunward of the heliopause is a termination shock where the solar wind flow becomes subsonic so that it ultimately can be turned to flow roughly parallel to the heliopause. Direct observations of the termination shock by the two Voyager spacecraft reveal that the termination shock is unusual in the sense that the bulk solar wind plasma is

---

**FIGURE 12.11** Solar wind speed as a function of time as measured by Voyager 2 during a 1.5-year interval when the spacecraft was beyond 18 AU from the Sun. Because stream amplitudes are severely damped at large distances from the Sun, the solar wind speed there generally varies within a very narrow range of values. Compare with the speed variations evident in Figure 12.2 that were obtained at 1 AU during a comparable period of the solar cycle. *Adapted from Lazarus and Belcher (1988).*
significantly slowed as it crosses the shock but is only slightly heated. A large fraction of the energy associated with the bulk plasma motion lost at the shock is transferred to suprathermal and energetic particles, which dominate the solar wind’s internal pressure in the outer heliosphere.

The shape of the heliosphere is asymmetric because of its motion relative to the interstellar gas; it is compressed in the direction of that motion and is greatly elongated in the opposite direction. The shape is also affected by the draping of the interstellar magnetic field about the heliopause. Observations in the outer heliosphere indicate that the termination shock is constantly in motion relative to the Sun, owing at least in part to an ever-changing solar wind momentum flux. The size and shape of the heliosphere depend on the momentum flux carried by the solar wind, the dynamic pressure of the interstellar plasma, the strength and orientation of the interstellar magnetic field, and the motion of the heliosphere relative to the interstellar medium. Voyager 1 crossed the termination shock in December 2004 at a heliocentric distance of about 94 AU; Voyager 2 crossed it in August 2007 at a distance of about 84 AU. Both Voyager trajectories are directed within about 45° of the heliosphere’s motion relative to the interstellar medium.

It is currently believed that the nose of the heliopause lies at a heliocentric distance of about 120 AU.

Voyager 1 observed significant and, ultimately, long-lasting changes in the magnetic field and in energetic particle populations beginning on 26 July 2012 at a heliocentric distance of ~121 AU. A recent interpretation of those observations that invokes reconnection between the heliospheric and interstellar magnetic fields suggests that those changes signaled Voyager 1’s crossing of the heliopause. That interpretation is consistent with subsequent plasma wave measurements of large increases in plasma density after 26 July 2012.

10. KINETIC PROPERTIES OF THE PLASMA

10.1. The Solar Wind as a Marginally Collisional Plasma

On a large scale the solar wind behaves like a compressible fluid and is capable of supporting relatively thin structures such as shocks. It is perhaps not obvious why the solar wind
should exhibit this fluid-like behavior since the wind is a dilute plasma in which collisions are relatively rare. For example, using values given in Table 12.1, we find that the time between collisions for a typical solar wind proton at 1 AU is several days. (These collisions do not result from direct particle impacts such as colliding billiard balls, but rather from the long distance Coulomb interactions characteristic of charged particles.) The time between collisions is thus comparable with or less than the local Alfvén speed, suggesting that the streaming is limited by an ion beam streaming instability. Closer to the Sun, where the Alfvén speed is higher, relative streaming speeds between the beams and the main components can be as large as several 100 km/s. Secondary proton beams are common in the solar wind in both low- and high-speed flows and may play a role in the overall acceleration and heating of the wind; however, their origin in solar and/or heliospheric processes is presently uncertain. Figure 12.13(b) illustrates that solar wind ion distributions in the low-speed wind also commonly have extended nonthermal tails of uncertain origin. Particles in these extended tails are easily accelerated to much higher energy when they encounter shocks (see Section 12).

10.3. Kinetic Aspects of Solar Wind Electrons

Electron distributions in the solar wind consist of a relatively cold and dense thermal “core” population that is electrically bound to the solar wind ion population and a much hotter and freer running suprathermal population that becomes collisionless close to the Sun. At 1 AU the breakpoint between these populations typically occurs at an energy of \( \sim 70 \text{ eV} \) (Figure 12.14(a)). This breakpoint moves steadily to lower energies with increasing heliocentric distance as the core population cools. Typically the core contains about 95% of the electrons, and at 1 AU has a temperature of \( \sim 1.3 \times 10^5 \text{ K} \). The core electrons typically are mildly anisotropic, with the temperature parallel to the field exceeding the temperature perpendicular to the field by a factor of \( \sim 1.1 \) on average at 1 AU. However, the temperature anisotropy for core electrons varies systematically with density such that at very low densities (<2 cm\(^{-3}\)) the temperature ratio often exceeds 2.0, while at very high densities (>10 cm\(^{-3}\)) the temperature ratio is often slightly less than 1.0. Such systematic variations of core electron temperature anisotropy with plasma density reflect the marginally collisional nature of the thermal electrons and their nearly adiabatic expansion in the spiral magnetic field.

The suprathermal electrons consist of a beam of variable width and intensity, known as the “strahl,” directed outward from the Sun along the heliospheric magnetic field and a more tenuous and often roughly isotropic “halo” (Figure 12.14(b)). The angular width of the strahl results from a competition between focusing associated with conservation of an electron’s magnetic moment in the diverging heliospheric magnetic field and defocusing associated with particle scattering. The strahl carries the solar wind electron heat flux; variations in strahl intensity largely reflect spatial variations in the corona from which it arises. In addition, brief (hours) strahl intensifications

10.2. Kinetic Aspects of Solar Wind Ions

Collisional gases can usually be described by a single isotropic (i.e., the same in all directions) temperature, \( T \), with the distribution of particle speeds, \( v \), obeying the Maxwellian \( f(v) \sim \exp(-m(v-v_o)^2/2kT) \), where \( f \) is the number of particles per unit volume of velocity space, \( k \) is Boltzmann’s constant \((1.38 \times 10^{-23} \text{ J/K})\), \( m \) is the particle mass and \( v_o \) is the bulk speed of the gas. In contrast, proton distribution functions in the solar wind are usually anisotropic because of the paucity of collisions and because the magnetic field provides a preferred direction in space. Moreover, solar wind proton and alpha particle distributions often exhibit significant non-Maxwellian features such as the double-peaked distributions illustrated in Figure 12.13(a). The secondary proton and alpha particle peaks are associated with beams streaming relative to the main solar wind component along the heliospheric magnetic field. The relative streaming speed of such beams is usually comparable with or less than the
FIGURE 12.13 (a) A cut through a solar wind ion count spectrum parallel to the magnetic field. The first two peaks are protons and the second two peaks are alpha particles. (The velocity scale for the alpha particles has been increased by a factor of 1.4.) Both the proton and alpha particle spectra show clear evidence for a secondary beam of particles streaming along the field relative to the main solar wind beam at about the Alfvén speed. Such secondary beams, not always well resolved, are common in both the low and the high-speed wind (Asbridge et al., 1974). (b) Solar wind speed distributions of $\text{H}^+$, $\text{He}^{++}$ and $\text{He}^+$ observed in the low-speed solar wind at 1 AU, averaged over a 65-day period in 1998 and excluding intervals of shocks and other disturbances. Such extended suprathermal tails appear to be ubiquitous in the low-speed solar wind. The $\text{He}^+$ ions are primarily of interstellar origin. From Gloeckler et al. (2000).

FIGURE 12.14 (a) One-dimensional cut through a solar wind electron distribution showing the thermal and suprathermal populations. (b) Suprathermal electron pitch angle distribution (relative to the magnetic field) showing the field-aligned strahl and a nearly isotropic halo.
commonly occur during solar electron bursts associated with solar activity (see Section 12). The strahl serves as an effective tracer of magnetic field topology in the interplanetary medium since its usual unidirectional nature arises because field lines in the normal solar wind are “open” (see Section 7.5) and are thus effectively connected to the solar corona at only one end. In contrast, field lines threading ICMEs are often attached to the Sun at both ends (see Sections 7.4 and 7.5), and counterstreaming strahls are commonly observed there. Indeed, counterstreaming strahls are one of the more reliable signatures of ICMEs (see Figures 12.7(b) and 12.9 and Table 12.2). Finally, at 1 AU the electron halo results primarily from backscattered strahl electrons from distances beyond 1 AU. Those backscattered electrons are subsequently mirrored (i.e. magnetically reflected) inside 1 AU by the stronger magnetic fields that reside there, producing a halo population that often is roughly isotropic, as in the example shown in Figure 12.14(b).

11. HEAVY ION CONTENT

Although the solar wind consists primarily of protons (hydrogen), electrons, and alpha particles (doubly ionized helium), it also contains traces of ions of a number of heavier elements. Table 12.3 provides estimates of the relative abundances of some of the more common solar wind elements summed over all ionization states. After hydrogen and helium, the most abundant elements are carbon and oxygen. The ionization states of all solar wind ions are “frozen in” close to the Sun because the characteristic times for ionization and recombination are long compared with the solar wind expansion time. Commonly observed ionization states include He\(^{2+}\), C\(^{5+}\), C\(^{6+}\), O\(^{6+}\)--O\(^{8+}\), Si\(^{7+}\)--Si\(^{10+}\), and Fe\(^{8+}\)--Fe\(^{14+}\). Ionization state temperatures in the low-speed wind are typically in the range 1.4\(\times\)1.6 \(\times\) 10\(^6\) K, while ionization state temperatures in the high-speed wind are typically in the range 1.0\(\times\)1.2 \(\times\) 10\(^6\) K. Unusual ionization states such as Fe\(^{16+}\) and He\(^{1+}\), which are not common in the normal solar wind, are often abundant within ICMEs, reflecting the unusual coronal origins of those events.

The relative abundance values in Table 12.3 are long-term averages; however, abundances vary considerably with time. Such variations have been extensively studied for the He\(^{2+}\)/H\(^{+}\) ratio, A(He). The most probable A(He) value is \(\sim\)0.045, but the A(He) ranges from less than 0.01 to values of 0.35 on occasion. The average A(He) is about twice that commonly attributed to the solar interior, for reasons presently unknown. Much of the variation in A(He) and in the abundance of heavier elements is related to the large-scale structure of the wind. For example, Fe/O and Mg/O ratios are systematically lower in high-speed streams than in low-speed flows. A(He) tends to be relatively constant at \(\sim\)0.045 within the cores of high-speed streams from coronal holes, but tends to be highly variable within low-speed flows. Particularly low (<0.02) abundance values are commonly observed in the vicinity of the heliospheric current sheet. A(He) values greater than about 0.10 are relatively rare and account for less than 1% of all the measurements. At 1 AU enhancements in A(He) above 0.10 occur almost exclusively within ICMEs. The physical causes of these variations are uncertain for the most part, although thermal diffusion, gravitational settling, and Coulomb friction in the chromosphere and corona all probably play roles.

| TABLE 12.3 Average Elemental Abundances in the Solar Wind |
|-----------------|------------------|
| Element | Abundance Relative to Oxygen |
| H | 1900 ± 400 |
| He | 75 ± 20 |
| C | 0.67 ± 0.10 |
| N | 0.15 ± 0.06 |
| O | 1.00 |
| Ne | 0.17 ± 0.02 |
| Mg | 0.15 ± 0.02 |
| Si | 0.19 ± 0.04 |
| Ar | 0.0040 ± 0.0010 |
| Fe | 0.19 + 0.10 – 0.07 |

12. ENERGETIC PARTICLES

A proton moving with a speed of 440 km/s has an energy of \(\sim\)1 keV. Thus, by most measures solar wind ions are low-energy particles. The heliosphere is, nevertheless, filled with a number of energetic ion populations of varying intensities with energies ranging upward from \(\sim\)1 keV/nucleon to \(\sim\)10\(^8\) keV/nucleon. These populations include galactic cosmic rays, anomalous cosmic rays (see discussion that follows), and energetic particles associated with CIRs, CMEs, solar flares, and the planetary bow shocks. All but the galactic cosmic rays are energized within the heliosphere.

Shocks are particularly effective particle accelerators and all but one of the above populations have shock origins. The physical process by which a collisionless shock accelerates a small fraction of the ions it intercepts to high energy is reasonably well understood, although complex
Recent work indicates that shocks in the solar wind most easily accelerate ions that already exceed solar wind thermal energies when they encounter the shocks. These so-called “seed” particles include the suprathermal ion tails always present in the low-speed wind (Figure 12.13(b)), but also “pickup ions” (see discussion that follows), and energetic particles remaining in the heliosphere from previous solar flares and CME-driven disturbances.

Anomalous cosmic rays have energies per nucleon that are lower than that of galactic cosmic rays, are predominantly singly ionized H, He, N, O, and Ne and, like galactic cosmic rays, have an intensity that varies slowly with time. They are associated with a particularly interesting seed population—neutral atoms from the local interstellar cloud that penetrate deep into the heliosphere. As the neutrals approach the Sun some of them are ionized by solar EUV radiation, electron impact, or charge exchange with solar wind protons, are then picked up by the solar wind magnetic field (the pickup process accelerates them to ~4 keV/nucleon), and are swept into the outer reaches of the heliosphere by the solar wind flow. It long has been thought that the pickup ions are accelerated to high energies as they encounter the termination shock and then diffuse back into the interior of the heliosphere as anomalous cosmic rays. However, this idea was questioned when the Voyager 1 spacecraft observed that anomalous cosmic ray intensities continued to increase well after the spacecraft crossed the termination shock. One possible explanation for this unexpected result is that the pickup ions are accelerated primarily along the far flanks of the termination shock where the Archimedean spiral magnetic field line connection times are considerably longer than in the region around the nose of the shock where the Voyagers crossed it.

Of the energetic ion populations in the heliosphere that associated directly with solar flares appears to be the only population that is not obviously shock associated, although even in this case shock acceleration cannot be ruled out conclusively. Flare events are usually impulsive and short lived (hours), are overabundant in $^3$He, appear to originate relatively low in the solar atmosphere, occur at a rate of ~1000 events/year near solar activity maximum and generally occur in association with impulsive energetic solar electron bursts. The latter have energies ranging from several hundred eV up to several hundred keV. Recent work suggests that solar electron bursts originate at a variety of altitudes in the solar atmosphere and can be triggered by more than one process.

13. TURBULENCE AND MAGNETIC FIELD AND VELOCITY FLUCTUATIONS

Figure 12.15 illustrates that fluctuations in velocity and magnetic field are observed throughout the solar wind at 1 AU on a variety of spatial and temporal scales; however, fluctuation amplitudes tend to be greatest on the rising speed portions and within the cores of high-speed streams from coronal holes. As illustrated in Figure 12.16(a), velocity and magnetic field fluctuations in the high-speed wind are largely Alfvénic (coupled changes in velocity and magnetic field components). These Alfvénic fluctuations propagate predominantly anti-sunward in the solar wind rest frame, indicating that they are largely remnants of Alfvénic fluctuations present close to the Sun. In contrast, as illustrated in Figure 12.16(b), in the low-speed wind and on the rising speed portions of high-speed streams fluctuations in velocity and magnetic field often are not coupled to one another. Presumably, such non-Alfvénic fluctuations do not propagate in the solar wind rest frame. Particularly sharp changes in field orientation, such as that at 1853 UT in Figure 12.16(b), are where magnetic reconnection commonly occurs in the solar wind. Both Alfvénic and non-Alfvénic fluctuations typically appear to be stochastic in nature, reflecting the turbulent nature of the solar wind flow, and have amplitudes that decrease with increasing heliocentric distance. The turbulent nature of the fluctuations is graphically illustrated in Figure 12.17, which shows a

---

**FIGURE 12.15** Top three panels: 64-s samples of solar wind velocity in spacecraft centered $\mathbf{r}$, $\mathbf{t}$, $\mathbf{n}$ coordinates, where $\mathbf{r}$ is the Sun to spacecraft unit vector, $\mathbf{t}$ is the unit vector in the direction of $\Omega \times \mathbf{r}$, and $\mathbf{n}$ completes a right-handed system. Here $\Omega$ is the spin axis vector. Bottom panel: 64-s averages of the n-component of the magnetic field. The 42-day interval shown is the same as in Figure 12.2. Fluctuations in the other two field components are similar to those in $\mathbf{B}_n$. 

---
representative power spectrum of fluctuations in the heliospheric magnetic field obtained near Earth’s orbit. The spectral slope varies in a fashion characteristic of a turbulent cascade of energy via eddies from the lowest frequencies (the energy containing scales) through the so-called inertial range, down to the highest frequencies where the turbulent energy is dissipated. In addition to its probable, but poorly understood, role in heating and accelerating the solar wind, turbulence strongly affects energetic particle transport in the heliosphere and is an essential element of most current models of particle acceleration at shocks in the heliosphere.

14. CONCLUSION

The solar wind is a magnificent natural laboratory for studying and obtaining understanding of processes and phenomena that also occur in a variety of other astrophysical contexts. These include kinetic and fluid aspects of plasmas, plasma heating and acceleration, collisionless shock physics, particle acceleration and transport, magnetic reconnection, and turbulence and waves. Proof of the existence of the solar wind was one of the first great triumphs of the space age, and much has been learned about the physical nature of the wind and related processes in intervening years. Nevertheless, our understanding of the solar wind is far from complete. For example, we still do not know what physical processes heat and accelerate the solar wind or what determines its flow speed. We do not yet know if the low-speed wind arises primarily from quasi-stationary processes or from a series of transient reconnection events in the solar atmosphere. Likewise, the physical origins of CMEs are still being debated. We do not yet fully understand how a rough balance is obtained between new magnetic flux injected into the solar wind by CMEs and magnetic flux removed from the heliosphere by reconnection inside the Alfvén point. Nor do we understand how the magnetic topologies of ICMEs evolve with time. In general, our ideas about the structure of the heliospheric magnetic field are still developing and need testing with further observations. The generally weak heliospheric magnetic fields and reduced solar wind dynamic pressures that characterized the recent extended solar activity minimum have provided a new, and as yet not fully understood, perspective on the long-term variation of the solar wind. Ideas about the solar wind’s interaction with the interstellar medium, the physical nature of the heliopause, and the role of the termination shock in accelerating anomalous cosmic rays are presently being tested with both in situ and remote observations. The physical origin of variations in elemental abundances in the solar wind is beginning to be understood, as are temporal changes in the charge states of the heavier elements. Origins of double ion beams and suprathermal ion tails in the solar wind are still being debated, as are the mechanisms by which solar wind turbulence evolves and eventually dissipates. Moreover, we do not yet fully understand why different ionic species have different speeds and temperatures in the solar wind. Further analysis of existing data, new types of measurements, numerical...
simulations, and fresh theoretical insights should lead to understanding in these and other areas of solar wind research in the years ahead.

BIBLIOGRAPHY


1. EXPLORATION OF MERCURY

Mercury is one of the five planets known to the ancients (along with Venus, Mars, Jupiter, and Saturn), to whom it was the messenger of the gods. The planet’s small size and proximity to the Sun have historically made ground-based telescopic observations difficult, so until the 1970s very little was known about it. Since then, three phases of exploration have peeled back Mercury’s veil of mystery. The Mariner 10 spacecraft performed three flybys of Mercury on March 29 and September 21, 1974, and March 16, 1975. Mariner 10 imaged about 45% of the surface at an average resolution of about 1 km per pixel, but less than 1% at resolutions between 100 and 500 m per pixel, which are needed to see landforms diagnostic of a variety of surface processes. This coverage and resolution are comparable to telescopic Earth-based coverage and resolution of the Moon before the advent of space flight. Mariner 10 also discovered that Mercury is the only terrestrial planet besides Earth with an internally generated, global magnetic field, and it first detected Mercury’s thin atmosphere, or exosphere. Over the next 30 years, ground-based observations provided all the new measurements of the topographic, radar, and reflectivity characteristics of Mercury’s surface; measured its exosphere; and helped to constrain its surface composition. Finally, the MESSENGER spacecraft performed three flybys of its own on January 14, 2008, October 6, 2008, and September 29, 2009; on March 18, 2011, it entered Mercury orbit. Mercury thus became the last of the five “classical” planets to be studied by an orbiting spacecraft.

New data from MESSENGER have revolutionized knowledge of Mercury. This mission is one of the National Aeronautics and Space Administration’s Discovery series of planetary exploration missions. The name MESSENGER, besides reminding us of the ancients’ understanding of Mercury, is an acronym for MErcury Surface, Space ENvironment, GEochemistry, and Ranging. The spacecraft was launched on August 3, 2004. Over its 7-year cruise it used gravity assists from multiple flybys of terrestrial planets to slow its “fall” toward Mercury, so that the main rocket engine could finish slowing the spacecraft enough for orbit insertion. The flybys included an Earth flyby in August 2005, two flybys of Venus in October 2006 and June 2007, and the three Mercury flybys. MESSENGER’s initial science objectives were (1) to determine the nature of polar deposits including their composition; (2) to determine the
properties of Mercury’s core including its diameter and whether there is an outer fluid core; (3) to determine the geologic history of Mercury; (4) to determine the nature of the magnetic field and how it is generated; (5) to measure the composition of the surface, to help in understanding the planet’s formation and to test competing hypotheses for why it is so dense; and (6) to measure its exosphere and how it interacts with the magnetosphere and surface. As with any first orbital mission to a planet, some of MESSENGER’s initial findings were unexpected and drove new, more specific objectives that have been the mission’s new focus since the primary 1-Earth-year orbital investigation was completed.

Upon orbit insertion, MESSENGER was in a highly elliptical, 12-h orbit (Figure 13.1) — the main engine could slow the spacecraft enough to be captured into Mercury orbit, but it was not possible to carry enough fuel to make the orbit circular. The orbit is near-polar to provide global viewing and to measure the planet’s libration. As MESSENGER transitioned to its extended mission, the orbit apoapsis was lowered in altitude, while the periapsis stayed the same, changing the orbital period to 8 h and enabling more time to be spent at lower altitudes taking high-resolution measurements. There are eight science experiments onboard the spacecraft (Figure 13.1 and Table 13.1): (1) a dual imaging system, (2) a gamma-ray and neutron spectrometer, (3) a magnetometer, (4) a laser altimeter, (5) atmospheric (0.105–0.6 mm) and surface (0.3–1.45 mm) spectrometers, (6) an energetic particle and plasma spectrometer, (7) an X-ray spectrometer, and (8) a radio science experiment that uses the telecommunication system. These instruments address the science objectives discussed previously. More information on the mission and its findings at Mercury are available at http://messeger.jhuapl.edu.

For the next phase of Mercury’s exploration, the European Space Agency and the Japan Aerospace Exploration Agency will jointly conduct a Mercury orbital mission called BepiColombo, currently planned for a 2016 launch.

2. GENERAL PLANETARY CHARACTERISTICS

Mercury has been compared with the Moon because both are heavily cratered, but Mercury has had more extensive volcanism, and has an internal magnetic field and active surface processes that the Moon lacks. Mercury’s diameter is 4880 km, 28% smaller than the next largest planet, Mars, and its mass is 3.301 × 10²³ kg. Because of this large mass in relation to its volume, Mercury has an exceptionally high mean density of 5440 kg/m³, second among planets only to the Earth (5520 kg/m³). Without the effects of gravitational self-compression, however, Mercury’s density would be even greater than that of the Earth’s. Mercury’s brightness (albedo) is greater than that of iron-rich basaltic plains on the Moon, but less than that of low-iron lunar highlands. Like the Moon, Mercury is covered with a regolith consisting of fragmental material formed by impacting meteoroids over billions of years.

Global maps of Mercury’s surface from MESSENGER orbital imaging (Figure 13.2) show that the surface is divided into two major geologic units, regions that share similar measurable properties: mostly lighter colored smooth plains covering about a quarter of the planet, and darker, more heavily cratered intercrater plains (the most extensive terrain type) filling most regions between clusters of large craters. A third unit is ejecta, debris excavated from depth by impact craters and basins and deposited onto the surface. All these units are cut by tectonic features. Long lobate scarps traverse the surface for hundreds of kilometers, and arrays of graben occur in the interiors of some of the larger impact basins.

On Mercury, the prime meridian (0°) was chosen to coincide with the subsolar point during the first perihelion passage after January 1, 1950. In most maps constructed prior to MESSENGER, longitude increases to the west; map products generated from MESSENGER data use the convention of longitude increasing to the east, also widely used for modern spacecraft missions to the Moon and Mars. Craters are mostly named after famous authors, artists, and musicians such as Dickens, Michelangelo, and Beethoven, whereas valleys are named after prominent radio telescopes including Arecibo and Goldstone. Prominent ridges and scarps are named after ships associated with exploration and scientific research such as Discovery and Victoria. Plains are named after the name of planet Mercury in various languages, for example, Odin (Scandinavian) and Tir (Germanic). Borealis Planitia (Northern Plains) and Caloris Planitia (Plains of Heat) are exceptions. The largest well-preserved impact basin, named the Caloris Basin (Basin of Heat) because it nearly coincides with one of the “hot poles” of Mercury, figures conspicuously in the planet’s geology.

No natural satellite of Mercury has been discovered. Any satellite that might exist would likely be smaller than 1.6 km in size, to have escaped detection in surveys prior to MESSENGER.

3. MOTION AND TEMPERATURE

Mercury has the most eccentric (0.205) and inclined (7°) orbit of any major planet, although over periods of a few million years, eccentricity varies from about 0.1 to 0.28 and inclination varies from about 0° to 11°. Its average distance from the Sun is 0.387 AU (5.79 × 10⁸ km), but because of its large eccentricity, at present times, the distance varies from 0.308 AU (4.60 × 10⁸ km) at perihelion to 0.467 AU
FIGURE 13.1  (Left) Drawing of the MESSENGER spacecraft showing the placement of the science instruments listed in Table 13.1. (Right) Illustration of the orbit around Mercury shows the very elliptical orbit used for the primary mission and the lower orbits used during the extended mission. *These and other images in this article are thanks to NASA/Johns Hopkins University Applied Physics Laboratory/Carnegie Institution of Washington unless otherwise noted.*
(6.98 \times 10^7 \text{ km}) at aphelion. As a consequence, although Mercury’s orbital velocity averages 47.6 km/s, it varies from 56.6 km/s at perihelion to 38.7 km/s at aphelion. At perihelion, the Sun’s apparent diameter is over three times larger than its apparent diameter as seen from the Earth.

Mercury’s sidereal rotation period (relative to the stars) is 58.65 Earth days, and its orbital period is 87.97 Earth days. It has a unique 3:2 resonance between its rotational and orbital periods, making exactly three rotations on its axis for every two orbits around the Sun. This resonance was probably acquired as a consequence of the dissipative processes of tidal heating and the relative motion between a solid mantle and a liquid core. As a consequence of this resonance, a solar day (from sunrise to sunset) lasts exactly 2 Mercurian years, or 176 Earth days. The obliquity of Mercury is close to 0°; therefore, it does not experience seasons like those of the Earth and Mars. Some topographic depressions in the polar regions thus never receive direct rays of sunlight and can be permanently colder than \(-163 \text{ °C} (-262 \text{ °F}).\)

Another effect of the 3:2 resonance is that the same hemisphere always faces the Sun at alternate perihelion passages. This happens because the hemisphere facing the Sun at one perihelion will rotate one-and-a-half times by the next perihelion, so that it faces away from the Sun; after another orbit, it rotates another one-and-a-half times so that it directly faces the Sun again. Because the subsolar points at perihelion are the 0 and 180° longitudes, they are called hot poles. The 90° and 270° longitudes are called warm poles because they are the subsolar points at aphelion. Yet another consequence of the 3:2 resonance, combined with the large eccentricity and variable orbital velocity, is that the Sun briefly backtracks in the sky during perihelion passage. At the warm poles after sunrise, the Sun reverses direction, sets, and then rises again.

Although Mercury is the planet closest to the Sun, it is not the hottest planet. The surface of Venus is hotter because of its atmospheric greenhouse effect. However, Mercury experiences the greatest range (day to night) in surface temperatures (650 °C = 1170 °F) of any planet or satellite in the solar system, because of its close proximity to the Sun, its long solar day, and its lack of an insulating atmosphere. Near the equator, surface temperature reaches about 467 °C (873 °F) at perihelion, hot enough to melt zinc. At night just before dawn, surface temperature at the same location plunges to about \(-183 \text{ °C} (-297 \text{ °F}).\)

4. INTERNAL STRUCTURE AND MAGNETIC FIELD

Mercury’s large core is unique in the solar system and imposes severe constraints on the origin of the planet. Mercury’s mean density of 5440 kg/m³ is only slightly less than that of the Earth (5520 kg/m³) and larger than that of Venus (5250 kg/m³). Without Earth’s large internal pressures, however, Earth’s uncompressed density is only 4400 kg/m³ compared to Mercury’s uncompressed density of 5300 kg/m³. This means that Mercury contains a much larger fraction of iron than any other planet or satellite in the solar system. MESSENGER’s measurement of Mercury’s moment of inertia suggests that the core is about 4060 km in diameter, or about 83% of the planet’s diameter, and some 58% of its volume. The silicate mantle and crust together are only about 410 km thick. For comparison, Earth’s iron core is only 54% of its diameter and just 16% of its volume. Assuming an appropriate density for the thin silicate mantle, the density of the core must be less than that of pure iron−nickel metal, requiring that it contain lighter elements such as silicon and sulfur.

The thickness of the crust in the northern hemisphere has been estimated using MESSENGER’s low-altitude measurements of the gravitational field and topography, by subtracting the component of gravity due to topography and modeling the subsurface contours of a low-density crust that would be necessary to explain remaining variations in the gravitational field. A density contrast between the mantle and crust and an average crustal thickness have to be assumed: 200 kg/m³ based on geochemical results discussed below and 50 km based on tectonic models for the depth extent of faulting, respectively. Given these assumptions, the resulting crustal thickness is generally greater near the equator (50−80 km) and less toward the north polar region (20−40 km). The thinnest crust is
thought to be located beneath the northern lowlands, the Caloris basin, and other northern hemisphere impact basins. Crustal thickness in the southern hemisphere cannot be estimated in this way due to the lack of low-altitude MESSENGER measurements.

Aside from Earth, Mercury is the only terrestrial planet with an internally generated magnetic field. The field was first detected by Mariner 10 during its flybys, and later measured in detail by MESSENGER. The magnetic field can be modeled as a dipole field like a bar magnet, with a strength near Mercury’s surface of about 190 nanotesla (nT). The dipole is tilted less than 0.8° from the planet’s rotational axis, and its center is offset northward from the center of the planet by about 480 km (nearly 20% of the planet’s radius). In comparison, Earth’s dipole field is 25,000–65,000 nT at the surface (130–340 times stronger than Mercury’s field at the planet’s surface), tilted 11° from the axis of rotation, and offset only about 8% of the planetary radius.

The maintenance of a terrestrial planet dipole magnetic field is thought to require an electrically conducting, fluid...
outer core surrounding a solid inner core. Earth-based radar measurements of the magnitude of Mercury’s **librations** indicate that the mantle is detached from the core, confirming that the outer core is fluid. If the core were pure Fe and Ni metal, cooling of Mercury’s interior over the planet’s history would have allowed the core to solidify and shut down the magnetic field. The presence of the light elements, such as silicon or sulfur, in the core would lower its melting point, allowing the outer core to remain fluid to the present day.

5. **EXOSPHERE AND MAGNETOSPHERE**

The exosphere and magnetosphere of Mercury represent a highly dynamic, coupled system that is unique among the terrestrial planets.

5.1. **Exosphere**

The exosphere is an atmosphere so thin that its few atoms or molecules are unlikely to collide with one another. In Earth’s atmosphere, the exosphere is the highest part of the atmosphere where the density of gas molecules is very low. At Mercury the exosphere is the only atmosphere, so the planet has what is called a surface-bounded exosphere, whose gas molecules collide with the surface (or escape from the planet) rather than colliding with each other.

A primary method by which the exosphere of Mercury is studied is observation of **resonant emission** from atoms, in which solar photons of specific energies or wavelengths are absorbed and then reemitted at the same wavelength. Because the combinations of energies at which such emissions occur vary between elements, observed emission spectra provide unique spectral fingerprints for elements that are present. Mariner 10’s ultraviolet spectrometer discovered Mercury’s exosphere through observations of emission from both hydrogen (H) and helium (He) atoms. Mariner 10 measurements imply a surface pressure 1 trillion times smaller than that of the Earth’s atmosphere. Nearly a decade after Mariner 10’s flybys, advances in telescopes and instrumentation led to the discovery of sodium (Na) and potassium (K) in the exosphere; calcium (Ca) was detected in 2000. MESSENGER added magnesium (Mg) to the known elements in the exosphere during its second flyby.

In contrast to the denser atmospheres of Earth, Venus, and Mars, the contents of Mercury’s exosphere are transient and must be continually replenished. If the source processes for Mercury’s exosphere suddenly stopped, the exosphere would dissipate in just 2–3 days. Also in contrast to other terrestrial planet atmospheres, Mercury’s exosphere is composed almost entirely of atoms rather than molecules, a result primarily of the manner in which the exosphere is generated and maintained. Any molecules that are present in the exosphere are quickly photodissociated (i.e. broken apart) by sunlight, which is intense at Mercury owing to its proximity to the Sun and the lack of a thick upper atmosphere to absorb the sunlight.

Mercury’s exosphere originates from the planet’s surface, partly from material native to Mercury, partly from material implanted in Mercury’s surface by the stream of charged particles from the Sun known as the **solar wind**, and partly from the impacts of comets and meteoroids. The generation and maintenance of Mercury’s exosphere is summarized in **Figure 13.3**. The first of three major sources of exospheric atoms is sunlight striking the surface, releasing material in either of two ways. **Photon-stimulated desorption**, or PSD, occurs when solar photons hit the surface and release their energy, breaking the bonds that hold the surface materials together and ejecting atoms from the surface. **Thermal desorption**, or evaporation, occurs when sunlight heats the surface and loosely bound, volatile material is boiled off. Both processes are low-energy processes, so trajectories of the ejected atoms do not carry them very high or very far.

The second major source of exospheric atoms is a process known as **sputtering**, which occurs when ions from the solar wind or Mercury’s magnetosphere impact the surface. The energy in these impacts is higher than in the case of PSD or thermal desorption, so that atoms ejected by sputtering have larger velocities and their trajectories carry them higher and farther than atoms released through low-energy thermal processes. Ion sputtering can fracture the surface on atomic scales, liberating volatile species such as Na. This leads to greater release of material through PSD than in the absence of ion sputtering, by a process known as ion-enhanced PSD.

Meteoroid impacts are the third primary source of exospheric material. Although large impacts release much material, they are rare and an influx of small dust particles from interplanetary space that collide with Mercury’s surface is more responsible for day-to-day maintenance of the exosphere. The energy of these collisions vaporizes both the dust particles and some of the surface, releasing high-energy atoms to high altitudes.

Atoms released with low velocities follow ballistic trajectories under the influence of gravity. Because they do not go very high, these atoms mostly fall back to the surface where they either bounce or stick. Some atoms undergo multiple bounces before sticking (known as ballistic hops), and in this manner redistribute volatile material across Mercury’s surface, gradually transferring it from the hotter, equatorial regions to the colder, polar regions.

Atoms released with high velocities also follow ballistic trajectories; however, the longer time of residence of these atoms in the exosphere allows two other processes to affect them. The first is solar radiation pressure, in which solar photons push the atoms in the antisunward direction. If the solar radiation pressure pushes them far enough, atoms will
not return to the surface but will be pushed “behind” the planet to become part of a neutral, cometlike tail. Atoms pushed into the tail will escape the planet unless influenced by the second process redirecting exospheric atoms, ionization from solar photons striking the atoms (photoionization), and removing electrons. The positively charged atoms will be picked up by Mercury’s magnetic field and rapidly accelerated either toward or away from the planet, depending on the orientation of the local magnetic field lines. Those atoms accelerated toward the planet impact the surface and can drive sputtering; those accelerated away from the planet will be lost to interplanetary space. Thus, atoms in the exosphere of Mercury ultimately either return to the surface or are lost to space, explaining why the exosphere would dissipate so quickly if not resupplied.

The processes that release atoms to Mercury’s exosphere and that subsequently affect them differ in magnitude depending on the element. Ca and Mg are refractory elements (having strong chemical bonds requiring higher energies to release them from the surface), whereas Na is a volatile element (with weak bonds broken at lower energies). At the same time, Ca atoms have a lifetime against photoionization a factor of 10 smaller than for Na atoms and 100 smaller than for Mg atoms. Thus, neutral Ca atoms do not survive long in Mercury’s exosphere (typically 1 h), whereas Mg atoms last a much longer time (typically 2–3 days). In addition, the effects of radiation pressure vary with each element, Na being affected strongly, Ca more weakly, and Mg hardly at all. Radiation pressure is also proportional to solar flux, and the Sun’s spectrum contains deep absorptions, known Fraunhofer lines, at the wavelengths of most resonance emissions. In the parts of Mercury’s elliptical orbit where the planet is accelerating toward or away from the Sun, the Doppler shift between the Sun and Mercury shifts the exospheric resonance emissions away from the Fraunhofer lines, providing increased radiation pressure. This Doppler shifting creates “seasons” in the exosphere, as Mercury orbits the Sun. MESSENGER saw these seasonal variations clearly during its flybys of Mercury when it viewed the exosphere from a distance (Figure 13.4).

These varying effects on different elements lead to different distributions in Mercury’s exosphere. Sodium is seen everywhere in Mercury’s exosphere, a consequence of its volatile nature and relatively easy release from the surface. Because solar radiation pressure has a large effect on Na, it is also the primary constituent in Mercury’s tail and has been observed as far from Mercury as 2 million miles. At times there is so much Na in the tail that an observer on the nightside of Mercury might see a yellow-orange tinge to the night sky: the intensity of the Na emission, at the same wavelength as Na vapor street lamps, is similar in strength to a moderate aurora on Earth. The Na distribution is mostly symmetric about the Sun–Mercury line, indicating that it has a large PSD source; however, there are often local enhancements due to other processes,

FIGURE 13.3 Schematic illustration showing the source and loss processes responsible for the generation and maintenance of Mercury’s exosphere.
and the altitude distribution of Na exhibits a distinct two-component profile, consistent with its release from both low-energy and high-energy processes.

Calcium, on the other hand, has a very different distribution. MESSENGER observations reveal altitude profiles with only a high-energy component; they also show a persistent, strongly asymmetric distribution about the Sun—Mercury line, with peak densities near the equator at dawn. There are several possibilities for the difference between Ca and Na density with time of day. There may be more meteoroid impacts on the dawn side, which is the leading side of Mercury as it plows through the dust in the inner solar system. Alternatively, there may be differences in sputtering and photodissociation with time of day that affect the two elements differently.

Prior to MESSENGER, Mg had been predicted to be a part of Mercury’s exosphere, but was not discovered until MESSENGER’s second flyby. Mg has a distribution that contrasts with both Na and Ca. The overall Mg distribution is mostly isotropic about Mercury and characteristic of a high-energy release process; however, there is some evidence for localized enhancements. It remains a puzzle why Mg and Ca, both refractory species, are distributed so differently in the exosphere.

Other elements have also been observed in Mercury’s exosphere, including hydrogen, helium, potassium, and possibly oxygen and aluminum. They are more difficult to observe because their emissions are weak and/or they are not particularly abundant. What limited information we do have on these elements shows that there could be even more puzzling aspects to Mercury’s exosphere, whose understanding requires further observations.

5.2. Magnetosphere

The interaction of Mercury’s magnetic field with the solar wind creates a magnetosphere qualitatively similar in several ways to Earth’s magnetosphere, although roughly 20 times smaller (Figure 13.5). The field is strong enough to stand off the solar wind on the dayside of the planet under normal conditions, with the magnetopause—the boundary between the magnetosphere and the solar wind—generally located about 1000 km above the surface near the subsolar point. In the polar regions, funnel-shaped cusps form, exposing portions of Mercury’s surface to direct bombardment by the solar wind. Because the magnetic field is so asymmetric about the planet’s geographic equator, the southern polar region exposed to the solar wind is roughly four times larger than the corresponding region in the north (Figure 13.6). The solar wind sweeps a portion of the magnetic field downstream to form the magnetotail of Mercury. Reconnection events—releases of energy that occur when Mercury’s magnetic field and the interplanetary magnetic field (IMF) coalign and “splice” together—happen regularly. The magnetosphere is also populated with plasma, or hot, ionized gas.

The overall structures of Mercury’s and Earth’s magnetospheres are similar, whereas the time and spatial scales of their dynamic phenomena are very different. Reconnection at Mercury’s dayside magnetopause, which occurs...
when there is an IMF component antiparallel to the local planetary magnetic field, occurs at rates \( \approx 10 \) times the typical rate observed at Earth. Such rapid reconnection can significantly shrink the magnetosphere on the dayside and push the magnetopause toward the surface. Under extreme solar wind conditions, the two magnetospheric cusps in the north and south can migrate equatorward, merge, and expose large regions of the dayside directly to the solar wind. Such dayside reconnection events increase the overall energy levels of the magnetosphere, but whereas the fractional increase at Earth for similar events is \( 10^{\pm 30\%} \), the increase at Mercury is typically \( 200^{\pm 300\%} \)!

Reconnection also occurs in the magnetotail, where the magnetic field exhibits variations on time scales of seconds to minutes. Such fast magnetic field reconfiguration leads to vigorous heating of plasma trapped in the magnetic field, acceleration of ions to high energy levels, and other phenomena. One important phenomenon is “plasmoids”, quasi-loop-like magnetic “islands” that form in the tail during reconnections. Some plasmoids move away from the “X-line” where they form, in a sunward direction (i.e. toward the nightside of the planet), and deposit energetic particles on the planet’s surface; others move antisunward, down the tail away from the planet, and remove material from the magnetosphere. The circulation of plasma and energy from the X-line at the dayside magnetopause to the X-line in the magnetotail constitutes the “Dungey cycle” that powers Earth-type magnetospheres. At Earth, the cycle time is on the order of 1 h; at Mercury it is on the order of 1–2 min. Mercury’s magnetosphere is extremely dynamic, and in that respect has no equal in the solar system.

MESSENGER has studied Mercury’s magnetosphere not only by measuring the magnetic field but also by measuring ions and electrons in the magnetosphere, collectively called magnetospheric plasma. The plasma has two sources: the solar wind and the planet’s surface. Solar wind plasma enters the magnetosphere through the cusps or by “leaking” through the magnetopause, mostly in concert with reconnection events. Plasma originating from the planet’s surface not only derives from photoionization of neutral atoms in the exosphere but also includes ions produced directly from sputtering or micrometeoroid impacts.

The distribution of ions in Mercury’s magnetosphere is controlled by motion of charged particles within the magnetic field. Ions gyrate, or rotate in a helical fashion, around local magnetic field lines and drift in directions driven by the laws of electromagnetics. Motion of the ions within the magnetosphere is called magnetospheric convection. For heavier ions (i.e. those with atomic masses larger than \( \text{He}^+ \), such as \( \text{Na}^+ \)), the \text{gyro radius} around the field lines is large.
relative to the size of the magnetosphere. Many heavier ions thus collide with the planet during magnetospheric convection or cross the magnetopause and are picked up by the solar wind and swept away. The net result is that anisotropies are created in ion distributions—that is, differences in ion density as a function of latitude and local time of day. MESSENGER first mapped out these anisotropies and discovered three persistent features: (1) a large ion population at high northern latitudes on the dayside near the magnetospheric cusp, (2) an ion population near the equator on the nightside, and (3) an increase in ion abundance near the magnetopause, spanning the magnetopause boundary.

Ions observed by MESSENGER fall into five categories based on atomic mass and charge, that are distinguishable by the spacecraft’s plasma spectrometer: H\(^+\) (protons), He\(^+\) (ionized helium atoms), He\(^{2+}\) (alpha particles), O\(^+\)-group ions (ionized atoms or molecules having atomic masses close to oxygen), and Na\(^+\)-group ions (ionized atoms or molecules having atomic masses close to Na, including Mg and Al). Protons are the most abundant ions (average density, 10/cm\(^3\)), followed by He\(^{2+}\) ions (3.9 \times 10^{-2}/cm\(^3\)), Na\(^+\)-group ions (5.1 \times 10^{-3}/cm\(^3\)), O\(^+\)-group ions (8.0 \times 10^{-4}/cm\(^3\)), and finally He\(^+\) (3.4 \times 10^{-4}/cm\(^3\)). When ions are divided into these categories, anisotropies in plasma distribution are highlighted.

A look at Na\(^+\)-group ions illustrates the types of anisotropies discovered by MESSENGER. Two enhancements in ion density occur at high northern latitudes (Figure 13.7), with one feature (“feature 1”) centered at a local time of \(\sim 10.5\) h, corresponding to the northern magnetic cusp, and a second (“feature 2”) centered at a local time of \(\sim 19\) h. Enhancements are also evident at equatorial latitudes near the dawn terminator (“feature 3”, local time 6 h) and at premidnight local times (“feature 4”, centered around \(\sim 20\) h), at altitudes above \(\sim 2000\) km. This latter feature continues to high southern latitudes at an altitude of \(\sim 6000\) km (“feature 5”). Features 2, 4, and 5 appear to be part of a single larger structure that represents an asymmetry between the dawn and dusk hemispheres. O\(^+\)-group ions show the same major features as do the Na\(^+\)-group, but at lower densities. In contrast, He\(^+\) ions do not show these anisotropies, and are more uniformly distributed about the planet.

In contrast to the populations of positive ions, electrons exhibit no persistent distributions as a function of location or local time. Rather, their striking characteristic is localized energetic events. First observed by Mariner 10, and later studied in detail by MESSENGER, these events are recurring, intense bursts of high-energy electrons that increase in intensity by orders of magnitude above background in times as short as a few seconds. Individual events typically last for a few seconds to several minutes and come in groups spread over as many as several hours. Two regions of electron events are regularly observed: one at high northern latitudes on the nightside and another, less energetic one near the equator at most local times. An explanation for these electron bursts remains elusive.

6. GEOLOGIC FEATURES

Mercury’s surface is dominated by plains having various densities of superimposed craters that surround and fill impact basins. Younger smooth plains infill some impact basins and also cover vast expanses outside recognized basins. Older, more cratered intercrater plains occupy the space between well-preserved impact basins. The largest well-preserved impact basin, Caloris, is infilled with distinctly colored smooth plains (Figure 13.2). The north polar region is the largest topographic low and contains the

FIGURE 13.7 Average densities (cm\(^{-3}\)) for the Na\(^+\)-group ions. Left panel: average density projected onto the noon–midnight plane, binned 100 km \times 100 km. The Sun is to the right, and the red circle shows the approximate size of the planet in the projection. The shape of the sampled volume is driven by MESSENGER’s orbit around Mercury. Right panels: average density as a function of altitude (km) and local time (h) for three different latitude ranges. Mercury’s position in its orbit (“true anomaly angle”, or TAA) and its heliocentric distance (R, in astronomical units (AU)) at the times of the measurements are shown under the middle panel only, but apply to all three panels. The numbers indicate magnetospheric features discussed in the text. Figure courtesy of Jim Raines, University of Michigan.
largest expanse of relatively young smooth plains. Mercury’s surface is also traversed by compressional thrust faults that form lobate scarps. Locally within impact basins, radial and concentric patterns of tectonic ridges and graben occur.

Thermal infrared measurements from Mariner 10 showed that the surface is a good insulator and, therefore, consists of fine-grained, fragmental regolith, which is formed by fragmentation of surface rocks by eons of impact cratering. The highest resolution images from MESSENGER show small-scale landforms produced where the regolith has gradually moved down topographic slopes by a process called mass wasting, in which loose material migrates from topographically higher to lower regions (Figure 13.8). In the absence of a thick atmosphere and flowing water, mass wasting is a major process for gradually eroding Mercury’s surface, imparting a smoothed appearance to the surface at small scales.

6.1. Impact Craters and Basins

Mercury’s cratered surface records the late heavy bombardment by crater-forming projectiles. Heavily cratered regions on the Moon and Mars also record this bombardment, but volcanism and tectonics have removed its record from the surfaces of Venus and Earth. Based on radioactive dating of Apollo samples from the surface of the Moon and asteroid dynamical studies, the late heavy bombardment appears to have started about 4.0 billion years ago, peaked 3.9 billion years ago, and then declined rapidly for about 100-200 million years. After this time the impact rate by objects left over from the late heavy bombardment was much less, ending about 2 billion years ago. The same population of impacting bodies is thought to have affected the whole inner solar system, allowing dates from Apollo samples to be used to estimate ages of heavily cratered surfaces on Mercury and Mars.

Effects of young impacts are illustrated by the large, fresh crater Hokusai (Figure 13.9, left). Hokusai has an extensive system of crater rays extending nearly halfway around Mercury. Rays form where clouds of material
broken up during ejection from the parent crater create strings of reimpacting ejecta that form small secondary craters. The secondary craters in turn excavate brighter regolith from below a very thin surface layer that has been darkened and reddened by a process called space weathering. The tiny secondary craters in the rays will lose obvious association with Hokusai over millions of years as space weathering causes the rays to fade. On the Moon, space weathering is caused by iron in silicate minerals being reduced to metallic blebs only nanometers in size that coat the regolith grains. Micrometeoroid impacts and hydrogen from implanted solar wind drive the process. How the process works on Mercury is a subject of ongoing study.

Fresh impact craters on Mercury exhibit morphologies similar to those on the other terrestrial planets. Small craters are bowl shaped, but with increasing size, craters develop central peaks, flat floors, and terraces on their inner walls. The transition diameter from simple (bowl-shaped) craters to complex craters (with central peak and terraces) occurs at about 10 km. Beginning at a diameter of ~150 km and continuing through diameters up to ~350 km, Mercurian craters have an interior concentric ring, or peak ring, instead of a central peak. At even larger diameters, they have multiple, concentric rings, forming multiringed basins. For a given crater diameter, the radial extent of continuous ejecta outside the crater is uniformly smaller than on the Moon by a factor of about 0.65. The maximum density of secondary impact craters also occurs closer to the crater rim than for similarly sized lunar craters, at about 1.5 crater radii from the rim compared with 2–2.5 crater radii on the Moon. Both differences are due largely to the greater surface gravity of Mercury (3.70 m/s²) than that of the Moon (1.62 m/s²).

Mercury has 46 recognized impact basins 300 km or more in size. Per unit surface area, this population yields a lower density of impact basins than on the Moon, with a greater discrepancy between the two bodies at larger diameters: the Moon has an average of one 500-km basin per 2,700,000 km², but Mercury only has 1 per 4,300,000 km². The disparate population densities highlight two major differences between the superficially similar-looking cratered surfaces. First, it is generally thought that Mercury’s lower basin density results from the oldest basins being buried by intercrater plains, which are widespread on Mercury but not the Moon. Second, Mercury has three times as many peak-ring basins per unit area as the Moon, with a total of 110 recognized peak-ring basins, like Renoir (Figure 13.9, right), compared with only 17 on the Moon. Although the reason for this difference is debated, a leading hypothesis is that the higher mean velocity of Sun-orbiting impactors at Mercury versus at the Moon leads to greater melting at the center of impact basins, resulting in a larger central melt cavity. There is widespread evidence for impact melting being prevalent on Mercury in smaller craters as well. For example, at the young rayed crater Waters (15 km in diameter), a large and distinctly colored “tongue” of impact melt splashed out of the crater (Figure 13.10).

6.2. Volcanic Plains and Vents

Based on Mariner 10 images, there was strong evidence that some plains deposits originated as lavas, but uncertainty as to whether others may have formed as impact melt or impact basin ejecta. MESSENGER’s higher resolution images, multicolor imaging, and global coverage revealed that relatively young, smooth plains cover about 27% of Mercury’s surface. The smooth plains exhibit two characteristics of having formed from volcanic lavas. First, most large impact basins are at least partially infilled by smooth plains, and those smooth plains commonly partly fill impact craters superimposed on the basins, for example, in the interior of Renoir (Figure 13.9). This relation requires that a geologically long time transpired between the formation of the host basin and the plains, to allow time for the formation of the infilled craters, thus ruling out an impact melt origin. Second, the plains typically have a distinct color that indicates a compositional difference from the underlying basin. The large basins Caloris, Tolstoj, Rembrandt, and Rachmaninoff all have ejecta of dark-colored low-reflectance material (LRM) but are filled with smooth high-reflectance plains (HRP) (Figure 13.2).

In some places, the colors of materials excavated by craters suggest that different types of plains materials form
layers in Mercury’s upper crust, at least locally. For example, the crater Calvino (Figure 13.11) is formed in intermediate-colored plains that infilled an unnamed, highly degraded impact basin. That basin formed in a surface layer that contained LRM, which constitutes the basin’s ejecta. Small craters within the degraded basin expose more of the infilling intermediate plains. Calvino is the largest crater within the degraded basin and exposes HRP in its rim material; only material forming the central peak of Calvino excavated deeply enough to expose LRM. This suggests a three-layer sequence: LRM that the degraded basin excavated, overlain by HRP that Calvino excavated, overlain by intermediate material that infills the degraded basin.

Smooth plains have morphologic features suggesting that lavas forming them had very low viscosity and were comparable in age to, or older than, dark lava plains on the Moon. They mostly lack obvious flow fronts that sometimes mark the edges of Earth’s or the Moon’s lava flows. Instead, their emplacement appears to have eroded channels into preexisting rock, forming streamlined islands (Figure 13.12, left). In addition, their burial of craters in the northern plains suggests that at least those plains formed as massive floods up to 1 km or more in thickness, covering many hundreds of thousands of square kilometers. These characteristics indicate extremely fluid lavas, which is consistent with expectations based on the elemental compositions of the lavas (discussed below). Densities of superimposed craters suggest typical ages of 3.7—3.9 billion years, comparable to the oldest lava deposits that constitute the lunar maria. This means that Mercury’s youngest large volcanic deposits, forming the smooth plains, occurred only during the early phases of emplacement of the Moon’s widespread maria.

Mercury has at least 49 irregular, scallop-rimmed, steep-walled depressions surrounded by haloes of bright material having color properties like those of HRP. These depressions are interpreted to be pyroclastic vents, formed...
where ascending volatile-rich magma degassed and erupted explosively before falling back to mantle the surrounding surface (Figure 13.13). Similar features occur on the Earth and Mars, where water is the dominant volatile, and more rarely on the Moon, where the dominant volatile may instead be carbon monoxide. The driving volatile on Mercury is uncertain, but geochemical evidence discussed below suggests that sulfur or sulfur compounds may be important. The vents occur predominantly along basin rings, thrust faults, and at crater central peaks, suggesting that impact and tectonic fractures are important conduits for the eruptions.

Mercury’s youngest volcanic features are not uniformly distributed. Their mapping from MESSENGER images (Figure 13.14) shows that smooth plains are concentrated in the northern hemisphere, in the topographically low northern plains and within and around Caloris. In contrast, pyroclastic vents occur mostly outside the smooth plains.

Older, intercrater plains (Figure 13.12, right) are the most extensive terrain on Mercury. They differ from smooth plains mainly in having a greater density of superimposed small craters. They are probably responsible for burying a substantial number of impact basins, yet the ejecta of other large basins are also superimposed on the intercrater plains. These relations indicate that intercrater plains were emplaced over a range of ages contemporaneous with the late heavy bombardment. Craters in the intercrater plains also excavate layers of differently colored material from depth. Some intercrater plains infill extremely degraded large impact basins that are barely recognizable from peaks in their rings that poke through the plains. For these reasons, many intercrater plains are probably older, more cratered versions of the volcanic smooth plains.

FIGURE 13.13 The 36-km-long pyroclastic vent northeast of Rachmaninoff basin, surrounded by a diffuse, bright halo of material even higher in albedo and redder than HRP. The red, green, and blue image planes show images taken at 1.00, 0.75, and 0.43 μm. The vent is located at 35.8° N latitude, 63.7° E longitude.

FIGURE 13.14 Simple cylindrical map of Mercury’s smooth plains and pyroclastic vents. Crater materials that bury parts of the smooth plains are also shown. “Odin-type plains” are a special geologic unit ringing Caloris that may or may not be volcanic in origin. Courtesy of Brett Denevi, Applied Physics Laboratory.
6.3. Tectonics and Topography

No other planet or satellite in the solar system has tectonics so dominated by features formed by compression as does Mercury. On the Earth, for example, although compression occurs in folded mountain belts, extension occurs in rift zones and at the midocean ridges so that our planet’s surface area is conserved. Mercury has globally distributed compressional thrust faults called lobate scarps (Figure 13.15, left). Individual scarps vary in length from \(~20\) to \(>600\) km and have heights from a few hundred meters to about 3 km. They are nearly globally distributed, but tend to occur in broad belts separated by large regions having fewer scarps. In the equatorial region, the scarps have a predominantly north–south orientation. Lobate scarps are conspicuously absent from the vast northern expanse of smooth plains. There, the more common landform is smaller \textbf{wrinkle ridges} (Figure 13.15, right). This landform is also found in thick lava flows on Venus, the Earth, the Moon, and Mars and is thought to form by cooling and contraction of the lava. The dominance of compressional features on Mercury is widely thought to result from cooling and \textbf{global contraction} of Mercury throughout most of the planet’s history. The total length of the lobate scarps is \(~42,000\) km, and reduction in Mercury’s surface area associated with their formation is estimated to be \(~0.08–0.12\)%, corresponding to a decrease in planetary radius of \(~1–1.5\) km. The dominant orientation of lobate scarps has been speculated to be a consequence of the reuse of ancient fractures formed by tidal stresses during slowing of an earlier more rapid rotation rate, called \textbf{tidal despinning}. Mercury’s earlier, faster spin would have formed an equatorial bulge that collapsed as rotation slowed; at low latitudes, north–south compressional faults would be expected.

Mercury’s topography is also unlike that of the Moon and Mars. All three bodies have large impact basins, but on the Moon and Mars, large basins are also vast depressions that are obvious and unambiguous in topographic maps, despite being partially filled by lavas (Moon and Mars) or sediments (Mars). Only a few of the youngest large impact basins, including Rachmaninoff and Raditladi (Figure 13.16), are clearly recognizable in a topographic map of Mercury’s northern hemisphere (Figure 13.16, constructed using data from MESSENGER’s laser altimeter). Caloris Basin, the largest well-preserved basin on Mercury, does not even appear as a low area in the topographic map—parts of its floor stand above terrain outside the basin! This unexpected topography is thought to result from a combination of infilling by volcanic plains and subsequent contraction of the planet that warped Mercury’s
rigid, outer layer—the **lithosphere**—into broad, gentle dips and swells. Mercury also has comparatively low topographic relief; the range of elevations in the northern hemisphere measured by MESSENGER’s laser altimeter is 9.85 km, considerably less than the elevation range on the Moon (19.9 km) or Mars (30 km).

Interiors of impact basins are the locations where extensional tectonic features are common on Mercury. In smaller basins, such as Mozart (Figure 13.17, left) there are a few graben, but the larger basins Rembrandt and Caloris (Figure 13.17, right) contain complicated patterns of graben and ridges. Figure 13.18 shows maps of tectonic structures that illustrate the change in tectonic style with basin size. In Figure 13.18(a), an unnamed northern plains crater about 120 km in diameter is completely buried by plains materials, but its rim is traced by wrinkle ridges. In the crater’s interior, randomly oriented graben break the surface up into polygonal blocks a few kilometers in size. The larger, 235-km diameter Mozart basin (Figure 13.18(b)) has a few randomly oriented ridges at the center of smooth plains within its inner ring, surrounded by roughly concentric graben. Rachmaninoff (290 km) and Raditladi (257 km) (Figure 13.2) have a similar size and tectonic pattern. The even larger 720-km-diameter Rembrandt basin (Figure 13.18(c)), the second-largest well-preserved basin, has radial graben and ridges within its inner ring, concentric graben near the location of the basin ring, and randomly oriented ridges outside that. The largest well-preserved basin, 1550-km-diameter Caloris (Figure 13.18(d)), has breathtakingly complex tectonics. In the central part of the basin, a radial system of graben called Pantheon Fossae occurs together with roughly concentric ridges; they are surrounded by a ring of concentric graben. Outside this, as in Rembrandt, randomly oriented wrinkle ridges occur. The processes responsible for this transition in tectonic patterns with basin size are uncertain, but two plausible ones are cooling and sagging of the smooth plains fill and flow of the soft upper mantle toward the center of the basin early in Mercury’s history as early basin topography reached **isostasy**.

Impact basin formation is also thought to have formed tectonic features directly. Opposite the Caloris basin on the other side of Mercury (around the **antipode** of Caloris) is hilly and lineated terrain, also called weird terrain, that disrupts preexisting landforms including crater rims (Figure 13.19). The hills are 4–10 km wide and typically 0.1–1.8 km high. Linear depressions that are probably graben form a roughly orthogonal pattern. Geologic relationships suggest that the age of this terrain is comparable to that of Caloris. Similar terrains occur at the antipodes of the Imbrium and Orientale impact basins on the Moon. The hilly and lineated terrain is thought to be the result of shock waves generated by the Caloris impact and focused at the antipodal region. Computer simulations of shock wave propagation indicate that focused shock waves from an impact of this size can cause vertical ground motions of about 1 km or more and fracturing of crustal rock to depths of tens of kilometers below the antipode. The antipodal regions of the Rembrandt and Tolstoj basins lack hilly and

![FIGURE 13.17](image1.png) **FIGURE 13.17** The interior smooth plains of Mozart basin (left) are deformed by a relatively few, roughly concentric extensional graben, seen in this image covering an area approximately 150 km across. In contrast, the inner parts of plains infilling the much larger Caloris basin (right) are densely fractured by radial graben and roughly concentric ridges, seen in this image of an area approximately 280 km across.
lineated terrain, suggesting that formation of these smaller basins did not affect their antipodal regions as strongly as Caloris’ formation did.

6.4. Surface Composition

During exploration of most planetary bodies, the first information on surface composition typically comes from the manner in which the surface reflects the Sun’s visible and near-infrared light, measured using a technique called reflectance spectroscopy. Transition metal cations, especially ferrous iron in the common silicates olivine and pyroxene and traces of it in feldspar, cause absorptions—preferential absorption of light at specific energies—at wavelengths that are diagnostic of different minerals. Space weathering tends to mask these absorptions, but on the Moon, even a very low iron content of a few percent in feldspar-rich highland rocks creates an easily detected absorption. Mercury’s spectrum, in contrast, is utterly smooth at wavelengths where iron absorptions occur on other planetary bodies, even in crater rays where space weathering has had the least effect. The absence of iron absorptions shows that Mercury’s silicates probably contain no more than several tenths of a percent iron. A complementary technique called thermal emission spectroscopy uses reduced heat emission at specific wavelengths as the fingerprint of different minerals, with the wavelengths governed by vibrations of mineral lattices. This technique does not depend on the presence of transition metal cations. Thermal emission spectra of Mercury taken from the Earth suggest abundant pyroxene, but with a low iron abundance. The very low iron content in silicates from spectroscopic measurements was the first key evidence that Mercury’s surface is chemically highly reduced, with iron in chemical forms other than silicates.
Color imaging of Mercury, first by Mariner 10 and then by MESSENGER, does show that there are compositional variations among surface materials (Figure 13.20). However, those variations manifest themselves as differences in albedo and red slope—the steepness of the spectrum of reflected sunlight—leaving differences in mineralogy obscure. Many smooth plains consist of higher albedo, more red-sloped HRP material, whereas older plains contain interbedded HRP and intermediate-colored plains material. Low-albedo, less red LRM is concentrated in ejecta of impact basins, particularly Rembrandt and Tolstoj, suggesting that LRM is more prevalent at depth. The nature of the mineral phase(s) creating color variations is uncertain: Mercury’s low-iron silicates should be relatively bright, even after space weathering, whereas most of Mercury’s surface is low in albedo compared to low-iron lunar highlands.

The first quantitative information on surface composition comes from MESSENGER’s orbital measurements of elemental composition using X-ray and gamma-ray spectroscopy. These techniques measure interactions of high-energy radiation with different chemical elements and are not confounded by the reduced state of the surface. One surprising result, for a hot planet close to the Sun, is that the surface is not depleted and is even relatively rich in volatile elements. For example, on average, volatile potassium is more abundant relative to nonvolatile thorium than on the surface of Venus, Earth, and the volatile-depleted Moon and is comparable in abundance to that on volatile-rich Mars (Figure 13.21). Total contents of potassium and sodium are estimated to be up to 0.2 wt-% and 1–5 wt-% depending on location. Mercury’s surface also contains up to ~4 wt-% sulfur. Total iron content of ~1% is more than can be contained in silicates, with most probably occurring as sulfide or free metal; total iron content is very low compared to ~14% on Mars.

Spatial variations in major element abundances support the idea that Mercury’s plains were formed by different lava compositions. Figure 13.22 shows measured abundances of the major, nonvolatile elements magnesium, silicon, and aluminum measured from orbit in the upper centimeters of the surface. HRP plots close to a basaltic composition, although it is low in iron compared to terrestrial basalts. Less red units dominating the older, intercrater plains are higher in magnesium and lower in aluminum, suggesting
more pyroxene and/or olivine and less feldspar compared to basalt, more consistent with a type of volcanic rock called komatiite, which is rich in olivine and forms on Earth by melting of a larger fraction of its source region deep in the mantle than does basalt. These differences suggest that Mercury’s volcanism evolved over time, from komatiitic or a related composition to more basaltic. Mercury also exhibits spatial variations in abundance of the relatively volatile elements sodium and potassium in the upper centimeters of the surface. Both elements are lower in abundance closer to the equator, suggesting gradual evaporation of them over billions of years.

7. RECENT SURFACE FEATURES

Like the Earth and Mars, Mercury has ongoing surface processes related to migration of volatile materials, beyond the “dry”, gravitationally driven processes of mass wasting.

7.1. Radar-bright Polar Deposits

High-resolution radar images of Mercury from both the Arecibo and Goldstone radar facilities discovered patches with high radar reflectivity clustered around the poles. The reflectivity characteristics of the deposits are similar to those of outer-planet icy satellites and the residual polar water ice caps of Mars, so Mercury’s polar radar-bright patches have long been thought to be water ice, which in many places may be buried by a few centimeters of regolith cover. Water ice is stable in Mercury’s polar region inside some topographic depressions, especially craters, that remain in permanent shadow due to Mercury’s low obliquity; these areas are illuminated only indirectly by sunlight reflecting off crater walls. How much reflected sunlight reaches a crater floor depends on the shape of the crater interior, which varies with crater size. Larger craters have shallower depth-to-diameter ratios, so they have a lesser solid angle of illuminated crater wall radiating onto their floors. Thermal models suggest that large craters (>40 km in size) poleward of 82° latitude have shadowed regions so cold (colder than −163 °C, or −262 °F) that water ice can remain stable at the surface over the age of the solar system. In contrast, 10-km craters can preserve water ice at the surface only poleward of 88°, assuming typical fresh crater shapes. Ice can persist further equatorward under a thin cover of only a few centimeters of regolith, to insulate the ice from the warmest daytime temperatures.

MESSENGER images of Mercury’s polar regions show that craters with permanently shadowed portions enclose most of the radar-bright material. Radar-bright material around the south pole is largely confined within the 170-km crater Chao Meng-Fu (Figure 13.23). In the north polar region, the deposits reside within multiple craters (Figure 13.24). Close to both poles, nearly all permanently shadowed regions contain radar-bright material. However, radar-bright materials also occur in permanently shadowed regions equatorward of 80° latitude; those more equatorward deposits show a strong preference for longitudes near 90° E and 270° E, Mercury’s “warm poles”, where local noon occurs at aphelion, instead of at the “hot poles” at 0° E and 180° E, where local noon occurs at perihelion.

Differences in characteristics of the north polar radar-bright material have been revealed by MESSENGER orbital neutron spectrometer and laser altimeter measurements; presumably similar variations occur in south polar material but spacecraft altitude over high southern latitudes...
is too high for either instrument to resolve those deposits. The neutron spectrometer, which is highly sensitive to hydrogen, reveals concentrations of hydrogen centered on the pole. The strength of the neutron signal suggests that the deposits are nearly pure water ice. The laser altimeter measures reflectivity of the surface at 1064 nm, as well as spacecraft distance from the surface. Radar-bright materials in craters very close to the north pole have a high albedo, suggesting that water ice is present at the surface. Radar-bright materials farther from the pole are darker than surrounding terrain. This difference in reflectivity may indicate a darker, possibly organic component covering or mixed with the ice, at latitudes where a thin cover is required to preserve the ice. The total mass of ice preserved at the poles is estimated to be 20 to 2000 trillion kg—to within a factor of 10—comparable to the amount of water in Lake Tahoe.

The radar-bright deposits of frozen volatiles may have originated from impacting comets or water-rich asteroids, which released water and organic compounds that became cold-trapped in permanently shadowed craters. Comets and asteroids also impact the Moon. The neutron and gamma-ray spectrometers on the Lunar Prospector spacecraft discovered enhanced hydrogen signals at very high lunar latitudes that correspond with locations of permanently shadowed craters. These high-hydrogen regions have been interpreted as water ice with a concentration of only 1.5 ± 0.8% weight fraction, a much lower ice fraction than may exist in high-albedo, probably nearly pure ice very close to Mercury’s north pole. The Lunar Crater Observation and Sensing Satellite impacted into a permanently shadowed region in the crater Cabeus, creating a plume of cold-trapped volatiles and dust ejected into sunlight that was observed from the Earth and from the impactor’s shepherding spacecraft. This experiment confirmed the presence of water ice and suggested a variety of other volatiles. In the lunar south polar region, the crater Shackleton has a nearly permanently shadowed interior; measurements by the Lunar Reconnaissance Orbiter’s laser altimeter of the crater interior did not detect the same magnitude of brightening within it as within Mercury’s polar craters. Thus, Mercury’s polar regions resemble those of the Moon in being enriched with volatiles including water ice, but unlike at the Moon, there are large deposits of nearly pure ice and large exposures of it at the surface.

7.2. Hollows

An unexpected landform discovered in the highest resolution MESSENGER orbital images is “hollows”, flat-floored depressions hundreds of meters wide and tens of meters deep, which commonly coalesce into interconnected groups kilometers to tens of kilometers across. Many hollows are surrounded by bright, less red material. This surrounding bright material had been detected in Mariner 10 and MESSENGER flyby images taken at lower resolution and had been called “bright crater floor material”. Hollows occur where LRM has been exposed from depth, usually within craters and impact basins such as in the peak-ring and floor of Raditladi (Figure 13.25). They occur even in geologically young craters, and fresh-appearing hollows have few if any superimposed craters. These relations suggest that hollows may be actively forming even today. Hollows occur preferentially on equator- or hot-pole-facing slopes, and where impact melt or HRP-forming lava has contacted LRM, consistent with formation due to the loss of volatile material by heating; the surrounding bright material may be a residue of devolatilization. In the shallow...
subsurface under centimeters of regolith, temperature is buffered to a daily average value of 150 °C (303 °F) or less; on sunward facing slopes the surface can reach 467 °C (872 °F), and lavas and impact melts would be hundreds of degrees hotter. Although the volatile composition is uncertain, sulfur-containing phases are the leading candidates because of sulfur’s abundance and the temperature regime in which the hollows form.

8. HISTORY

8.1. Geologic History

The earliest well-preserved surface features on Mercury are large basins and intercrater plains that formed concurrently during the late heavy bombardment ≥3.9 billion years ago. Older basins exist, but are recognizable only from basin-ring massifs that protrude through the intercrater plains, from remnants of ejecta with sculptured textures radial to their basins, and from regions of thinned crust. Near the end of late heavy bombardment, the youngest large impact basin, Caloris, was formed. From 3.7 to 3.9 billion years ago further eruption of lava occurred within and surrounding Caloris and other smaller basins and in the northern lowlands to form smooth plains. The global system of thrust faults formed after the intercrater plains, but how long after is unclear; there are no recognized examples of smooth plains partially burying lobate scarps that would constrain the beginning of the scarps’ formation. However, the scarps do cross-cut relatively fresh craters; from estimates of the age of such scarps, lobate scarps probably were still forming ~1.5 billion years ago. Rayed craters are estimated to be about 100—200 million years old, and there are no examples of lobate scarps disrupting crater rays, suggesting that lobate scarpe formation has either stopped or greatly decreased. Smaller peak-ring basins continued to form relatively late into Mercury’s history, including Raditladi, which may have formed as recently as 1 billion years ago. Present-day surface processes include impact cratering, mass wasting, cold trapping of volatiles in permanently shadowed craters, space weathering, and formation of hollows.

8.2. Thermal History

Thermal history models of planetary interiors depend on compositional assumptions and are only as good as those assumptions. Mercury thermal models were generated during the time between Mariner 10 and MESSENGER and examined (among other factors) how the amount of sulfur in the core would affect the amount of global contraction that occurred as Mercury cooled from an initially molten state and the inner core solidified. The models have yet to be updated with new interpretations of other light elements occurring in the core, principally Si. With this important caveat in mind, the models predicted a total amount of planetary radius decrease due to cooling between 6 and 10 km: ~6 km due to cooling of the mantle, and up to 4 km due to cooling of the core, with greater amounts corresponding to lower contents of sulfur in the core. The 6 km of contraction due to mantle cooling is expected to have occurred mostly before the late heavy bombardment and should not be evident in Mercury’s present geology; contraction due to core cooling is thought to have occurred subsequently, and thus could be preserved. To the extent that the models remain valid, the ~1 km radius decrease inferred from lobate scarps may be consistent with a core sulfur abundance of ~5%.

Thermal models also suggest that Mercury has been contracting throughout its history, creating compression in the lithosphere that would tend to close fractures and make it difficult for magmas to ascend to the surface. However, large impacts would be expected to strongly fracture the lithosphere, providing egress for lavas to reach the surface.

8.3. Origin

A major question about the origin of Mercury is how it acquired such a large fraction of core-forming metal compared to the other terrestrial planets. Prior to MESSENGER’s orbital measurements, four recent hypotheses had been put forward to explain Mercury’s enrichment in iron. One (selective accretion) invoked an enrichment of iron due to mechanical and dynamical processes in the innermost part of the solar nebula during Mercury’s accretion. Two more (postaccretion vaporization and giant impact) invoked removal of a large fraction of the silicate mantle from a once larger proto-Mercury. A final one (accretion from carbon-rich dust) invoked carbon-bearing dust desiccated of water by high inner nebular temperatures modifying the chemistry of accreting Mercury. In the selective accretion model, the different responses of iron and silicates to impact fragmentation, and aerodynamic sorting of fragments in the nebula, led to iron enrichment owing to the higher gas density and shorter dynamical timescales in the innermost part of the solar nebula. In this model, the removal process for silicates from Mercury’s present position is more effective than for iron, leading to iron enrichment. The postaccretion vaporization hypothesis proposed that intense bombardment by solar electromagnetic and particle radiation in the earliest phases of the Sun’s evolution vaporized and drove off much of the silicate fraction of Mercury leaving the core intact. In the giant impact hypothesis, a planet-sized object impacted Mercury and blasted away much of the planet’s silicate mantle leaving the core largely intact. In the carbon-rich dust hypothesis, as the solar nebula cooled, abundant carbon bonded with silicon and kept it in a gaseous phase.
until iron had condensed and accreted to form Mercury’s core. As a result, once silicates condensed and formed the mantle, they were nearly iron free; some silicon and oxygen were lost to more distant parts of the nebula, enriching Mercury in iron. The very reducing environment caused sulfur to be concentrated in early forming calcium and magnesium sulfide, enriching Mercury in sulfur compared to Earth.

These models predict somewhat different chemical compositions for the silicate part of Mercury, due to the different processes experienced by the early planet. For the selective accretion model, there is no reason for Mercury’s silicate portion to be highly depleted in either alkali oxides (Na and K) or FeO. In contrast, postaccretion vaporization should lead to depletion of more volatile elements like sulfur and alkali oxides. For the giant impact model, the prediction depends on the composition of crust stripped away late in accretion: if there were a primordial crust analogous to the feldspar-rich lunar highlands, then alkali oxides may be depleted. However, there is no reason to expect a depletion of FeO. The carbon-rich dust hypothesis predicts extremely low iron in silicates, and unique among the four hypotheses, specifically predicts a high content of sulfur, around 4%.

MESSENGER’s findings provide tests for these hypotheses. Mercury’s high contents of alkali oxides and sulfur are inconsistent with postaccretion vaporization. The high alkali oxides are also inconsistent with impact stripping of a lunarlike primordial crust, but are consistent with the selective accretion model. However, this hypothesis offers no explanation for the high sulfur content or low iron in silicate. The carbon-rich dust hypothesis explains both the low iron content of silicates and the high content of sulfur, and at present, seems best able to explain Mercury’s composition. The origins of Mercury’s unique and surprising composition will be better understood as analysis of MESSENGER results continues and, eventually, new results are obtained by BepiColumbo.

**BIBLIOGRAPHY**


Venus possesses a dense, hot atmosphere, composed primarily of carbon dioxide. A surface pressure of nearly 100 bars sustains the mean surface temperature of 740 K, which is essentially globally uniform except for topographic effects. The surface is totally hidden at visible wavelengths by multiple cloud decks extending from about 48 km altitude to about 65 km above the surface, above which the particle concentration falls off gradually with a scale height of about 3 km. The clouds are approximately bounded by the evaporation temperature of H$_2$SO$_4$ below and the top of the convectively mixed troposphere above. Their composition is primarily liquid droplets of concentrated sulfuric acid, with an additional ultraviolet (UV) absorber in the upper layers and large, possibly solid, particles near the base level, both of unknown composition. The middle atmosphere (stratosphere and mesosphere) extends from 65 to about 95 km and the upper atmosphere (thermosphere and exosphere) from 95 km up. Although the rotation period of the solid planet is 243 Earth days (sidereal), tracking of pronounced markings visible in the clouds at UV wavelengths (Figure 14.1) shows that the atmosphere in the cloud region rotates in about 4 days in the same retrograde direction. Evidence for complex and very active dynamics and meteorology is seen globally in variations in these markings and in the temperature, composition, and cloud density distributions in the upper troposphere up to the thermosphere. Giant polar vortices with complex, variable
morphologies (most commonly dominated by wavenumber 2) are found at high latitudes (>60°) in both hemispheres.

1. INTRODUCTION AND OBSERVATIONS

1.1. Earth-Based Observations

The study of Venus through telescopes was frustrating for centuries due to the complete cloud cover, until Earth-based radars began to penetrate the cloud to map the surface in the 1960s. The presence of CO₂ in the atmosphere had been established in 1932, as soon as spectrometers equipped with infrared (IR)-sensitive photographic plates could be applied to the problem. Determining the abundance proved difficult because the radiation observed is scattered among the cloud particles and also because it was assumed that nitrogen would be abundant, as it is on Earth. This gas cannot be detected in the spectral range available from the ground, so its proportion remained speculative for many years.

Following the detection of CO₂, relatively little was achieved until the mid-1960s, when improved techniques including the development of Fourier spectroscopy led to the discovery of small proportions of water vapor and carbon monoxide in Venus’ atmosphere in and above the clouds. Traces of hydrogen chloride and hydrogen fluoride were also found, and a tight upper limit was set on the amount of O₂. Similar studies of the atmosphere below the clouds became possible from 1983 onward, following the discovery of several narrow spectral “windows” in the near-IR region. These are parts of the spectrum that fall between the strong absorption bands of CO₂ and H₂O, so that the radiation from deep layers can be detected from above. Their existence had not been predicted theoretically because it was assumed that absorption in the clouds, and the far wings of strong lines, would obscure the windows. In fact, the line wings are weaker than simple theory predicts, and the sulfuric acid clouds are conservatively scattering with little absorption at short IR wavelengths (below ~2.3 μm).

In the shortest wavelength windows, as at microwave radio wavelengths, radiation from the planet's surface can escape to space. At longer IR wavelengths (greater than ~5 μm), the emission from the night side is characteristic of the temperature of the cloud tops, about 240 K. In the windows, the brightness, and therefore the temperature of the emitting region, is characteristic of the lower atmosphere and surface and therefore considerably higher. Images taken in a window reveal horizontally banded structures formed of silhouettes of the lowest part of the cloud (around 50 km) against the hotter atmosphere below (Figure 14.2).

The two most prominent near-IR windows are at 1.74 and 2.3 μm (Figure 14.3), and others are at 1.10, 1.18, 1.27, and 1.31 μm. From spectra of the absorption lines and bands in these windows, like the example in Figure 14.3, inferences about the composition at various levels all the way to the surface are possible. Some of the composition data in Table 14.1 were obtained by this technique, and the rest mainly by mass spectrometer measurements from entry probes. Earlier, millimeter-wave spectroscopy had provided estimates of the water and carbon monoxide abundances in the deep atmosphere, initially using radiation from the whole disk. In the early 1990s, modest spatial resolution became available for ground-based radio work by use of interferometric techniques in which the signals from several antennas are combined. The breakthrough in the near IR produced improved spatial resolution on the subcloud atmosphere, although only at equatorial- and midlatitudes since the polar regions are virtually inaccessible due to the small obliquity of Venus’ spin axis.
Cloud patterns detected in blue and near-UV images were used from the 1930s onward to establish the presence of the 4-day rotation at the cloud tops. Measurements of the Doppler shifting of spectral lines confirmed that this is indeed due to bulk motions, corresponding to winds of the order of 100 m/s at pressure levels similar to those near the Earth’s surface. Later, IR imaging disclosed the 6-day period of a deeper region. The same patterns, seen in more detail from spacecraft, have revealed wave motions and other meteorological activity in the cloudy layers on a variety of scales.

The composition of the clouds was another important question that was answered first from analysis of ground-based observations of the polarization of light reflected from the planet. Although such measurements were first made in the 1930s, the computers and programs to carry out the analysis did not exist until the mid-1970s. The results pinned down the refractive index and showed that the upper cloud particles are spherical; these two properties eventually led to the identification of supercooled droplets of concentrated sulfuric acid (H₂SO₄). The nature of the compositional and microstructure variations in the clouds that gives rise to the UV and near-IR patterns remains obscure, however.

Radio astronomers, observing Venus’s emission at the microwave wavelength of 3.15 cm, discovered in 1958 that it appears to be much hotter than expected, and this was confirmed by later results at other wavelengths. The radiation at this wavelength is expected to originate at the surface, which should be warmed, as Earth is, by the greenhouse effect, but the warming required to explain the new microwave data was so extreme compared to expectations that other hypotheses, such as auroral emissions, were debated. Spacecraft measurements, including radiometry from close flybys and then direct measurements from the early landers, finally settled the issue in favor of the greenhouse effect and showed that the pressure at the mean surface is 93 bars.

1.2. Space Missions

A large number of experiments on 25 spacecraft have been devoted to studies of the atmosphere. United States
missions, starting in 1962, were the flybys Mariner 2, 5, 10 (which went on to Mercury); Pioneer Venus Multiprobe and Orbiter in 1978; the radar mapper Magellan; and Galileo en route to Jupiter, followed by the Saturn-bound Cassini. The Soviet Union had success with Venera 4–14, which included entry and descent probes as well as flybys or orbiters; Venera 15 and 16, which were radar mappers; and Vega 1 and 2, which dropped both probes and balloon-borne payloads during Venus encounters en route to Halley’s comet. Early missions were devoted to reconnaissance, in particular to confirmation of the high surface pressure and temperature inferred from the microwave radio measurements, and basic composition measurements. The Pioneer Venus probes confirmed the cloud composition and gave unprecedented detail on the densities, sizes, and layering of the particles. The European Space Agency sent its first mission to Venus in 2006, a polar orbiter based on the design of its successful Mars Express, which had reached the red planet 3 years earlier. The payload for Venus Express focused on high-resolution imagery and spectroscopy of the atmosphere and surface, measuring winds, detecting lightning, and mapping the escape of water from the planet’s exosphere and magnetosphere.

Many of the same techniques used from the Earth, principally spectroscopy, radiometry, and imaging, have been applied from flyby and orbiting spacecraft. An important addition is the radio occultation experiment, which tracks the effect of the atmosphere on the telemetry carrier as the spacecraft disappears behind the atmosphere or reappears from behind it. On Venus, the neutral atmosphere can be observed in this way from about 34 to 90 km and the ionosphere from 100 to 400 km. At greater depths, the refraction of the waves by the atmosphere is so great that the beam strikes the surface and never reappears. In addition to using radio occultation and carrying several instruments for remote sensing, Pioneer Venus Orbiter actually physically penetrated the upper atmosphere once per orbit down to as low as 135 km above the surface, carrying a suite of instruments to make measurements in situ. Two mass spectrometers measured individual gases and positive ions; a Langmuir probe and a retarding potential analyzer measured electron and ion densities, temperatures, and velocities; and a fifth instrument measured plasma waves. Higher energy ions and electrons, both near the planet and in the solar wind, were measured by a plasma analyzer, and important auxiliary information was provided by a magnetometer. In addition, the atmospheric drag on the spacecraft gave an excellent measure of the density as a function of height. This drag experiment was repeated by Venus Express in 2014, and stellar and solar occultation techniques, which have allowed spectroscopic observations up to 160 km altitude, added to new radio occultation measurements.

A number of probes have descended part or all the way through the atmosphere, and the Vega balloons carried out measurements in the middle of the cloud region. All of them have carried an atmospheric structure package measuring pressure, temperature, and acceleration as a function of height. The height was determined on the early Venera probes by radar and on all probes by integration of the hydrostatic equation. Gas analyzers have increased in sophistication from the simple chemical cells on Venera 4 to mass spectrometers and gas chromatographs on later Soviet and US missions. In some cases, however, there are suspicions that the composition was significantly altered in passage through the sampling inlets, especially below 40 km, where the temperature is high.

A variety of instruments measured the clouds and their optical properties, most recently the visible-infrared thermal imaging spectrometer (VIRTIS) on the Venus Express orbiter. Previously, radiometers on entry probes had observed the attenuation of solar energy during their descent through the atmosphere, and others measured the thermal IR fluxes as well. Winds were obtained by tracking the horizontal drifts of the probes as they descended and the balloons as they floated, following direct measurement of wind, pressure, and temperature at the surface by the Venera landers. Venera 11–14 carried radio receivers to seek evidence of lightning activity and Venus Express detected “whistler”-mode emissions attributed to lightning bursts.

Venera, Pioneer Venus, and Venus Express orbiters obtained a great deal of information on the dynamics of the atmosphere with UV and near-IR imaging and thermal IR mapping. Pioneer Venus discovered the north polar vortex with its double “eye”, the polar warming trend in the middle atmosphere, and the principal wave modes and tides in the atmosphere at cloud-top level. Spectacular images and movies were obtained of the corresponding vortex over the South Pole by Venus Express, revealing its detailed structure and its dynamical variations. Still operating at the time of writing, the European mission built on the earlier explorations to paint a picture of an Earthlike planet that apparently lost its oceans of water and gained a dense, volcano-fueled atmosphere as a result.

2. ATMOSPHERIC TEMPERATURES

A representative mean temperature profile is illustrated in Figure 14.4, with an indication of the names given to the different regions and the levels occupied by the cloud layers.

2.1. Surface

The air temperature at the surface of 740 K and the apparent absence of any significant diurnal or latitudinal variation, are consistent with energy balance calculations
for Venus’ distance from the Sun and an overlying atmosphere with the observed density and composition. The high surface temperature is produced by the small percentage of solar energy that reaches the surface, trapped by the opacity of the overlying atmosphere in an extreme version of the phenomenon we know on Earth as the greenhouse effect. Radiation from the lower atmosphere in the thermal IR is ineffective for cooling to space because of the opacity of the atmospheric gases at long IR wavelengths. The species principally responsible are CO$_2$, SO$_2$, H$_2$O, and the solids and liquids in the clouds; H$_2$SO$_4$, in particular, is strongly absorbing at long IR wavelengths, in contrast to its behavior in the near IR where scattering dominates over absorption, allowing around 2% of the solar energy incident on the planet to reach the surface.

### 2.2. Lower Atmosphere

The lower atmosphere or troposphere extends from the surface up to the level of the visible cloud tops at about 65 km. The latter marks the tropopause, defined as the upper boundary of the region where vertical heat transport by convection dominates over radiative exchange with space. The lower atmospheric temperature profile has been measured in detail by numerous descent probes, with most results in close agreement, and above 35 km also by radio occultation. The gradient of temperature versus height is basically consistent with simple thermodynamic theory, being close to the dry adiabatic lapse rate of $\sim 10$ K/km from the surface to a few kilometers below the tropopause, where it tends toward the constant value with height that characterizes the overlying stratosphere (Figure 14.4). Small deviations from this lapse rate exist and have significance for the atmospheric dynamics, as discussed below.

Venus and Earth actually have rather similar vertical atmospheric temperature profiles if the comparison is restricted to the range of pressures common to both (Figure 14.5). The main difference, where they overlap, is the stratospheric heating on Earth due to the ozone layer.

![FIGURE 14.4](image-url) Temperature and pressure profiles, with their day–night variations, from the surface to 180 km altitude. The approximate locations of the main layers of cloud are also shown.

![FIGURE 14.5](image-url) Temperature profiles for Venus and Earth, on a common pressure scale (From Taylor and Grinspoon (2009)). The main difference, where they overlap, is the stratospheric heating on Earth due to the ozone layer.
solar and IR net fluxes can be closely reproduced without considering them.

The uniformity with both latitude and longitude of the surface temperature, for a given elevation, is consistent with the high time constant of around 100 years for the dense near-surface atmosphere and very slow (<1 m/s) near-surface winds, again as observed. In this regard, the deep Venusian atmosphere is somewhat analogous to the oceans on Earth. The vertical relief of the surface is about 15 km and, taking into account the vertical temperature gradient, is responsible for surface temperature differences of around 150 K across the planet. Again, this has an effect on the general circulation.

2.3. Middle Atmosphere

The middle atmosphere (stratosphere and mesosphere) extends from the tropopause near 65 km to the temperature minimum or mesopause at about 95 km (see Figure 14.4). Here, as expected for a region in radiative balance between the hot tropopause below and cold space above, the temperature is approximately constant with height. The distinction between the stratosphere and mesosphere is not as useful as it is on Earth, where the middle atmosphere temperature profile is strongly affected by the ozone layer, and most authors use “middle atmosphere” or mesosphere (which of course means the same thing) for the combined 30-km-deep layer.

Unlike the lower atmosphere, there is considerable temperature structure and variability in the middle atmosphere (Figure 14.6). The latitudinal gradient above the clouds is such that the polar region is warmer by 15–20 K than the equator, that is, in the opposite sense to what would be expected from radiative balance. This is in contrast to the upper troposphere, where a large equator-to-pole temperature contrast of ~30 K in the other sense is found at the 1000 hPa level. Model studies indicate that the polar warming is a feature of the rapid global superrotation of the atmosphere, being associated with the pressure gradients required to balance the flow.

In the circumequatorial (zonal) direction, the most remarkable feature of the temperature field is a prominent solar-fixed wavenumber-2 structure with an eastward, i.e. upwind, tilt with increasing altitude at a mean rate of 6 K/km. At the cloud tops, temperature maxima are found just before local noon and midnight, and these have moved eastward by more than 180° in longitude by the 100-km level. This clearly is an upward-propagating solar tide, which is to be expected, although the dominance of wavenumber-2 is less obvious as a response to predominantly wavenumber-1 forcing. In fact, wave 1 does dominate inside the clouds where most of the solar energy is deposited, and the cloud tops themselves rise and fall by about 2 km during the solar day, but tidal models show that this component propagates vertically much less efficiently than wave-2.

2.4. Upper Atmosphere

The upper atmosphere (thermosphere and exosphere) lies above the mesopause. Here, temperatures can no longer be measured directly, but are inferred from the

**FIGURE 14.6** Time-averaged temperature fields in the middle atmosphere of Venus (Schofield & Taylor, 1982). (a) The zonal mean field and (b) the variations around a latitude belt from 0 to 30° N, both plotted against pressure and approximate height. The horizontal stepped line represents the retrieved mean cloud-top height.
scale heights of various gases with use of the hydrostatic equation. On Earth, this region is called the thermosphere because temperatures as high as 1000 K are reached in the outermost layer, or exosphere. The exospheric temperature is much more modest on Venus, no more than 350 K on the day side and not far above 100 K on the night side, a result of strong cooling by the IR bands of carbon dioxide (Figure 14.4). The large temperature difference translates into a pressure difference that drives strong winds from the day to the night side at all levels above 100 km.

On Earth, the exospheric temperature changes markedly with solar activity, being around 700 K at sunspot minimum and 1400 K at maximum. The corresponding change at Venus is much more modest, perhaps 50 K. These differences are again traceable to the fact that CO₂, the principal radiator of heat, is just a trace constituent of Earth’s atmosphere but is the major constituent for Venus (and also Mars). Venus’s slow rotation also contributes to the very cold temperatures on the night side, although the atmosphere does rotate substantially faster than the solid planet.

3. COMPOSITION

A summary of the bulk composition of Venus’ atmosphere from all measurements is given in Table 14.1 and model vertical profiles are given in Figure 14.7.

3.1. Carbon Dioxide and Nitrogen

Confirmation that carbon dioxide is indeed the major gas came from a simple chemical analyzer on the Venera 4 entry probe. The mole fraction was found to be about 97%, in reasonable agreement with the currently accepted value. The next most abundant gas is nitrogen; although it is only 3.5% of the total, the absolute quantity is about three times that in the Earth’s atmosphere. Many of the differences between Venus and Earth can be traced to a relative scarcity of water in the atmosphere of Venus and the total absence of liquid water on the surface. On Earth, carbon dioxide and sulfuric, hydrochloric, and hydrofluoric acids are all carried down by precipitation, a process that is absent in the hot, dry lower atmosphere of Venus. They all then react and are incorporated in geological deposits; estimates of the total amount of carbonate rocks in the Earth give a quantity of CO₂ almost equal to that seen in the atmosphere of Venus.

3.2. Oxygen and Ozone

Free oxygen is undetectable at the Venus cloud tops where one molecule in 10 million could have been seen. At higher levels, the oxygen in carbon dioxide is liberated by dissociation by sunlight and is readily detected (along with CO) by spacecraft instruments orbiting through the upper atmosphere and in airglow measurements. The free oxygen is removed before it can diffuse down to the cloud level by the action of a strong mechanism in the middle atmosphere that converts O₂ and CO back to CO₂. The nature of this mechanism is revealed by observations of around 1 ppm of hydrogen chloride in the region. The HCl molecules are also subject to dissociation by solar radiation, yielding a chlorine concentration that is nearly 1000 times greater than that on Earth. This participates in a catalytic cycle that can completely eliminate free oxygen from the middle atmosphere and prevent the formation of ozone in significant quantities.
On Venus the chlorine-based chemistry is closely coupled to the sulfur cycle that maintains the clouds, as further discussed below.

3.3. Water Vapor

Measurements of water abundance above, within, and below the clouds show a great deal of variance, some attributable to difficulties with the measurement technique but some probably real. Overall, it is clear that Venus has between ten and one hundred thousand times less water in its atmosphere than exists in the oceans and atmosphere of the Earth.

The fact that water is depleted while, at the same time, deuterium is more than 100 times more abundant on Venus than Earth (see next section) suggests that Venus had much more water initially but that most of it has been lost by dissociation to form hydrogen and oxygen which then escape from the planet. Models suggest that Venus could have lost an ocean of present-day terrestrial proportions in only a few hundred million years in this way. The ratio of heavy to light hydrogen (D/H) on Venus would be much larger if all of the deuterium in a primordial Venusian ocean had been retained; however, deuterium as well as hydrogen can escape from the atmosphere, only more slowly due to its greater mass which affects the mixing ratio of vapor in the upper atmosphere falls and leads to fractionation of the two isotopes.

Studies of the water vapor abundance in the middle atmosphere of Venus, both within and above the main cloud deck, have yielded very diverse results. Earth-based observations, made from high-altitude sites and with high-resolution airborne spectrometers, resulted in mixing ratios between 0 and 40 ppm and evidence of higher values in localized and temporary wet spots. The infrared radiometer experiment on board the Pioneer Venus Orbiter discovered the presence of a wet area (~100 ppm) in the afternoon equatorial region, presumed to be produced by the heating maximum at the subsolar point, with very small amounts (a few ppm) of water vapor at other times of the day and at higher latitudes.

Such large variations do not seem to occur in the lower atmosphere. Although in situ measurements by the Venera and Pioneer Venus entry probes and the two Vega balloons reported mixing ratios ranging from 30 to 5000 ppm, direct sampling measurements of water are notoriously difficult and the higher value probably can be discounted. Since the discovery of the near-IR windows, high-resolution Earth-based spectroscopic observations of the night side have found a near-constant global abundance of 40 ± 20 ppm over a range of subcloud altitudes. The near-IR mapping spectrometer operating during the Galileo flyby of Venus confirmed this with a deep atmosphere water vapor mixing ratio of 30 ± 15 ppm, subsequently endorsed by Venus Express.

3.4. Deuterium to Hydrogen (D/H) Ratio

The D/H ratio was first measured on ions in the ionosphere then confirmed in the middle and lower atmosphere by mass spectrometer data and by analysis of spectra taken from Earth in the near-IR windows. For the direct sampling instruments, deuterium provided a valuable signature for distinguishing Venus water vapor in the mass spectrometer from any contaminants carried along from Earth. Values for the D/H isotopic ratio from the accumulated measurements averaged around 0.019, 120 times larger than Earth.

The very high spectral resolution obtained by the Spectroscopy for Investigation of Characteristics of the Atmosphere of Venus (SPICAV)/Solar Occultation at Infrared (SOIR) spectrometer on Venus Express allows the measurement of simultaneous vertical profiles of H2O and HDO above the clouds and D/H fractionation. The averaged HDO/H2O ratio equals a factor of 240 ± 25 times the ratio in Earth’s ocean, or around two times the bulk atmospheric value measured in the lower atmosphere. This could be due to preferential destruction of H2 relative to HD, preferential escape of H relative to D, or possibly selective condensation, a process that has recently been found to be important for fractionating D and H on Mars and Earth.

3.5. Sulfur Dioxide

The high sulfur content of the atmosphere, including the H2SO4 clouds, is a powerful indicator of recent volcanic activity, since gases like sulfur dioxide have a short lifetime in the atmosphere before they are removed by interaction with the surface. The measured abundance of SO2 in the deep atmosphere is about 180 ppm, which is more than 100 times too high to be at equilibrium with the surface. The time constant for the decline of the sulfur abundance in the atmosphere if the source were removed is a few million years, indicating that the atmospheric sulfur must be of recent origin. Pioneer Venus UV spectra showed a decline by more than a factor of 10 in sulfur dioxide abundance at the cloud tops over a 5-year period, and more recently, Venus Express has also detected very large, long- and short-term variations in SO2 at all altitudes from the clouds to the thermosphere.

The high level of SO2 in the atmosphere is the source for the concentrated sulfuric acid that is the dominant component of the clouds (see Section 4.4 below). Although less well understood, it is probably the nonuniform distribution of SO2 and the formation of trace amounts of elemental sulfur and possibly other sulfur compounds that gives rise to the UV markings in the clouds that the visible face of Venus. Apart from forming the highly reflective clouds that tend to cool the planet, sulfur dioxide is a
greenhouse gas contributing to the warming of the surface (Section 6).

### 3.6. Carbon Monoxide

The principal source of carbon monoxide in the atmospheres of both Venus and Mars is the dissociation of CO$_2$ by solar UV radiation, which occurs primarily at high levels on Venus since the energetic photons required do not penetrate far into the atmosphere. During its encounter with Venus in February 1990, the near-infrared mapping spectrometer NIMS on the Galileo spacecraft used observations of the CO 2-0 vibration-rotation band in the 2.3-μm spectral region to obtain the concentration of CO in the deep atmosphere, around 30 km above the surface. These data indicated relatively high concentrations of CO at high northern latitudes, a mean increase of $\sim 50\%$ from 23 ± 2 ppm to 32 ± 2 ppm, behavior which was shown to occur in a symmetric manner in the other hemisphere by observations from Venus Express in 2006. The findings raise the question of what is the source of the lower atmospheric CO and why its proportion in the well-mixed atmosphere increases systematically from equator to pole in both hemispheres. Vertical diffusion from the lower atmosphere increases downward transport from the upper atmosphere at high latitudes, as discussed below.

### 3.7. Noble Gases and Isotopes

Most of the advanced studies and proposals for future missions to Venus emphasize the potential value of obtaining more accurate and comprehensive measurements of the isotopic ratios of the chemical elements within the gases in the atmosphere. Some current values are listed in Table 14.2. These can give important clues about the early history of the Solar System, the formation of the planets, and the evolution of their climates. Particularly interesting cases, in addition to the deuterium to hydrogen ratio ($^2$H/$^1$H or D/H) discussed above, are the isotopic ratios of helium ($^3$He/$^4$He), nitrogen ($^{15}$N/$^{14}$N), carbon ($^{13}$C/$^{12}$C), and oxygen ($^{16}$O/$^{18}$O). The noble gases argon ($^{36}$Ar/$^{38}$Ar/$^{40}$Ar), krypton, neon, and xenon are all both heavy (hence less likely to escape) and inert (hence not prone to removal by chemical reactions with the crust) and so the ratios of their abundances on different planets should provide particularly good tests of evolutionary models. In this way we would hope, for example, to distinguish between initial capture from the solar nebula or later from the solar wind, radioactive decay in the interior, followed by release into the atmosphere, and the contribution from the collision of volatile-rich bodies such as comets with the planet.

Interpretations rely on quite complex models, the integrity of which is hampered by the inadequacy of present data. For example, xenon, the heaviest of the relatively abundant noble gases, has not been measured dependably on Venus at all. Still, some important inferences are possible. Primordial argon and neon are more than an order of magnitude less abundant on Earth than on Venus and less common on Mars than Earth by a similar large factor. This observation is interpreted to mean that the present atmospheres accreted along with the planetary bodies and were later outgassed, with the differences in argon abundance explained in terms of differences in the density of the solar nebula at the time of formation. At the same time, the ratios between the noble gas abundances on Venus are more similar to the ratios found in the Sun than those of Earth and Mars. This suggests less modification and fractionation than the highly processed atmospheres of Earth and Mars.

The isotopic ratios of carbon and oxygen are approximately the same on Venus, Mars, and Earth, but $^{15}$N/$^{14}$N in the Martian atmosphere is about 1.7 times that on Earth. If the difference on Mars is a result of the preferential loss of the lighter isotope, models require that nitrogen was initially of the order of 100 times more abundant on Mars than it is now. Venus has roughly the same (actually about three times as much) total amount of atmospheric nitrogen as Earth and a similar $^{15}$N/$^{14}$N ratio; this tends to suggest that both planets may have retained most of their original inventory.

### 3.8. Ionosphere

The principal heat source for the thermosphere is the production of ions and electrons by far-UV solar radiation. The most abundant positive ions are O$_2^+$, O$^+$, and CO$_2^+$. As part of these processes, CO$_2$ is dissociated into CO and O and N$_2$ into N atoms. All of these ions, molecules, and atoms
have been observed or directly inferred (Figure 14.8). Some of the O\(^+\) ions (with an equal number of electrons) flow around to the night side and help to maintain a weak ionosphere there. Venus lacks any detectable magnetic field, and the dayside ionosphere is therefore impacted by the solar wind, a tenuous medium of ions (mostly H\(^+\)) and electrons flowing from the Sun at about 400 km/s. Electrical currents are induced in the ionosphere, and they divert the solar wind flow around the planet. The boundary between the two media, called the ionopause, is typically at an altitude of a few hundred kilometers near the subsolar point, flaring out to perhaps 1000 km above the terminators and forming a long, tail-like cavity behind the planet (See The Solar Wind).

4. CLOUDS AND HAZES

4.1. Appearance and Motions

The view of the planet from Earth or from an orbiting spacecraft shows Venus completely and permanently shrouded in cloud. The surface apparently formed by the cloud is not a discrete upper boundary, but extends upward as a gradually thinning haze to least 80 km. The “cloud top” is usually defined as the level at which the optical depth reaches unity at some specified wavelength, and for visible light occurs near 65 km altitude. At this level, the range of visibility (the horizontal distance within which objects are visible) is still several kilometers.

Studies of sequences of images like those in Figure 14.1 reveal that the cloud-top region is rotating with a period of about 4 days, corresponding to an equatorial east–west wind speed of about 100 m/s. This varies with latitude, usually reaching around 140 m/s in a localized high-speed jet near 60\(^\circ\) in each hemisphere, although secular variations have also been observed and sometimes the jet is absent, leaving near solid-body rotation speeds. Near-IR images like Figure 14.3 show a longer period consistent with the idea that the dark features are silhouettes of the lower cloud, where entry probes have measured wind speeds of 70–80 m/s.

4.2. Vertical Layering

Several entry probes have made measurements of cloud scattering as they descended, and spectroscopic and other optical measurements add information about what is obviously a very complex and variable regime. Three regions (upper, middle, and lower) can be distinguished in the main cloud, and there is also a thin haze extending down to 30 km. Descent probe and spectroscopic and polarimetric data suggest that the particle size distributions form four distinct “modes”, which have different mean particle radii and different distributions with height, and possibly different constituents. Figure 14.9 shows a simplified interpretative model that shows a mean vertical structure and inferred composition for the clouds.

The upper cloud, the one that can be most readily studied from the Earth or from orbit, is in a region of high convective stability and intense solar UV irradiation, indicating a photochemical production regime. Here, most of the opacity is due to “Mode 2” particles with radii around 1 \(\mu\)m. The same particles extend throughout the clouds, but become somewhat larger in the middle and lower clouds (Mode 2’). In the deepest layer, which the vertical stability data from temperature profile measurements suggest is in a tropospheric convective regime, a third population of large Mode 3 (6–35 \(\mu\)m in diameter) particles is also found. The existence of multiple, distinct modes is still neither
understood nor well delineated and depends to a considerable extent on the interpretation of the single profile obtained by the Pioneer Venus Large Probe. However, the dominance in the upper cloud of a “monodispersion” of spherical sulfuric acid droplets with a tight size distribution around a radius of $w_1$ mm is well established from polarimetry measurements. There is some evidence, also from Pioneer, that the Mode 3 particles in the lower cloud are nonspherical and this, and their large size, suggests that they might contain solid crystals.

4.3. Global Variability

Ground-based observations first revealed enormous variations in the optical thickness of the lower cloud deck, which were later studied in detail from spacecraft. The clustered appearance of the deep clouds is consistent with tropospheric cumulus, in contrast to the muted variability and small contrasts seen in the upper clouds. The spatial variation of sulfuric acid concentration in the cloud particles has been estimated from Venus Express VIRTIS spectral maps and found to be higher in regions of optically thick cloud. The retrieved cloud base altitude varies with latitude, reaching a maximum height near $-50^\circ$ before falling by several kilometers toward the pole, along with a similar fall in the cloud-top height (Figure 14.10). The cloud particles in the polar region have different scattering properties from elsewhere on the planet and this has been interpreted as being due to an increase in average particle size, and possibly a difference in composition, near the pole.

The latitudinally variable CO abundance at 35–40 km altitude originally found by Galileo has been confirmed by Venus Express. An increase in CO collocated with a decrease in tropospheric H$_2$O abundance is observed at high latitudes, which is probably due to strong downwelling between $-60^\circ$ and $-75^\circ$ latitude marking the poleward extent of the Hadley cell circulation. In addition, tentative evidence for long-term secular change, over a period of 2...
Earth years, was observed in the acid concentration and the CO and H$_2$O abundances.

### 4.4. Cloud Chemistry

The high abundance of sulfur dioxide in the atmosphere leads to the formation of the concentrated sulfuric acid cloud layers via a chemical system involving the photolytic destruction of carbon dioxide by solar UV radiation, summarized by

$$\text{CO}_2 \rightarrow \text{CO} + \text{O}$$

followed by reactions equivalent to

$$\text{SO}_2 + \text{H}_2\text{O} + \text{O} \rightarrow \text{H}_2\text{SO}_4.$$  

This sequence forms the acid near the visible cloud tops, where it combines with other H$_2$O molecules to produce the hydrated acid droplets that are the main constituent of the clouds. The degree of hydration varies between perhaps 10% and 25%, with 20% (4H$_2$SO$_4$.H$_2$O) typical.

A cloud particle of the observed mean radius (~1 μm) has a sedimentation velocity of 7.5 m/day at 60 km; this velocity varies as the square of the size. Although small, these velocities, aided by coagulation, eventually carry the particles out of the cloud to lower altitudes and higher temperatures, where they will evaporate and, at still lower heights, decompose back into water and sulfur dioxide. Atmospheric mixing carries these gases back upward where they can again contribute to the formation of H$_2$SO$_4$. An important intermediate is the reactive free radical SO, and probably some elemental sulfur is produced. UV spectra (pertaining to the region above the clouds) reveal the presence of small amounts of SO$_2$ shown in Table 14.1, but much less than the amounts that have been measured below the clouds.

Sulfuric acid is perfectly colorless in the blue and near-UV regions, and the yellow coloration that provides the contrasts of Figure 14.1 must be caused by something else. The most likely thing is elemental sulfur, but yellow compounds are abundant in nature, and the identification remains tentative. The photochemical models do predict production of some sulfur, but it is a minor by-product, and the amount produced is uncertain. It is also unclear what constitutes the large Mode 3 particles in the lower cloud. Optical data suggests solid, irregular particles coated with sulfuric acid; the most likely candidate for the solid material is volcanic ash.

### 4.5. Lightning

Electromagnetic pulses, attributed to lighting bursts, have been observed by several entry probes and orbiters. Most recently, *Venus Express* confirmed the existence of “whistler” mode waves with burst durations of about 100 ms and properties similar to signals generated by atmospheric lightning in the terrestrial ionosphere. The frequency of occurrence, mapped over more than 4 years, suggests a level of lightning activity on Venus that is also similar to Earth’s. However, searches for the corresponding optical flashes have been negative, except for one ambiguous inference from *Venera 9* and a few optical events reported from Earth-based observations. A close flyby by the *Cassini* spacecraft saw no evidence of any impulses with a sensitive instrument that, in a later Earth encounter, found them in abundance. Theoreticians have opined that conditions on Venus do not seem propitious for large-scale charge separation, noting that on Earth, lightning is seen during intense precipitation and in volcanic explosions. In thunderstorms, large drops are efficient at carrying charge of one sign away from the region where it is produced, and the gravitational force is large enough to resist the strong electric fields, but this is not the case for small particles. The evidence for large, precipitating particles on Venus remains incomplete, as indeed are direct observations of volcanoes, although both are difficult to detect and either or both may turn out to be the source of the lightning behind the radio bursts.

### 5. GENERAL CIRCULATION AND DYNAMICS

#### 5.1. Surface and Lower Atmosphere Wind Profiles

The Soviet landers *Venera 9* and *10* used a simple cup anemometer, i.e. a rotating vane device similar to those seen on most earthbound meteorological stations, and found velocities of ≤1 m/s. As noted above, such slow winds are in line with expectations based on the atmospheric density and small thermal contrasts at these levels. It is also possible to track the drift of descent probes as they pass through the atmosphere, and so to obtain vertical coverage and directional data, as well as wind strength. The *Venera* landers were tracked by the measurement of the Doppler shift in the radio signal from the spacecraft, while the *Pioneer Venus* probes used an interferometric technique involving more than one receiving station.

The results (Figure 14.11) confirm the low surface winds but show a dramatic increase with height to speeds near 100 m/s near the cloud tops, more than 50 times faster than the rotation rate of the surface below. This zonal “superrotation” also manifests itself in the observed cloud structure, which moves rapidly around the planet in a direction parallel to the equator. The cloud markings, which appear with high contrast through an UV filter, have their origin at heights near 60 km above the surface (where the pressure is of the order of 100 mb). Motions in the deeper cloud layers can be observed by near-IR imaging on the
night side of the planet in the “windows” at wavelengths from 1 to 3.5 \( \mu m \). These originate in the main cloud deck, illuminated from behind by the hot lower atmosphere. The typical velocities inferred near the equator were about half as fast as those from UV markings, consistent with the vertical profiles of wind and cloud opacity measured by the Pioneer and Venera probes. Because the density increases by a large factor over this height range, the angular momentum is a maximum at 20 km.

5.2. Zonal Winds and Superrotation

Superrotation is observed in many atmospheres, usually at the low pressures found in thermospheres, and superposed on a rapid planetary rotation. Venus exhibits the most dramatic example, with zonal winds in excess of 100 m/s seen in the circulation of UV cloud markings at relatively high pressure levels of around 100 mb. The time-averaged global maps of the Venustian middle atmosphere temperature field obtained by remote sensing (Figure 14.6) show several features clearly related to the general circulation, including the temperature increase from pole to equator over a broad altitude range, against the trend in radiative heating, and the variation with local time of day of the air temperature. A model-dependent analysis, requiring the parameterization of viscosity, particularly that due to eddies, and assumptions of cyclostrophic balance that break down near the equator, has shown that the temperature field is consistent with the observed winds.

Attempts have been made to explain the high zonal wind speeds in the relatively dense atmospheric regions on Venus by mechanisms that convert the slow motion of the Sun, relative to a fixed point on Venus, into a much more rapid motion of the atmosphere. Currently prevailing opinion favors a mechanism in which momentum from the solid planet is transported by eddies whose interaction with the main flow is complex and in which the mean meridional circulation plays an important role. Experiments with general circulation models based on this principle suggest that global superrotation is always a characteristic of optically thick atmospheres on slowly rotating planets, with the predicted wind speed depending critically on the detailed energy deposition profile of the atmosphere. Such models can also explain the observed superrotation of Titan’s atmosphere. Their validation requires more information about wind speed versus latitude and height, cloud variability and wave modes in the atmosphere below the visible cloud tops, and the role of the surface topography in maintaining or opposing the superrotation, some of which is currently being provided by UV and near-IR mapping from Venus Express.

5.3. Meridional Wind Field

The cloud motions which trace the zonal winds also reveal the pattern of the meridional circulation on Venus, although with much larger proportional errors, since the poleward component is much slower than that parallel to the equator (Figure 14.11). Despite the uncertainties, the data fairly consistently show mean poleward motions of up to 10 m/s at most latitudes, tending to confirm the theoretical expectation that Hadley cells exist in each hemisphere, i.e. global-scale circulation cells characterized by rising motion at low latitudes and descending motion nearer to the poles. It has been suggested that the Pioneer Venus probe tracking data, showing alternations in the direction, as well as the magnitude, of the meridional wind, marks the passage of the probe through the different components of a stack of Hadley cells, each extending from the equator to high latitudes. This notion is supported by the fact that the Hadley cell seen at the cloud tops appears to be thermally indirect, that is to say, carries heat from the equator to the pole against the observed temperature gradient. It may be driven by a stronger, direct cell underneath, and the layered eddy sources and sinks which could drive the zonal superrotation may be related to the cell interfaces, although the current data is inadequate to establish this with any certainty.
5.4. Meteorology

The tracking of meteorological features on Venus was, for many decades, limited to the transient and quasi-permanent features seen in the UV images of the cloud-top region, where they revealed structures identified with Rossby and gravity wave activity, and the measurement of temperature anomalies by remote sensing. The spectral imaging instruments on *Venus Express* have exploited the fact that the near-IR windows permit imaging of the deep cloud structure to investigate the meteorological activity that is clearly present in the deep atmosphere of Venus. Along with high-resolution UV images, these reveal chaotic convective and wave activity near the equator where most of the solar energy is deposited in the clouds, with an abrupt switch to a more laminar flow at midlatitudes, and then finally a further transition to the polar vortex complex near the poles.

5.5. Polar Vortex

Vortex behavior occurs in the polar region of any terrestrial planet, due to the subsidence of cold, dense air and the propagation of zonal angular momentum in the meridional flow. On Venus, the small obliquity and the equatorial superrotation lead to an extreme version of this effect, manifest by a sharp transition in the circulation regimes in both hemispheres at a latitude of about 65°. There, a complex instability develops, resulting in dramatic long-lived wave structures. The *circumpolar collar* takes the form of a belt of very cold air that surrounds the pole at a radial distance of about 2500 km and has a predominantly wavenumber-1 structure locked to the Sun. The vertical extent of the collar must be much less than its 5000 km diameter, and the indications from *Pioneer Venus* and *Venus Express* data are that it may be only about 10 km deep, with a complex vertical structure. The cloud-top temperatures that characterize the collar are about 30–40°C colder than at the same altitude outside, so the feature generates pressure differences that would cause it to dissipate rapidly were it not continually forced.

Poleward of the collar, the air at the center of the vortex must descend rapidly to conserve mass, and we expect to find a relatively cloud-free region at the pole, analogous to the eye of a terrestrial hurricane but much larger and more permanent. Interestingly, however, the “eye” of the Venus polar vortex is not circular but elongated, and with typically two brightness maxima (possibly corresponding to maxima in the downward flow) at either end of a quasi-linear feature connecting the two (Figure 14.12). This wave-2 characteristic gives the polar atmosphere a “dumbbell” appearance in IR images that use the thermal emission from the planet as a source and has led to the name *polar dipole* for the feature. A dipole was first seen at the North Pole by *Pioneer Venus*, and a similar feature has been discovered and extensively studied at the South Pole as well by *Venus Express*. In particular, it has become clear that the “dipole” description is too simplistic: more complicated shapes, as well as monopoles and tripoles, also occur, with remarkably rapid (for such a large feature) morphing between them, although wave-2 does seem to dominate as some theories predict.

The northern dipole was observed in successive images obtained in 1979–1980 to be rotating about the pole with a period whose dominant component, among several, was 2.7 Earth days, i.e. with about twice the angular velocity of the equatorial cloud markings. If angular momentum were being conserved by a parcel of air as it migrated from equator to pole, the dipole might be expected to rotate five or six times faster. In fact, the UV markings are observed to keep a roughly constant zonal velocity (solid-body rotation) from the equator to at least 60° latitude, and must be accelerating poleward of this if the rotation of the dipole represents the actual speed of mass motions around the pole.

![Figure 14.12](image-url) The Venus polar “dipole”, left the North Pole, by *Pioneer Venus*, and right, the South Pole, by *Venus Express*. The bright feature is approximately 2000 km across in both cases.
and not simply the phase speed of a wavelike disturbance superimposed on the polar vortex. Southern dipole observations by Venus Express found a rotation rate that varied from 2.2 to 2.5 days, and confirmed that, in addition to a variable rotation rate, the position of the apparent center of the vortex can wander several degrees away from the rotation pole of the planet.

5.6. Tides

A particularly important form of wave motion is the solar tide, the temperature and density cycle induced by the apparent motion of the Sun overhead. This contains a whole spectrum of Fourier components, because the forcing is nonsinusoidal; the actual atmospheric response depends on the mean wind and the interference between the various components. The Earth’s atmosphere has a small wavenumber-2 component superposed on the familiar early afternoon maximum to postmidnight minimum cycle, but this component dominates on Venus. In fact, the dynamical theory of atmospheric tides, as developed for Earth, shows when applied to Venus that the observed state of affairs can be explained as primarily a consequence of the long solar day on Venus, provided that a realistic representation of the zonal wind is incorporated.

5.7. Upper Atmosphere

There are no direct wind measurements above the cloud tops, but deductions from temperature measurements and the measurement of Doppler-shifted emission lines from atmospheric gases suggest a slowing of the 100 m/s flow up to perhaps the 100-km level. At still greater heights, the dominant circulation switches to a rapid day—night flow, first suggested on theoretical grounds and confirmed by the large observed temperature difference. But the flow is not quite symmetrical; maxima in the hydrogen and helium concentrations, and in several airglow phenomena, are systematically displaced from the expected midnight location toward morning. The thermospheric winds carry the photochemical products O, CO, and N from the day side to the night side, where they descend into the middle atmosphere in a region ~2000 km in diameter and generally centered near the equator at 2 a.m. local time. This region can be observed by the airglow emitted during the recombination of N and O atoms into NO molecules, which then radiate in the UV, and O2 molecules, which radiate in the near IR. The light gases hydrogen and helium are also carried along and accumulate over the convergent point of the flow; for these gases, the peak density is observed at about 4 a.m. These offsets are the principal evidence that this part of the atmosphere rotates with a 6-day period, a rotation that is superposed on the rapid day-to-night flow.

6. EVOLUTION OF THE ATMOSPHERE AND CLIMATE

Earth and Venus have a common origin if, as is generally believed, the Sun, the planets, and their atmospheres condensed, about 4.6 billion years ago, from the same solar nebula (See The Origin of the Solar System). An intermediate stage was the formation of planetesimals, typically Moon-sized objects in noncircular orbits that collided and merged, so that those in the inner solar system might end up as part of any of the terrestrial planets. One would therefore expect them to begin with similar atmospheric compositions, and indeed those of Venus and Earth have many similarities and the same gases are present in each, although in different proportions.

The key question is whether the large divergences in the surface climate we now see have arisen from different evolutionary paths, starting from quite similar early atmospheric conditions, and to what extent the two planets may always have been very different. In either case, the amounts of the various gases in the atmosphere are expected to vary in time, including those that contribute to heating the surface, especially carbon dioxide. Water vapor and sulfur dioxide are also “greenhouse” gases, and are the main precursors of the clouds, which reflect some of the solar energy that would otherwise heat the planet. These and other gases like carbon monoxide and argon are released from the interior through volcanic activity and outgassing, while water in particular may also be brought in by icy debris from space.

At the same time as gases are being added, they are lost by various processes. Carbon dioxide and sulfur dioxide have well-studied reactions with rocky minerals that are likely to be common on the surface of Venus. Water and carbon dioxide are dissociated by solar UV radiation, but the CO2 tends to recombine, while the light H atoms from H2O escape relatively easily into space from the top of the atmosphere. Venus Express recently discovered that oxygen atoms also escape in large quantities, assisted by the solar wind flux impinging directly onto the planet with no interceding magnetic field. Current knowledge of the various key processes and their effect on the climate of Venus over time are reviewed by Taylor and Grinspoon (2009) and summarized below.

6.1. Sources of Atmospheric Gases: Volcanism

The existence of the atmosphere on Venus, as on Earth, depends primarily on the venting of gases from the interior, which has continued at some level up to the present, accompanied by an unknown contribution from the infall of icy material as cometary (i.e. volatile rich) and meteoritic dust and larger fragments. The main evidence for recent
volcanism on Venus is the abundance of volcanic structures seen on the surface, many of them of pristine appearance with apparently fresh lava flows, and the high level of reactive, and therefore, relatively short-lived, volcanic gases in the atmosphere, especially SO₂. Searches for “hot spots” on the surface, and plumes of volcanic effluvia, which would confirm current active volcanism and help establish its level compared to Earth, have been attempted with tantalizingly suggestive, but so far inconclusive, results.

If Venus has the same total heat flux as Earth and, in the absence of tectonics, it is all accounted for by volcanoes, then we would expect 1000 times as much gas, in particular sulfur dioxide, to be released. In fact, there is approximately 100,000 times as much SO₂ in Venus’ atmosphere compared to that of the Earth’s. The difference of a factor 100 could conceivably be explained by the fact that this is approximately the ratio of the lifetimes of the gas on the two planets when the efficient rainout mechanism that applies near the surface on Earth is taken into account. Such crude attempts at balancing the budget must admit additional unknowns; for instance, volcanism on Venus may release a mix of gases different from that on the Earth, where each volcano is at least slightly different from every other in any case.

The most that can be said with reasonable certainty at present is that the existence of SO₂ and H₂SO₄ in the amounts observed must imply substantial outgassing in recent geological time. The present flux of volcanic gases into the atmosphere remains unknown and could conceivably be as low as zero, although the latter seems very unlikely. The time constant for the decline of the sulfur abundance in the atmosphere if the source was removed is of the order of 1 Myr, much shorter than the period since the hypothetical global resurfacing event 750—500 Myr ago (see Venus Surface), indicating that the atmospheric sulfur must be of recent origin. The very high variability in sulfur dioxide abundance measured at and above the cloud tops has been extensively interpreted as indicating current volcanism. While it is not possible at present to associate these with specific eruptions on the surface, and transport effects due to local meteorology need also to be considered, large SO₂ variations are seen at comparable pressure levels in the terrestrial upper atmosphere following large eruptions. Short-lived but very bright plumes in the stratosphere, seen in UV images of reflected sunlight, and dark plumes in deep tropospheric clouds in near-IR images are also possible evidence for active volcanoes.

### 6.2. Surface Pressure and the CO₂ Budget

A plausible first-order explanation for the apparent superabundance of CO₂ on Venus relative to Earth is not particularly difficult to find. It has been estimated that the carbonate rocks on the Earth, of which the White Cliffs of Dover are a noted example, hold in total the equivalent of roughly 90 bars of CO₂ which would in the distant past have been in the atmosphere, although not necessarily all at once. The fact that the conversion of atmospheric to crustal carbonate occurs much more efficiently in the presence of liquid water to dissolve the CO₂ first suggests that the relatively water-depleted state of Venus may be responsible for so much of the gas remaining in the atmosphere there but not on Earth.

The evidence for such processes in Earth’s history, presumably accompanied by large changes in surface pressure (although this is hard to verify), raises the crucial question of whether the current surface pressure on Venus is stable. The possible effects of even quite small variations in the CO₂ abundance on the climate on Earth have received a great deal of recent attention and clearly similar or very much larger changes can be contemplated on past or future Venus, resulting from changes in the volcanic CO₂ output, for example. If the climate on Venus is in fact stable in the long term, then it is likely that some mechanism provides a buffer that stabilizes the atmospheric carbon dioxide content. It was suggested by Urey as long ago as 1953 that the exchange between atmospheric CO₂ and common minerals in the surface, viz.

\[
\text{CaCO}_3 \text{(calcite)} + \text{SiO}_2 \text{(quartz)} \leftrightarrow \text{CaSiO}_3 \text{(wollastonite)} + \text{CO}_2,
\]

might perform this function. Intriguingly, it has since been shown that this reaction reaches equilibrium at precisely the temperature and pressure found on the surface of Venus. However, the equilibrium is unstable, so additional factors and a more complex model of surface—atmosphere interactions are required to make a model of the climate on Venus that moves to, and stays at, the status quo. It is worth noting that we have no such model for Earth either.

### 6.3. Escape Processes and Loss of Water

Venus is generally taken to have essentially the same internal structure as the Earth, with sufficient differences to account for the apparent absence of dynamo action, as evidenced by the lack of an intrinsic magnetic field. The history of the field, if there was one in the past, is crucial for determining the rate at which volcanic gases including water vapor have escaped due to solar wind erosion.

The loss processes involve dissociation to form hydrogen and oxygen ions (Figure 14.7) which then escape from the planet. Hydrogen, but not oxygen, can escape by purely thermal processes (Jeans escape) but several other mechanisms are active for both. For the heavier atoms like oxygen, the dominant processes are probably sputtering and electric fields produced by charge exchange processes.
that accelerate the exospheric ions. Each of these has a different solar cycle average loss rate and depends ultimately on the abundance of water in the middle atmosphere. The Earth is protected to some extent from water loss by the very low temperatures near the tropopause, which form a cold trap. On Venus, the enhanced solar heating of its presumed primordial ocean would have evaporated additional water into the atmosphere, increasing greenhouse warming and raising the humidity still higher. This feedback may have continued until the oceans were gone and the atmosphere contained up to several hundred bars of steam. Whether it was ever cool enough for this to form a liquid water ocean is a matter of ongoing debate.

Either way, water vapor was probably for a long time the major atmospheric constituent, extending to high altitudes where it would be efficiently dissociated into hydrogen and oxygen by UV sunlight. Rapid escape of hydrogen would ensue, accompanied by the slower escape of the heavier deuterium, helium, and oxygen. This could explain why Venus now has 100,000 times less water, but 150 times more deuterium relative to hydrogen, in its atmosphere than exists in the oceans and atmosphere of the Earth. The difference in D/H is attributable to differences between the escape rates of H and D under given conditions, plus the fact that the mixing ratio of vapor in the upper atmosphere is different for the two water isotopes.

If total water on Venus is currently in a steady state between source (volcanic outgassing and cometary infall) and loss (dissociation and escape) processes, then the escape flux measures the time-averaged sum of these sources. All three components of the equation can, in principle, be measured, although recent progress is mainly limited to measuring the escape rates. While sampling the hydrogen (and helium) fluxes, the Analyser of Space Plasmas and Energetic Atoms (ASPERA) experiment on Venus Express found surprising evidence that a large flux of oxygen ions is also escaping from the upper atmosphere of Venus. Provisional estimates suggest a planetary average column escape flux in roughly the ratio of 2:1 for hydrogen and oxygen, seemingly indicating water as the parent. If confirmed, this would solve an earlier puzzle about the fate of the large quantities of oxygen that were thought to have been left behind following massive hydrogen escape.

The rates of atmospheric water loss that the ASPERA data imply are two orders of magnitude too small to balance the water budget of the planet even if the level of current volcanic activity on Venus is no more than that on Earth. As we saw above, estimates based on imaging data, SO2 abundances, and heat fluxes would place it several orders of magnitude higher. The explanation may be that the volcanic effluvia are much drier on Venus than they are here. Terrestrial ocean waters seep continuously into the hot crust, to be subsequently expelled as vapor, a process that is not available on Venus. Sulfur dioxide may be the dominant volatile driving explosive volcanism on Venus, rather than water as on Earth.

6.4. Evolutionary Climate Models

Attempts have been made to trace the origins and evolution of Venus’ atmosphere using simplified one-dimensional evolutionary climate models. These incorporate the large-scale processes and their interrelations in a globally averaged sense, neglecting or simplifying dynamics so that they can model the complex set of time-dependent feedbacks that control the planetary climate. Bullock and Grinspoon (2001) used a radiative transfer code to calculate the radiative–convective equilibrium temperature structure as a function of atmospheric composition and coupled this to models of the chemistry and microphysics of Venus’ clouds, volcanic outgassing, heterogeneous reactions of atmospheric gases with surface minerals, and the escape of hydrogen from the exosphere.

A key feature of such models is how they simulate the clouds, since these have two important effects on climate. They alter the visual albedo of the planet, changing the magnitude and profile of solar heating, and they alter the thermal IR opacity of the atmosphere, affecting the cooling at and above the surface. The formation of clouds depends on the abundance of sulfur dioxide, which in turn depends on the volcanic source and the loss processes at the surface. The current measured mixing ratio of SO2 is 180 ppm, which is more than two orders more abundant than that required for equilibrium with calcite, but it is close to equilibrium with pyrite and magnetite. The actual reaction rate will depend on the abundances of these and other potential reactants in the surface, on the relevant chemical kinetics, and on the ability of the gas to diffuse to new reaction sites on buried grains once the easily available surface has reacted. The choice of diffusion coefficient requires assumptions about soil porosity and the effectiveness with which forming sulfate rings will reduce pore space.

A case study that modeled the rapid supply of gases from the putative massive volcanic event responsible for depositing the extensive lava plains found that this could raise the surface temperature by as much as 100 °C for half a billion years, before relaxing to conditions similar to present ones. A second study that ran the model into the future predicted that the present situation is unstable and that after a billion years the clouds may disappear altogether. The surface temperature is predicted to fall by about 50 K since the reduced opacity more than offsets the increased absorption of incoming solar energy. A large number of other scenarios can be imagined, depending on the rate and timing of the events that might supply extra gases and the ratio of water to sulfur dioxide in each event.
For example, the impact of a large comet would supply mostly water vapor, with relatively little sulfur dioxide.

Figure 14.13 shows two simple radiative—convective models and a measured temperature profile for the middle atmosphere of Venus from the Magellan radio occultation experiment. The simple models have a stratosphere in radiative equilibrium with the Sun, overlying a deep atmosphere in which the profile follows a dry adiabat. The solid line is such a model calculated assuming present-day conditions; the dashed line is an imaginary scenario in which the surface pressure on Venus falls to 1 bar and the planetary albedo falls to 0.52, that is, to a less cloud-reflective state, as pictured by Arrhenius a century ago. This may yet occur, perhaps as the result of continued exospheric loss and chemical erosion of the atmosphere following a cessation of volcanic activity at some distant point in the future. If so, a surface temperature in the region of 320 K is predicted, the Earthlike “tropical” scenario envisaged by Arrhenius and other early visionaries long ago.

BIBLIOGRAPHY


1. INTRODUCTION

Venus plays a pivotal role in understanding the evolution of the terrestrial planets, the four rocky bodies closest to the sun. Venus is the planet most similar to Earth in terms of radius and density, implying a very similar bulk composition. Since the terrestrial or inner planets have all formed via the same process, condensing out of the solar nebula, the primary factor that distinguishes them is their size, and to a lesser extent, distance from the sun. The energy available to drive geologic evolution comes from the heat of accretion and from decay of radiogenic isotopes. The majority of geologic activity on the smaller bodies of the inner solar system, Mars, Mercury, and the Moon, occurred during the first $1-2$ billion years. Larger planets like Earth and Venus have a greater volume of radiogenic elements and can be expected to be geologically active longer. Earth has abundant geologic activity today. We are uncertain about the present-day level of activity on Venus, but it has clearly been extremely active within the last billion years since current data reveal no evidence of impact craters that must have accumulated previously. However, Venus does not currently have a system of \textit{plate tectonics} that governs the pattern of geologic activity on Earth (see Earth as a Planet: Surface and Interior). Clearly, size is important in determining the duration and extent of geologic activity, but other factors must affect the overall style of geologic evolution. The atmospheric conditions on Venus are also wildly different from those on Earth. Venus’ atmosphere contains a runaway greenhouse effect, in which abundant carbon dioxide causes the atmosphere to heat up, trapping more water vapor. The water vapor increases the infrared opacity, which further heats up the atmosphere (see Venus: Atmosphere). The resulting thick, dense atmosphere gives
Venus a surface temperature of about 468 °C (874 °F), and a pressure 90 times greater than that of Earth’s. For this reason, Venus has been called Earth’s “evil twin”.

Volatile, such as water and carbon dioxide, are essentially the link between the atmosphere, the surface, and the interior, as well an essential element in the habitability of a planet. A planet’s atmosphere forms primarily through the outgassing of volatiles from the interior. Outgassing results from the eruption and degassing of lava onto the surface. Volcanic and tectonic resurfacing on a planet is driven by heat loss from the interior, which is primarily fueled by decay of radioactive elements. The interiors of the larger terrestrial planets are hot enough to convect, allowing hot material to rise and cold material to sink on timescales of millions of years. On Earth, convection is linked to surface processes via the process of plate tectonics. The presence of water in the interior of Earth acts to reduce the strength of the rock, which in turn allows the exterior shell of the Earth to be broken up into plates. As plates are pushed back into the interior, water is recycled back into the interior. Volatiles on Earth are also strongly affected by both the hydrosphere and the biosphere, both lacking on Venus.

Although plate tectonics has controlled the evolution of Earth for at least 3 billion years, the surface of Venus shows no clear evidence of such a process. Venus lacks the interconnected network of crustal spreading zones, where new crust is continually created as at the midocean ridges of Earth, and on the other side subduction troughs, where crust is pushed back into the mantle. Although it lacks the clear plate boundaries evident in Earth’s topography, there are major rifts and even possible subduction sites. Most explanations of why Venus does not have plate tectonics point to the very low amounts of water currently present on Venus. The water in the atmosphere is equivalent to a surface layer less than 10 cm thick. The abundance of heavy hydrogen, or deuterium, in the atmosphere relative to the normal hydrogen population indicates that a huge amount of water was lost from Venus atmosphere early in its history. The dry atmosphere could imply a dry interior for Venus, which is believed to make the outer shell on Venus too stiff to break into the plates observed on Earth.

Although plate tectonics does not operate on Venus, it is clearly an active planet with a relatively young surface and a wealth of volcanic and tectonic features. The majority of the planet is covered with volcanic features such as shield volcanoes and lava plains directly analogous to Earth’s volcanic features. Many of the highland areas appear to form over mantle plumes, where hot material from the interior rises to the surface creating “hotspots” on the surface similar to Hawaii. In contrast, many of the tectonic features are unique to Venus. Examples include coronae, which are believed to result from small-scale plumes deforming the surface, and tessera, which are intensely deformed regions with multiple intersecting fracture sets.

2. HISTORY OF VENUS EXPLORATION

Venus has been long observed as one of the brightest objects in the evening or morning sky. Transits of Venus had been used to determine a value for the astronomical unit, and thus for the absolute scale of the solar system. During the transit of 1761, Lomonosov discovered that Venus had an atmosphere. But it was not until the 1960s that the modern exploration of the surface of Venus began, with observation by Earth-based radio telescopes. Radio telescopes at Arecibo in Puerto Rico and at Goldstone in California were used to accurately measure the rotation period and diameter of Venus. They also produced images of the surface that showed large, continent-size regions. However, Earth-based radio telescopes were hindered by only being able to image the same side of Venus that faced Earth at inferior conjunction (see Solar System at Radio Wavelengths).

Spacecraft observation of Venus began in 1962 with a flyby by the Mariner 2 spacecraft. It observed Venus from 34,833 km, determining a 468 °C (874 °F) surface temperature and that Venus lacked a magnetic field. In 1967, Mariner 5 flew by Venus at an altitude of 4023 km returning data on atmospheric composition and surface temperature. Also in 1967, the first probe entered the Venus atmosphere, when the Soviet Union’s Venera 4 returned data for 93 min. The Venera 5 and 6 probes followed in 1969, sending back more atmospheric measurements. Two more Venera probes followed in 1970 and 1972 making soft landings on the surface, with Venera 8 in 1972 transmitting data on surface temperature, pressure, and composition. The Venera 8 measurements were initially thought to be consistent with a granitic composition (see Section 5 for more discussion).

The next US mission to observe Venus was Mariner 10 in 1973, which was on its way to Mercury. Mariner 10 provided observations of the atmospheric circulation of Venus with both visible and ultraviolet wavelengths. In 1975, the Soviet Union landed two more probes on the surface of Venus, Veneras 9 and 10, sending back panoramas of the surface for the first time (see Section 2), and making detailed geochemical measurements. These landers measured surface compositions similar to terrestrial basalts.

The US Pioneer Venus mission in 1978 consisted of an orbiter plus four atmospheric probes. The probes returned data on atmospheric circulation, composition, pressure, and temperature. The orbiter provided radar images of the surface, as well as a detailed global topographic map with a resolution of about 150 km. Major topographic regions such as Aphrodite Terra and Bell Regio were mapped, as
well as the 11-km-high Maxwell Montes. The spacecraft was also used to map the gravity field of Venus.

The Soviets followed with four more soft landers between 1978 and 1981, with three of the landers returning surface panoramas and surface compositional information. The last two soft landers (Veneras 13 and 14) returned color panoramas (see Section 2), and drilled into the surface for samples. The next two Soviet missions were orbiters, Veneras 15 and 16, and returned synthetic aperture radar (SAR) images of the northern hemisphere of Venus in 1983, with resolutions of about 5–10 km. This rich data set revealed new types of features on the surface of Venus, including tessera terrain and coronae (discussed below). Vega 1 and Vega 2 in 1984 carried balloon probes into the atmosphere, and were the Soviet Union’s last missions to Venus.

The National Aeronautics and Space Administration (NASA)’s Magellan mission to Venus was launched in 1989 from the space shuttle and arrived at Venus in August of 1990. It obtained SAR images and altimetry of the surface between 1990 and 1994, mapping over 98% of the surface. The spacecraft also obtained high-resolution gravity field measurements, especially after the orbit was lowered and circularized in 1993. The 120 to 250 m-resolution SAR images and 10 to 27-km-resolution altimetry data completely unveiled the surface of Venus and provided a global data set that could be used to test models of the interior and surface evolution of the planet.

In 2005, the European Space Agency launched Venus Express. The primary mission objectives are the composition and circulation of the atmosphere of Venus. In addition, this mission has provided a valuable new data set for the surface of Venus. Ground-based observations and infrared images acquired during the Venus flyby of the NASA mission Galileo in 1991 showed that some spectral windows permit thermal emission of the surface to escape to space. The visible and infrared thermal imaging spectrometer (VIRTIS) was able to image most of the southern hemisphere in a window near 1 μ. The derived near-infrared surface emissivity provides information on differences in surface composition. Additionally, local maps of surface emissivity in more equatorial regions were derived from the images of the Venus Monitoring Camera.

The Japanese Space Agency, JAXA, launched Akatsuki to study the atmosphere and surface of Venus in 2010. However, it did not achieve orbit as planned. Another attempt at orbit insertion is planned for 2016.

3. GENERAL CHARACTERISTICS

3.1. Orbital Rotations and Motions

Venus orbits the Sun in a nearly circular path once every 224.7 Earth days. It is the second planet from the sun, located between Mercury and Earth. The plane of Venus’ orbit is inclined to that of the Earth by 3.4°. Analysis of the obliquity of Venus reveals that it has a liquid core, similar to that of Earth. One day on Venus lasts 116.7 Earth days. The rotation of Venus on its axis is not only extremely slow, but occurs in the opposite direction from all the other planets (retrograde rotation), so that the sun rises in the west.

The location of topographic features based on their surface brightness in Venus Express VIRTIS data has indicated that the length of day on Venus averaged over 16 years is ~0.002 days longer than when it was estimated over 20 years ago by Magellan. This change in the length of day could indicate angular momentum exchange between the Venus’ dense, superrotating atmosphere and the solid body.

When visible, Venus is the brightest planet in the night sky due to its size, albedo, and proximity to both the Sun and the Earth. Its easy visibility and the unusual pattern it makes in the night sky have given Venus a special place in astrology and the mythology of ancient civilizations, as well as made it an easy target for stargazers. Its proximity to the Sun means that it never rises very high in the sky, but can often be seen as either the “evening star” in the west, or as the “morning star” in the east.

3.2. Radius, Topography, and Physiography

The radius of Venus is 6052 km, only 5% less than the equatorial radius of the Earth, 6378 km. The average density of Venus is 5230 kg/m³, somewhat higher than Earth’s density. Thus, the acceleration of gravity at the surface is 8.87 m/s², 90% of Earth’s. The radius of the Earth measured at the poles is approximately 21 km less than the radius at the equator. This difference is called the “rotational bulge”. The Earth’s spin accelerates the equator more than the pole, causing the pole to be flattened and the equator to bulge out. The very slow rotation of Venus means that no such flattening occurs, making it, on average, nearly spherical.

The topography on Venus is dominated by plains, which cover at least 80% of the planet. There are also major highlands, including plateaus and topographic rises, as well as rifts and ridge belts that stand out from the background plains (see Figure 15.1). Venus and Earth both have a large topographic range due to the intense geologic activity that the two planets have experienced. However, the distribution of elevations on the two planets is very different (see Figure 15.2). Earth’s topography is bimodal, while Venus’ topography is unimodal. The two peaks on Earth reflect the division between oceans and continents. Venus has no ocean, and as we will discuss below, arguably no continents. The topography on Venus differs from that on Earth in other significant ways. Most importantly, Venus lacks the
interconnected system of narrow midocean ridges and long linear mountain belts that are the hallmark of plate tectonics on Earth (see Figure 15.1). The absence of these features on Venus reflects fundamental differences in evolution between the two planets, and will be discussed in greater detail below.

3.3. Surface Conditions

The surface conditions on Venus can best be described as hellish. The surface temperature at the mean planetary elevation is 437 °C (867 °F), with most of this high temperature due to the Venus greenhouse, not proximity to the sun. The surface temperature at the highest elevations is approximately 10 °C less. The surface pressure is 95 bars, equivalent to the pressure under almost 1 km of water. The surface temperature varies by only about 1 °C over the course of a day (243 Earth days) due to the dense, insulating atmosphere and the very low obliquity. Note that a year on Venus (225 Earth days) is actually shorter than 1 day. The atmosphere is 96.5% carbon dioxide, with lesser amounts of nitrogen, sulfur dioxide, argon, carbon monoxide, and water. The clouds are largely sulfuric acid (see chapter on Venus Atmosphere).

3.4. Views of the Surface

Four Soviet landers have returned views of the surface of Venus, Veneras 9, 10, 13, and 14. These panoramas showed relatively similar sites: rocky surfaces with varying amounts of sediment (Figure 15.3). Rocks at each site tend to be relatively angular, suggesting minimal erosion and possible ejection from an impact crater. All the sites are consistent with a volcanic origin, showing platy lava flows that have been covered to varying extents by sediments. The sediments may be of impact origin, produced by aeolian erosion or by chemical weathering.

4. IMPACT CRATOR S AND RESURFACING HISTORY

There are approximately 940 identified impact craters on the surface of Venus. They range in diameter from approximately 1.5–268.7 km. The dense atmosphere on
Venus causes impactors \( \leq 1 \) km in diameter to break up before impacting the ground, reducing the number of craters \( \leq 30 \) km in diameter. The shock waves that travel through these small objects can cause them to explode in a manner analogous to the Tunguska event on Earth (see Planetary Impacts). Atmospheric breakup and explosion, or other dynamic effects in the atmosphere, is believed to produce both radar-bright and radar-dark splotches on the surface (Figure 15.4). The brightness of a radar image is primarily a function of how rough the surface is at the scale of the radar wavelength (for the Magellan radar, 12.6 cm).

**FIGURE 15.2** Histogram of the elevation in 0.5 km bins for Earth and Venus, normalized by area.

**FIGURE 15.3** These photographs of the surface of Venus were obtained by the Soviet Venera 13 spacecraft. Venera13 was the first of the Venera lander missions to include a color camera. The Venera 13 lander touched down on March 3, 1982, near 305\(^\circ\) E 5\(^\circ\) S, in the plains east of Phoebe Regio. The arm on the surface in the top image is a soil mechanics experiment. A color bar for calibration is visible in each image, as well as other spacecraft parts.

**FIGURE 15.4** This radar image (approximately 125 by 140 km in size) shows an impact splotch with a dark center and a bright halo. The splotch is superimposed on a set of predominantly NW-trending wrinkle ridges. The spacing between major ridges is roughly 10–20 km. These wrinkle ridges are part of the set of ridges that wraps around Western Eistla Regio.
The darker the image is, the smoother the surface. Very rough areas appear very bright. Rough areas reflect the signal back to the spacecraft while smooth areas allow the radar waves to bounce off in a direction away from the spacecraft. Approximately 400 of these “splotch” regions have been identified. These regions are believed to be either areas where fine-grained material has been scoured away (radar-bright areas), or regions where relatively fine-grain material has settled out of the atmosphere (smooth, radar-dark areas). Additionally, most impact craters have dark parabolas associated with them, which are also part of fine-grained ejecta that are deposited out of the atmosphere.

In the absence of samples returned from planetary bodies, the only means of dating the surface is the analysis of the impact crater population. A great deal of work has been done on assessing the population of comets and asteroids available to impact the larger planetary bodies (see chapters on Comets, Asteroids, and Planetary Impacts). Dating of samples returned from the Moon has been used to tie the record of lunar craters to an absolute age. The estimated flux of impactors on the Moon must be extrapolated to other bodies in the solar system, which have different dynamical environments and thus different expected rates for impacts. This introduces a major uncertainty into the estimated age of a surface based on impact crater counts. Another major factor is the history of the surface itself. Modification of a surface by erosion, deposition, volcanic resurfacing, or tectonism can decrease the number of identifiable craters. Erosion can also remove deposits that had covered a crater. Additionally, secondary craters can form when large blocks of material are ejected during an impact. Impactors can break up during entry to the atmosphere, producing multiple smaller impacts rather than a single large impact. Despite these issues and the resulting uncertainties, estimated surface age is a very important clue in deciphering the geologic history of a planet.

The estimated age of resurfacing on Venus is \( \sim 750 \) My. Given all the possible uncertainties in this age, estimates between 300 My. and 1 billion years are permissible. This age is in contrast to ages of 3–4 Gy on average for Mars and the Moon. On Earth, new crust is continually forming along spreading centers in the oceanic crust. Continental crust can be old as 4 Gy, but craters are erased by water and wind erosion much more rapidly than on the other terrestrial planets.

There are two highly intriguing characteristics of the Venus crater population. The first is that the distribution of craters cannot be distinguished from a random population. The second is that apparently very few of the craters are modified by either volcanism or tectonism. Only \( \sim 17\% \) of the total population is either volcanically embayed and/or tectonized. An example of a crater that is both embayed and tectonized is Baranamtarra (Figure 15.5). The low number of modified craters on the surface of Venus means that there is little record of the process or processes that reset the surface age to be less than 1 billion years. If volcanic flows had covered the surface of Venus at a uniform rate, there would be more partially buried craters. This observation of the crater population initially led to the hypothesis of global, catastrophic resurfacing. Subsequent detailed modeling of resurfacing showed that the population is consistent with a wide range of resurfacing models, allowing for different size areas to resurface at different rates. Analysis of both crater density and modification of craters provides additional information that is most consistent with resurfacing occurring as small diameter regions, on the order hundreds of kilometers in diameter rather than on global scale. Another potential clue to the resurfacing history is whether or not craters with dark floors may have been volcanically flooded. Ultimately, more data are needed to provide constraints on the nature of processes that have resurfaced the planet.

We can estimate the rate of volcanic resurfacing if we assume that craters have been removed by burial under volcanic flows. Crudely, if one takes the characteristic resurfacing age to be 750 million years and the average crater rim height to be 0.5 km, then the rate of volcanic production is \( \sim 0.3 \) km\(^3\)/yr. Alternatively, if we consider the hypothesis that Venus resurfaced more quickly, in perhaps 50 Ma, the production rate is \( \sim 4.6 \) km\(^3\)/yr. The relative volume of lava extruded on the surface is believed to be small compared to the volume intruded into the subsurface, perhaps 10\% of the total. For comparison, the

![FIGURE 15.5](image)
estimated rate of volcanism for combined intrusive and extrusive volcanism on Earth is 20 km³/yr.

On planets with large numbers of craters, such as Mars, the surface age of local regions can be estimated from the crater populations. On Venus, some attempts have been made to determine the relative ages of either populations of specific types of geologic features or large areas on Venus. However, statistical analysis of this approach indicates that the very small number of craters on Venus makes attempts at dating particular landforms or even large areas not reliable. Although traditional crater-counting methods are not very useful, both the distribution of modified craters and the distribution of dark crater parabolas suggest some variation in surface age. In particular, the region with the highest density of volcanoes, coronae, and rifts appears to have a lower density of haloes and more modified craters, suggesting a younger age.

Overall, the crater population on Venus indicates it is a comparatively active planet, completely resurfaced within the last 1 billion years, possibly with resurfacing ongoing today. Volcanic resurfacing rates are likely on the same order of magnitude as those on Earth, but are a function of the poorly constrained rate of resurfacing, which could be either constant or variable. The distribution and modification of the craters imply that there are limited differences in the ages of large regions on Venus, unlike the dichotomy between the age of oceanic and continental crust on the Earth. The small number of modified impact craters leaves few clues as to the processes that obliterated the earliest surface of Venus. Below we discuss the implications of resurfacing for the overall geologic evolution of Venus.

5. INTERIOR PROCESSES

One of the greatest curiosities about Venus is that global-scale geologic processes are totally unlike that of Earth. The system that shapes the Earth's large-scale physiography and the majority of geologic features is plate tectonics. The surface of the Earth is broken into dozens of plates that move over the surface of the Earth at rates of up to a few centimeters per year. The plates are tens to hundreds of kilometers thick. Mountain belts form where plates meet, such as where they collide, slide at an angle past each other, or where one plate is pushed into the mantle beneath another at subduction zones. Hot material wells up from the mantle below along narrow ridges in the ocean crust, creating new oceanic crust. These characteristic features are easily seen in the topography of the Earth, even at the relatively low resolution available for Venus (Figure 15.1). Venus clearly does not have plate tectonics. There is no evidence for this type of geologic process in the topography or in the radar images (see Earth Surface and Interior).

The energy that drives plate tectonics and other geologic processes is dominantly generated by the decay of radioactive elements. For the terrestrial planets, the primary contributors to radioactive decay are uranium (U), thorium (Th), and potassium (K). Based on estimates of the abundance of these elements on Earth and in chondrites (see Meteorites), radioactive decay cannot account for the total amount of energy. In addition, a significant amount, perhaps 25%, of the heat lost from the interior results from cooling of the planet over time, with some additional contribution from the heat of initial planetary accretion. The heat in the interior of the planet is dominantly transmitted to the surface via convection in the interior. Convection in the mantle brings hot, low-density material from the interior to the surface, or near the surface, allowing it to cool.

Generally speaking, the larger a planet, the longer it will continue to lose energy and be geologically active. However, the details of the thermal evolution are complex. Numerous factors affect thermal evolution, including accretion, differentiation, composition, convective style, and amounts of volcanism. Venus and Earth provide perhaps the quintessential example of variations in evolution. Most explanations of how Venus and Earth ended up on different geologic paths have to do with the history of volatiles. Volatiles, mainly in the form of water, play a key role in enabling plate tectonics on Earth. The presence of even a small amount of water in rock has a major effect on its strength and on the temperature at which it will melt. The water in the lithosphere is believed to be essential to making it weak enough to break into plates and subduct in response to the motions of convection in the interior. The asthenosphere is the upper part of the mantle, directly below the lithosphere, which has a lower viscosity than the rest of the mantle and acts to lubricate the motion of the plates at the surface of the Earth. The low viscosity of the asthenosphere may be a result of small amounts of melt. Melt would not be expected in the asthenosphere unless at least a small percentage of water is present. Thus, water appears to be an essential ingredient in the development of plate tectonics.

Measurements made to date indicate that the atmosphere of Venus has very little water, on the order of tens of parts per million. The upper crust is inferred to be dry as well, although Ar isotopes in the atmosphere indicate that only about 25% of interior volatiles have been lost. In terms of the strength of the crust, the extremely high surface temperatures might be expected to offset the lack of water, making the crust extremely weak. However, laboratory studies of rock strength at Venus temperatures have shown that dry basalt (see Section 5 below) is stronger than wet basalt at Earth temperatures. This extreme strength of the crust on Venus likely contributes to the apparent lack of lithosphere scale breaks that are required to form plates. As we discuss below, there is also evidence suggesting that Venus has no asthenosphere.
Convection studies have proposed that Venus exists in a "stagnant lid" mode rather than the "active lid" mode predicted for Earth. When convective stresses exceed the lithospheric strength, an active lid such as the terrestrial system of plate tectonics is predicted. On Earth, conditions such as weak, narrow fault zones, or the presence of a low-viscosity asthenosphere, allow the convective stresses to exceed the lithospheric strength. On Venus, the present-day lithospheric strength is apparently too high to allow plates to develop. This model is consistent with the loss of volatiles as key to differences on Venus and Earth.

Given the similarity in heat-producing elements and size between Earth and Venus and the absence of plate tectonics on Venus, how does Venus lose its heat? Venus must be convecting in its interior. Although there is no evidence for present-day plate tectonics, some numerical simulations suggest that Venus may have had plate tectonics in the past, or may have experienced multiple transitions active and stagnant lid regimes. The greater heat loss during plate tectonics causes sufficient cooling of the mantle and lithosphere to initiate a stagnant lid regime. The insulating lid then causes the mantle to heat up and revert to plate tectonics. The lack of plate tectonics today suggests that Venus may be heating up today.

In addition to the deformation of surface plates, hot plumes from the interior can affect surface geology. On Earth, hot blobs of material form within the overall convecting pattern in the interior. These plumes form "hotspots", such as the Hawaiian Island chain. The hot mantle material pushes up on the lithosphere, creating a broad topographic swell. The heat causes the lithosphere and crust to melt locally, thickening the crust and forming surface volcanoes. On Earth, the majority of the heat is lost where the upwelling mantle creates new crust at midocean rises and cold lithosphere is pushed back into the mantle at subduction zones. Hotspots account for <10% of Earth's heat loss. Venus appears to have a similar number of hotspots as Earth, providing evidence of current convection and contributing to heat loss.

There are approximately 10 such hotspot features on Venus. These rises are Atla, Bell, Beta, Dione, W. Eistla, C. Eistla, E. Eistla, Imdr, Themis, and Laufey Regiones (Figure 15.6). Those features believed to be active today, such as Atla, Beta, and Bell Regiones, have broad topographic swells, abundant volcanism, and strong, positive gravity signatures. New evidence from surface emissivity gives further evidence of their present-day activity (see Composition). Several rises also have rifts, such as Guor Linea at W. Eistla Regio (Figure 15.7). These features are characteristic of hotspots above a mantle plume. However, there are too few hotspot features on Venus (~10 on Venus vs 10—30 on Earth) to account for a major portion of Venus’ heat budget. In addition to the large-scale (1000—2000 km diameter) hotspots on Venus, there are also smaller scale (mean diameter of ~250 km) features called coronae (see Section 7). There are ~515 of these features, which are unique to Venus. There is considerable evidence that many of these features form above small-scale plumes. Some may be a result of cold lithosphere becoming gravitationally unstable and sinking back into the interior, which also acts to cool the planet. However, all these heat loss mechanisms together would not be able to account for more than about one-quarter of the interior heat loss on Venus.

The relationship between the gravity and topography provides evidence that Venus does not have a low-viscosity asthenosphere. On Earth, a mantle plume must pass through the asthenosphere before reaching the lithosphere. (Note that there is not an asthenospheric layer beneath the very thick continental lithosphere on Earth.) The plume tends to spread out in the relatively weak asthenospheric layer, resulting in a reduced amount of topographic uplift for a given plume size. Comparing the observed amount of uplift to the estimated size and depth of the low-density plume provides evidence for this behavior on Earth but not on Venus. On Venus, plumes strike the lithosphere directly, thus causing more uplift for a given plume size.

The relationship between the gravity and the topography provide some insight into interior structure and convection. The magnitude of variations in the gravity field as compared to a given topographic feature is an indication of the interior structure that supports a given topographic feature. The strength of the lithosphere can support topography. Variations in density in the interior can also support topography. A mountain can be supported by a thick “root” of low-density crust, analogous to an iceberg floating in denser water. Variations in the mantle temperature associated with convection can also support topography. The gravity field of Venus has been carefully studied to estimate the thickness of the strong, or elastic, part of the lithosphere, the thickness of the crust, and the location of low-density, relatively hot regions in the mantle. Clearly, some highlands, such as tessera plateaus (see Section 7), are compensated by crustal roots. Many other highlands appear to be compensated by mantle plumes.

In addition to plumes, conduction through the lithosphere must contribute to the heat loss on Venus. The thinner the lithosphere, the more rapidly the planet loses heat. Estimates of the thickness of the lithosphere on Venus, derived from gravity and topography, are typically 100 to 200+ km. This is comparable to the lithospheric thickness on Earth, and is too large to account for the majority of Venus’ heat loss. There is growing evidence that the recycling of the lower lithosphere back into the mantle may help cool Venus, just as subduction helps cool the Earth. New models for corona formation show that at least some coronae may form above sites where the thickening, cold lithosphere becomes too dense and breaks off into the
mantle. Estimates of lithospheric thickness variations also suggest that the lower lithosphere may thicken and become unstable locally. Although possibly important, such a process is not going to be nearly as efficient a cooling mechanism as subduction.

Volcanism, resulting from melting of the mantle and/or lithosphere and the rise of hot magma, can contribute to heat loss. As discussed above, Venus was completely resurfaced, most likely by volcanism within the last billion years. Present-day rates of volcanism are not constrained, but are unlikely to be sufficient to be a major contributor to heat loss.

Another constraint on interior processes is the absence (or very low level) of a magnetic field. The Mariner flyby missions measured no magnetic field, indicating that, if present, the field must be <500 nT at the surface. More recent data from Venus Express suggest a limit of 10 nT on the strength of the magnetic field at the surface. Most models of interior dynamos indicate that a planet must be losing large amounts of heat from the planet’s metal core to provide enough energy for a dynamo. Some models have suggested that relatively rapid heat loss through plate tectonics is a good method of driving a dynamo. Thus, one possible scenario is that Venus had early plate tectonics and an active dynamo, but eventually lost much of its water from the crust through volcanism to the atmosphere, where it was subsequently lost to space. This decrease in water increased the strength of the lithosphere to the point that tectonics ceased and the dynamo shut down. Heat is then lost primarily by conduction through the lithosphere, causing the mantle to heat up and increase the rate of volcanism, causing the planet to resurface. This idea is

![Four views of Venus, with centers at 0°, 90°, 180°, and 270° E. Topography is in color, with Magellan radar images overlain on top.](image)
speculative, as there is no direct evidence for an early plate tectonic period. The possibility of a low-level magnetic field provides a new challenge for understanding dynamo processes and could indicate that the core of Venus is losing heat but at a lower rate than that of Earth.

The unusual cratering record on Venus indicates that the first 3.5 billion years of geologic history has somehow erased, possibly with a lower rate of resurfacing occurring subsequently. In contrast, Mars, Mercury, and the Moon have surfaces that preserve the large impact basins from early bombardment and reflect a gradual loss of heat and decline in geologic activity. Some models have proposed that resurfacing on Venus occurs episodically. In one scenario, the lithosphere thickens and becomes denser due to both cooling and chemical phase transitions. The lithosphere is predicted to founder, or get mixed into the mantle, when it becomes gravitationally unstable. However, how the lithosphere actually breaks and initiates this process is unclear. In another scenario, the stagnant lid heat insulates the mantle, causing it to heat up to the point that widespread melting occurs, eventually erupting on the surface. Other models show that volcanism that is globally distributed and resurfaces small regions in each event can produce the observed distribution. High mantle temperatures could facilitate this kind of widespread volcanism.

6. COMPOSITION

6.1. Global Implications

The similarity between Venus and Earth in terms of size and location in the solar system indicates that their bulk compositions should be comparable. The exact composition of the crust is related to the composition and temperature in the interior of the planet when the rock melts, as well how much of the original rock is melted. The typical rock type that forms on Earth when the interior melts and erupts is basalt. Thus, it is not surprising that geochemical measurements on the surface of Venus have a gross composition similar to terrestrial basalts, with some variation. On Earth, basalts make up the majority of the oceanic crust and are found in volcanic regions of continents. When processes such as subduction remelt basalts the resulting rocks are enriched in silica (SiO₂). Continental rocks are a result of billions of years of remelting of a basaltic crust driven by convective and plate tectonic processes. They are of lower density than basalt due to the enrichment of silica relative to iron and magnesium. The presence of at least small amounts of water may be essential to the formation of such silica-rich rocks. Continents stand higher than the oceanic crust due to both their lower density and the greater thickness of continental crust. As we will discuss, there is possible evidence for silica-rich rock on Venus.

The abundances of primary mineral-forming and radiogenic elements were measured by spectrometers on Venera landers. Venera landers 8, 9, and 10 and Vega landers 1 and 2 measured the amounts of uranium (U), thorium (Th), and potassium (K) using a gamma-ray spectrometer (Table 15.1). The Venera landers 13 and 14 measured these elements as well as the major-element forming minerals (see Table 15.2). Due to the orbital dynamics of delivering probes to the surface of Venus in any given time period, Venera landers 8–14 are all located in a relatively small region on Venus within 27°–33° E and 15° S to 30° N. This area includes the eastern flank of Beta Regio, a major hotspot, and the plains to the east of Beta and Pheobe Regiones. The Vega 1 and 2 landers, sent at an earlier time, are located near 170° E, 10° N, and 180° E, 10° S, to the west of Atla Regio.

The silica content and the relative abundances of iron and magnesium for rocks at the Venera lander sites

<table>
<thead>
<tr>
<th>Lander</th>
<th>U (ppm)</th>
<th>Th (ppm)</th>
<th>K (wt%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venera 8</td>
<td>2.2 ± 0.7</td>
<td>6.5 ± 0.2</td>
<td>4.0 ± 1.2</td>
</tr>
<tr>
<td>Venera 9</td>
<td>0.6 ± 0.2</td>
<td>3.6 ± 0.4</td>
<td>0.5 ± 0.1</td>
</tr>
<tr>
<td>Venera 10</td>
<td>0.5 ± 0.3</td>
<td>0.7 ± 0.3</td>
<td>0.3 ± 0.2</td>
</tr>
<tr>
<td>Vega 1</td>
<td>0.68 ± 0.47</td>
<td>1.5 ± 1.2</td>
<td>0.5 ± 0.3</td>
</tr>
<tr>
<td>Vega 2</td>
<td>0.68 ± 0.38</td>
<td>2.0 ± 1.0</td>
<td>0.4 ± 0.2</td>
</tr>
</tbody>
</table>
rock experiences changes in pressure and/or temperature. Through chemical weathering or metamorphism when the example, the amount of Al, Ti, Ca, or Si may change Earth’s. Some variation may occur after the rock forms. For subsequent analyses have discounted this idea and uncertainties in the measurements mean that numerous variations such as active lava flows. Making the assumption that lava flows on Venus have a similar size and effusion rate as typical terrestrial flows, a thermal signature that could be resolved at a scale of 100 km is likely to last only hours to days. VIRTIS data show no definitive evidence of such transient anomalies.

Thermal emissivity anomalies that are significantly lower and higher than the average are found to correlate with tessera terrains (see Figure 15.8). Low emissivity is consistent with a high silica composition, such as granite. There are two significant uncertainties in this interpretation. One is the uncertainty in the Magellan altimetry measurement in rugged areas, like the tesserae (see below). Another is the nonuniqueness of the interpretation of a single spectral band. However, if this interpretation is correct, it is extremely significant. On Earth, granite forms when basalt melts in the presence of water, such as when subducted plates melt at high temperatures. As discussed further below, individual tessera plateaus are up to ~2000 km across, indicating significant water would have been present during their formation if they are granitic in composition.

The highest emissivity anomalies are correlated with volcanic flows at hotspots, where gravity data indicate a plume at depth. These anomalies are interpreted to indicate relatively recent flows. The contrast between the average emissivity of most of the southern hemisphere and the high emissivity flows is consistent with the signatures of minerals expected to form when basalt is exposed to Venus surface conditions versus fresh basalt (see Weathering below). Although rates of weathering are not well constrained, the lack of weathering indicates that the flows formed in geologically recent times (<2 My) and may even be currently active. Although a difference in the iron content of the initial magma eruption could also produce an emissivity variation, higher initial iron contents would also

<table>
<thead>
<tr>
<th>Constituent</th>
<th>Venera 13 (wt%)</th>
<th>Venera 14 (wt%)</th>
<th>Vega 2 (wt%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SiO₂</td>
<td>45.1 ± 3.0</td>
<td>48.7 ± 3.6</td>
<td>45.6 ± 3.2</td>
</tr>
<tr>
<td>TiO₂</td>
<td>1.59 ± 0.45</td>
<td>1.25 ± 0.41</td>
<td>0.2 ± 10.1</td>
</tr>
<tr>
<td>Al₂O₃</td>
<td>15.8 ± 3.0</td>
<td>17.9 ± 2.6</td>
<td>16.0 ± 1.8</td>
</tr>
<tr>
<td>FeO</td>
<td>9.3 ± 2.2</td>
<td>8.8 ± 1.8</td>
<td>7.74 ± 1.1</td>
</tr>
<tr>
<td>MnO</td>
<td>0.2 ± 0.1</td>
<td>0.16 ± 0.08</td>
<td>0.14 ± 0.12</td>
</tr>
<tr>
<td>MgO</td>
<td>11.4 ± 6.2</td>
<td>8.1 ± 3.3</td>
<td>11.5 ± 3.7</td>
</tr>
<tr>
<td>CaO</td>
<td>7.1 ± 0.96</td>
<td>10.3 ± 1.2</td>
<td>7/5 ± 0.7</td>
</tr>
<tr>
<td>K₂O</td>
<td>4.0 ± 0.63</td>
<td>0.2 ± 0.07</td>
<td>0.1 ± 0.08</td>
</tr>
<tr>
<td>S</td>
<td>0.65 ± 0.4</td>
<td>0.35 ± 0.31</td>
<td>1.9 ± 0.6</td>
</tr>
<tr>
<td>Cl</td>
<td>&lt;0.3</td>
<td>&lt;0.4</td>
<td>&lt;0.3</td>
</tr>
</tbody>
</table>
weather to minerals with a lower emissivity, indicating that the flows are likely young in that case as well.

In addition to direct measurements of the composition, morphology can be used as a very crude indication of composition. For example, lavas with a basaltic composition tend to be very fluid, forming long, narrow flows, and broad, low volcanoes. As the silica content increases, the viscosity of the lava increases. The thickness of flows increases, their length decreases, and the slopes of volcanoes formed increases. Terrestrial examples are Mauna Loa in Hawaii (basaltic) and Mt. St. Helens in Washington (more silica rich). On Venus, the morphology of flows is generally consistent with low-viscosity basaltic compositions. There are some features that appear to represent much thicker, shorter flows (see description of “pancakes” in Section 6 below). However, these morphologies cannot be considered diagnostic of composition as factors such as the volume and rate of material erupting, the atmospheric pressure during eruption, and the amount of gas in the lava also shape the morphology of the flow.

6.2. Surface Weathering

Although there is little evidence of weathering of the surface by wind consistent with Venus’ dense atmosphere, the environment for chemical weathering is extremely harsh. In addition to the searing temperature and high pressure, the atmosphere contains highly corrosive and chemically active gases such as SO₂ (sulfur dioxide), CO, OCS, HCl (hydrochloric acid), and CO₂. A variety of minerals form in laboratory experiments that simulate Venus conditions, such as calcite, dolomite, anhydrite, and hematite, but no landers have measured actual minerals. Measurement of the specific minerals present and their abundances is highly desirable as they provide insight into the nature of the chemical interaction between the surface and the

FIGURE 15.8  Surface emissivity at 1.02 μm derived from VIRTIS surface brightness. This map is derived in the same manner as that in Mueller et al. (2008) but using an improved topographic corrections. An average surface emissivity of 0.62 is used to provide a physical range of values. Contours show the number of VIRTIS images acquired in each area. The data are reduced using first-order corrections for stray light, clouds, and topography based on data statistics. A greater number of looks therefore mean not only a reduction of instrumental noise but also a lower likelihood of deviations due to variable observing effects. White areas are data gaps.
atmosphere. This information is a critical piece of understanding the larger problem of how Venus arrived at the hellish climate that now exists.

One of the key questions is how much CO₂ is trapped in minerals on the surface of Venus. Most of the CO₂ found on Earth is trapped as carbonates via biological processes, specifically the formation and accumulation of seashells (although carbonates would precipitate out of sea water even without biological processes). This process is an important element of the overall balance that makes Earth habitable. Available information from surface composition and laboratory experiments suggests that significant amounts of carbonates could be present on the surface of Venus, perhaps up to 10%. If so, this would mean that CO₂ in surface rocks is an important part of determining the atmospheric pressure and composition. Another key question is how atmospheric SO₂ interacts with the surface. On Earth, most of the SO₂ is dissolved in the oceans. Rates of chemical reactions involving SO₂ are known for the conditions in the atmosphere of Venus and predict that the SO₂ in the present-day sulfuric acid clouds on Venus should disappear over time. Note this reaction requires abundant surface calcite, which may or may not be present. If this reaction is occurring, atmospheric SO₂ is not in equilibrium and must be resupplied. The fact that sulfuric acid clouds are present today implies that new sulfur gases have been added to the atmosphere. Both the Pioneer Venus Orbiter and Venus Express have observed significant changes in the concentration of SO₂ in the atmosphere over the timescale of years. One possible explanation is the release of SO₂ due to recent volcanism. Atmospheric dynamics could also be responsible. Thus, determination of surface mineralogy is key to understanding the surface–atmosphere interaction and any associated contribution to atmospheric stability and climate change.

7. VOLCANISM

With the exception of Jupiter’s moon Io, Venus is the most volcanic world in the solar system. Volcanic features of a broad range in morphology cover the surface, from sheetlike expanses of lava flows to volcanoes shaped like pancakes and ticks, as illustrated below. The high surface temperature and pressure on Venus make explosive volcanism less likely, although some possible deposits produced by explosive volcanism have been mapped. Magellan data illustrated that volcanic features do not occur in chains or specific patterns, indicating the lack of plate tectonics on Venus.

The plains or low-lying regions on Venus are covered by sheet and digitate deposits that are interpreted to be volcanic in origin (Figure 15.9). These extensive deposits are likely to be flood basalts, formed in similar ways to the Columbia River Basalts or the Deccan Traps on Earth. In some plains regions, the surface is clearly built up of multiple, superposed lava flow deposits, while other regions are more featureless. Lava flows have varying brightness in the Magellan SAR images. Most lava flows are of intermediate brightness. Comparisons to radar images of lava flows in Hawaii indicate that the Venusian flows have similar roughness, although some flows on Venus are unusually smooth.

Large volcanoes on Venus (those with diameters >100 km) are found at topographic rises, along rift zones,
and concentrated in the region bounded by Beta Regio, Atla Regio, and Themis Regiones. Over 100 large volcanoes have been identified. Large volcanoes have average heights of about 1.5 km and aprons of lava flows that extend hundreds of kilometers from their summits. Maat Mons, the largest volcano on Venus is about 8.5 km high and 400 km across (Figure 15.11). In comparison, Mauna Loa, the largest volcano on Earth, is about 9 km high and 100 km across. Detailed studies of individual large volcanoes have revealed their complex histories. Many volcanoes show evidence of multiple eruptions from their summits as well as sites on their flanks. Some large volcanoes have calderas at their summits similar to volcanoes on Earth and Mars, formed by collapse of the underlying magma chamber. Others have radially fractured summits, with the radial fractures interpreted as the surface expression of subsurface dikes. These dike sets provide evidence that many large volcanoes have undergone multiple episodes of intrusion and extrusion.

At the smaller end of the scale, volcanoes 5–50 km across are also abundant on the surface of Venus (Figure 15.12). Many of the volcanoes resemble their terrestrial counterparts, with summit calderas and radiating digitate flows. Venus also has several types of volcanic features that differ from those on Earth and other planets. Steep-sided or pancake domes are flat-topped, steep-sided features (Figure 15.13), similar to flat-topped domes like the Inyo domes in California that are formed by silicic lavas. The Venus domes may have a different composition, however, as they are much larger, and have smooth rather
than blocky surfaces in comparison to the terrestrial domes. Other unusual volcanoes on Venus resemble ticks, or bottle caps. These small domes have scalloped margins, and are interpreted to be steep-sided domes whose margins have collapsed.

The Magellan radar also imaged channels, a few kilometers wide and hundreds of kilometers long. The channels are found in many places within the plains, tend to be very sinuous, and in places show evidence of levees and flow breakouts. The channels have formed by lava of some unusual composition, so fluid that it behaved like water and is able to flow long distances without cooling. A number of compositions have been proposed, including carbonate- or sulfur-rich lavas and ultramafic silicate melts. Others have suggested that the channels were formed by erosion of the surface by lava, similar to lunar rilles on the Moon (See The Moon). Some of the channels extend for long distances allowing them to be used as a time marker, as it can be assumed that the channel formed over a relatively short period of time. For example, the channel may superpose one feature, but be overlain or cut by another. Also, a few channels now trend uphill, indicating that the surface deformed after they formed.

Surface thermal emissivity data create a new way to monitor volcanic activity. Gravity studies indicate that plumes exist under ~10 hotspot regions, suggesting that the associated volcanoes could be active. However, a thermal signature can take on the order of 100 My to conduct through the lithosphere. Interpreting the emissivity data for four hotspot regions as indicating unweathered basalts suggests that these volcanoes were active recently, perhaps up to the present day. Venus Express mapped only the southern hemisphere, but all the hotspots observed showed high emissivity anomalies associated with flows. The implication is that hotspots in the northern hemisphere are also active given the similarity in their gravity signatures.

The resolution of the thermal emissivity data is on the order of 100 km. Thus, it is entirely possible that smaller, unresolved volcanic features are also active. However, if volcanism today is confined to hotspots, and plains volcanism is inactive, this could have very important implications for interior dynamics. For example, it is possible that prolonged melting of the upper mantle due to high mantle temperatures induced by an insulating stagnant lid could have dried out the upper portion of the mantle. Thus, melting might occur more easily in hotspot settings where new material is brought up from depth in the mantle, localizing new melting.

Unfortunately, currently available data preclude a strong constraint on the rate of volcanism. Thus, it is not possible to resolve the resurfacing debate. However, the size of the observed areas of recent volcanism is consistent with ongoing resurfacing.

8. TECTONICS

For the larger terrestrial planets, Venus, Earth, and Mars, mantle convection is the primary driving force for tectonic processes. On Mars, most tectonic structures are associated with either the gigantic Tharsis rise or the global dichotomy (See Mars Surface and Interior). The global dichotomy divides the smoother northern lowlands from the heavily cratered southern highlands. On Earth, plate tectonics is clearly dominant. Tectonic features on Venus are highly variable and enigmatic. Tessera terrains are unique to Venus, and are defined as having multiple intersecting deformational structures with different directions. One possible factor in creating these highly deformed regions is that Venus experiences very little surface erosion. In contrast, most continental regions have experienced multiple episodes of deformation but surface structures are often eroded between events, leaving evidence of only the most recent occurrence. Many of the tectonic features on Venus are continuous for thousands of kilometers, and likely reflect underlying mantle processes including upwelling, downwelling, and horizontal flow. Tectonic deformation takes many forms and is distributed across the surface of Venus, in contrast to the concentration of deformation along plate boundaries on Earth. Below we describe the characteristics and likely origins of the key types of tectonic features on Venus.
8.1. Tessera and Crustal Plateaus

Tessera terrains are highly deformed and thus stand out as very bright in radar images (Figure 15.14). They are made up of both extensional and compressional deformational features. In some cases the sequence of events can be determined, but more often it is ambiguous. Tesserae occur both as isolated fragments embayed by later plains material and in major plateaus. There are six major crustal plateaus: Alpha, Ovda, Pheobe, Thetis, and Tellus Regiones plus Ishtar Terra. Figure 15.15 shows Alpha Regio, one of the smaller highland plateaus. Western Ovda Regio may be a relaxed crustal plateau. These plateaus are 1000–3000 km in diameter and 0.5–4 km higher than the surrounding plains. Their gravity signature indicates that they are supported by crustal roots rather than active mantle processes, thus the name “crustal” plateaus.

Ishtar Terra is unique among the highland plateaus. It is the largest of the crustal plateaus, and is surrounded by significant mountain belts on three sides, with large areas of tesserae occurring on their exterior flanks. They are Venus’ only real mountain belts. Lakshmi Planum makes up the interior of Ishtar Terra. This smooth plateau is elevated 3–4 km above the surrounding plains, and is covered by volcanic flows that emanated primarily from two large calderas. The Maxwell Montes to the east of Lakshmi Planum contain the highest point on Venus, at approximately 11 km above the mean planetary radius (see Figures 15.1 and 15.6). Although other crustal plateaus tend to have relatively flat interiors and rims of higher topography, no other crustal plateau is as extreme as Ishtar Terra in terms of its diameter, elevation, undeformed interior volcanic plains, and circumferential deformation features.

Crustal plateaus have been proposed to form over mantle upwellings and over mantle downwellings. In the mantle upwelling scenario, a plume creates a crustal plateau through decompression melting above the plume head, analogously to plateaus formed on the terrestrial seafloor. Deformation occurs as the topography viscously relaxes. The alternative model forms the plateaus above a cold, sinking mantle downwelling. On Earth, both subduction zones and local sites of downwelling form below cold mountain roots. Venusian crustal plateaus are proposed to form as a downwelling causes sinking of the lower lithosphere and accumulation and compression of the crust at the surface. The mechanism for forming small, local regions of tesserae is not clear. In many cases these regions are embayed and thus appear to be old and possibly inactive. There are few clues as to original processes that cause deformation. One possibility is that these areas represent sections of tessera plateaus that were once elevated but have topographically relaxed. If plateaus formed in an earlier, hotter time period, relaxation may have proceeded more rapidly, allowing for complete relaxation of plateaus.
The semicircular rim of Western Ovda Terra could be the remnant of a relaxed plateau. Alternatively, small tessera terrains may be a result of an entirely different type of tectonic event, such as ridge belt formation.

A key question for tesserae is their composition. Some researchers have suggested that crustal plateaus may be analogous to terrestrial continents due to their size, their compensation by crustal roots, their highly deformed surfaces, and their local embayment by the surrounding plains. Tesserae are relatively old, and could be made of more silica-rich crust. As discussed above, the low thermal emissivity anomalies observed by both Venus Express and the Galileo Orbiter flyby support this interpretation. If tesserae are silica-rich, perhaps granitic rocks like Earth’s continents, it implies that they are remnants of an earlier, wetter time period on Venus.

8.2. Chasmata and Fracture Belts

**Chasmata** (chasma means canyon) are regions of extensional deformation, as indicated by their locally low topography and graben or grabenlike morphology. There are five major chasmata on Venus that extend for thousands of kilometers and are several kilometers deep: Parga, Hecate (see Figure 15.16), Dali/Diana, Devana, and Ganis Chasmata. The fracture zones in these regions are typically ~200 km wide, with topographic troughs that are generally narrower, with widths of ~50–80 km. There are seven smaller chasmata, with lengths of hundreds of kilometers and proportionately narrower fracture belts and troughs. Several of the chasmata occur on the flanks of hotspot rises and may be a result of topographic uplift above a plume. The majority of other chasmata form synchronously with coronae, as discussed above. Although chasmata are not required for coronae to form, or vise-versa, it is clear that the presence of one increases the likelihood of the other. Both extension and upwelling plumes can thin the lithosphere, which may focus additional extension and upwelling in an area. Chasmata are analogous to terrestrial continental rift zones, which have relatively small amounts of crustal extension, in the range of several percent.

Fracture belts appear similar to minor chasmata, but are less intensely fractured, implying less extension. A curious feature of fracture belts is that they are topographically broad swells rather than topographic lows. The positive relief suggests that they went through a compressional stage, and that the fractures may be due to topographic uplift rather than regional extension.

8.3. Coronae

Coronae are large (>100 km across) circular features surrounded by concentric ridges and fractures (Figure 15.17).
Over 500 coronae have been identified on Venus; the largest one is Artemis Corona at 2500 km across. Coronae often have volcanoes in their interiors and many are surrounded by extensive lava flows. Coronae tend to be raised at least 1 km above the surrounding plains, but others are depressions, rimmed depressions, or rimmed plateaus. Most coronae are located along rift or chasmata systems, although some are at topographic rises and others occur in the plains away from other features. Coronae are thought to form over thermal plumes or rising hot blobs, smaller in scale and probably rising from shallower depths than the plumes that form topographic rises. The wide range in corona topographic shapes indicates that corona evolution also involves delamination or sinking of lithospheric material in its later stages. Studies of the gravity signatures of some large coronae indicate that many coronae are likely to be isostatically compensated, and thus probably inactive. The fact that we do not see coronae on Earth may be due to the lack of an asthenosphere on Venus.

8.4. Ridge Belts and Wrinkle Ridges

Ridge belts occur in a variety of morphologies and are distributed around the planet. Based on the morphology of individual fractures and the long, narrow topographic highs that comprise individual ridges, ridge belts are interpreted to be a result of compressional stresses (see Figure 15.18). Individual ridges are typically less than 0.5 km in height, 10–20 km wide, and 100–200 km long with a spacing of ~25 km. The two largest concentrations of ridge belts occur in Atalanta/Vinmara Planitiae and Lavinia Planitia. The belts in Atalanta/Vinmara Planitiae are roughly an order of magnitude larger than those elsewhere. Belts in Lavinia are unusual in that they have extensional fractures roughly parallel to compressional features within the same belt, possibly due to topographic uplift along the ridge. Larger belts are believed to result from mantle downwelling, similar to the proposed downwelling origin for crustal plateaus, but with lower strain. Smaller belts may be associated with more local-scale tectonics.

Wrinkle ridges are extremely common features on Venus and are also interpreted as simple compressional folds and/or faults but are much narrower (~1 km or less in width) than ridges. They have positive relief, based on the fact that lava flows can be seen to pond against some wrinkle ridges, but that relief is too small to be seen in Magellan altimetry. Most ridges occur in evenly spaced set, 20–40 km apart. These sets of wrinkle ridges can be local in nature, associated with a corona, for example, but more commonly cover thousands of kilometers. These larger sets are likely to be gravitational spreading of high topography into lower regions and can be seen to form rings around some large topographic features (see Figure 15.4). Other sets cannot be clearly associated with topographic highs.

One hypothesis is that these features result from thermal contraction due to climate change-driven atmospheric temperature changes. In some regions there are two sets of wrinkle ridges, although one set is usually better developed.

8.5. Plains Fractures, Grids, and Polygons

A wide range of long, narrow, approximately straight fractures occur in the plains. Some fractures are wide enough to be resolved as graben, but most are too narrow (less than 0.5 km) to be resolved as more than fractures. Most are interpreted as extensional fractures because they parallel resolvable graben and because of their shape. Some are clearly associated with local features such as volcanoes or corona, and are probably due to extension above dikes. In some locations there are either single sets or intersecting grids of fractures that cover hundreds of kilometers (Figure 15.19). They are very regularly spaced, with separations of 1–2.5 km. The narrow spacing suggests that a thin layer is involved in the deformation. It is not obvious how a uniform stress can be transmitted to such a thin layer over such a broad region. Shear deformation is required to produce grids of intersecting lineations.

Another type of extensional feature observed on Venus is polygons, which are found in over 200 locations on Venus.
These features are analogous to mud cracks in that they form in a uniform, extensional stress field. However, they form not as water is lost but instead when rock cools and contracts. The typical diameter is \( \approx 2 \) km, but some are up to 25 km across. Some areas have multiple scales of deformation. Again, some of these features can be associated with local events such as volcanoes, but others cover very broad regions and do not have an obvious origin. Polygons are most commonly associated with small volcanic edifices, and frequently appear to form synchronously (Figure 15.10). Some may form by actual cooling of lava flows. Such basaltic columns are common on Earth, but the scales of the features found on Venus are orders of magnitude larger, implying that the flow thickness on Venus would probably be too large to be plausible. Another mechanism, as proposed for wrinkle ridges, is the possible heating and cooling of the upper crust due to climate change.

9. SUMMARY

Venus provides a unique window into the evolution of terrestrial planets. It is essentially identical to Earth in size and bulk composition, yet its geologic history is entirely different. Venus’ level of geologic activity over the last billion years is comparable to that of Earth and exhibits many of the same geologic processes. The convecting interior drives geologic activity at the surface, creating a dozen major highlands. These highlands include “hot-spots”, which form above mantle plumes, and the more enigmatic and intensely deformed highland plateaus. Venus’ hotspots appear to be sources of recent or even active volcanism today, providing clues about the interior, surface, and atmospheric processes. The majority of the surface is composed of vast volcanic plains along with nearly ubiquitous tectonic features. There are tens of thousands of volcanic features from small-scale (hundreds of meters) flows, vents, and shields, to hundreds of large-scale (>100 km) shield volcanoes that blanket the surface. The pervasive volcanism may have buried the earliest, heavily cratered surfaces, or they may have been destroyed through tectonic processes. Tectonic features range in scale from pervasive linear fractures and polygons at the limit of resolution to highland plateaus composed of tessera terrain 1000—2000 km in diameter.

Despite the similarities between Venus and Earth, Earth is the only body in our solar system that developed the system of plate tectonics that has so shaped the geologic and environmental evolution of our planet. The atmosphere of Venus lost nearly all its water early in its evolution. The loss appears to have affected the interior as well, causing the lithosphere to be too strong to break into the plates observed on Earth, and the asthenosphere to be too strong to facilitate rapid horizontal plate motion. This same loss of water has contributed to the dominance of CO\(_2\) in the atmosphere and the resulting greenhouse effect that created the scorching surface conditions. Why Venus lost its water is not understood, but as with Mars, the absence of a strong magnetic field exposes the atmosphere to erosion by solar wind. In turn a planet must be losing heat rapidly enough to drive the formation of a magnetic dynamo. The interior volatile content affects the processes through which planets lose heat, and appears to be the key to whether or not plate tectonics develops. Was Venus originally on the same evolutionary path as Earth? What was the pivotal event or process that sent Venus down an alternate path to the hellish, uninhabitable planet we observe today? We can begin to address these questions, thus better understanding the evolution of our own planet, through future missions to understand the coupled evolution of the atmosphere, surface, and interior.

BIBLIOGRAPHY


Books


Journal Articles Special Issues


Web Sites

Venus data from US missions are available through the Planetary Data System at pds.nasa.gov and from ESA missions at www.sciops.esa.int.
A fundamental question about the surface of Mars is whether it was ever conducive to life in the past, which is related to the broader questions of how the planet’s atmosphere evolved over time and whether past climates supported widespread liquid water. Taken together, geochemical data and models support the view that much of the original atmospheric inventory was lost to space prior to about 3.7 billion years ago. Before and around this time, the erosion of valley networks by liquid water suggests a past climate that was warmer. But exactly how the early atmosphere produced warmer conditions and the extent to which it did so remain open questions. Suggestions include an ancient greenhouse effect enhanced by various gases, impacts that created temporary wet climates by turning ice to vapor and rainfall, and periodic melting of ice under moderately thicker atmospheres as Mars’ orbit and axial tilt changed. For the last 3.7 billion years, it is likely that Mars has been predominantly cold and dry so that outflow channels that appeared later were probably formed by fluid release mechanisms that did not depend on a warm climate. Very recent gullies and narrow, summertime dark lineae that form on steep slopes are features that form in the current cold climate. In addition, wind erosion, dust transport, and dust deposition have been modulated by changes in Mars’ orbital elements over time, which complicates the interpretation of climate and volatile history. In the past, surface modification by winds in a denser atmosphere may have been significant.

1. INTRODUCTION

The most interesting and controversial questions about Mars revolve around the history of liquid water. Because temperatures are low, the current, thin Martian atmosphere only contains trace amounts of water as vapor or ice clouds. In larger quantities, water is present as ice and hydrated minerals near the surface. Some geological structures resemble dust-covered glaciers or rock glaciers, while others strongly suggest the flow of liquid water relatively recently as well as in the distant past. But the present climate does not favor liquid water near the surface. Surface temperatures range from about 140 to 310 K. Temperatures above freezing occur only under highly desiccating conditions in a thin layer at the interface between the soil and atmosphere. Also, the surface air pressure over much of the planet is below the triple point of water (611 Pa or 6.11 mbar); under these conditions, at temperatures above freezing, liquid water would boil away. If liquid water is present near the surface of Mars today, it is confined to thin adsorbed layers on soil particles or highly saline solutions. No standing or flowing...
liquid water, saline or otherwise, has been unambiguously proven.

Conditions appear to have been more favorable for liquid water in the ancient past. The landscape has a number of fluvial (stream-related) features, of which the most important for climate are the valley networks, which are dried-up riverlike depressions fed by treelike branches of tributaries. Deltas exist at the end of a small fraction of valleys. Fluvial features that occurred later than the valleys are giant outflow channels. The valley networks indicate wetter past climates, while the outflow channels are commonly interpreted as massive release of liquid water from subsurface aquifers or the melting of underground ice (although a minority opinion argues in favor of runny lavas as the primary erosive agent). In addition to fluvial features, the soil and sedimentary rocks incorporate hydrous (water-containing) minerals that are interpreted to have formed in the presence of liquid water. The extent and timing of the presence of liquid water are central to the question of whether microbial life ever arose and evolved on Mars.

Atmospheric volatiles are substances that tend to form gases or vapors at the temperature of a planet’s surface and so could have influenced the past climate and the occurrence of liquid water. Here we review the current understanding of volatile reservoirs, the sources and sinks of volatiles, the current climate, and evidence for different climates in the past. We consider the hypothesis that there have been one or more extended warm and wet climate regimes in the past, the problems with that hypothesis, and the alternative possibility that Mars has had a cold, dry climate similar to the present climate over nearly all of its history, while still allowing for some fluid flow features to occur on the surface. The possible relevance of very large orbital variations (Milankovitch cycles) for Mars’ climate history is also examined.

Whether or not extended periods of warm, wet climates have occurred in the past, wind is certainly an active agent of surface modification at present and has probably been even more important in the past. Consequently, we also discuss evidence for how the surface has been changed by wind erosion, burial, and exhumation, and the resulting complications for interpreting Mars’ surface history. We conclude with a brief overview of open questions.

2. VOLATILE INVENTORIES AND THEIR HISTORY

2.1. Volatile Abundances

Mars’ thin atmosphere is dominated by carbon dioxide (CO₂), and in addition to the major gaseous components listed in Table 16.1, the atmosphere contains a variable amount of water vapor (H₂O) up to 0.1%, minor concentrations of photochemical products of carbon dioxide and water vapor (e.g. CO, O₂, H₂O₂, and O₃), and trace amounts of the noble gases neon (Ne), argon (Ar), krypton (Kr), and xenon (Xe). Methane (CH₄), averaging about 10 parts per billion by volume (ppbv), has been reported based on spectra from ground-based telescopes (which are complicated by having to remove the effect of viewing Mars through the Earth’s atmosphere) and relatively low resolution spectra obtained by European Space Agency’s (ESA’s) Mars Express orbiter. However, in situ measurements by National Aeronautics and Space Administration’s (NASA’s) Mars Science Lab rover have found no methane with an upper limit of about 1 ppbv, which must be taken as more definitive.

Volatiles that can play important roles in climate are also stored in the regolith and near-surface sediments. The regolith is a geologic unit that includes fine dust, sand, and rocky fragments comprising the Martian soil together with loose rocks, but excluding bedrock. Approximate estimates of the inventories of water, carbon dioxide, and sulfur are given in Table 16.2.

Water is stored as ice in the permanent north polar cap and its surrounding layered terrains, in layered terrains around the South Pole, and as ice, hydrated minerals, or adsorbed water in the regolith. The 5-km-deep residual northern polar cap consists of a mixture of ≥95% water ice and fine soil or dust, while layered south polar terrains contain water ice and about 15% dust. Taking account of their volumes and ice fractions, each cap and associated layered terrain contains water ice equivalent to a global ocean about 10 m deep.

Measurements of the energy of neutrons emanating from Mars into space by NASA’s Mars Odyssey orbiter has also provided evidence for abundant water ice, adsorbed water, and/or hydrated minerals in the upper 1–2 m of regolith at high latitudes and in some low-latitude regions (Figure 16.1). Cosmic rays enter the surface of Mars and cause neutrons to be ejected with a variety of energies depending on the elements in the subsurface and their

<table>
<thead>
<tr>
<th>TABLE 16.1 Basic Properties of the Present Atmosphere</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average surface pressure</td>
</tr>
<tr>
<td>Surface temperature</td>
</tr>
<tr>
<td>Major gases</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Significant atmospheric isotopic ratios relative to the terrestrial values</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
distribution. Abundant hydrogen serves as a proxy for water and/or hydrated minerals. In 2008, the robotic arm and thrusters on NASA’s Phoenix Lander exposed ice at 68°N some 5–10 cm below the surface, verifying the inferences from Mars Odyssey’s neutron measurements. However, water ice probably extends to no more than ~20–30 m depth in the mid- to high-latitude regolith based on radar and the morphology of small craters. Consequently, the total water inventory appears to be dominated by hydrated minerals rather than ice and has a depth of 200–1000 m of a global equivalent ocean.

The inventory of CO₂ mainly depends on how much is locked up in carbonates hidden in the subsurface. Weathering of dust has occurred over billions of years even in the prevailing cold dry climate, and as a consequence some CO₂ appears to have been irreversibly transferred from the atmosphere to carbonate minerals in dust particles. The total amount depends on the global average depth of dust. Some CO₂ is likely to be adsorbed in the soil also, but the quantity is limited by competition for adsorption sites with water. From orbital spectra, some carbonate sedimentary rock outcrops have been identified but with an area of only 10⁵ km² assuming a subsurface extent underneath associated geologic units. Taking an average thickness of 100 m, the CO₂ inventory in these outcrops is only 3 mbar. However, the carbonate inventory in the subsurface remains unknown. Also, carbonate outcrops that are smaller than orbital resolution can detect are likely present. Indeed, instruments on NASA’s Spirit Rover identified a small carbonate outcrop in the Columbia Hills region of Gusev Crater.

Table 16.2 also lists sulfates. Although there are presently no detectable sulfur-containing gases in the atmosphere, sulfur gases should have existed in the atmosphere in the past when Mars was volcanically active. Measurements by NASA’s landers and rovers show that sulfur is a substantial component in soil dust (~7–8% by mass) and surface rocks. Hydrated sulfate salt deposits have been identified in numerous layered deposits from near-infrared spectral data collected by Mars Express and NASA’s Mars Reconnaissance Orbiter (MRO). About two-thirds of these deposits are within 10° latitude of the equator. Several sulfate minerals have been detected and the total sulfur abundance can be estimated. Notable sulfate minerals include kieserite (MgSO₄·H₂O) and gypsum (CaSO₄·2H₂O). Jarosite (XFe₃(SO₄)₂(OH)₆, where “X” is a singly charged species such as Na⁺, K⁺, or hydronium (H₃O⁺)) has been identified by the Opportunity rover in Meridiani Planum. Additional sulfate as gypsum is present in northern circum-polar dunes. Anhydrous sulfates, such as anhydrite (CaSO₄), are probably present but would give no signature in near-infrared spectra.

The total sulfur in visible deposits on Mars is around 10¹⁸ kg SO₃, which is within an order of magnitude of Earth’s oceanic sulfate of 3.2 × 10¹⁸ kg SO₃ but well below Earth’s sulfur inventory of 2.4 × 10¹⁹ kg SO₃ that includes sedimentary sulfur in the form of pyrite and sulfates. Thus, even accounting for its lower surface area, Mars’ surface apparently has a smaller sulfur inventory than the Earth, which is presumably because of less extensive volcanic outgassing.

Evidence of volatile abundances also comes from Martian meteorites [See Meteorites]. These meteorites are known to be from Mars because of their igneous composition, unique oxygen isotope ratios, spread of ages, and gaseous inclusions whose elemental and isotopic compositions closely match the present Martian atmosphere. Ages of crystallization of these basaltic rocks, i.e. the times when the rocks solidified from melts, range from 4.4 to 0.15 billion years, which implies a parent body with active volcanism during this entire interval. Many of the Martian meteorites contain salt minerals, up to 1% by volume,

### Table 16.2 Volatile Reservoirs

<table>
<thead>
<tr>
<th>Water (H₂O) Reservoir</th>
<th>Equivalent Global Ocean Depth:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atmosphere</td>
<td>10⁻⁵ m</td>
</tr>
<tr>
<td>Polar caps and layered terrains</td>
<td>20 m</td>
</tr>
<tr>
<td>Ice, adsorbed water, and/or hydrated salts stored in the regolith</td>
<td>&lt;100 m</td>
</tr>
<tr>
<td>Alteration minerals in 10-km crust assuming 1–3 wt% hydration</td>
<td>150–900 m</td>
</tr>
<tr>
<td>Deep aquifers</td>
<td>None found by radar</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Carbon Dioxide (CO₂) Reservoir</th>
<th>Equivalent Surface Pressure:</th>
<th>Global Mass:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atmosphere</td>
<td>6 mbar</td>
<td>0</td>
</tr>
<tr>
<td>Carbonate in weathered dust</td>
<td>~200 mbar/100 m global average layer of weathered dust</td>
<td>&lt;0.9 × 10¹⁶ kg SO₃ per m of global average soil (assuming &lt;8 wt% sulfur as SO₄)</td>
</tr>
<tr>
<td>Adsorbed in regolith</td>
<td>&lt;40 mbar</td>
<td>&lt;10¹⁷ kg SO₃ (assuming 20 vol % SO₃ in observed volumes of sulfate deposits)</td>
</tr>
<tr>
<td>Carbonate sedimentary rock</td>
<td>~3 mbar (from known outcrops)</td>
<td>~10¹⁷ kg SO₃ (assuming 20 vol % SO₃ in observed volumes of sulfate deposits)</td>
</tr>
<tr>
<td>Crustal subsurface carbonates</td>
<td>&lt;250 mbar per km depth</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sulfur Dioxide (SO₂) Reservoir</th>
<th>Global Mass:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atmosphere</td>
<td>0</td>
</tr>
<tr>
<td>Sulfate in weathered dust</td>
<td>&lt;0.9 × 10¹⁶ kg SO₃ per m of global average soil (assuming &lt;8 wt% sulfur as SO₄)</td>
</tr>
<tr>
<td>Sulfate sedimentary rock reservoirs</td>
<td>~10¹⁷ kg SO₃ (assuming 20 vol % SO₃ in observed volumes of sulfate deposits)</td>
</tr>
</tbody>
</table>
which can include halite (NaCl), gypsum, anhydrite, and carbonates of magnesium, calcium and iron. The bulk meteorite compositions are generally dry, 0.05—0.3 wt% water, compared to terrestrial H$_2$O contents from 0.1 wt% in midocean ridge basalts to 2 wt% in basaltic magmas from subduction zones. There is debate about the extent to which Martian magmas may have degassed on eruption and lost their water. Consequently, estimates for the preeruptive volatile contents of Martian magmas vary from nearly anhydrous to about 2 wt% H$_2$O, which is a range from lunarlike to Earth-like. On the other hand, the Martian mantle is generally inferred to be sulfur rich, with 0.06—0.09 wt% S compared to 0.025 in Earth’s mantle.

One important Martian meteorite, ALH84001, is a sample of 4.1 billion-year-old crust and contains about 1% by volume of distributed, 3.9-billion-year-old carbonate. ALH84001 has been heavily studied because of a controversial investigation in which four features associated with the carbonates were considered of possible biological origin: the carbonates themselves, traces of organic compounds, 0.1 μm-scale structures identified as microfossils, and crystals of the mineral magnetite (Fe$_3$O$_4$) (McKay et al., 1996). However, the biological nature of all these features has been strongly disputed and alternative abiotic origins have been proposed.

2.2. Sources and Losses of Volatiles

Volatile acquisition began during the formation of Mars. Planetary formation models indicate that impacting bodies that condensed from the evolving solar nebula near Mars’ orbit were highly depleted relative to solar composition in the atmospheric volatiles: carbon, nitrogen, hydrogen, and noble gases. Nonetheless, formation of Jupiter and the outer planets would have gravitationally deflected volatile-rich asteroids from the outer asteroid belt and Kuiper Belt comets into the inner solar system. Analyses of the compositions of the Martian meteorites indicate that Mars acquired a rich supply of the relatively volatile elements during its formation. However, carbon, nitrogen, and noble gases are severely depleted in Mars’ atmosphere and surface compared with Earth and Venus, apparently because loss processes efficiently removed these elements from Mars, as they did for hydrogen.

Two processes, hydrodynamic escape and impact erosion, must have removed much of any early Martian atmosphere. Hydrodynamic escape is pressure-driven escape that occurs when a planet’s upper atmosphere is sufficiently warm to expand, accelerate through the speed of sound, and attain escape velocity en masse. Because this process is easiest for hydrogen-rich atmospheres, the general conception of hydrodynamic escape on Mars is of an early hydrogen-rich atmosphere flowing outward in a planetary wind (analogous to the “solar wind”) that entrains and removes other gases. Heavy atoms are carried upward by collisions with hydrogen faster than they diffuse down under gravity, and the downward diffusion gives only weak selectivity to atomic mass. Nonetheless, the high $^{38}$Ar/$^{36}$Ar ratio (Table 16.2) could be a sign of early hydrodynamic escape, although later escape processes can also drive this ratio high (see below).

Intense solar ultraviolet radiation and soft X-rays provide the energy needed to drive hydrodynamic escape. These fluxes would have been at least two orders of magnitude larger than at present during the first $\sim 10^7$ years after the solar system formed as the evolving sun moved toward the

---

**FIGURE 16.1** Water-equivalent hydrogen content of subsurface water-bearing soils derived from the Mars Odyssey Neutron Spectrometer. From Feldman et al. (2004).
main sequence. Although the early Sun was 25–30% less luminous overall, studies of early stars suggest that the early Sun was rotating more than 10 times faster than present, which would have caused more magnetic activity, associated with over a 100 times more emission in the extreme ultraviolet portion of the spectrum than today. Consequently, hydrodynamic escape would have been a very efficient atmospheric removal mechanism if hydrogen had been a major atmospheric constituent during this period.

The amount of hydrogen in the early atmosphere of a terrestrial planet depends on the chemical reaction of iron and water during accretion and the segregation of the core and mantle. If water brought in by impactors could mix with free iron during this period, it would oxidize free iron, releasing large amounts of hydrogen to the atmosphere and fostering hydrodynamic escape. Interior modeling constrained by Mars’ gravitational field and surface composition together with analyses of the composition of the Martian meteorites indicates that Mars’ mantle is rich in iron oxides relative to the Earth, consistent with the hypothesis that a thick hydrogen-rich atmosphere formed at this early stage. It has been suggested that hydrodynamic escape removed the equivalent of an ocean at least 1 km deep together with most other atmospheric volatiles from Mars, although this estimate is based on extrapolation from the current value of the deuterium−hydrogen ratio (D/H), which is uncertain because D/H may reflect geologically recent volatile exchange rather than preferential loss of hydrogen compared to deuterium over the full history of Mars. Comets arriving after the completion of hydrodynamic escape may have brought in much of the atmospheric volatiles in the current inventory.

Early Mars was also potentially vulnerable to impact erosion—the process where atmospheric gases are expelled as a result of the large-body impacts. Big impacts release enough energy to accelerate atmospheric molecules surrounding the impact site to speeds above the escape velocity. A large fraction of these fast molecules escape. Since escape is easier with a smaller gravitational acceleration of the planet, impact erosion would have been far more efficient on Mars than on Earth. The early history of the inner solar system is characterized by a massive flux of large asteroids and comets, and most models suggest net atmospheric erosion for early Mars rather than accumulation of volatiles. Based on dating of lunar rocks and impact features, bombardment by massive objects is known to have declined rapidly with time after planet formation. The interval from 4.1 to 3.7 billion years ago is the Noachian eon in Martian geologic time, so that massive bombardment effectively ceased around the end of the Noachian.

The base of the Noachian is defined by the time of formation of the Hellas impact basin around 4.1 billion years ago, before which is the Pre-Noachian. Interestingly, the lunar record suggests a spike in the impact flux 4.0–3.8 billion years ago, called the Late Heavy Bombardment, when most impact melt rocks formed. Thus, the late Noachian was probably a time of particularly intense impacts.

Bombardment by massive bodies during the Noachian has left an imprint in the form of large impact craters that are obvious features of the southern hemisphere (Figure 16.2). More subtle “ghost” craters and basins that have been largely erased by erosion and/or filling in the relatively smooth northern plains provide further evidence of Noachian impact bombardment. Calculations suggest that impact erosion should have removed all but ~1% of an early CO2-rich atmosphere (e.g., see Carr, 1996, p. 141). Water in ice and carbon in carbonates would have been relatively protected, however, and the exact efficiency of the removal of volatiles by impact erosion is unknown.

What was the size of Mars’ volatile reservoirs after Late Heavy Bombardment, some 3.7 billion years ago? The isotopic ratios $^{13}C/^{12}C$, $^{18}O/^{16}O$, $^{38}Ar/^{36}Ar$, and $^{15}N/^{14}N$ are heavy compared with the terrestrial ratios (see Table 16.1). This has been interpreted to indicate that 50–90% of the initial reservoirs of CO2, N2, and cosmogenic argon have been lost over the past 3.7 billion years by mass-selective nonthermal escape from the upper atmosphere (mainly sputtering produced by the impact of the solar wind on the upper atmosphere). Considering the possible current reservoirs of CO2 in Table 16.2, the resulting CO2 available 3.7 billion years ago could have been as much as ~1 bar or as little as a few tens of millibars.

Another approach to estimating the CO2 abundance at the end of the Noachian is based on the abundance of $^{85}$Kr in the present atmosphere. Since this gas is chemically inert and too heavy to escape after the end of massive impact bombardment, its current abundance probably corresponds closely to the abundance at the end of the Noachian. Impact erosion would have effectively removed all gases independent of atomic mass, so the ratio of $^{85}$Kr abundance to C in plausible impactors (Kuiper Belt comets or outer solar system asteroids) can then yield estimates of the total available CO2 reservoir at the end of the Noachian. The corresponding atmospheric pressure, if all CO2 were in the atmosphere, would be only ~0.1 bar, in the lower range of estimates from the isotopic and escape flux analysis. This low estimate is consistent with the low modern nitrogen abundance after allowing for mass-selective escape as indicated by the high $^{15}N/^{14}N$ ratio (Table 16.1).

If ~0.1 bar was left at the end of the Noachian, besides nonthermal escape, slow carbonate weathering of atmospheric dust could also have removed CO2 from the atmosphere (as mentioned previously). This irreversible mechanism may account for the fate of a large fraction of the CO2 that was available in the late Noachian, along with adsorbed CO2 in the porous regolith (Table 16.2). It has long been speculated that much of the CO2 that was in the
Noachian atmosphere got tied up as carbonate sedimentary deposits beneath ancient water bodies. However, so far only small outcrops and no significant quantities of carbonate sedimentary rocks have been found (see further discussion below).

Escape of water in the form of its dissociation products H and O takes place now, and must have removed significant amounts of water since the Noachian. Isotopic ratios of D/H and $^{18}$O/$^{16}$O in the atmosphere and in Martian meteorites and escape flux calculations provide rather weak constraints on the amount that has escaped over that period. Upper bounds on the estimates of water loss are 30–50 m of equivalent global ocean. These amounts are roughly comparable to estimates of the water currently stored as ice in the polar caps and regolith (Table 16.2).

Sulfur is not stable in the current Martian atmosphere as either sulfur dioxide (SO$_2$) or hydrogen sulfide (H$_2$S) because both gases oxidize and ultimately produce sulfate aerosols that fall to the surface; however, significant sulfur gases must have been introduced into the atmosphere by volcanism. Estimated ages of volcanic surfaces on Mars indicate that rates of volcanism declined and became more intermittent after about 3.5 billion years ago. Formation of the Tharsis ridge volcanic structure, believed to have occurred in the late Noachian eon, must have corresponded with outgassing of large amounts of sulfur as well as water from the mantle and crust. The probable quantity of sulfur released is consistent with the relatively high mass fraction of sulfur in the soil and the presence of large deposits of sedimentary sulfates. Martian meteorites are $\sim$5 times as rich in sulfur as in water and it is likely that the regolith contains more sulfur than water. The volatile elements chlorine and bromine are also abundant in rocks and soils, but more than an order of magnitude less than sulfur.

An important observation in the sulfates found in Martian meteorites is that sulfur and oxygen isotopes are found in relative concentrations that are mass-independently fractionated. Most kinetic processes fractionate isotopes in a mass-dependent way. For example, the mass difference between $^{34}$S and $^{32}$S means that twice as much fractionation between these isotopes is produced as between $^{33}$S and $^{32}$S in a mass-dependent isotopic discrimination process such as diffusive separation. Mass-independent fractionation (MIF) is a deviation from such proportionality. MIF is found to arise when ultraviolet radiation interacts with certain atmospheric gases in photochemistry. On Earth, the MIF of oxygen in sulfates in the extraordinarily dry Atacama Desert is taken to prove that these sulfates were deposited by photochemical conversion of atmospheric SO$_2$ to submicron particles and subsequent dry deposition. The MIF signature in sulfates in Martian meteorites suggests that a similar process produced these sulfates on Mars, and implies that the sulfur cycled through the atmosphere at some ancient time.

3. PRESENT AND PAST CLIMATES

3.1. Present Climate

The thin, predominantly carbon dioxide atmosphere produces a small greenhouse effect, raising the average
surface temperature of Mars only 5–8 K above the 210 K temperature that would occur in the absence of an atmosphere. Carbon dioxide condenses out during winter in the polar caps, causing a seasonal range in the surface pressure of about 30%. There is a small residual CO₂ polar cap at the South Pole, which persists all year round in the current epoch; it represents a potential increase in the CO₂ pressure of 4–5 mbar if it were entirely sublimated into the atmosphere. The atmospheric concentration of water vapor is controlled by saturation and condensation, and so varies seasonally and daily. Water vapor exchanges with the polar caps over the course of the Martian year, especially with the North Pole. After sublimation of the winter CO₂ polar cap, the summertime central portion of the cap surface is water ice. Water vapor sublimates from this surface in northern spring to early summer, and is transported southward, but most of it is precipitated or adsorbed at the surface before it reaches southern high latitudes.

In addition to gases, the atmosphere contains a variable amount of dust as well as icy particles that form clouds. Dust loading can become quite substantial, especially during northern winter. Transport of dust from regions where the surface is being eroded by wind to regions of dust deposition occurs in the present climate. Acting over billions of years, wind erosion, dust transport, and dust deposition strongly modify the surface (see Section 3.5). Visible optical depths can reach ~5 in global average and even more in local dust storms. A visible optical depth of 5 means that direct visible sunlight is attenuated by a factor of 1/e^5, which is roughly 1/150. Much of the sunlight that is directly attenuated by dust reaches the surface as scattered diffuse sunlight. Median dust particle diameters are ~1 μm, so this optical depth corresponds to a column dust mass ~3 mg/m². Water ice clouds occur in a “polar hood” around the winter polar caps and over low latitudes during northern summer, especially over uplands. Convective carbon dioxide ice clouds occur at times over the polar caps, and they occur rarely as high-altitude cirrus clouds of CO₂ ice particles.

Orbital parameters cause the cold, dry climate of Mars to vary seasonally in somewhat the same way as intensely continental climates on Earth. The present tilt of Mars’ axis (25.2°) is similar to that of the Earth (23.5°), and a Martian year is 687 Earth days long or about 1.9 Earth years. Consequently, seasonality bears some similarity to that of the Earth but Martian seasons last about twice as long on average. However, the eccentricity of Mars’ orbit is much larger than that of the Earth’s (0.09 compared with 0.015) and perihelion (the closest approach to the sun) currently occurs near northern winter solstice. As a consequence, asymmetric between northern and southern seasons are much more pronounced than on the Earth. Mars’ rotation rate is similar to that of the Earth’s, and like the Earth, the atmosphere is largely transparent to sunlight so that heat is transferred upward from the solid surface into the atmosphere. These are the major factors that control the forces and motions in the atmosphere, i.e. atmospheric dynamics, which is similar on Mars and Earth. Both planetary atmospheres are dominated by a single meandering mid-latitude jet stream, strongest during winter, and a Hadley circulation in lower latitudes. The Hadley circulation is strongest near the solstices, especially northern winter solstice, which is near perihelion, when strong rising motion takes place in the summer (southern) hemisphere and strong sinking motion occurs in the winter (northern) hemisphere.

Mars lacks an ozone layer, and the thin, dry atmosphere allows very short wavelength ultraviolet radiation to penetrate to the surface. In particular, solar ultraviolet radiation in the range 190–300 nm, which is shielded on Earth by the ozone layer and oxygen, can reach the lower atmosphere and surface on Mars. This allows water vapor dissociation close to the Martian surface (H₂O + ultraviolet photon → H + OH). As a consequence of photochemical reactions, oxidizing free radicals (highly reactive species with at least one unpaired electron, such as OH or HO₂) are produced in near-surface air. In turn, any organic material near the surface rapidly decomposes and the soil near the surface is oxidizing. These conditions as well as the lack of liquid water probably preclude life at the very surface on present-day Mars.

Although liquid water may not be completely absent from the surface, even in the present climate, it is surely very rare. This is primarily because of the low temperatures. Even though temperatures of the immediate surface rise above freezing at low latitudes near midday, above-freezing temperatures occur only within a few centimeters or millimeters on either side of the surface in locales where the relatively high temperatures would be desiccating. A second factor is the relatively low pressure. Over large regions of Mars, the pressure is below the triple point for which exposed liquid water would rapidly boil away.

Because the present atmosphere and climate of Mars appear unsuitable for the development and survival of life, at least near the surface, there is great interest in the possibility that Mars had a thicker, warmer, and wetter atmosphere in the past.

### 3.2. Past Climates

Several types of features suggest that fluids have shaped the surface during all eons—the Noachian (4.1–3.7 billion years ago), Hesperian (3.7 to 3.5–3.0 billion years ago), and Amazonian (from 3.5 to 3.0 billion years ago to the present). In terrains whose ages are estimated on the basis of crater distributions and morphology to be Noachian, “valley network” features are abundant (Figure 16.3). The morphology of valley networks is very diverse, but most consist of dendritic networks of small valleys, often with V-shaped profiles in their upper reaches becoming more U-shaped downstream. Their origin is
attributed to surface water flows or groundwater sapping. The latter is when underground water causes erosion and collapse of overlying ground. Although often much less well developed than valley network systems produced by fluvial erosion on Earth, Martian valley networks are suggestive of widespread precipitation and/or groundwater sapping that would have required a much warmer climate, mainly but not entirely, contemporaneous with termination of massive impact events at the end of the Noachian (~3.7 billion years ago). In Figure 16.3, we show two examples of valley network features. Figure 16.3(a) is a high-resolution image that shows a valley without tributaries in this portion of its reach (although some tributary channels are found farther upstream), but its morphology strongly suggests repeated flow events. Figure 16.3(b) shows fairly typical valley networks that are incised on Noachian terrain. A relatively small number of valleys on Mars terminate in deltas, which provides strong evidence of liquid water. However, the deltas are undissected, which implies an abrupt end to the era of valley formation. Some deltaic deposits also contain clay minerals such as the one in Jezero crater (18.4° N, 282.4° W) (Figure 16.4), which is fed by a valley network northwest of Isidis.

Valley networks are incised on top of a Noachian landscape of craters with heavily degraded rims and infilling or erosion (Figure 16.3(b)). Such crater morphologies indicate relatively high erosion rates. Some models suggest that erosion and deposition was caused by fluvial activity, at least in part. However, the interpretation is complex because the image data suggests that craters were also degraded or obscured by impacts, eolian transport, mass wasting and, in some places, airfall deposits such as volcanic ash or impact ejecta.

Besides valley networks, a second class of geomorphic features suggesting liquid flow is a system of immense
channels that formed during the late Hesperian (Figure 16.5). These features, referred to as “outflow channels” or “catastrophic outflow channels”, are sometimes more than 100 km in width, up to \( \sim 1000 \) km in length, and as much as several kilometers deep. They occur mainly in low latitudes (between 20° north and south) around the periphery of major volcanic provinces such as Tharsis and Elysium, where they debouch northward toward the low-lying northern plains. The geomorphology of these channels has been compared with the scablands produced by outwash floods in Eastern Washington State from ice age Lake Missoula, but if formed by flowing water, flow volumes must have been larger by an order of magnitude or more. It has been estimated that the amount of water required to produce them is equivalent to a global ocean at least a few hundred meters deep. Many of these channels originate in large canyons or jumbled chaotic terrain that was evidently produced by collapse of portions of the plateau surrounding Tharsis. The origin of outflow channels is unknown, but the dominant hypothesis is that they were generated by catastrophic release of water from subsurface aquifers or rapidly melting subsurface ice. Alternatively, volcanic or impact heating caused catastrophic dehydration of massive hydrated sulfate deposits and resulting high-volume flows of liquid water. If water was released by these flows, its fate is unknown, although a number of researchers have proposed that water pooled in the northern plains and may still exist as ice beneath a dust-covered surface. A minority view is that the outflow channels were not carved by water but runny lavas. Some outflow channels (e.g. Mangala Valles, Athabasca Valles) have lava flows on their floors (Figure 16.6) and the source of some outflow channels are also sources of lava, e.g. Cerberus Fossae for Athabasca Valles or Memnonia Fossae for Mangala Valles. However, water and lava are not mutually exclusive. Magmatic intrusions could have melted subsurface ice, triggering floods of water in association with lava flows.

Unlike the valleys and channels, gullies are very recent features that have been interpreted to imply fluvial flow, but evidence from MRO images suggests that gullies can form today in association with carbon dioxide ice rather than liquid water. Gullies are incisions of tens to hundreds of meters length commonly found on poleward facing sloping walls of craters, plateaus, and canyons, mainly at southern midlatitudes (\( \sim 30°-55° \)) (Figure 16.7). The gullies typically have well-defined alcoves above straight or meandering channels that terminate in debris aprons. Their setting and morphology has led to comparisons with debris flows in terrestrial alpine regions that are produced by rapid release of meltwater and consist typically of \( \sim 75\% \) rock and silt carried by \( \sim 25\% \) water.

High-resolution images from MRO show that gullies are forming on Mars today when carbon dioxide frost turns to vapor at the end of winter (Figure 16.7). Sublimation of CO2 frost presumably causes a fluidlike, dry flow of rock and soil with the CO2 gas providing lubrication. This mechanism explains the predominant location of gullies in the southern hemisphere where there is a greater seasonal accumulation and distribution of carbon dioxide frost than in the north because of prolonged southern winters. Consequently, gullies do not require a warm climate or low-latitude reservoirs of subsurface water or ice. In fact, gullies...
induced by CO$_2$ sublimation illustrate how features that look very similar to fluvially eroded analogs on Earth can actually have other causes.

Other curious features that develop on Mars today are Recurring Slope Lineae (RSL), which may be associated with seasonal melting of water ice. RSL are dark, narrow (0.5—5 m) lineaments that develop during warm seasons on steep rocky slopes that are equator facing in equatorial or mid latitude regions (Figure 16.8). The RSL fade and disappear during the cold season. Their formation is not understood, but an association with peak surface temperatures ranging from 250 to 300 K supports the idea that melting of water ice is probably involved. Salts dissolved in the water would also allow flow below 273 K.

Minerals on very ancient surfaces provide evidence for a wetter early Mars that bolsters the inference of liquid water from valley networks and outflow channels (if water eroded). Much of the surface on Mars is basalt, a dark-colored igneous rock rich in iron and magnesium silicate minerals. When basalt reacts with water, “alteration minerals” are produced, such as clays. So alteration minerals can be diagnostic of the past presence of liquid water and sometimes they suggest its pH. For example, clay minerals tend to be produced when alkaline water reacts with basaltic minerals. Notably, clay minerals have been detected in mudstone by NASA’s Curiosity Rover in an area interpreted as an ancient dried-up lake bed in Gale Crater.

Alteration minerals on Mars have a broad trend in their distribution in geologic time. Orbital spectroscopy suggests that hydrous alteration minerals cover about 3% of Noachian surfaces in the form of clays and some carbonates. Sulfate minerals tend to be found on late Noachian or Hesperian surfaces, while younger Amazonian surfaces have reddish, dry iron oxides. To some, this broad pattern suggests three environmental epochs, starting in Noachian when alkaline or neutral pH waters made clay minerals from basalt. In the second epoch, sulfates were presumably derived from sulfur-rich volcanic gases. The third epoch is the cold, dry environment with rust-colored surfaces that continues today.

3.3. Mechanisms for Producing Past Wetter Environments

Despite extensive investigation, the causes of early warm climates in the Noachian or Hesperian remain to be identified. Here we review several possibilities.

3.3.1. Carbon Dioxide Greenhouse

A suggestion put forward after the Mariner 9 orbiter mission in 1972 is that the early atmosphere contained much more CO$_2$ than it does now. The idea is that substantial CO$_2$ caused an enhanced greenhouse effect through
its direct infrared radiative effect and the additional greenhouse effect of increased water vapor that the atmosphere would have held at higher temperatures. Applied to the late Noachian period of valley network formation, this theory runs into difficulty because of the lower solar output at 3.7 billion years ago (about 75% of the present flux), and consequent large amount of CO$_2$ required to produce an adequate CO$_2$–H$_2$O greenhouse effect. At least several bars of CO$_2$ would have been required to produce widespread surface temperatures above freezing. However, such thick atmospheres are not physically possible because CO$_2$ condenses into clouds at $\sim$1 bar and also into permanent ice caps at pressures exceeding $\sim$3 bar. CO$_2$ ice clouds can have some greenhouse effect because they scatter infrared radiation, but three-dimensional climate models show that the clouds are never opaque enough in the infrared to warm the surface to a mean temperature above freezing anywhere on the planet.

Apart from the difficulties with climate simulations, if a massive CO$_2$ atmosphere ever existed at the same time as abundant liquid water, it would have eventually collapsed due to removal of the CO$_2$ by dissolving in the water and subsequently forming carbonate sediments. However, despite extensive efforts, few outcrops of carbonate sediments have been found and their equivalent global pressure of CO$_2$ is insubstantial (Table 16.2). Carbonates are absent even in areas in which water is interpreted to have flowed, such as valley networks, and in areas of extensive erosion where we would expect exposures of carbonate sediments buried beneath regolith. In contrast, sulfate sedimentary deposits are widespread at low latitudes, some in terrains that have been exhumed by wind erosion. In retrospect, it is not surprising that carbonate reservoirs have not been found. In the presence of abundant sulfuric acid, carbonate would be quickly converted to sulfate with release of CO$_2$ to the atmosphere, where it would be subject to various loss processes discussed earlier.

Although a future discovery of a large carbonate sediment reservoir cannot be ruled out, it seems doubtful, and the amount of CO$_2$ available seems inadequate to have produced a warm enough climate to account by itself for the valley networks by surface in the late Noachian.

### 3.3.2. Impact Heating

The largest asteroid or comet impacts would vaporize large quantities of rock. Vaporized rock would immediately spread around the planet, condense, and, upon reentry into the atmosphere, would flash heat the surface to very high temperature. This would quickly release water from surface ice into the atmosphere. Upon precipitation, this water could produce flooding and rapid runoff over large areas. Water would be recycled into the atmosphere as long as the surface remained hot, anywhere from a few weeks to thousands of years depending on impact size. It has been proposed that this is an adequate mechanism for producing most of the observed valley networks and that the drop off in impact flux explains why valley formation declines after the early Hesperian. Although a very extended period of warm climate would not be produced this way, repeated short-term warm climate events could have occurred during the late Noachian to early Hesperian. Detailed questions of timing of large impact events and formation of the valley network features needed to test this hypothesis remain to be resolved. The effect of water clouds on the postimpact climate also needs further study. However, impact heating of ice must have released water to the atmosphere and caused subsequent precipitation at some times during the Noachian.

### 3.3.3. Sulfur Dioxide Greenhouse

The high abundance of sulfur in surface rocks and dust as well as in the Martian meteorites suggests that Martian volcanism may have been very sulfur rich. In contrast to Earth, Martian volcanoes may have released sulfur in amounts equal to or exceeding water vapor. In an atmosphere in the presence of water vapor, reduced sulfur would rapidly oxidize to SO$_2$ and then form aerosols of sulfate and, in more reducing atmospheres, elemental sulfur also. Sulfur dioxide is a powerful greenhouse gas, but in the presence of liquid water, it dissolves and is removed from the atmosphere by precipitation very rapidly. More importantly, the sulfate and sulfur aerosols that form reflect sunlight so that the net effect of sulfur gases would be to cool Mars. Such cooling has been measured as an effect of volcanic eruptions on Earth that inject SO$_2$ into the stratosphere where the resulting sulfate aerosols increase the albedo of the Earth.

### 3.3.4. Methane-Aided Greenhouse

Methane is also a greenhouse gas, but because of its long-term instability in the atmosphere resulting from photolysis and oxidation, it is not an attractive option for contributing to an early warm climate. To have a role of any significance, early Mars would require a global methane flux similar to that produced by the present-day biosphere on Earth. Even so, calculations suggest that the net warming would be limited and inadequate to solve the early climate problem.

### 3.3.5. Hydrogen-Aided Greenhouse

In an atmosphere that is thick and hydrogen-rich, hydrogen behaves as a greenhouse gas and this might have relevance for early Mars. Although hydrogen (H$_2$) is a nonpolar molecule, collisions with other molecules can cause it to acquire a temporary dipole or absorb infrared photons in
transitions that are normally forbidden; the effect is “collision-induced absorption” (CIA). Indeed, hydrogen is an important greenhouse gas in all the giant planet atmospheres because of CIA. If a thick atmosphere of early Mars had five to tens of percent H₂, the greenhouse warming might be significant. However, hydrogen easily escapes from Mars into space and so very large volcanic fluxes of H₂ would be required. Furthermore, such a solution to the early Mars climate problem, if true, would disfavor life on the early planet because hydrogen is a food for primitive microbes and so its high abundance in the early air would imply an absence of biological consumption.

3.3.6. Mechanisms for Producing Fluvial Features in Cold Climates

Although some precipitation must have occurred due to impacts and short-lived greenhouse warming is plausible, other factors may have been conducive to widespread fluvial features. One proposal concerns seasonal snowmelt under atmospheres that have surface pressures of a few hundred millibars. A second idea is that many of the fluids were low-temperature brines that can exist at temperatures far below 0 °C.

On Mars today, at a particular latitude with the same solar heating, there is a little variation of temperature with altitude, but when the pressure exceeds a few hundred millibars, vertical convection maintains a temperature gradient in the atmosphere that influences the temperature of topography, so that the tops of mountains become much colder than low altitudes. As a result, the southern highlands can become ice or snow covered and periodic melting from changes in orbital characteristics (see below) might be sufficient to erode valleys. If salty fluids rather than pure water were prevalent on early Mars, melting could have occurred at temperatures many tens of degrees below 0 °C. For example, the eutectic temperature (at which a salty solution freezes into ice and solid salt) is −50 °C for calcium chloride (CaCl₂), −75 °C for calcium perchlorate (Ca(ClO₄)₂), and −57 °C for magnesium perchlorate (Mg(ClO₄)₂). In addition, laboratory experiments show that brines of some salts (such as the aforementioned perchlorates) can supercool tens of degrees Celsius below the eutectic temperature for periods exceeding a Martian day.

The outflow channels remain enigmatic because of the large flows of water that are required relatively late in geologic time, in the late Hesperian, when the atmosphere was probably thin. The idea that very fluid lava flows might have been responsible for channel erosion is the only concept that has no need of a warmer climate to allow flows to persist for hundreds of kilometers. Extensive outflow channels, some of which strongly resemble Martian outflow channel features, are found on Venus and are explained by low-viscosity lava flows. The spatial relationship between the Martian outflow channels and the major volcanic constructs is suggestive of the idea that very fluid lavas may have played some role in the formation of outflow channels.

3.4. Milankovitch Cycles

As on Earth, Mars’ orbital elements (obliquity, eccentricity, argument of perihelion) exhibit oscillations known as Milankovitch cycles at periods varying from 50,000 to several million years. The obliquity and eccentricity oscillations are much larger in amplitude on Mars than on Earth (Figure 16.9). Milankovitch cycles cause climate variations in two ways. First, they control the distribution of incoming solar radiation (insolation) on both an annual average and seasonal basis as functions of latitude. Second, because Milankovitch-driven changes of insolation force variations of annual average surface temperature, they can cause exchanges of volatiles between various surface reservoirs and the atmosphere. Water vapor can migrate between polar cap ice deposits, and ice and adsorbed water in the regolith. Carbon dioxide can move between the atmosphere, seasonal residual polar caps, and the surface adsorption reservoir. Milankovitch variations are believed to be responsible for complex layered structure in both the north polar water ice cap and terrains surrounding the south polar residual carbon dioxide ice cap. Also, rhythmic variations in the layering of some sedimentary deposits in low latitudes have been interpreted as a probable sign of influence from orbital cycles, perhaps because such cycles shifted ice to the tropic in times of high obliquity that were followed by periods of melting.

In general, annual average polar cap temperatures increase relative to equatorial temperatures as obliquity increases. At very low obliquity (<10°–20° depending on the precise values of polar cap albedo and thermal emissivity), the carbon dioxide atmosphere collapses onto permanent carbon dioxide ice polar caps. Orbital calculations indicate that this collapse could occur ~1–2% of the time. At high obliquity, atmospheric pressure may increase due to warming and release of adsorbed carbon dioxide from high-latitude regolith. Calculations indicate, however, that the maximum possible pressure increase is likely to be small, only a few millibars, so Milankovitch cycles are unlikely to have been responsible for significant climate warming.

Evidence for the active influence of Milankovitch-type cycles includes a thin, patchy mantle of material, apparently consisting of cemented dust, that has been observed within a 30°–60° latitude band in each hemisphere, corresponding to places where near-surface ice has been stable in the last few million years due to orbital changes. The material is interpreted to be an atmospherically deposited ice—dust mixture from which the ice has sublimated.
3.5. Wind Modification of the Surface

Orbital and landed images of the surface show ubiquitous evidence of active wind modification of the surface, which complicates the interpretation of climate and volatile history. The action of wind erosion, dust transport, and dust deposition is modulated by Milankovitch cycles and must have strongly changed the surface over the last few billions of years and during the Noachian, as we discuss below.

Today, dunes, ripples, and other bedforms are widespread. Wind-modified objects, known as ventifacts, are very evident in the grooves, facets, and hollows produced by the wind in rocks at the surface. Yardangs are also common, which are positive relief features in coherent materials sculpted by wind on scales from tens of meters to kilometers. Strong winds exert stress on the surface that can initiate saltation (hopping motion) of fine sand grains (diameter \( \sim 100-1000 \mu m \)) and creep of larger particles. Saltating grains can dislodge and suspend finer dust particles (diameters \( \sim 1-10 \mu m \)) in the atmosphere, thereby initiating dust storms. Minimum wind speeds required to initiate saltation are typically \( \sim 30 \text{ m/s} \) at the level 2 m above the surface, but this saltation threshold wind speed decreases with increasing surface pressure.

Strong winds needed for saltation are rare. Wind observations at the Viking Lander sites and computer simulations of the atmospheric circulation suggest that they occur at most sites <0.01% of the time. Nevertheless, over the planet as a whole, dust storms initiated by saltation are common; they tend to occur with greater frequency in low-elevation regions than uplands because the relatively high surface pressure in lowlands lowers the saltation threshold wind speed. They are favored by topographic variations, including large and small-scale slopes and are common over ice-free surfaces near the edges of the seasonally varying polar caps and in “storm track” regions where the equator-to-pole gradient of atmospheric temperature is strong. Dust storms generated by strong winds and saltation are common in some tropical lowland regions, especially close to the season of perihelion passage when the Hadley
circulation is strong (near the southern summer solstice at the current phase of the Milankovitch cycle). During some years, these perihelion season storms expand and combine to such an extent that high dust opacity spreads across almost the entire planet. These planetwide dust events are fostered by positive feedbacks between dust-induced heating of the atmosphere, which contributes to driving wind systems, and the action of the wind in picking up dust.

Dust can also be raised at much lower wind speeds in dust devils, which are small-scale quasivertical convective vortices called dust devils. Because the atmosphere is so thin, convective heating per unit mass of atmosphere is much greater on Mars than anywhere on Earth, and Martian dust devils correspondingly tend to be much larger sizes (diameters up to several hundred meters and depths up to several kilometers). Since the winds required to raise dust in dust devils are lower than salination threshold winds, dust devils are common in some regions of Mars during the early afternoon and summer when convective heating is strongest. They are often associated with irregular dark tracks produced by the removal of a fine dust layer from an underlying darker stratum. The relative importance of large saltation-induced dust storms and dust devils to the overall dust balance is unclear, but modeling studies suggest that the former are substantially more important.

Over the past 4 billion years, there must have been substantial systematic wind transport of fine soil particles from regions in which erosion is consistently favored to regions of net deposition. Models of Martian atmospheric circulation and the saltation process suggest that net erosion must have taken place in lowland regions, particularly in the northern lowlands, the Hellas basin, and some tropical lowlands (e.g. Isidis Planitia and Chryse Planitia), with net deposition in upland regions and in some moderate elevation regions where the regional slope is small and westward facing, such as portions of Arabia Terra and southern portions of Amazonis Planitia. The distribution of surface thermal inertia inferred from the measured surface diurnal temperature variation supports these inferences. Regions of high thermal inertia, corresponding to consolidated or coarse-grained soils, exposed surface rocks, and bedrock patches, are found where the circulation-saltation models predict net erosion over Milankovitch cycles, and regions of very low thermal inertia corresponding to fine dust are found where net deposition is predicted by the models.

There are no terrestrial analogs of surfaces modified by wind erosion and deposition over 4 billion years, so it is difficult to fully comprehend the modifying effect of Martian winds over such a long time. However, it is clear from the surface imagery that in some areas repeated burial and exhumation events have taken place. Based on the heights of erosionally resistant mesas, the Meridiani Planum site of the Opportunity rover activities appears to have been exhumed from beneath at least several hundred meters and perhaps as much as several kilometers of soil. Many of the sulfate layer deposits described above appear to be undergoing exhumation. Since surface features can be repeatedly buried, exposed, and reburied over time, inferences of event sequences and surface ages from crater size distributions are rendered complex.

Because the salination process operates on the extreme high velocity tail of the wind speed distribution, it is very sensitive to surface density or pressure changes. Model results indicate that an increase in surface pressure up to only 40 mbar would increase potential surface erosion rates by up to two orders of magnitude. If, as is likely, Mars had a surface pressure ~ 100 mbar or higher during the late Noachian, rates of surface modification by wind should have been orders of magnitude greater than today. Indeed, it has long been observed that late Noachian surfaces were undergoing much more rapid modification than during later periods. This has generally been attributed to precipitation and runoff under a warmer climate regime, as discussed earlier. But surface modification by winds under a denser atmosphere should also have contributed to the observed rapid modification of late Noachian age surfaces.

4. CONCLUDING REMARKS

Although ice is now known to be widespread near the surface and there is considerable evidence that liquid water once flowed across the surface in dendritic valley networks, a major outstanding problem is that we still do not know the exact conditions responsible for releasing liquid water on early Mars or what controlled the early ancient climate (see Figure 16.10 for a timeline). As the sophistication of climate models for early Mars have grown, it has perhaps become increasingly difficult (rather than easier) to explain how Mars could have had a sustained, warm, and wet climate during the Noachian or Hesperian. The basic problem is that the early Martian surface needs about 80 °C of greenhouse warming to raise its mean global temperature above freezing, which is more than double the greenhouse warming of 33 °C of the modern Earth. The discovery of only a few outcrops of sedimentary carbonates provides little support for the idea that a large CO2 reservoir exists on Mars today and was derived from an earlier thick atmosphere; in any case, three-dimensional climate models are unable to raise global mean surface temperatures above freezing for any CO2–H2O atmosphere. More generally, no widely accepted solution to the climate of early Mars has been found despite numerous modeling permutations of an enhanced greenhouse effect with various gases on early Mars.

In view of the new data and theoretical constraints, other candidate mechanisms for the release of fluids at the surface to form valley networks and outflow channels need to be considered. During the Noachian, large impacts
would have provided sufficient heat to vaporize subsurface volatiles, such as water and CO$_2$ ice. Consequently, impacts may have generated many temporary warm, wet climates, which would be accompanied by erosion from rainfall or the recharge of aquifers sufficient to allow groundwater flow and sapping. Such a scenario might explain why valley networks appear to peak in occurrence around the era of the Late Heavy Bombardment (the late Noachian, early Hesperian) and why there is subsequently an apparently large drop in erosion rates. The periodic release of meltwater from high-altitude ice and snow that occurs under moderately thicker atmospheres is another mechanism that seems likely to have produced fluvial erosion on early Mars. Also, the potential of thicker Noachian atmospheres to cause wind erosion requires further consideration.

Geochemical data and models suggest that most of Mars’ original volatile inventory was lost early by hydrodynamic escape and impact erosion (Figure 16.10). However, we still do not know the degree to which some volatiles were sequestered into the subsurface as minerals and protected. Future landed and orbital missions can refine our understanding of the distribution and properties of hydrated minerals and subsurface ices. However, determining the amount of sulfate and carbonate that has been sequestered into the subsurface will require drilling into the subsurface. Further study of the geology of Mars from orbit and the surface will also help establish the amount of fluvial erosion, its duration, or episodicity. Finally, resolving questions of timing ultimately requires absolute (radiometric) dating of Martian surfaces.

**FIGURE 16.10** An overview of the Martian geologic timescale and its relationship to geologic features, predominant minerals, and events affecting the atmosphere.

**BIBLIOGRAPHY**


Mars, the outermost of the four terrestrial planets—Mercury, Venus, Earth, and Mars—is intermediate in size between the Earth and the Moon. The terrestrial planets all have solid surfaces, each of which preserves a partial record of how each planet has evolved. Successive events, such as volcanic eruptions or meteorite impacts, both create a new record and partly destroy the old. The task of the geologist is to reconstruct the history of the planet from what is preserved at the surface. Both Mercury and the Earth’s moon appear to have become geologically inactive early in their history, so most of the preserved record dates from very early in the history of the solar system, 3.5 Gyr ago. The geologic record on Venus is relatively young, most of the surface apparently having formed in the past 1 Gyr. The record on Earth is also mostly young, although ancient records are preserved on some continents. On Mars we have a record that spans almost the entire history of the solar system. While much of the Martian surface dates back to the first billion years, volcanism, tectonism, fluvial activity, glaciation, and impact processes appear to have continued at a low rate until the recent geologic past, allowing us to follow the evolution of the planet for almost its entire history.

Our knowledge of the geologic evolution of the Earth has been largely derived from the study of the lithology, chemistry, mineralogy, and distribution of rocks at the surface.Geomorphology has played a relatively minor role. On Mars, however, much of what we know about the geology is derived from the morphology of the surface. While geomorphic data are being increasingly supplemented by orbital mineralogy measurements as well as data from Martian meteorites and landers/rovers, our global perspective is still largely based on the appearance of the surface from orbit, which is the main subject of this chapter.

1. MARS EXPLORATION

The modern era of Mars exploration began on July 14, 1965, when the Mariner 4 spacecraft flew by the planet and transmitted to the Earth 22 close-up pictures of the surface with resolutions of several kilometers per pixel
Prior to this, we depended on telescopic observations, with resolution at best of 100–200 km. The observations revealed no topography, only surface markings. Nonetheless, prior to the space age we knew that Mars has a thin CO₂ atmosphere, polar caps that advance and recede with the seasons, and surface markings that undergo annual and secular changes. Geologic studies of the planet, however, could realistically begin only when we acquired spacecraft data.

The Mariner 4 pictures, covering only a small fraction of the southern hemisphere, revealed only an ancient surface that resembled the lunar highlands. These results were disappointing because it had been speculated that Mars, having an atmosphere and being larger than the Moon, might be more Earth-like than Moon-like. Mariner 4 was followed by two more Mariner spacecraft in 1969 (Table 17.1), which seemed to confirm Mars’ lunar-like characteristics. However, our perception of Mars changed dramatically in 1972 when systematic mapping by the Mariner 9 orbiter spacecraft revealed the planet that we know today. As mapping progressed, huge volcanoes, deep canyons, enormous dry riverbeds, and extensive dune fields came into view and a complex geologic history became apparent. Exploration of Mars continued in the 1970s as both the USSR and the United States sent landers to the surface and other vehicles to the planet. Exploration in the 1970s culminated with the Viking mission, which successfully placed two landers on the surface and two other spacecraft into orbit. By the end of the Viking mission, almost all the surface had been photographed from orbit at a resolution of about 250 m/pixel and small fractions with resolutions as good as 10 m/pixel. In addition, the Viking landers had carried out a variety of experiments directed mostly toward detecting life and understanding the chemistry of the soil.

In the early 1980s, our understanding of Mars was further enhanced when it became clear that we had samples of Mars in our meteorite collections here on the Earth. A group of meteorites, called SNCs (pronounced snicks and standing for “Shergottites, Nahklites, and Chassigny”) were initially suspected to be of Martian origin because they were basaltic and had ages of ~1.3 billion years. These meteorites could not have come from the Earth because their oxygen isotope ratios are distinctly different from terrestrial ratios. The only plausible body that could have been volcanically active so recently and supplied the meteorites was Mars. Martian origin was later confirmed by

<table>
<thead>
<tr>
<th>Mission</th>
<th>Nation</th>
<th>Launch Date</th>
<th>Fate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mariner 4</td>
<td>US</td>
<td>November 18, 1964</td>
<td>Flew by July 15, 1965; first close-up images</td>
</tr>
<tr>
<td>Mariner 6</td>
<td>US</td>
<td>February 24, 1969</td>
<td>Flew by July 31, 1969; imaging and other data</td>
</tr>
<tr>
<td>Mariner 7</td>
<td>US</td>
<td>March 27, 1969</td>
<td>Flew by August 5, 1969; imaging and other data</td>
</tr>
<tr>
<td>Mars 3</td>
<td>USSR</td>
<td>July 25, 1973</td>
<td>Into orbit February 12, 1974; imaged surface</td>
</tr>
<tr>
<td>Viking 1</td>
<td>US</td>
<td>August 20, 1975</td>
<td>Landed on surface July 20, 1976; orbiter mapping</td>
</tr>
<tr>
<td>Viking 2</td>
<td>US</td>
<td>September 9, 1975</td>
<td>Landed on surface September 3, 1976; orbiter mapping</td>
</tr>
<tr>
<td>Phobos2</td>
<td>USSR</td>
<td>July 12, 1988</td>
<td>Mars and Phobos remote sensing</td>
</tr>
<tr>
<td>Pathfinder</td>
<td>US</td>
<td>December 4, 1996</td>
<td>Landed July 4, 1997; lander and rover data</td>
</tr>
<tr>
<td>Global Surveyor</td>
<td>US</td>
<td>November 7, 1996</td>
<td>Into orbit September 11, 1997; imaging and other data</td>
</tr>
<tr>
<td>Mars Odyssey</td>
<td>US</td>
<td>April 7, 2001</td>
<td>In orbit October 24, 2001; imaging, remote sensing</td>
</tr>
<tr>
<td>Spirit Rover</td>
<td>US</td>
<td>June 10, 2003</td>
<td>Landed in Gusev January 3, 2004</td>
</tr>
<tr>
<td>Opportunity Rover</td>
<td>US</td>
<td>July 7, 2003</td>
<td>Landed in Meridiani January 24, 2004</td>
</tr>
<tr>
<td>Mars Express</td>
<td>Europe</td>
<td>June 2, 2003</td>
<td>In orbit December 25, 2003; imaging, remote sensing</td>
</tr>
<tr>
<td>Mars Reconnaissance Orbiter</td>
<td>US</td>
<td>August 12, 2005</td>
<td>In orbit March 10, 2006; imaging, remote sensing</td>
</tr>
<tr>
<td>Curiosity Rover</td>
<td>US</td>
<td>November 26, 2011</td>
<td>Landed in Gale Crater August 6, 2012</td>
</tr>
<tr>
<td>MAVEN</td>
<td>US</td>
<td>November 18, 2013</td>
<td>Mars orbiter, en route</td>
</tr>
</tbody>
</table>
finding gasses trapped within the meteorites that are identical in composition to gasses in the Martian atmosphere as measured by the Viking landers. The meteorites are believed to have been ejected from Mars by large impacts and subsequently captured by the Earth after spending several million years in space. We have since added to the collection, and there are now about 50 known Martian meteorites.

All these meteorites are basaltic and all but one are aged significantly less than the presumed ~4.6 Gyr age of the planet. The one exception known so far, ALH84001, has a crystallization age of ~4 Gyr. In 1996, it was tentatively suggested that carbonate globules within this meteorite, together with some disequilibrium mineral assemblages, polycyclic aromatic hydrocarbons and a number of different types of very small segmented rods that resemble some terrestrial nanofossils, might all be the result of biologic activity. This suggestion has, however, received little support from subsequent investigations by the general science community.

More recently, the exploration of Mars has continued with a series of landers (Pathfinder, Phoenix), rovers (Spirit, Opportunity, Curiosity) and orbiters (Mars Global Surveyor, Mars Odyssey, Mars Express, and Mars Reconnaissance Orbiter), as listed in Table 17.1. These missions have documented the mineralogical diversity of the planet; mapped the global chemistry and topography, the distribution of water ice, and the strength of the remanent magnet field; as well as provided higher resolution imaging. The scientific analyses of these data sets has confirmed the active role that water has played in the planet’s evolution and has documented numerous ongoing changes on the surface, some possibly related to water action.

2. GENERAL CHARACTERISTICS

2.1. Orbital and Rotational Constants

The Martian day is almost the same as the Earth’s day, but the year is almost twice as long (Table 17.2). Because its rotational axis is inclined to the orbit plane, Mars, like the Earth, has seasons. But the Martian orbit has significant eccentricity, causing the pole that is tilted toward the Sun at perihelion to have warmer summers than the other pole. At present, the south has the warmer summers, but, because of a slow change in the direction of tilt of the rotational axis and a slow change in the orientation of perihelion, the hot and cold poles change on a 51,000-year cycle. The eccentricity also causes the seasons to have significantly different lengths (see Table 17.2).

### Table 17.2 Earth and Mars: General Characteristics Compared

<table>
<thead>
<tr>
<th></th>
<th>Earth</th>
<th>Mars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean equatorial radius (km)</td>
<td>6378</td>
<td>3396</td>
</tr>
<tr>
<td>Mass (×10²⁴ km³)</td>
<td>5.98</td>
<td>0.624</td>
</tr>
<tr>
<td>Mean distance from the Sun (10⁶ km)</td>
<td>150</td>
<td>228</td>
</tr>
<tr>
<td>Orbit eccentricity</td>
<td>0.017</td>
<td>0.093</td>
</tr>
<tr>
<td>Obliquity</td>
<td>23.5°</td>
<td>25.2°</td>
</tr>
<tr>
<td>Length of day</td>
<td>24 h</td>
<td>24 h, 39 m, 35 s</td>
</tr>
<tr>
<td>Length of year (Earth days)</td>
<td>365.3</td>
<td>686.9</td>
</tr>
<tr>
<td>Seasons (Earth days)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northern spring</td>
<td>92.9</td>
<td>199</td>
</tr>
<tr>
<td>Northern summer</td>
<td>93.6</td>
<td>183</td>
</tr>
<tr>
<td>Northern fall</td>
<td>89.7</td>
<td>147</td>
</tr>
<tr>
<td>Northern winter</td>
<td>89.1</td>
<td>158</td>
</tr>
<tr>
<td>Atmosphere</td>
<td>78% N₂, 21% O₂, 1% Ar</td>
<td>95% CO₂, 3% N₂, 2% Ar</td>
</tr>
<tr>
<td>Surface pressure (mbar)</td>
<td>1013</td>
<td>7</td>
</tr>
<tr>
<td>Mean surface temperature (K)</td>
<td>288</td>
<td>215</td>
</tr>
<tr>
<td>Surface gravitational acceleration (cm/s²)</td>
<td>981</td>
<td>371</td>
</tr>
<tr>
<td>Moons</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 17.2). At present, the Martian obliquity is similar to that of the Earth. Yet while the Earth experiences only minor changes in its obliquity, the obliquity of Mars changes significantly, and relatively rapidly, over long timescales. For example, during the past 10 Myr it has been as low as 15° and as high as 45°. It has been estimated that there is a 63% probability that the obliquity reached 60° in the past 1 Gyr. At low obliquities the atmosphere thins because most of the CO$_2$ in the atmosphere condenses out onto the poles. At high obliquities the water ice polar caps dissipate in summer and ice condenses at lower latitudes.

2.2. Surface Conditions

Mars has a thin atmosphere that provides almost no thermal blanketing. As a result, temperatures at the surface have a wide diurnal range, controlled largely by latitude, the reflectivity of the surface, and the thermal properties of the surface materials. Typically, surface temperatures in summer at latitudes ±60° range from 180 K at night to 290 K at midday, but can range more widely if the surface consists of unusually low-density, fine-grained material. However, these temperatures are somewhat deceiving because at depths of just a few centimeters below the surface, temperatures at the equator are at the diurnal mean of 210–220 K. At the poles in winter, temperatures drop to 150 K, at which point CO$_2$ condenses out of the atmosphere to form a decimeter- to meter-thick dry ice seasonal cap. The atmospheric pressure at the surface ranges from about 14 mbars at the bottom of the Hellas basin to about 3 mbars at the top of the tallest volcanoes, and it changes seasonally as a result of the formation and sublimation of the polar caps. Winds typically have a velocity of a few meters per second but there may be gusts up to 50 m/s. Dust devils and local dust storms are common, and almost every year, regional- or global-scale dust storms can occur.

The stability of water is of profound importance for understanding Martian geology. Under the conditions just described, the planet has a thick permafrost layer that extends a few kilometers deep at the equator and several kilometers deep at the poles. Any unbound water present will exist as ice in this zone. There may be liquid water beneath the permafrost, depending on the magnitude of the planet’s largely unknown geothermal heat flux and other subsurface thermophysical properties. Water ice caps roughly 3 km thick are present at both poles, although at the South Pole the water ice cap is largely masked by a remnant summer CO$_2$ cap. At latitudes between about 40° and the edge of the water ice cap, abundant ice has been detected just below a dehydrated zone a few tens of centimeters thick. At latitudes less than about 40°, ice is unstable at all depths (i.e. a block of pure water ice placed in the ground at these latitudes will slowly sublime into the atmosphere). The small amounts of water that have been detected at low latitudes may be water bound in minerals or water inherited from an earlier era of higher obliquity when water ice was stable at these latitudes. Under present conditions, liquid water, although unstable near the surface, can exist transiently, particularly if very salty.

2.3. Planet Formation and Global Structure

Like the other planets, Mars formed from materials that condensed out of the early solar nebula, a disk of gas and dust that surrounded the early Sun. A class of meteorites called carbonaceous chondrites, which are almost identical in composition to the photosphere of the Sun, are believed to closely resemble the early nebula in composition. Radioisotopes date the formation of the nebula at 4.567 billion years ago. The planets formed as the dust and gas accumulated into discrete bodies, and gravitational attraction favored growth of larger bodies over smaller bodies. Mars appears to have formed remarkably quickly based on evidence from short-lived radioisotopes. The high rate of accretion resulted in global melting, which enabled settling of heavy iron-rich melts to the center of the planet to form a core separated from the silicate-rich mantle. During this differentiation process, siderophile elements (those which dissolve preferentially in iron-rich melts over coexisting silicate-rich melts) became depleted in the mantle and enriched in the core. This enables formation of the core to be dated because the daughter products of some short-lived, strongly siderophile elements are present in the mantle, as indicated by the composition of Martian meteorites. For example, $^{182}$Hf decays to $^{182}$W with a half-life of 9 Myr. W is highly siderophilic so should mostly enter the core, yet there is an excess of $^{182}$W in the mantle, implying that not all the Hf had decayed before the core formed. This and other isotopic evidence indicates that the core formed within 20 Myr of the formation of the elements that comprise the solar system. Isotopic evidence indicates that some crust formed very early but that new crust continued to form through Mars’ history, as indicated by volcanoes and extensive volcanic plains.

The Earth’s core is inferred to be iron-rich from (1) the core’s density as deduced from the core’s size and the planet’s moment of inertia, (2) computer models of the bulk composition of the Earth and compared with the composition of the chondritic meteorites from which the Earth formed, and (3) depletion of siderophile elements in mantle-derived rocks as compared with chondritic meteorites. Similar reasoning can be used for Mars except that, while the size of the Earth’s core is accurately known from seismic data, the size of Mars’ core must be inferred indirectly. The best estimate is that the core radius is between 1300 and
1500 km. In addition, the Martian core may be more sulfur rich than the Earth’s core because Mars’ mantle is more depleted of chalcophile elements (those that preferentially dissolve in sulfur-rich melts) than is the Earth’s.

One of the more surprising results of the Mars Global Surveyor mission was the discovery of large magnetic anomalies in the crust despite the absence of a global planetary magnetic field today. Their presence indicates that Mars had a magnetic field in the past, but that it switched off at some time. The size of the anomalies suggests that they must result from sources in the outer few tens to several tens of kilometers of the crust and that their magnetizations are higher by an order of magnitude than magnetizations typically encountered in terrestrial rocks. The anomalies probably formed when rocks, containing magnetically susceptible Fe-bearing minerals, crystallized in the presence of a strong magnetic field. Most of the anomalies, and all of the largest ones, are in the southern highlands. They are particularly prominent on either side of the 180° longitude where there are several broad, east—west stripes. One interpretation of the linear anomalies is that they result from injection of dikes or dike swarms several tens of kilometers wide and hundreds of kilometers long in the presence of a strong magnetic field. Anomalies are mostly absent around the youngest large-impact basins (Utopia, Hellas, Isidis, and Argyre). The simplest explanation is that there was no longer a magnetic field when these basins formed, formation of the basins melted and reset/destroyed any preexisting anomalies, and no new ones formed when the affected materials cooled after the basin-forming events. The ages of the basins are not known, but, by analogy with the Moon, they are likely to have formed toward the end of the so-called late heavy bombardment period in solar system formation history, around 3.8—4 Gyr ago. Thus, the magnetic field may have turned off by around 4 Gyr ago.

The Earth’s magnetic field is generated by convection within its core. Mars’ early dynamo probably had a similar cause. Possible causes for cessation of the dynamo are loss of core heat, solidification of most of the core, and/or changes in the mantle convection regime. Magnetization of minerals within 3.9- to 4.1-Gyr-old carbonates in the Martian meteorite ALH84001 suggests that there was still a magnetic field at this time. If true, it implies that Mars had a magnetic field for the first 500 My of its history and that the field turned off around 4 Gyr ago, just before formation of the youngest impact basins.

Like the Earth’s mantle, Mars’ mantle is chondritic in composition except for the depletion of siderophile and chalcophile elements as noted above and depletion of volatile elements which would have been largely lost from the interior during the early global melting phase. It consists mainly of iron—magnesium silicates. One difference between the mantles of the two planets that is suggested by the compositions of the Martian meteorites is that the Fe/Mg ratio is higher in the Martian mantle.

The crust of Mars appears to be essentially a melt extract from the mantle and is mostly basaltic in composition. The thickness of the crust varies considerably, ranging from 5 to 100 km, as estimated from the relations between the global gravity field and the global topography. The thickest crust is under the high-standing cratered terrain in the southern hemisphere, whereas the thinnest is under the large impact basins of Isidis and Hellas.

2.4. Global Topography and Physiography

The topography and physiography of Mars have a marked north—south asymmetry, which is referred to as the global dichotomy (See Figure 17.1). The dichotomy is expressed in three ways: as a change in elevation, a change in crustal thickness, and a change in crater density. The southern uplands have an average elevation 5.5 km higher than the northern plains, the crust is roughly 25 km thicker in the uplands, and most of the upland terrain is heavily cratered, dating back to the period of the late heavy bombardment. The northern plains are mostly younger, but remnants of buried craters poke through the surface indicating that there is an older surface at some depth beneath the younger plains. The low-lying plains constitute roughly one-third of the planet and are mostly in the north. The cause of the dichotomy is not known. Suggestions include a very large impact, soon after the planet formed, or internal convection sweeping most of the light, crustal material into one half of the planet.

Superimposed on the global dichotomy is the Tharsis bulge, over 5000 km across and 10 km high, centered on the equator at 260 °E. Most of the planet’s volcanic activity has been centered on the bulge, which has the five largest volcanoes (Montes Olympus, Alba, Arsia, Ascreus, and Pavonis) on its northwest flank. Tharsis is also at the center of a vast array of radial faults and circumferential ridges that affect over half the planet’s surface. The bulge, an accumulation of $3 \times 10^8$ km$^3$ of volcanic rocks, deformed the lithosphere, thereby affecting slopes around the bulge. Ancient valley networks incised into these slopes indicate that most of the bulge formed very early in the planet’s history. To the east of the center of the bulge are a series of vast canyons thousands of kilometers long and up to 10 km deep. They are roughly radial to the bulge and appear to have formed largely by faulting, although they also have been extensively modified by fluvial and mass-wasting processes. At the east end of the canyons, extensive areas of terrain have seemingly collapsed to form chaotic terrain from which emerge large dry river beds that extend for thousands of kilometers downslope into the northern plains. A much smaller bulge centered in Elysium at 25 °N, 213 °W has also been a center of volcanic, tectonic, and
fluvial activity. Other prominent topographic features are large impact basins; the largest are Hellas (2600 km diameter), Isidis (1600 km), and Argyre (1500 km).

The physiography of the poles is distinctively different from that of the rest of the planet. At each pole, extending out to the 80° latitude circle, is a stack of finely layered deposits a few kilometers thick. In the north they rest on plains; in the south they rest on cratered uplands. The small number of superimposed impact craters suggests that these sediments are only a few tens of millions of years old.

3. IMPACT CRATERING

3.1. Cratering Rates and the Martian Timescale

All solid bodies in the solar system are subject to impact by asteroidal and cometary debris. Current cratering rates are low. On Earth, in an area the size of the United States, a crater larger than 10 km across is expected to form every 10–20 million years and one larger than 100 km across every billion years. The rates on the other terrestrial planets are likely to be within a factor of two or three of these rates. On the Moon, surfaces are either densely covered by large craters (lunar highlands) or sparsely affected by large craters (maria) with no surfaces of intermediate crater densities. This contrast arises because of the Moon’s cratering history. Very early, cratering rates were much higher, but around 3.7 billion years ago they declined rapidly to roughly the present rate. Accordingly, surfaces that formed more than 3.7 billion years ago are heavily cratered, and those that formed afterward are much less cratered. Mars has had a similar cratering history, hence the contrast between the heavily cratered uplands and the sparsely cratered plains.

Craters provide a means of estimating the ages of surfaces. The solar system’s most densely cratered surfaces formed more than 3.7 billion years ago, and the cratering rate has been roughly constant since that time. Consequently, a 3 billion-year-old surface will have roughly three times more craters on it than a 1 billion-year-old surface. However, there is considerable uncertainty in estimating absolute ages in this way because we do not know exactly what the cratering rate on Mars has been for the past few billion years. Nevertheless, by counting craters over key geologic areas, we can put surfaces in a time-ordered relative age sequence and perhaps even make rough estimates of their absolute ages. Craters have thus been used to divide the history of Mars into different epochs. The Noachian refers to the period of heavy bombardment that ended around 3.7 billion years ago. The rest of the planet’s history is divided into the Hesperian, roughly 3.7 to 3.0 billion years ago, and the Amazonian, roughly 3.0 billion years ago to the present. The Noachian period is characterized by high cratering rates, formation of valley networks, and the presence of hydrated minerals such as...
phyllosilicates. The Hesperian period is characterized by large outflow channel floods and extensive lava plains and the presence of abundant sulfate deposits. During the Amazonian, most of the processes that occurred earlier continued, but at much lower rates, enabling less-energetic processes such as wind erosion to exert a strong influence on the preserved landforms.

3.2. Crater Morphology

Impact craters have similar morphologies on different planets. Small craters are simply bowl-shaped depressions with constant depth-to-diameter ratios. With increasing size, the craters become more complex as central peaks appear, terraces form on the walls, and the depth-to-diameter ratio decreases. At very large diameters, the craters become multiringed, and it is not clear which ring is the equivalent of the crater rim of smaller craters. On Mars, the transition from simple to complex takes place at 6–7 km, and the transition from complex craters to multiringed basins takes place at 130–150 km diameter.

Although impact craters on Mars resemble those on the Moon, the patterns of ejecta are quite different. Lunar craters generally have continuous hummocky ejecta near the rim crest, outside of which is a zone of radial or concentric ridges, which merge outward into string or loops of secondary craters, formed by material thrown out of the main crater. In contrast, the ejecta around most fresh-appearing Martian craters, especially those in the 5–100 km size range, occur in discrete, clearly outlined lobes (See Figure 17.2). Various patterns are observed. The ejecta around craters smaller than 15 km in diameter are enclosed in a single, continuous lobate ridge or rampart, situated about one crater diameter from the rim. Around larger craters, there may be many lobes, some superimposed on others, but all surrounded by a rampart. The distinctive Martian ejecta patterns have been attributed to two possible causes. The first suggestion, based on experimental craters formed under low atmospheric pressures, is that the patterns are formed by interaction of the ejecta with the atmosphere. The second is that the ejecta contained water and had a mudlike consistency and so continued to flow along the ground after ejection from the crater and ballistic deposition. This view is supported by the resemblance of Martian craters to those produced by impacts into mud.

The previous discussion refers to fresh-appearing craters. Erosion rates at low latitudes for most of Martian history are very low—typically 0.1–10 nm/year, although rates may be higher locally. However, early in the planet’s history erosion rates were much higher. As a consequence, in the cratered uplands, craters range in morphology from fresh-appearing craters to barely discernible, rimless depressions. In contrast, on volcanic plains in equatorial regions, almost all the craters are fresh-appearing, even though they may be billions of years old. Obliteration rates have been higher at high latitudes. This has been attributed to ice-abetted creep of the near-surface materials, but other factors such as repeated burial and removal of possibly ice-rich materials by sublimation and the wind, may have contributed to modification of the craters. Such processes have been invoked to explain the so-called pedestal craters that are particularly common at high latitudes. These craters are inset into a platform or pedestal that has the same or a slightly larger aerial extent as the ejecta. The simplest explanation is that the region in which these craters are found was formerly covered with a layer of loose material or ice that has since been removed except around craters where the surface was protected by the ejecta.

4. VOLCANISM

Mars has had a long and varied volcanic history. Crystalization ages of Martian meteorites as young as 150 million years, and the scarcity of impact craters on some volcanic surfaces, suggest that the planet is still volcanically active, although the rates must be very low compared with the Earth. The tectonic framework within which Martian volcanism occurs is very different from that in which most volcanism occurs on the Earth. Most terrestrial volcanism takes place at plate boundaries, but these have no Martian
equivalents, there being no plate tectonics on Mars. Perhaps the closest terrestrial analogs to Martian volcanoes are those, such as the Hawaiian volcanoes, that occur within plates rather than on the boundaries. Most Martian volcanism is **basaltic**, but basaltic volcanism expresses itself somewhat differently on Mars because of the lower heat flow, gravity, and atmospheric pressure. Eruptions are expected to be larger and less frequent and more likely to produce ash, and ash clouds are more likely to collapse and produce ash-rich surface flows.

The large **shield volcanoes** of Tharsis and Elysium present the most spectacular evidence of volcanism (See Figure 17.3). Shield volcanoes, such as those in Hawaii, are broad domes with shallow sloping flanks that form mainly by eruption of fluid basaltic lava. Each has a summit depression formed by collapse following eruptions on the volcano flanks or at the summit. In contrast, stratovolcanoes such as Mt Fujiyama, tend to be much smaller and have steeper flanks and a summit depression that is a true volcanic vent. Explosive, ash-rich eruptions tend to be more common in the building of a stratovolcano and the lava tends to be more volatile rich, more siliceous, and more viscous than that which forms shields. In Tharsis, three large shield volcanoes form a northeast—southwest trending line and 1500 km to the northwest of the line stands the largest shield of all, Olympus Mons, 550 km across and reaching a height of 21 km above the Mars datum. The three aligned Tharsis Montes shields are only slightly smaller. Olympus Mons has a summit caldera 80 km across and the flanks have a fine striated pattern caused by long linear flows, some with central channels. The main edifice is surrounded by a cliff, in places 8 km high. Outside the main edifice is the aureole that consists of several huge lobes with a distinctively ridged texture. The lobes are thought to have formed as a result of successive collapses of the periphery of a previously much larger Olympus Mons. The collapses, which could have been catastrophic or gradual, left a cliff around the main edifice. The largest lobe has roughly the same area as France. The edifice is thought to have been built slowly over billions of years by large eruptions, widely spaced in time, and fed from a large magma chamber within the edifice that was itself fed by a magma source deep within the mantle. Although huge, Olympus Mons is not the largest volcano in aerial extent. Alba Patera, at the north end of Tharsis is $2000 \times 3000$ km across, almost the size of the United States. The large size of the Martian shields results partly from the lack of plate tectonics. The largest shield volcanoes on the Earth, those in Hawaii, are relatively short lived. They sit on the Pacific plate, and the source of the lava is below the rigid plate. As a Hawaiian volcano grows, movement of the Pacific plate carries it away from the lava source, so it becomes extinct within a few 100,000 years. A trail of extinct volcanoes across the Pacific attests to the long-term supply of magma from the mantle source presently below Hawaii. On Mars, a volcano remains stationary and will continue to grow as long as magma continues to be supplied, so the volcanoes are correspondingly larger.

The Elysium province is much smaller than Tharsis, having only three sizeable volcanoes. One unique attribute of the Elysium province is the array of large channels that start in **graben** around the volcanoes and extend thousands of kilometers to the northwest. They may have been formed by dikes injected into ice-rich frozen ground. Other volcanoes occur near Hellas and in the cratered uplands. Not all the volcanoes are formed by fluid lava. Some appear to be surrounded by extensive ash deposits and some have

**FIGURE 17.3** View looking southeast across Tharsis. Olympus Mons, in the foreground, is 550 km across; 21.2 km high; and surrounded by a cliff 8 km high. Lobes of the aureole can be seen extending from the base of the cliff; 10 vertical exaggeration (MOLA).
densely dissected flanks as though they were composed of easily erodible materials such as ash.

Lava plains may constitute the bulk of the planet’s volcanic products. There are several kinds of volcanic plains. On some plains, found mostly between the volcanoes in Tharsis and Elysium, volcanic flows are clearly visible. On others, mostly found around the periphery of Tharsis and in isolated patches in the cratered uplands, ridges are common but flows are rare. Others with numerous low cones may have formed when lava flowed over water-rich sediments. Finally, some young, level plains, such as those in Cerberus, estimated to be only a few million years old, appear to consist of thin plates that have been pulled apart, for they can be reconstructed like a jigsaw puzzle. The plates may indicate rafting of pieces of crust on a lava lake. Geomorphic evidence of volcanic activity that occurred prior to the end of the late heavy bombardment has been largely destroyed by the effects of impacts.

5. TECTONICS

Most of the deformation of the Earth’s surface results from the movement of the large lithospheric plates with respect to one another. Linear mountain chains, transcurrent fault zones, rift systems, and oceanic trenches all result directly from plate tectonics. There are no plate tectonics on Mars, so most of the deformational features familiar to us here on the Earth are absent. The tectonics of Mars are instead dominated by the Tharsis bulge. The enormous pile of volcanics that constitute the Tharsis bulge has stressed the lithosphere and caused it to flex under the load. Modeling suggests that around the bulge tensional stresses should be circumferential and compressional stresses should be radial. This is entirely consistent with what is observed.

The bulge is surrounded by arrays of radial, tensional fractures and circumferential compressional ridges. Some of the tensional fractures, particularly those to the southwest of the bulge, extend for several thousand kilometers. Development of some of the fractures may have been accompanied by emplacement of dikes. The fractures clearly started to form very early in the planet’s history, since many of the young lava plains are only sparsely fractured, whereas the underlying plains, visible in windows through the younger plains, may be heavily fractured.

Not all the deformational features result from the Tharsis load, however. Ridges, suggestive of compression, are common on ridged plains, such as Hesperia Planum and Syrtis Major, which are far removed from Tharsis. Some arcuate faults around Isidis and Hellas clearly result from the presence of the large basins. Circular fractures around large volcanoes, such as Elysium Mons and Ascleps Mons, have formed as a result of bending of the lithosphere under the volcano’s load. Finally, large areas of the northern plains are cut by fractures that form polygonal patterns at a variety of scales. Polygonal fracture patterns are common in the terrestrial arctic where they form as a result of seasonal contraction and expansion of ice-rich permafrost. Some of the polygonal patterns on Mars, those with polygons up to a few tens of meters across, may have also formed in this way. However, polygons that are several kilometers across are unlikely to have formed in this way and may be the result of regional warping of the surface. Despite these examples, the variety of deformation features is rather sparse compared with the Earth because of the lack of plate tectonics. In particular, folded rocks, although present, are rare.

6. CANYONS

On the eastern flanks of the Tharsis bulge is a vast system of interconnected canyons (See Figure 17.4). They extend just south of the equator from Noctis Labyrinthus at the crest of the Tharsis bulge eastward for about 4000 km until they merge with some large channels and chaotic terrain. The characteristics of the canyons change from east to west. Noctis Labyrinthus at the western end consists of numerous intersecting closed, linear depressions. The depressions are generally aligned with faults in the surrounding plateau. Further east the depressions become deeper, wider, and more continuous to form roughly east—west trending canyons. Still further east the canyons become shallower; fluvial features become more common; both on the canyon floor and on the surrounding plateau; and finally, the canyons end as the canyon walls merge into walls containing chaotic terrain and evidence of floods. The canyons almost certainly formed largely by faulting and not by fluvial erosion, as is the case with the Grand Canyon in Arizona.

FIGURE 17.4 The middle section of the canyons. In the upper left is the completely enclosed Hebes Chasma, within which a mound of layered, sulfate-bearing sediments has been detected. The main part of the canyon consists of three parallel canyons each 200 km across, also partly filled with mounds of sulfate-bearing sediments. Some of the sediments may have been deposited in lakes which drained catastrophically to the east (MOLA and THEMIS).
Faulting is indicated by the partial merger of numerous closed depressions in the western end of the canyons and by straight walls in the east. While faulting created the initial relief, the canyons have been subsequently enlarged by failure of the walls in huge landslides and by fluvial action. The faulting was on such an enormous scale that it probably involved the entire lithosphere.

Thick sequences of layered (and unlayered) deposits are present in many places throughout the canyons, including some closed canyons completely isolated from the main depression. One possibility is that the canyons formerly contained lakes and that the layered sediments were deposited in these lakes. The lakes drained to the east, hence the continuity eastward from the canyons into several large flood channels. Orbital detection of sulfates within the canyons and some of the outlying depressions supports the lake hypothesis. If climatic conditions were similar to the present, such lakes would have frozen over, although the lake beneath the ice could have been sustained for extended periods if fed by groundwater. While the lake hypothesis is plausible, there are many unanswered questions, such as: Were the lakes fed mostly by surface runoff or groundwater? Where did the sediment in the layered deposits come from? What caused the layering?

7. WATER

Water-formed features present some of the most puzzling problems of Martian geology. Valley networks likely formed when the climate was significantly warmer than at present, yet how the climate might have changed is unclear. Huge floods have episodically moved across the surface, yet there is little trace left of the vast amounts of water that must have been involved, and gullies are forming on steep slopes during the present epoch despite the cold conditions. Perhaps most puzzling of all is the question of whether there were ever oceans present, and if so how big they were; when did they form; and where did all the water go?

7.1. Erosion and Weathering

All Noachian terrain has undergone extensive erosion, whereas younger terrains are only barely eroded for the most part. It thus appears that there were continuously high or episodically high rates of erosion during the Noachian, that the rates fell precipitously at the end of the period, around 3.7 Gyrs ago, and that then the rates have remained low for the rest of the planet’s history. Average erosion rates during the Noachian are estimated to be around \(10^{-6} - 10^{-5}\) m/yr, at the low end of continental denudation rates on Earth.

Weathering rates may also have been higher during the Noachian. Hydrated secondary silicate minerals occur throughout the Noachian terrains but are generally not found in younger terrains. A strong possibility, consistent with the presence of valley networks, is that warm and wet conditions at the surface resulted in weathering of primary igneous rocks to produce the hydrated minerals (mainly clays) that we see. An alternative explanation is that the hydrous minerals were formed at depths below the surface and were brought to the surface as a result of excavation by impacts.

7.2. Valley Networks, Lakes, Deltas

Much of the ancient cratered uplands are dissected by branching valley networks that in plan resemble terrestrial river valleys (See Figure 17.5). However, although all the uplands are highly eroded, not all are densely dissected. The valleys mostly have rectangular to U-shaped cross-sections, are a few kilometer across, 50\(–\)300 m deep, and tens to hundreds of kilometers long, although a few extend for thousands of kilometers (See Figures 17.6 and 17.7). Drainage densities vary considerably by location but many areas have such high densities that precipitation and surface runoff is implied. Precipitation could have been rain or snow; in either case, climatic conditions significantly warmer than the present are needed. The distribution of the valleys, their excellent preservation, and the poor development of drainage basins suggest that there was a period of enhanced incision at the end of the Noachian. The extent to which valley formation contributed to the universal degradation of the earlier Noachian terrain remains unclear.

Lakes were likely common throughout the poorly graded Noachian terrain while it was undergoing fluvial erosion. Most valleys terminate in closed depressions such as craters or low areas between craters, where at least transient, closed lakes almost certainly formed. Many of these areas are underlain by seemingly fine-grained, horizontally layered, easily erodible sediments. Possible chlorine-bearing (salt) deposits and sulfates in local depressions within the Noachian uplands may be the result of evaporation from such lakes. Many depressions in the highlands have both inlet and outlet valleys indicating the former presence of a lake that overflowed, as might be expected from a flood event. Other lakes have no outlets indicating that infiltration and evaporation kept pace with runoff supply. Comparisons of lake volumes with drainage basin areas suggest that most lakes formed by modest-sized fluvial events spread over extended periods.

Most valleys terminate at grade in local depressions with little or no deposits at their mouths, which suggests that in most cases the materials eroded to form the valleys were either too fine grained to form a delta or were distributed across the depression to form alluvial fans. If a lake was present, its level may have fluctuated. An alternative explanation is that the hydrous minerals were formed at depths below the surface and were brought to the surface as a result of excavation by impacts.
which would have resulted when the lake level dropped, indicates that fluvial activity terminated abruptly. The valley and delta dimensions, coupled with lake volumes and drainage basin areas suggest that the fluvial episode at the end of the Noachian involved peak discharges comparable to terrestrial rivers. Failure to overflow most closed basins implies that the valley was not formed by a few deluge events but by intermittent modest-sized events spread over an extended period.

No satisfactory explanation has been proposed for how early Mars could have been warmed to allow precipitation and stream flow. The output of the Sun is thought to have been perhaps as much as 20–30% less than at present during this early era. Greenhouse models suggest that even a very thick CO₂–H₂O atmosphere could not warm the surface enough. The possible role of other greenhouse gases is being explored. One possibility is that large
impacts injected massive amounts of water into the atmosphere, which then precipitated out as hot acid rain. The idea is attractive in that it might explain why the valley networks formed mainly in the old terrains when impact rates were high but it does not appear to be consistent with the pattern of precipitation implied by the geometry of the fluvial features. Other possibilities include massive injection of greenhouse gases during volcanic eruptions, or melting of snow deposited at low latitudes during high obliquity periods.

A major uncertainty is whether large bodies of water were present while the valley networks were forming. If rainfall was involved in their formation, then large bodies of water must have been present as a source for the precipitation. But if the valleys formed by melting of snow during periods of high obliquities, then ice at the poles could have been the source. Observational evidence for ocean-sized bodies of water remains ambiguous, particularly for this early era.

### 7.3. Outflow Channels

**Outflow channels** are very different from valley networks. They are tens of kilometers wide, thousands of kilometers long, have streamlined walls and scoured floors, and contain teardrop-shaped islands (See Figures 17.9 and 17.10). Most start full size and have few if any tributaries. They closely resemble large terrestrial flood features and have almost universally been accepted to be the result of massive floods. Most start around the Chryse basin, emerging either from the canyons or from closed rubble-filled depressions, and extending northward for thousands of kilometers until all traces are lost in the northern plains. While the largest flood features are in the Chryse region, others occur in Elysium, Hellas, and elsewhere, commonly starting at faults. As already indicated above, the channels that merge with the canyons may have formed by catastrophic drainage of lakes within the canyons. Other outflow channels appear to have formed by massive eruptions of groundwater that may have been stored under pressure beneath kilometers-thick permafrost. Release occurred when the permafrost seal was broken such as by impact, volcanic activity, or faulting. Most of the outflow channels formed in the Hesperian period (middle Mars history), well after the time of formation of most of the valley networks, but some formed much more recently. Cold surface conditions and thick permafrost were probably required for their formation.

Major issues are how much water was involved and where did it all go. The size of the channels suggests that the discharges were enormous: 1000 to 10,000 times the discharge of the Mississippi. But we do not know how long the floods lasted and so do not know the total volume of each flood. Nevertheless, large bodies of water, or seas, must have been left in low-lying areas when the floods were over. Efforts to find evidence for these seas has led to mixed
results. Some researchers claim that Mars must have had oceans as extensive as those on the Earth; others claim that seas larger than the Mediterranean were unlikely. Under the present conditions, such seas would have frozen and the ice would have slowly sublimed, thereby adding to the ice at the poles. However, estimates of the amount of water currently in the polar ice caps falls far short of even the lowest estimates of the amounts of water involved in the floods, so a mystery remains as to where the water went. Most likely, it is in the subsurface as either ground ice or groundwater.

7.4. Gullies

Gullies are by far the most common fluvial feature that have formed in the past few billion years of Martian history (See Figure 17.11). They are common on steep slopes in the 30°—60° latitude belts with a preference for poleward-facing slopes. They typically consist of an upper theater-shaped alcove that tapers downslope to converge on a channel that extends further downslope to terminate in a debris fan. The channels are mostly several meters wide and hundreds of meters long. They appear to be forming today. Their origin is controversial. Although initially attributed to groundwater seeps, this origin now seems unlikely given the probable thick cryosphere during the second half of Mars’ history and the common presence of gullies at locations where groundwater is unlikely, as at crater rim crests and on slopes around mesas and central peaks. Dry mass-wasting may contribute to their formation but this also seems to be an unlikely cause since many of the gullies cut through bedrock ledges. All the morphologic attributes are consistent with water erosion, and the broad consensus is that this is their cause. One possibility, consistent with their preference for pole-facing slopes, is that they formed mainly by summer melting of snow during periods of high obliquity. But that does not explain the changes that are occurring today. Another possibility may be slope failure aided by the volatilization of interstitial H₂O or CO₂ ice when heated by the summer Sun.

7.5. Dark Streaks

Dark streaks occur on many slopes at low latitudes in dust-covered areas (See Figure 17.12). They are forming today, are dark when first formed, brighten with age, and ultimately disappear over decadal timescales. Although they likely
result from slope failure, precisely how they form is uncertain. Hypotheses proposed for their formation include subsurface liquid water seepage and flow, disturbance and flow of surface materials as a result of devolatilization of ground ice by solar illumination, and dry granular flow of dust- or sand-sized grains initiated by rockfalls resulting from nearby small impacts or minor internal tectonic activity.

8. ICE

As discussed in more detail below, at each pole is a 3-km-thick water ice cap that extends out to about the 80° latitude circle. Water ice is also present within 2 m of the surface at latitudes down to about 60°, but how deep the ice-rich zone extends is uncertain. Most of the terrain in the 30°–55° latitude belt is covered with a 10 m-thick veneer of ice-rich material that appears to be undergoing removal. The veneer is thought to have been deposited within the past few millions years during a period of high obliquity. Other ice-rich deposits are found at latitudes as low as 30°, despite the present stability relations. Between latitudes 30° and 55°, flows called lobate debris aprons occur at the base of many steep slopes (See Figure 17.13). They have long been suspected to contain significant fractions of ice, but recently, ground-penetrating radar has shown that they consist almost entirely of ice. They have been interpreted as remanents of a more extensive ice cover that formed at midlatitudes during high-obliquity periods. Glacier-like flows on volcanoes and elsewhere may have formed similarly. A wide range of other observations, particularly in the low-lying northern plains, have been interpreted to be the result of ground ice or glaciers. These include polygonally fractured ground (analogous to arctic-patterned ground?); closely spaced, curvilinear, parallel ridges (moraines?); local hollows (left by removal of ice?); branching ridges (sites of former subglacial streams?); and striated ground (glacial scour?).

9. WIND

We know that the wind redistributes material across the Martian surface. We have observed dust storms from the orbits of the Earth and Mars and the changing patterns of surface markings that they cause. The 2004 Mars Exploration Rovers has made movies of dust devils, and tracks made by dust devils are visible on many high-resolution images taken from the orbit. Dust can be seen draped over rocks in many lander and rover images. Wind streaks
caused by eolian deposition or erosion in the lee of craters and other topographic features are common. Dunes are visible in almost all orbiter images with resolutions of a few meters per pixel or better (See Figure 17.14), and in some areas such as around the North Pole, dunes cover vast areas. Given all this evidence, it is somewhat surprising that wind erosion is not more widespread. Fine details of lava flows and impact ejecta, even though billions of years old, are generally well preserved. The wind appears to mostly move loose material around the surface. Additions to the inventory of loose material by erosion of consolidated primary igneous rocks or impact ejecta appears be proceeding only slowly.

Although the average large-scale effects of wind erosion in most places are trivial, locally the effects may be substantial. This is particularly true where friable deposits are at the surface. Fine details of lava flows and impact ejecta, even though billions of years old, are generally well preserved. The wind appears to mostly move loose material around the surface. Additions to the inventory of loose material by erosion of consolidated primary igneous rocks or impact ejecta appears be proceeding only slowly.

Although the average large-scale effects of wind erosion in most places are trivial, locally the effects may be substantial. This is particularly true where friable deposits are at the surface. Fine details of lava flows and impact ejecta, even though billions of years old, are generally well preserved. The wind appears to mostly move loose material around the surface. Additions to the inventory of loose material by erosion of consolidated primary igneous rocks or impact ejecta appears be proceeding only slowly.

10. POLES

During fall and winter, CO$_2$ condenses onto the polar regions to form a seasonal cap that can extend as far equatorward as 40° latitude. In summer the CO$_2$ cap sublimes. The seasonal north polar cap sublimates. The CO$_2$ cap does not dissipate completely, but water ice has still been detected under the seasonal cap. At both poles, kilometers-thick layered deposits extend down to roughly the 80° latitude. Individual layers are best seen in the walls of valleys cut into the sediments, where layering is observed at a range of scales down to the resolution limit of our best pictures. The frequency of impact craters on the upper surface of the deposits suggests that the sediments are young, of the order of 10$^8$ years or less. The poles act as a cold trap for water. Any water entering the atmosphere as a result of geologic processes such as volcanic eruptions or floods will ultimately freeze out at the poles. The poles may also be a trap for dust, in that dust can be scavenged out of the atmosphere as CO$_2$ freezes onto the poles each fall and winter. The layered deposits are, therefore, probably mixtures of dust and ice, with ice predominating. The layering is thought to be caused in some way by periodic changes in the thermal regimes at the poles, induced by variations in the planet’s orbital and rotational motions (see Section 2.1). These cyclical motions affect temperatures at the poles, the stability of CO$_2$ and H$_2$O, the pressure and circulation of the atmosphere, the incidence of dust storms, and so forth, hence the belief that they are responsible in some way for the observed layering.

11. THE VIEW FROM THE SURFACE

At the time of this writing, we had successfully landed at seven locations on the Martian surface: two Viking lander spacecraft in 1976, Mars Pathfinder in 1997, the Mars Exploration Rovers Spirit and Opportunity in 2004, the Phoenix lander in 2008, and the Mars Science Laboratory Curiosity rover in 2012. Viking 1 landed on a rolling, rock-strewn plain partly covered with dunes in the Chryse basin. Viking 2 landed on a level, rocky plain in Utopia. The main goal of the Viking landers was life detection. They carried a complex array of experiments designed to detect metabolism in different ways and to determine what organics there might be in the soil. Neither metabolism nor organics was detected. The lack of organics was somewhat surprising since organics should have been there from meteorite infall. However, the soil turned out to be oxidizing and the solar ultraviolet radiation environment on the surface harsh, which probably caused decomposition of any organics that might at one time have been present. Mars Pathfinder also landed on a rock-strewn plain in Chryse and deployed a small microrover named Sojourner. The site is at the mouth of one of the large outflow channels. It was hoped that evidence of floods might be observed there. However, the only sign of floods were some rocks stacked on edge and terraces on nearby hills that could have been shorelines. Phoenix landed at 68°N on a plain where ground ice was expected to be found, and indeed water ice was found just below the surface.
The two rovers Spirit and Opportunity, launched in 2003, provided the first solid evidence from the surface for pooling of water and aqueous alteration. Spirit landed on the flat floor of the 160-km-diameter crater Gusev. The site was chosen because the southern wall of Gusev is breached by a large channel called Ma‘adim Vallis. Water from the channel must at one time have pooled in Gusev, and it was hoped that the rover would be able to sample sediments from the postulated Gusev lake. The floor of Gusev turned out to be another rock-strewn plain. The rocks are basalts but they have alteration rinds with varying amounts of water-soluble components such as S, Cl, and Br. The alteration is minor and has been attributed to the action of acid fogs. Erosion rates estimated from craters superimposed on the plains indicate that the rates have been several orders of magnitude less than typical terrestrial rates. These somewhat disappointing results spurred a move to drive to and explore some nearby hills, where it was hoped different materials would be found, and indeed they were. Most of the rocks on the Columbia Hills are very different from those on the plains (See Figure 17.15). Many different classes of rocks have been identified, ranging from almost unaltered olivine basalts like those on the plains to almost completely altered, soft rocks enriched throughout with mobile elements such as S, Cl, and Br. In these altered rocks, primary basalt minerals are almost absent, having been replaced by secondary minerals such iron oxides and oxyhydroxides that have high Fe\(^{3+}/Fe^{2+}\) ratios compared with the unaltered rocks. Hydrated sulfates, opaline (hydrated) silica, or carbonates may also be present. A sulfate cement in some rocks suggests evaporation of sulfate-bearing (low pH) waters. Some rocks have soft interiors that have been hollowed out by the wind to leave only the hard outer shell. Layered rocks are common, and a coarse stratification appears to follow the contours of the hills. The origin of the Columbia Hills rocks and minerals is still being debated. Some may have formed by aqueous alteration of newly deposited impact or volcanic debris. Some may have been hydrothermally altered either during or perhaps even long after deposition. For others, waters from the postulated Gusev lake may have been implicated. Whatever the cause, aqueous processes were involved. Regolith deposits of almost pure silica are also indicative of hydrothermal activity.

Opportunity landed in Meridiani Planum on a thick stack of layered rocks that had been observed from orbit (See Figures 17.16 and 17.17). The site was chosen because a particular coarse-grained form of the iron oxide mineral hematite, which often forms in aqueous environments, had been detected there. The number of impact craters superimposed on the layered rocks suggests that they formed at the end of the heavy bombardment period around 3.8 billion years ago. The rover data demonstrated unequivocally that the local rocks are reworked evaporitic sandstones with roughly equal proportions of basaltic debris and evaporitic minerals such as Mg, Ca, Fe, Na sulfates, and chlorides. Although most of the rocks were initially deposited by the wind, there had to be a nearby source for the evaporites, which form by evaporation of bodies of water. The source had to be substantial because the layered sequence on which the rover landed extends for several hundred kilometers. A small fraction of the rocks have depositional textures that indicate that they were deposited in standing water. The environment in which the Meridiani sequence accumulated is thus thought to be one in which there were wind-blown dunes with interdune ponds.

Both Spirit and Opportunity, although landing on very different geologic materials, are telling a somewhat similar story. The oldest rocks, those that formed during or near the end of the late heavy bombardment, have abundant evidence for aqueous processes, but the evidence for
FIGURE 17.16 View of Endurance crater from the Opportunity Rover in Meridiani. The impact crater formed in a sequence of horizontally layered rocks, which are exposed in the foreground and in the walls of the crater. The horizon in the background gives an indication of how level the rock sequence is. The rover entered the crater and made measurements down section, almost to the center of the crater. Burns Cliff, seen in the next figure, is on the far wall (MER/Pancam).

FIGURE 17.17 View of Burns Cliff from the Opportunity rover in Meridiani. The rocks consist of a mixture of evaporites, such as sulfates and chlorides, and basaltic debris. The bedding patterns indicate that they were mostly deposited by the wind. However, the evaporites must originally have been derived by evaporation of a nearby lake or sea (MER/Pancam).
such processes well after the end of the late heavy bombardment is sparse or absent. Designed to last just 90 Martian days, Spirit’s mission lasted more than 2200 days and Opportunity’s mission continues as of this writing, more than 3100 days after landing. The longevity and mobility of these remote-controlled vehicles has enabled significantly more diverse scientific discoveries than originally envisioned. Indeed, Opportunity is now extensively exploring phyllosilicate-bearing terrains within the rim of an ancient Noachian crater, pushing our understanding of Martian geology and geochemistry back even further into the past.

Most recently, the Curiosity rover landed in August 2012 and began its mission within the 150-km-diameter crater named Gale, along the boundary between the southern highlands and northern lowlands. Gale was chosen because it is a deep and closed depression that may once have hosted a crater lake (like Gusev) and which also contains an enormous (4 km tall) central mound of finely layered sedimentary rocks that appear to span a large fraction of early Martian history (See Figure 17.18). The floor of the crater and the lower layers of the mound contain evidence for fluvial transport as well as hydrated phyllosilicate and sulfate minerals. It is hoped that over the course of the mission Curiosity will be able to traverse up the mound, going from older to younger deposits, looking for evidence of changing environmental conditions. Already, as of this writing, solid evidence of stream-deposited sediments has been found (See Figure 17.19), confirming the former presence of liquid water on the crater floor.
12. SUMMARY

Mars is a geologically heterogeneous planet on which have operated many of the geologic processes familiar to us here on the Earth. It has been volcanically active throughout its history; the crust has experienced extensive tectonic deformation, largely as a result of massive surface loads; and the surface has been eroded by wind, water, and ice. Despite these similarities, the evolution of Mars and Earth has been very different. The lack of plate tectonics on Mars has prevented the formation of linear mountain chains and cycling of crustal material through the mantle, and climatic conditions that hindered the flow of water across the surface have limited erosion and deposition to almost negligible levels for most of the planet’s history. As a consequence, a geologic record is preserved on the surface that spans almost the entire history of the planet. For the late heavy bombardment period we have compelling chemical and mineralogic evidence for aqueous alteration and compelling geomorphologic evidence of widespread fluvial erosion and transient lakes. The following Hesperian period was characterized by large floods, widespread volcanism, and accumulation of sulfate minerals indicative of lower pH weathering conditions. In the second half of Mars’ history, geologic activity declined significantly, although there were still occasional floods and other fluvial and glacial events. The climatic implications of the geologic observations remain uncertain. While early Mars must have had at least episodic warm, wet climatic episodes, any warm episode after the end of the late heavy bombardment must have been very short, because the cumulative amounts of erosion and weathering are so small. Significant mysteries remain. What caused the early warm conditions? What processes were responsible for the early massive amounts of erosion and deposition indicated by the geologic record? Where did all of the water go? How much of the Martian surface and subsurface was (or perhaps still is) habitable, by terrestrial standards? Most important of all, did some form of life ever evolve on the planet, and if so is it extant today?

BIBLIOGRAPHY

1. INTRODUCTION

The planet Mars is situated further from the Sun than the other three terrestrial planets of the solar system and has the lowest average surface temperature. Because its surface layers are, nevertheless, much warmer than the interplanetary medium, Mars, like the other terrestrial planets, loses heat and, as its decreasing internal heat sources cannot balance the loss, slowly cools. The rate at which heat is transferred from the deep interior to space is currently unknown but is much smaller than the energy emitted by a blackbody in thermal equilibrium with the surface temperature of Mars. The surface temperature is mainly determined from the balance between the absorbed solar energy and the reemitted blackbody radiation. Although the energy lost by Mars is small, the thermal evolution of the planet and the slow transfer of heat from the deep interior to the surface is the major driver for the general evolution and dynamics of the planet. The internal dynamics of Mars is affected by the internal temperature and internal heat transfer because heat is often transported by macroscopic mass motion in terrestrial planets. For example, convection is the main mechanism for redistributing heat in the mantle of Mars and is ultimately responsible for a wide range of phenomena including volcanism. Convective motions inside the core have even caused a dynamo to operate in the early history of Mars and explain the existence of large regions of magnetized rock on the surface of Mars.

All four terrestrial planets are thought to be principally composed of only four elements: iron, oxygen, silicon, and magnesium. For the Earth, the first two elements constitute about 30% each of the total planetary mass, whereas the latter two contribute about 15% each. Information on the bulk composition of Mars is much more limited than for the Earth, with main data obtained from analyses of meteorites from Mars, in situ data from landers, remote sensing data from orbiters, and cosmochemistry. In terms of bulk chemical composition it is thought that Mars is rather similar to the Earth. This may, however, not be taken as an indication that all terrestrial planets would have such a bulk composition. In the case of Mercury, more than half of its mass is in the form of iron, suggesting a different formation history (see the chapter on Mercury).
The chemical elements are distributed unevenly in three main reservoirs of the terrestrial planets: the core consists mainly of iron and the mantle and crust consist of silicate rock, which consists primarily of the four elements mentioned above (see Figure 18.1). In the Earth, the core is subdivided into a solid inner core and a liquid outer core. The physical state and structure of the core of the other terrestrial planets is not known with certainty. For Mars, recent geodesy data suggest that the core is entirely liquid.

Compared to the Earth and Venus with a radius over 6000 km, Mars is much smaller with a mean radius of 3389.5 ± 0.2 km (Figure 18.2). In terms of mass, Mars is almost 10 times less massive than its larger sister planets (9.3 times for the Earth and 7.6 times for Venus, see Table 18.1 for an overview of general characteristics of Mars). However, compared to Mercury, Mars is somewhat larger and almost twice as massive. Because Mars is smaller than the Earth, it is expected to form at a lower temperature. One might also expect that smaller planets like Mars cool faster than the Earth because the heat content of a planet is proportional to the planetary volume (or to the third power of the radius) and the heat loss is proportional to the surface (or to the square of the radius) so that the heat source to heat loss ratio is proportional to the radius. However, heat transport in planets is a complex phenomenon and can be organized in different ways in different planets, causing, for example, Mars to cool somewhat slower than the Earth.

The mean density of Mars is 3933 kg/m$^3$, which is about 28% less than that of the Earth (5515 kg/m$^3$, see Figure 18.2). This smaller value, however, does not mean that Mars’ bulk chemical composition is depleted in heavier elements and enriched in lighter chemical elements with respect to the Earth, instead that the lower density results from the planet’s smaller size. The pressure in smaller planets, like Mars, is on average lower than that in the larger planets Venus and the Earth. Therefore, planetary material in small planets is on average less compressed than in large planets, and small planets are expected to be less dense than large terrestrial planets if they were to have the same bulk composition. Since this compression effect is important in terrestrial planets, the interior of terrestrial planets cannot be properly understood without good knowledge of how much planetary materials can be compressed at typical planetary pressure conditions. For Mars, internal pressures increase with depth up to about 40 GPa. At those pressures iron is compressed by about 20%. Moreover, planetary materials expand with increasing temperatures, which in Mars increases with depth to about 2000 K, but the effect on the material density is smaller than the compression effect of pressure. For example, at a pressure of 40 GPa and a temperature of 2000 K, iron is about 12% denser than at standard conditions of 298 K and $10^5$ Pa. Material properties at the high internal pressures and temperatures of Mars can currently be measured in the laboratory. Moreover, theoretical calculations based on the basic principles of quantum mechanics (ab initio methods) are currently being used to study how planetary material behaves at high pressures and temperatures.

This chapter is organized as follows. We present a brief overview of the formation process and internal differentiation of Mars into a core, mantle, and crust in Section 2. General properties of these three main reservoirs of Mars are provided in the next three sections. Section 6
explains the general physical principles governing the interior structure and evolution of terrestrial planets. Recent results about the global interior structure of Mars and the thermochemical evolution are described in Sections 7 and 8.

2. FORMATION AND DIFFERENTIATION OF MARS

Terrestrial planets like Mars form when kilometer-sized planetesimals, which originate from the accumulation of dust grains in less than about 10,000 years, collide under the influence of the gravitational attraction between them and gas drag, a process called accretion. In less than a million years, tens to hundreds of planetary embryos are formed with masses between those of the Moon and Mars. By bringing planetesimal material with a Mars mass \( M \) from a far distance to a small region of space with Mars radius \( R \), the gravitational potential energy decreases enormously by about \( 0.6 \frac{GM^2}{R} \), where \( G \) is the universal gravitational constant. This change in energy is mostly converted into heat. The associated temperature rise is about \( 0.6 \frac{GM^2}{RC} \approx 6000 \text{ K} \), where \( C \approx 1200 \text{ J/K/kg} \) is the specific heat of Mars, suggesting that Mars formed hot.

However, part of the massive amount of energy liberated by accretion will be radiated to space. For slow accretion over about 1 My of small planetesimals, even most of the gravitational energy could be lost by radiation and planets would not form hot but would be cold with temperatures of a few hundred Kelvin close to the temperature of the protoplanetary disk. The key to the hot origin of planets lies in how deep the energy can be deposited in a planet. If energy can be brought to the deep interior, a very efficient internal heat transport mechanism would be required for a cold formation to occur. The general consensus is that in particular the large impacts at the end of formation strongly heat the deep interior of a planet in a fraction of geological time and can melt at least part of the planet interior producing a magma ocean.

The final formation of Mars took less than about 10 My. This is faster than for the Earth, for which the last large impact that formed the Moon is thought to have occurred 30–50 My after the formation of the solar system. The much smaller mass of Mars and fast formation compared to the Earth suggest that Mars could be a remnant embryo. This embryo status of Mars could be explained in the accretion scenario if the outer edge of the planetesimal disk of the initial solar system was at about 1 AU (an Astronomical Unit is the distance between the Sun and the Earth). Such a small inner disk possibly formed by inward migration of Jupiter to the Sun in the early phases of the solar system to a distance of only about 1.5 AU.

Because of its fast formation, Mars probably suffered less violent impacts with respect to the Earth resulting in a more limited heating. Nevertheless, the decay of the short-lived radioactive isotope \(^{26}\text{Al}\) with a half-life of 0.72 My could produce up to half the energy of accretion, more than sufficient for mantle melting and the formation of a magma ocean (half-life is the length of time needed for half of the parent atoms to decay into daughter atoms). Iron droplets in the magma ocean could then descend to form the core and at the same time mantle material crystallized at the cold

<table>
<thead>
<tr>
<th>TABLE 18.1 General Characteristics of Mars</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quantity</strong></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Mass</td>
</tr>
<tr>
<td>Radius</td>
</tr>
<tr>
<td>Mean density</td>
</tr>
<tr>
<td>Mean moment of inertia factor</td>
</tr>
<tr>
<td>Mean surface gravitational acceleration</td>
</tr>
<tr>
<td>Semimajor axis</td>
</tr>
<tr>
<td>Orbital period</td>
</tr>
<tr>
<td>Rotation period</td>
</tr>
<tr>
<td>Obliquity</td>
</tr>
<tr>
<td>Eccentricity</td>
</tr>
</tbody>
</table>
surface layers to form the primordial crust. In the core formation process, additional gravitational energy is released and converted into heat, further increasing the internal temperature of Mars and facilitating core formation. Mars, therefore, quickly differentiated into a core, a mantle, and a crust, a process that already started during the formation. After solidification of the magma ocean, at least part of the mantle is thought to be gravitationally unstable because the lightest material most rich in magnesium solidifies first at the bottom of the magma ocean and the more iron-rich and denser silicates solidify later. As a result, the mantle overpools: the denser materials sink and the lighter materials rise producing a gravitationally stable stratification, a process that could have latest up to 100 My after core formation.

The chronology of the planetary formation has been possible to unravel thanks to the study of radioactive parent—daughter systems with half-lives of the order of 10 My. In particular, the hafnium—tungsten system is widely used to constrain the accretion timescale and to date core formation of planetary bodies. $^{182}$Hf is a short-lived isotope that decays to $^{182}$W with a half-life of 9 My. Hafnium is a lithophile, or “rock-loving”, element meaning that it will stay in the mantle when the iron core forms. Tungsten on the other hand is a siderophile (“iron-loving”) element and will follow the iron to the core on core formation. Therefore, part of the $^{182}$W produced from $^{182}$Hf will be in the core if the core formation time is comparable to the half-life of hafnium. By comparing the Hf/W ratio from the Martian mantle (estimated from Martian meteorites) with the initial ratio (derived from chondrite meteorites), the age of core formation of less than 10 My can be deduced.

3. CORE

Seismic data have shown that the core of the Earth has an outer liquid part and a solid inner part and that both parts have densities that are only several percent lower than iron at core pressure and temperature. Iron is the most likely constituent element of planetary cores since it is the only element abundant enough in the universe with a density and elastic properties closely agreeing with seismic data. Besides iron, analyses of iron meteorites—thought to be representative for cores of planetary bodies—suggest that the Earth’s core also contains a few percent of the somewhat denser element nickel. Since the core is less dense than pure iron, it also contains light elements. The identity of those elements is still debated, but the most prominent candidates are sulfur, silicon, and oxygen. The light elements not only lower the density and change the elastic properties of pure iron but also decrease the melting temperature of the core material significantly compared to the melting temperature of pure iron. About 1220 km of the 3480 km of the core radius have solidified due to cooling of the Earth over the past 4.5 By. On cooling, iron-rich material solidifies at the boundary of the solid inner core and liquid outer core. Seismic data have revealed that the solidified inner core material is denser than solid material of outer core composition at inner core pressures and temperatures. This implies that some of the light elements do not partition in equal amounts into the solid and liquid parts of the core. As a result, material immediately above the inner core—outer core boundary becomes enriched in light elements. The buoyancy force acting on this less dense material forming at the bottom of the outer core generates convective motions in the whole liquid outer core and provides the major part of the energy necessary to drive the dynamo responsible for the Earth’s global magnetic field.

Similarly as for the Earth, the core of Mars is thought to be principally made of iron with unknown but small fractions of nickel and light elements. Direct information on the density, elastic properties, and size of the core is lacking because of absence of seismic data. Based on the chemical affinity of light elements to iron—nickel mixtures, the main candidate light elements for the Martian core are sulfur, silicon, oxygen, carbon, and hydrogen, in binary, ternary or more complicated systems with iron and nickel. Analyses of rock samples representative of the bulk composition of the mantle of Mars, such as Martian meteorites, allow further constraining the core composition, although the results depend on core formation scenarios. They indicate that iron, nickel, and the light element sulfur are the principal constituents of the core. Additional evidence for the presence of sulfur comes from the fact that it has been found in many nickel—iron meteorites. Moreover, all of the light elements, sulfur and silicon are most easily incorporated into planetary cores of terrestrial planets as their solubility in molten iron is high over an extended pressure range. In contrast, for example, oxygen solubility in liquid iron is below 1wt% at ambient pressure and increases with increasing pressure and sulfur concentration. At the pressure and temperature conditions at the bottom of the Martian mantle, at most a few weight percent of oxygen could partition in the liquid core. Silicon and carbon can only have been incorporated in the core if Mars formed under reducing conditions which is, however, rather unlikely given the highly oxidized state of the Martian surface. Finally, hydrogen might also enter the core if the magma ocean contained water and the pressure at the bottom of the magma ocean was higher than about 5 GPa.

Given the presumed formation conditions and the above data for liquid iron systems, it is unlikely that Mars’ core contains significant amounts of light elements other than sulfur. Moreover, sulfur has the important property that it strongly reduces the melting temperature of the iron system with respect to pure iron. For example, at 21 GPa, the melting temperature of Fe decreases by about 60 K for each
additional weight percent of S as long as the concentration of S is below the eutectic concentration, at which the melting temperature is the lowest for a Fe—FeS system. Low melting temperature of the core might be important since the outer part of the core is liquid since the solid tides of Mars, estimated by observing orbital variations of spacecraft motion, are several times larger than for an entirely solid Mars. Whether or not the inner part of the core is solid like in the Earth depends on the composition and temperature of the core. If the melting temperature of the core is below the core temperature, as is possible for a large sulfur concentration in the core, the core will be entirely liquid. A fully liquid core is consistent with the observation that Mars does currently not possess a global magnetic field as the Earth but only localized magnetized rocks. Dynamo action in the core, which is responsible for the global magnetic field of the Earth, is not possible at the present day without a growing inner core. In Section 7, it will be shown that tidal observations can be used to put further constraints on the size, density, physical state, and composition of the core.

4. MANTLE

The silicate part of the Earth consists mainly of rocks (~21wt% Si and ~44wt% O) that are rich in magnesium (~23wt%) and poor in iron (~7wt%). Together with the elements Al (~2.3wt%) and Ca (2.5wt%) these four elements are responsible for more than 98% of the total mass of the silicate Earth. The upper mantle rocks are principally made of the minerals olivine and pyroxenes. With increasing depth, or increasing pressure and temperature, these minerals experience phase transitions to denser polymorphs or dissociate to other minerals that are stable at those pressure and temperature conditions. The principal lower mantle minerals are magnesium perovskite and magnesiowüstite. At the bottom of the mantle close to the core, the still denser mineral post-perovskite dominates. The distribution of the minerals as a function of depth in the mantle is not accurately known but can be constrained from the known seismic velocities and density in the Earth’s mantle and from measurements of the electrical conductivity of the mantle.

Unlike for the Earth, seismic data are not yet available for Mars. As Mars is also a terrestrial planet it is assumed that the principal minerals in the Martian mantle are the same as those in the terrestrial mantle. Since the pressures in the mantle of the smaller planet Mars are much smaller than those in the Earth (maximum pressure of about 20 GPa in the Martian mantle compared to about 135 GPa at the bottom of the mantle of the Earth), the densest mineral phases in the Earth’s mantle are not present (post-perovskite) or only marginally (perovskite and magnesiowüstite) present. Therefore, the minerals in the mantle of Mars are likely the same as those in the upper mantle of the Earth: olivine and its high-pressure polymorphs, pyroxenes and garnet (Figure 18.3). The chemical composition of the silicate part of Mars has been constrained by analyzing Martian meteorites, performing surface monitoring from orbiting spacecraft, and analyzing surface rocks in situ by means of robotic rovers. One of the most striking differences compared to the Earth is the significantly higher concentration of iron (>20wt%) in the minerals. This difference could be explained by a difference in the chemical composition of the materials that formed the planets and by the significantly lower pressure at the bottom of the magma ocean on Mars compared to the Earth, which strongly reduces the dissolution of oxidized iron in the iron melt forming the core.

**FIGURE 18.3** Phase diagram for the mantle of Mars derived from the chemical analysis of Martian meteorites (Dreibus and Wänke, 1985) assuming a cold end-member mantle temperature profile. The upper mantle is rich in olivine (ol) and pyroxenes (clinopyroxene LP (Lcpx), clinopyroxene HP (Hcpx), orthopyroxene (opx), and Ca-pyroxene (Ca-px)). With increasing pressure the pyroxenes dissociate to majorite (maj) and olivine transforms to its higher pressure polymorphs wadsleyite (wad) and ringwoodite (ring). At the highest possible pressure, close to the core—mantle boundary, ringwoodite starts to dissociates to (Mg,Fe)-perovskite (pv) and (Mg,Fe)-wüstite (mw). At those high pressures, majorite first takes the akimotoite (ak) structure, and subsequently transforms into perovskite.
In terms of oxides, the mantle of Mars is thought to consist of 45% SiO₂, 30% MgO, 17% FeO, 3% Al₂O₃, 2% CaO, and small contributions from other elements. For the Earth, we have of 47% SiO₂, 37% MgO, 8% FeO, 4% Al₂O₃, and 3.5% CaO.

5. CRUST

The crust is the thin upper layer of a terrestrial planet. It consists of silicate rocks like the mantle but is chemically different from the mantle. The crust is more silica rich (SiO₂ contributes 50% or more in mass) and has a lower density than the underlying mantle as a result of its formation from the mantle by melting and crystallization. Studies of radioactive parent—daughter isotopes in Martian meteorites that separate differently between the liquid and the solid silicate phases show that the bulk of the crust was created within 100 My after formation of the planet. This primordial crust presumably formed from crystallization of the cooling magma ocean. Additional crust is formed later in Mars’ evolution when ascending hot material from the mantle partially melts and rises to the crust and surface by volcanism (see Section 8). The crust is basaltic in composition with the older crust being more silica rich. However, rocks with the highest silica content on the Earth, felsic rocks such as granite, are almost absent on Mars. On the Earth, plate tectonics, which is absent on Mars, continuously changes the crust. A division in continental crust and oceanic crust as for the Earth can, therefore, not be made for Mars.

Mars’ crust exhibits a very notable crustal dichotomy: the northern hemisphere is almost flat and covered with volcanic rocks, whereas the southern hemisphere is a few kilometers higher and cratered by ancient impacts (Figure 18.4). The crust is about 25 km thicker in the southern highlands than in the northern lowlands. The thickness of the crust has been estimated from the topography data measured by the laser altimeter onboard the Mars Global Surveyor (operational between 1997 and 2006) and from the gravity field determined by radio tracking orbiting spacecraft. The average crustal thickness is estimated to be between 38 and 62 km, and the average crust density is about 2700–3100 kg/m³. The dichotomy was formed shortly after formation of Mars within the first half billion year and maybe even within the first 50 My. The dichotomy could have been created by large impactors but could also be due to internal processes such as the crystallization of the magma ocean, an early phase of plate tectonics, or convective motions in the mantle characterized by a large upwelling plume beneath the southern hemisphere.

One of the most prominent topographic features on the surface of Mars besides the dichotomy is the volcanic plateau of Tharsis which is located near the equator in the western hemisphere. The plateau or bulge is about 5000 km across and up to 7 km high and harbors the largest and highest volcanoes of the solar system. It is thought that Tharsis is created by a volcanic plume, quite similar to the one found beneath the island of Hawaii. The idea is that a hot column of mantle material rose from the core—mantle boundary through the mantle and delivered substantial volumes of basaltic lava to the surface. Because of the absence of plate tectonics, high volcanoes could develop over long timescales, with the highest of all, Olympus Mons, rising 22 km above the reference surface. The lower gravity on Mars (see Table 18.1) explains why higher mountains can exist on Mars than on the Earth.

Analyses of the mineralogy of surface rocks and water erosion features such as outflow channels on Mars’ surface indicate that Mars had large amounts of liquid water on its
6. PRINCIPLES OF GLOBAL INTERIOR STRUCTURE AND EVOLUTION

6.1. Basic Equilibrium Equations

Although a terrestrial planet such as Mars is not exactly spherically symmetric, its global interior structure in the radial direction can very well be described by assuming that the physical quantities do not depend on the angular variables but only on the radial distance to the planet’s mass center. A major mathematical advantage of this approximation is that the problem of the description of the interior structure is reduced from a three-dimensional problem to a one-dimensional (1D) problem.

Although the cold and stiff upper layers of Mars can support shear stresses on very long timescales, the bulk of the interior cannot and behaves as a viscous fluid. The nearly spherical shape of Mars is a clear indication of the fluid-like behavior of Mars on very long timescales. Even on timescales of the order of 10,000 years mantle material of terrestrial planets is known to behave as a viscous fluid as convincingly follows from studies of postglacial rebound on Earth. A basic physical model that exhibits elastic behavior on short timescales and viscous behavior on long timescales is the Maxwell model, which can be described in terms of a spring and a dashpot in series. When an outward force is applied at both sides of such a system, the spring immediately responds elastically and stretches. When the system is subsequently kept at the same length, the dashpot gradually pulls apart, and the spring decreases in length and relieves stress. After a long time, the spring returns to its original length without stress.

Because of the fluid-like behavior on geological timescales terrestrial planets are close to hydrostatic equilibrium. Hydrostatic equilibrium expresses that the downward gravitational force is balanced by the upward differential pressure force at any point in the planet at a radial distance \( r \) to the planet’s mass center:

\[
\frac{dP(r)}{dr} = -\rho(r)g(r), \quad (18.1)
\]

where \( P(r) \) is the pressure, \( \rho(r) \) is the mass density, and \( g(r) \) is the gravity.

Newton’s theory of gravitation shows that anywhere in the planet, gravity and density must also satisfy Poisson’s equation:

\[
\frac{dg(r)}{dr} + \frac{2}{r}g(r) = 4\pi G\rho(r), \quad (18.2)
\]

Besides pressure and gravity, Eqns (18.1) and (18.2) depend on the density. A third equation is therefore needed to solve for these three variables. This is given by an equation of state (EoS) specifying the dependence of the density on pressure, temperature, and composition. For solids and liquids, EoSs are more complicated than the ideal gas law. An often used EoS for solids is the Birch–Murnaghan equation.

For a given temperature profile and composition, the radial profiles of density, pressure, and gravity can then be calculated from the center to the surface (Figure 18.5). The temperature profile depends on the heat sources and on the method of heat transport, which are discussed below. Although Mars loses heat at its surface to interplanetary space and globally cools, the heat loss of Mars is small and the average mantle temperature of Mars only decreases by \( 30–50 \) K per billion year. It is then justified to assume that, on short timescales, Mars is in thermal equilibrium expressing a balance between the energy lost by outward energy flux and the internal energy generation. This energy balance can be expressed as

\[
\frac{dg(r)}{dr} + \frac{2}{r}g(r) = \rho(r)\varepsilon(r), \quad (18.3)
\]

where \( \varepsilon(r) \) is the specific heat production rate per unit mass and unit time and \( q(r) \) is the outward heat flux per unit area and unit time. Expressions for \( q(r) \) are given below.

6.2. Heat Sources

Several radioactive isotopes occur in nature but only a few of them play an important role in the global energy budget of terrestrial planets on long timescales. The most important are the uranium isotopes \(^{235}\text{U}\) and \(^{238}\text{U}\), thorium \(^{232}\text{Th}\), and the potassium isotope \(^{40}\text{K}\). All these isotopes have a long half-life, of \( 7.04 \times 10^8 \) years, \( 4.47 \times 10^9 \) years, \( 1.40 \times 10^{10} \) years, and \( 1.25 \times 10^{10} \) years, respectively, so that they still contribute to the heat budget of Mars up to the present day. Currently on the Earth, natural uranium consists of 99.28 wt\% \(^{238}\text{U}\) and 0.71 wt\% \(^{235}\text{U}\) and natural potassium contains only 0.0119 wt\% \(^{40}\text{K}\). As the energy liberated by a unit of mass differs between all these isotopes by at most one order of magnitude, \(^{238}\text{U}\) and \(^{232}\text{Th}\) are the most important radiogenic heat sources of Mars in the current era, but both \(^{235}\text{U}\) and especially \(^{40}\text{K}\) were more important in the early phases of Mars (see Figure 18.6).

Radioactive isotopes can be found in mantle and crust rocks. As they are incompatible elements, meaning that on partial melting of rocks they will be concentrated in the melt phase, radiogenic elements are expected to be more abundant in crustal rocks than in mantle rocks. On Earth, the concentration of radioactive elements in the crust is...
more than 10 times higher than in the mantle. Little information is available to constrain the amount of heat-producing elements in Mars. Based on analyses of meteorites from Mars, it is usually assumed that their concentration is close to that in chondritic meteorites, the most primitive type of meteorites. Data on the occurrence of radioactive elements in the crust are compatible with this assumption. For an essentially chondritic concentration the radioactive elements have up to now produced an energy of about $7 \times 10^{29}$ J in Mars.

Although radiogenic isotopes keep on producing energy, the dominant source of energy for Mars is the release of gravitational energy on formation (see Section 2). When a total mass $M$ is accreted to form Mars, a gravitational energy of about $5 \times 10^{30}$ J is released and mostly converted into heat. In addition to the two major energy sources, several other phenomena contribute to Mars’ heat balance. In Mars’ early history, core formation released about 5% of the gravitational energy production of formation and short-lived radioactive isotopes such as $^{26}$Al also...
delivered an important amount of energy. In the later evolution, contraction on cooling and further differentiation of the planet, in particular the formation of an inner core, release additional gravitational energy.

6.3. Heat Transport by Conduction and Convection

Heat is transported in Mars by two different mechanisms. Like in any medium with a spatial variation in temperature, heat is conducted from hot to colder regions. Conduction is a diffusive process wherein molecules transmit their kinetic energy to other molecules by colliding with them. Besides conduction, macroscopic motions in the medium can also transfer heat when flows carry material of a certain temperature into a region with a different temperature.

6.3.1. Conduction

According to Fourier’s law, the conductive heat flux $q(r)$ (the flow of heat through a unit area per unit of time) is linearly proportional to the temperature gradient:

$$q(r) = -k(r) \frac{dT(r)}{dr},$$  \hspace{1cm} (18.4)

where $k$ is the coefficient of thermal conductivity. The minus sign indicates that heat flows in the direction of decreasing temperature. The temperature distribution in the crust and the upper part of the mantle of terrestrial planets like Mars is controlled by conduction. Equation (18.4) shows that the heat flux at the surface can be determined from the surface temperature gradient. Accurate measurements of the surface temperature gradient require deep drill holes. On the Earth, such measurements combined with measurements of the coefficient of thermal conductivity of crustal rocks have shown that the mean continental heat flow is 65 $\pm$ 1.6 mW/m$^2$. The heat flow out of Mars has up to now not been measured, but will be by the National Aeronautics and Space Administration (NASA) InSight mission planned for launch to Mars in 2016. From theoretical models of the thermal evolution of Mars (see Section 8) the heat flow is estimated to be about 20 mW/m$^2$ and the temperature increase with depth in the upper crust to be about 7 K/km.

Lord Kelvin (1862) assumed that conduction was the only heat transport mechanism in the Earth and tried to calculate the age of the Earth by considering a simple model. He approximated heat transfer in the Earth by that in an infinite half-space and assumed that surface heat flow results from the cooling of an initially hot Earth. The temperature in such a system obeys the diffusion equation

$$\frac{\partial T(x, t)}{\partial t} = \kappa \frac{\partial^2 T(x, t)}{\partial x^2},$$  \hspace{1cm} (18.5)

where $\kappa = k(\rho C)$ is the thermal diffusivity and $x$ is the distance from the surface. A large thermal diffusivity indicates a fast adjustment of temperature to the surrounding temperature. Assuming that the Earth was initially at a hot homogeneous temperature $T_i$ and the surface at a cold temperature $T_s$, a solution of Eqn (18.5) for the temperature can be expressed as

$$T(x, t) - T_i = (T_s - T_i) \text{erfc} \left( \frac{x}{2\sqrt{\kappa t}} \right),$$  \hspace{1cm} (18.6)

where erfc is the complementary error function. Since erfc(z) decreases rapidly with increasing argument $z$—its value decreases from 1 at the origin to 0.1 for an argument of about 1.16, and is nearly equal to 0 for larger arguments—only the surface layer cools significantly (Figure 18.7). The physical reason for this behavior is that the characteristic time for conductive propagation of a temperature change over the radius of the planet, which can be expressed as $R^2/\kappa$, is much larger than the age of the solar system. For Mars, the characteristic time for cooling over its entire radius is about 100 times longer than the age of Mars. The thermal boundary layer over which Mars can efficiently cool over its lifetime is less than 1000 km. As a result, Mars, like the Earth, cannot efficiently change its deep interior temperature by conduction. A purely conductive planet can, therefore, not maintain a large heat flux over the age of the solar system. From the high observed surface heat flux, Lord Kelvin estimated the age
of the Earth to be of the order of 100 My, much smaller than
the age of the solar system as we now know. Addition of
heat-producing radioactive isotopes can increase the age
estimate, but the main reason why the Earth can maintain a
high geothermal gradient is that the principal means of heat
transfer in the deeper Earth layers is not by conduction but
by convection. For Mars too, convection is the dominant
heat transport mechanism in the mantle.

6.3.2. Convection

Convection in the mantle is driven by radiogenic heat
sources and the cooling of Mars. A fluid that is heated from
below and within and cooled from above is denser (cooler)
at the top than at depth. A gravitationally unstable situation
then develops and the cool fluid tends to sink and the hot
fluid tends to rise due to buoyancy. As explained above, the
mantle behaves as a viscous fluid on long timescales and
approximately as an elastic solid on short timescales. On
microscopic level, the fluidlike behavior of the mantle of
terrestrial planets is due to solid-state creep processes. The
dominant creep process at low stress levels is the diffusion
of ions and vacancies through the crystal lattice. For a
material with a Maxwell rheology, the characteristic
timescale that separates predominantly elastic from
predominantly viscous behavior is given by the Maxwell
timescale \( t = 2\eta/E \), where \( \eta \) is the dynamic viscosity and \( E \)
is the Young’s elastic modulus. For a viscosity of \( 10^{20} \) Pa s
and \( E = 70 \) GPa, typical estimates for Earth mantle rocks
expected in Mars, the Maxwell timescale is about
900 years. Mantle convection occurs on much longer
timescales so that a viscous fluid behavior is a very good
description for the mantle rocks. As a result viscosity will
tend to resist the convective motion driven by buoyancy.
Typical convection velocities are on the order of a few
centimeters per year only, so it takes on the order of
10 My or more for a cold mass element to descend to the
core—mantle boundary.

Mantle convection can be described by the laws of fluid
mechanics: the equation of motion, the continuity equation
expressing conservation of mass, and an energy equation
expressing the change in internal energy due to heat
sources, fluid flow, and conduction. These equations cannot
be solved analytically for planets and are therefore solved
numerically. Nevertheless, analytical solutions have been
derived for some simplified cases. These studies, as well as
laboratory measurements and numerical models, show that
an isoviscous layer will be convecting if a certain quantity
called the Rayleigh number \( Ra \) exceeds a critical value, the
critical Rayleigh number \( Ra_{cr} \). The Rayleigh number is
defined for a plane-parallel layer without internal heat
sources and heated from below and cooled from above as

\[
Ra = \frac{\rho g \alpha \Delta T b^3}{\eta \kappa},
\]

where \( g \) is gravity and \( \alpha \) is the coefficient of thermal
expansion, describing the relative volume change with
changing temperature of a material. The thickness of the
layer is denoted by \( b \) and \( \Delta T \) is the nonadiabatic
temperature difference over the layer. The Rayleigh
number can be thought of as a ratio between two
timescales: a cooling timescale due to conduction and a
timescale of motion due to buoyancy and viscosity. If the
cooling timescale is large with respect to the timescale of
motion, convection can develop (\( Ra > Ra_{cr} \)). If, on the
other hand, conductive cooling is fast enough, the mantle
can cool before convection can set in (\( Ra < Ra_{cr} \)). The
critical Rayleigh number \( Ra_{cr} \) depends on the wavelength \( \lambda \)
in the horizontal direction of the velocity and temperature
perturbations resulting from the convective motions:

\[
Ra_{cr} = \left( \frac{\pi^2 + 4\pi^2 \frac{\lambda^2}{b^2}}{4\pi^2 \frac{\lambda^2}{b^2}} \right)^3.
\]

The minimum \( Ra_{cr} \) for this situation with free surfaces is
about 658 and is obtained at a wavelength of \( 2^{3/2} b \). The
most unstable horizontal wavelength is therefore a few
times the thickness of the layer. When convection first
sets in at a Rayleigh number equal to the critical Ray-
leigh number, convection cells of that size will therefore
develop. For larger Rayleigh number convection cells
with both larger and smaller wavelengths than the most
unstable will appear. For a mantle thickness of about
1500 km for Mars, a temperature increase of about
1500 K over the mantle, and the values \( \eta = 10^{20} \) Pa s,
$k = 4 \text{ W/(mK)}, \quad \kappa = 10^{-6} \text{ m}^2/\text{s}, \quad \alpha = 3 \times 10^{-5} \text{ K}^{-1}, \quad \rho = 3500 \text{ kg/m}^3$, we have $\text{Ra} \approx 2 \times 10^7 \gg \text{Ra}_{\text{cr}}$, strongly suggesting that the mantle of Mars is convecting.

In the convective mantle, the temperature increases approximately adiabatically with depth. An adiabatic temperature change describes how a material element adapts its temperature to the surrounding temperature when rapidly rising or sinking in the layer without exchanging energy with the surrounding fluid. For convection to set in, the temperature gradient in the layer must be larger than the adiabatic temperature gradient, as can be understood by considering the dynamics of a fluid parcel that is locally heated. The parcel expands and starts rising due to buoyancy. If it does not exchange heat with the surrounding, the fluid parcel changes its temperature and density adiabatically due to the different pressure it encounters. If the temperature gradient is larger than the adiabatic temperature gradient, the adiabatic density change of the fluid parcel will be smaller than the change in density of the environment and the parcel will have the tendency to keep on rising. Viscosity and conduction have a stabilizing effect on convective motions, as expressed by the Rayleigh number. However, due to the efficient energy transport of convection only a very small temperature difference of a few Kelvin on top of the adiabatic temperature gradient is needed for convection. As a result, the temperature gradient of convection is nearly adiabatic for a layer in convection. In the convecting part of the Martian mantle, we therefore have approximately

$$\frac{dT}{dr} = -\frac{\alpha g T}{C}. \quad (18.9)$$

This adiabatic temperature increases only slightly with depth by about 1.2 K per 10 km, much slower than in the crust where convection is absent (Figure 18.8). Over the entire range of about 1200 km in the convective mantle, the temperature increases by only about 140 K.

Heat flow by convection is much more efficient than by conduction. The heat flow for a convective planet is customarily described in terms of the Nusselt number, which is defined as the ratio of the total heat flux and the heat flux that would be transported by conduction only. It is generally expressed in terms of the Rayleigh number as

$$\text{Nu} = c \left( \frac{\text{Ra}}{\text{Ra}_{\text{cr}}} \right)^\beta, \quad (18.10)$$

where $c$ and $\beta$ are constants. Experimental and advanced numerical studies show that typically $c$ is smaller than 1 and that $\beta$ is approximately 1/3. Since the conducted heat flux is linearly proportional to the temperature difference over the convecting layer (Fourier’s law), the total heat flux at the surface of Mars is approximately proportional to that temperature difference to the power 4/3. The larger the temperature difference is over the convective system, the larger the heat flux.

Since the surface temperature is much lower than the temperatures in the convecting mantle and the temperature change in the convecting region is very small, a transition layer, or boundary layer, will develop at the top of the mantle in which temperature increases much faster than the adiabatic temperature and heat is transported by conduction. At the bottom of the mantle, a second boundary layer exists above the core (Figure 18.8). These boundaries have thicknesses of the order of several 10 km in Mars. The change in temperature in the mantle occurs mainly over the boundary layers. Since convection is only driven by a temperature difference on top of the difference in adiabatic temperature, the temperature difference considered in the Rayleigh number is approximately the sum of the temperature differences over the boundary layers. As heat in the boundary layers diffuses upward by conduction, the heat flux out of the mantle can be expressed by Fourier’s law as $q = k_m \Delta T/\delta$, where $k_m$ is the coefficient of thermal conductivity of the mantle and $\Delta T$ is the temperature difference over the boundary layer with thickness $\delta$.

In Mars, in contrast to the Earth, the outer layers of the planet do not participate in the convection and are essentially rigid. This type of mantle convection is usually

![Figure 18.8 Sketch of the temperature in the mantle and crust of Mars. Here, $R_c$, $R_u$, $R_l$, $R_i$, and $R$ denote the radial coordinate of the core–mantle boundary, of the top of the boundary layer at the bottom of the mantle, of the bottom of the upper boundary layer, of the bottom of the stagnant lid, of the crust–mantle boundary, and of the outer surface, respectively. Temperatures at the core–mantle boundary, at the top of the boundary layer at the bottom of the mantle, at the bottom of the upper boundary layer, at the bottom of the stagnant lid, at the crust–mantle boundary, and at the outer surface are indicated by $T_{\text{CMB}}$, $T_u$, $T_i$, $T_{\text{cr}}$, and $T_e$, respectively.](image)
referred to as stagnant lid convection, as the upper lid is immobile or stagnates, and heat transport in the outer lid is through conduction only. In an almost isoviscous fluid with limited viscosity contrast, the whole fluid takes part in the convection. The stagnant lid convection regime is typical for planets with a large viscosity ratio between the cold surface layers and the hot mantle interior of the order of at least $10^4$ between the surface layers and the deep mantle. Viscosity decreases with depth in Mars because of the increase in temperature. The decrease can be large from a cold surface to a hot interior since the dependence on temperature is exponential (viscosity can be expressed as $\eta = \eta_0 \exp((A/R)(1/T - 1/T_r))$, where $T_r$ is a reference temperature, $\eta_0$ the viscosity at the reference temperature, $R$ is the gas constant, and $A$ is the activation energy of the rocks; the small dependence of viscosity on pressure for Mars is not made explicit here). For example, the viscosity in the mantle of Mars at a temperature of 1500 K is already a factor 40 smaller than at a temperature of 1300 K. Therefore, a stagnant lid will develop above the convective layer. Most of the radial viscosity variation occurs in the stagnant lid, whereas the convective layer below is approximately isoviscous as a result of its small temperature gradient. Since the stagnant lid does not belong to the convection system and the temperature increases significantly over it, the temperature difference to be considered in the Rayleigh number for the convective layer below the stagnant lid is much smaller than for the larger convective system including the upper layers if the lid would be mobile (Figure 18.8). As a result, the Rayleigh number and the Nusselt number are smaller for stagnant lid convection than for mobile lid convection and the heat flux of the mantle is also smaller for a given temperature profile. Stagnant lid convection therefore cools a planet less effectively than mobile lid convection. For most of its history, Mars did not have plate tectonics, as the old crust convincingly shows.

Both numerical and laboratory experiments have shown that the temperature difference over the thermal boundary layer beneath the stagnant lid is of the order of 100 K. For a mantle temperature of about 1500 K, this means that the stagnant lid extends to a depth at which the temperature is about 1400 K. The thickness of the stagnant lid can then be estimated from the temperature gradient to be about a few hundred kilometers thick. The stagnant lid thus consists of the crust and an upper part of the mantle.

7. GLOBAL INTERIOR STRUCTURE OF MARS

7.1. Introduction

Most information on the interior structure of the Earth is provided by seismic data. Since the propagation of seismic waves depends on material properties like density and elastic parameters, observation of seismic wave arrival times and seismic normal-mode frequencies informs on the Earth’s interior. Since normal modes can involve motion of the whole Earth and seismic waves can penetrate down to the Earth’s center, the whole interior can be sampled. For Mars, no seismic data are yet available, but NASA plans to send a lander called InSight equipped with a seismometer to Mars in 2016. In the absence of these seismic data, geodesy data provide the strongest constraints on the deep interior structure of Mars.

7.2. Global Geodesy Data

Besides surface data and data from Mars meteorites, which are related to the interior but on itself cannot determine the interior structure, the best current data available to constrain Mars’ deep interior are global geodesy data derived from the gravity field, rotation, and tides of Mars. Four main data constrain the interior structure of Mars: the radius, the mass, the mean moment of inertia, and the tidal Love number $k_2$, which describes the reaction of Mars to the tidal forcing of the Sun (see below).

The mean radius of Mars $R = 3389.5 \pm 0.2$ km, defined as the radius of a sphere defining the same volume as Mars, has been determined by measuring the time needed for laser pulses to travel from the Mars Global Surveyor to the surface of Mars and back. The distance to the surface can then be calculated and by subtracting it from the spacecraft’s location with respect to the center of mass of Mars, the distance of the surface from the mass center can be determined. A spherical volume with the mean Martian radius is used to describe global 1D interior structure models of Mars, although Mars itself is only nearly spherical. A better approximation for the shape of Mars is a biaxial ellipsoid flattened at the poles due to rotation, as is the case for the Earth. The mean polar radius of Mars is $3376.2 \pm 0.1$ km, whereas the mean equatorial radius is $3396.2 \pm 0.1$ km.

The gravitational field of Mars can nowadays be determined very accurately by tracking orbiting spacecraft. By precisely measuring the Doppler shift on radio links between the spacecraft and the Earth the orbital motion of the spacecraft can be modeled. From the knowledge of the orbit the gravitational field of Mars can be determined since the spacecraft’s orbit depends essentially on the gravitational attraction from Mars. The most accurate representation of the external gravitational field of Mars reaches an average spatial resolution of 120 km at the surface and yields an estimate of the product $GM$ of Mars with 10 significant digits. Since the universal gravitational constant is only known with limited precision, the mass of Mars can be determined with five significant digits only.
For a value of $G = (6.67259 \pm 0.00085) \times 10^{-11}$ m$^3$/s$^2$/kg, the mass $M = (6.4186 \pm 0.0008) \times 10^{23}$ kg.

The moment of inertia of Mars can be determined by combining data from the gravitational field of Mars and the rotation. It can in general be expressed by a $3 \times 3$ matrix $I_{ij}$ as

$$I_{ij} = \int_{V} \rho(\vec{r})(r^2 \delta_{ij} - x_i x_j) \, dV,$$

where $\vec{r}$ is the position vector of a point P in Mars with respect to the origin of the Cartesian coordinate system, chosen here to be at the mass center of Mars, $r$ is the distance to the mass center, $x_i$ and $x_j$ are the i and j coordinates, and $\delta_{ij}$ is the delta Kronecker, which is equal to 1 when $i = j$ and else is equal to 0. Three principal moments of inertia exist which express the moment of inertia of Mars with respect to the three principal axes of inertia. When the three coordinate axes are chosen along these three axes of inertia, the moment of inertia matrix $I_{ij}$ can be expressed entirely in terms of the three principal moments of inertia $A$, $B$, and $C$, describing the moment of inertia with respect to the three axes. We have

$$A = \int_{V} \rho(\vec{r})(y^2 + z^2) \, dV,$$

$$B = \int_{V} \rho(\vec{r})(x^2 + z^2) \, dV,$$

$$C = \int_{V} \rho(\vec{r})(x^2 + y^2) \, dV.$$ (18.11)

The mean moment of inertia $I = (A + B + C)/3$ is the third quantity used to constrain the interior structure.

The gravitational field can be related to differences in moments of inertia if it is described with respect to the inertia axes defined above. If we restrict ourselves to a model that can describe a triaxial ellipsoidal shape, the external gravitational field $\Phi$ can be expressed as

$$\Phi(\vec{r}) = -\frac{GM}{r} \left[ 1 + \left( \frac{a}{r} \right)^2 \left( \frac{(A + B)/2 - C}{Ma^2} \cos(m\phi) \times P_{20}(\cos \theta) + \frac{B - A}{4Ma^2} \sin(m\phi)P_{22}(\cos \theta) \right) \right].$$ (18.13)

where $a$ is the length of the longest equatorial principal axis, $\theta$ is colatitude, $\phi$ is longitude, and $P_{20}$ and $P_{22}$ are two associated Legendre functions. Since the gravitational field is accurately known, the moment of inertia differences in Eqn (18.13) can be determined precisely. In order to be able to determine the three principal moments of inertia and the mean moment of inertia, a third equation is needed and can be obtained from the rotation of Mars.

Like the Earth, Mars performs a precessional motion in space on a long timescale. Its orientation in space slowly changes, like a spinning top, and the tip of its polar axis would describe a large cone in space every 171,000 years if the orbital plane of Mars were to be fixed in space. During this motion, the obliquity, or the angle between the polar axis and the normal to the orbital plane, remains constant. Precession is due to the gravitational torque of the Sun on Mars, which differs from zero because Mars is flattened at its poles. The precession rate is related to the three moments of inertia in the following way. It is proportional to the difference of the polar and mean equatorial moment of inertia since this difference is a measure of the polar flattening and it is inversely proportional to the polar moment of inertia because that quantity describes the rotational inertia for Mars rotating around its polar axis.

Since a change in the orientation of Mars also changes its external gravitational field, precession can be measured by tracking the orbital motion of a spacecraft around Mars. More directly it can be measured by recording the Doppler shifts on radio signals between a lander on Mars and the Earth. The best current value for the precession rate is $(7594 \pm 10) \times 10^{-3}$ arcsec/year. This estimate has been determined from the Doppler and range data to the Mars Pathfinder lander and the Viking landers and from radio tracking Mars Global Surveyor, Mars Odyssey, and the Mars Reconnaissance Orbiter. The polar moment of inertia factor $C/(MR^2)$ of Mars derived from the precession rate is then $C/(MR^2) = 0.3644 \pm 0.0005$. Since the polar moment of inertia is much less precisely known than the moment of inertia differences from the gravitational coefficients, the mean moment of inertia factor is known with similar precision as the polar moment of inertia factor: $I/(MR^2) = 0.3645 \pm 0.0005$.

The Sun raises tides on Mars because the acceleration induced on each mass element due to the Sun’s gravitational force is not equal to the acceleration of the mass center of Mars responsible for the orbital motion on account of the different distance to the Sun for different positions in Mars. As Mars is not a rigid body, it deforms in response to this differential gravitational force. The deformations or solid body tides are smaller than for the Earth because Mars is further away from the Sun and is smaller than the Earth. Moreover, the Moon raises larger tides on the Earth than the Sun. The radial tidal displacement on the surface of Mars is of the order of 1 cm only compared to about 40 cm for the Earth. Currently, this is too small to be measured directly by either a laser altimeter on board an orbiter or by radio tracking of a lander. Nevertheless, tides have been observed indirectly by their effect on the orbital motion of spacecraft. As tides slightly change the mass distribution in Mars, the external gravitational potential changes and so does the spacecraft.
orbit. The Love number $k_2$ describes the reaction of Mars to the tidal forcing and is defined by

$$\delta \Phi(r, t) = \left[ 1 + k_2 \left( \frac{R_s}{r} \right)^3 \right] \Phi_t(r, t). \quad (18.14)$$

Here, $t$ is time, $\delta \Phi$ is the change in the external gravitational potential due to the tides, and we have restricted the tidal potential $\Phi_t$ to its degree-two part, which is a factor $dR_s$, where $d$ is the distance between the Sun and Mars, larger than the degree-three part. By measuring small secular changes in the orbital inclination of Mars Global Surveyor, Mars Odyssey, and the Mars Reconnaissance Orbiter, the Love number has been determined to be $k_2 = 0.164 \pm 0.009$.

### 7.3. Model Results

Viable models of the interior structure of Mars must be consistent with the four observed quantities mass, radius, mean moment of inertia, and Love number $k_2$. Obviously, four quantities are not sufficient to determine all details of the interior structure and many properties cannot be constrained by them. The latter property can be used to our advantage by allowing to simplify the model. For example, the detailed structure of the crust has only a small effect on the moment of inertia and Love number. As a result, in modeling the global interior structure of Mars, the crust can be described by two parameters only: its mean thickness and mean density.

In the lowest order approach useful to understand the Martian interior, it is even possible to neglect the crust and to consider that Mars consists of only two layers with a homogeneous density: the mantle and the core. Since the radius is well known, such a two-layer model is described by three parameters: the radius $R_c$ and density $\rho_c$ of the core and the density $\rho_m$ of the mantle. The mean density $\rho$ and moment of inertia $I$ of such a Mars model can then be expressed as

$$\rho = \rho_m + (\rho_c - \rho_m) \left( \frac{R_c}{R} \right)^3 \quad (18.15)$$

$$\frac{I}{MR^2} = \frac{\rho_m}{\rho} + \frac{\rho_m - \rho_m}{\rho} \left( \frac{R_c}{R} \right)^5 \quad (18.16)$$

The Love number $k_2$ is not uniquely determined for a basic two-layer model because it also requires knowledge of the rigidity of the layers. The rigidity for silicate rocks in the Martian mantle is thought to be in the range 70–80 GPa. Since both the moment of inertia and the Love number $k_2$ are determined with a certain error and the mantle rigidity is also not exactly known, the geodesy data cannot uniquely determine the interior structure of Mars, not even for a basic two-layer model. Nevertheless, the data show that the radius of the core of Mars is $1840 \pm 65$ km and that the density of the core is $6715 \pm 270$ kg/m$^3$ (at 1σ, see Figure 18.9). The Love number puts the strongest constraint on the core because tides are highly sensitive to the distance of the liquid core from the surface. The results show that the core of Mars is about equally large as that of the Earth relative to the total size of the planet.

The core density is much smaller than the density of pure iron at core pressure and temperature. At a pressure of 30 GPa and a temperature of 2000 K, the density of pure iron is 8670 kg/m$^3$, which is 29% higher than the mean estimated density, suggesting a large amount of light elements in the core. More realistic and sophisticated models confirm the above findings for a basic two-layer model. They indicate that the core radius is $1794 \pm 65$ km (at 1σ) and that the core density is $6265 \pm 200$ kg/m$^3$. The smaller core radius and density than for the two-layer model is mainly due to the neglect of the crust and the assumption of homogeneous layers in the simplified model. If it is assumed that sulfur is the only light element in the core, the sulfur concentration is $16 \pm 2$ wt%, which is a larger light element concentration than that for the Earth. These high sulfur concentrations confirm results obtained from compositional studies of Mars, although much smaller concentrations have also been proposed for Mars’ core. As a result of the large sulfur concentration in the core, the melting temperature of the
core is strongly reduced with respect to a pure iron composition and is below the expected temperatures at the core—mantle boundary for current models of the thermal evolution of Mars (see next section). As a consequence, the core of Mars is not expected to be differentiated into a solid inner core and a liquid outer core as for the Earth, but to be entirely liquid. A lower temperature in Mars than expected or a different chemical composition of the core, with a large amount of other light elements than sulfur, which has the strongest effect on the reduction of the melting temperature of an iron alloy, might, however, be consistent with an inner core. Nevertheless, the absence of an inner core is in agreement with the absence of a global magnetic field, which is thought to require a growing inner core.

The mantle density and composition cannot be determined from the global data, although the estimate of the core radius suggests that the mantle of Mars does most likely not contain minerals occurring in the lower mantle of the Earth. Mars’ core seems to be too large to have a lower mantle like the Earth, although a thin lower mantle cannot be excluded.

8. EVOLUTION OF MARS

8.1. Thermal Evolution

Mars formed hot and differentiated into a mantle, a core, and a crust (see Section 2). The subsequent evolution is much less dramatic: during the longest part of its evolution Mars slowly cools altering only slightly and gradually its core, mantle, and outer layers. Obviously, Mars can only cool since it loses more energy than it produces in its interior, implying that its internal energy transport is effective in carrying heat from the deep interior to the surface. If we only consider the major energy sources, the energy lost from the solid surface of Mars per second, or total surface heat flow, $L$, is equal to the difference of the total energy production per second and the sum of the internal energy $U$, consisting of the heat content and the strain energy, and the gravitational energy $E_G$:

$$L = \int_0^M \frac{\varepsilon(m)}{d} \left( U + E_G \right) \, dm$$  \hspace{1cm} (18.17)

Here, $t$ is time and $\frac{dm}{dt}$ an infinitesimal mass element. Compression of Mars as a result of cooling increases the internal energy and decreases the gravitational energy, but it can be shown that these contributions almost cancel. Therefore, the total surface heat flow is approximately equal to the sum of the energy production per second by heat-producing radioactive isotopes and the loss of thermal energy.

If radiogenic heating were the only heat source of the planet and were to supply energy at a constant rate, a balance between the internal energy source and the loss of energy through the surface could exist and Mars would not need to cool. However, the radioactive isotopes decay so that the amount of heat-producing elements decreases with time. Since the internal heat produced in Mars declines with time, the surface heat flux also declines with time. The heat transported by convection thus diminishes, and therefore the temperature decreases with time, as follows from the relation between the Nusselt number and temperature. Mars therefore necessarily cools. About 30% of the heat flux of Mars is estimated to be due to cooling. A value of 25–30% is characteristic for planets with stagnant lid convection and is smaller than for the Earth, which has plate tectonics and cools more efficiently.

The cooling of Mars can be described by considering the changes in energy in the core and the mantle. Since the specific heat describes the amount of energy needed to raise the temperature of 1 kg of a material by 1 K, the thermal energy of the core is equal to the product of the mass $M_c$, the mean specific heat $C_c$, and the mean temperature $T_c$ of the core. The change in this energy with respect to time is due to heat loss to the mantle, which can be expressed as

$$M_c C_c \frac{dT_c}{d\tau} = 4\pi R_c^2 q_c,$$  \hspace{1cm} (18.18)

where $q_c$ is the heat flux out of the core. If Mars was to have an inner core, this equation would have to be extended in order to account for the latent heat and gravitational energy release of inner core formation.

For the mantle several additional energy sources and sinks have to be taken into account in the energy equation. The mantle gains energy from the decay of radioactive isotopes and the flow of heat from the core into the mantle. When mantle rock melts and solidifies or undergoes mineral phase changes, latent heat is also consumed and released. Moreover, the mantle loses heat through its upper surface. The energy equation for the mantle expresses that the change in internal energy is equal to the sum of the energy gains and the energy losses. If we neglect the contribution of the latent heat, the energy equation for the mantle can be expressed as

$$M_m C_m \frac{dT_m}{d\tau} = M_m \varepsilon + 4\pi R_m^2 q_m - 4\pi R_m^2 \varepsilon_m,$$  \hspace{1cm} (18.19)

where $T_m$ is the mean mantle temperature, $C_m$ is the mean specific heat of the mantle, and $M_m$ is the mass of the mantle. The two energy equations for the core and the mantle are the basic equations for the study of the thermal evolution of Mars. For a given initial temperature in the core and the mantle, they describe the evolution of the temperature with time.

The current heat flux out of Mars has not yet been observed and can therefore not be used to constrain the thermal evolution. However, this situation will improve in
2016 when the NASA InSight mission is foreseen to be launched to Mars. This mission carries a heat probe that will measure the heat flow in the upper few meters of Mars at the landing location. For the Earth, the current mean total surface heat flow is known to be about 45 TW. From this value, an upper limit to the decrease in temperature over recent times can easily be estimated by neglecting the contribution of the radioactive isotopes. The total change in internal energy of the Earth over the last billion year, \( \Delta E \), where \( \Delta T \) is the change in temperature, must be equal to the product of 1 billion year and the total surface heat flow. We here assume that the heat flow has not changed much, as thermal evolution models indicate. It then follows that the internal temperature of the Earth has decreased approximately by about 200 K over the last billion years. The real temperature decrease is somewhat less than half that value since about half or somewhat less of the total surface heat flow comes from the radioactive elements. For Mars, the heat flux from the surface is not yet known, but the temperature decrease can be estimated from the energy produced by the radioactive elements. For an essentially chondritic concentration, the radioactive elements currently produce about 2 TJ per second. Since about 30% of the heat flow, or about 0.86 TW, is due to cooling, the temperature reduction over the past billion years is about 35 K/Gyr, which corresponds to estimates of cooling rates from geochemical observations of the melting history in the mantle.

Since the initial temperature of Mars was probably high (see Section 2), both the mantle and the core were vigorously convecting shortly after core formation. Even if the mantle was not fully molten, the viscosity of the mantle rocks was low because of the high temperature. The Rayleigh number is then high and convection can support a high heat flux (high Nusselt number). As a result, the planet initially cools rapidly (Figure 18.10). In case of a mantle overturn after magma ocean solidification, the onset of subsolidus mantle convection may be delayed as a result of the gravitationally stable configuration of the mantle produced by the overturn. The phase of fast cooling lasted only a few hundred million years until the lower temperature led to a larger viscosity resulting in a lower Rayleigh number and less vigorous convection. Analyses of meteorites from Mars have shown that chemical heterogeneities in the silicate part are still present indicating that mantle convection has not been sufficiently vigorous to homogenize the mantle.

In the phase of slow cooling, which is still ongoing, the mantle is in a state in which temperature and viscosity take values that facilitate the removal of the heat produced by radioactivity plus part of the primordial heat. If less heat would be transported than is created internally, the mantle temperature would rise, viscosity would decrease, and more heat would be transported until at least as much heat would be transported as is produced. It is often argued that the exact initial temperature is not very important for the present-day evolution of Mars since, for any hot initial temperature, Mars will have cooled rapidly and reached the slow cooling regime in a short timescale. The higher the initial temperature is and the lower the initial viscosity, the faster will Mars cool in its early phase. Mars therefore has little memory of its initial thermal state. Nevertheless, this principle does not always apply, and temperature evolutions starting from different initial temperatures do not always converge and lead to different present-day temperatures, for example, for low Rayleigh numbers or low initial temperatures.

Mantle viscosity is very important for the thermal evolution of Mars. However, it is not well known and the reference mantle viscosity \( \eta_0 \) may vary over several orders of magnitude. In particular, viscosity increases with increasing grain size of the rocks and water reduces significantly the viscosity of mantle rocks. The grain size of the upper mantle of the Earth is about 1 mm but is not well constrained for Mars. For dry olivine, an increase in grain size from 1 mm to 1 cm increases the reference viscosity by a factor of about 300. The viscosity of rock with a water content of the Earth’s mantle is about two orders of magnitude smaller than the viscosity of dry rock. Studies of Martian meteorites show water contents ranging from as low as 1–1000 ppm of water per silicon atom. Higher reference viscosities, for example, due to less water or larger grain size, imply more resistance to convection and therefore less vigorous convection and a lower heat flux.

![FIGURE 18.10 Temporal evolution of the average mantle (\( T_m \)) and core temperature (\( T_c \)) of Mars for cold (continuous curves) and hot (dotted curves) initial temperatures. The reference viscosity of the hot initial temperature model is 20% larger than of the cold model.](image_url)
(smaller Rayleigh number and smaller Nusselt number). In order for Mars to be able to transfer the heat generated by the radioactive elements to the surface, a higher mantle temperature is needed. For an increase in the reference viscosity of one order of magnitude, the expected current mantle temperature is about 100 K higher, independent of the initial temperature. The cooling rate in the slow cooling phase is, however, almost independent of the reference viscosity because the same energy produced by the radioactive isotopes must be transferred and the relative contribution of cooling to the heat flux has a characteristic value of about 30%.

If the core cools sufficiently, the local temperature somewhere in the core will reach the freezing temperature of the core material. At that depth in the core material will start to solidify, initiating the formation of a solid iron core. Models of the thermal evolution of Mars indicate that the present-day temperature at the core—mantle boundary could be as low as 1500 K, but higher values up to over 2000 K are also not excluded. For a core composed of iron and sulfur and assuming that sulfur makes up about 16% of the mass of the core as indicated by the geodesy measurements, the core—mantle boundary of Mars still has to cool at least by about 200 K before an inner core will form. Since the cooling rate is only at most 50 K per billion year, the formation of an inner core is expected to start only in at least about 4 Gyr from now. However, if the core composition is different and other light elements are present or even contribute more to the core mass than sulfur, the melting temperature of the core can be several hundred Kelvin higher. In that case, the core could already have formed an inner core, although that might be difficult to reconcile with the observations that Mars does not contain a global magnetic field since such a field is supposed to develop once an inner core starts growing because the production of less dense liquid core material enriched in light elements at the inner core boundary is a driver for convection in the outer core. Nevertheless, a sufficiently large increase in light element concentration of the liquid at the inner core boundary is required for the maintenance of a dynamo, and this depends on how light elements partition between the solid and the liquid. At the Mars core pressures and temperatures, almost all sulfur remains in the liquid on solidification for sulfur concentrations below the eutectic concentration, but this is not the case for all other light elements.

When Mars cools, the stagnant lid thickens because less heat flows into the stagnant lid. The layers below the base of the stagnant lid will then cool, thereby thickening the lid. The upper mantle layers that become part of the stagnant lid during the evolution of Mars cool much more than the deeper mantle layers, which stay relatively hot. During the past 4 Gyr, the stagnant lid of Mars has steadily grown and may now be several hundred kilometers thick.

8.2. Early Dynamo

Based on first observations of magnetic patterns in the Southern highlands with the magnetometer onboard the Mars Global Surveyor, it has been suggested that plate tectonics might have existed on early Mars. A consequence of early plate tectonics is that it could have driven a core dynamo. Early plate tectonics is, however, difficult to reconcile with the age of the crust. During the transition from plate tectonics to stagnant lid convection the mantle would heat, leading to increased partial melting and substantial crust formation. This peak in crust production would happen at a later time than the bulk of the crust is thought to have formed, rendering the hypothesis of early plate tectonics unlikely.

Without plate tectonics, the deep interior of Mars does not cool sufficiently fast to sustain a purely thermally driven dynamo. An early magnetic field on Mars must, however, have been present since otherwise the observation of remanent magnetization of Martian surface rocks cannot be explained. The magnetization was acquired in the first 500 My of Mars' history when rocks crystallized in the presence of a magnetic field. Younger rocks do not show any magnetization as the dynamo had already shut down. A more likely possibility is that the release of gravitational energy associated with core formation heated the core to temperatures a few hundred Kelvin higher than the temperature at the bottom of the mantle. In that case, the core could cool rapidly enough to be convecting, providing a sufficiently large source of kinetic energy to drive the dynamo. Alternatively, a large impactor penetrating to the core might have superheated the core and triggered core convection and a dynamo. Such dynamos could only operate for a short time, at most a few hundred million years because of the fast energy transport and cooling. Once the decreasing heat flux out of core reaches the value that can be conducted through the iron core at the temperature gradient corresponding to an adiabatic profile, the dynamo shuts down.

8.3. Chemical Evolution

8.3.1. Crust Formation

During its cooling history, the mantle of Mars chemically evolves because of partial melting. When mantle material ascends beneath the stagnant lid as a result of mantle convection, both the temperature and the melting temperature experienced by the upwelling material decrease because of the lower pressure exerted on the rocks. Since the temperature in the rising material element decreases slower than the melting temperature, the rocks can reach a region where their temperature is higher than the melting temperature and partially melt. Only part of a rock melts because rock is
composed of several minerals, each with a different melting temperature. Through volcanism the molten rock can rise above the base of the upper thermal boundary layer and will recrystallize in the crust or at the surface, thereby contributing to the growth of the crust.

For rocks to partially melt, the temperature in Mars must be higher than the solidus temperature. As a result of cooling the occurrence of partial melting therefore decreases with time. Moreover, the thickening of the stagnant lid also moves the melt zone to greater depths. As the melting temperature increases faster with depth than the mantle temperature, melting is less likely in deeper mantle layers. Because most of the crust was formed in the first 100 My, partial melting must have been strongly reduced from then on. Initial mantle temperatures well above 2000 K and high Rayleigh numbers are thought to be implausible as they would lead to a thicker crust than observed and might even destroy the crustal dichotomy. The state of the crust therefore sheds light on the initial conditions of Mars.

Locally, volcanism has, nevertheless, continued to create new crust, in particular in the Tharsis region, with evidence of lava flows not older than a few million years. Volcanism requires melting and penetration of melt into the stagnant lid, which becomes increasingly unlikely with increasing stagnant lid thickness but local melting due to hot upwelling mantle plumes seems not excluded. Local magmatism observed in the Tharsis region might also be a consequence of the locally thick crust. If the crust is sufficiently enriched in radioactive elements and has a significantly lower thermal conductivity than the mantle, the crust locally better insulates the mantle leading to higher local temperatures in the upper mantle and partial melting.

8.3.2. Extraction of Water

Besides crustal formation, partial melting also leads to the extraction of incompatible elements, such as the radioactive elements and also water. About 40–80% of the water of the mantle is thought to be removed in the first 500 million years and probably outgassed through extrusive volcanism. The loss of water from the mantle increases the mantle viscosity, resulting in less vigorous convection. The chemical evolution of the mantle, which is driven by the cooling, therefore also affects the thermal evolution.

Although the initial water concentration is not known, estimates of the volume of water outgassed from the mantle correspond to a global water layer of tens to hundreds of meters on the surface of Mars. This water extracted in the first half billion year may have been important for the early surface conditions and atmosphere of Mars, and therefore for the habitability of early Mars.

BIBLIOGRAPHY


1. INTRODUCTION TO MARS EXPLORATION

Most of our detailed information about the materials that make up the Martian surface comes from in situ investigations accomplished by seven successfully landed spacecraft (Table 19.1). The focus of these spacecraft and the era in which they explored Mars have varied, but all have been preceded by orbiters that acquired remote sensing data that helped frame the questions they addressed and the locations where they landed. The first successful landings were the Viking spacecraft in 1976, part of two orbiter/lander pairs that were launched in 1975. These landers were preceded by a number of failed Soviet and United States spacecraft, several successful “flyby” missions, and the Mariner 9 orbiter that provided basic imaging and spectral information that gave an early view of the surface and atmosphere of Mars. The overriding impetus for the Viking landers was to determine if life existed on Mars. Both immobile, legged landers carried sophisticated life detection experiments as well as imagers, seismometers, atmospheric science packages, and magnetic and physical properties experiments. The Viking mission was done in the post-Apollo era (after 1972) and involved a massive mobilization of engineering and scientific talent (as well as a budget befitting a major mission). Both landers carried arms with scoops that collected soil and fed them into the life detection experiments. No unequivocal evidence for life was found in the soil (although gases released from the soil suggested a significant oxidizing component), but the spacecraft imaged the landing sites, determined the chemistry of the soils, and provided a long record of surface meteorology.

The successful landings and operations of the orbiters (that lasted years) set the stage for the systematic study of Mars that fueled our modern view of the Red Planet. The spacecraft also left a legacy for landing using aeroshields and supersonic parachutes that have been
employed by all subsequent landers. Orbiter and lander data defined the atmosphere and basic geology of Mars. By mapping morphologic units and crater density, three main eras of Martian geologic history were defined: Noachian (>3.6 billion years ago), Hesperian (3.6–3.0 billion years ago), and Amazonian (since 3.0 billion years ago). The mappings showed that Mars was very geologically and tectonically active during the Noachian, with decreasing activity into the Hesperian and Amazonian. The Viking orbiters returned images of valley networks and eroded ancient craters in Noachian terrain that suggested an earlier wetter and possibly warmer environment and the onset of freezing conditions in the Hesperian, leading to the present climate in the Amazonian that is generally too cold and thin (and dry) to support liquid water (current atmospheric pressure and temperature are so low that water is typically stable only in solid and vapor states).

The Mars Pathfinder (MPF) mission, launched 20 years later in 1996, was an engineering demonstration of a low-cost lander and small mobile rover and on landing on July 4, 1997, ushered in our modern era of Mars exploration. The spacecraft was a small free flyer that used a Viking-derived aeroshell and parachute, but employed newly developed robust airbags surrounding a tetrahedral lander, rather than retrorockets and legged landers as did Viking. The lander carried a stereoscopic color imager (Imager for Mars Pathfinder [IMP]), which included a magnetic properties experiment and wind sock and an atmospheric structure and meteorology experiment. The 10-kg, microwave-size rover (Sojourner) carried engineering cameras, 10 technology experiments, and an Alpha Proton X-ray Spectrometer (APXS) for measuring the chemical composition of surface materials, and conducted 10 technology experiments. The MPF lander and rover operated on the surface for about 3 months (well beyond their design lifetime) and the rover traversed about 100 m around the lander, exploring the landing site and characterizing surface materials in a couple of hundred square meter area. Rocks analyzed by the APXS appeared relatively high in silica, similar to andesites; tracking of the lander fixed the spin pole and polar moment of inertia that indicates a central metallic core and a differentiated planet, and the atmosphere was observed to be quite dynamic with water ice clouds, abruptly changing near-surface morning temperatures, and the first measurement of small wind vortices or dust devils. The mission captured the imagination of the public, garnered front-page headlines during the first week of operations, and became one of NASA’s most popular missions as the largest Internet event in history at the time. Much of the flight system, lander, and rover design were used for the next two successful landings.

Launching before MPF, but arriving later, was the Mars Global Surveyor (MGS) orbiter, which was a partial relift of instruments on the Mars Observer orbiter that was lost when attempting to enter into orbit around Mars in 1993. This spacecraft defined the global topography and magnetic field and identified different rock types and minerals that make up the surface. It also identified layered sedimentary rocks in high-resolution images suggesting deposition in standing bodies of water and fresh gullies suggesting recent flow of liquid water. Mars Odyssey (2001) followed MGS and the failed Mars Climate Orbiter and Mars Polar Lander launched in 1999. Instruments on Odyssey identified ground ice at high latitudes, produced the highest resolution global image mosaic (100 m/pixel) to date, and with MGS improved our knowledge of the atmosphere and global physical and mineralogical properties of surface materials by measuring their thermal properties and infrared spectral characteristics.

### Table 19.1 Landing Sites on Mars

<table>
<thead>
<tr>
<th>Site</th>
<th>Latitude (Degree +N)</th>
<th>Longitude (Degree +E)</th>
<th>Elevation (km, MOLA)</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viking Lander 1 (VL1)</td>
<td>22.27</td>
<td>311.81</td>
<td>−3.6</td>
<td>Chryse Planitia</td>
</tr>
<tr>
<td>Viking Lander 2 (VL2)</td>
<td>47.67</td>
<td>134.04</td>
<td>−4.5</td>
<td>Utopia Planitia</td>
</tr>
<tr>
<td>Mars Pathfinder (MPF)</td>
<td>19.09</td>
<td>326.51</td>
<td>−3.7</td>
<td>Ares Vallis</td>
</tr>
<tr>
<td>MER Spirit (SPI)</td>
<td>−14.57</td>
<td>175.47</td>
<td>−1.9</td>
<td>Gusev Crater</td>
</tr>
<tr>
<td>MER Opportunity (OPP)</td>
<td>−1.95</td>
<td>354.47</td>
<td>−1.4</td>
<td>Meridians Plantum</td>
</tr>
<tr>
<td>Phoenix (PHX)</td>
<td>68.22</td>
<td>234.25</td>
<td>−4.1</td>
<td>Vastitas Borealis (high northern plains)</td>
</tr>
<tr>
<td>MSL Curiosity (MSL)</td>
<td>−4.59</td>
<td>137.44</td>
<td>−4.5</td>
<td>Gale Crater</td>
</tr>
</tbody>
</table>

MOLA, Mars Orbiter Laser Altimeter on Mars Global Surveyor.
The Mars Exploration Rover (MER) mission landed twin golf cart-sized rovers in early 2004 that have explored over 40 km of the surface at two locations. Each rover carried a payload that contains multiple imaging systems including the color, stereo Panoramic Camera (Pancam) and Miniature Thermal Emission Spectrometer (Mini-TES) for determining mineralogy. The rovers also carried an arm that can brush and grind away the outer layer of rocks (the Rock Abrasion Tool (RAT)) and can place an Alpha Proton X-ray Spectrometer (APXS), Mössbauer Spectrometer (MB), and Microscopic Imager (MI) against rock and soil targets (Table 19.2). The rover and payload partially mimics a field geologist, being able to identify interesting targets using the remote sensing instruments (a field geologist’s eyes), rove to those targets (legs), remove the outer weathering rind of a rock (equivalent to a rock hammer), and identify the rock type (equivalent to a geologist’s hand lens and analysis in the laboratory) using the chemical composition (APXS), iron mineralogy (MB), and rock texture (MI). These rovers have lasted years (well beyond their 3-month design lifetime) and returned a treasure trove of basic field observations along their traverses as well as sophisticated measurements of the chemistry, mineralogy, and physical properties of the rocks and soils encountered. They have returned compelling information that indicates an early wet and likely warm environment on Mars.

Mars Express, the first European Space Agency mission, also carried the British Beagle 2 exobiology lander to Mars, arriving in late 2003. Although the lander was not successful, the orbiter has observed Mars for almost 10 years. Mars Express carries imagers, imaging spectrometers, radar sounders, and atmosphere and exosphere sensors. Stereo color images have refined the geologic history of Mars and the first visible to near-infrared imaging system discovered clay minerals that formed by alteration of primary volcanic minerals in neutral waters in the ancient terrains in agreement with an early warmer and wetter Mars.

The Mars Reconnaissance Orbiter (MRO) was launched in 2005 and carries imagers capable of resolving meter-size features on the surface (25 cm/pixel), images that cover broad regions at 6 m/pixel, and a higher resolution (18 m/pixel) visible and near-infrared spectral imager. It has confirmed widespread deposits of clay minerals in the ancient highlands, refining our understanding of water activity on Mars. It has also sounded the atmosphere to provide a much better understanding of its temperature, pressure, and density variations with

<table>
<thead>
<tr>
<th>TABLE 19.2 Instruments Used to Process and Analyze Rocks and Soils at Spacecraft Landing Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alpha Particle X-ray Spectrometer (APXS)</strong> on Mars Exploration Rovers and MSL: measures rock elemental chemistry using interactions of alpha particles with the target</td>
</tr>
<tr>
<td><strong>Alpha Proton X-ray Spectrometer (APXS)</strong> on MPF: measured rock elementary chemistry, using interactions of alpha particles and protons with the target</td>
</tr>
<tr>
<td><strong>ChemCam on MSL</strong>: fires a laser and analyzes the elemental abundances of vaporized areas on rocks and soils</td>
</tr>
<tr>
<td><strong>ChemMin on MSL</strong>: a powder X-ray diffraction instrument used to identify minerals</td>
</tr>
<tr>
<td><strong>Gas Chromatograph/Mass Spectrometer (GCMS)</strong> on Viking: instruments that analyzed chemical compounds in soils</td>
</tr>
<tr>
<td><strong>IMP</strong>: a lander-mounted digital imaging system for stereo, color images and visible near-infrared reflectance spectra of minerals</td>
</tr>
<tr>
<td><strong>Mars Hand Lens Imager (MAHLI)</strong> on MSL: a camera that provides close-up views of the textures of rocks and soil</td>
</tr>
<tr>
<td><strong>Mast Camera (MASTCAM)</strong> on MSL: a digital imaging system for stereo color images and visible near-infrared reflectance spectra of minerals</td>
</tr>
<tr>
<td><strong>MI on Mars Exploration Rovers</strong>: a high-resolution camera used to image textures of rocks and soil</td>
</tr>
<tr>
<td><strong>Microscopy, Electrochemistry, and Conductivity Analyzer (MECA)</strong> on Phoenix: includes a wet chemistry laboratory, optical and atomic force microscopes, and a thermal and electrical conductivity probe</td>
</tr>
<tr>
<td><strong>Mini-TES on Mars Exploration Rovers</strong>: identifies minerals via thermal infrared spectral characteristics produced by crystal lattice vibrations</td>
</tr>
<tr>
<td><strong>MB on Mars Exploration Rovers</strong>: identifies iron-bearing minerals and distribution of iron oxidation states by measuring scattered gamma rays</td>
</tr>
<tr>
<td><strong>Pancam on Mars Exploration Rovers</strong>: digital imaging system for stereo color images and visible near-infrared reflectance spectra of minerals</td>
</tr>
<tr>
<td><strong>RAT on Mars Exploration Rovers</strong>: brushes or grinds rock surfaces to reveal fresh interiors</td>
</tr>
<tr>
<td><strong>SAM on MSL</strong>: suite of three instruments (mass spectrometer, gas chromatograph, tunable laser spectrometer) used to identify carbon compounds and to analyze hydrogen, oxygen, and nitrogen</td>
</tr>
<tr>
<td><strong>Sampling System (SA/SPaH)</strong> on MSL: includes a drill, brush, soil scoop, and sample processing device</td>
</tr>
<tr>
<td><strong>Surface Stereo Imager (SSI)</strong> on Phoenix: digital imaging system for stereo color images and visible near-infrared reflectance spectra of minerals</td>
</tr>
<tr>
<td><strong>Thermal and Evolved Gas Analyzer (TEGA)</strong> on Phoenix: furnace and mass spectrometer to analyze ice and soil</td>
</tr>
<tr>
<td><strong>X-ray Fluorescence Spectrometer (XRFS)</strong> on Viking: instrument that analyzed elemental composition of soils</td>
</tr>
</tbody>
</table>
altitude, which has dramatically improved our knowledge of the atmosphere that is important in landing spacecraft. The Phoenix lander was a low-cost refly of a lander originally developed to be launched in 2001 that landed in the high northern plains in 2008. It carried a variety of imagers and meteorology instruments, but its main goal was to measure the chemistry of the soil and shallow ground ice believed to be in equilibrium with the present-day climate. It did find ice several centimeters beneath the surface and found a surface that is heavily modified by the ice. The instruments discovered low levels of calcium carbonate and perchlorate salts in the soils, both arguing for aqueous processes in the past.

The Mars Science Laboratory (MSL) rover is a major mission designed to determine if Mars was habitable in the past. MSL is a mobile laboratory with remote sensing instruments and in situ instruments that can be placed against rocks and surface materials. MSL carries a drill designed to feed material to sophisticated laboratory instruments that measure the mineralogy and geochemistry of surface materials and, for the first time since Viking, organic molecules. It landed on Mars in 2012 in Gale crater and is designed to last several years and traverse tens of kilometers. It is the first spacecraft that used aeromaneuvering and entry guidance during flight on the aeroshell to dramatically reduce the size of the landing ellipse (25 km compared to >100 km for all previous landers). The small landing ellipse (the uncertainty from entry, descent, and landing to a targeted location) and long roving capability make this mission the first to consider “go to” landing sites in which landing occurs in smooth, flat terrain next to areas of prime scientific interest (that are too hazardous to land). As of this writing, the Curiosity rover is in the middle of its surface exploration, but has already discovered conglomerates that formed in surface running water, sandstones and mudstones deposited in streams and lakes, and clays, indicative of a habitable environment.

Two missions are presently under development that will continue the exploration of Mars. The low-cost MAVEN orbiter, launched in 2013, will study the upper Martian atmosphere to determine atmospheric escape rates as a clue to how the atmosphere evolved from a possibly warmer and wetter (thicker) state early on to its current cold and dry (thin) state. Finally, the low-cost Interior Exploration Using Seismic Investigations, Geodesy, and Heat Transport (InSight) mission will land a seismometer, heat flow probe, and precision tracking station in 2016 to measure the overall structure of the interior to better understand the accretion and differentiation of the rocky planets.

2. LANDING SITES ON MARS

The seven landing sites (Table 19.1) that constitute the “ground truth” for orbital remote sensing data on Mars were all selected primarily on the basis of science and safety considerations. Because a safe landing is required for a successful mission, the surface characteristics must meet the engineering constraints based on the designed entry, descent, and landing system. The most important factor controlling the selection of the seven landing sites is elevation, as all landers used an aeroshell and parachute to slow them down and sufficient atmospheric density and time are required to carry out entry and descent. This favored landing at low elevations is shown in Figure 19.1, which illustrates the locations of the landing sites on a topographic map of Mars.

The map shows that the southern hemisphere is dominated by ancient heavily cratered terrain estimated to be more than 3.6 billion years old (Noachian). The northern hemisphere is dominated by younger (Hesperian and Amazonian), smoother, less-cratered terrain that is on an average 5 km lower in elevation. Astride the hemispheric dichotomy is the enormous Tharsis volcanic province, which rises to an elevation of 10 km above the datum, covers one quarter of the planet, is surrounded by tectonic features that cover the entire western hemisphere, and is topped by five giant volcanoes and extensive volcanic plains (active during the Hesperian and Amazonian). The elevated Tharsis province and the cratered highlands have been too high for landing of existing spacecraft. The Viking landers landed in the northern lowlands, as did MPF and Phoenix; the Mars Exploration Rovers and MSL landed at relatively low elevations in the transition between the highlands and lowlands. The next most important factor in landing site selection is latitude, with low latitudes (±30°) favored for greater solar power (Pathfinder, Spirit, and Opportunity) and thermal management (Curiosity).

Landing site selection for the seven landers included intensive periods of data analysis of preexisting and incoming information. The Viking lander/orbiter pairs were captured into Mars orbit and the orbiter cameras started a concentrated campaign to image prospective landing sites (at tens to hundreds of meters per pixel) selected on the basis of previous Mariner 9 images. A large site selection science group assembled mosaics (using paper cutouts pasted together by hand) in real time and, after waiving off several landing sites on the basis of rough terrain and radar scattering results (and missing the intended July 4th landing), Viking 1 landed on ridged plains in Chryse Planitia. The site is downstream from Maja and Kasei Valles, giant catastrophic outflow channels that originate north of Valles Marineris, the huge extensional rift or canyon that radiates from Tharsis (Figure 19.1). The site’s low elevation and proximity to the channels suggested that water and near-surface ice might have accumulated there, possibly leading to organic molecules and life. Viking 2 was sent to the middle northern latitudes where larger amounts of atmospheric water vapor were detected, thereby ostensibly improving the chance for life. Landing was deferred for Viking 2 as well, as the site...
selection team analyzed images and thermal observations before landing in the midnorthern plains, just west of the crater Mie (Figure 19.1). Although predictions of the surfaces and materials present at the Viking landing sites were incorrect (likely due to the newness of the data and the coarse resolution of the orbital images), the atmosphere was within specifications and both landed successfully.

The MPF site selection effort involved little new data since the Viking mission 20 years earlier, but there was a much better understanding of how the two Viking landing sites related to the remote sensing data acquired by the Viking orbiters. The site selection effort took place over a two-and-a-half-year period prior to launch and included extensive analysis of all existing data as well as the acquisition of Earth-based radar data. An Earth analog in the Ephrata fan near the mouth of a catastrophic outflow channel in the Channeled Scabland of western and central Washington State was identified as an analog and studied as an aid to understanding the surface characteristics of the selected site on Mars. Important engineering constraints, in addition to the required low elevation, were the narrow latitudinal band 15° N ± 5° for solar power and the large landing ellipse (300 km by 100 km), which required a relatively smooth flat surface over a large area. This and the requirement to have the landing area covered by high-resolution Viking Orbiter images (<50 m/pixel) severely limited the number of possible sites to consider (~approximately seven). The landing site selected for MPF was near the mouth of a catastrophic outflow channel, Ares Vallis, that drained into the Chryse Planitia lowlands from the highlands to the southeast (Figure 19.2). Ares Vallis formed during the Hesperian (after the early warm and wet period) and involved outpourings of huge volumes of water (roughly comparable to the water in the Great Lakes) in a relatively short period of time (a few weeks).

The surface appeared acceptably safe, and the site offered the prospect of analyzing a variety of rock types from the ancient cratered terrain and intermediate aged ridged plains. Surface and atmospheric predictions were correct and Pathfinder landed safely.

Landing site selection for the Mars Exploration Rovers took place over a two-and-a-half-year period involving an unprecedented profusion of new information from the MGS (launched in 1996) and Mars Odyssey (launched in 2001) orbiters. These orbiters supplied targeted data of the prospective sites that made them the best-imaged, best-studied locations up to that time in Mars exploration history. For comparison, most of the ellipses were covered by ~3 m/pixel Mars Orbiter Camera (MOC) images, whereas the MPF ellipse was covered by ~40 m/pixel Viking images.

All major engineering constraints were addressed by data and scientific analyses that indicated that the selected sites were safe. Important engineering requirements for landing sites for these rovers included relatively low elevation, a latitude band of 10° N to 15° S for solar power, and landing ellipse sizes that were ultimately less than 100 km long and 15 km wide. Because of the smaller
ellipse size compared to Pathfinder, ~150 sites were initially possible from which high-science-priority sites were selected for further investigation. Both sites selected showed strong evidence for surface processes involving water to determine the aqueous, climatic, and geologic history of sites where conditions may have been favorable to the preservation of prebiotic or biotic processes. The site selected for the Spirit rover was within Gusev crater, an ancient 160-km-diameter impact crater at the edge of the cratered highlands in the eastern hemisphere. The southern rim of Gusev is breached by Ma’adim Vallis, an 800-km-long branching valley network that drains the ancient cratered highlands to the south (Figure 19.3). The smooth flat floor of Gusev was interpreted as sediments deposited in a crater lake, so that the rover could analyze fluvial sediments deposited in a lacustrine environment (Figure 19.4).

The site selected for the Opportunity rover is in Meridiani Planum in which thermal infrared spectra from orbiting Thermal Emission Spectrometer (TES) instrument indicated an abundance (somewhat unique) of dark, gray coarse-grained hematite, a mineral that typically forms in the presence of liquid water. Layers associated with the hematite deposit in Meridiani Planum suggested a sequence of sedimentary rocks that could be interleaved by the rover. Meridiani Planum is a unique portion of the ancient heavily cratered terrain in western Arabia Terra that was downwarped in response to the formation of Tharsis and heavily eroded early in Mars history and thus stands at a lower elevation than the adjacent southern highlands (Figure 19.5). The atmospheric and surface characteristics inferred from the extensive remote sensing data were correct for both, and Spirit and Opportunity landed safely.

The Phoenix lander was designed to land at the northern polar region (65°–72° N) where ground ice overlain by a few centimeters of soil had been detected by instruments on the Mars Odyssey orbiter. In addition to the low elevation of the northern plains, landing sites were initially identified (before MRO was operational) that had low rock abundance, low slopes, and a calm atmosphere. Once MRO became operational, about 1 year before Phoenix launched, the High-Resolution Imaging Science Experiment (HiRISE) began imaging these areas at about 25 cm/pixel. To everyone’s surprise and dismay, most of the high northern plains are covered by areas with dense boulder fields that could not be avoided with the large landing ellipse of Phoenix (~100 km long) and were far too rocky to safely land. Because of the low Sun angle of the HiRISE images, large rocks cast long shadows, which could be easily measured. An anxious search for rock-free areas using HiRISE ultimately identified a suitable landing site just before launch. Phoenix landed safely in an ellipse completely covered by HiRISE images and the site matched expectations with few rocks, a smooth flat plain, and a few centimeters of soil over ground ice. The mission lasted several months until the darkness and cold of polar winter encased the lander in ice.

The selection of Gale crater as the MSL landing site took over 5 years with prospective landing sites heavily targeted by MRO instruments. Engineering constraints included low latitude for thermal management of the rover and instruments, low elevation and relief, and low rock abundance. Science criteria important for the selection included the ability to assess past habitable environments, which include diversity, context, and potential biosignature
(including organic molecules) preservation. Over 50 prospective landing sites were studied and downselected to four finalists (three of which were “go to” sites), all of which have layered sedimentary rocks with spectral evidence for clays. All four sites were covered with unprecedented imaging data, dominantly from MRO, including spectral and stereo images and derived high-resolution HiRISE-derived topographic and rock maps that were

![FIGURE 19.3 Viking regional color mosaic of Ma’adim Valles and Gusev crater. The 800-km-long Ma’adim Valles, one of the largest branching valley networks on Mars, drains the heavily cratered terrain to the south and breaches the southern rim of Gusev crater. Gusev crater, which formed much earlier, is 160 km in diameter and the smooth flat floor strongly suggests that it was a crater lake that filled with water and sediments. Spirit did not identify any sediment associated with Ma’adim Valles. The cratered plains are composed of basalt flows modified by impact and eolian processes and so represent a late volcanic cover. Rocks in the Columbia Hills have been altered by water, but cannot be related to deposition in a lake associated with Ma’adim discharge.](image1)

![FIGURE 19.4 Mosaic of Gusev crater showing the landing ellipse, landing location for the Spirit rover, and the extensive data sets that were obtained to evaluate the Mars Exploration Rover landing sites. Ma’adim Valles breaches the southern rim and hills immediately downstream have been interpreted as delta deposits. The blue ellipse is the final targeted ellipse and the red X is the landing location. Background of mosaic is Viking 230 m/pixel mosaic, overlain by MOLA elevations in color. Thin image strips mostly oriented to the north—northwest are MOC high-resolution images typically at 3 m/pixel. Wider image strips mostly oriented to the north—northeast are Mars Odyssey THEMIS-visible images at 18 m/pixel. Mosaic includes 13° S—16° S latitude and 174° E—177° E longitude; solid black lines are 0.5° (~30 km) and dashed black grid is 0.1° (~6 km).](image2)

![FIGURE 19.5 Regional setting of Meridiani Planum in Mars Orbiter Laser Altimeter shaded relief map (~850 km wide). Note that smooth lightly cratered plains on which Opportunity landed (cross), which bury the underlying heavily cratered (ancient) terrain with valley networks to the south. Note that large degraded craters in the smooth plains indicate the sulfate rocks below the basaltic sand surface are very old (>3.6 billion years). In contrast, the lightly cratered basaltic sand surface that Opportunity has traversed is young. Opportunity has traversed 35 km to the large crater, and Endeavour to the southeast of the landing location.](image3)
used to run detailed landing simulations that indicated all four sites were safe. In addition, the traversability of the landing sites and target areas outside of the ellipse were evaluated, indicating that all are trafficable and that “go to” sites could be accessed within the lifetime of the mission. The Gale crater site (Figure 19.6) has a 5-km-high mound adjacent to the landing site that has layered strata that contains clays and sulfates at its base, which will be studied by the Curiosity rover after traversing out of the landing ellipse. The landing site explored so far is consistent with expectations from remote sensing data.

3. MARS LANDING SITES IN REMOTELY SENSED DATA

3.1. Surface Physical Properties

Understanding the relationship between orbital remote sensing data and the surface is essential for safely landing spacecraft and for correctly interpreting the surfaces and kinds of materials globally present on Mars. Safely landing spacecraft on the surface of Mars is obviously critically important for future landing missions. Understanding the surfaces and kinds of materials globally present on Mars is also fundamentally important to deciphering the erosional, weathering, and depositional processes that create and affect the Martian surface layer. This surface layer or regolith, composed of rocks and soils, although likely relatively thin (of order meters thick), represents the key record of geologic processes that have shaped it, including the interaction of the surface and atmosphere through time via various alteration (weathering) and eolian (wind-driven) processes.

Remote sensing data available for selecting landing sites have varied for each of the landed missions, but most used visible images of the surface as well as thermal inertia and albedo. Thermal inertia is a measure of the resistance of surface materials to a change in temperature and can be related to particle size, thermal conductivity, bulk density, and cohesion. Albedo is a measure of the solar reflectance of a surface in which the viewing geometry has been taken into account. A surface composed mostly of rocks will change temperature more slowly, remaining warmer in the evening and night, than a surface composed of fine-grained loose material that will change temperature rapidly, thereby achieving higher and lower surface temperatures during the warmest part of the day and the coldest part of the night, respectively. As a result, surfaces with high thermal inertia will be composed of more rocks or cohesive, cemented material than surfaces with low thermal inertia. Thermal inertia can be determined by measuring the surface temperature using a spectrometer that measures the thermal infrared radiance at several times during the day or by fitting a diurnal thermal model to a single radiance-derived temperature measurement. Thermal observations of Mars have been made by many orbiters, including the Mariners, Viking, MGS, and Mars Odyssey, with increasingly higher spatial resolution. Thermal inertia data have been used to map areas of the surface covered by high-inertia materials or rocks from areas covered by lower inertia materials or soil.

Global thermal inertia and albedo data combine in ways that reveal several dominant surface types. One has high albedo and very low thermal inertia and is likely dominated by substantial thicknesses (centimeters to a meter or more) of high albedo, reddish dust that is neither load bearing nor trafficable. These areas have very few rocks and have been eliminated for landing solar-powered or surface missions interested in investigating rocks or outcrop. Regions with moderate to high thermal inertia and low albedo are likely relatively dust free and composed of dark eolian sand and/or rock. Regions with moderate to high thermal inertia and intermediate to moderately high albedo are likely dominated by cemented crusty, cloddy, and blocky soil units that have been referred to as duricrust with some dust and various abundances of rocks. Coarse-resolution global abundance of rocks on Mars, derived by thermal differencing techniques that remove the high-inertia (rocky) component, shows that the high-albedo, low-inertia type of surface has almost no rocks and the other two types of
surfaces have rock abundances that vary from about 5% (the global mode of rock abundance of Mars) to a maximum of about 50% of the surface covered by rocks.

The seven landing sites sample the latter two types of surfaces in thermal inertia and albedo combinations that cover most of Mars. Along with variations in their rock abundance, they sample the majority of likely safe surfaces that exist and are available for landing spacecraft on Mars. The Viking landing sites both have intermediate to relatively high albedo, high rock abundance (~17%), and intermediate thermal inertia. On the surface, these sites are consistent with these characteristics, with both being rocky and somewhat dusty plains with a variety of soils, some of which are cohesive and cemented (Figures 19.7 and 19.8).

Prior to landing, the MPF site was expected to be a rocky plain composed of materials deposited by the Ares Vallis catastrophic flood that was safe for landing and roving and was less dusty than the Viking landing sites based on the intermediate to high thermal inertia, high rock abundance (18%), slightly lower albedo, and relation to an analogous catastrophic outflow depositional plain in the Channel Scabland. All these predictions were confirmed by data gathered by the MPF lander and rover (Figure 19.9).

The Spirit landing site in Gusev crater has comparable thermal inertia and fine-component thermal inertia and albedo to the two Viking sites and so was expected to be similar to these locations, but with fewer rocks (8%). Dark dust devil tracks in orbital images suggested that some of the surfaces would be lower albedo, where the dust has been preferentially removed (Figure 19.10). Spirit landed across both dusty (Figure 19.11) and dust devil track surfaces. It found that the average rock abundance is similar to expectations. In darker dust devil tracks,
Spirit found that the albedo is low and the surface is relatively dust free (at the landing site) compared to areas outside of dust devil tracks, where the albedo is higher and the surface is more heavily coated with bright atmospheric dust that has fallen from the sky (Figure 19.11).

The Meridiani Planum site has moderate thermal inertia, very low albedo, and few rocks. This site was expected to look very different from the three landing sites with a dark surface, little high albedo dust, and few rocks. Opportunity has traversed across a dark, basaltic sand surface with very few rocks and almost no dust (Figure 19.12). The Phoenix landing site has moderate thermal inertia and intermediate albedo and was expected to be slightly dusty with low rock abundance (from HiRISE images), all of which were confirmed at the surface (Figure 19.13). The MSL landing site (Figure 19.14) has intermediate albedo and relatively high thermal inertia (comparable to Pathfinder). The landing site is as expected, slightly dusty with low rock abundance and cemented surface materials.
The slopes and relief at various length scales that are important to landing safely were also estimated at the seven landing sites using a variety of altimetric, stereo, shape-from-shading, and radar backscatter remote sensing methods. Results estimated from these data are in accord with what was found at the surface. Of the seven landing sites, Meridiani Planum was judged to be the smoothest, flattest location ever investigated at 1 km, 100 m, and several meter length scales, which is in agreement with the incredibly smooth flat plain traversed by Opportunity (Figure 19.12). On the other extreme, the MPF landing site (Figure 19.9) was expected at the time of landing to be the roughest at all three of these length scales, which agrees with the undulating ridge and rough terrain and the more distant streamlined islands visible from the lander. The other five landing sites are in between these extremes at the three length scales, with Viking 2 (Figure 19.8) and portions of Gusev (Figure 19.10) fairly smooth at the 100-m and 1-km scale, Viking 1 slightly rougher at all three length scales, Viking 2 and portions of Gusev in between in roughness, and the Columbia Hills in Gusev roughest at the several meter length scale. Phoenix (Figure 19.13) is comparable to Viking 2 at all three scales and Gale crater (Figure 19.14) is the roughest site at the two longer scales and similar to Pathfinder in roughness at the several meter scale. All these observations are consistent with the relief observed at the surface.

The close correspondence between surface characteristics inferred from orbital remote sensing data and that found at the landing sites argues that future efforts to select safe landing sites will be successful. Linking the seven landing sites to their remote sensing signatures indicates that surface types with moderate to high thermal inertia and moderate to low albedo are both suitable for landing on Mars. Such surfaces constitute almost 60% of the planet, suggesting that to first order most of Mars is likely safe for suitably engineered landers. These results show that basic engineering parameters important for safely landing spacecraft such as elevation, atmospheric profile, bulk density, rock distribution and slope can be well constrained using available and targeted remote sensing data.

3.2. Global Compositional Units

The compositions of surface materials on Mars can be determined from infrared measurements of the planet’s surface. The TESs on the MGS and THEMIS (Thermal Emission Imaging System) on Mars Odyssey orbiting spacecraft revealed two broad spectral classes representing different compositional units. Based on spectral similarity to rocks measured in the laboratory on Earth, “Surface Type 1” material is interpreted as basaltic rock and/or sand derived from basalt (Figure 19.15). Basalt consists mostly of silicate minerals—pyroxene, feldspar (plagioclase) and olivine—and forms by partial melting of the upper mantle producing a mafic (magnesium- and iron-rich) magma that erupts on the surface as a dark lava flow (or shallow intrusion). Basalt is the most abundant type of rock on Earth, comprising the floors of the oceans and significant flooded...
areas of the continents, and it is no surprise that it is common on Mars as well. The giant shield volcanoes of Olympus Mons and the Tharsis Montes are likely composed of basalts based on their similar morphology to shield volcanoes as well as many plains that resemble basalt plains on Earth. “Surface Type 2” material is variously interpreted as either andesite or partly weathered basalt. Earth. Surface Type 2 material is variously interpreted as canoes as well as many plains that resemble basalt plains on of basalts based on their similar morphology to shield volcanoes. Olympus Mons and the Tharsis Montes are likely composed on Mars as well. The giant shield volcanoes of areas of the continents, and it is no surprise that it is common.

Most of these are secondary minerals, formed by aqueous conditions suggesting subsurface hydrothermal alteration rather than surface weathering, so they would not necessarily be indicators of climate change. Alteration of primary igneous rocks in neutral pH conditions (pH refers to hydrogen ion concentration). Orbital surveys reveal that these alteration minerals are geographically widespread but concentrated in ancient, mostly Noachian terrains of the southern highlands. Some occur in layered sequences, often showing mineralogical changes within succeeding strata (like Gale crater). It has been suggested that the Martian sedimentary record consists of distinct mineralogical epochs reflecting changes in aqueous conditions from wet, neutral pH conditions in the Noachian to highly acidic or low pH conditions later, to cold and dry in the Amazonian. In this scenario, the neutral aqueous conditions would lead to the production of clay minerals, and the highly acidic conditions would lead to the deposition of sulfates via evaporation, with cold and dry conditions in the past ~3 billion years. However, some of the clays and secondary minerals apparently formed under conditions suggesting subsurface hydrothermal alteration rather than surface weathering, so they would not necessarily be indicators of climate change.

The global distribution of Type 1 and 2 spectrally identified units on Mars is distinctive (Figure 19.17). The heavily cratered, ancient southern hemisphere of Mars is mapped mostly as Surface Type 1. In contrast, the younger northern lowlands are mapped mostly as Surface Type 2 materials. The distribution of global geochemical units is illustrated in Figure 19.17. About half of the surface of Mars is covered with a thin layer of dust, which precludes the infrared spectrometers from mapping the compositions of the rocks that underlie the dust. Some of the spacecraft landing sites on Mars are located in dusty regions. Consequently, it is difficult to compare interpretations of orbital spectra with rocks actually on the ground. The two MER
landed in a region mapped as Surface Type 1 and the Opportunity site in Meridiani was selected because of its hematite spectral signature. The Curiosity landing site was chosen for its alteration minerals (dominantly clays and sulfates), identified from visible to near-infrared spectra, and the Phoenix landing site in the high northern plains in Type 2 materials was selected for its ground ice.

### 4. LANDING SITE GEOLOGY

#### 4.1. Introduction

The geology of the seven landing sites has been investigated from color, stereo, panoramic imaging that provides information on the morphology of the landing sites; on the lithology, texture, distribution, and shape of rocks and eolian soil deposits; and on other local geologic features and landforms that are present. Landing sites on Mars are composed of rocks, outcrops, eolian bedforms, and soils, many of which are cemented. Craters and eroded crater forms are also observed at almost all the landing sites and other hills have been observed at some of the landing sites.

Our knowledge of how the surfaces at the different landing sites developed and the important geological processes that have acted on them is directly related to the mobility of the lander (arm) or rover and the ability of the lander or rover to make basic field geologic observations. The lack of mobility of the two Viking landers and their inability to analyze rocks at these sites hampered our ability to constrain their geologic evolution. Because the Phoenix lander was able to dig down into the ground ice and observe variations (Figure 19.13), it was better able to characterize the materials and important geological processes that shaped the surface. The mobility of the MPF Sojourner rover and its ability to make basic field observations over a couple of hundred square meter area resulted in a much better understanding of the geology and the events that shaped the Ares Vallis surface. The two Mars Exploration Rovers that collectively traversed over 40 km have amassed a robust suite of geologic observations over a wide area that have resulted in much better knowledge of the geologic evolution of the rocks and surfaces investigated (and the same is true for MSL). This section will review the basic geological materials found at the seven landing sites and discuss the landforms present.

#### 4.2. Rocks

Rocks are common at all the landing sites (except Meridiani). At most sites, they are distinct dark, angular to subrounded clasts that range in size from several meters in diameter down to small pebbles that are a centimeter or less in diameter. Most appear as float or individual rocks not associated with a continuous outcrop or a body of rock. Many appear dust covered and there is evidence at Gusev and Meridiani for some surface chemical alteration as is common on the Earth (see next section), where rocks exposed to the atmosphere develop an outer rind of weathered material. Although the composition of rocks could not be measured at the Viking 1 and 2 and Phoenix landing sites (Figures 19.7, 19.8 and 19.13), their dark angular and occasionally pitted appearance is consistent with a common igneous rock known as basalt. Rocks making up the cratered plains on which Spirit landed and traversed (Figures 19.10 and 19.11) for the first few kilometers are clearly basalts (see next section). The distribution and shape of many of the rocks at the Viking 1 and 2 landing sites and the Gusev cratered plains are all consistent with a surface that has experienced impact cratering with the rocks constituting the ejected fragments. Many subrounded rocks at the MPF (Figure 19.9) and Viking 1 landing site have been attributed to deposition by catastrophic floods in which motion in the water partially rounded the clasts. Some rocks at the Pathfinder site had textures that looked like layers (perhaps sedimentary or volcanic), one resembled a pillow basalt in which hot lava cools rapidly in the presence of water, and several rocks resembled conglomerates, in which rounded pebbles and cobbles were embedded in a rock. The cobbles were rounded by running water and later cemented in a finer grained matrix. Curiosity has also observed conglomerates near the surface of its landing site (Figure 19.18). Some rocks at most of the landing sites appear polished, fluted, and grooved. These are interpreted as ventifacts in which sand-sized grains, entrained by the wind, have impacted and eroded the rocks. Rocks at the Phoenix landing site are size sorted within polygonal troughs formed by thermal contraction processes in the ground ice (Figure 19.13).
4.3. Outcrops

Continuous expanses of rocks typically referred to as outcrop (or bedrock) have been observed at three of the landing sites. An area of continuous jointed rocks has been observed at the Viking 1 landing site, but little else is known about it (Figure 19.7). Outcrop has been discovered in the Columbia Hills by Spirit where there appear to be coherent stratigraphic layers in and near the Cumberland Ridge on the flank of Husband Hill (Figure 19.19). These rocks, described in the next section, appear to be layers of ejecta or explosive volcanics deposited early in Mars history. In some places the rocks are finely layered and in other places they appear massive. At Meridiani Planum light-toned outcrops are exposed in crater walls and areas where the covering dark, basaltic sand sheet is thin (Figure 19.20). These outcrops appear to be thinly laminated evaporites that formed via evaporation of subaerial salt water (see next section) early in Mars history. The layers are composed of sand-sized grains of fairly uniform composition that appear to have been reworked by the wind in sand dunes before being diagenetically altered by acid groundwater of differing compositions (see next section). Finally, continuous layers or strata of sedimentary rocks have been found and investigated by the Curiosity rover.

4.4. Soils

All the landing sites have soils composed of generally small fragments of granules, sand, and finer materials. Except where they have been sorted into bedforms by the wind, they have a variety of grain sizes and cohesion, even though their composition appears remarkably similar at all the landing sites (dominantly basaltic). Crusty to cloddy and blocky soils are also present at most of the landing sites and are distinguished as more cohesive and cemented materials. These materials appear to be the duricrust inferred to be present over much of Mars based on higher thermal inertia, but generally low rock abundance. Strong cemented light-toned duricrust was uncovered at the MPF site by Sojourner and may contribute to the higher thermal inertia at this site than the others. Some bright soil deposits outside the reach of the arm at the Viking 1 landing site (Figure 19.7(a)) show layers and hints of coarse particles that could be fluvial materials deposited by the Maja or Kasei Valles floods. Most fine particles (roughly sand sized) appear rounded to subrounded suggesting that they have been entrained in the wind and rounded when they impact the surface (see next section). Ground ice has been observed beneath the soil at the Phoenix landing site (Figure 19.13).
4.5. Eolian Deposits

Most of the landing sites have examples of eolian bedforms, or materials that have been transported and sorted by the wind. Sand-sized particles that average several hundred microns in diameter can be moved by saltation in which they are picked up by the wind and hop in parabolic arcs across the surface. Because these particles can be preferentially moved by the wind, they are effectively sorted into bedforms. Sand dunes form when sand-sized particles are sorted into a large enough pile to move across the surface. Sand dunes take a variety of forms such as barchan or crescent shaped (horns pointing downwind), star shaped from reversing winds, transverse to the wind, and longitudinal or parallel to the wind, all of which are generally diagnostic enough to be identified from orbit. Sand dunes have been identified at the MPF landing site where a small barchan dune was discovered in a trough by the rover and at Meridiani Planum where star dunes were found at the bottom of Endurance (Figure 19.21) and Victoria craters.

Ripples are eolian bedforms formed by saltation-induced creep of granules, which are millimeter-sized particles. They typically have a coarse fraction of granules at the crest and poorly sorted interiors indicating a lag of coarser grains after the sand-sized particles have been removed (Figure 19.22). Ripples have been found at the MPF, Spirit, Opportunity, and Curiosity sites. Drifts of eolian material have also been identified at many of the landing sites behind rocks as wind tails and in other configurations.

Finally, the reddish dust on Mars is only several microns in diameter and is carried in suspension in the atmosphere, giving rise to the omnipresent reddish color. Although it takes high winds to entrain dust-sized particles in the atmosphere, once it is in the atmosphere it takes a long time to settle out. Dust has been identified on the surface at all the landing sites (in addition to being in the atmosphere) giving everything a reddish color and has fallen steadily on the solar panels decreasing solar power at a similar rate.

**FIGURE 19.21** False-color mosaic of star sand dunes in the bottom of Endurance crater. Dark bluish surface is basalt with a surface lag of hematite spherules. Lighter sides of dunes are likely covered by dust that has settled from the atmosphere. Light-toned outcrop in the foreground.

**FIGURE 19.22** Large ripple called Serpent that was studied by Spirit on the cratered plains. (A) A hazard camera image showing the rover front wheels and the tracks produced by a wheel wiggle maneuver to section the drift. (B) Color image of the dusty (reddish) surface and darker more poorly sorted interior. (C) MI image of the brighter (dust cover) granule-rich surface (millimeter-sized particles) and poorly sorted, but generally finer grained basaltic sand interior. The dusty, granule-rich surface indicates that the eolian feature is an inactive (dust covered) ripple formed by the saltation-induced creep of granules, which are left as a lag.
Dust devils, or wind vortices, have been observed at the MPF and Spirit sites and appear to be an important mechanism for lifting dust into the atmosphere.

### 4.6. Craters

Impact craters are ubiquitous on Mars, so it is no surprise that craters have been imaged at most of the landing sites. At Viking 1 (Figure 19.7) and the MPF landing sites, the uplifted rims of craters have been imaged from the side. At Gusev (Figure 19.11) and Meridiani (Figure 19.23), the rovers have investigated a number of craters of various sizes during their traverses, including the interiors of some. Because impact craters resemble nuclear explosion craters and because many fresh craters have been characterized on the Moon, much is known about the physics of impact cratering and the resulting shape and characteristics of fresh craters (see chapter on Planetary Impacts). Fresh primary impact craters less than 1 km in diameter have well-understood, bowl-shaped interiors whose depth is about 0.2 times their diameter; they also have uplifted rims and ejecta deposits (Figures 19.11 and 19.20) that get less rocky and thin with distance from the crater. As a result, imaging impact craters provides clues to the geomorphologic changes that have occurred at the site such as the amount of erosion and/or deposition.

### 5. Landing Site Mineralogy and Geochemistry

#### 5.1. Rocks

Based on their appearance, rocks at the Viking and Phoenix landing sites (Figures 19.7, 19.8, and 19.13) were inferred to be basalts, but the Viking lander arms could not reach and collect rocks small enough to analyze and Phoenix could not analyze rocks, so little is known about their composition. Rocks at the MPF, Spirit, Opportunity, and Curiosity landing sites have been analyzed by a variety of rover-mounted instruments, as described in Table 19.2.

Pathfinder rock chemical compositions were analyzed by the APXS (Figure 19.24), and partial mineralogy was inferred from IMP spectra on the lander. The APXS analyzes only the outer surface (generally just a few tens of microns) of rocks. IMP images showed that the rocks were variably coated with dust. Plots of different elements versus sulfur yield straight lines, with soils plotting at the sulfur-rich end, best interpreted as mixing lines between the compositions of rocks and soil. The composition of the dust-free rock interior was inferred by extrapolating the rock composition trends to zero sulfur. The dust-free rocks have concentrations of alkalis and silica that would classify them as andesite (two different calibrations of the APXS instrument data are shown in Figure 19.16), and it was inferred from the rocks’ appearance that these were volcanic rocks. However, because the APXS analyzes only the rock surface, it is also possible that this andesitic composition represents a silica-enriched weathering rind beneath the dust rather than the composition of the rock interiors. The IMP spectra indicate the presence of iron oxides, but a more comprehensive spectral interpretation was hampered by the dust coatings.

Rocks at the Spirit landing site in Gusev crater were analyzed using a greater variety of analytical instruments (Table 19.2), aided by the RAT that can brush or grind the outer rock surface. Rocks on the plains in the vicinity of the Spirit lander are clearly basalts, in agreement with the location of Gusev crater within an area mapped by TES as Surface Type 1. Some of these rocks are vesicular—pocked with holes that were once gas bubbles exsolved from magma, and most rocks are coated with dust (Figure 19.19). Spectra from Pancam, Mini-TES, and MB of relatively dust-free or RAT-abraded rocks provide a consistent picture.
of the minerals that comprise these basalts—olivine, pyroxene, and iron oxides. All the spectra from these instruments are dominated by minerals containing iron and magnesium. Chemical compositions of basalts of plains measured by APXS not only support the presence of olivine, pyroxene, and oxides but also suggest abundant feldspar (plagioclase) and phosphate, which cannot be seen by other spectra. The APXS analyses confirm that the rocks on the plains of Gusev (Figure 19.16) are basalts (called the Adirondack class), especially rich in olivine (and hence lower in silica). Abundant dark crystals interpreted to be olivine can be seen in MI images of RAT holes in the rocks (Figure 19.25). Surface alteration rinds and veins cutting through the interiors of these rocks can also be clearly seen in some MI images, suggesting limited interactions of the rocks with water.

After analyzing rocks near the landing site, the Spirit rover traversed about 3 km across the plains and climbed Husband Hill, a promontory within Gusev crater (one of the Columbia Hills in Figure 19.11). The Hills outcrops are distinct from the basalts of the plains. Some are massive, others are laminated, and most are altered and deeply weathered (Figure 19.26). Pancam, Mini-TES, and MB spectra suggest highly varying mineralogy. Some rocks contain combinations of olivine, pyroxene, feldspar, and iron oxides (as on the plains), whereas others contain large amounts of glass, sulfate, ilmenite, and phosphate. APXS analyses have been used to divide the rocks into several different classes according to their chemistry, but the mineralogy can vary considerably even within a class. Some rocks appear to be relatively unaltered, but most show very high contents of sulfur, phosphorus, and chlorine, suggesting a high degree of alteration. The chemical compositions of these rocks are not illustrated in Figure 19.16, because this classification is only applicable to unaltered igneous rocks. The textures of Hills rocks, as revealed by the MI, are also highly variable but commonly indicate alteration of rocks composed of angular particles and clasts (Figure 19.26). RAT grinding indicates that these rocks are much softer than the basalts of the plains. They have been interpreted as mixtures of materials formed by impacts or explosive volcanic eruptions, and subsequently altered by fluids. Two classes of rocks on the northwest flank of Husband Hill have what appear to be roughly concordant dips to the northwest suggesting a stratigraphy (Figure 19.19). The lower rock has layered materials and angular to rounded clasts in a matrix that compares favorably to impact ejecta that has been altered by water to various extents. The upper rock is a finely layered sedimentary rock that has been cemented by sulfate, but the aqueous alteration did not affect the basaltic character of the sediment. A few distinctive rock types found as loose stones (geologists call these “float”) in the Hills include Backstay, Irvine, and Wishstone, which are dark, fine-grained basaltic rocks with compositions distinct from the basalts of the plains (Figure 19.16) and only limited signs of alteration by water. These rocks appear to have formed by removal of crystals from magmas similar in composition to basalts of the plains.

Once Spirit gained the crest of Husband Hill, it traveled down the south face, encountering olivine-rich rocks of the Algonquin class (Figure 19.16). Although not recognized until later, carbonate-bearing rocks were also discovered on Husband Hill. Upon reaching the bottom, Spirit traversed an area containing highly vesicular rocks (scoria) to Home Plate, tentatively interpreted as a small volcanic edifice formed of ash. Outcrops of silica-rich rocks at Home Plate are thought to have formed by precipitation under hydrothermal conditions. Similar environments on Earth are habitable and the deposition of silica provides a ready mechanism for preservation of fossils. The compositions of all the relatively unaltered igneous rocks in Gusev crater are rich in alkalis and low in silica (Figure 19.16), allowing their classification as alkaline rocks. These are the first alkaline rocks recognized on Mars.
Rocks at the Opportunity landing site are mostly exposed in the walls of impact craters and where the sand is thin. Outcrops in Eagle crater were studied extensively after landing (Figure 19.12), and thicker stratigraphic sections in Endurance and Victoria craters were analyzed later in the mission (Figure 19.20). Pancam and MI images (Figure 19.27) show that the rocks are finely laminated and sometimes exhibit cross-bedding (Figure 19.28), and RAT grinds indicate that they are very soft. At the microscopic scale, they consist of sand grains bound together by fine-grained cement. Small gray spherules, called “blueberries” (Figures 19.27 and 19.28), are embedded within the rock (the spherules are actually gray, but appear bluish in many false-color images). Some parts of the outcrop also exhibit tabular voids (Figure 19.27). APXS analyses of these rocks indicate very high concentrations of sulfur, chlorine, and bromine (highly water-soluble elements), demonstrating that the cement and sand (partially) consists of sulfate and halide salts. MB spectra reveal the presence of iron sulfate, and Mini-TES spectra suggest that magnesium and calcium sulfates also occur. The spherules are at least half hematite, the mineral seen from orbital TES spectra of the Meridiani region. The rocks are interpreted as sandstones composed of dirty evaporites of basaltic and sulfate composition formed by the evaporation of brines. Their textures suggest repeated cycles of flooding, exposure, and desiccation. Exposure and desiccation allowed some of the sediments to be mobilized into sand dunes (Figure 19.29). After deposition, the rocks underwent a number of different phases of diagenesis by groundwater of varying composition that circulated through the rocks, mobilizing and reprecipitating iron in the form of hematite spherules (concretions) and dissolving highly soluble minerals to leave the voids. Subsequently studied rocks in Victoria crater, 6 km from Endurance crater, are very similar in composition, indicating that the same stratigraphic sequence occurs over a wide area. Certain iron-bearing minerals present in the sulfates indicate that the water involved in their formation was highly acidic.

Several unusual rocks discovered by Opportunity deserve special mention. Bounce Rock, so named because the lander bounced on it as it rolled to a stop, was discovered on the Meridiani plains as the rover exited Eagle crater. Its chemical composition, as measured by APXS, is remarkably like the compositions of a group of Martian basaltic meteorites called shergottites (Figure 19.16). Its mineralogy is dominated by
FIGURE 19.27 Images of Meridiani outcrops acquired by the Opportunity rover. (A) Pancam image of Guadalupe in Eagle crater, after RAT grinding. Notice slightly redder, dustier surface around the circular RAT hole and small hematite spherules protruding from the outcrop. (B) Microscopic image of Guadalupe RAT hole, showing blueberries (dark circles) and tabular voids produced by dissolution of soluble minerals. (C) Pancam image of Ontario in Endurance crater, after RAT grinding (circular smooth area). (D) Mosaic of microscopic images of Ontario, showing fine laminations, tabular voids, and a few blueberries (dark circles).

FIGURE 19.28 MI image mosaic of the Upper Dells in Eagle crater showing fine sand-sized particles making up the laminations, blueberries, and cuspate or curved cross-laminations that indicate that the sand-sized particles were deposited by running water.
pyroxenes and plagioclase, as are shergottites. This rock is obviously not in place and was probably lofted in as ejecta from a large impact crater to the south. Marquette Island is an unusual mafic rock that also was probably ejected from elsewhere on Mars. Heat Shield Rock, named for its proximity to the heat shield discarded during descent of the Opportunity lander, is likewise an interloper in this terrain. The Opportunity instruments revealed that it is an iron meteorite, composed of iron–nickel alloys, similar to some iron meteorites that fall to the Earth (see Meteorites chapter). Several other iron meteorites, some large in size and displaying ablation features formed as they came through the atmosphere as well as subsequent weathering features, were also found by Opportunity along its extended traverse.

Rocks in the rim of Endeavour crater studied by Opportunity (Figure 19.23) are impact breccias, similar to suevites in terrestrial impact craters. Localized enrichments in zinc suggest that some breccia materials were affected by hydrothermal alteration, and veins of gypsum arise from low-temperature aqueous fluids in these deposits. Light-toned clay bearing rocks were found beneath the breccias, pointing to low-pH aqueous conditions in the Noachian.

The Curiosity landing site is littered with small rocks that appear similar to a desert pavement or lag, similar to the Spirit landing site, left by winnowing away smaller sand-sized particles. Rover images show subsurface layers of strong conglomerates composed of subrounded particles cemented together that are related to the alluvial fan observed in orbital images (Figure 19.18). An angular, dark rock analyzed with APXS has a composition similar to an alkali-rich volcanic rock (mugearite). Clay minerals, composed of at least 20% of a mudstone outcrop in a network of stream channels, have been identified definitively by the ChemMin instrument. Curiosity’s Sample Analysis at Mars (SAM) instrument also found a mixture of chemicals representing different oxidation states in the rock implying a habitable environment. Orbital spectra indicate that more clay- and sulfate-bearing sedimentary rocks will be found at the base of the Gale crater mound.

5.2. Soils

In addition to numerous soil analyses by the MPF, Spirit, Opportunity, and Curiosity rovers, soils were collected by scoops and analyzed at the two Viking landing sites and the Phoenix landing site. As defined by soil scientists on the Earth, “soil” usually contains a component of organic matter formed by decayed organisms. Soils on Mars do not contain measurable organic materials, but the term “soil” is nonetheless commonly used in planetary science (“regolith” is also used for the surface layer formed by the destruction of rocks).

Soil and dust on Mars are distinguishable based on particle size and spectral and thermal properties, although these materials are often comingled. Soil, normally dark, represents deposited materials, commonly of sand-sized grains (Figure 19.22). Bright reddish dust is much finer grained (several microns in size), and can either be suspended in the atmosphere or deposited on the ground. The top surface of soil is usually a thin layer of reddish dust, as seen by the color change when it is disturbed in rover tracks (Figure 19.22) or airbag bounces. Most measurements of soil mineralogy or chemistry represent a mixture of soil and dust, sometimes with an admixture of small particles of the local rocks.

At all these sites, the soils have broadly similar compositions, consisting of basaltic sands mixed with fine-grained dust and salts. Pancam and MB spectra of bright dust are dominated by nanophase ferric oxides, especially hematite, while Mini-TES spectra show evidence for plagioclase, minor carbonate, and an unidentified hydrous phase. MB spectra of dark soils indicate abundant olivine, pyroxene, and magnetite at the MER landing sites. The degree of alteration appears to be limited, but fractionation of chlorine and bromine in some soils suggest some mobilization by water. APXS chemical analyses show that plagioclase is also an important component of soils, and that their compositions resemble basalts with extra sulfur, chlorine, and bromine. At the Pathfinder site, local andesitic rock fragments are present in varying amounts, and at the Opportunity site, hematite spherules occur abundantly at the surface as a lag of granules. Trenches dug by the Spirit and Opportunity rovers reveal clods, suggesting greater
proportions of salts that precipitated in the subsurface have sand bound into weakly cohesive near-surface layers or clumps, and APXS analyses of some subsurface soils show high concentrations of magnesium sulfate salt. Soils also contain significant amounts of nickel, which may reflect admixture of meteorite material into the regolith. Dust appears similar in composition to the soil (basaltic). Analysis of dust adhered to magnets on the rovers indicates that it contains olivine, magnetite, and a nanophase iron oxide (likely hematite) that suggests the dust is an oxidation or alteration product of fine-grained basalt. The presence of olivine in the dust suggests that liquid water was not heavily involved in its formation as it would have readily changed to other minerals (especially serpentine) in the presence of water.

An unusual silica- and titanium-rich soil, likely the result of abrasion of a silica-rich outcrop by the Spirit rover’s stuck wheel, was discovered at Home Plate in Gusev crater. Soil at the Phoenix polar landing site is also distinct from low-latitude surface sediments. Although Phoenix did not have the capability to analyze bulk soil chemistry, it did measure unusual ratios of water-soluble elements, the presence of calcium carbonate, and the surprising occurrence of perchlorate ions, possibly formed by photochemical reactions in the atmosphere.

6. IMPLICATIONS FOR THE EVOLUTION OF MARS

6.1. Origin of Igneous Rocks

Igneous rocks form by partial melting of the planet’s deep interior. The significance of the olivine-rich basaltic compositions found by Spirit on the Gusev cratered plains is that they appear to represent “primitive” magmas formed by melting in the mantle. Most magmas partly crystallize as they ascend toward the surface, losing the crystals in the process, so that the liquid progressively changes composition. Primitive magmas retain their original compositions and thus reveal the nature of their mantle source regions.

It is unlikely that rocks with andesitic composition at the MPF landing site formed by partial melting of the mantle, unless the mantle contains large quantities of water-bearing minerals. More likely, andesite lavas would form by partial melting of previously formed basaltic crustal rocks (the crust forms an outer layer above the mantle). A more likely alternative, previously mentioned, is that these rocks are not really andesites at all, but instead are basalts with silica-rich weathering rinds. The latter idea seems especially plausible considering that Surface Type 2 (andesitic) rocks are found primarily in places (like the northern lowlands) where surface waters would have collected and the sediments they carried would have been deposited. If this is correct, the orbital data and the samples of rocks at the various landing sites strongly argue that Mars is a basalt-covered world. Basalts, sediments derived from basalts, and dust derived from mildly weathered basalts are confirmed or suspected of the landing sites. Adding the thermal emission spectra of Type 1 and Type 2 materials as basalt and weathered basalt would suggest that most of Mars is made of this primitive volcanic rock.

6.2. Chemical Evolution and Surface Water

The iron-bearing sulfate jarosite is one of the minerals that formed in outcrops of evaporites at the Opportunity landing site. This mineral could only have precipitated from highly acidic water. Any sea at Meridiani was more like battery acid than drinking water. Given the abundance of basaltic lavas on the Martian surface, it is surprising that these waters would be so acidic. Reactions between water and basalt on the Earth tend to produce neutral to basic solutions. On Mars, huge volumes of sulfur and chlorine emitted from volcanoes must have combined with water to make sulfuric and hydrochloric acids. Only a few locations on the Earth mimic this kind of fluid—mostly areas devastated by acidic waters released by weathering of sulfides that drained from mines. Acidic water dissolves and precipitates different minerals than the neutral waters we are more familiar with on the Earth. Carbonates are not precipitated, and iron sulfates are more common in acidic solutions.

The presence of significant amounts of sulfate and chloride in soils from all the landing sites further suggests that acidic waters may have been common at one time in many places on Mars. Either evaporites like those at Meridiani Planum were abundant and have been redistributed as small particles throughout the planet’s regolith or they occur as cements formed by groundwater leaching all over Mars. Results from the visible to near-infrared spectrometers on Mars orbiters support the finding of abundant sulfates elsewhere on Mars.

The occurrence of clay minerals in ancient highlands rocks is inconsistent with acidic fluids and instead suggests neutral to slightly basic water. This could signify a change in environmental conditions. Because the clay minerals generally appear in terrains older than the sulfates, this change in chemistry could be associated with decreasing amounts of water that became more acid at the tail end of a possibly wet period, prior to the dry modern era. The presence of carbonates in the soil, carbonate discovered by Spirit in the Columbia Hills, and alteration carbonates in the ancient, heavily cratered terrain suggest that acidic fluids were not present and/or pervasive everywhere (or the carbonates would have been destroyed).

6.3. Weathering on Mars

There is considerable controversy about the degree to which Mars rocks are weathered. Weathering by acidic
water preferentially attacks olivine, and the surface layers and rinds of rocks at the MER sites appear to be depleted in olivine. However, remote sensing indicates that olivine is a common mineral in many places on Mars, and olivine appears to be a ubiquitous constituent of Martian soils and dust. Perhaps weathering was more common in the distant past, when acidic waters were abundant and produced outcrops like those found by Opportunity. Then the acid waters disappeared, and since that time, the lavas that were erupted and the soils that formed have only experienced limited weathering.

Visible and near-infrared spectral data indicate that clay minerals occur in some localities in the ancient terrains of Mars. Clays have been suggested to be present in some rocks on Husband Hill (Gusev crater) and at Endeavour crater (Meridian Planum), based on aspects of their chemistry. Their occurrence was measured directly by ChemMin in Gale crater. Clay minerals can form by weathering, but they may also form by subsurface hydrothermal activity. Weathering processes clearly occurred at the Phoenix landing site. The soils there are clearly chemically altered, and soil particles are bound together with subsurface ice.

6.4. Eolian Processes

The remarkable uniformity of soil compositions at all the landing sites, some separated by thousands of kilometers, suggests an efficient homogenization process, although soils nearly everywhere on Mars were probably made from similar (basaltic) rocks. Eolian transport of rock particles along the surface by the wind has apparently mixed these materials very efficiently, so that the soil everywhere on Mars represents a rather homogeneous stratigraphic layer.

In contrast to sand, tiny dust particles can be suspended in the atmosphere and circulated globally, which has also created a homogeneous material that is distributed globally. A dust cycle can be inferred from the omnipresent dusty atmosphere being supplied by dust devils and other processes that occasionally lead to globe-encircling storms. Dust deposition has been observed on most of the landed spacecraft at a rate that is so high that it must be removed at a similar rate (or the surfaces would be quickly buried by thick accumulations of dust). Dust currently may be or previously has been deposited at a higher rate overall in broad areas of the planet that have very low thermal inertia and very high albedo.

Sand-sized particles created by impact and other processes have been harnessed by the wind to form sand dunes and other eolian bedforms. The consistent basaltic composition of the soil and dust all over Mars further argues that Mars is dominated by basaltic rocks and that the}

soil and dust form by physical weathering and minor oxidation without large quantities of water. This further argues that these weathering products have formed and been mobilized by the wind in the current dry and desiccating environment.

6.5. Geologic Evolution of the Landing Sites and Climate

Study of the geology, geomorphology, and geochemistry of the seven landing sites in context with their regional geologic setting allows some constraints to be placed on the environmental and climatic conditions on Mars through time. The Viking 1 landing site shows sedimentary drift and soil deposits over angular, dark, presumably volcanic rocks with local outcrops (Figure 19.7). The location of this site on the ridged plains terrain downstream from the mouth of Maja and Kasei Valles suggests that the site is on layered basalts (the preferred interpretation of the ridged plains) with rocks, soils, and drifts derived from impact ejecta, flood, and eolian processes. The rocks at the Viking 2 landing site (Figure 19.8), in the midnorthern plains, are angular and pitted consistent with being volcanic rocks as part of the distal ejecta from Mie crater. High-resolution orbiter images show that the surface has a small-scale hummocky character and lander images show small polygonal sediment-filled troughs, both suggesting that the surface has been partially shaped by the presence of ground ice. The density of craters observed from orbit at both sites places them as Hesperian in age and constraints on the geomorphologic development of the sites suggest very little erosion or change of the surfaces since they formed.

Many characteristics of the MPF landing site (Figures 19.9 and 19.24) are consistent with its being a plain composed of materials deposited by catastrophic floods as suggested by its location near the mouth of the Ares Vallis catastrophic outflow channel. Some of the rocks potentially identified (conglomerate, pillow basalt) are suggestive of a wetter past. However, given that the surface still appears similar to that expected for a fresh depositional fan, any erosional and/or depositional process appears to have been minimal since it formed around 3 billion years ago.

The cratered surface of Gusev that Spirit has traversed (exclusive of the Columbia Hills) is generally low-relief, moderately rocky plains dominated by hollows, which appear to be small craters filled with soil (Figure 19.10). The plain formed by basaltic volcanism with impacts producing an unconsolidated regolith greater than 10 m thick (Figure 19.11). The observed gradation and deflation of ejected fines and deposition in craters to formollows thus provides a measure of the rate of erosion or redistribution of mobile sediment since the plains formed about 3.5 billion
years ago. These rates of erosion are so slow that they provide a broad indicator of a climate that has been cold and dry. Taken together, the slow rates of change inferred from the Viking-, Pathfinder-, and Gusev cratered plains landing sites argue for a dry and desiccating climate similar to today’s for the past ~3.6 billion years.

Rocks in the Columbia Hills (Figure 19.19) sampled by the Spirit rover reveal an earlier period in which liquid water was present. The Columbia Hills appear to be older materials that were either uplifted or eroded before deposition of the basalts responsible for the cratered plains. The basalts of the cratered plains are Hesperian in age and so the Columbia Hills rocks are likely older (Noachian in age). These rocks record impact and explosive volcanic processes, but many have been heavily altered or deeply weathered by water. In contrast, soils in the Columbia Hills are similar to basaltic soils elsewhere, suggesting that these formed and were deposited later in the cold and dry Martian climate.

The geology and geomorphology of the Meridiani Planum landing site explored by the Opportunity rover shows clear evidence for an earlier warmer and wetter environment followed by a drier period dominated by eolian activity. The layered rocks examined by Opportunity are older than 3.6 billion years based on the density of highly eroded large craters observed in orbital images (Figure 19.5). These rocks are dirty evaporites composed of materials that have precipitated from salty water and been mobilized and moved by the wind (Figures 19.20 and 19.29) before being deposited and altered by groundwater. On Earth, this sequence of events and resulting rocks is common in hot and dry saltwater playa or sabkha environments such as the Persian Gulf, the Gulf of California, and some inland enclosed basins. By analogy, the environment on Mars was warmer and wetter when these rocks were deposited more than 3.6 billion years ago. Because the evaporites are part of a sedimentary sequence that outcrops throughout the broad Meridiani region, these climatic conditions were operative over an area that was at least 1000 km wide, arguing that the environment was both warmer and wetter and the atmosphere was thicker. Later in Mars’ history, the environment changed and Meridiani Planum was dominated by eolian activity that eroded and filled in impact craters and concentrated the hematite spherules as a lag on the top of the layer of basaltic sand. The presence of olivine in the basaltic sand suggests that these materials were not weathered by liquid water and that the saltation of the sand appears to have efficiently eroded the weak sulfates.

The Phoenix landing site clearly reveals that the polar environment is distinct from other regions. The presence of perchlorate in soils also reveals atmospheric processes that are not so common elsewhere. The surface has pervasive polygonal troughs and very few impact craters suggesting a near surface that is constantly being modified by thermal contraction processes in the ground ice. The polar ice caps are even more dynamic, with annual cycles of sublimation and condensation of ices, and probably represent a much more geologically active environment than the rest of present-day Mars.

### 6.6. Implications for a Habitable World

The Meridiani Planum evaporites and Columbia Hills rocks in Gusev crater indicate a warmer and wetter environment before about 3.6 billion years ago. This is consistent with a variety of coeval geomorphic indicators such as valley networks, degraded and filled ancient craters, highly eroded terrain, and layered sedimentary rocks that point to an early warm and wet climate. The presence of clays and other alteration products that formed in neutral aqueous conditions are consistent with this early wet period. The highly acidic conditions that the sulfates formed in are consistent with a drying environment with less water. A warmer and wetter environment would also imply a thicker atmosphere capable of supporting liquid water. In contrast, the surficial geology of the landing sites younger than about 3.6 billion years all indicate a dry and desiccating environment in which liquid water was not stable and eolian and impact processes dominated. This further indicates that a major climatic change occurred around 3.6 billion years ago.

A warmer and wetter environment before 3.6 billion years ago suggests that Mars was possibly habitable at a time when life started on the Earth. The highly acidic nature of water at some Mars landing sites may not have been conducive to the appearance of early organisms, but clays and carbonates, which formed in earlier neutral aqueous conditions, may indicate more habitable conditions. In any case, the earliest chemical evidence for life on Earth is about 3.6 billion years old and the most important ingredient for life on Earth is liquid water. If liquid water was stable on Mars when life began on Earth, could a second genesis on Mars have occurred? Is it possible that life actually started on Mars earlier when it was more clement than Earth, which was subject to early giant possibly sterilizing impacts, and was later transported to the Earth via meteorites ejected off the Martian surface? Will life form at any place where liquid water is stable or is it a rare occurrence? These are the compelling questions that can be addressed by missions in our ongoing exploration of Mars.

### BIBLIOGRAPHY


Golombek, M., Grant, J., Kipp, D., Vasavada, A., Kirk, R., et al. (2012). Selection of the Mars science laboratory landing site. *Space Science Reviews*, 170, 641–737. http://dx.doi.org/10.1007/s11214-012-9916-y, which describes the selection of the landing site. See also other papers in this volume (pp. 1–860) that describe the pre-landing mission, instruments and various aspects of the landing site.


Smith, P. H., et al. (2009). H2O at the phoenix landing site. *Science*, 325, 58–61. and the next three papers (pp. 61–70) in which the initial results from the phoenix landing were reported.


Chapter Outline

1. Overview of Planetary Characteristics 424
   1.1. Length of Day 424
2. Vertical Structure of the Atmosphere  425
   2.1. Troposphere 426
   2.2. Stratosphere 426
   2.3. Mesosphere 427
   2.4. Thermosphere 427
   2.5. Exosphere and Ionosphere 427
3. Atmospheric Circulation 428
   3.1. Processes Driving the Circulation 428
   3.2. Influence of Rotation 428
   3.3. Observed Global-Scale Circulation 429
   3.4. Insights from Other Atmospheres 432
4. Oceans 433
   4.1. Oceanic Structure 433
   4.2. Ocean Circulation 434
   4.3. Salinity 435
   4.4. Atmosphere—Ocean Interactions 436
   4.5. Oceans on Other Worlds 436
5. Climate 436
   5.1. Basic Processes: Greenhouse Effect 437
   5.2. Basic Processes: Feedbacks 437
   5.3. Recent Times 438
   5.4. Ice Ages 439
   5.5. Volatile Inventories of Terrestrial Planets 441
6. Life in the Atmosphere—Ocean System 441
   6.1. Interplanetary Spacecraft Evidence for Life 441
      6.1.1. Radio Emissions 442
      6.1.2. Surface Features 443
      6.1.3. Oxygen and Methane 443
7. Conclusions 444

Bibliography 444

Earth is the only planet that orbits the Sun in the distance range within which water occurs in all three of its phases at the surface (as solid ice caps, liquid oceans, and atmospheric water vapor), which results in several unusual characteristics. Earth is unique in the solar system in exhibiting a global ocean at the surface, which covers almost three-quarters of the planet’s area (such that the total amount of dry land is about equal to the surface area of Mars). The ocean exerts a strong control over the planet’s climate by transporting heat from equator to pole, interacting with the atmosphere chemically and mechanically, and, on geological timescales, influencing the exchange of volatiles between the planet’s atmosphere and interior. The Earth’s atmosphere follows the general pattern of a troposphere at the bottom, a stratosphere in the middle, and a thermosphere at the top. There is the usual east—west organization of winds, but with large north—south and temporal fluctuations. Many of the atmospheric weather patterns (jet streams, Hadley cells, vortices, thunderstorms) occur on other planets too, but their manifestation on the Earth is distinct and unique. The Earth’s climate has varied wildly over time, with atmospheric CO2 and surface temperature fluctuating in response to ocean chemistry, planetary orbital variations, feedbacks between the atmosphere and interior, and a 30% increase in solar luminosity over the past 4.6 billion years (Ga). Despite these variations, the Earth’s climate has remained temperate, with at least partially liquid oceans, over the entire recorded 3.8-Ga geological record of the planet. Life has had a major influence on the ocean—atmosphere system, and as a result it is possible to discern the presence of life from remote spacecraft data. Global biological activity is indicated by the presence of atmospheric gases such as oxygen and methane that are in extreme thermodynamic disequilibrium, and by the widespread presence of a red-absorbing pigment (chlorophyll) that does not match the spectral
signatures of any known rocks or minerals. The presence of intelligent life on Earth can be discerned from stable radio-wavelength signals emanating from the planet that do not match naturally occurring signals but do contain regular pulsed modulations that are the signature of information exchange.

1. OVERVIEW OF PLANETARY CHARACTERISTICS

Atmospheres are found on the Sun, eight planets, and seven of the 60-odd satellites, for a total count of 16—in addition to the atmospheres that exist around the ~1000 known gas giant planets orbiting other stars. Each has its own brand of weather and its own unique chemistry. They can be divided into two major classes: the terrestrial planet atmospheres, which have solid surfaces or oceans as their lower boundary condition, and the gas giant atmospheres, which are essentially bottomless. Venus and Titan form one terrestrial subgroup that is characterized by a slowly rotating planet, and interestingly, both exhibit a rapidly rotating atmosphere. Mars, Io, Triton, and Pluto form a second terrestrial subgroup that is characterized by a thin atmosphere, which in large measure is driven by vapor—pressure equilibrium with the atmosphere’s solid phase on the surface. Both Io and Triton have active volcanic plumes. Earth, along with Mars and the giant planets, is in the rapidly rotating regime where the Coriolis force plays a dominant role. And although regional lakes of methane—ethane mixtures exist near the poles of Titan, Earth is also the only planet with a global (planet-encircling) ocean at the surface (see Venus: Atmosphere; Io: The Volcanic Moon; Triton; and Pluto).

Earth has many planetary attributes that are important to the study of its atmosphere and oceans, and conversely, there are several ways in which its physically and chemically active fluid envelope directly affects the solid planet. Earth orbits the Sun at a distance of only 108 times the diameter of the Sun. The warmth from the Sun that the Earth receives at this distance, together with a 30 K increase in surface temperature resulting from the atmospheric greenhouse effect, leads to temperatures allowing H$_2$O to appear in all three of its phases. This property of the semimajor axis of Earth’s orbit is the most important physical characteristic of the planet that supports life.

Orbiting the Sun at just over 100 Sun diameters is not as close as it may sound; a good analogy is to view a basketball placed just past first base while standing at home plate on a baseball diamond. For sunlight, the Sun-to-Earth trip takes 499 s or 8.32 min. Earth’s semimajor axis, $a_3 = 1.4960 \times 10^{11}$ m = 1 AU (astronomical unit), and orbital period, $\tau_3 = 365.26$ days = 1 year, where the subscript 3 denotes the third planet out from the Sun, are used as convenient measures of distance and time. When the orbital period of a body encircling the Sun, $\tau$, is expressed in years, and its semimajor axis, $a$, is expressed in astronomical units, then Kepler’s third law is simply $\tau = a^{3/2}$, with a proportionality constant of unity (see Solar System Dynamics: Regular and Chaotic Motion).

1.1. Length of Day

The Earth’s rotation (see Solar System Dynamics: Rotation of the Planets) has an enormous effect on the motions of its fluid envelope that accounts for the circular patterns of large storms like hurricanes, the formation of western boundary currents like the Gulf Stream, the intensity of jet streams, the extent of the Hadley cell, and the nature of fluid instabilities. All these processes are discussed in Sections 2–5. Interestingly, the reverse is also true: The Earth’s atmosphere and oceans have a measurable effect on the planet’s rotation rate. For all applications but the most demanding, the time the Earth takes to turn once on its axis, the length of its day, is adequately represented by a constant value equal to 24 h or 1440 min or 86,400 s. The standard second is the Système International (SI) second, which is precisely 9,192,631,770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the $^{133}$Cs atom. When the length of day is measured with high precision, it is found that Earth’s rotation is not constant. The same is likely to hold for any dynamically active planet. Information can be obtained about the interior of a planet, and how its atmosphere couples with its surface, from precise length-of-day measurements. Earth is the only planet to date for which we have achieved such accuracy, although we also have high-precision measurements of the rotation rate of pulsars, the spinning neutron stars often seen at the center of supernova explosions.

The most stable pulsars lose only a few seconds every million years and are the best-known timekeepers, even better than atomic clocks. In contrast, the rotating Earth is not an accurate clock. Seen from the ground, the positions as a function of time of all objects in the sky are affected by Earth’s variable rotation. Because the Moon moves across the sky relatively rapidly and its position can be determined with precision, the fact that Earth’s rotation is variable was first realized when a series of theories that should have predicted the motion of the Moon failed to achieve their expected accuracy. In the 1920 and the 1930s, it was established that errors in the position of the Moon were similar to errors in the positions of the inner planets, and by 1939, clocks were accurate enough to reveal that Earth’s rotation rate has both irregular and seasonal variations.

The quantity of interest is the planet’s three-dimensional angular velocity vector as a function of time, $\Omega(t)$. Since the 1970s, time series of all three components of $\Omega(t)$ have been generated by using very long baseline
interferometry for purposes ranging from accurately determining the positions of quasars and laser ranging to accurately determining the positions of man-made satellites and the Moon, the latter with corner reflectors placed on the Moon by the Apollo astronauts (see Planetary Exploration Missions; The Moon).

The theory of Earth’s variable rotation combines ideas from geophysics, meteorology, oceanography, and astronomy. The physical causes fall into two categories: those that change the planet’s moment of inertia (like a spinning skater pulling in her arms) and those that torque the planet by applying stresses (like dragging a finger on a spinning globe). Earth’s moment of inertia is changed periodically by tides raised by the Moon and the Sun, which distort the solid planet’s shape. Nonperiodic changes in the solid planet’s shape occur because of fluctuating loads from the fluid components of the planet, namely, the atmosphere, the oceans, and, deep inside the planet, the liquid iron-rich core. In addition, shifts of mass from earthquakes and melting ice cause nonperiodic changes. Over long timescales, plate tectonics and mantle convection significantly alter the moment of inertia and hence the length of day.

An important and persistent torque that acts on the Earth is the gravitational pull of the Moon and the Sun on the solid planet’s tidal bulge, which, because of friction, does not line up exactly with the combined instantaneous tidal stresses. This torque results in a steady lengthening of the day at the rate of about 0.0014 s/century and a steady outward drift of the Moon at the rate of 3.7 ± 0.2 cm/year, as confirmed by lunar laser ranging. On the top of this steady torque, it has been suggested that observed 0.005-s variations that have timescales of decades are caused by stronger, irregular torques from motions in Earth’s liquid core. Calculations suggest that viscous coupling between the liquid core and the solid mantle is weak, but that electromagnetic and topographic coupling can explain the observations. Mountains on the core—mantle boundary with heights around 0.5 km are sufficient to produce the coupling and are consistent with seismic tomography studies, but not much is known about the detailed topography of the core—mantle boundary. Detailed model calculations take into account the time variation of Earth’s external magnetic field, which is extrapolated downward to the core—mantle boundary. New improvements to the determination of the magnetic field at the surface are enhancing the accuracy of the downward extrapolations (see Earth as a Planet: Surface and Interior).

Earth’s atmosphere causes the strongest torques of all. The global atmosphere rotates faster than the solid planet by about 10 m/s on average. Changes in the global circulation cause changes in the pressure forces that act on mountain ranges and changes in the frictional forces between the wind and the surface. Fluctuations on the order of 0.001 s in the length of day, and movements of the pole by several meters, are caused by these meteorological effects, which occur over seasonal and interannual timescales. General circulation models (GCMs) of the atmosphere routinely calculate the global atmospheric angular momentum, which allows the meteorological and non-meteorological components of the length of day to be separated. All the variations in the length of day over weekly and daily timescales can be attributed to exchanges of angular momentum between Earth’s atmosphere and the solid planet, and this is likely to hold for timescales of several months as well. Episodic reconfigurations of the coupled atmosphere—ocean system, such as the El Niño-Southern Oscillation, cause detectable variations in the length of day, as do changes in the stratospheric jet streams.

2. VERTICAL STRUCTURE OF THE ATMOSPHERE

The Earth may differ in many ways from the other planets, but not in the basic structure of its atmosphere (Figure 20.1). Planetary exploration has revealed that essentially every atmosphere starts at the bottom with a troposphere, where temperature decreases with height at a nearly constant rate up to a level called the tropopause, and then has a stratosphere, where temperature usually increases with height or, in the case of Venus and Mars, decreases much less quickly than in the troposphere. It is interesting to note that atmospheres are warm both at their bottoms and their tops, but do not get arbitrarily cold in their interiors. For example, on Jupiter and Saturn there is significant methane gas throughout their atmospheres, but nowhere does it get cold enough for methane clouds to form, whereas in the much colder atmospheres of Uranus and Neptune, methane clouds do form. Details vary in the middle atmosphere regions from one planet to another, where photochemistry is important, but each atmosphere is topped off by a high-temperature, low-density thermosphere that is sensitive to solar activity and an exobase, the official top of an atmosphere, where molecules float off into space when they achieve escape velocity (see Venus: Atmosphere; Mars Atmosphere: History and Surface Interactions; Atmospheres of the Giant Planets).

Interestingly, the top of the troposphere occurs at about the same pressure, about 0.1—0.3 bar, on most planets (Figure 20.1). This similarity is not coincidental but instead results from the pressure dependence of the atmospheric opacity on solar and especially infrared radiation. In the high-pressure regime of tropospheres, the gas is relatively opaque at infrared wavelengths, which inhibits heat loss by radiation from the deep levels and hence promotes a profile where temperature decreases strongly with altitude. In the low-pressure regime of stratospheres, the gas becomes relatively transparent at infrared wavelengths, which allows
the temperature to become more constant—or in some cases even increase—with altitude. This transition from opaque to transparent tends to occur at pressures of \(0.1\) to \(0.3\) bar for the compositions of most planetary atmospheres in our solar system.

In the first 0.1 km of a terrestrial atmosphere, the effects of daily surface heating and cooling, surface friction, and topography produce a turbulent region called the planetary boundary layer. Right at the surface, molecular viscosity forces the “no slip” boundary condition and the wind reduces to zero, such that even a weak breeze results in a strong vertical wind shear that can become turbulent near the surface. However, only a few millimeters above the surface, molecular viscosity ceases to play a direct role in the dynamics, except as a sink for the smallest eddies.

Up to altitudes of about 80 km, Earth’s atmosphere is composed of 78% \(\text{N}_2\), 21% \(\text{O}_2\), 0.9% \(\text{Ar}\), and 0.002% \(\text{Ne}\) by volume, with trace amounts of \(\text{CO}_2\), \(\text{CH}_4\), and numerous other compounds. Water exists in abundances up to \(\sim 1\)% at the surface in the tropics, less at the poles, and dropping to a few parts per million in the stratosphere. Diffusion, chemistry, and other effects substantially alter the composition at altitudes above \(\sim 90\) km.

2.1. Troposphere

The troposphere is the lowest layer of the atmosphere, characterized by a temperature that decreases with altitude (Figure 20.1). The top of the troposphere is called the tropopause, which occurs at an altitude of 18 km at the equator but only 8 km at the poles (the cruising altitude of commercial airliners is typically 10 km). Gravity, combined with the compressibility of air, causes the density of an atmosphere to fall off exponentially with height, such that Earth’s troposphere contains 80% of the mass and most of the water vapor in the atmosphere, and consequently most of the clouds and stormy weather. Vertical mixing is an important process in the troposphere. Temperature falls off with height at a predictable rate because the air near the surface is heated and becomes light and the air higher up cools to space and becomes heavy, leading to an unstable configuration and convection. The process of convection relaxes the temperature profile toward the neutrally stable configuration, called the adiabatic temperature lapse rate, for which the decrease of temperature with decreasing pressure (and hence increasing height) matches the drop-off of temperature that would occur inside a balloon that conserves its heat as it moves, that is, moves adiabatically. In reality, latent heating due to water vapor—and horizontal heat transports—causes the temperature profile to decrease slightly less with height than such an adiabat. As a result, the troposphere is slightly stable to convection. Nevertheless, the adiabat provides a reasonable reference for the troposphere.

In the troposphere, water vapor, which accounts for up to \(\sim 1\)% of air, varies spatially and decreases rapidly with altitude. The water vapor mixing ratio in the stratosphere and above is almost four orders of magnitude smaller than that in the tropical lower troposphere.

2.2. Stratosphere

The nearly adiabatic falloff of temperature with height in the Earth’s troposphere gives way above the tropopause to an increase of temperature with height. This results in a rarified, stable layer called the stratosphere. Observations of persistent, thin layers of aerosol and of long residence times for radioactive trace elements from nuclear explosions are direct evidence of the lack of mixing in the stratosphere. The temperature continues to rise with altitude in the Earth’s stratosphere until one reaches the stratosphere at about 50 km. The source of heating in the Earth’s stratosphere is absorption of solar ultraviolet (UV) light by ozone. Ozone itself results from photochemistry, and exhibits abundances that peak at about 25 km. The Sun’s UV
radiation causes stratospheres to form in other atmospheres, but instead of the absorber being ozone, which is plentiful on the Earth because of the high concentrations of \( \text{O}_2 \) maintained by the biosphere, other gases absorb the UV radiation. On the giant planets, methane, hazes, and aerosols do the job.

The chemistry of Earth’s stratosphere is complicated. Ozone is produced mostly over the equator, but its largest concentrations are found over the poles, meaning that dynamics is as important as chemistry to the ozone budget. Some of the most important chemical reactions in Earth’s stratosphere are those that involve only oxygen. Photodissociation by solar UV radiation involves the reactions \( \text{O}_2 + \text{hv} \rightarrow \text{O} + \text{O} \) and \( \text{O}_3 + \text{hv} \rightarrow \text{O} + \text{O}_2 \), where \( \text{hv} \) indicates the UV radiation. Three-body collisions, where a third molecule, \( M \), is required to satisfy conservation of momentum and energy, include \( \text{O} + \text{O} + \text{M} \rightarrow \text{O}_2 + \text{M} \) and \( \text{O} + \text{O}_3 + \text{M} \rightarrow \text{O}_3 + \text{M} \), but the former reaction proceeds slowly and may be neglected in the stratosphere. Reactions that either destroy or create “odd” oxygen, \( \text{O} \) or \( \text{O}_3 \), proceed at much slower rates than reactions that convert between odd oxygen. The equilibrium between \( \text{O} \) and \( \text{O}_3 \) is controlled by fast reactions that have rates and concentrations that are altitude dependent. Other reactions that are important to the creation and destruction of ozone involve minor constituents such as \( \text{NO}, \text{NO}_2, \text{H}, \text{OH}, \text{HO}_2, \) and \( \text{Cl} \).

An important destruction mechanism is the catalytic cycle
\[
\text{X} + \text{O}_3 \rightarrow \text{XO} + \text{O}_2 \quad \text{followed by} \quad \text{XO} + \text{O} \rightarrow \text{X} + \text{O}_2,
\]
which results in the net effect \( \text{O} + \text{O}_3 \rightarrow 2\text{O}_2 \). On the Earth, human activity has led to sharp increases in the catalysts \( \text{X} = \text{Cl} \) and \( \text{NO} \) and subsequent sharp decreases in stratospheric ozone, particularly over the polar regions. The Montreal Protocol is an international treaty signed in 1987 that is designed to stop and eventually reverse the damage to the stratospheric ozone layer; regular meetings of the parties, involving some 175 countries, continually update the protocol.

### 2.3. Mesosphere

Above Earth’s stratosphere, temperature again falls off with height, although at a slower rate than in the troposphere. This region is called the mesosphere. Earth’s stratosphere and mesosphere are often referred to collectively as the middle atmosphere. Temperatures fall off in the mesosphere because there is less heating by ozone and emission to space by carbon dioxide is an efficient cooling mechanism. The mesopause occurs at an altitude of about 80 km, marking the location of a temperature minimum of about 130 K.

### 2.4. Thermosphere

As is the case for ozone in Earth’s stratosphere, above the mesopause, atomic and molecular oxygen strongly absorb solar UV radiation and heat the atmosphere. This region is called the thermosphere, and temperatures rise with altitude to a peak that varies between about 500 and 2000 K depending on solar activity. Just as in the stratosphere, the thermosphere is stable to vertical mixing. At about 120 km, molecular diffusion becomes more important than turbulent mixing, and this altitude is called the homopause (or turbopause). Rocket trails clearly mark the homopause—they are turbulently mixed below this altitude but mixed primarily by molecular diffusion above it, causing the rocket trails to appear differently above and below the interface. Molecular diffusion is mass-dependent and each species falls off exponentially with its own scale height, leading to elemental fractionation that enriches the abundance of the lighter species at the top of the atmosphere.

For comparison with Earth, the structure of the thermospheres of the giant planets has been determined from Voyager spacecraft observations, and the principal absorbers of UV light are \( \text{H}_2, \text{CH}_4, \text{C}_2\text{H}_2, \) and \( \text{C}_2\text{H}_6 \). The thermospheric temperatures of Jupiter, Saturn, and Uranus are about 1000, 420, and 800 K, respectively. The high temperature and low gravity on Uranus allow its upper atmosphere to extend out appreciably to its rings (see Atmospheres of the Giant Planets).

### 2.5. Exosphere and Ionosphere

At an altitude of about 500 km on the Earth, the mean free path between molecules grows to be comparable to the density scale height (the distance over which density falls off by a factor of \( e \approx 2.7128 \)). This defines the exobase and the start of the exosphere. At these high altitudes, sunlight can remove electrons from atmospheric constituents and form a supply of ions. These ions interact with a planet’s magnetic field and with the solar wind to form an ionosphere. On Earth, most of the ions come from molecular oxygen and nitrogen, whereas on Mars and Venus most of the ions come from carbon dioxide. Because of the chemistry, however, ionized oxygen atoms and molecules are the most abundant ions for all three atmospheres. Mercury and the Moon have exospheres right down to the planetary surface, with ions supplied from the surface crust and the solar wind.

Mechanisms of atmospheric escape fall into two categories, thermal and nonthermal. Both processes provide the kinetic energy necessary for molecules to attain escape velocity. When escape velocity is achieved at or above the exobase, such that further collisions are unlikely, molecules escape the planet. In the thermal escape process, some fraction of the high-velocity wing of the Maxwellian distribution of velocities for a given temperature always has escape velocity; the number increases with increasing temperature. An important nonthermal escape process is dissociation, both chemical and photochemical. The energy
for chemical dissociation is the excess energy of reaction, and for photochemical dissociation, it is the excess energy of the bombarding photon or electron, which is converted into kinetic energy in the dissociated atoms. A common effect of electrical discharges of a kilovolt or more is “sputtering”, where several atoms can be ejected from the spark region at high velocities. If an ion is formed very high in the atmosphere, it can be swept out of a planet’s atmosphere by the solar wind. Similarly, at Io, ions are swept away by Jupiter’s magnetic field. Other nonthermal escape mechanisms involve charged particles. Charged particles get trapped by magnetic fields and therefore do not readily escape. However, a fast proton can collide with a slow hydrogen atom and take the electron from the hydrogen atom. This charge exchange process changes the fast proton into a fast, hydrogen atom that is electrically neutral and hence can escape.

Nonthermal processes account for most of the present-day escape flux from Earth, and the same is likely to be true for Venus. They are also invoked to explain the $62 \pm 16\%$ enrichment of the $^{15}$N/$^{14}$N ratio in the Martian atmosphere. If the current total escape flux from thermal and nonthermal processes is applied over the age of the solar system, the loss of hydrogen from the Earth is equivalent to only a few meters of liquid water, which means that Earth’s sea level has not been affected much by this process. However, the flux could have been much higher in the past, since it is sensitive to the structure of the atmosphere (see Mars Atmosphere: History and Surface Interaction).

3. ATMOSPHERIC CIRCULATION

3.1. Processes Driving the Circulation

The atmospheric circulation on Earth, as on any planet, involves a wealth of phenomena ranging from global weather patterns to turbulent eddies only centimeters across and varies over periods of seconds to millions of years. All this activity is driven by absorbed sunlight and loss of infrared (heat) energy to space. Of the sunlight absorbed by the Earth, most (~70%) penetrates through the atmosphere and is absorbed at the surface; in contrast, the radiative cooling to space occurs not primarily from the surface but from the upper troposphere at an altitude of 5–10 km. This mismatch in the altitudes of heating and cooling means that, in the absence of air motions, the surface temperature would be much hotter than temperatures in the upper troposphere. However, such a trend produces an unstable density stratification, forcing the troposphere to overturn. The hot air rises, the cold air sinks, and thermal energy is thus transferred from the surface to the upper troposphere. This energy transfer by air motions leads to surface temperatures cooler than they would be in radiative equilibrium (while still being significantly hotter than the upper troposphere). This vertical mixing process is fundamentally responsible for near-surface convection, turbulence, cumulus clouds, thunderstorms, hurricanes, dust devils, and a range of other small-scale weather phenomena.

At global scales, much of Earth’s weather results not simply from vertical mixing but from the atmosphere’s response to horizontal temperature differences. Earth absorbs most of the sunlight at low latitudes, yet it loses heat to space everywhere over the surface. Hot equatorial air and cold polar air results. This configuration is gravitationally unstable—the hot equatorial air has low density and the cold polar air has high density. Just as the cold air from an open refrigerator slides across your feet, the cold polar air slides under the hot equatorial air, lifting the hotter air upward and poleward while pushing the colder air downward and equatorward. This overturning process transfers energy between the equator and the poles and leads to a much milder equator-to-pole temperature difference (about 30 K at the surface) than would exist in the absence of such motions. On average, the equatorial regions gain more energy from sunlight than they lose as radiated heat, while the reverse holds for the poles; the difference is transported between equator and pole by the air and ocean. The resulting atmospheric overturning causes many of Earth’s global-scale weather patterns, such as the 1000-km-long fronts that cause much midlatitude weather and the organization of thunderstorms into clusters and bands. Horizontal temperature and density contrasts can drive weather at regional scales too; examples include air—sea breezes and monsoons.

3.2. Influence of Rotation

The horizontal pressure differences associated with horizontal temperature differences cause a force (the “pressure gradient force”) that drives most air motion at large scales. However, how an atmosphere responds to this force depends strongly on whether the planet is rotating. On a nonrotating planet, the air tends to directly flow from high to low pressure, following the “nature abhors a vacuum” dictum. If the primary temperature difference occurs between equator and pole, this would lead to a simple overturning circulation between the equator and pole. On the other hand, planetary rotation (when described in a non-inertial reference frame rotating with the solid planet) introduces new forces into the equations of motion: the centrifugal force and the Coriolis force. The centrifugal force naturally combines with the gravitational force and the resultant force is usually referred to as simply the gravity. For rapidly rotating planets, the Coriolis force is the dominant term that balances the horizontal pressure gradient force in large-scale circulations (a balance called geostrophy). Because the Coriolis force acts perpendicular
to the air motion, this leads to a fascinating effect—the horizontal airflow is perpendicular to the horizontal pressure gradient. A north—south pressure gradient (resulting from a hot equator and a cold pole, for example) leads primarily not to north—south air motions but to east—west air motions! This is one reason why east—west winds dominate the circulation on most planets, including the Earth. For an Earth-sized planet with Earth-like wind speeds, rotation dominates the large-scale dynamics as long as the planet rotates at least once every 10 days.

Two other important effects of rapid rotation are the suppression of motions in the direction parallel to the rotation axis, called the Taylor—Proudman effect, and the coupling of horizontal temperature gradients with vertical wind shear, a three-dimensional relationship described by the thermal wind equation.

3.3. Observed Global-Scale Circulation

As described earlier, the atmospheric circulation organizes primarily into a pattern of east—west winds, and perhaps the most notable feature is the eastward-blowing jet streams in the midlatitudes of each hemisphere (Figure 20.2). In a longitudinal and seasonal average, the winter hemisphere wind maximum reaches 40 m/s at 30° latitude, and the summer hemisphere wind maximum reaches 20—30 m/s at 40°—50° latitude. In between these eastward wind maxima, from latitude 20 °N to 20 °S, the tropospheric winds blow weakly westward. The jet streams are broadly distributed in height, with peak speeds at about 12 km altitude. Although the longitudinally and seasonally averaged winds exhibit only a single tropospheric eastward wind maximum in each hemisphere, instantaneous three-dimensional snapshots of the atmosphere illustrate that there often exist two distinct jet streams, the subtropical jet at ~30° latitude and the so-called eddy-driven jet at ~50° latitude. These jets are relatively narrow—a few 100 km in latitudinal extent—and can reach speeds up to 100 m/s. However, the intense jet cores are usually less than a few thousand kilometers in longitudinal extent (often residing over continental areas such as eastern Asia and eastern North America), and the jets typically exhibit wide, time-variable wavelike fluctuations in position. When averaged over longitude and time, these variations in the individual jet streams smear into the single eastward maximum evident in each hemisphere in Figure 20.2.

Although the east—west winds dominate the time-averaged circulation, vertical and latitudinal motions are nevertheless required to transport energy from the equator to the poles. Broadly speaking, this transport occurs in two distinct modes. In the tropics exists a direct thermal overturning circulation called the Hadley cell, where, on average, air rises near the equator, moves poleward, and descends. This is an extremely efficient means of transporting heat and contributes to the horizontally homogenized temperatures that exist in the tropics. However, planetary rotation prevents the Hadley cell from extending all the way to the poles (to conserve angular momentum about the rotation axis, equatorial air would accelerate eastward to extreme speeds as it approached the pole, a phenomenon that is dynamically inhibited). On Earth, the Hadley cell extends to latitudes of ~30°. The subtropical jet lies at the poleward edge of the Hadley cell at ~30°. Poleward of ~30°, the surface temperatures decrease rapidly toward the pole. Although planetary rotation inhibits the Hadley cell in this region, north—south motions still occur via a complex three-dimensional process called baroclinic instability. Meanders on the jet stream grow, pushing cold high-latitude air under warm low-latitude air in confined regions ~1000—5000 km across. These instabilities grow, mature, and decay over ~5-day periods; new ones form as old ones disappear. These structures evolve to form regions with a sharp thermal gradient called fronts, as well as 1000—5000-km-long arc-shaped clouds and precipitation that dominate much of the winter weather in the United States, Europe, and other midlatitude regions.

Water vapor in Earth’s troposphere greatly accentuates convective activity because latent heat is liberated when moist air is raised above its lifting condensation level, and this further increases the buoyancy of the rising air, leading to moist convection. Towering thunderstorms get their energy from this process, and hurricanes are the most dramatic and best-organized examples of moist convection. Hurricanes occur only on the Earth because only the Earth provides the necessary combination of high humidity and surface friction. Surface friction is required to cause air to spiral into the center of the hurricane, where it is then forced upward past its lifting condensation level.

The Hadley cell exerts a strong control over weather in the tropics. The upward transport in the ascending branch of the Hadley circulation occurs almost entirely in localized thunderstorms and cumulus clouds whose convective towers cover only a small fraction (perhaps ~1%) of the total horizontal area of the tropics. Because this ascending branch resides near the equator, equatorial regions receive abundant rainfall, allowing the development of tropical rainforests in Southeast Asia/Indonesia, Brazil, and central Africa.

On the other hand, this condensation and rainout of water dehydrates the air, so the descending branch of the Hadley cell, which occurs in the subtropics at ~20°—30° latitude, is relatively dry. Because of the descending motion and dry conditions, little precipitation falls in these regions, which explains the abundance of arid biomes at 20°—30° latitude, including the deserts of the African Sahara, southern Africa, Australia, central Asia, and the southwestern United States. However, the simple Hadley cell is to some degree a theoretical idealization, and many regional three-dimensional
time-variable phenomena—including monsoons, equatorial waves, El Niño, and longitudinal overturning circulations associated with continent-ocean and sea-surface-temperature contrasts—affect the locations of tropical thunderstorm formation and hence the climatic rainfall patterns.

Satellite images (Figure 20.3) dramatically illustrate the signature of the Hadley cell and midlatitude baroclinic instabilities as manifested in clouds. In Figure 20.3, the east-west band of clouds stretching across the disk of the Earth just north of the equator corresponds to the rising branch of the Hadley cell (this cloud band is often called the intertropical convergence zone). These clouds are primarily the tops of thunderstorm anvils. In the midlatitude regions of both hemispheres (30°–70° latitude), several arc-shaped clouds up to 3000–5000 km long can be seen. These are associated with baroclinic instabilities. These clouds, which can often dominate midlatitude winter precipitation, form when large regions of warm air are forced upward over colder air masses during growing baroclinic instabilities. In many cases, the forced ascent associated with these instabilities produces predominantly sheetlike stratus clouds and steady rainfall lasting for several days, although sometimes the forced ascent can trigger local convection events (e.g. thunderstorms).
What causes the jet stream? This is a subtle question. At the crudest level, poleward-moving equatorial air deflects eastward due to the **Coriolis acceleration**, so the formation of eastward winds in the midlatitudes is a natural response to poleward-moving air in the upper troposphere. Because the low latitudes are warm and the high latitudes are cold, the horizontal pressure gradients at the top of the troposphere point on average from the equator to pole, and in steady state, are balanced by a Coriolis force (associated with atmospheric winds) that points toward the equator. Such an equatorward-pointing Coriolis force can only occur if the upper tropospheric winds in midlatitudes flow to the east. However, these processes alone would tend to produce a relatively broad zone of eastward flow rather than a narrow jet. Nonlinear turbulent motions, in part associated with baroclinic instabilities, pump momentum upgradient into this eastward-flowing zone and help to produce the narrow jet streams.

Identifying the particular mechanisms that cause the jet streams is aided by examining the force balance in the longitudinal direction. For example, in the Hadley cell, air moves toward the poles in the upper troposphere, and as it does so, the Coriolis force acts on the air to accelerate it in the eastward direction—causing the **subtropical jet**. On the other hand, forces due to waves and turbulence cause a westward acceleration at this latitude. Thus, in the subtropical jet, the Coriolis forces accelerate the jet, and turbulent forces act as a drag that tries to slow it down. The balance between these two opposing forces leads to a jet stream whose speeds remain relatively steady, on average, over time.

On the other hand, at the latitudes of the baroclinic instabilities, the situation is the exact opposite. Baroclinic instabilities lead to complex three-dimensional wave structures that transport momentum from their surroundings into the latitudes of the baroclinic instabilities. As this wave-transported momentum builds up at the instability latitude, it leads to an eastward acceleration—causing a jet stream called the **eddy-driven jet**. This is the latitude of the **Ferrel cells**, where air in the upper troposphere flows equatorward, causing a westward Coriolis force. Thus, in the eddy-driven jet, waves and turbulence act to accelerate the jet, and Coriolis forces act as a drag that slows it down. This is the exact opposite force balance as occurs in the subtropical jet. Again, the balance between these two opposing forces leads to a jet stream whose speeds remain relatively steady, on average, over time.

Although the Earth’s equator is hotter than the poles at the surface, it is noteworthy that, in the upper troposphere and lower stratosphere (\(\sim 18 \text{ km altitude}\)), the reverse is
true. This seems odd because sunlight heats the equator much more strongly than the poles. In reality, the cold equatorial upper troposphere results from a dynamical effect: large-scale ascent in the tropics causes air to expand and cool (a result of decreased pressure as the air rises), leading to the low temperatures despite the abundant sunlight. Descent at higher latitudes causes compression and heating, leading to warmer temperatures. Interestingly, this means that, in the lower stratosphere, the ascending air is actually denser than the descending air. Such a circulation, called a thermally indirect circulation, is driven by the absorption of atmospheric waves that are generated in the troposphere and propagate upward into the stratosphere. There is a strong planetary connection because all four giant planets—Jupiter, Saturn, Uranus, and Neptune—are also thought to have thermally indirect circulations in their stratospheres driven by analogous processes.

3.4. Insights from Other Atmospheres

Planetary exploration has revealed that atmospheric circulations come in many varieties. The goal of planetary meteorology is to understand what shapes and maintains these diverse circulations. The Voyager spacecraft provided the first close-up images of the atmospheres of Jupiter, Saturn, Uranus, and Neptune and detailed information on the three satellites that have atmospheres thick enough to sport weather—Io, Titan, and Triton. Galileo, Cassini, and New Horizons have visited Jupiter, and Cassini has obtained a wealth of information about Saturn and Titan. The atmospheres of Venus and Mars have been sampled by entry probes, landers, orbiting spacecraft, and telescopic studies. Basic questions like why Venus’ atmosphere rotates up to 60 times faster than does the planet, or why Jupiter and Saturn have superrotating equatorial jets, do not have completely satisfactory explanations. However, by comparing and contrasting each planet’s weather, a general picture has begun to emerge.

Theoretical studies and comparative planetology show that planetary rotation rate and size exert a major control over the type of global atmospheric circulations that occur. When the rotation rate is small, Hadley cells are unconfined and stretch from the equator to the pole. Venus, with a rotation period of 243 days, seems to reside in such a state. Titan rotates in 16 days and, according to circulation models, its Hadley cell extends to at least $\sim 60^\circ$ latitude, a transitional regime between Venus and Earth. On the other hand, fast rotation confines the Hadley cell to a narrow range of latitudes ($0^\circ$–$30^\circ$ on Earth) and forces baroclinic instabilities to take over much of the heat transport between low latitudes and the poles. Increasing the rotation rate still further—or making the planet larger—causes the mid-latitudes to break into a series of narrow latitudinal bands, each with their own east–west jet streams and baroclinic instabilities. The faster the rotation rate, the straighter and narrower are the bands and jets. This process helps explain the fact that Jupiter and Saturn, which are large and rapidly rotating, have $\sim 30$ and 20 jet streams, respectively (as compared to only a few jet streams for the Earth). Fast rotation also contributes to smaller structures because it inhibits free movement of air toward or away from pressure lows and highs, instead causing the organization of vortices around such structures. Thus, a planet identical to the Earth but with a faster or slower rotation rate would exhibit different circulations, equatorial and polar temperatures, rainfall patterns, and cloud patterns, and hence would exhibit a different distribution of deserts, rainforests, and other biomes.

The giant planets Jupiter and Saturn exhibit numerous oval-shaped windstorms that superficially resemble terrestrial hurricanes. However, hurricanes can generate abundant rainfall because friction allows near-surface air to spiral inward toward the low-pressure center, providing a source of moist air that then ascends inside thunderstorms; in turn, these thunderstorms release energy that maintains the hurricane’s strength against the frictional energy losses. In contrast, vortices like Jupiter’s Great Red Spot and the hundreds of smaller ovals seen on Jupiter, as well as the dozens seen on Saturn and the couple seen on Neptune, do not directly require moist convection to drive them and hence are not hurricanes. Instead, they are simpler systems that are closely related to three types of long-lasting, high-pressure “storms”, or coherent vortices, seen on the Earth: blocking highs in the atmosphere and Gulf Stream rings and Mediterranean salt lenses (“meddies”) in the ocean. Blocking highs are high-pressure centers that stubbornly settle over continents, particularly in the United States and Russia, thereby diverting rain from its usual path for months at a time. For example, the serious 1988 drought in the US Midwest was exacerbated by a blocking high. Gulf Stream rings are compact circulations in the Atlantic that break off from the meandering Gulf Stream, which is a “river” inside the Atlantic Ocean that runs northward along the eastern coast of the United States and separates from the coast at North Carolina, where it then jets into the Atlantic in an unsteady manner. Seen in three dimensions, the Gulf Stream has the appearance of a writhing snake. Similar western boundary currents occur in other ocean basins, for example, the Kuroshio Current off the coast of Japan and the Agulhas Current off the coast of South Africa. Jet streams in the atmosphere are a related phenomenon. When Gulf Stream rings form, they trap phytoplankton and zooplankton inside them, which are carried large distances. Over the course of a few months, the rings dissipate at sea, are reabsorbed into the Gulf Stream, or run into the coast, depending on which side of the Gulf Stream they formed. The ocean plays host to another class of long-lived vortices,
Mediterranean salt lenses, which are organized high-pressure circulations that float under the surface of the Atlantic. They form when the extrasalnty water that slips into the Atlantic from the shallow Mediterranean Sea breaks off into vortices. After a few years, these meddies eventually wear down as they slowly mix with the surrounding water. The mathematical description of these long-lasting vortices on the Earth is the same as that used to describe the ovals seen on Jupiter, Saturn, and Neptune (see Atmospheres of the Giant Planets).

Given that we know that atmospheric motions are fundamentally driven by sunlight, and we know that the problem is governed by Newton’s laws of motion, why then are atmospheric circulations difficult to understand? Several factors contribute to the complexity of observed weather patterns. In the first place, fluids move in an intrinsically nonlinear fashion that makes paper-and-pencil analysis formidable and often intractable. Laboratory experiments and numerical experiments performed on high-speed computers are often the only means for making progress on problems in geophysical fluid dynamics. Second, meteorology involves the intricacies of moist thermodynamics and precipitation, and we are only beginning to understand and accurately model the microphysics of these processes. And for the terrestrial planets, a third complexity arises from the complicated boundary conditions that the solid surface presents to the problem, especially when mountain ranges block the natural tendency for winds to organize into steady east–west jet streams. For oceanographers, even more restrictive boundary conditions apply, namely, the ocean basins, which strongly affect how currents behave. The giant planets are free of this boundary problem because they are completely fluid down to their small rocky cores. However, the scarcity of data for the giant planets, especially with respect to their vertical structure beneath the cloud tops, provides its own set of difficulties (see Interiors of the Giant Planets).

4. OCEANS

Earth is the only planet in the solar system with a global ocean at the surface. The oceans have an average thickness of 3.7 km and cover 71% of Earth’s surface area; the greatest thickness is 10.9 km, which occurs at the Marianas Trench. The total oceanic mass—$1.4 \times 10^{21}$ kg—exceeds the atmospheric mass of $5 \times 10^{18}$ kg by nearly a factor of 300, implying that the oceans dominate Earth’s surface inventory of volatiles (one way of visualizing this fact is to realize that, if Earth’s entire atmosphere condensed as ices on the surface, it would form a layer only $\sim 10$ m thick).

The Earth therefore sports a greater abundance of fluid volatiles at its surface than any other solid body in the solar system. Even Venus’ 90 bar CO$_2$ atmosphere contains only one-third the mass of Earth’s oceans. On the other hand, Earth’s oceans constitute only 0.02% of Earth’s total mass; the mean oceanic thickness of 3.7 km pales in comparison to Earth’s 6400 km radius, implying that the oceans span only 0.06% of Earth’s width. The Earth is thus a relatively dry planet, and the oceans truly are only skin deep.

It is possible that Earth’s solid mantle contains a mass of dissolved water (stored as individual water molecules inside and between the rock grains) equivalent to several oceans’ worth of water. Taken together, however, the total water in Earth probably constitutes less than 1% of Earth’s mass. In comparison, most icy satellites and comets in the outer solar system contain $\sim 40\text{–}60\%$ $\text{H}_2\text{O}$ by mass, mostly in solid form. This lack of water on Earth in comparison to outer solar system bodies reflects the relatively dry conditions in the inner solar system when the terrestrial planets formed; indeed, the plethora of water on Earth compared to Venus and Mars has raised the question of whether even the paltry amount of water on Earth must have been delivered from an outer solar system source such as impact of comets onto the forming Earth.

The modern oceans can be subdivided into the Pacific, Atlantic, Indian, and Arctic Oceans, but these four oceans are all connected, and this contiguous body of water is often simply referred to as the global ocean.

4.1. Oceanic Structure

The top meter of ocean water absorbs more than half of the sunlight entering the oceans; even in the sediment-free open ocean, only 20% of the sunlight reaches a depth of 10 m and only $\sim 1\%$ penetrates to a depth of 100 m (depending on the angle of the Sun from vertical). Photosynthetic single-celled organisms, which are extremely abundant near the surface, can thus only survive above depths of $\sim 100$ m; this layer is called the photic zone. The much thicker aphotic zone, which has too little light for photosynthetic production to exceed respiration, extends from $\sim 100$ m to the bottom of the ocean. Despite the impracticality of photosynthesis at these depths, the deep oceans, nevertheless, exhibit a wide variety of life fueled in part by dead organic matter that slowly sediments down from the photic zone (see Astrobiology).

From a dynamical point of view, the ocean can be subdivided into several layers. Turbulence caused by wind and waves homogenizes the top 20–200 m of the ocean (depending on weather conditions), leading to profiles of density, temperature, salinity, and composition that vary little across this layer, which is therefore called the mixed layer. Below the mixed layer lies the thermocline, where the temperature generally decreases with depth down to $\sim 0.5\text{–}1$ km. The salinity also often varies with depth between $\sim 100$ and 1000 m, a layer called the halocline. For example, regions of abundant precipitation but lesser evaporation, such as the North Pacific, have relatively fresh
surface waters, so the salinity increases with depth below the mixed layer in those regions. The variation of temperature and salinity between ~100 and 1000 m implies that density varies with depth across this layer too; this is referred to as the pycnocline. Below the thermocline, halocline, and pycnocline lies the deep ocean, where temperatures are usually relatively constant with depth at a chilly 0–4 °C.

The temperature at the ocean surface varies strongly with latitude, with only secondary variations in longitude. Surface temperatures reach 25 °C–30 °C near the equator, where abundant sunlight falls, but plummet to 0 °C near the poles. In contrast, the deep oceans (>1 km) are generally more homogeneous and have temperatures between 0 °C and 4 °C all over the world (when enjoying the bathtub-temperature water and coral reefs during a summer vacation to a tropical island, it is sobering to think that if one could only scuba dive deep enough, the temperature would approach freezing). This latitude-dependent upper ocean structure implies that the thermocline and pycnocline depths decrease with latitude: They are about ~1 km near the equator and reach zero near the poles.

Because warmer water is less dense than colder water, the existence of a thermocline over most of the ocean implies that the top ~1 km of the ocean is less dense than the underlying deep ocean. The implication is that, except for localized regions near the poles, the ocean is stable to vertical convective overturning.

4.2. Ocean Circulation

Ocean circulation differs in important ways from atmospheric circulation, despite the fact that the two are governed by the same dynamical laws. First, the confinement of oceans to discrete basins separated by continents prevents the oceanic circulation from assuming the common east–west flow patterns adopted by most atmospheres (topography can cause substantial north–south deflections in an atmospheric flow, which may help explain why Earth’s atmospheric circulation involves more latitudinal excursions than that of the topography-free giant planets; nevertheless, air’s ability to flow over topography means that atmospheres, unlike oceans, are still fundamentally unbounded in the east–west direction.) The only oceanic region unhindered in the east–west direction is the Southern Ocean surrounding Antarctica, and, as might be expected, a strong east–west current, which encircles Antarctica, has formed in this region.

Second, the atmosphere is heated from below, but the ocean is heated from above. Because air is relatively transparent to sunlight, sunlight penetrates through the atmosphere and is absorbed primarily at the surface, where it heats the near-surface air at the bottom of the atmosphere. In contrast, liquid water absorbs sunlight extremely well, so that 99% of the sunlight is absorbed in the top 3% of the ocean. This means, for example, that atmospheric convection—thunderstorms—predominates at low latitudes (where abundant sunlight falls) but is rare near the poles; in contrast, convection in the oceans is totally inhibited at low latitudes and instead can occur only near the poles.

Third, much of the large-scale ocean circulation is driven not by horizontal density contrasts, as in the atmosphere (although these do play a role in the ocean), but by the frictional force of wind blowing over the ocean surface. In fact, the first simple models of ocean circulation developed by Sverdrup, Stommel, and Munk in the 1940s and the 1950s, which were based solely on forcing caused by wind stress, did a reasonably good job of capturing the large-scale horizontal circulations in the ocean basins.

As in the atmosphere, the Earth’s rotation dominates the large-scale dynamics of the ocean. Horizontal Coriolis forces nearly balance pressure gradient forces, leading to geostrophy. As in the atmosphere, this means that ocean currents flow perpendicular to horizontal pressure gradients. Rotation also means that wind stress induces currents in a rather unintuitive fashion. Because of the existence of the Coriolis force, currents do not simply form in the direction of the wind stress; instead, the three-way balance between Coriolis, pressure gradient, and friction forces can induce currents that flow in directions distinct from the wind direction.

Averaged over time, the surface waters in most mid-latitude ocean basins exhibit a circulation consisting of a basin-filling gyre that rotates clockwise in the Northern Hemisphere and counterclockwise in the Southern Hemisphere. This circulation direction implies that the water in the western portion of the basin flows from the equator toward the pole, while the water in the eastern portion of the basin flows from the pole toward the equator. However, the flow is extremely asymmetric: The equatorward flow comprises a broad, slow motion that fills the eastern 90% of the ocean basin; in contrast, the poleward flow becomes concentrated into a narrow current (called a western boundary current) along the western edge of the ocean basin. The northward-flowing Gulf Stream off the US eastern seaboard and the Kuroshio Current off Japan are two examples; these currents reach speeds up to ~1 m/s in a narrow zone 50–100 km wide. This extraordinary asymmetry in the ocean circulation results from the increasing strength of the Coriolis force with latitude; theoretical models show that in a hypothetical ocean where Coriolis forces are independent of latitude, the gyre circulations do not exhibit western intensification. These gyres play an important role in Earth’s climate by transporting heat from the equator toward the poles. Their clockwise (counterclockwise) rotation in the northern (southern) hemisphere helps explain why the water temperatures tend to be colder along continental west coasts than continental east coasts.
In addition to the gyres, which transport water primarily horizontally, the ocean also experiences vertical overturning. Only near the poles does the water temperature become cold enough for the surface density to exceed the deeper density. Formation of sea ice helps this process, because sea ice contains relatively little salt, so when it forms, the remaining surface water is saltier (hence denser) than average. Thus, vertical convection between the surface and deep ocean occurs only in polar regions, in particular in the Labrador Sea and near parts of Antarctica. On average, very gradual ascending motion must occur elsewhere in the ocean for mass balance to be achieved. This overturning circulation, which transports water from the surface to the deep ocean and back over ~1000 year timescales, is called the thermohaline circulation.

The thermohaline circulation helps explain why deep ocean waters have near-freezing temperatures worldwide: all deep ocean water, even that in the equatorial oceans, originates at the poles and thus retains the signature of polar temperatures. Given the solar warming of low-latitude surface waters, the existence of a thermocline is thus naturally explained. However, the detailed dynamics that control the horizontal structure and depth of the thermocline are subtle and have led to major research efforts in physical oceanography over the past four decades.

Despite the importance of the basin-filling gyre and thermohaline circulations, much of the ocean’s kinetic energy resides in small eddy structures only 10–100 km across. The predominance of this kinetic energy at small scales results largely from the natural interaction of buoyancy forces and rotation. Fluid flows away from pressure highs toward pressure lows, but Coriolis forces short circuit this process by deflecting the motion so that fluid flows perpendicular to the horizontal pressure gradient. The stronger the influence of rotation relative to buoyancy, the better this process is short circuited, and hence the smaller are the resulting eddy structures. In the atmosphere, this natural length scale (called the deformation radius) is 1000–2000 km, but in the oceans, it is only 10–100 km. The rings and meddies described earlier provide striking examples of oceanic eddies in this size range.

4.3. Salinity

When one swims in the ocean, the leading impression is of saltiness. The ocean’s global mean salinity is 3.5% by mass but varies between 3.3% and 3.8% in the open oceans and can reach 4% in the Red Sea and Persian Gulf; values lower than 3.3% can occur on continental shelves near river deltas. The ocean’s salt would form a global layer 150 m thick if precipitated into solid form. Sea salt is composed of 55% chlorine, 30% sodium, 8% sulfate, 4% magnesium, and 1% calcium by mass. The ~15% variability in the salinity of open ocean waters occurs because evaporation and precipitation add or remove freshwater, which dilutes or concentrates the local salt abundance. However, this process cannot influence the relative proportions of elements in sea salt, which therefore remain almost constant everywhere in the oceans.

In contrast to seawater, most river and lake water is relatively fresh; for example, the salinity of Lake Michigan is ~200 times less than that of seawater. However, freshwater lakes always have both inlets and outlets. In contrast, lakes that lack outlets—the Great Salt Lake, the Dead Sea, and the Caspian Sea—are always salty. This provides a clue about processes determining saltiness.

Why is the ocean salty? When rain falls on continents, enters rivers, and flows into the oceans, many elements leach into the water from the continental rock. These elements have an extremely low abundance in the continental water, but because the ocean has no outlet (unlike a freshwater lake), these dissolved trace components can build up over time in the ocean. Ocean—seafloor chemical interactions (especially after volcanic eruptions) can also introduce dissolved ions into the oceans. However, the composition of typical river water differs drastically from that of sea salt—typical river salt contains ~9% chlorine, 7% sodium, 12% sulfate, 5% magnesium, and 17% calcium by mass. Although sodium and chlorine comprise ~85% of sea salt, they make up only ~16% of typical river salt. The ratio of chloride to calcium is 0.5 in river salt but 46 in sea salt. Furthermore, the abundance of sulfate and silica is much greater in river salt than in sea salt. These differences result largely from the fact that processes act to remove salt ions from ocean water, but the efficiency of these processes depends on the ion. For example, many forms of sea life construct shells of calcium carbonate or silica, so these biological processes remove calcium and silica from ocean water. Much magnesium and sulfate seems to be removed in ocean water—seafloor interactions. The relative inefficiency of such removal processes for sodium and chlorine apparently leads to the dominance of these ions in sea salt despite their lower proportion in river salt.

It is often suggested that ocean salinity has been stable over the past billion years. If so, this would imply that the ocean is near a quasisteady state where salt removal balances salt addition via rivers and seafloor—ocean interactions. Nevertheless, evidence from fluid inclusions in marine halites, among other sources, suggests significant changes of oceanic chemistry (including salinity) over time. Although the salinity in early Earth history is not well known, indirect evidence suggests that it may have exceeded the current ocean salinity by up to a factor of two. These temporal fluctuations in salinity result from imbalances in salinity input and removal processes to the ocean over time. Salt removal processes include biological sequestration in shell material, abiological seafloor—ocean water chemical interactions, and physical processes such as...
formation of evaporate deposits when shallow seas dry up, which has the net effect of returning the water to the world ocean while leaving salt behind on land.

4.4. Atmosphere—Ocean Interactions

Many weather and climate phenomena result from a coupled interaction between the atmosphere and ocean and would not occur if either component were removed. Two major examples are hurricanes and El Niño.

Hurricanes are strong vortices, 100–1000 km across, with warm cores and winds often up to \( \sim 70 \text{ m/s} \); the temperature difference between the vortex and the surrounding air produces the pressure differences that allow strong vortex winds to form. In turn, the strong winds lead to increased evaporation off the ocean surface, which provides an enhanced supply of water vapor to fuel the thunderstorms that maintain the warm core. This enhanced evaporation from the ocean must continue throughout the hurricane’s lifetime because the thermal effects of condensation in thunderstorms inside the hurricane provide the energy that maintains the vortex against frictional losses. Thus, both the ocean and atmosphere play crucial roles. When the ocean component is removed—say, when the hurricane moves over land—the hurricane rapidly decays.

El Niño corresponds to an enhancement of ocean temperatures in the eastern equatorial Pacific at the expense of those in the western equatorial Pacific; increased rainfall in western North and South America result, and drought conditions often overtake Southeast Asia. El Niño events occur every few years and have global effects. At the crudest level, “normal” (non-El Niño) conditions correspond to westward-blowing equatorial winds that cause a thickening of the thermocline (hence producing warmer sea surface temperatures) in the western equatorial Pacific; these warm temperatures promote evaporation, thunderstorms, and upwelling there, drawing near-surface air in from the east and thus helping to maintain the circulation. On the other hand, during El Niño, the westward-blowing trade winds break down, allowing the thicker thermocline to relax eastward toward South America, hence helping to move the warmer water eastward. Thunderstorm activity thus becomes enhanced in the eastern Pacific and reduced in the western Pacific compared to non-El Niño conditions, again helping to maintain the winds that allow those sea surface temperatures. Although El Niño differs from a hurricane in being a hemispheric-scale long-period fluctuation rather than a local vortex, El Niño shares with hurricanes the fact that it could not exist where either the atmosphere or the ocean component prevented from interacting with the other. To successfully capture these phenomena, climate models need accurate representations of the ocean and the atmosphere and their interaction, which continues to be a challenge.

4.5. Oceans on Other Worlds

The Galileo spacecraft provided evidence that subsurface liquid-water oceans exist inside the icy moons Callisto, Europa, and possibly Ganymede (see Ganymede and Callisto; Europa; Titan; Planetary Satellites). The recent detection of a jet of water molecules and ice grains from the south pole of Enceladus raises the question of whether that moon has a subsurface reservoir of liquid water. Theoretical models suggest that internal oceans could exist on a wide range of other bodies, including Titan, the smaller moons of Saturn and Uranus, Pluto, and possibly even some larger Kuiper Belt objects. These oceans of course differ from Earth’s ocean in that they are ice covered; another difference is that they must transport the geothermal heat flux of those bodies and hence are probably convective throughout. Barring exotic chemical or fluid dynamical effects, then, one expects that such oceans lack thermoclines. In many cases, these oceans may be substantially thicker than Earth’s oceans; estimates suggest that Europa’s ocean thickness lies between 50 and 150 km.

The abundant life that occurs near deep-sea vents (“black smokers”) in Earth’s oceans has led to suggestions that similar volcanic vents may help power life in Europa’s ocean. (In contrast to Europa, any ocean in Callisto and Ganymede would be underlain by high-pressure polymorphs of ice rather than silicate rock, so such silicate—water interactions would be weaker.) However, much of the biological richness of terrestrial deep-sea vents results from the fact that Earth’s oceans are relatively oxygenated; when this oxidant-rich water meets the reducing water discharged from black smokers, sharp chemical gradients result, and the resulting disequilibrium provides a rich energy source for life. Thus, despite the lack of sunlight at Earth’s ocean floor, the biological productivity of deep-sea vents results in large part from the fact that the oceans are communicating with an oxygen-rich atmosphere. If Europa’s ocean is more reducing than Earth’s ocean, then the energy source available from chemical disequilibrium may be smaller. Nevertheless, a range of possible disequilibrium reactions exist that could provide energy to drive a modest microbial biosphere on Europa (see Astrobiology).

5. CLIMATE

Earth’s climate results from a wealth of interacting physical, chemical, and biological effects, and an understanding of current and ancient climates has required a multidisciplinary research effort by atmospheric physicists, atmospheric chemists, oceanographers, glaciologists, astronomers, geologists, and biologists. The complexity of the climate system and the interdisciplinary nature of the problem have made progress difficult, and even today many aspects remain
poorly understood. “Climate” can be defined as the mean conditions of the atmosphere/ocean system—temperature, pressure, winds/currents, cloudiness, atmospheric humidity, oceanic salinity, and atmosphere/ocean chemistry in three dimensions—when time-averaged over intervals longer than those of typical weather patterns. It also refers to the distribution of sea ice, glaciers, continental lakes and streams, as well as coastlines, and the spatial distribution of ecosystems that result.

5.1. Basic Processes: Greenhouse Effect

Earth as a whole radiates with an effective temperature of 255 K, and therefore, its flux peaks in the thermal infrared part of the spectrum. This effective temperature is 30 K colder than the average temperature on the surface, and quite chilly by human standards.

What ensures a warm surface is the wavelength-dependent optical properties of the troposphere. In particular, infrared light does not pass through the troposphere as readily as visible light. The Sun radiates with an effective temperature of 5800 K and therefore, its peak flux is in the visible part of the spectrum (or stated more correctly in reverse, we have evolved such that the part of the spectrum that is visible to us is centered on the peak flux from the Sun). The atmosphere reflects about 31% of this sunlight directly back to space, and the rest is absorbed or transmitted to the ground. The sunlight that reaches the ground is absorbed and then reradiated at infrared wavelengths.

Water vapor (H₂O) and carbon dioxide (CO₂), the two primary greenhouse gases, absorb some of this upward infrared radiation and then emit it in both the upward and downward directions, leading to an increase in the surface temperature to achieve balance. This is the greenhouse effect. Contrary to popular claims, the elevation of surface temperature by the greenhouse effect is not a situation where “the heat cannot get out”. Instead, the heat must get out, and to do so in the presence of the blanketing effect of greenhouse gases requires an elevation of surface temperatures.

The greenhouse effect plays an enormous role in the climate system. A planet without a greenhouse effect, but otherwise identical to Earth, would have a global mean surface temperature 17°C below freezing. The oceans would be mostly or completely frozen, and it is doubtful whether life would exist on Earth. We owe thanks to the greenhouse effect for Earth’s temperate climate, liquid oceans, and abundant life.

Water vapor accounts for between one-third and two-thirds of the greenhouse effect on Earth (depending on how the accounting is performed), with the balance resulting from CO₂, methane, and other trace gases. Steady increases in carbon dioxide due to human activity seem to be causing the well-documented increase in global surface temperature over the past ~100 years. On Mars, the primary atmospheric constituent is CO₂, which together with atmospheric dust causes a modest 5 K greenhouse effect. Venus has a much denser CO₂ atmosphere, which, along with atmospheric sulfuric acid and sulfur dioxide, absorbs essentially all the infrared radiation emitted by the surface, causing an impressive 500 K rise in the surface temperature. Interestingly, if all the carbon held in Earth’s carbonate rocks were liberated into the atmosphere, Earth’s atmospheric CO₂ abundance and greenhouse effect would approach that on Venus (see Mars Atmosphere: History and Surface Interaction; Venus: Atmosphere).

5.2. Basic Processes: Feedbacks

The Earth’s climate evolves in response to volcanic eruptions, solar variability, oscillations in Earth’s orbit, and changes in internal conditions such as the concentration of greenhouse gases. The Earth’s response to these perturbations is highly nonlinear and is determined by feedbacks in the climate system. Positive feedbacks amplify a perturbation and, under some circumstances, can induce a runaway process where the climate shifts abruptly to a completely different state. In contrast, negative feedbacks reduce the effect of a perturbation and thereby help maintain the climate in its current state. Some of the more important feedbacks are as follows.

Thermal feedback: Increases in the upper tropospheric temperature lead to enhanced radiation to space, tending to cool the Earth. Decreases in the upper tropospheric temperature cause decreased radiation to space, causing warming. This is a negative feedback.

Ice-albedo feedback: Ice caps and glaciers reflect visible light easily, so the Earth’s brightness (albedo) increases with an increasing distribution of ice and snow. Thus, a more ice-rich Earth absorbs less sunlight, promoting colder conditions and growth of even more ice. Conversely, melting of glaciers causes Earth to absorb more sunlight, promoting warmer conditions and even less ice. This is a positive feedback.

Water vapor feedback: Warmer surface temperatures allow increased evaporation of water vapor from the ocean surface, increasing the atmosphere’s absolute humidity. Because water vapor is a greenhouse gas, it promotes an increase in the strength of the greenhouse effect and hence even warmer conditions. Cooler conditions inhibit evaporation, lessen the greenhouse effect, and cause additional cooling. This is a positive feedback.

Cloud feedback: Changes in climate can cause changes in the spatial distribution, heights, and properties of clouds. Greater cloud coverage means a brighter Earth
(higher albedo), leading to less sunlight absorption. Higher altitude clouds have colder tops that radiate heat to space less well, promoting a warmer Earth. For a given mass of condensed water in a cloud, clouds with smaller particles reflect light better, promoting a cooler Earth. Unfortunately, for a specified climate perturbation (e.g. increasing the CO₂ concentration), the extent to which the coverage, heights, and properties of clouds will change remains unclear. In the current Earth climate, clouds cause a net cooling effect (relative to an otherwise similar atmosphere with no clouds). Sophisticated GCMs suggest that the cloud feedback for the modern Earth climate is positive, although significant uncertainties remain.

The sum of these and other feedbacks determine how Earth’s climate evolved during past epochs and how Earth will respond to current human activities such as emissions of CO₂. Much of the uncertainty in current climate projections results from uncertainty in these feedbacks. A related concept is that of thresholds, where the climate undergoes an abrupt shift in response to a gradual change. For example, Europe enjoys temperate conditions despite its high latitude in part because of heat transported poleward by the Gulf Stream. Some climate models have suggested that increases in CO₂ due to human activities could suddenly shift the ocean circulation in the North Atlantic into a regime that transports heat less efficiently, which could cause widespread cooling in Europe (although this might be overwhelmed by the expected global warming that will occur over the next century). The rapidity with which ice ages ended also suggests that major reorganizations of the ocean/atmosphere circulation occurred during those times. Although thresholds play a crucial role in past and possibly future climate change, they are notoriously difficult to predict because they involve subtle nonlinear interactions.

### 5.3. Recent Times

A wide range of evidence demonstrates that Earth’s global mean surface temperature rose by about 0.6 °C between 1900 and 2000 (see Figure 20.4). Over the past 50 years, the global mean rate of temperature increase has been ~0.13 °C per decade (with a greater rate of warming over land than ocean). As of 2006, 20 of the hottest years measured since good instrumental records started in ~1860 have occurred within the past 25 years, and the past 25 years has been the warmest 25-year period of the past 1000 years.

There is widespread consensus among climate experts that the observed warming since ~1950 has been caused primarily by the release of CO₂ due to human activities, primarily the burning of oil, coal, natural gas, and forests: the greater CO₂ concentration has increased the strength of the greenhouse effect, modified by the feedbacks discussed in Section 5.2. Before the Industrial Revolution, the CO₂ concentration was ~280 ppm (i.e. a mole fraction of $2.8 \times 10^{-4}$), as shown in Figure 20.5. Starting approximately in 1800, however, the atmospheric CO₂ abundance began rising rapidly, and in 2012, the CO₂ concentration was 391 ppm—a 40% increase over pre-Industrial Revolution values. Evidence indicates that the sharp rise of CO₂ since 1800 is not a natural climate cycle but the result of human activity. Interestingly, only half of the CO₂ released by human activities each year remains in the atmosphere; the remainder is currently absorbed by the biosphere and especially the oceans.

Superposed on top of the mean rise of temperature with time since ~1950 are numerous short-term fluctuations associated with weather and short-period regional or global climate cycles such as El Niño, temporary shifts in the latitudes and strengths of the jet streams, and other effects. These short-term, year-to-year fluctuations are visible as the jittery year-to-year variation of the blue points in Figure 20.4. This necessarily means that, in some years, the mean climate is warmer than the previous year, while in other years, it is cooler than the previous year. This fact is often quoted by climate skeptics in the popular press as being evidence against global warming. Figure 20.4, however, shows that this argument is specious. Despite the year-to-year fluctuations, the overall long-term trend is clearly toward a warmer climate.

This increase in mean surface temperature has been accompanied by numerous other climate changes, including retreat of glaciers worldwide, thawing of polar...
permafrost, early arrivals of spring, late arrivals of autumn, changes in the Arctic sea ice thickness, approximately 0.1–0.2 m of sea level increase since 1900, and various effects on natural ecosystems. These changes are expected to accelerate in the twenty-first century.

### 5.4. Ice Ages

The repeated occurrence of ice ages, separated by warmer interglacial periods, dominates Earth’s climatic record of the past 2 million years. During an ice age, multikilometer-thick ice sheets grow to cover much of the high-latitude land area, particularly in North America and Europe; most or all of these ice sheets melt during the interglacial periods (however, ice sheets on Antarctica and Greenland have resisted melting during most interglacials, and these two ice sheets still exist today). The sea level varies by up to 120 m between glacial and interglacial periods, causing migration of coastlines by hundreds of kilometers in some regions. The time history of temperature, ice volume, and other variables can be studied using stable isotopes of carbon, hydrogen, and oxygen as recorded in glacial ice, deep-sea sediments, and land-based records such as cave calcite and organic material. This record shows that glacial/interglacial cycles over the past 800,000 years have a predominant period of ~100,000 years (Figure 20.6). During this cycle, glaciers gradually increase in volume (and air temperature gradually decreases) over most of the 100,000-year period; the glaciers then melt, and the temperature increases over a relatively short ~5000-year interval. The cycle is thus extremely asymmetric and resembles a sawtooth curve rather than a sinusoid. The last ice age peaked 18,000 years ago and ended by 10,000 years ago; the modern climate corresponds to an interglacial period. Analysis of ancient air trapped in air bubbles inside the Antarctic and Greenland ice sheets shows that the atmospheric CO₂ concentration is low during ice ages—typically about 200 ppm—and rises to ~280 ppm during the intervening interglacial periods (Figure 20.6).

Ice ages seem to result from changes in the strength of sunlight caused by periodic variations in Earth’s orbit, magnified by several of the feedbacks discussed in Section 5.2. A power spectrum of the time series in Figure 20.6 shows that temperature, ice volume, and CO₂ vary predominantly on periods of 100,000, 41,000, 23,000, and 19,1000 years (ka; the summation of sinusoids at each of these periods leads to the sawtooth patterns in Figure 20.6).

Interestingly, these periods match the periods over which Northern Hemisphere sunlight varies due to orbital oscillations. The Earth’s orbital eccentricity oscillates on periods of 100 ka, the orbital obliquity (the tilt of Earth’s rotation axis) oscillates on a period of 41 ka, and the Earth’s rotation axis precesses on periods of 19 and 23 ka. These variables affect the difference in sunlight received at Earth between winter and summer and between the equator and pole. In turn, these sunlight variations determine the extent...
to which snowpack accumulates in high northern latitudes during winter, and the extent to which this snowpack resists melting during summer; glaciers build up when snow that falls during winter cannot melt the following summer. The idea that these orbital variations cause ice ages has become known as the Milankovitch theory of ice ages.

By themselves, however, orbital variations are only part of the story. Sunlight variations due to the 100-ka eccentricity variations are much weaker than sunlight variations due to the 41-, 23-, and 19-ka obliquity and precession variations. Thus, if the orbit-induced sunlight variations translated directly into temperature and ice variations, ice ages would be dominated by the 41-, 23-, and 19-ka periods, but instead, the 100-ka period dominates (as can be seen in Figure 20.6). This means that some nonlinearity in the climate system amplifies the climatic response at 100 ka much better than at the shorter periods. Furthermore, the observed oscillations in CO₂ between glacial and interglacial periods (Figure 20.6) indicates that ice ages are able to occur partly because the greenhouse effect is weak during ice ages but strong during interglacial periods. Most likely, atmospheric CO₂ becomes dissolved in ocean water during ice ages, allowing the atmospheric CO₂ levels to decrease; the ocean then rejects this CO₂ at the end of the ice age, increasing its atmospheric concentration. Recent analyses of Antarctic ice cores show that, at the end of an ice age, temperature rise precedes CO₂ rise in Antarctica by about 800 years, indicating that CO₂ variation is an amplifier rather than a trigger of ice age termination. Interestingly, however, both of these events precede the initiation of deglaciation in the Northern Hemisphere. These observations suggest that the end of an ice age is first triggered by a warming event in the Antarctic region; this initiates the process of CO₂ rejection from the oceans to the atmosphere, and the resulting increase in the greenhouse effect, which is global, then allows deglaciation to commence across the rest of the planet. The ice-albedo and water vapor feedbacks (Section 5.2) help amplify the

FIGURE 20.6 CO₂ concentrations (top) and temperature variations (bottom) over the last 420,000 years as obtained from ice cores at Vostok, Antarctica (data from Petit et al., 1999). The approximate 100,000-year period of the ice ages is evident, although many shorter period fluctuations are superimposed within the record. Prominent ice age terminations occurred at ~410, 320, 240–220, 130, and 15 ka in the past. Also note the correlation between temperature and CO₂ concentration during these cycles, which shows the influence of changes in the greenhouse effect on ice ages. The vertical line at the right side of the top plot shows the increase in CO₂ caused by humans between ~1800 and 2012.
transition. However, many details, including the exact mechanism that allows CO₂ to oscillate between the ocean and atmosphere, remain to be worked out.

Figure 20.6 shows how the increase in CO₂ caused by human activities compares to the natural variability in the past. The saw-toothed variations in CO₂ between 200 and 280 ppm over 100,000-year periods indicate the ice age/interglacial cycles, and the vertical spike in CO₂ at the far right of Figure 20.6 (from 280 to 391 ppm, also visible in the last ~200 years of the time series in Figure 20.5) shows the human-induced increase. The current CO₂ concentration far exceeds that at any previous time over the past 420,000 years, and is probably the greatest CO₂ level the Earth has seen since 20 million years ago. The fact that CO₂ rises by 30–40% at the end of an ice age indicates that very large magnitude climate changes can accompany modest CO₂ variations; it is noteworthy that human activities have so far increased CO₂ by an additional 36% beyond pre-industrial levels. The relationship between CO₂ and global temperature during ice ages may differ from the relationship these quantities will take over the next century of global warming; however, it is virtually certain that additional CO₂ will cause global temperature increases and widespread climate changes. Current economic and climate projections indicate that, because of continued fossil fuel burning, the atmospheric CO₂ will reach 500–1000 ppm by the year 2100 unless drastic measures are adopted to reduce fossil fuel use.

5.5. Volatile Inventories of Terrestrial Planets

Venus, Earth, and Mars have present-day atmospheres that are intriguingly different. The atmospheres of Venus and Mars are both primarily CO₂, but they represent two extreme fates in atmospheric evolution: Venus has a dense and hot atmosphere, whereas Mars has a thin and cold atmosphere. It is reasonable to ask whether Earth is ultimately headed toward one or the other of these fates, and whether these three atmospheres have always been so different.

The history of volatiles on the terrestrial planets includes their origin, their interactions with refractory (nonvolatile) material, and their rates of escape into space. During the initial accretion and formation of the terrestrial planets, it is thought that most or all of the original water reacted strongly with the iron to form iron oxides and hydrogen gas, with the hydrogen gas subsequently escaping to space. Until the iron cores in the planets were completely formed and this mechanism was shut down, the outflow of hydrogen probably took much of the other solar-abundance volatile material with it. Thus, one likely possibility is that the present-day atmospheres of Venus, Earth, and Mars are not primordial, but have been formed by outgassing and by cometary impacts that have taken place since the end of core formation.

The initial inventory of water that each terrestrial planet had at its formation is a debated question. One school of thought is that Venus formed in an unusually dry state compared with Earth and Mars; another is that each terrestrial planet must have started out with about the same amount of water per unit mass. The argument for an initially dry Venus is that water-bearing minerals would not condense in the high-temperature regions of the protoplanetary nebula inside of about 1 AU. Proponents of the second school of thought argue that gravitational scattering caused the terrestrial planets to form out of materials that originated over the whole range of terrestrial planet orbits, and therefore that the original water inventories for Venus, Earth, and Mars should be similar.

An important observable that bears on the question of original water is the enrichment of deuterium (D) relative to hydrogen. A measurement of the D/H ratio yields a constraint on the amount of hydrogen that has escaped from a planet. For the D/H ratio to be useful, one needs to estimate the relative importance of the different hydrogen escape mechanisms and the original D/H ratio for the planet. In addition, one needs an idea of the hydrogen sources available to a planet after its formation, such as cometary impacts. The initial value of D/H for a planet is not an easy quantity to determine. A value of 0.2 × 10⁻⁴ has been put forward for the protoplanetary nebula, which is within a factor of 2 or so for the present-day values of D/H inferred for Jupiter and Saturn. However, the D/H ratio in Standard Mean Ocean Water (a standard reference for isotopic analysis) on Earth is 1.6 × 10⁻⁴, which is also about the D/H ratio in hydrated minerals in meteorites, and is larger by a factor of 8 over the previously mentioned value. At the extreme end, some organic molecules in carbonaceous chondrites have shown D/H ratios as high as 20 × 10⁻⁴. The enrichment found in terrestrial planets and most meteorites over the protoplanetary nebula value could be the result of exotic high-D/H material deposited on the terrestrial planets, or it could be the result of massive hydrogen escape from the planets early in their lifetimes through the hydrodynamic blowoff mechanism (which is the same mechanism that currently drives the solar wind off the Sun).

6. LIFE IN THE ATMOSPHERE–OCEAN SYSTEM

6.1. Interplanetary Spacecraft Evidence for Life

An ambitious but ever-present goal in astronomy is to detect or rule out life in other solar systems, and in
planetary science that goal is to detect or rule out life in our own solar system apart from the Earth. Water in its liquid phase is one of the few requirements shared by all life on Earth, and so the hunt for life is focused on the search for liquid water. We know that Mars had running water on its surface at some point in its history because we can see fluvial channels in high-resolution images, and because the Mars rovers Spirit and Opportunity have discovered aqueous geochemistry on the ground; there is even some evidence suggesting present-day seepage in recent orbiter images. Farther out in the solar system, we know that Europa, a satellite of Jupiter, has a smooth icy surface with cracks and flow features that resemble Earth’s polar ice fields and suggest a liquid water interior, while its larger sibling, Ganymede, exhibits a conductive reaction to Jupiter’s magnetic field that is most easily explained by a salty liquid water interior (see Mars: Surface and Interior; Meteorites; Planetary Satellites).

However, to date we have no direct evidence for extraterrestrial life. This includes data from landers on Venus, Mars, and the Moon, and flyby encounters with eight planets, a handful of asteroids, a comet (Halley in 1986), and over 60 moons. Are the interplanetary spacecraft we have sent out capable of fulfilling the goal of detecting life? This question has been tested by analyzing data from the Galileo spacecraft’s two flyby encounters with the Earth, which, along with a flyby encounter with Venus, were used by the spacecraft’s navigation team to provide gravity assists to send Galileo to Jupiter. The idea was to compare ground-truth information to what we can learn solely from Galileo (see Atmospheres of the Giant Planets; Io: the Volcanic Moon; Planetary Satellites; Planetary Exploration Missions).

Galileo’s first Earth encounter occurred on December 8, 1990, with closest approach 960 km above the Caribbean Sea; its second Earth encounter occurred on December 8, 1992, with closest approach 302 km above the South Atlantic. A total of almost 6000 images were taken of Earth by Galileo’s camera system. Figure 20.7 shows the Earth–Moon system as seen by Galileo. Notice that the Moon is significantly darker than the Earth. The spacecraft’s instruments were designed and optimized for Jupiter; nevertheless, they made several important observations that point to life on Earth. These strengthen the null results encountered elsewhere in the solar system. The evidence for life on Earth includes complex radio emissions, non-mineral surface pigmentation, disequilibrium atmospheric chemistry, and large oceans.

6.1.1. Radio Emissions

The only clear evidence obtained by Galileo for intelligent life on Earth was unusual radio emissions. Several natural radio emissions were detected, none of which were unusual, including solar radio bursts, auroral kilometric radiation, and narrowband electrostatic oscillations excited by thermal fluctuations in Earth’s ionospheric plasma. The first unusual radio emissions were detected at 1800 UT and extended through 2025 UT, just before closest approach. These were detected by the plasma wave spectrometer on

FIGURE 20.7 The Earth–Moon system as observed by the Galileo spacecraft.
the nightside, in-bound pass, but not on the day side, out-bound pass. The signal strength increased rapidly as Earth was approached, implying that Earth itself was the source of the emissions. The fact that the signals died off on the dayside suggests that they were cut off by the day side ionosphere, which means we can place the source below the ionosphere.

The unusual signals were narrowband emissions that occurred in only a few distinct channels and had average frequencies that remained stable for hours. Naturally occurring radio emissions nearly always drift in frequency, but these emissions were steady. The individual components had complicated modulations in their amplitude that have never been detected in naturally occurring emissions. The simplest explanation is that these signals were transmitting information, which implies that there is advanced technological life on Earth. In fact, the radio, radar, and television transmissions that have been emanating from Earth over the past century result in a nonthermal radio television emissions that have been emanating from technological life on Earth. In fact, the radio, radar, and television transmissions that have been emanating from Earth over the past century result in a nonthermal radio emission spectrum that broadcasts our presence out to interstellar distances (see The Solar System at Radio Wavelengths).

### 6.1.2. Surface Features

During its first encounter with the Earth, the highest resolution mapping of the surface by Galileo’s Solid-State Imaging System (SSI) covered Australia and Antarctica with 1–2 km resolution. No usable images were obtained from Earth’s nightside on the first encounter. The second encounter netted the highest resolution images overall of Earth by Galileo, 0.3–0.5 km/pixel, covering parts of Chile, Peru, and Bolivia. The map of Australia from the first encounter includes 2.3% of Earth’s total surface area, but shows no geometric patterns that might indicate an advanced civilization. In the second encounter, both the cities of Melbourne and Adelaide were photographed, and yet no geometric evidence is visible because the image resolution is only 2 km. The map of Antarctica, 4% of Earth’s surface, reveals nearly complete ice cover and no signs of life. Only one image, taken of Southeastern Australia during the second encounter, shows east–west and north–south markings that would raise suspicions of intelligent activity. The markings in fact were caused by boundaries between wilderness areas, grazing lands, and the border between South Australia and Victoria. Studies have shown that it takes nearly complete mapping of the surface at 0.1-km resolution to obtain convincing photographic evidence of an advanced civilization on Earth, such as roads, buildings, and evidence of agriculture.

On the other hand, many features are visible in the Galileo images that have not been seen on any other body in the solar system. The SSI camera took images in six different wavelength channels. A natural-color view of Earth was constructed using the red, green, and violet filters, which correspond to wavelengths of 0.670, 0.558, and 0.407 μm, respectively. The images reveal that Earth’s surface is covered by enormous blue expanses that specularly reflect sunlight, and end in distinct coastlines, which are both easiest to explain if the surface is liquid. This implies that much of the planet is covered with oceans. The land surfaces show strong color contrasts that range from light brown to dark green.

The SSI camera has particular narrowband infrared filters that have never been used to photograph Earth before, and so they yielded new information for geological, biological, and meteorological investigations. The infrared filters allow the discrimination of H_2O in its solid, liquid, and gaseous forms; for example, clouds and surface snow can be distinguished spectroscopically with the 1-μm filter. False-color images made by combining the 1-μm channel with the red and green channels reveal that Antarctica strongly absorbs 1-μm light, establishing that it is covered by water ice. In contrast, large regions of land strongly reflect 1 μm without strongly reflecting visible colors, which conflicts with our experience from other planetary surfaces and is not typical of igneous or sedimentary rocks or soil. Spectra made with the 0.73- and 0.76-μm channels reveal several land areas that strongly absorb red light, which again is not consistent with rocks or soil. The simplest explanation is that some nonmineral pigment that efficiently absorbs red light has proliferated over the planet’s surface. It is hard to say for certain if an interstellar explorer would realize that this is a biological mechanism for gathering energy from sunlight, probably so, but certainly we would recognize it on another planet as the signature of plant life. We know from ground truth that these unusual observations are caused by the green pigments chlorophyll a (C_{55}H_{72}MgN_{4}O_{5}) and chlorophyll b (C_{55}H_{70}MgN_{4}O_{6}), which are used by plants for photosynthesis. No other body in the solar system has the green and blue colorations seen on Earth (see The Solar System at Ultraviolet Wavelengths; Infrared Views of the Solar System From Space).

### 6.1.3. Oxygen and Methane

Galileo’s Near-Infrared Mapping Spectrometer (NIMS) detected the presence of molecular oxygen (O_2) in Earth’s atmosphere with a volume mixing ratio of 0.19 ± 0.05. Therefore, we know that the atmosphere is strongly oxidizing. (It is interesting to note that Earth is the only planet in the solar system where one can light a fire.) In light of this, it is significant that NIMS also detected methane (CH_4) with a volume mixing ratio of 3 ± 1.5 × 10^{-6}. Because CH_4 oxidizes rapidly into H_2O and CO_2, if thermodynamical equilibrium holds, then there should be no detectable CH_4 in Earth’s atmosphere.
The discrepancy between observations and the thermodynamic equilibrium hypothesis, which works well on other planets (e.g., Venus), is an extreme 140 orders of magnitude. This fact provides evidence that Earth has biological activity and that it is based on organic chemistry. We know from ground truth that Earth’s atmospheric methane is biological in origin, with about half of it coming from nonhuman activity like methane bacteria and the other half coming from human activity like growing rice, burning fossil fuels, and keeping livestock. NIMS also detected a large excess of nitrous oxide (N₂O) that is most easily explained by biological activity, which we know from ground truth comes from nitrogen-fixing bacteria and algae.

The conclusion is that the interplanetary spacecraft we have sent out to explore our solar system are capable of detecting life on planets or satellites, both the intelligent and primitive varieties, if it exists in abundance on the surface. On the other hand, if there is life on a planet or satellite that does not have a strong signature on the surface, as would probably be the case if Europa or Ganymede harbor life, then a flyby mission may not be adequate to decide the question. With regard to abundant surface life, we have a positive result for Earth and a negative result for every other body in the solar system.

7. CONCLUSIONS

Viewing Earth as a planet is the most important change of consciousness that has emerged from the space age. Detailed exploration of the solar system has revealed its beauty, but it has also shown that the home planet has no special immunity to the powerful forces that continue to shape the solar system. The ability to remotely sense Earth’s dynamic atmosphere, oceans, biosphere, and geology has grown up alongside our ever-expanding ability to explore distant planetary bodies. Everything we have learned about other planets influences how we view Earth. Comparative planetology has proved in practice to be a powerful tool for studying Earth’s atmosphere and oceans. The lion’s share of understanding still awaits us, and in its quest we continue to be pulled outward.

BIBLIOGRAPHY


1. INTRODUCTION: THE EARTH AS A GUIDE TO OTHER PLANETS

The surface of the Earth is perhaps the most geochemically diverse and dynamic among the planetary surfaces of our solar system. Uniquely, it is the only one with liquid water oceans under a stable atmosphere and—as far as we now know—it is the only surface in our solar system that has given rise to life. The Earth’s surface is a dynamic union of its solid crust, its atmosphere, its hydrosphere, and its biosphere, all having acted in concert to produce a constantly renewing and changing symphony of form (Figure 21.1).

The unifying theme of the Earth’s surficial system is water—in liquid, vapor, and solid phases—which transfers and dissipates solar, mechanical, chemical, and biological energy throughout global land and submarine landscapes (e.g., Ritter, Kochel, & Miller, 2011). The surface is a window to the interior processes of the Earth, as well as the putty that atmospheric processes continually shape. It is also the Earth’s interface with extraterrestrial processes and, as such, has regularly borne the scars of impacts by meteors, comets, and asteroids, and will continue to do so. The February 15, 2013, impact of an approximately 17- to 20-m-diameter (12–13 × 10^3 kg) body over Chelyabinsk, Russia, is a potent reminder that impacts are still modifying the earth and other bodies in the solar system. Much larger impacting bodies (e.g., >1 km diameter), although now rare, are still possible, and pose a serious threat to society. (See Planetary Impacts.)

Our solar system has a variety of terrestrial planets and satellites in various hydrologic states with radically differing hydrologic histories. Some appear nearly totally desiccated, such as the Moon, Mercury, and Venus. Even in those places, water may yet prove to be more prominent. For instance, recently, evidence of probable magmatic water in the form of hydroxyl molecules found in spacecraft observations of equatorial lunar impact craters, combined with reanalyses of Apollo lunar samples, may argue for a source of magmatic water within the moon. Isotopic data from these samples are also somewhat problematic. Such findings may cause revision of theories of the origin and early evolution of the earth–moon system.
In some places water is very abundant now at the surface, such as on the Earth, on the Jovian Galilean satellite Europa (solid at the surface and possibly liquid underneath, with strong indications of surface water eruption plumes at its south pole), on the Saturnian satellite Enceladus (erupting water vapor into space through an icy surface), and on Titan, Saturn’s largest moon (where a 94K surface temperature makes water ice at least as hard as granite). In other places, such as Mars and Ganymede, it appears that water may have been very abundant in liquid form on the surface in the distant past. Also, in the case of Mars, water may yet be abundant in solid and/or liquid form in the subsurface today. Thus, for understanding geological (and, where applicable, biological) processes and environmental histories of terrestrial planets and satellites within our solar system, it is crucial to explore the geomorphology of surface and submarine landforms and the nature and history of the land—water interface where it existed. Such an approach and “lessons learned” from this solar system will also be key in future reconnaissance of extrasolar planets. (See Mars: Surface and Interior.)

For the planets, remote sensing techniques, especially as implemented at optical (e.g. visible, near-infrared, and thermal infrared) and at radio frequencies (e.g. radars), have been a primary exploration tool. Starting with telescopic observations from earth-based observatory, through the phase of flyby and orbital observations, and now by landers and rovers, the planets and their satellites have become very real places to interested observers on this planet. Over the last 20 years, a number of planetary data archives have been created, and are now conveniently online, accessible over the Internet. It is thus particularly ironic, that effective, accessible, systematically organized digital data archives containing earth remote sensing image (and other) data, have only come to the fore within about the last 10 years. A discussion of how such data are archived, accessed, and standardized, both within the planetary and terrestrial contexts, is an important one, but beyond the scope of this treatment. Nevertheless, it is significant to point out that for many reasons, some historical, some technical, and some psychological, because of our intimate familiarity with the Earth, practitioner’s of Earth Science, have tended to separate the various realms of scientific investigations of our planet, as the realms themselves seem separate to us humans from the vantage point of our existence and history on the Earth’s land surface. The lack of emphasis on studying the Earth as a system—land, ocean, atmosphere, and near space—has hampered our full understanding of the complex interplay, history, and evolution of processes within these milieus.

Now, however, and especially within the last decade, new systematically organized archives of earth data, acquired using remote sensing techniques and predominately time-series image data, including topographic and seismic data, have emerged. This is, in large part, because of a conscious attempt of national scientific and space agencies to address the Earth as a system. Such has been particularly true in the United States within the decades-long orbitally focused National Aeronautics and Space Administration (NASA) Earth Observing System (EOS) project. Currently the EOS data are warehoused within the EOS Data Information System with a variety of subarchives representing general discipline areas, such as the Land Surface Data Acquisition and Analysis Center (LPDAAC; Sioux Falls, South Dakota, USA, operated under an agreement between NASA and the US Geological Survey). The LPDAAC provides online access to data from a variety of US earth orbital missions, acquired over many years.

For workers in a specific discipline, data access can be challenging within such a generalized large volume “Big Data” archive as the LPDAAC, so smaller more focused specialty archives have developed. One specific archive example that draws on data from the United States—Japan Advanced Spaceborne Thermal Emission and Reflection (ASTER) radiometer mission (1999—present) is the ASTER Volcano Archive at NASA’s Jet Propulsion Laboratory in Pasadena, California, USA (http://ava.jpl.nasa.gov). It puts data that reflect ASTER’s unique capabilities to see volcanic activity worldwide (14 optical channels between 0.4 and 12 μm; Pieri & Abrams, 2004) for over
Chapter | 21 Earth as a Planet: Surface and Interior

1500 volcanoes in the Smithsonian Global Volcanism catalog for the period 2000—present, at the fingertips of online professional and general users. Archive holdings include basic and higher level data products (e.g. topography and thermal emission analyses) in convenient formats that can be displayed in Google Earth™ and Google Maps™ clients. Also included in such archives is instrument-specific in situ data (e.g. atmospheric profiles and ground-based and low-altitude surface temperature and emissivity measurements), acquired by both manned aircraft and newly employed unmanned aerial vehicles (e.g. Pieri et al., 2013).

The ready accessibility of such data, worldwide, tends to militate against geoscientific parochialism, and allows a more unified planetary perspective from which to assess earth science data, in this case volcanological data, more consistent with a comparative planetology approach advocated by researchers in solar system studies—examples of which are rife throughout this volume. Sister Federal agencies in the United States, and in other space-faring countries, significantly within Europe, Japan, and India, have also similarly pursued this approach. The net effect is an increasing emphasis on interdisciplinary investigations that necessarily cross traditional subject boundaries and promote a more general understanding of how the Earth system has responded in the past to various external stimuli (e.g. the solar Milankovic cycles) and how it will respond in the future (e.g. increased anthropogenic CO2 inputs). It is becoming more and more clear that such understanding will be a requirement as mankind moves forward through the current century, with profound forecast impacts on terrestrial land surfaces, on our atmosphere, and on our oceans, and thus on the habitability of our global environment. (See Earth as a Planet: Oceans and Atmospheres.)

2. PHYSIOGRAPHIC PROVINCES OF EARTH

2.1. Basic Divisions

From a geographic and geomorphologic point of view, especially when seen from space, the surface of the Earth is dominated by its oceans of liquid water; approximately 75% of the Earth’s surface is covered by liquid or solid water. The remaining 25% of nonmarine subaerial land, the subject of nearly all historical geological and geomorphological study, lies mainly in its Northern Hemisphere, where most of the world’s population lives. The Southern Hemisphere is dominated by oceans, some subaerial continental and archipelago land masses (mainly parts of Africa, South America, southeast Asia, and Australia), and the large, currently subglacial, island continent of Antarctica (Figure 21.2(a)).

Remarkably, despite the fact that geological and geographical sciences have been practiced on the Earth for about 200 years, it has been mainly since the Second World War that scientists have begun detailed mapping and geophysical explorations of the submarine land surface. Subsea remote-sensing technology has provided one of the most profound discoveries in the history of geological science: the paradigm of “plate tectonics”. The extent, morphology, and dynamics of the Earth’s massive tectonic plates were only realized after careful topographic and geomagnetic mapping of the intensely volcanic mid-oceanic ridges and their associated parallel-paired geomagnetic domains.

Similar topographic mapping of the corresponding submarine trenches along continental or island-arc margins was equally revealing. The mid-oceanic ridges were found to be sites of accretion of new volcanically generated plate material, and the trenches the sites of deep subduction, where oceanic crust is consumed beneath overriding crustal plates. Tectonic plates represent the most fundamental and largest geomorphic provinces on Earth.

The Earth’s crustal plates come in two varieties: oceanic and continental (Figure 21.3(a)).

Oceanic plates comprise nearly all of the Earth’s ocean floors, and thus most of the Earth’s crustal area. They are composed almost exclusively of iron- and magnesium-rich rocks derived from volcanic processes (called “basalts”). Oceanic plates are created by volcanic eruptions along the apices of the Earth’s mid-oceanic ridges, 1000-km-long sinuous ridges that rise from the flat ocean floor (called “abyssal plains”) in the middle of oceans. Oceanic plates are typically less than 10 km thick. Here, nearly continuous volcanic activity from countless submarine volcanic centers (far more than the 1000 or so active subaerial volcanoes) provides a steady supply of new basalt, which is accreted and incorporated into the interior part of the plate.

At plate edges, roughly the reverse occurs, where the outer, oldest plate margins are forced below overriding adjacent plate edges. Usually, when two oceanic plates collide, the resulting subduction zone forms an island arc along the trace of the collision. The islands, in this case, are the result of the eruption of lighter, more silica-rich magmas generated as part of the subduction process. The subducted plate margin is consumed along the axis of the resulting trench. Because the more silicic island arcs tend to be less dense and thus more resistant to subduction, they can be accreted onto plate margins and can thus increase the areal extent at the edges of oceanic plates or can enlarge the margins of existing continental plates.

Continental plates tend to consist of much more silicic material, and are thus lighter, as compared with oceanic plates. Because of their lower density and the fact that they are isostatically compensated, they are much thicker than oceanic plates (30–40 km thick) and tend to “float” over the denser, more mafic (ferromagnesian—consisting of mostly of the metals iron and magnesium) subjacent
material in the Earth’s upper mantle. When continental plates collide with oceanic plates, deep subduction trenches, such as the Peru—Chile trench along the west coast of South America, occur, as the oceanic plate is forced under the much thicker and less dense continental plate. Usually, the landward side of the affair is marked by so-called Cordilleran belts of mountains, including andesitic-type volcanoes, which parallel the coastline. The Andes Mountains are an example of this type of tectonic arrangement.

When continental plates collide, a very different tectonic and geomorphic regime ensues. Here, equally buoyant and thick continental plates crush against each other, resulting in the formation of massive fold belts and towering mountains, as long as the tectonic zone is active (e.g. the Himalayan Range in Asia). When aggregate stresses are
FIGURE 21.3  (a) Tectonic plate interactions. Tectonic plate interactions and the three fundamental kinds of plate boundaries. (Left) A convergent boundary caused by the subduction of oceanic material as it is overridden by another oceanic plate. (Center left) A subplate hot spot capped by a shield volcano (e.g. Hawaiian Islands). (Center right) A divergent plate boundary, in particular, a mid-oceanic spreading ridge. (Right) Another kind of convergent plate boundary, where the oceanic crust is being subducted by overriding continental crust, producing a chain of volcanic mountains (e.g. Andes Mountains). (Far right) A continental rift zone, another kind of divergent plate boundary (e.g. East African Rift). Finally, a transform plate boundary is shown at the upper middle of the scene, where two plates are sliding past each other without subduction. The three relationships are shown as block diagrams at the top of the figure. (Courtesy of the US Geological Survey.)

(b) Emperor seamount chain spans Pacific plate. Perspective view of the Emperor seamount chain that spans the central and northwest sector of the Pacific Basin. The southeastern end of the hot spot track terminates in the Hawaiian Islands, and the predominating trend of plate motion has been to the northwest over time. Deeply rooted, persistent hot spots are the result of persistent hot upwelling plumes of lower density material from the upper mantle. "Petit spot" subsea volcanoes may form as hot spots in oceanic plates above or near the intersection of flexure cracks. Courtesy of Google Earth®.
tensional rather than compressive, extensional mountain ranges can form, as tectonic blocks founder and rotate. The western US Basin and Range Province is a good example of that type of mountain terrane. Another large subaerial extensional tectonic landform is the axial rift valley and associated inward-facing fault scarps, which form when aggregate tensional stresses tend to pull a continental plate apart (e.g. the East African Rift Valley). Such rift valleys are often characterized by ubiquitous mafic volcanism (e.g. Afar Triangle).

The geomorphic provinces just discussed generally tend to be very dynamic, with lifetimes that are intrinsically short (100–200 million years) relative to the age of the Earth (4.56 billion years). Some of the stable interior areas of continental plates, or cratons, however, do possess landforms and associated lithologic regions with ages comparable within a factor of two or three to the age of the Earth (2–3 billion years). The interior of the Canadian Shield and the Australian continent are two such special areas. Despite having been scoured repeatedly by continental ice sheets, the granitic craton of the Canadian Shield possesses a record of giant asteroidal and cometary impacts that are about 2 billion years old. (See Planetary Impacts.) These interior cratonic areas, in contrast to most of the rest of the Earth, which is mobile and active, provide a chance to view a part of the long sweep of the Earth’s surface history. They are thus important, particularly in trying to understand how the environmental history of the Earth compares to that of the other terrestrial planets.

The distribution of the earth’s landscape altitudes, relative to the mean geoid, is bimodal—continental and seafloor (Figure 21.4(a)). Although limited in percentage of surface area coverage, the interface between the two modes is a relatively high-energy place called the littoral or tidal zone. Ocean tides in this zone generate frequent (twice daily) environmental stresses on its residents that profoundly encourage evolution and natural selection, and may have been a key influence on the origin and early evolution of life here. It is interesting that Mars is another planet with a global bimodal highland/lowland dichotomy and may have had early oceans, although the absence of large lunar tides may be significant in this context. (See Planets and the Origin of Life.)

2.2. Landform Types

2.2.1. Submarine Landforms

Geomorphically, submarine oceanic basins comprise the areally dominant landform of the Earth, but ironically, they are probably less well explored than the well-imaged surfaces of Mars, Venus, and the satellites of the outer planets. Dominant features of oceanic basins are the oceanic ridge and rise systems, which have a total length of about 60,000 km (~1.5 times the equatorial circumference of the Earth), rise to 1–3 km above the average depth of the ocean, and can be locally rugged. In the Atlantic Ocean, oceanic rises exhibit a central rift valley that is at the center of the rise, whereas in the Pacific Ocean this is not always present (Figure 21.2(a)).

Older crust within oceanic basins can have gently rolling abyssal hills, which are generally smoother than the ridge and rise systems. These may have been much more rugged originally, but are now buried beneath accumulated sediment cover. Perhaps the most areally dominant feature of ocean basins (with the largest ones occurring in the Atlantic Ocean) is the predominantly flat abyssal plains that stretch for thousands of kilometers, usually also covered with accumulated marine sediments. Generally characterized by little topographic relief, in places they are punctuated by seamounts (Figure 21.4(b)), which are conical topographic rises sometimes topped by coral lagoons, or which sometimes do not reach the oceans’ surface. These features are subsea volcanoes associated with island arcs or with midplate hot spots, such as the famous Emperor seamount chain, the southeastern end of which terminates in the Hawaiian Islands (Figure 21.3(b)). Such large hot spots are probably the result of persistent hot upwelling plumes from the upper mantle. Smaller “petit spot” subsea volcanoes may form above flexure cracks in oceanic plates. (See Planetary Volcanism.)

Oceanic margins represent another important, although more areally restricted, submarine landform province (Figure 21.4(b) and (c)). Because nearly half of the world’s people live within 100 km of them and because seafood is a major food source for most of the world’s population, they comprise a suite of landforms especially critical to the health and well-being of humanity.

“Atlantic style” continental margins tend to exhibit substantial ancient sediment accumulations and a shelf–slope–rise overall morphology, which probably represents submerged subaerial landscapes remnant from the last Ice Age, when the sea level was lower (about 135 m below current sea level, worldwide). Nevertheless, many such margins, and those of related basins (e.g. Hudson’s Bay) appear to us now as “emergent shorelines” (e.g. Figure 21.4(d)), as they undergo postglacial rebound (PGR).

Ice ages were manifested by expansion, then contraction, of the Earth’s ice sheets and mountain glaciers in most high-latitude and high-altitude zones. The most recent global deglaciation event was essentially complete by 6000 years ago, but relative sea levels have continued to change. This continuing change is generally thought to be the result of the earth’s latent viscoelastic response to deglaciation (PGR), as its surface mass was redistributed. Regions that were most heavily glaciated (e.g. Canada and Northwestern Europe) show relative sea level falling at a rate controlled by postglacial crustal isostatic upward
FIGURE 21.4  (a) Global altitude diagrams. At left are histograms of land altitudes and seafloor depth as a percentage of the Earth’s surface area (50 m intervals), illustrating the classic continent—seafloor dichotomy. The interface between the two, subject to tidal and climatic fluctuation stress, is thought to have provided, in part, stimuli for biological evolutionary adaptations. At right is the global hypsometric curve, showing cumulative frequency of global topographic heights. (b) Ocean basin schematic. Principal features of the ocean floor shown in schematic form—height is greatly exaggerated. (c) Topography of the submarine Monterey Canyon, California, USA. The continental shelf offshore of Monterey California showing the Monterey and other canyons. Such canyons are common on shelves on both Atlantic and Pacific margins, often cutting through the shelf and down the continental slope to deep water. (Figures used with permission of the Monterey Bay Aquarium Research Institute (MBARI).) (d) Rebounding Canadian beach. A systematically striated sand beach near Nunavut, Canada (68° 05′ 50.74″ N, 108° 16′ 54.97″ W) seen on July 01, 2013. Each striation marks an episode of isostatic uplift, illustrating how the Arctic Ocean coastline has continued to rebound after the last glacial period. Tides in this area are weak enough such that the strands are preserved. (With permission of P.D. Tillman.) (e) Map of Gravity Recovery and Climate Experiment (GRACE)-derived global postglacial isostatic rebound. Shown is the distribution of global postglacial isostatic rebound as derived from GRACE data, expressed as changes in the surface mass distribution that would cause the changes in gravity if the mass were concentrated at the surface. It is expressed here in millimeters per year of equivalent water thickness. The mass estimates are provided on a 1 × 1° grid and have an estimated ±20% accuracy. With permission of NASA and the GRACE Team. (From Gerou et al., (2013); http://grace.jpl.nasa.gov.) (f) Postglacial sea rise plotted as a function of time. Rise in sea level since the most recent global glaciations. (From Fleming et al. (1998), Fleming (2000), and Milne et al. (2005).) The existence of significant short-term fluctuations versus smooth and gradual change is disputed, although rapid deglaciation, “meltwater pulse 1A”, by consensus is indicated on the plot. Lowest sea level occurred at about the last glacial maximum. Before this, waxing ice sheets resulted in almost continuously decreasing sea level during an approximately 100,000 year interval. (With permission of Robert A. Rohde.) (g) Most recent global sea level rise plotted as a function of time—1870 to present day. Sea level increases illustrated here indicate an average of approximately 0.15 cm/year during the term 1879—2008; however, since 2008 the rate has increased to about 0.30 cm/year. Regional and local trends may be variable depending on postglacial land movement and coastal current variations. (With permission of the U.S. Environmental Protection Agency.)
FIGURE 21.4 (Continued).
adjustment, greater than 1 cm/year in some places (e.g. Hudson Bay). Even in zones distant from where past glaciations occurred, rates of relative sea level adjustment are substantial (Peltier, 1999). The current distribution of areas affected by isostatic adjustment are strikingly displayed in the data from the Gravity Recovery and Climate Experiment twin satellite orbital gravimeter mission, launched in 2002 and operated jointly by NASA and the German Aerospace Center (e.g. Figure 21.4(e); Hanna et al., 2013). Along shorelines where PGR is observed, it somewhat offsets the eustatic (i.e. ocean volume is increasing as land ice melts) sea level rise now being observed as a result of global climate change (Figure 21.4(f); Fleming, 2000; Fleming et al., 1998; Ivins et al., 2013; Milne, Long, & Bassett, 2005). Recently absolute ocean levels have generally risen dramatically and systematically, but observed trends can be variable, with local relative sea levels not uniformly increasing due to idiosyncratic changes in land isostatic adjustment, as well as changes in coastal circulation patterns over extended periods due to long-term changes in weather patterns, and possibly climate.

Such costal changes, as related to a systematic rise in sea level (e.g. Figure 21.4(g)), are some of the biggest environmental challenges facing mankind in the twenty-first century. In countries that are not landlocked, populations tend to be most concentrated near shorelines, due to benefits derived from harbors and ocean transport, harvesting from fisheries, and recreational uses and tourism. Current global estimates of 40% population within 100 km of the world’s shorelines, suggest that with additionally increased population densities and economic activity will come increasingly severe pressures on coastal human infrastructure and natural habitats. Changes in indigenous land cover, introduction of exotic species, and general increase in pollution will lead to narrower biodiversity, destruction of coral reefs, as well as a range of microbiota changes that will result in negative impacts on human health. Increases in pathogens, particularly cholera and hepatitis A, are associated with the expected decreased potability of drinking water sources.

Continental shelves are usually less than about 100 km in width and have very shallow (∼0.1°) topographic slopes. They typically end in a slope break that merges into the continental slope (∼4° slope, about 50 km wide), which in turn merges into a gentle continental rise (∼0.2° slope, about 50 km wide), which then typically transitions into an abyssal plain. Submarine canyons (also probably remnant from the last Ice Age, e.g. Hudson Canyon of the coast of New York) can deeply cut the continental shelf and slope and terminate in broad submarine sediment fan deposits at the seaward canyon outlet. “Pacific style” oceanic margins can be even narrower. Along the margins of continents of the Pacific Rim, a short shelf and slope can terminate into deep submarine trenches, manifested by subduction zones (e.g. South America and Kamchatka), up to 10 km depth. Similar fore-arc submarine morphology is observed along the margins of Pacific island arcs (e.g. Aleutians and Kurile Is). Much shallower “back-arc” basins occur behind the arcs, on the overriding plate (e.g. Sea of Okhotsk). (See Earth as a Planet: Atmosphere and Oceans.)

2.2.2. Subaerial Landforms

The subject of classic geomorphological investigations, and historically far more well studied because they are where people on Earth live, are the “subaerial” landscapes—the quarter of the earth’s surface that is not submerged. These terranes exist almost exclusively on continents; however, some important subaerial landscapes (particularly volcanic ones, e.g. Hawaii and Galapagos Islands) exist on oceanic islands. Most continental landscapes are predominately Cenozoic to late Cenozoic in age, because over that timescale (65 million years or so), the combined action of plate tectonics, constructive landscape processes (e.g. volcanism and sedimentary deposition), and destructive landscape processes (e.g. erosion and weathering) have tended to rearrange, bury, or destroy preexisting continental landscapes at all spatial scales. Thus, while often retaining the imprint of preexisting forms, subaerial landscapes on the Earth are constantly being reinvented.

Because the Earth’s crust is so dynamic, one must realize from the planetary perspective that any geomorphic survey of the Earth’s surface may be representative only of the current continental plate arrangement, and currently associated climatic and atmospheric circulation regimes. Plate tectonics is a powerful force in setting scenarios for continental geomorphology. For instance, during early Cenozoic times the global continental geography was characterized by the warm circumglobal Tethys Sea and higher sea levels than now (possibly linked to higher rates of mid-oceanic spreading), which strongly biased the overall terrestrial climate toward the tropical range (Figure 21.5).

The rearrangement of continental landmasses in the later Cenozoic closed the Tethys Sea, produced a circum-Antarctic ocean, and set up predominantly north—south circulation regimes within the Atlantic and Pacific Oceans. This global plate geography, combined with greater ocean basin volume (linked to lower ridge spreading rates) and the onset of continental glaciation, lowered sea levels, exposing large marine continental self-environments to subaerial erosion. Our current global surface environment reflects a kind of “oceanic recovery” after the last Ice Age, with somewhat higher sea levels. Thus, our current perception of the Earth’s subaerial geomorphic landform inventory is strongly biased by our temporal observational niche in its
environmental history. Hypothetical interstellar visitors who arrived here 50 million years ago or may arrive 50 million years in the future would likely have a much different perception because of this distinctive dynamic character.

Terrestrial subaerial landform suites are the classic landscapes studied in geomorphology. These are listed in Table 21.1 (modified from Baker (1986), and Bloom (1998)). Currently, on Earth, globally dominant subaerial geomorphic regimes are related to the surface transport of liquid water and sediment due to the action of rainfall. Thus drainage basins dominate terrestrial landscapes at nearly all scales, from the continental scale to sub-100 m scales. These include currently active drainage basins in humid and semiarid climatic zones, to only occasionally active or relict drainages in arid zones. Drainage basin topographies

FIGURE 21.5 Continental geography through time. Modern plate tectonic theory is consistent with, and provides the scientific framework for observations of continental drift. Geologic evidence records the breakup of the supercontinent Pangaea about 225–200 million years ago, eventually fragmenting over time to create our familiar continental geography. The Tethys Sea referred to in the text is labeled. Courtesy of the US Geological Survey.
and network topologies, however, are strongly influenced by the interplay of the orogenic aspects of plate tectonics (i.e., mountain building) and prevailing climatic regimes, including the biogenic aspects of climate (e.g., vegetative ground cover). Clearly, areas of rapid uplift (e.g., San Gabriel Mountains, California) have characteristically steep bedrock drainages, where gravitational energies are high enough to scour stream valleys, generally have parallel or digitate (handlike) drainage patterns, have high local flood potentials, and respond strongly to local weather (e.g., spatial scales $10^2$ to $10^3$ km in characteristic dimension). At the other spatial extreme, major continental drainages (e.g., Amazon River, Mississippi River, and Ob River in Siberia—Table 21.1), with highly dendritic (tree-like) overall pattern organization, are low average gradient systems that integrate the effects of a variety of climatic regimes at different spatial scales and tend to respond to mesoscale and larger climatic and weather events (e.g., $10^4$ to $10^5$ km scale).

Subaerial volcanic processes produce characteristic landforms in all terrestrial climate zones (Figure 21.2b). They tend to occur in belts, mainly at plate boundaries, with a few notable oceanic (e.g., Hawaiian Islands) and continental (e.g., the San Francisco volcanic field in Northern Arizona, the Columbia and Snake River volcanic plains in the US Pacific Northwest, and the Deccan Traps in India) exceptions that occur within plate interiors. Although not as massive or as topographically high as their planetary counterparts (e.g., Martian volcanoes such as Olympus Mons), they provide some of the most spectacular and graceful landforms on the Earth’s surface (e.g., Mt Fujiyama, Japan, and Mt Kilamanjaro, Kenya). Our planet’s central vent volcanic landforms range from the majestic stratocone volcanic structures just mentioned to large collapse and resurgent caldrons or caldera features (e.g., Valles Caldera, New Mexico; Yellowstone Caldera, Wyoming; Campi Flegrei, Italy; and Krakatau, Indonesia). More areally extensive and lower subaerial shield volcanoes, formed by more fluid lavas (and thus with topographic slopes generally less than 5°) exist in the Hawaiian Islands, at Piton de la Fournaise (Reunion Island), in Sicily at Mt Etna (compound shield with somewhat higher average slopes, up to $\sim 20^\circ$), and the Galapagos Islands (Equador), for example. Often their areal extent corresponds strongly to the rate of their effusion. Subaerial and submarine volcanoes occur on the Earth at nearly all latitudes. Indeed

### Table 21.1 Classification of Terrestrial Geomorphological Features by Scale

<table>
<thead>
<tr>
<th>Order</th>
<th>Approximate Spatial Scale ($\text{km}^2$)</th>
<th>Characteristic Units (with Examples)</th>
<th>Approximate Timescale of Persistence (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$10^7$</td>
<td>Continents, ocean basins</td>
<td>$10^6$ to $10^9$</td>
</tr>
<tr>
<td>2</td>
<td>$10^5$ to $10^6$</td>
<td>Physiographic provinces, shields, depositional plains, continental-scale river drainage basins (e.g., Amazon, Mississippi rivers, Danube, and Rio Grande)</td>
<td>$10^8$</td>
</tr>
<tr>
<td>3</td>
<td>$10^4$</td>
<td>Medium-scale tectonic units (sedimentary basins, mountain massifs, domal uplifts)</td>
<td>$10^7$ to $10^8$</td>
</tr>
<tr>
<td>4</td>
<td>$10^2$</td>
<td>Smaller tectonic units (fault blocks, volcanoes, troughs, sedimentary subbasins, individual mountain zones)</td>
<td>$10^5$</td>
</tr>
<tr>
<td>5</td>
<td>$10$ to $10^3$</td>
<td>Large-scale erosional/depositional units (deltas, major valleys, and piedmonts)</td>
<td>$10^6$</td>
</tr>
<tr>
<td>6</td>
<td>$10^{-1}$ to $10$</td>
<td>Medium-scale erosional/depositional units or landforms (floodplains, alluvial fans, moraines, smaller valleys, and canyons)</td>
<td>$10^5$ to $10^6$</td>
</tr>
<tr>
<td>7</td>
<td>$10^{-2}$</td>
<td>Small-scale erosional/depositional units or landforms (ridges, terraces, and dunes)</td>
<td>$10^4$ to $10^5$</td>
</tr>
<tr>
<td>8</td>
<td>$10^{-4}$</td>
<td>Larger geomorphic process units (hillslopes and sections of stream channels)</td>
<td>$10^3$</td>
</tr>
<tr>
<td>9</td>
<td>$10^{-6}$</td>
<td>Medium-scale geomorphic process units (pools and riffles, river bars, solution pits)</td>
<td>$10^2$</td>
</tr>
<tr>
<td>10</td>
<td>$10^{-8}$</td>
<td>Microscale geomorphic process units (fluvial and eolian ripples and glacial striations)</td>
<td>$10^{-1}$ to $10^4$</td>
</tr>
</tbody>
</table>

Modified from Baker (1986).
some of the world’s most active volcanoes occur along the
Kurile-Kamchatka-Aleutian arc, in subarctic to arctic en-
vironments, often with significant volcano–ice interaction.
High-altitude volcanoes that occur at more humid, lower
latitudes (e.g. Andean volcanoes like Nevado del Ruiz in
Columbia) can also have significant magma or lava–ice
interactions. Volcanoes also occur in Antarctica, Mt Erebus
being the most active, with a perennial lava pond. (See
Planetary Volcanism.)

2.3. Summary: Terrestrial vs Planetary
Landscapes

Overall, the Earth’s geomorphic or physiographic prov-
inces, as compared to those of the other planets in our solar
system, are distinguished by their variety, their relative
youth, and their extreme dynamism. Many of the other
terrestrial-style bodies, such as the Moon, Mars, and
Mercury, are relatively static, with landscapes more or less
unchanging for billions of years. Although this may not
have been the case early in their histories, as far as we can
tell from spacecraft exploration, this is the case now. Other
landscapes, such as those on Venus and Europa and a few of
the other outer planets’ satellites, appear younger and
appear to be the result of very dynamic planetwide pro-
cesses, and possibly for Venus, a planetwide volcanic
“event”. Currently most of these bodies appear relatively
static, although this point may be credibly debated. For
instance, the Jovian satellite Io has vigorous on-going
volcanic activity as was first discovered in Voyager
spacecraft imaging, and the Saturnian satellite Enceladus
appears to be erupting water from relatively warm spots in
its southern hemisphere, as seen in recent Cassini space-
craft data. Nevertheless, it seems that the crusts of all these
bodies are currently somewhat less variegated than that of
the Earth. Be aware, however, that this last statement may
turn out to be just another example of “Earth chauvinism”,
and will be proved wrong once we eventually know the
lithologies and detailed environmental histories of these
bodies as well as we know the Earth’s. (See Venus: Surface
and Interior.)

3. Earth Surface Processes

The expenditure of energy in the landscape is what sculpts a
planetary surface. Such energy is either “interior” (endo-
genic) or “exterior” (exogenic) in origin. The combined
gravitational and radiogenic thermal energy of the Earth
(endogenic processes) powers the construction of terrestrial
landscapes. Thus, the Earth’s main constructional land-
scape processes, plate tectonics and resulting volcanism,
are endogenic processes.

Destructional processes, such as rainfall-driven runoff
and streamflow, are essentially exogenic processes. That is,
the energy that drives the evaporation of water that event-
ually results in precipitation, and the winds that transport
water vapor, comes from an exterior source—the Sun (with
the possible exception of very local, but often hazardous,
weather effects near explosive volcanic eruptions, and
endogenic energy source). In familiar ways, such destruc-
tional geomorphic processes work to reduce the “gravi-
tational disequilibria” that constructive landscapes represent.
For instance, the relatively low and ancient Appalachian
Mountains, pushed up during one of the collisions between
the North American and European continental landmasses,
were probably once as tall as the current Himalayan chain.
Their formerly steep slopes and high altitudes represented a
great deal of gravitational disequilibria, and thus a great
deal of potential energy that was subsequently expended as
kinetic energy by erosive downhill transport processes
(e.g. rainfall runoff and streamflow). Once the processes of
continental collision ebbed and tectonic uplift ceased,
continuing erosion and surface transport processes (such as
rainfall, associated runoff, snowfall, and glaciation) over
only a few tens of millions of years reduced the proto-
Appalachian Mountains to their present gently sloping
and relatively low-relief state.

Volcanic landforms provide myriad illustrations of the
competition between destructive and constructive pro-
cesses in the landscape. For example, Mt Fuji, the most
sacred of Japanese mountains, is actually an active volcano
that erupts on the order of every 100–150 years. Its
perfectly symmetrical conical shape is the result of volca-
ic eruptions that deposit material faster than it can be
transported away, on average. If Fuji stopped erupting, it
would become deeply incised by stream erosion and it
would lose its classic profile over a geologically short time
interval (Figure 21.6).

3.1. Constructive Processes in the
Landscape

Over the geologic history of the Earth, volcanism has been
one of the most ubiquitous processes shaping its surface.
Molten rock (lava) erupts at the Earth’s surface as a result of
the upward movement of slightly less dense magma. Its
melting and upward migration are triggered by convective
instabilities within the upper mantle. Volcanic processes
very likely dominated the earliest terrestrial landscapes
and competed with meteorite impacts as the dominant
surface process during the first billion years of Earth his-
tory. With the advent of plate tectonics, multiphase melting
of ultramafic rocks tended to distill more silicic lavas.
Because silicate-rich rocks tend to be less dense than more
mafic varieties, they tend to “float” and resist subduction,
thus continental cores (cratons) were generally created and
enlarged by island-arc accretion.
Most volcanism tends to occur on plate boundaries. Subaerial plate boundary volcanism tends to produce island arcs (e.g. Aleutian Islands and Indonesian archipelago) when oceanic plates override one another or subaerial volcanic mountain chains (e.g. Andes) underride more buoyant continental plates. Such volcanism tends to be relatively silica-rich (e.g. andesites), producing lavas with higher viscosities, thus tending to produce steeper slopes. Rough lava flows on these volcanoes tend to be classified as aa or blocky lavas. High interior gas pressures contained by higher viscosity magmas can produce very explosive eruptions, some of which can send substantial amounts of dust, volcanic gas, and water vapor into the stratosphere.

Another kind of volcanic activity tends to occur within continental plates. As is thought to have been widespread on the Moon, Mars, and Venus and to a lesser degree within impact basins on Mercury, continental flood eruptions have erupted thousands of cubic kilometers of layered basalts (e.g. Deccan and Siberian Traps in India and Russia; Columbia River Basalt Group in the United States). These are among the largest single subcontinental landforms on the Earth. Such lavas were mafic, of relatively low viscosity, and are thought to have erupted from extended fissure vents at very high eruption rates over relatively short periods (1–10 years). Recent work on the 100-km-long Carrizozo flow field in New Mexico, however, suggests that such massive deposits may have formed at much lower volume effusion rates over much longer periods than previously thought (10–100 years or more). The same may be true for lava flows of similar appearance on other planets. (See Planetary Volcanism.)

Perhaps the most familiar kind of subaerial volcanism is the well-behaved, generally nonexplosive, Hawaiian-style low-viscosity eruptions of tholeiitic basalts that form shield volcanoes, erupting in long sinuous flows. Typically such flows are either very rough (“aa”) (Figure 21.7(a)) with well-defined central channels and levees or very smooth, almost glassy (“pahoehoe”) (Figure 21.7(b)).

These lavas are thought to be comparable to lavas observed in remote sensing images of Martian central vent volcanoes (e.g. Alba Patera and Olympus Mons). Shield volcanoes on both planets tend to exhibit very low slopes (i.e. \( \sim 5^\circ \)). Active submarine basaltic volcanoes tend to occur along mid-oceanic ridges. Often the hot sulfide-rich waters circulating at erupting submarine venting sites provide habitats for a wide variety of exotic chalcophile (sulfur-loving) biota found nowhere else on Earth and proposed as a model for submarine life on Europa.

The transport of water across the land surface also has a hand in forming constructional landforms. Sediment erosion, transportation, and deposition can set the stage for a variety of landscapes, especially in concert with continental-scale tectonic (“epirogenic”) uplift. The Colorado Plateau in the southwestern United States is perhaps the best example of this type of landscape. The Grand Canyon of the Colorado River slices through the heart of the Colorado Plateau and exposes over 5000 vertical feet of sedimentary layers, the oldest of which date to the beginning of the Cambrian era (Figure 21.8(a)).

Water itself can form constructive landforms on the Earth. In its solid form, water can be thought of as another solid component of the Earth’s crust, essentially as just another rock. Under the present climatic regime, the Earth’s great ice sheets—Antarctica and Greenland—along with numerous valley glaciers scattered in mountain ranges across the world in all climatic zones, compose a distinct suite of landforms. Massive (up to kilometers thick) deposits of perennial ice form smooth, crevassed, plastically deforming layers of glacial ice. Continental ice sheets depress the upper crust upon which they reside and can scour the subjacent rocky terrains to bedrock, as during the Wisconsin Era glaciation in Canada (i.e. last Ice Age in North America). Valley glaciers, mainly by mechanical and chemical erosion in concert, tend to carve out large hollows (cirques) in their source areas and have large outflows of meltwater at their termini (Figure 21.8(b)).

3.2. Destructive Geomorphic Processes

Friction probably represents the largest expenditure of energy as geologic materials move through the landscape: friction of water (liquid or solid) on rock, friction of the
wind, friction of rock on rock, or friction of rock on soil. All these processes are driven by the relentless force of gravity and generally express themselves as transport of material from a higher place to a lower one. Erosion (removal and transport of geologic materials) is the cumulative result, over time reducing the average altitude of the landscape and often resculpting or eliminating preexisting landforms of positive relief (e.g. mountains) and incising landforms of negative relief (e.g. river valleys or canyons). Overall, the source of potential energy for these processes (e.g. the height of mountain ranges) is provided by the tectonic activity of plates as they collide or subduct.

Subaerial landscapes on the Earth are most generally dominated by erosive processes, and subaqueous landscapes are generally dominated by depositional processes. Thus, from a planetary perspective, it is the ubiquitous availability and easy transport of water, mostly in liquid form, that makes it the predominant agent of sculpting terrestrial landscapes on Earth. Based on the geologic record of ancient landscapes, it appears that this has been the case for eons on the Earth. Such widespread and constant erosion does not appear to have happened for such a long time on any other planet in the solar system, although it appears that Mars may have had a period of time when aqueous erosion was important and even prevalent.

Fluvial erosion and transport systems (river and stream networks) dominate the subaerial landscapes of the Earth, including most desert areas. Even in deserts where aeolian (wind-driven, e.g. sand dunes) deposits dominate the current landscape, the bedrock signature of ancient river
FIGURE 21.8  (a) Classic view of the Grand Canyon, Colorado, USA. Classic view of the Grand Canyon of the Colorado River in Arizona. The massive layering records the local geologic history for at least the last 500 million years. Comparable layering has also been observed recently in canyons on Mars. This simulated true color perspective view over the Grand Canyon was created from Advanced Spaceborne Thermal Emission and Reflection (ASTER) radiometer data acquired on May 12, 2000. The Grand Canyon Village is in the lower foreground; the Bright Angel Trail crosses the Tonto Platform, before dropping down to the Colorado River and then to the Phantom Ranch (green area across the river). Bright Angel Canyon and the North Rim dominate the view. At the top center of the image the dark blue area with light blue haze is an active forest fire. (Courtesy NASA/GSFC/METI/ERSDAC/JAROS, and U.S./Japan ASTER Science Team.) (b) Bhutan Glaciers, Himalayan Mountains, Asia. Classic Himalayan valley glaciers in Bhutan, showing theaterlike “cirque” source areas, long debris-covered ice streams, and terminal meltwater lakes. ASTER data have revealed significant spatial variability in glacier flow, with velocities from 10 to 200 m/year. Meltwater volumes have been increasing in recent years and threaten to breach terminal moraine deposits with consequent dangerous downstream flooding. This ASTER scene acquired on November 20, 2001, is centered near 28.3° N latitude, 90.1° E longitude, and covers an area of 32.3 × 46.7 km. Courtesy NASA/GSFC/METI/ERSDAC/JAROS, and U.S./Japan ASTER Science Team.