1 A BRIEF HISTORY OF SOLAR-TERRESTRIAL PHYSICS

C. T. Russell

SOLAR-TERRESTRIAL PHYSICS is principally concerned with the interaction of energetic charged particles with the electric and magnetic fields in space. In the vicinity of the earth, most of these charged particles derive their energy ultimately from the sun or from the interaction of the solar wind with the earth's magnetosphere. These interactions are complex, because the magnetic and electric fields that determine the motion of the particles are affected in turn by the motion of these charged particles. Some solar-terrestrial research is carried out on the surface of the earth with cameras, photometers and spectrometers, and magnetometers and other devices sensitive to the processes occurring high in the upper atmosphere and magnetosphere, but today most of this research is carried out using rockets and satellites that enable measurements to be obtained directly in the regions of strongest interactions. In recent years, these in situ data have resulted in explosive growth in our knowledge and understanding of solar-terrestrial processes. Nevertheless, the field has had a long history of investigation, starting well before the advent of satellites and rockets. We shall briefly review that history in order to provide a context for our later, more physically oriented presentation of the processes occurring in the solar-terrestrial environment.

1.1 ANCIENT AURORAL SIGHTINGS

The emerging field of solar-terrestrial physics began with a growing appreciation of two phenomena: the aurora and, later, the geomagnetic field. Because it can be observed visually, the aurora was the first of these phenomena to be recorded. Most other phenomena of this emerging field awaited the advent of new technology, such as the compass in the case of geomagnetism, before they were discovered. References to the aurora are contained in the ancient literature from both East and West. Several passages in the Old Testament appear to have been inspired by auroral sightings, and Greek literature includes references to phenomena most likely to have been auroral phenomena. For example, Xenophanes, in the sixth century B.C., mentions "moving accumulations

FIG. 1.1. Early drawing of the aurora, 12 January 1570. (Original print in Crawford Library, Royal Observatory, Edinburgh.)

of burning clouds." Chinese literature also describes possible auroral sightings, of which several occurred prior to 2000 B.C.

Because the phenomenon was not understood, much fear and superstition surrounded those early sightings of the aurora. Figure 1.1, inspired by an auroral display in 1570, illustrates the lack of scientific understanding prevalent at that time. The seventeenth century marked the beginning of scientific theories concerning the origin of the lights in the north. Galileo Galilei, for example, proposed that the aurora was caused by air rising out of the earth's shadow to where it could be illuminated by sunlight. He also appears to have coined the term *aurora borealis,* meaning "dawn of the north." Pierre Gassendi, a French mathematician and astronomer, at about the same time, deduced that auroral displays must be occurring at great heights, because they were seen to have the same configuration when observed at places quite remote from one another. His contemporary, René Descartes, seems to have been the originator of the idea that auroras were caused by reflections from ice crystals in the air at high latitudes. From about 1645 to about 1715, both solar activity and auroral sightings declined, although neither were completely lacking.

Edmund Halley, after finally, at the age of 60, having personally observed an auroral display, seems to have been the first to suggest that the auroral phenomenon was ordered by the direction of the earth's magnetic field. In 1731, the French philosopher de Mairan ridiculed the currently popular idea that the aurora was a reflection of polar ice and snow, and he also criticized Halley's theory. He suggested that the aurora was connected to the solar atmosphere, and he suspected a connection between the return of sunspots and the aurora. After that

time, studies of geomagnetism and the aurora became more firmly linked.

1.2 EARLY MEASUREMENTS OF THE GEOMAGNETIC FIELD

The earliest indication of the existence of the geomagnetic field was the direction-finding capability of the compass. As compasses were improved, more and more was learned about the geomagnetic field. The earliest reliable evidence of Chinese knowledge that a compass points north or south dates from the eleventh century. The encyclopedist Shon-Kau (A.D. 1030–93) stated that "fortune-tellers rub the point of a needle with the stone of the magnet in order to make it properly indicate the south." In the European literature, the earliest mention of the compass and its application to navigation appeared in two works by Alexander Neekan, a monk of St. Albans (A.D. 1157–217), entitled *De Untensilibus* and *De Rerum*. In the former he described the use of the magnetic needle to indicate north and noted that mariners used that means to find their course when the sky was cloudy. In the second, he described the needle as being placed on a pivot, a second-generation form of the compass. In neither work did he describe the instrument as a novelty; it was in common use at that time. Official records indicate that by the fourteenth century, many sailing ships carried compasses.

The direction of magnetic north and that of geographic or true north differ over most of the globe. The measure of this difference is called the declination. It is not clear when magnetic declination was actually discovered. However, a letter written by Georg Hartmann, vicar of St. Sebald's at Nürnberg, to Duke Albrecht of Prussia in 1544 showed that he had observed the declination of Rome in 1510 to be 6° east, whereas it was 10° at Nürnberg. Also, it is known that between the years 1538 and 1541, João de Castro made 43 determinations of declination during a voyage along the west coast of India and in the Red Sea.

The geomagnetic field is also inclined to the horizontal. To measure this inclination, one must pivot a needle about a horizontal axis. Georg Hartmann's letter also discussed such an observation, but the angle of inclination was incorrect for his point of observation. William Gilbert ascribed the discovery of the magnetic dip or inclination to an Englishman, Robert Norman, who in 1576 published a work with the title *The newe Attractiue containyng a short discourse of the Magnes or Lodestone, and amongest other his vertues, of a newe discouered secret and subtill propertie, concernyng the Declinyng of the Needle, touched there with onder the plaine of the Horizon. Now first found out by ROBERT NORMAN Hydrographer. Here onto are annexed certaine necessarie rules for the art of Nauigation, by the same R.N. Imprinted at London by John Kyngston, for Richard Ballard, 1581.*

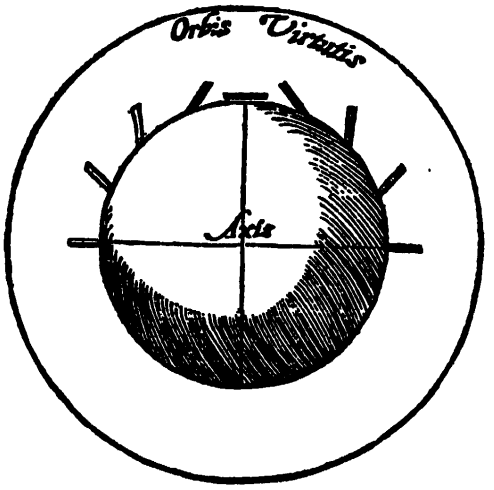


FIG. 1.2. Illustration of magnetic-dipole character of the earth's main magnetic field, as shown in Gilbert's *De Magnete.*

The year 1600 saw the publication of the famous treatise *De Magnete,* by William Gilbert, who in 1601 was appointed chief physician in personal attendance to Queen Elizabeth. This treatise consists of six books containing a total of 115 chapters. The central theme of the book is also the title of Chapter 17, Book 1: "That the globe of the earth is magnetic, a magnet; how in our hands the magnet stone has all the primary forces of the earth, while the earth by the same powers remains constant in a fixed direction in the universe." Figure 1.2 is Gilbert's woodcut showing the distribution of magnetic inclination or dip over the earth, and over a small spherical lodestone, which he called a "terrella." Gilbert believed that the terrestrial magnetic field was constant, but it is not. Henry Gellibrand, professor of astronomy at Gresham College, discovered that magnetic declination changed with time, and he published his discovery in a work entitled *A discourse mathematical on the variation of the magneticall needle. Together with its admirable diminuation lately discovered. London 1635.*

Another early pioneer in the study of geomagnetism was Edmund Halley, who published in 1683 and 1692 two works on the theory of geomagnetism, but needed to test his theory further. King William III put at his disposal the ship *Paramour Pink,* on which Halley made two voyages: in October 1698 to the North Atlantic Ocean, and in September 1700 to the South Atlantic Ocean. Those voyages were the first purely scientific expeditions, and they returned measurements of great value both for practical navigation and for the theory of navigation. Those investigations led to the publication of two geomagnetic charts: "New and Correct Chart showing the Variations of the Compass in the Western and Southern Oceans, as observed in year 1700 by his Majesty's Command by Edm. Halley" and "Sea Chart of the whole world, showing the Variations of the Compass," published in 1701 and 1702, respectively.

1.3 THE EMERGENCE OF A SCIENTIFIC DISCIPLINE

Despite the fact that the sun is the most luminous object we can see, the solar side of solar-terrestrial physics awaited technological change as surely as the study of geomagnetism awaited the development of the compass and its successor the magnetometer. Sunspots, magnetized cool spots in the solar photosphere, are generally too small to be resolved by the naked eye. Thus the study of sunspots did not begin until the invention of the telescope. Galileo Galilei was one of the first to use this new invention to study them. Sunspot studies proceeded slowly, perhaps because very few sunspots occurred during the period called the Maunder minimum, from about 1645 to 1700. The now-familiar 11-yr periodicity in sunspot number illustrated in Figure 1.3 was not discovered until 1851. The sunspot or solar cycle is discussed in greater depth in Chapter 3, which reviews our current understanding of the physics of the sun, in which magnetism plays a significant role that is only gradually being understood.

Perhaps the first discovery in the emerging discipline we now call solar-terrestrial physics was the observation in 1722 by George Graham, a famous London instrument maker, that the compass is always in motion. Graham's discovery was confirmed in 1740 by Anders Celsius in Uppsala, Sweden. His observations were continued by O. Hiorter to a total of over 20,000 observations made on more than 1,000 different days. From those data Hiorter discovered the diurnal variation of the geomagnetic field. Magnetic perturbations vary systematically with local time, which is determined by the longitudinal separation between the meridian of the observer and that containing the sun, which is called the noon meridian. These perturbations are due to the rotation of one's observation station under current systems flowing in the upper atmo-

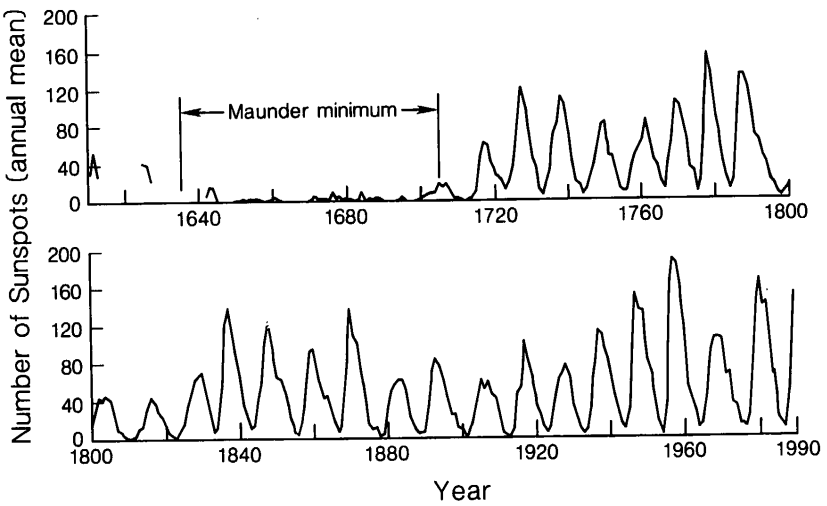


FIG. 1.3. Sunspot cycle since A.D. 1610.

sphere that are fixed with respect to the sun. Perhaps even more important, on April 5, 1741, Hiorter discovered that geomagnetic and auroral activities were correlated. Simultaneous observations in London by Graham confirmed the occurrence of strong geomagnetic activity on that day. In 1770, J. C. Wilcke noted that auroral rays extend upward along the direction of the magnetic field. That was the same year that Captain James Cook first reported the southern counterpart of the aurora borealis, the aurora australis, or "dawn of the south." Twenty years later, the English scientist Henry Cavendish used triangulation to estimate the height of auroras as between 52 and 71 miles. Earlier attempts at triangulation by Halley and Mairan had been much less accurate.

The great advance of the early nineteenth century was the development of a network to make frequent simultaneous observations with widely spaced magnetometers. C. F. Gauss was one of the leaders of that effort and one of the foremost pioneers in the mathematical analysis of the resulting measurements, which allowed contributions to the geomagnetic field from below the surface of the earth to be separated from those contributions arising high in the atmosphere. Meanwhile, Heinrich Schwabe, on the basis of his sunspot measurements taken between 1825 and 1850, deduced that the variation in the number of sunspots was periodic, with a period of about 10 yr. Magnetic observatories had spread to the British colonies by 1839. Edward Sabine was assigned to supervise four of those observatories (Toronto, St. Helena, Cape of Good Hope, and Hobarton). Using the data from those observatories, he was able to show in 1851 that the intensity of geomagnetic disturbances varied in concert with the sunspot cycle. Chapter 13 discusses our modern understanding of these disturbances.

The next discovery to provide a link between the sun and geomagnetic activity was Richard Carrington's sighting of a great flare of white light on the sun, on September 1, 1859. Carrington, who was sketching sunspot groups at the time, was startled by the flare, and by the time he was able to summon someone to witness the event a minute later, he was dismayed to find that it had weakened greatly in intensity. Fortunately, it had been simultaneously noted by another observer some miles away. Furthermore, at the moment of the flare, the Kew Observatory (London) measurements of the magnetic field had been disturbed. Today we realize that that disturbance of the magnetic field was caused by an increase in the electric currents flowing overhead in the earth's ionosphere. Such currents flow in response to electric fields in the ionosphere. The extreme radiation by ultraviolet rays and x-rays from the flare increased the ionization and hence the electrical conductivity of the ionosphere, causing more current to flow in response to the unaltered electric field. Finally, 18 h later, one of the strongest magnetic storms ever recorded broke out. Auroras were seen as far south as Puerto Rico.

To have arrived that quickly, the disturbance would have had to

travel from the sun at over 2,300 km·s^{-1}. As discussed in Chapter 4, we know today that the sun and the earth are linked by the supersonic solar wind, but 2,300 km · s^{-1} is a high velocity even for the disturbed solar wind. When such disturbances arrive at the earth, the terrestrial field jumps quite abruptly, indicating that the discontinuity in the interplanetary medium that is flowing by the earth is quite thin. The thinness of these disturbance fronts strongly suggests that they are caused by shocks in the interplanetary medium, despite the collisionless nature of the gas there. Usually, collisions are needed to account for the dissipation and heating that occur at a shock. That was the first indication of the existence of collisionless shocks, which since the advent of interplanetary exploration have been found to be ubiquitous in the solar system, as discussed in Chapter 5. Shortly after those observations, in 1861, Balfour Stewart noted the occurrence of pulsations in the earth's magnetic field, with periods of minutes. We now know that the magnetosphere pulsates at a wide variety of periods. These pulsations are discussed in further detail in Chapter 11.

The nineteenth century also brought another simple but important observation about the aurora. Captain John Franklin, the ill-fated English Arctic explorer whose party perished in 1845 attempting to discover the Northwest Passage, noted that auroral frequency did not increase all the way to the pole, according to observations made during his 1819–22 journeys. In 1860, Elias Loomis of Yale was one of the first to plot the zone of maximum auroral occurrence, which roughly corresponds to what today we call the auroral zone. The auroral zone is an oval band around the magnetic pole, roughly 20–25° from the pole.

Precursors to our modern understanding of the aurora began to appear in the late nineteenth century. About 1878, H. Becquerel suggested that particles were shot off from the sun and were guided by the earth's magnetic field to the auroral zone. He believed that sunspots ejected protons. A similar theory was espoused by E. Goldstein. In 1897, the great Norwegian physicist Kristian Birkeland made his first auroral expedition to northern Norway. However, it was not until after his third expedition in 1902–3, during which he obtained extensive data on the magnetic perturbations associated with auroras, that he concluded that large electric currents flowed along magnetic-field lines during aurora. The invention of the vacuum tube led to the understanding that the aurora was in some way similar to the cathode rays in those devices. Soon Sir William Crookes demonstrated that cathode rays were bent by magnetic fields, and shortly thereafter J. J. Thomson showed that cathode rays consisted of the tiny, negatively charged particles we now call electrons. Birkeland adopted those ideas for his auroral theories and attempted to verify his theories with both field observations and laboratory experiments. Specifically, he conducted experiments with a magnetic dipole inside a model earth, which he called a terrella. Figure 1.4 shows Birkeland in his laboratory beside his terrella experiment. Those

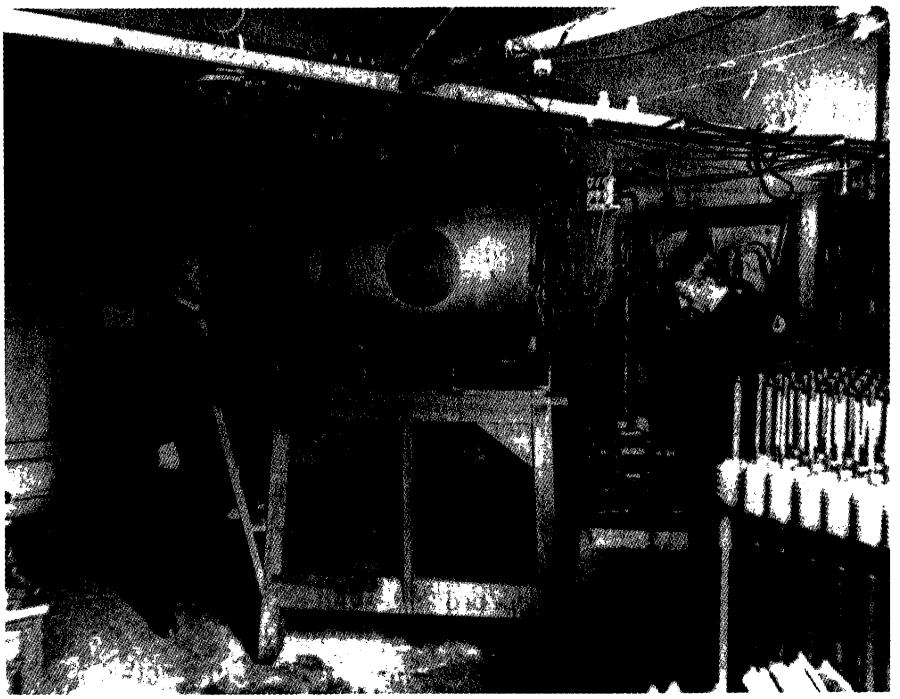


FIG. 1.4. Kristian Birkeland (left) in his laboratory with his terrella and with his assistant, O. Devik (right), about 1909. (Photo courtesy of A. Egeland.)

experiments showed that electrons incident on the terrella would produce patterns quite reminiscent of the auroral zone. He believed, as we do today, that those particles came from the sun.

K. Birkeland's work inspired the Norwegian mathematician Carl Størmer, whose subsequent calculations of the motion of charged particles in a dipole magnetic field in turn supported Birkeland. Figure 1.5 shows Størmer and his assistant Olaf (not Kristian) Birkeland. As is evident from this photograph, the advent of the camera was an important advance in the study of the aurora. It was through measurements such as these that Størmer accurately determined the height of the aurora. Figure 1.6 illustrates one of Størmer's charged-particle orbit calculations in a forbidden region to which charged particles from the sun would not have direct access. In such a region, charged particles would spiral around the magnetic field and bounce back and forth along it, reflected by the converging magnetic-field geometry. Størmer's contributions became much more relevant and appreciated after the discovery of the earth's radiation belts, whose particle motions resemble those of Figure 1.6. Birkeland's work was not appreciated until even later. A more detailed discussion of the trajectories of charged particles in the earth's magnetic field can be found in Chapter 10, and more about the aurora can be found in Chapter 14.

All that work proceeded despite Lord Kelvin's 1882 argument that he had provided absolutely conclusive evidence against the supposition that terrestrial magnetic storms were due to magnetic action in the sun

FIG. 1.5. Auroral physicists C. F. Størmer, standing, and Olaf Birkeland, seated, in northern Norway, ca. 1910. (Photo courtesy of A. Egeland.)

or to any kind of dynamic action taking place within the sun. Lord Kelvin also claimed "that the supposed connection between magnetic storms and sunspots is unreal, and the seeming agreement between periods has been a mere coincidence." More telling was the criticism of A. Schuster that a beam of electrons from the sun could not hold together against their mutual electrostatic repulsion.

1.4 THE IONOSPHERE AND MAGNETOSPHERE

The electrically conducting region above about 100 km altitude that we now call the ionosphere may rightly be claimed to have been discovered by Balfour Stewart. In his 1882 *Encyclopaedia Britannica* article entitled "Terrestrial Magnetism" he concluded that the upper atmosphere was the most probable location of the electric currents that produce the solar-controlled variation in the magnetic field measured at the surface of the earth. He noted that "we know from our study of aurora that there are such currents in these regions – continuous near the poles and occasional in lower latitudes." He proposed that the primary causes of the daily variations in the intensity of the surface magnetic field were "convective currents established by the sun's heating influence in the upper regions of the atmosphere." These currents "are to be regarded as conductors moving across lines of magnetic force and are thus the vehicle of electric currents which act upon the magnetic field." Those statements are very close to modern atmospheric-dynamo theory. However, it was left to A. Schuster to put the dynamo theory into quantitative form.

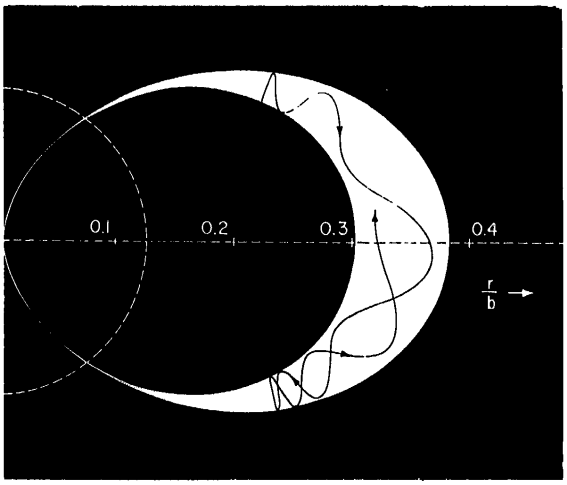


FIG. 1.6. Trajectory of an energetic charged particle in the "forbidden" zone of a dipole magnetic field, as drawn by Størmer. (From Rossi and Olbert, 1970.)

The turn of the century brought another new invention that was used to probe the solar-terrestrial environment: the radio transmitter and receiver. In 1902, A. E. Kennelly and O. Heaviside independently postulated the existence of a highly electrically conducting ionosphere to explain G. Marconi's transatlantic radio transmissions. Verification of the existence of the ionosphere did not come until much later, in 1925, when E. V. Appleton and M. A. F. Barnett in the United Kingdom, and shortly thereafter G. Breit and M. A. Tuve in America, established the existence and altitude of the Kennelly-Heaviside layer, as it was known then. The original method of Breit and Tuve, using short pulses of radio energy at vertical incidence and timing the arrival of a reflected signal in order to infer the altitude of the electrically reflecting layer, is still used today for sounding the ionosphere. In drawing diagrams of the electromagnetic waves reflected by the ionosphere, Appleton used the letter E for the electric vector of the downcoming wave. When he found reflections from a higher layer, he used the letter F for the electric vector of those reflected waves, and when he occasionally got reflections from a lower layer, he naturally used the letter D. When it came time to name these layers, he chose the same letters, leaving the letters A, B, and C for possible later discoveries that never came. So now the ionospheric layers are called the D, E, and F layers, as illustrated in Figure 1.7. We now know that all planets with atmospheres have electrically conducting ionospheres like that of the earth. Chapter 7 discusses how these are formed.

At about that same time, progress was being made in understanding the auroral glows. Spectroscopy, together with photography, permitted first the determination of the wavelength and then the identity of the excited molecule that was radiating. There were initial successes, beginning with Lars Vegard's work in Norway relating auroral emissions to emission bands from known atmospheric gases such as nitrogen. How-

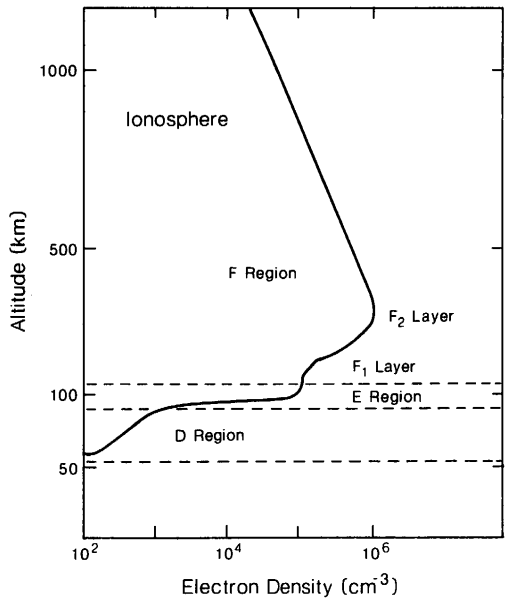


FIG. 1.7. Electron density of the earth's ionosphere as a function of altitude.

ever, identification of the yellow-green line at 557.7 nm was elusive. Finally, H. Babcock's precise measurements in 1923 allowed John McLennan to identify it as a metastable transition of atomic oxygen. At atmospheric pressures close to that at the surface of the earth, collisions between molecules de-excite the molecules before they have a chance to radiate if they happen to become excited into one of the metastable states. However, at the altitude of the aurora, collisions are so rare that the time between collisions is longer than the lifetimes of the metastable states, and the excitation energy of the state can be released by radiation. A similar line in the auroral spectrum is the 630.0-nm red line of atomic oxygen. This metastable transition has a lifetime of 110 s and can radiate only above some 250 km. Those discoveries led to the realization that the varied colors of the aurora were simply related to height. In low-altitude auroras, below 100 km, where collisions quench even the oxygen green line, the blue and red nitrogen bands predominate. From 100 to 250 km, the oxygen green line is strongest. Above 250 km, the red line is most important.

Although most of the auroral forms are associated with electrons, some aurora are due to precipitating protons. The first observations of the proton aurora were made in 1939. Measurements of the Doppler shifts of the proton emissions permitted estimates of the energy of the precipitating particles from the ground. Chapter 14 contains more detailed discussion of the aurora and auroral ionosphere.

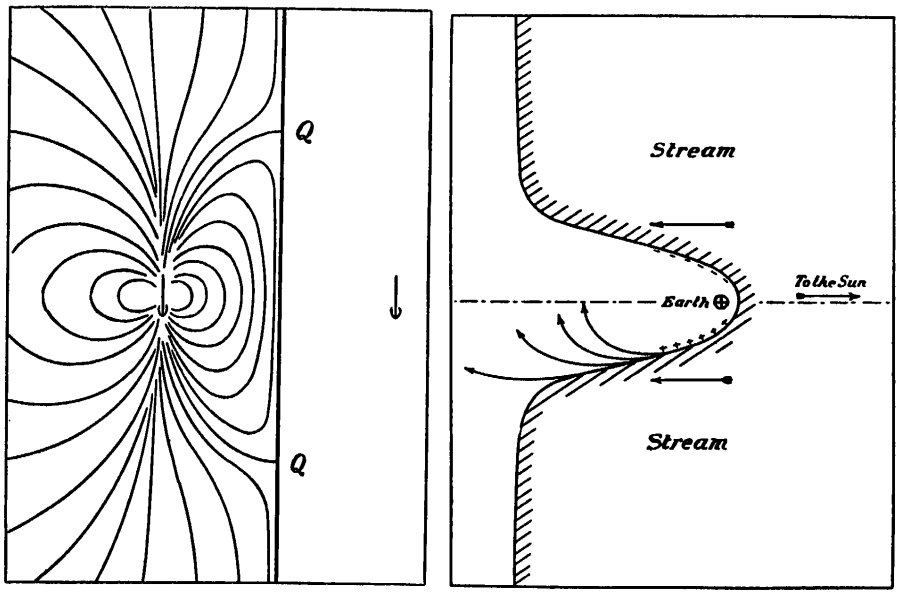
With the concept of the ionosphere firmly established, scientists began to wonder about the upper extension of the ionosphere, linked magnetically to the earth, which region we today call the magnetosphere. In 1918, Sydney Chapman postulated a singly-charged beam

from the sun as the cause of worldwide magnetic disturbances. That was a revival of an old idea that had previously been criticized by Schuster. Chapman was soon challenged by Frederick Lindemann, who pointed out that mutual electrostatic repulsion would destroy such a stream. Lindemann, instead, suggested that the stream of charged particles contained particles of both signs in equal numbers. We would now call such a stream a "plasma." That proposal was a breakthrough, and it permitted Chapman and his co-workers, in a series of papers beginning in 1930, to lay the foundations for our modern understanding of the interaction of the solar wind with the magnetosphere.

In the rarefied conditions of outer space, where collisions between particles are infrequent, the ion-electron gas, or plasma, is highly electrically conducting. Thus, Chapman and Ferraro proposed that as the plasma from the sun approached the earth, the earth would effectively see a mirror magnetic-dipole moment advancing on the earth, as illustrated in Figure 1.8. The net result of that advancing mirror field would be to compress the terrestrial field. Eventually, as sketched in Figure 1.9, the plasma would surround the earth on all sides, and a cavity would be carved out of the solar plasma by the terrestrial magnetic field. That is very similar to our modern concept of the geomagnetic cavity, which is discussed in greater detail in Chapter 6.

After the compression of the magnetosphere, which is detected by ground-based magnetometers as a sharp increase in the magnetic field, the magnetosphere becomes inflated. Chapman and Ferraro correctly interpreted that subsequent decrease in the magnetic field at the surface of the earth as the appearance of energetic plasma deep inside the magnetosphere, forming a ring of current around the earth in the near-equatorial regions. The development of this ring current in what we now call a geomagnetic storm is discussed at greater length in Chapters 10 and 13.

At the same time that the ionosphere was being discovered by virtue of its effects on man-made radio signals, natural radio emissions were also being explored, and the magneto-ionic theory developed for the man-made signals was being applied to those natural emissions. The first report of those electromagnetic signals in the audio-frequency range was an observation of what have become known as "whistlers," coming from a 22-km telephone line in Austria in 1886. Whistlers are short bursts of audio-frequency radio noise of continuously decreasing pitch. In 1894, British telephone operators heard "tweeks," possibly whistlers generated by lightning, and a "dawn chorus" generated deep in the magnetosphere during a display of aurora borealis. Little work was done on those observations because of the lack of suitable analysis equipment at the time. During World War I, equipment installed to eavesdrop on enemy telephone conversations picked up whistling sounds. Soldiers at the front would say, "You could hear the grenades fly." H. Barkhausen reported on those observations in 1919 and suggested that they were



(*Left*) **FIG. 1.8.** Compression of a dipole field by an advancing infinite, superconducting slab. The magnetic field is due to the original dipole plus an image dipole an equal distance behind the front, as shown by the right-hand arrow. (From Chapman and Bartels, 1940.)

(*Right*) **FIG. 1.9.** Expected evolution of the front of superconducting plasma as it passes the earth. This model was proposed by Chapman and Ferraro in the 1930s to explain the phenomena of the geomagnetic storm. (From Chapman and Bartels, 1940.)

correlated with meteorological influences. However, he could not duplicate the phenomenon in laboratory experiments. In 1925, T. L. Eckersley also described that phenomenon and ascribed it to the dispersion of an electrical impulse in a medium loaded with free ions. Eventually, after much work and several incorrect explanations, in 1935 Eckersley concluded that the distinctive swooping sound of whistlers was due to the dispersion of a burst of electromagnetic noise traveling through the ionosphere. Very little work was done on whistlers until the early 1950s, at which time L. R. O. Storey, with a homemade spectrum analyzer, conducted a thorough study of whistlers. He found that whistlers are caused by lightning flashes, whose electromagnetic energy then echoes back and forth along field lines in the upper ionosphere, as illustrated in Figure 1.10. A major implication of those findings was that the electron density in the outer ionosphere, which is now called the plasmasphere, was unexpectedly high. Storey also found other types of audio-frequency, or very low frequency (VLF), emissions that are not associated with lightning and are now known to be generated within the magnetospheric plasma. Chapter 12 discusses the generation and propagation of these waves.

1.5 THE SOLAR WIND

If the auroras were caused by electrons, and if those electrons came from the sun, as was commonly believed among solar-terrestrial researchers in the first half of the twentieth century, then those electrons would have to travel in the company of an equal number of ions, or else the beam would disrupt. That idea can be considered the first model of

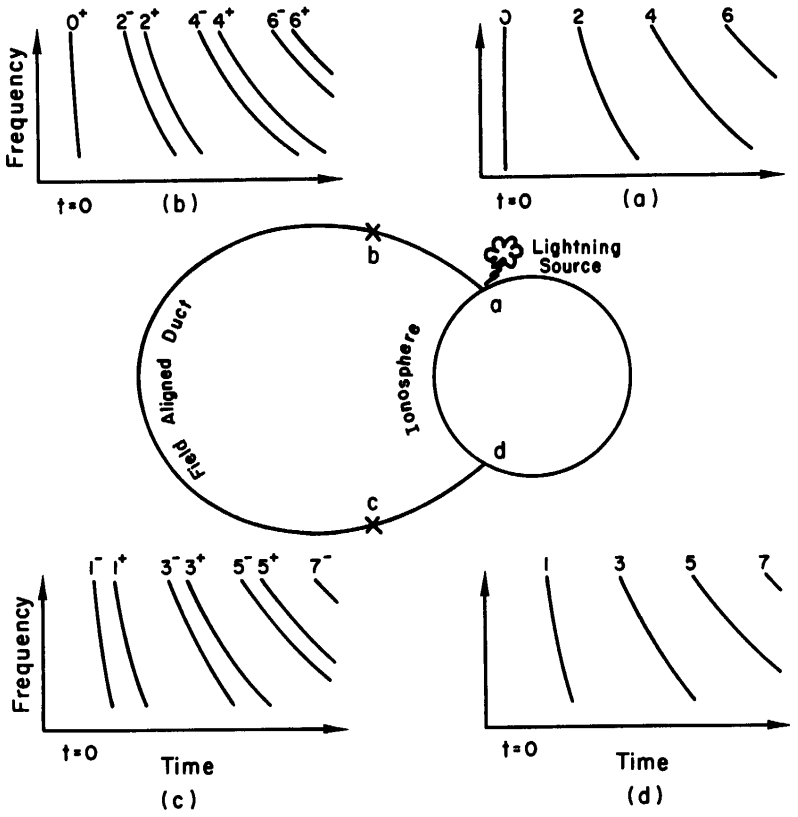


FIG. 1.10. Dispersion of whistler-mode waves generated by lightning, as seen at four different locations. The different velocities of propagation as functions of wave frequency (dispersion) cause the wave arrival to be delayed by a different amount at each frequency. The delay depends on the distance traveled and the properties of the plasma traversed by the wave. (From Russell, 1972.)

the streaming ionized plasma of interplanetary space that we call the solar wind. It was an essential element of the geomagnetic-storm model of Chapman and Ferraro, but in their model the solar wind was intermittent. It flowed only at active times. However, in 1943, C. Hoffmeister noted that a comet tail was not strictly radial, but lagged behind the comet's radial direction by about 5°. In 1951, L. Biermann correctly interpreted that lag in terms of an interaction between the comet tail and a solar wind. That wind was said to flow at about 450 km $\cdot$ s^{-1} at all times and in all directions from the sun, although he assumed that the electron density was about 600 cm^{-3}, two orders of magnitude too high. Several years later, in 1957, Hannes Alfvén postulated that the solar wind was magnetized and that the solar-wind flow draped that magnetic field over the comet, forming a long magnetic tail downstream in the antisolar direction, as illustrated in Figure 1.11. The cometary ions were confined by that tail in a narrow ribbon between the two tail "lobes." In 1958, E. W. Parker provided the theoretical underpinning for such a flow of magnetized plasma, and in 1962 he showed that in order to be consistent with the geomagnetic records, the electron density of the solar wind should seldom exceed 30 cm^{-3}. Confirmation was not long in coming. That was the dawn of the space age, and soon observations were being returned by both Soviet and American space probes that

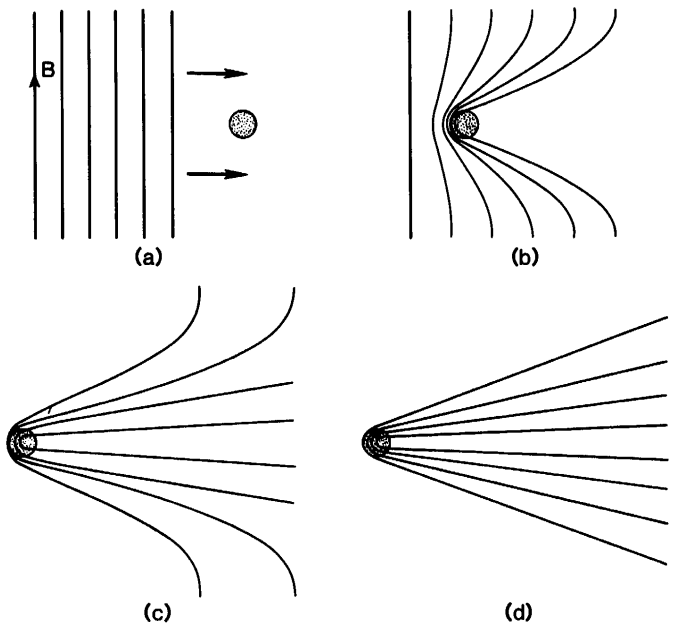


FIG. 1.11. Original model of the formation of a type I (plasma) cometary tail, according to H. Alfvén. In this model the solar-wind magnetic field is draped over the comet by the motion of the plasma from left to right. (From Alfvén, 1957.)

clearly confirmed the existence of the solar wind and its entrained magnetic field, measured its properties, and demonstrated its pivotal role in controlling geomagnetic activity and the aurora. Chapter 4 discusses the solar wind in greater detail.

1.6 MAGNETOSPHERIC EXPLORATION

Rockets provided the opportunity to begin to explore the magnetosphere. In the early and middle 1950s, James Van Allen and his colleagues launched a series of rocket flights into the Arctic and Antarctic ionosphere, reaching heights up to 110 km. Those flights detected either energetic electrons or the bremsstrahlung radiation from such electrons. The year 1957 marked the beginning of an International Geophysical Year (IGY), an 18-month period of worldwide geophysical studies. It also marked the launch of *Sputnik 1*. The attendant space race began a period of explosive growth in our knowledge of the terrestrial magnetosphere and its interaction with the solar wind. In 1958, *Explorer 1* carried a Geiger counter that enabled Van Allen to discover the trapped radiation belts. Instrumentation developed by Konstantin I. Gringauz for the Soviet *Luna* probes provided the first measurements of the solar wind, and instrumentation developed by Conway Snyder and Marcia Neugebauer for *Mariner 2* in 1962 provided the first detailed study of this plasma.

Battery-powered *Explorer 10,* launched in 1961, was the first spacecraft to provide measurements across the magnetopause, the boundary

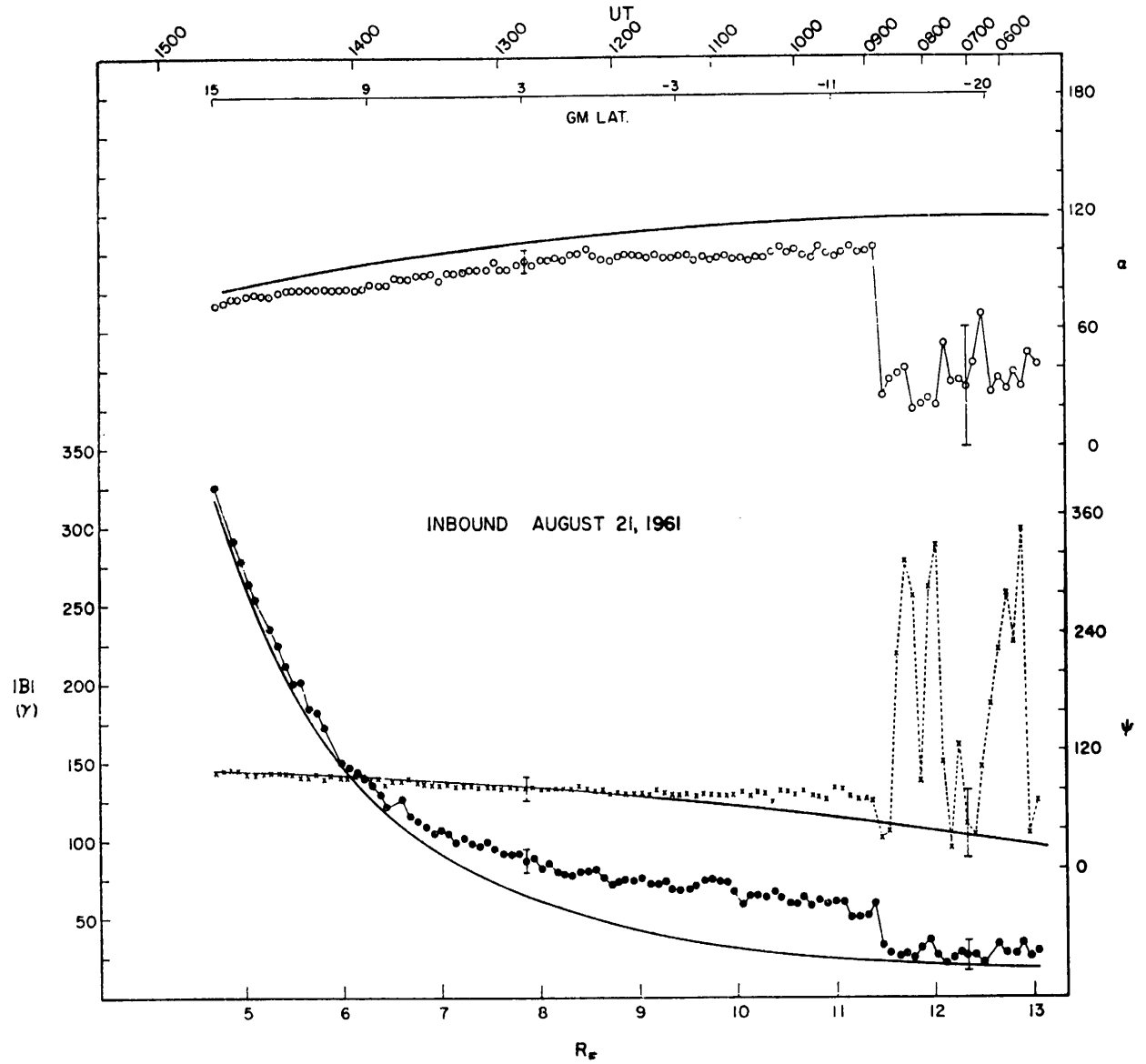


FIG. 1.12. Measurements of magnetic-field strength in the outer magnetosphere and through the magnetopause by the *Explorer 12* spacecraft. Two angles and the field magnitude are shown. Smooth lines are dipole values. (From Cahill and Patel, 1967.)

between the flowing solar wind and the earth's magnetic field, but the first detailed examination of that boundary awaited measurements with a solar-powered spacecraft, *Explorer 12,* which provided 4 months of data and coverage from the noon meridian to the dawn meridian. Figure 1.12 shows magnetic measurements obtained by *Explorer 12* during a traversal of the outer magnetosphere, through the magnetopause, and out into the magnetosheath. It was clear from the data provided by the many spacecraft that were launched into the solar wind during those

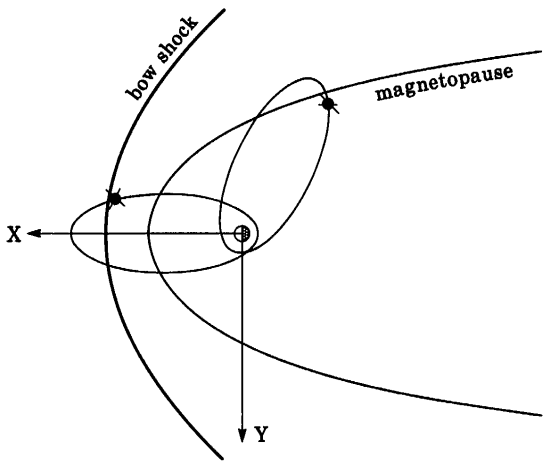


FIG. 1.13. Mapping of the magnetopause and bow shock as the earth goes around the sun. During the course of a year, the magnetopause and bow-shock surfaces maintain their orientation about the earth–sun line, while the spacecraft orbit is fixed in inertial space. Thus the orbits appear to sweep through these boundaries in the course of a year.

early years that the solar wind undergoes an abrupt transition prior to reaching the magnetopause. This transition is a shock produced as the high-speed solar wind encounters the earth as an obstacle in its flow. The existence of a shock in a collisionless plasma surprised many physicists. In the intervening years it has become clear that the electric and magnetic fields in the plasma can alter the motion of the particles in a manner similar to ordinary collisions. These changes provide the dissipation needed to form a shock. The physics of this process is discussed in Chapter 5. The shock allows the solar wind, which flows faster than the speed of compressional waves in a plasma, to be slowed, heated, and deflected around the planet.

It was not until the launch of the first Orbiting Geophysical Observatory (OGO) in 1964 that scientists obtained time-resolution measurements of sufficient accuracy to study the bow shock. The *OGO 1, 3,* and *5* spacecraft in highly eccentric orbits mapped the locations of both boundaries as the earth orbited the sun and the orbits precessed relative to the magnetosphere, as shown in Figure 1.13. Those measurements and data from other spacecraft launched in the 1960s, such as the Interplanetary Monitoring Platform (IMP) spacecraft and the VELA spacecraft, revealed that the structure of the bow shock was very sensitive to the conditions in the plasma, the ratio of the speed of the solar wind to the speed of compressional waves (Mach number), and the ratio of the thermal pressure to the magnetic pressure (beta). It was also found to be sensitive to the direction of the interplanetary magnetic field. When the magnetic field is almost aligned with the direction of propagation of the shock, the shock normal, the shock is said to be quasi-parallel. When the field is more nearly perpendicular to the normal, it is referred to as being quasi-perpendicular. Ion beams are found upstream of the quasi-parallel shock, as illustrated in Figure 1.14, and these ion beams interact with the incoming solar-wind plasma to produce copious large-amplitude waves, called upstream waves, shown in Figure 1.15.

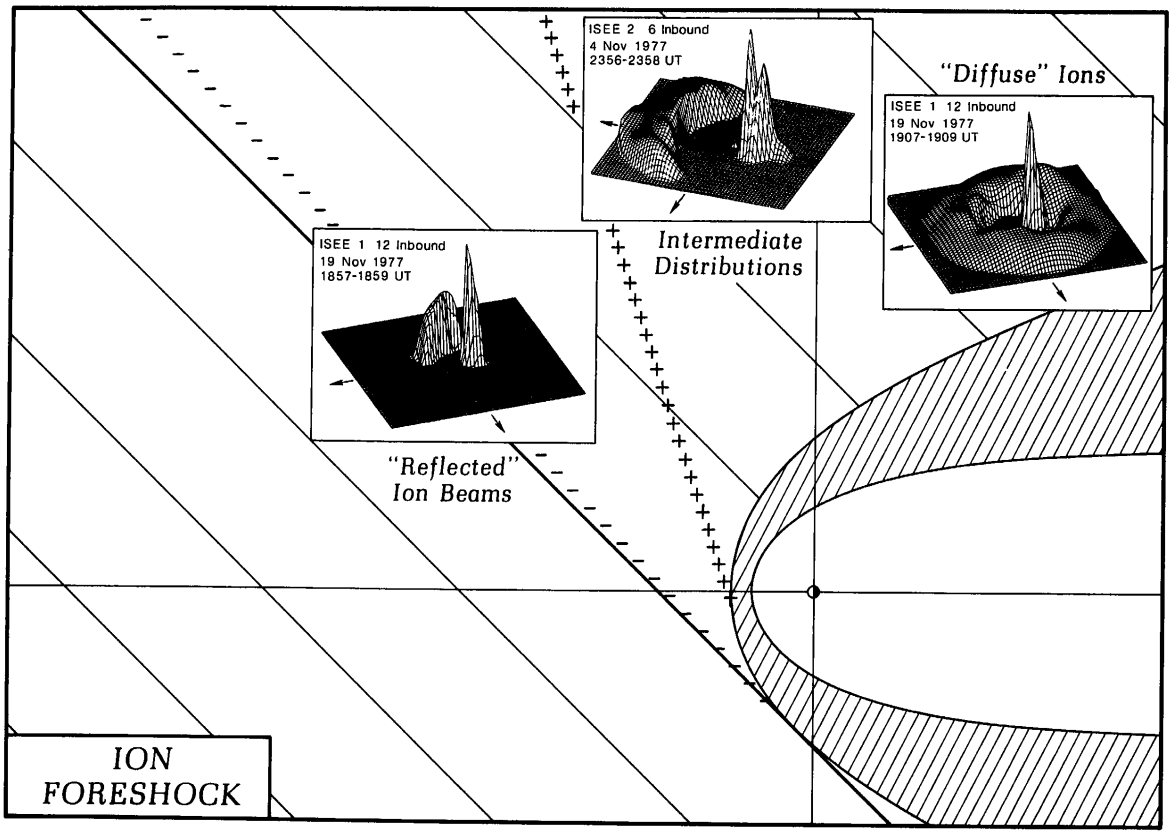


FIG. 1.14. Ion foreshock. On field lines that touch the bow shock, charged particles can spiral upstream along the magnetic field. The solar-wind electric field causes these particles to drift antisunward. The fastest particles (electrons) are affected the least, and the slowest particles the most, by this drift. Representative distribution functions of ions are shown for various locations. The narrow peak represents the unperturbed solar-wind beam. The broader distributions represent the back-streaming ions. (From Russell and Hoppe, 1983.)

The shock is important because it modifies the properties of the solar-wind flow before the flow interacts with the earth's magnetic field, but the processes acting at the magnetopause are the ones finally responsible for determining how much energy the magnetosphere receives from the solar-wind flow. One can imagine a very inviscid interaction in which the solar wind is completely diverted by the magnetosphere and there is very little drag and hence little momentum transfer across the boundary. In fact, this situation does occur when the interplanetary magnetic field is northward, but when the interplanetary field is southward, the momentum transfer from the solar wind increases markedly.

The most important clue that the nature of the processes at the magnetopause changes with variations in the properties of the solar wind is the observation that geomagnetic activity is controlled by the north–south component of the interplanetary magnetic field. The availability of

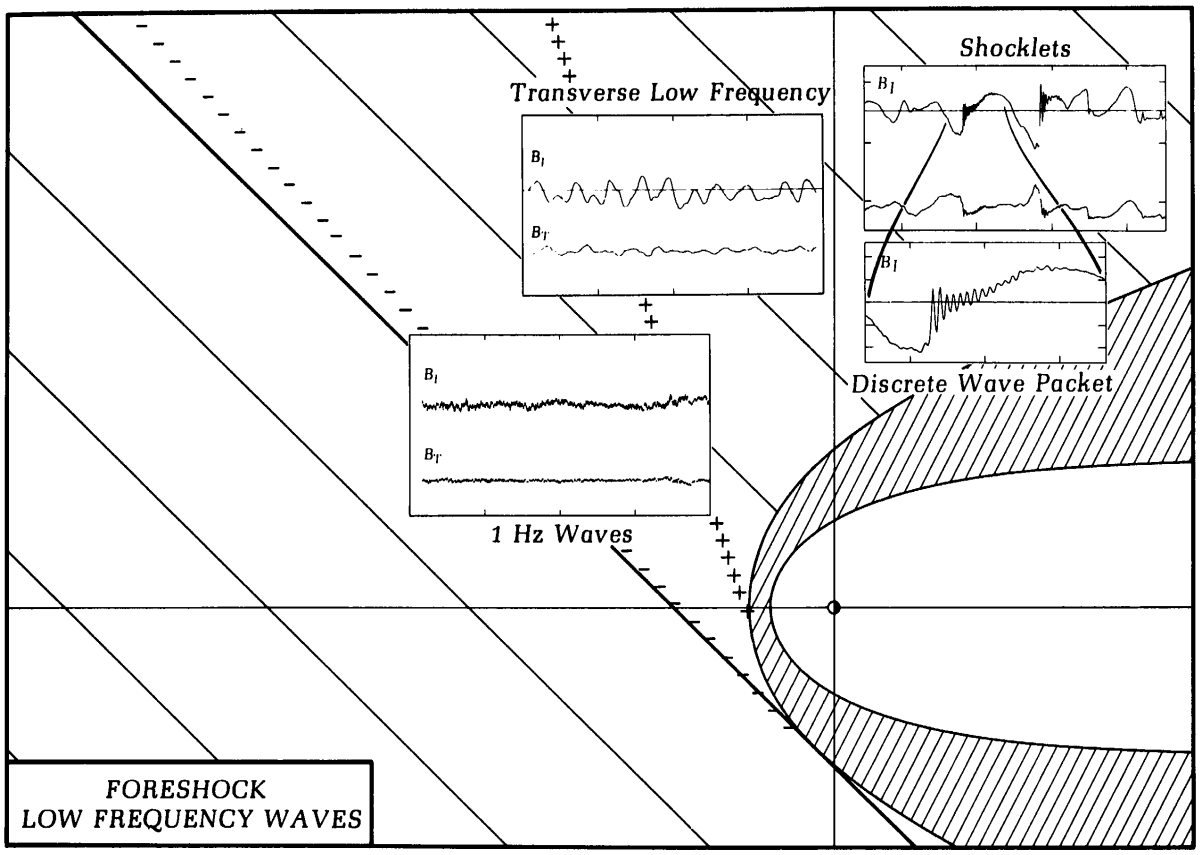


FIG. 1.15. Back-streaming ions and electrons can stimulate a variety of low-frequency waves (with periods of seconds to many tens of seconds). Waves representative of different regions of the near-earth solar wind are illustrated here. (From Russell and Hoppe, 1983.)

extended measurements in the solar wind, especially from *Explorer 33* and *35*, allowed researchers such as Roger Arnoldy and Joan Hirshberg to study this control. The mechanism by which the interplanetary magnetic field exerts this control is known as "reconnection," which was postulated by James Dungey in 1961. As illustrated in Figure 1.16, interplanetary and planetary magnetic fields become linked. As a result, magnetic flux is transported from the dayside of the magnetosphere to the nightside. This magnetic flux builds up in the tail until reconnection occurs there too and returns the magnetic flux to the magnetosphere proper. Spacecraft such as *OGO 5*, launched in 1968, showed the erosion of the dayside magnetosphere and the corresponding activity in the magnetotail. This process, which also leads to activation of the aurora, is called a "substorm." As implied by the name substorm, often there are occasions of more major activity covering the entire magnetosphere, which are called geomagnetic storms. The most popular model of substorms is called the near-earth neutral-point model, where the neutral

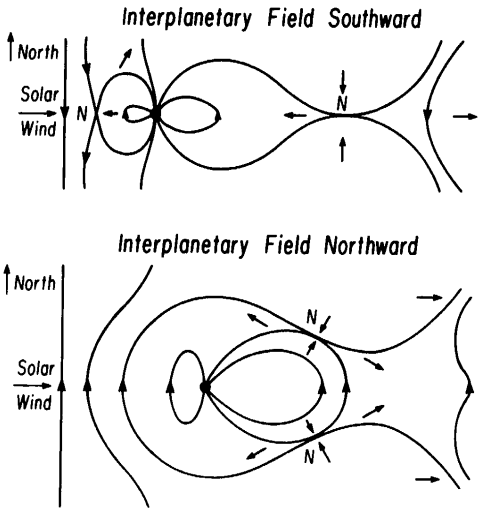


FIG. 1.16. Topology of the magnetosphere for northward and southward interplanetary fields, according to J. W. Dungey in the early 1960s. In the steady state, the plasma flows as indicated by the short arrows. (From Dungey, 1963a.)

point, a location where the magnetic field vanishes, is the site of the reconnection. It was not until the late 1970s, after the launch of the dual co-orbiting satellites *ISEE 1* and *2*, that the reconnection mechanism gained general acceptance. Those satellites returned plasma data of sufficiently high resolution to show the accelerated flows at the magnetopause and in the magnetotail. Chapter 9 provides further details about these processes. However, even today there is debate as to where reconnection occurs and how important it is relative to other processes. Furthermore, it has been found that three-dimensional structures are present, and their explication will require not just two spacecraft, but a whole cluster. Thus there is much to be done in future magnetospheric exploration.

Laboratory plasma measurements have also been useful in understanding the magnetosphere. Figure 1.17 shows a wire model of the magnetosphere developed by Igor Podgorny and his colleagues at the Space Research Institute in Moscow, based on laboratory experiments undertaken in the 1960s, illustrating the development of cusp-shaped openings in the field pattern on the dayside and a long tail at night. Figure 1.18 shows a three-dimensional sketch of the magnetosphere, representing the structure that has been inferred from spacecraft observations. It is deep within this magnetosphere that the radiation-belt particles bounce and drift, as illustrated in Figure 1.19. As noted earlier, the converging magnetic-field lines of the dipole magnetic field stop the forward motion of the particles and accelerate them back toward the equatorial regions. While gyrating and bouncing, these particles also drift, because their gyroradii are greater when they gyrate into weaker fields than when they are in the stronger fields on the inner part of their trajectories. The dipole magnetic field of the earth can confine particles over a wide range of energies, the more energetic of which Van Allen encountered when he and his colleagues discovered the radiation belts.

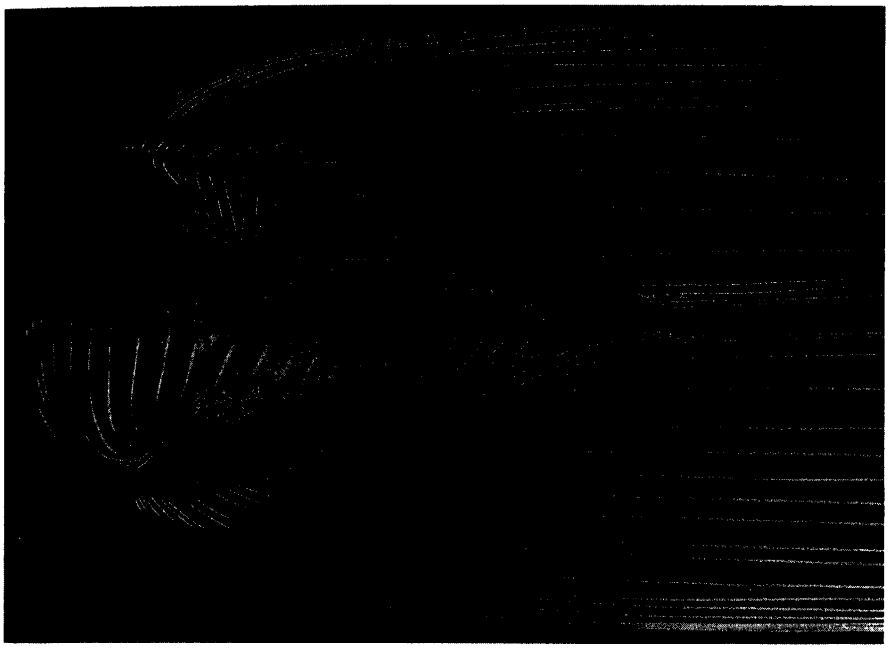


FIG. 1.17. Three-dimensional wire model of the magnetosphere based on the laboratory experimental data of Podgorny and colleagues. (From Podgorny, 1976.)

Figure 1.20 shows the intensities of very energetic protons and electrons in the inner magnetosphere. These particles enter the radiation belts through a variety of means, including radial diffusion from more distant regions, with accompanying acceleration and the decay of neutrons from the sputtering of the atmosphere by cosmic rays. Whereas energetic protons form a single belt, as illustrated in the top panel of Figure 1.20, electrons, as illustrated in the bottom panel, form two belts separated by a region called the slot. This slot in the flux of electrons of fixed energy is formed when naturally occurring electromagnetic waves interact with the gyromotion of the electrons, causing them to spiral into the atmosphere, where they are lost through collisions. The inner electron belt is quite stable and the outer belt quite variable. The radiation belts and the motions of charged particles are discussed further in Chapter 10.

1.7 PLANETARY AND INTERPLANETARY EXPLORATION

The earth is but one test bed for the physical processes occurring in space plasmas. There is much occurring at the other planets and in the interplanetary plasma itself. The solar-wind properties evolve with heliocentric distance, and on the way through the solar system the solar wind encounters a variety of obstacles to its flow, including magnetized bodies, unmagnetized bodies, some with atmospheres and others without, and eventually the heliopause, where the solar wind and the interstellar wind meet.

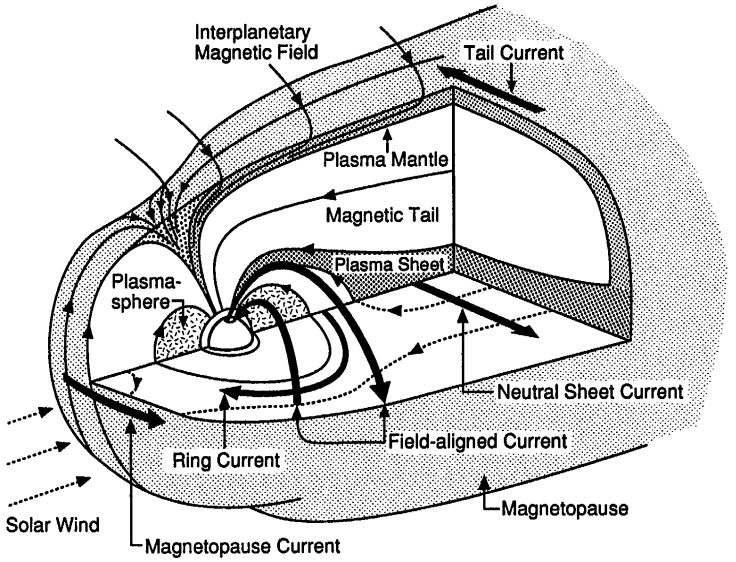


FIG. 1.18. Three-dimensional cutaway view of the magnetosphere showing currents, fields, and plasma regions.

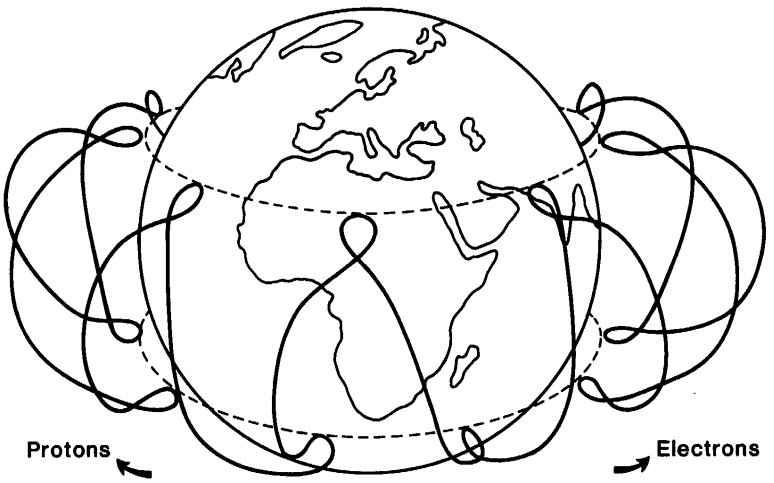


FIG. 1.19. Longitudinal drift of energetic charged particles in the earth's dipolar magnetic field. Electrons drift eastward in the direction of the earth's rotation, and protons westward.

The earliest deep-space probes were the *Mariner 2, 4,* and *5* spacecraft that in the early 1960s went to Venus, Mars, and back to Venus again. Those missions showed that both Venus and Mars were quite different from the earth in their interactions with the solar wind, because neither planet had any significant magnetic moment. However, it was not until after the *Venera 9* and *10* orbiters in 1975, the *Pioneer Venus* orbiter in 1978, and the *Phobos* mission to Mars in 1989 that the details of the solar-wind interactions with these planets became fully understood. At these two planets, the extreme ultraviolet radiation from the sun ionizes the upper atmosphere. It also creates a hot neutral atmosphere that extends into the solar wind. As shown in Figure 1.21, the ionospheric pressure, consisting of both thermal and magnetic components, balances the dynamic pressure of the solar-wind flow. The neutral

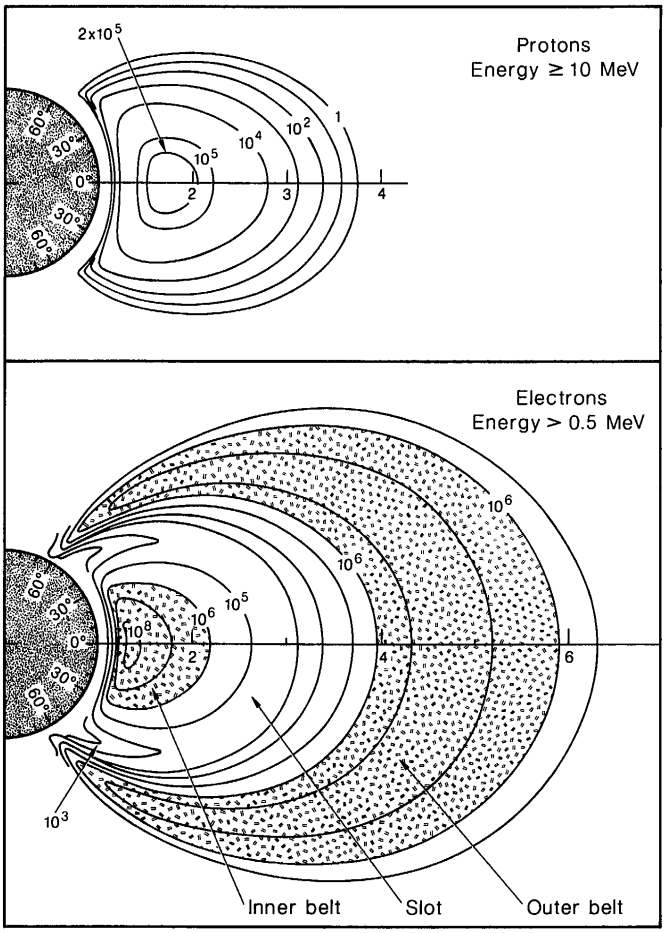


FIG. 1.20. Earth's radiation belts. The top panel shows the contours of the omnidirectional flux (particles per square centimeter per second) of protons with energies greater than 10 MeV. The bottom panel shows the contours of the omnidirectional flux of electrons with energies greater than 0.5 MeV.

atmosphere that extends into the solar wind becomes ionized and adds to the solar-wind flow, further decelerating it.

The slowing down of the magnetized flow around the planet leads to the draping of magnetic-field lines over the obstacle and the formation of a long tail. In this respect, the solar-wind interaction with Venus and Mars resembles that with a comet. This interaction was probed by the International Cometary Explorer (ICE) spacecraft at comet Giacobini-Zinner in 1985 and by the *VEGA 1* and *2, Giotto, Suisei,* and *Sakigake* spacecraft at the comet Halley in 1986. Chapter 8 describes in greater detail the solar-wind interaction with such unmagnetized bodies.

Mariner 10, with a gravitational assist from Venus, made three passes by Mercury in 1974 and 1975. As illustrated in Figure 1.22, *Mariner 10* found a minimagnetosphere very much like that of the earth. Mercury, however, has almost no atmosphere, so that the ionospheric current systems that we believe to be so important for the terrestrial magnetosphere must be absent from Mercury. Thus we expect Mercury's magnetosphere to be quite different in some respects from that of the earth. There has been little investigation of even the meager data from the

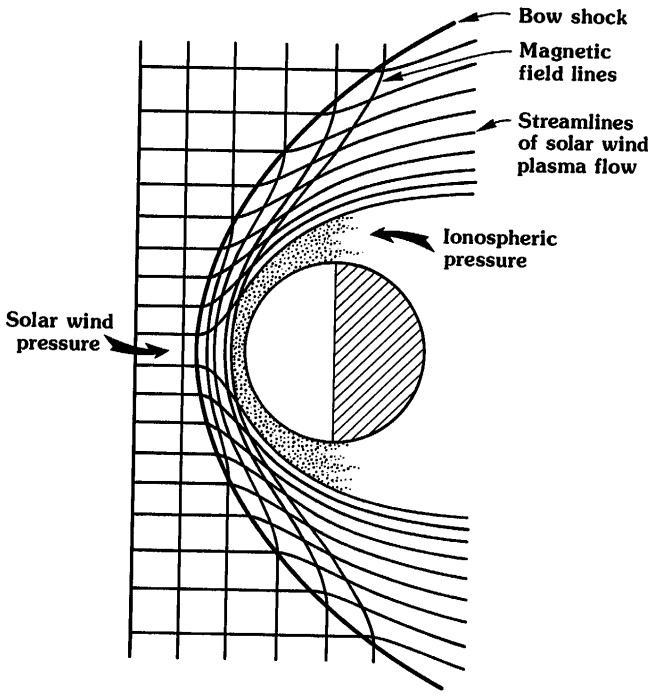


FIG. 1.21. Solar-wind interaction with an unmagnetized planet. The ionospheric pressure stands off the solar-wind flow, so that the streamlines going from left to right flow around the planet. The magnetic field, which is shown here perpendicular to the flow in the solar wind, is bent around the obstacle by this interaction (From Luhmann, 1986.)

Mariner 10 mission, and currently there are no plans to return to Mercury. Thus the Mercury magnetosphere will remain mysterious for many years to come.

In 1972 and 1973, the first spacecraft to the outer solar system, *Pioneer 10* and *11*, were launched, reaching Jupiter in December 1973 and 1974, with *Pioneer 11* going on to Saturn in 1979. Now *Pioneer 10* and *11* are heading out of the solar system, with *Pioneer 10* going downwind relative to the interstellar medium, and *Pioneer 11* upwind. *Voyager 1* and *2* were launched in 1977 and reached Jupiter in 1979 and Saturn in 1980 and 1981. *Voyager 2* then went on to successful encounters with Uranus in 1986 and Neptune in 1989. Both *Voyager 1* and *2* are now heading upwind toward the heliopause.

Those missions revealed well-developed magnetospheres at all the outer planets, each with a bow shock, magnetopause, and magnetotail. The magnetosphere of Jupiter distinguishes itself because its rapid rotation, coupled with a strong plasma source at the moon Io, causes the magnetosphere to be distorted into a disklike geometry. Jupiter is also a source of intense radio waves. Saturn has a simpler magnetosphere, with no strong mass source. Its magnetic field is almost perfectly aligned with its rotation axis.

Uranus and Neptune have unusual magnetospheres because of the unusual orientations of their planetary magnetic fields. Both fields are very complex, and when each is fitted with a dipole moment, the best-fit

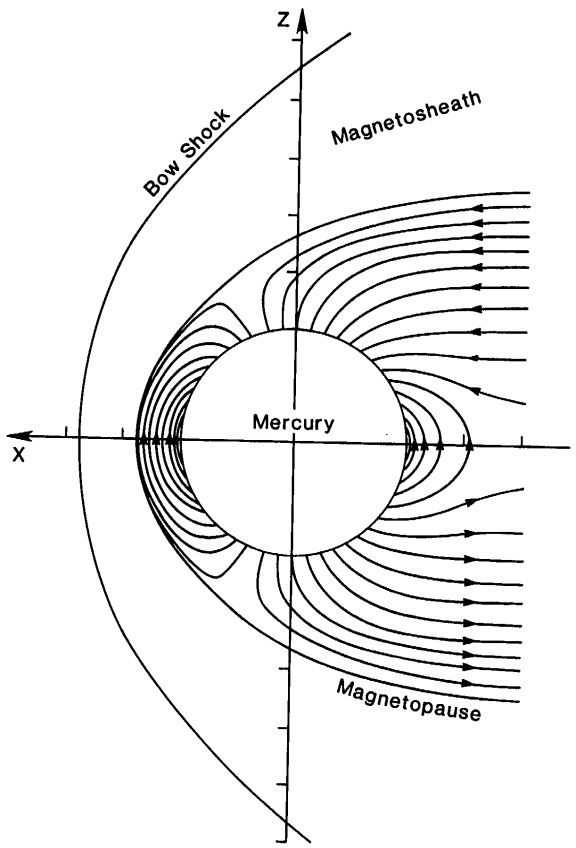


FIG. 1.22. Magnetic-field configuration and noon–midnight cross section of Mercury's magnetosphere. (From Russell et al., 1988.)

dipole is at a large angle to the rotation axis, and the dipole moment is offset from the center of the planet.

Uranus's spin axis is nearly in its orbital plane and is currently nearly pole-on to the sun. Because its magnetic axis is at such a large angle to its rotation axis, its magnetosphere undergoes large oscillations during the course of a day. However, currently the angle between the magnetic moment and the solar-wind flow is not much larger than the maximum angles possible at the earth. Neptune has a more customary rotation axis, roughly perpendicular to its orbit around the sun, but its planetary magnetic field is even more complex than that of Uranus. In both magnetospheres, the radiation belts are much more benign than those of the earth. For Uranus and Neptune we shall have to rely on the findings provided by *Voyager 2* for the foreseeable future, but plans are being made for a Saturn orbiter, called *Cassini,* to arrive in 2004, and a Jovian orbiter, *Galileo,* has been launched and is on its way to Jupiter, to arrive in 1995. Chapter 15 describes the phenomena occurring in the magnetospheres of these outer planets.

1.8. CONCLUDING REMARKS

The field of solar-terrestrial physics has advanced greatly since the earliest recorded sightings of the aurora. We now have physical models for almost all the observed phenomena. Sometimes several competing models exist. The discipline has evolved from one of remote sensing to in situ observations, theory, and computer modeling. In fact, it is now perhaps more appropriate to refer to the field as "space physics," as one of the major journals in the field does, rather than solar-terrestrial physics.

ADDITIONAL READING

Brekke, A., and A. Egeland. 1983. *The Northern Light*. Berlin: Springer-Verlag.
Chapman, S., and J. Bartels. 1940. *Geomagnetism*. Oxford University Press.
Eather, R. H. 1980. *Majestic Lights*. Washington, DC: American Geophysical Union.
Gilbert W. 1893. *De Magnete*, trans. P. Fleury Mottelay. Reprinted 1958, New York: Dover.
Helliwell, R. A. 1965. *Whistlers and Related Ionospheric Phenomena*. Stanford University Press.

PROBLEMS

1.1. Discuss the role of new technology in the development of solar-terrestrial physics.

1.2. At its typical velocity of 440 km $\cdot$ s^{-1}, how long does it take the solar wind to arrive at Mercury, Earth, Jupiter, and Pluto?