Fundamentals of Space Physics

J. M. Hughes, PhD
Physical Sciences Department
Embry-Riddle Aeronautical University

To the students of EP410 - Space Physics

Contents

Preface

This text grew out of course notes for a senior-level class in Space Physics taken by Engineering Physics and Space Physics majors at Embry-Riddle Aeronautical University in Daytona Beach, FL. Most of these students go on to become either engineers working in the general field of aerospace or graduate students across a wide variety of disciplines. Notably, most do not become space physicists or space science researchers and, largely because of this fact, this text is intended to provide a broad overview of the space environment and an introduction to the fundamental physics at work in it. Until the end, almost no attempt is made to discuss or describe topics of current research; the focus of this text is on the fundamental processes governing the environment and on the phenomena they produce.

The presentation here assumes no previous knowledge of plasma physics and Chapter 2 is intended to provide an introduction to the topic at the minimum level necessary for later discussions. Following that chapter, Chapters 3-**??** trace the flow of mass and energy from the Sun to Earth, through it's magnetosphere and atmosphere, and ends with a discussion of auroral physics. All of this subject matter then leads into Chapter **??** which discusses some outstanding questions and areas of current research. **Z: do we want to do this?**

Acknowledgments

We who labor in academia are blessed with pleasant and challenging work. We are also blessed by the history, presence and support of our typically excellent colleagues and research communities. Our investigations and research activities are almost always community endeavours and any appreciations I am in a position to express are due to the Space Physics community as a whole. But perhaps a few from whom I have learned most directly may be mentioned. Thanks therefore go to the following gaggle of folk: Drs. James

LaBelle and Simon Shepherd of Dartmouth College; Mr. Mike Trimpi of Dartmouth College; Dr. Gulamabas Sivjee of Embry-Riddle Aeronautical University; Drs. William Bristow, Antonius Otto and Dirk Lummerzheim of the Geophysical Institute at the University of Alaska Fairbanks; Dr. Mike Kosch of Lancaster University; and the students of EP410 who have asked many thought-provoking questions.

John M. Hughes
Datyona Beach, FL, 2014

Chapter 1

Introduction and Overview

1.1 Introduction

It is good to begin with a definition. The term *Space Physics*, as we use it here, is meant to be synonymous with the phrase *Physics of the Solar-Terrestrial Environment*, where by Solar-Terrestrial Environment we mean that part of Earth's environment that begins at Earth's surface and extends out to several tens of Earth radii. Given this definition, it may seem odd to retain the word "solar" because the environment as we define it does not extend to the Sun itself. Perhaps it is odd, but we will see many times throughout the text that the Sun is the main source of mass, energy and momentum that drives almost all of the exciting and important physics and engineering considerations in this environment. To borrow a phrase, the Sun is the "elephant in the room" of the near-Earth space environment, even though it is not there physically. And so, although we will principally be concerned with the near-Earth space environment, we will be ever mindful that the Sun's presence, if not the Sun itself, is never far away. Part of Chapter 3 addresses solar physics.

The solar-terrestrial environment is filled with plasmas and radiations of different types, with electric and magnetic fields, with highly energetic charged particles that interact with and modify these fields, with neutral particles and, it must not be omitted, with Earth's satellite fleet and ground-based technologies that can be, and indeed continually are, impacted by all these things. This environment is complex and important, not only due to its effects on the untold worth of human technologies but also due to the fundamental and compelling physics involved in understanding the observed and expected interactions, boundaries, and phenomena. It is this environment with which our study of space physics is concerned.

It could be argued that the field of space physics had its genesis in early theories and considerations related to the northern lights, or *aurora borealis*[1]. As one of nature's most spectacular phenomena, the aurora has captivated the human mind for eons, probably, we may suppose, ever since it was first observed[2]. Yet even with this ancient history, the field of space physics must be considered relatively young. Its scientific investigation as we understand it in the modern sense began less than 100 years ago when technology progressed to a point where humans began to have significant interactions with the solar-terrestrial environment. In particular, the development of radio, radars and satellites have provided, and continue to provide, major impetuses to the study of this environment.

We will begin our investigation of this environment in Chapter 3 with an introduction to the Sun and proceed away from it, chapter by chapter, through the interplanetry medium in the direction of Earth. In Chapter 4 we will first encounter a region known as Earth's magnetosphere and we will spend two chapters there considering its formation, structure and dynamics. We then continue towards Earth's surface and in Chapters 6-7 encounter Earth's neutral and ionoized atmospheres. An introduction to and overview of auroral physics is given in Chapter **??**. Finally, our investigation concludes in Chapter **??** with a summary disucssion leading to outstanding questions and areas of current research **(Z: do we want to do this?)**

[1] The phrase *aurora borealis* is composed of the names for the Roman goddess of the dawn, *Aurora* and the latinized Greek name for the north wind, *Boreas*. The same phenomena occurs in the southern hemisphere where it is called the *aurora australis* or southern lights.

[2] The arctic peoples have a rich folklore surrounding the aurora. Interestingly, in most of this folklore the aurora is taken as a bad omen or as respresentative of evil spirits. Those who have observed the aurora's sometimes violent undulations under the deep darkness and silence of an arctic night may empathize with these associations. For an overview of auroral folklore, the interested reader is referred to any of the excellent popular books on the subject (for example, *Aurora: The Northern Lights in Mythology, History, and Science* by Harlald Falck-Ytter.)

1.2 Overview

1.2.1 The Sun and the Solar Wind

The Sun is a star and, as astronomers tell us, a rather unremarkable one at that. Perhaps the Sun *is* unremarkable when compared to other stars, but no one would disagree that it is most certainly remarkable in its relation to Earth. It has rightly been called our “engine of life” and we may think of it as one of the few constants in our ever-changing world[3]. From day to day and decade to decade, it appears the same to our naked eye. More sensitive instruments and the space environment itself reveal, however, that the Sun is by no means constant. It not infrequently exhibits violent storms that emit photon and particle radiations capable of seriously impacting and even destroying Earth-based technologies. Even during its relatively calm periods, its temperatures are so extreme that its outer atmosphere is continuously “boiling off” into space, bathing Earth (and the other planets as well as everything else in the solar system’s space environment) in a continuous flow of charged particles and magnetic fields. This continuous flow of charged particles is known as the *solar wind* and it, with the *interplanetary magnetic field*, or IMF, it carries forms a vital connection between the Sun and Earth. The solar wind and the IMF do not penetrate to Earth’s surface but, as illustrated in Figure 1.1, instead interact with Earth’s own magnetic field to form what is known as the magnetosphere.

1.2.2 The Magnetosphere

Earth’s main magnetic field is generated by electric currents flowing at great depths beneath its surface and to a first approximation may be fictitiously conceptualized as the field due to an exceedingly strong bar magnet located at the center of our planet. This geomagnetic field extends out into space where it encounters the solar wind and IMF, against which it forms a sort of barrier not entirely unlike a rock in a stream’s flow of water. As the solar wind flows around this barrier, it confines Earth’s magnetic field inside a boundary known as the magnetopause that identifies the outer reaches of our magnetosphere. The magnetosphere protects both life on the surface and most satellites in orbit from continuous bombardment by the solar wind

[3] The quoted phrase and essence of this sentence is taken from the excellent IMAX movie *Solar Max*, 2002.

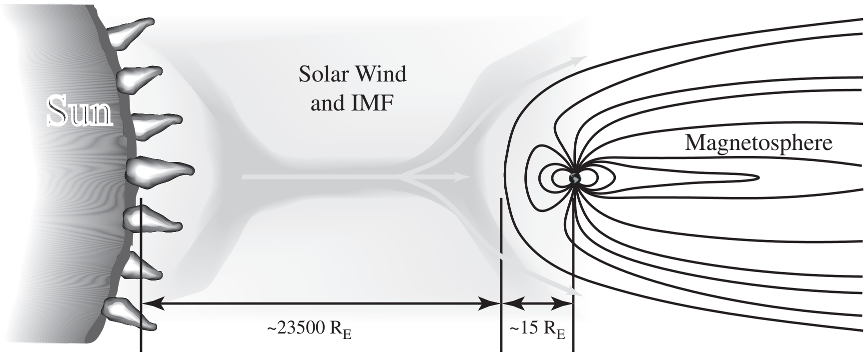

Figure 1.1: A schematic illustration (not to scale) of the Sun, the interplanetary medium filled by the solar wind and IMF, and Earth's magnetosphere.

and IMF. Although under ordinary condititions the magnetosphere is large enough to provide protection to even those satellites orbiting at the large altitudes required by geostationary orbits[4], solar storms may push, and in the past have pushed, the boundary of the magnetosphere so near Earth that these satellites are exposed to the full force of the solar wind. In such cases, system failures and even destruction are unfortunate likelihoods.

The magnetosphere is not a static object and the electric fields, magnetic fields, and plasmas inside it are continually evolving. Among many other important effects, the dynamics of this evolution are responsible for producing the aurora mentioned earlier.

1.2.3 The Neutral Atmosphere

We live in a region known as the troposphere, the lowest layer of Earth's neutral atmosphere, and we are all familiar with the fact that, as one ascends in altitude from the surface, the temperature decreases rather rapidly, as do the atmospheric density and pressure. As it turns out, approximately two thirds of the atmosphere's mass lies below 8 km (∼5 miles or ∼26,000 ft)

[4] Geostationary orbits are those for which a satellite always remains above a fixed point on Earth's surface. To achive this, the orbit must be circular with zero inclination (orbiting above the equator) and have a period equal to one sidereal day (the time it takes Earth to complete one rotation relative to inertial space). Such orbits have an orbital radius of $6.6R_E$ where R_E is Earth's radius.

and over 99% lies below 50 km. We live and breathe within a very thin shell of suitable atmosphere!

While the density and pressure essentially fall off monotonically, the temperature profile of the atmosphere is not so simple. In the troposphere, temperature decreases with increasing altitude but, at some point, the trend reverses so that it begins to increase with increasing altitude. Amazingly, the trend reverses itself two more times within the span of several hundred kilometers so that the profile of temperature as a function of altitude is a complicated set of curves with both positive and negative slopes. This temperature profile and several other considerations make the upper neutral atmosphere a very complicated part of the solar-terrestrial environment. Among these complications is the fact that, embedded within the neutral atmosphere, there is another atmosphere that consists of ionized constituents.

1.2.4 The Ionosphere

Photon radiation and, most importantly, extreme ultra-violet radiation from the Sun is undeflected by Earth's magnetosphere and penetrates to the upper regions of Earth's neutral atmosphere. These photons have energy sufficient to ionize a portion of the neutral atmosphere so that electrons and ions are continually produced. At certain altitudes the density of the electrons and ions and the types of ions produced are such that their recombination into neutral constituents is relatively slow and a net concentration of ions and free electrons result. This ionized part of the atmosphere is known as the ionosphere. It is conductive, can interact strongly with radio transmissions and supports the flow of strong electric currents. For these and other reasons, the ionosphere plays a very important role in the overall solar-terrestrial environment.

1.2.5 Summary

The reader may now appreciate that the solar-terrestrial environment is complex, highly coupled, significant to our technologies, and compelling as a field of study. In the remainder of this text, we aim to introduce the student to the field primarily through a first-principles approach. It must be noted that this text is not intended to be a treatment of current research or of outstanding problems in the field. We aim instead to provide the student with both qualitative and quantitative understandings of basic phenomena and interactions. The text is intended to be foundational and to provide students with a competent overview in preparation for future careers as engineers, physicists, and space scientists.

Before we begin our study of the solar-terrestrial environment, a significant point must be made. Of all the regions mentioned above (the Sun, the interplanetary medium containing the solar wind and IMF, the magnetosphere, the ionosphere and the neutral atmosphere), all but the neutral atmosphere are *plasmas*. Consequently, the field of space physics is extensively grounded in both the physics and the nomenclature of plasmas. And so we begin in the next chapter by discussing some important and fundamental topics from plasma physics.

Exercises

1.1: Research the other planets in our solar system to determine which have magnetospheres. Briefly discuss ways in which these other magnetospheres are similar and/or different than Earth's magnetosphere.

Chapter 2

Plasma Physics Fundamentals

2.1 Plasma Definition

2.1.1 General Plasma Properties

Before proceeding to study plasmas in some detail, we must first answer the basic question: What is a plasma? That is, before diving into details, let us as before first define the thing that we will be investigating.

There are four states of matter: solid, liquid, gas, and plasma. In general terms[1], a material is in the solid state when chemical bonding provides for a rigid structure wherein each molecule (or atom - here the single term molecule will be used for convenience) is held in a fixed position relative to the rest of the material. A solid cannot, therefore, easily change its shape or volume. As a solid is heated, the added thermal energy increases the random motion[2] of these molecules until, at some point, a phase transition occurs and the random thermal motion overcomes the chemical bonding holding each molecule rigidly in place. The material has melted and is now a liquid wherein its molecules maintain approximate physical positions relative to the bulk of the material but are in no definite, fixed positions. The liquid may therefore, for example, change shape to assume that of its container.

As a liquid is heated, the added thermal energy continues to increase the random motion of the material's molecules until, at some point, molecules

[1] By this I mean to caution you: Don't take the two following paragraphs as strictly and universally accurate!

[2] Here and below, this random motion should be thought of as a measure of the material's temperature.

have sufficient kinetic energy to break free of the bulk material and, so to say, burst freely into space. Another phase transition has taken place and the material evaporates into a gas. In the gasseous state, each molecule is, to some order of approximation, independent of all others and the material is now free to change both its shape and its volume. As we continue to add energy to the material, it now becomes possible that, in addition to increasing the temperature, we may provide enough energy to strip electrons from some of the otherwise neutral molecules, resulting in a mixture of neutrals, ions and electrons. The material has undergone yet another phase transition and is now a plasma. In general, everyday terms then, we can describe a plasma as an *ionized gas.*

"A plasma is an ionized gas" is a fine definition when speaking with the man or woman on the street or with colleagues who are non-specialists. But we seek a more precise and accurate definition. To aid us in efforts towards obtaining such a definition, let us first note a few interesting properties of plasmas.

Suppose a substance that may properly be called a plasma is created in some container. Each electron that was freed in the phase transition was liberated when a neutral was ionized, so it must be the case that there are equal numbers of electrons and ions, regardless of what fraction of the neutrals were ionized[3]. A plasma created in this way is therefore electrically *neutral.* That is, the sum of all charges equals zero and, at least over large- and long-enough spatial and temporal averages, there should be same number of electrons in any given volume as there are ions. Let us denote the ion density as n_i and the electron density as n_e. It should be that $n_i = n_e$. Now it must be admitted that on some spatial and time scales, there may be slight departures from neutrality (more on this later), but we can with confidence ammend our definition and say that a plasma is a *quasineutral gas of charged particles.*

Another striking and vastly important property of plasmas is that they exhibit *collective behaviour.* Consider first a group of 10 billiard balls scattered randomly on a pool table. I may move one or two of those ball without affecting the positions of any others, so long as I cause no collision (as is too often the case when I play pool!). But not so with plasmas. If instead of 10 billiard balls, the pool table was populated with 4 "neutrals", 3 "ions" and

[3] Unless otherwise noted, we will assume throughout this text that ions are singly ionized.

3 "electrons", moving one or two of the ions or electrons will result, through the actions of the Coulomb force, in the movement of all other charged balls and, quite possibly, motion of the neutrals due to collisions. We therefore say that plasmas exhibit collective behaviour.

A useful, if not precise, definition: "A plasma is a quasineutral gas of charged and neutral particles which exhibits collective behavior." [Chen, 1983, p.3]

So far we have seen that plasmas are quasineutral and that they exhibit collective behavior. A third property of plasmas is that the charges must be mobile. Suppose, as an extreme example, the temperature of a plasma was reduced to zero. That is, suppose all random thermal motion was eliminated. What would happen to the ions and electrons in the plasma? The Coulomb force would as always be effective and would cause ion/electron pairs to come, and to stick, together. They would neutralize each other and the plasma would, after a very short time, no longer be ionized - it would return to the neutral gas from whence it came. Of course, the temperature need not be zero for this neutralization to occur. It will substantially occur whenever and wherever the electric potential energy exceeds the thermal energy. It must be then that, to have a plasma, the following condition is required: $kT >> q\phi$ where k is Boltzman's constant[4], ϕ is the electric potential, and q is the ion charge.

Plasmas must contain random thermal motion sufficient to overcome the attractive Coulomb potential energy gradient between ions and electrons.

2.1.2 Debye Length and Plasma Parameter

Let us take a further step in the direction of a proper definition by considering the effects of introducing a test ion into an otherwise neutral plasma. The excess positive charge provided by the test ion will set up an electric field that imparts a force on all other charged particles. They will be accelerated and will move in response. But ions with their heavy nuclei are much more massive and consequently accelerate much less than electrons and it is a

[4] $k = 1.3806488 \times 10^{-23}$ $\mathrm{m}^2 \cdot \mathrm{kg/s}^2/\mathrm{K}$, named the Boltzmann constant after the Austrian physicist Ludwig Boltzmann (1844-1906).

useful assumption to take the ions as stationary while the electrons alone move in response to the field. The electrons are attracted to the test charge but recall that their thermal energy greatly exceeds their electric potential energy. They therefore move towards the test charge but do so in a very random manner. In effect, they will swarm around the test charge as a cloud of hungry mosquitoes on the Alaskan tundra would swarm around an intrepid backpacker wearing mosquito repellent. They are attracted to the test ion, but not strongly enough to overcome their random thermal motion and so a swarming cloud of electrons gathers around the imposed test charge.

The "size" of this cloud is an important and fundamental plasma parameter and will form the basis for the first of what will become three quantative conditions that serve as an acceptable technical definition of a plasma. Before we derive the size of this cloud, let us imagine the result at two insightful limits: when the electron temperature goes to zero and to infinity. At zero temperature, electrons have no thermal energy and we have already assumed that the ions are stationary. The electron/ion pairs will stick together as described before and the size of the cloud will therefore be zero in this case. In the case of infinite electron temperature, the random motion will be infinitely more dominant than electrical attraction and the size of the cloud will be infinite. Clearly then, the size of the cloud will depend on the electron temperature such that as temperature decreases, the size of the cloud decreases.

We can appreciate another parameter that impacts the size of this electron cloud by thinking a little more carefully about it. When the test charge is initially placed in the plasma, *many* electrons cloud around it because no single electron with its dominant thermal energy is able to completely neutralize the created electric field. But it does seem reasonable that, as the density of electrons (before introducing the test charge) is increased, the size of the cloud should decrease as more electrons per unit volume are available to assist in neutralizing the electric field. So we suppose that the size of our electron cloud should grow with increasing temperature and decrease with increasing electron density.

To obtain an expression for the size of the cloud, let us begin with Poisson's equation:

$$\nabla \cdot \mathbf{E} = \rho/\epsilon_0$$

where $\mathbf{E}$ is the electric field and ρ is the charge density in the plasma[5]. Note

[5] $\epsilon_0 = 8.854187817... \times 10^{-12}\ \mathrm{N \cdot m^2/C^2}$ is the permittivity of free space. Its value is

that in the same average sense described before, both of these quantities would be zero in the absence of the imposed test charge. The electric potential[6] is given by $\mathbf{E} = -\nabla\phi$ so that we may rewrite the above equation as $\nabla \cdot (-\nabla\phi) = \rho/\epsilon_0$ or $\nabla^2\phi = -\rho/\epsilon_0$[7]. The charge density results from the presence of ions and electrons and we may write $\rho = en_i - en_e$ where e is the electron charge[8] and it is assumed as before that each ion is singly ionized. The potential is then

$$\nabla^2\phi = -\frac{e}{\epsilon_0}(n_i - n_e). \tag{2.1}$$

To solve Equation 2.1, which involves partial derivatives in space, we must identify how n_i and n_e vary in space. We have assumed that the massive ions are stationary and it is reasonable to further assume that their density is constant throughout the plasma. Let us call that constant n_0, the average density of charged particles (of a given sign) in the plasma. That is, we assume $n_i = n_0$ which states that the density of ions is equal to the overall average density of charged particles and that it is independent of position[9]. The density of electrons, on the other hand, will not be constant over space. There will be more electrons per unit volume near the test charge and fewer as distance increases from the test charge. Further, assuming thermodynamic equilibrium, we can state without proof that the distribution of electrons follows the Boltzman distribution[10] so that

$$n_e(\phi) = n_0 e^{\left(\frac{e\phi}{kT_e}\right)}$$

where T_e is the electron temperature. Here, the gradient in the electric potential energy creates an attraction between an electron and the test ion

defined as $\epsilon_0 = \frac{1}{\mu_0 c^2}$ where $c = 2.99792458 \times 10^8$ m/s (by definition) is the speed of light in vacuum and $\mu_0 = 4\pi \times 10^{-7}$ T · m/A (also by definition).

[6] Notice that the electric potential we are about to write down assumes we are dealing with an electro*static* situation. That is, we will assume that whatever adjustments the plasma makes in response to the test charge have already happened and, on average, the electron density, and thus the electric field, is static.

[7] If $\mathbf{A}$ is the magnetic vector potential (so that $\mathbf{B} = \nabla \times \mathbf{A}$) and one chooses the Coulomb gauge so that $\nabla \cdot \mathbf{A} = 0$, this relation holds even without the electrostatic assumption.

[8] $e = 1.602176565(35) \times 10^{-19}$ C is the measured value of the electron charge.

[9] That is, the ions are homogeneous.

[10] If you are uncomfortable with this result stated without proof (and you should be!), see [Chen, 1983, p.9] for a few more steps.

but this attraction is, in some sense, countered by the random thermal motion in such a way that a Boltzman distribution results.

We may substitute these results for n_i and n_e into Equation 2.1 to get

$$\nabla^2\phi = -\frac{e}{\epsilon_0} n_0 \left(1 - e^{\frac{e\phi}{kT_e}}\right).$$

Now this is still a difficult equation to solve! A further simplification can be achieved by recalling that, in a plasma, the thermal energy greatly exceeds the electric potential energy. Because $e\phi << kT_e$, we may Taylor expand about $e\phi = 0$ to find that $e^{\left(\frac{e\phi}{kT_e}\right)} \approx 1 + \frac{e\phi}{kT_e}$ which yields

$$\nabla^2\phi = \frac{n_0 e^2}{\epsilon_0 k T_e}\phi. \tag{2.2}$$

Solving Equation 2.2 assuming spherical symmetry and applying a boundary condition requring the potential to go to zero as the radial coordinate r approaches infinity, we arrive at the solution for the potential as a function of distance from the test charge:

$$\phi(r) = C \exp\left(-r\left(\frac{n_0 e^2}{\epsilon_0 k T_e}\right)^{\frac{1}{2}}\right) = Ce^{-r/\lambda_D}. \tag{2.3}$$

where λ_D is known as the *Debye length.*

For our purposes here, it is not important to evaluate the constant C in Equation 2.3. What is important is that we have found how the electric potential, and thus the electron density, varies with distance from the test charge. The electric potential and electron density fall off exponentially with a scale length of

$$\boxed{\lambda_D = \left(\frac{\epsilon_0 k T_e}{n_0 e^2}\right)^{\frac{1}{2}}.} \tag{2.4}$$

As we initially supposed, the cloud size characterized by λ_D increases with increasing temperature and decreases with increasing electron density. These are in fact the only two variables on which it depends.

Some numbers may be instructive here. Combining the constants in Equation 2.4 yields the useful approximation that

$$\boxed{\lambda_D \approx 69\,(T_e/n_0)^{1/2}}$$

where T_e is in Kelvins and n_0 is in electrons per cubic meter.[11] Table 2.1 shows representative temperatures, densities and Debye lengths for several plasmas.

Plasma	Electron Temperature (K)	Electron Density (m^{-3})	Debye Length
Earth's Ionosphere	1000	10^{12}	2 mm
Interstellar Gas	6000	10^{5}	20 m
VanAllen Radiation Belts	10^{6}	10^{9}	2 m
Fusion Reactor	2×10^{8}	10^{20}	0.1 mm
Sun's Core	10^{7}	10^{32}	10^{-11} m
Solar Wind	10^{5}	10^{6}	20 m

Table 2.1: Typical Debye lenghts for selected plasmas.

The Debye length is of fundamental importance becuase it determines the scale over which a plasma is able to neutralize the effects of any imposed charge disturbances or electric fields. Moving to a distance of λ_D away from a charge disturbance, the potential (or electric effect of that charge) has decreased to approximately 37% of its maximum value. Moving a distance of $2\lambda_D$ away, the potential has decreased to approximately 14% of it maximum, and moving to a distance of $5\lambda_D$, the potential is less than 1% of its maximum value. That is, at distances of $\sim 5\lambda_D$ and farther, it is as if the disturbance was not there at all; any effects have been "screened" by the intervening cloud of swarming electrons.

As a brief aside, it may be worthwhile to point out one important implication of this screening effect. When a rocket or satellite passes through the plasma of Earth's ionosphere or magnetosphere, the vehicle often acquires a net charge and, as a result, essentially flies around carrying with it a Debye cloud of the opposite polarity to the charge acquired by the vehicle. Should an instrument on the vehicle be intended to measure the ambient plasma density, it is essential that the sensor be physically separated from the vehicle by at least several Debye lengths in order to avoid sampling the density inside the cloud rather than the ambient density. As shown in Table 2.1,

[11] This approximation is accurate to about 1 part in 10^4.

Debye lengths in the near-Earth space environment vary over several orders of magnitude so that this problem requires careful engineering and science considerations.

This ability to screen potentials (or, in essence, short out electric fields) is a key defining feature of plasmas and we make it the first of our three conditions a plasma is required to satisfy: The physical size of a plasma must be much larger than the Debye length. Furthermore, we can impose a second condition and require that, before the introduction of any test charges, a sphere with radius λ_D (a so-called Debye sphere) contains *many* electrons. The number of electrons contained in a Debye sphere is called the *plasma parameter* and is given by $\Lambda = \left(\frac{4}{3}\pi\lambda_D^3\right) n_0$. The factor of $\frac{4}{3}\pi$ is often dropped in practice in which case the plasma parameter gives the number of electrons in a Debye *cube*.

So we are now in a position to state the first two of three conditions that must be satisfied by a plasma:

1. $\lambda_D << L$

2. $\Lambda >> 1$

where L is a parameter characterizing the physical extent of the plasma. That is, a plasma must be many times larger than the Debye length (so that screening of imposed potentials will be effective) and there must be many electrons in a Debye sphere (or cube). The remaining one of our three conditions can be stated after considering another fundamental property of plasmas.

2.1.3 Plasma frequency

Figure 2.1a shows, as we considered before, an equilibrium state in which a uniform background of ions (the black dots) is surrounded by a swarm of mobile electrons (the gray rectangle). Suppose we somehow stopped time for a moment, grabbed every electron in the plasma and displaced them all to right as indicated in Figure 2.1b. If we then let go of those electrons and restarted time, what motion do you suppose would result (stop and think before reading on!)?

First, we again realize that no matter what the electron motion is, a good first approximation would be that the much more massive ions remain essentially stationary. So we turn our attention to the electrons. Will they

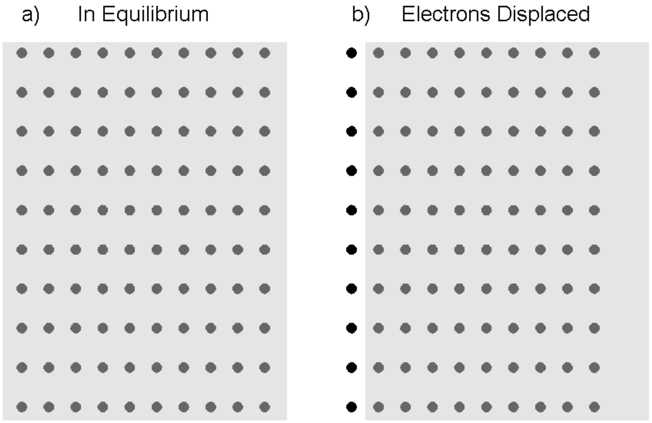

Figure 2.1: a) An plasma in equilibrium indicated by stationary ions (the dots) and mobile electrons (the rectangle). b) A perturbed plasma where all electrons have been shifted to the right while holding the ions stationary.

move or will they also remain stationary? If they are to move, what force causes them to accelerate from rest?

Consider the electric field that would be present in Figure 2.1b. It will be directed from regions of positive charge (the left side of the figure) to regions of negative charge (the right side of the figure). Electrons from the right will then accelerate to the left in response to this field and, by the time they have returned to their equilibrium positions at which point the electric field and force has vanished, they will have gained some momentum and will overshoot until the mirror image of the initial perturbed condition is present. At this point they will be accelerated to the right, will overshoot the equilibrium again and the situation will return to that of Figure 2.1b. The process will repeat and it should be clear that a periodic motion will result.

The frequency of this oscillation is known as the *plasma frequency* and it is another fundamental plasma characteristic. It will, in fact, form the basis for our third and final condition that a plasma must satisfy. This plasma frequency can be derived in a variety of ways and here we take the opportunity to introduce a very useful technique known as linear perturbation analysis.

As the electrons are oscillating, we can expect three quantities to vary in space and/or time: the electric field, the electron density, and the electron velocity. Treating the problem in one dimension, we therefore have three unknowns and require three independent equations to define the motion. Our three equations will be Newton's second law (in essence), conservation of charge, and Poisson's equation. We will also make five simplifying assumptions:

1. $\mathbf{B} = 0$ (there is no background magnetic field so the plasma is unmagnetized)
2. $kT_e = 0$ (we are interested in the bulk motion of the electrons, not their random motion)
3. ions are stationary (they are much more massive than electrons)
4. the plasma is infinite in extent (we perturb only a part of it)
5. the problem is one-dimensional (as stated above).

Newton's second law applied to an electron is $\sum \mathbf{F} = m_e \frac{d\mathbf{v}_e}{dt}$ where m_e and $\mathbf{v}_e$ are the electron mass and velocity, respectively. The only force acting on an electron is the Coulomb force so that

$$m_e \frac{d\mathbf{v}_e}{dt} = -e\mathbf{E}.$$

The time derivative in the above equation implicitely contains terms due to variations in time and variations in space. To separate these dependences, we can expand the derivative using the chain rule as

$$\begin{aligned}\frac{d\mathbf{v}_e}{dt}i &= \frac{\partial \mathbf{v}_e}{\partial t} + \frac{\partial \mathbf{v}_e}{\partial x}\frac{dx}{dt} + \frac{\partial \mathbf{v}_e}{\partial y}\frac{dy}{dt} + \frac{\partial \mathbf{v}_e}{\partial z}\frac{dz}{dt} \\ &= \frac{\partial \mathbf{v}_e}{\partial t} + (\mathbf{v}_e \cdot \nabla)\,\mathbf{v}_e\end{aligned}$$

to obtain [12]

[12] Here we are considering variations in the electron's velocity. Suppose, for simplicity, we were instead considering variations in temperature. The total time derivative of temperature T is $\frac{dT}{dt} = \frac{\partial T}{\partial t} + (\mathbf{v} \cdot \nabla)\,T$. The first term on the RHS results from a time change in temperature at a fixed point in space (perhaps there is heater near the thermometer's

$$m_e n_e \left[\frac{\partial \mathbf{v}_e}{\partial t} + (\mathbf{v}_e \cdot \nabla)\, \mathbf{v}_e \right] = -e n_e \mathbf{E}. \tag{2.5}$$

This, the first of our three needed equations, is generally known as a *momentum equation.* If the notation $(\mathbf{v}_e \cdot \nabla)\, \mathbf{v}_e$ is unfamiliar to you, it will be worth pointing out that the parentheses contain a scalar *operator* that acts from the left on the remaining vector velocity. That is, you should not violate the parentheses when evaluating the expression[13].

To obtain the second of our required equations, we need to derive an expression of charge conservation. To this end, consider a closed surface $\mathbf{S}$ (of any shape you like - a rectangular box may be convenient) with an electron current density $\mathbf{J} = -e n_e \mathbf{v}_e$ leaving it. The stationary ions do not contribute to the current density. The integral of $\mathbf{J} \cdot d\mathbf{S}$ over the surface yields the current leaving the enclosed volume so that

$$I_{out} = -\oiint e n_e \mathbf{v}_e \cdot d\mathbf{S} = -\frac{dQ_{enc}}{dt} \tag{2.6}$$

where I_{out} is the current leaving the volume and Q_{enc} is the charge enclosed in it. But the charge enclosed is just the integral of the charge density over the enclosed volume, to wit $Q_{enc} = \iiint (-e n_e + e n_i) dV$. Upon making this substitution, we may pull the time derivative inside the integrals on the right hand side (RHS) of Equation 2.6, realize that the time derivative of the ion charge density is zero (because they are assumed stationary), and employ the divergence theorem on the left hand side (LHS) to obtain

$$\iiint \nabla \cdot (n_e \mathbf{v}_e) dV = -\iiint \frac{\partial n_e}{\partial t} dV.$$

Now, because this result must hold true for any volume (we did not specify any particular volume in the derivation), the integrands must be equal and

$$\frac{\partial n_e}{\partial t} + \nabla \cdot (n_e \mathbf{v}_e) = 0 \tag{2.7}$$

location). The second term on the RHS results from moving the thermometer in a direction along which there is a gradient to the temperature at some fixed time (perhaps by moving the thermometer from inside to outside on a hot day in Florida.) Particularly with satellite observations, it is often difficult in practice to separate these two types of contributions to the time derivative.

[13] Note that this order of operation is not necessary, only convenient. It can be shown that $(\mathbf{v} \cdot \nabla)\mathbf{v} = \mathbf{v} \cdot (\nabla \mathbf{v})$. The expression on the RHS involves the tensor $\nabla \mathbf{v}$ which it is convenient here to avoid.

which is the required expression of conservation of charge known as the *continuity equation*. This equation states that, if the electron density at a given location is changing in time, then that change is given by the divergence of current density from the location. Charge is not allowed to simply appear or disappear, it can only move from one place to another.

We have already met with Poisson's equation, the third of the three we require. Again, Poisson's equation is

$$\nabla \cdot \mathbf{E} = \frac{en_i - en_e}{\epsilon_0}. \tag{2.8}$$

Equations 2.5, 2.7 and 2.8 completely specify the motion of our plasma but, in order to determine the oscillation frequency, they must be solved simultaneously and it is here that linear perturbation analysis comes in handy. We begin by assuming a "perturbed" situation. That is, we assume that our three quantities of interest (electric field, electron density and electron velocity) can be written as the sum of two parts: an equilibrium part and a perturbed part. The perturbed part is what we are interested in because, in equilibrium, no oscillations occur. We therefore take

$$\mathbf{E} = \mathbf{E}_0 + \delta\mathbf{E}, \quad n_e = n_{e_0} + \delta n_e, \quad \mathbf{v}_e = \mathbf{v}_{e_0} + \delta\mathbf{v}_e$$

where quantities with a subscript of 0 are the equilibrium parts (the values before things were perturbed) and the quantities preceeded by a δ are the perturbed parts. Before the situation was perturbed by moving all the electrons to the right, there was no electric field and there was no electron motion so that $\mathbf{E}_0 = 0$ and $\mathbf{v}_{e_0} = 0$. Furthermore, in equilibrium, the electron density equals the ion density which both equal, as we have labelled it, the average charged particle density n_0. We have therefore a significantly simplified system that can be substituted into our three equations.

The continuity equation becomes

$$\frac{\partial}{\partial t}(n_0 + \delta n_e) + \nabla \cdot [(n_0 + \delta n_e)\delta\mathbf{v}_e] = 0$$

and can be *linearized* by assuming that any product of perturbed quantities is negligible compared to terms involving no more than one perturbed quantity. Linearizing and realizing that the time derivative of equilibrium quantities is zero yields

$$\frac{\partial \delta n_e}{\partial t} + \nabla \cdot (n_0 \delta\mathbf{v}_e) = 0. \tag{2.9}$$

Applying the same procedure to the momentum and Poisson's equations respectively give

$$m_e n_0 \frac{\partial \delta \mathbf{v}_e}{\partial t} = -e n_0 \delta \mathbf{E} \tag{2.10}$$

and

$$\nabla \cdot \delta \mathbf{E} = -\frac{e \delta n_e}{\epsilon_0}. \tag{2.11}$$

Now, based on our intuition and physical understanding of the situation, let us assume a set of oscillating solutions given by

$$\delta \mathbf{v}_e = \delta v_e \exp\left(i(kx - \omega t)\right) \hat{\mathbf{x}}$$

$$\delta n_e = \delta n_e \exp\left(i(kx - \omega t)\right)$$

$$\delta \mathbf{E} = \delta E \exp\left(i(kx - \omega t)\right) \hat{\mathbf{x}}$$

where we hope the reader will excuse the redundant notation in the second equation that we retain for convenience. In these equations, $i = \sqrt{-1}$, k is the wavenumber given by 2π over the oscillation wavelength and ω is the oscillation angular frequency. One advantage of assuming solutions of this form is that time and spatial derivatives simplify nicely into algebraic operations by making the following substitutions[14]:

$$\frac{\partial}{\partial t} \rightarrow -i\omega$$

$$\nabla \rightarrow ik\hat{\mathbf{x}}.$$

Applying these substitutions to Equations 2.9, 2.10 and 2.11 yields the following set of algebraic equations:

$$-i\omega m_e n_0 \delta v_e = -e n_0 \delta E$$

$$-i\omega \delta n_e = -ik n_0 \delta v_e$$

$$ik \delta E \epsilon_0 = -e \delta n_e$$

that the student may solve in the usual way of dealing with three equations with three unknowns. The result is an oscillation frequency given by

$$\boxed{\omega = \left(\frac{n_0 e^2}{\epsilon_0 m_e}\right)^{\frac{1}{2}} \equiv \omega_{pe}} \tag{2.12}$$

[14] The student should verify that these substitutions are valid.

where we have introduced a new variable, ω_{pe}, the *electron plasma frequency.*

This plasma frequency is the frequency at which a perturbed, unmagnetized plasma naturally tends to oscillate and it is remarkable that it depends on only a single variable: the background (unperturbed) electron density. The more dense the plasma, the faster it tends to oscillate. The linear frequency given by $f_{pe} = \omega_{pe}/2\pi$ can be nicely approximated by

$$\boxed{f_{pe} \approx 9\sqrt{n_e}}$$

where n_e is the density of electrons per cubic meter.[15] Table 2.2 shows values of the plasma frequency typical for several plasmas.

Plasma	Electron Density (m^{-3})	Plasma Frequency (Hz)
Earth's Ionosphere	10^{12}	9×10^6
Interstellar Gas	10^5	3×10^3
VanAllen Radiation Belts	10^9	3×10^5
Fusion Reactor	10^{20}	9×10^{10}
Sun's Core	10^{32}	9×10^{16}
Solar Wind	10^6	9×10^3

Table 2.2: Linear plasma frequencies for selected plasmas

So far we have made only passing mention of the fact that plasmas are not usually fully ionized. Instead, they are usually a mixture of neutrals, ions and electrons. When a plasma is perturbed and attempts to execute oscillations at the plasma frequency, the electrons will be impeded in their motion as they collide with neutrals. As the density of neutrals increases, the time between electron/neutral collisions decreases and at some point, the collision frequency exceeds the plasma frequency. The "plasma" is therefore not able to oscillate at its natural frequency. In such a case, we are loath to call the substance a plasma. In fact, taking τ to be the electron/neutral collision time, we can formulate the third and last requirement that a plasma must satisfy: $\omega_{pe}\tau >> 1$. That is, the natural oscillation frequency must greatly

[15] This approximation is valid to about 1 part in 300.

exceed the impeding collision frequency so that the plasma executes many oscillations between electron/neutral collisions.

We can then formally say that a plasma must meet the following three conditions:

1. $\lambda_D << L$
2. $\Lambda >> 1$
3. $\omega_{pe}\tau >> 1.$

To put our definition of a plasma into words, the Debye length must be much smaller than the physical extent of the plasma so that effective screening of imposed charges or electric fields can be accomplished, there must be many electrons in a Debye sphere, and the plasma must be able to execute its natural oscillation at the plasma frequency.

2.2 Single-Particle Motions

Figure 1.1 shows a sketch of the Sun-Earth system and we may imagine some of the magnetic field geometries encountered by plasmas in that environment. Near the Earth, the magnetic field lines are curved and, following a particular field line from the equator to the polar regions, we notice that the field strength increases as indicated by converging field lines. If we remain on the equator and cross field lines moving toward the Earth, we notice that, again, the field strength increases. Plasmas in the solar-terrestrial environment often encounter magnetic fields that, as we noticed here, are curved and have gradients in directions parallel and/or perpendicular to the field lines. It is a complicated system, made even more complicated by the sometimes present electric fields that are not shown in Figure 1.1. If we are to understand the flow of plasma through this environment, we must understand how these various field geometries affect the motion of plasma particles. This is the subject we now take up.

Sometimes it is appropriate to model plasmas as a fluid (or a set of fluids), and sometimes it is appropriate to model them as a collection of single particles. Here, let us treat the plasma as a collection of single particles and investigate the motion that results from the action of the Lorentz force on each particle.

2.2.1 Cyclotron Motion

To begin, let us first investigate the motion of charged particles in the presence of a magnetic field **B** with no electric field present. We will make the simplifying assumption that **B** is constant and is therefore not affected by any currents resulting from the particle motions.[16] The Lorentz force on a particle is

$$\mathbf{F} = q\left(\mathbf{E} + \mathbf{v} \times \mathbf{B}\right) \tag{2.13}$$

where q is the particle charge and, in this case, $\mathbf{E} = 0$. Without loss of generality we may take $\mathbf{B} = B\hat{\mathbf{z}}$ [17] so that

$$\mathbf{v} \times \mathbf{B} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ v_x & v_y & v_z \\ 0 & 0 & B \end{vmatrix} = v_y B\hat{\mathbf{x}} - v_x B\hat{\mathbf{y}}$$

and, from Newton's second law,

$$m\dot{v}_x = qv_yB \;\; , \;\; m\dot{v}_y = -qv_xB \;\; , \;\; m\dot{v}_z = 0 \tag{2.14}$$

where the particle mass is m and we employ the usual dot notation to indicate time derivatives.

Equations 2.14 already reveal some interesting features. Notice that there is no acceleration in the direction of the magnetic field. This is because the Lorentz force is perpendicular to **B** and the component of velocity along the field vector is therefore *constant.* Motion in the plane perpendicular to the magnetic field is apparently more complicated. The acceleration along the $x-$axis is dependent on the velocity along the $y-$axis and the acceleration along the $y-$axis is dependent on the velocity along the $x-$axis. These equations of motion are *coupled* and to find the resulting motion, we must

[16] Under this assumption, our solution to the single-particle motions will not be self-consistent but they will be be useful and insightful.

[17] A constant magnetic field has some fixed direction. It greatly simplifies the solution to take the direction of **B** as one of the $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, or $\hat{\mathbf{z}}$ directions so that **B** has only one component. For whatever reason, it is traditional to take **B** along the $\hat{\mathbf{z}}$ axis. This assumption does not restrict the physics of the solution in any way, so we say the assumption is made without loss of generality.

decouple them. This can be done by taking the time derivative of the first two equations:

$$\ddot{v}_x = \frac{qB}{m}\dot{v}_y = -\left(\frac{qB}{m}\right)^2 v_x \tag{2.15}$$

and

$$\ddot{v}_y = -\frac{qB}{m}\dot{v}_x = -\left(\frac{qB}{m}\right)^2 v_y. \tag{2.16}$$

These equations should be familiar and, in particular, it should be obvious that the solutions are oscillatory with a frequency $\omega = qB/m$. However, there is a bit of awkwardness here that we wish to avoid. The charge q can be either positive (for ions) or negative (for electrons) and so the frequency can be either positive or negative. Seeking a frequency that is always greater than or equal to zero, we introduce the quantity

$$\boxed{\omega_c = \frac{|q|B}{m}} \tag{2.17}$$

known as the *cyclotron frequency* or *gyrofrequency* that is the absolute value of the oscillation frequency such that $\omega = \pm\omega_c$ where the top and bottom signs correspond to positive and negative charges, respectively.

A general solution to the equations of motion is then

$$v_x = v_\perp \exp\left(\pm i\omega_c t + i\delta_x\right)$$

and

$$v_y = v_\perp \exp\left(\pm i\omega_c t + i\delta_y\right)$$

where δ_x and δ_y are phase constants to be determined and $v_\perp$ is the magnitude of velocity in the plane perpendicular to the magnetic field (the $xy-$plane in our case). The physically significant part of the above equations is the real part so, for example, what we really mean (no pun intended) is that the general solution for the $x-$component is given by

$$v_x = \Re\left(v_\perp \exp\left(\pm i\omega_c t + i\delta_x\right)\right) = v_\perp \cos(\pm\omega_c t + \delta_x).$$

We retain the complex notation for algebraic simplicity[18] and understand that each expression contains an assumed $\Re$ operator.[19]

[18] The exponential function is normally easier to deal with the trigonometric functions.

[19] These manipulations require the often-handy Euler's formula: $e^{ix} = \cos x + i \sin x$.

Now, we are free to choose either δ_x or δ_y which amounts to choosing the instant at which we set $t = 0$ and, for convenience, we choose δ_x such that

$$v_x = \dot{x} = v_\perp \exp(i\omega_c t). \tag{2.18}$$

Note carefully that we have *not* simply chosen $\delta_x = 0$, but rather have made a choice that depends on the sign of the charge such that the $x-$component of the velocity takes its maximum positive value at $t = 0$. With this choice,

$$v_y = \dot{y} = \frac{m}{qB}\dot{v}_x = \pm\frac{1}{\omega_c}\dot{v}_x = \pm i v_\perp \exp(i\omega_c t) \tag{2.19}$$

and we may solve Equations 2.18 and 2.19 to find the positions as a function of time. Carrying out the integrations and letting x_0 and y_0 be the locations of the particle at $t = 0$ yields

$$x - x_0 = -i\frac{v_\perp}{\omega_c}\exp(i\omega_c t)$$

and

$$y - y_0 = \pm\frac{v_\perp}{\omega_c}\exp(i\omega_c t).$$

The form of these solutions is perhaps move apparent after taking the real part:

$$x - x_0 = r_L \sin \omega_c t \tag{2.20}$$

and

$$y - y_0 = \pm r_L \cos \omega_c t. \tag{2.21}$$

A charged particle moving in the presence of a constant magnetic field will execute uniform circular motion known as *cyclotron motion* with radius

$$\boxed{r_L = \frac{v_\perp}{\omega_c} = \frac{m v_\perp}{|q|B}} \tag{2.22}$$

where r_L is known as the *Larmour radius* or *gyroradius.* The radius of the "gyration" will increase with the particle mass and with the component of velocity perpendicular to the field. It will decrease with an increase in the magnitude of the charge or with the field strength. Also note that the sense of rotation is different for positive (the top sign) and negative (the bottom sign) charges. Positive charges gyrate in one direction and negative charges gyrate in the other. Use of the right-hand-rule for currents shows that the magnetic field generated by this cyclotron motion is always opposite to the magnetic field that causes the motion. That is, a magnetized plasma is inherently *diamagnetic.*

Figures 2.2a) and b) show the motion of an ion and electron moving in the presence of a constant magnetic field directed into the page as indicated by the circled crosses. Each particle of charge equal to one electron charge (either positive or negative) is shown with the same gyroradius, which requires that the electron's perpendicular velocity is a factor of m_i/m_e larger than the ion's. This would be unexpected in reality and one typically finds that the electron gyroradius is very much smaller than the ion gyroradius.

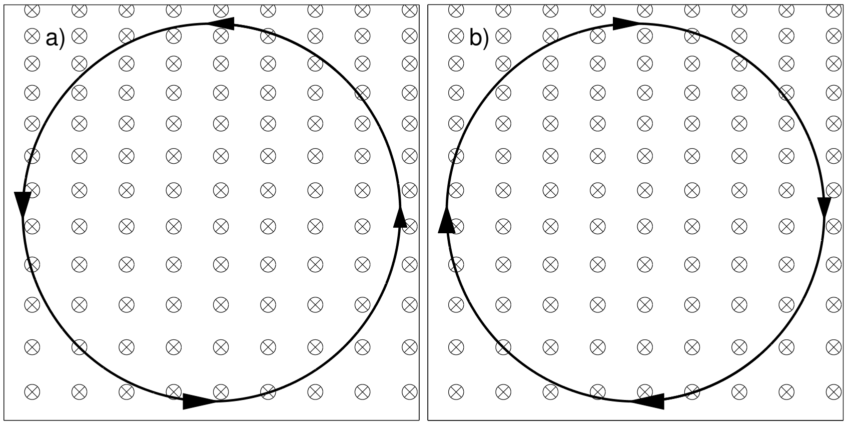

Figure 2.2: a) The motion of an ion in the plane perpendicular to a background magnetic field (directed into the page). b) The motion of an electron in the plane perpendicular to the same magnetic field.

Note from Figure 2.2 that the acceleration is in the $\pm(\mathbf{v} \times \mathbf{B})$ direction for the ion and electron respectively and that the direction points toward the center of the circle. The circulating charges constitute a current that flows counter-clockwise for both particles. The magnetic field produced by these currents within the circular paths points out of the page and so the background field is weakened by the gyromotion. Thus, as was stated above, the plasma is inherently diamagnetic.

If a particle of either charge has a component of velocity in the direction parallel to the background magnetic field, this motion will be unaffected by the field because the force is in the perpendicular plane. The resulting motion will be a helix.

2.2.2 The $\mathbf{E} \times \mathbf{B}$ Drift

Suppose we now allow for the existence of a finite constant electric field $\mathbf{E}$ in addition to the constant magnetic field $\mathbf{B}$. These two fields may be oriented in any directions relative to each other. Now, to establish a convenient coordiate system, let us as before take the $z-$axis to be parallel to the magnetic field so that $\mathbf{B} = B\hat{\mathbf{z}}$. Further, without loss of generality, we are free to rotate our coordinate system about the $z-$axis and it is helpful to rotate such that the electric field lies in the $xz-$plane so that $E_y = 0$.

The force on a charged particle is, as before, given by Equation 2.13. Integrating the z-component to find the velocity in the direction of the magnetic field as a function of time yields

$$v_z = v_{z_0} + \frac{qE_z}{m}t \tag{2.23}$$

where v_{z_0} is the $z-$component of the velocity at $t = 0$. The charged particle experiences constant acceleration in the parallel[20] direction with a magnitude that scales with the parallel component of the electric force (qE_z). The acceleration in the parallel direction is completely independent of the magnetic field. This is expected because the magnetic force term is perpendicular to both the particle's velocity and the magnetic field through the $\mathbf{v} \times \mathbf{B}$ term.

Before deriving the equations of motion in the perpendicular (xy) plane, let us think through the situation to gain some physical understanding. For simplicity, take $\mathbf{E}$ and $\mathbf{B}$ to be perpendicular to each other with $\mathbf{E}$ directed up the page and $\mathbf{B}$ into the page. Let the particle have a positive charge and be held at rest until its release at $t = 0$. As soon as it is released, it will accelerate in the direction of the electric field under the action of the electric force $q\mathbf{E}$ but it will initially feel *no* magnetic force $q(\mathbf{v} \times \mathbf{B})$ becuase it was released from rest. It will therefore begin to move up the page. Once the particle gains some speed, it will feel the magnetic force that attempts to turn it to the left in a circular path, but the intended circle will be distored by the continued action of $q\mathbf{E}$ that always forces it up the page. The gyroradius will

[20] The terms *parallel* and *perpendicular* will be used in reference to the direction of the magnetic field.

increase with the particle's motion up the page due to the speed acquired from the electric force. Once the particle reaches its highest point one-quarter of the way around its path, its speed will begin to decrease as it is decelerated by the electric force and the gyroradius will shrink in response. We will then have a distorded "circle" (actually not a circle at all!) with a larger gyroradius at the top than at the bottom. The effect will be that, when the particle returns to the level from which it started, it will have shifted to the left. After another cycle, the particle will be further to the left. In fact, each cycle takes the particle further to the left and we can say that it is *drifting* in the direction of $\mathbf{E} \times \mathbf{B}$. A negatively charged particle will execute motion in the opposite sense but, as you can appreciate by thinking it through, it will drift in the same direction as the positive charge. Figure 2.3 shows the paths of an ion and an electron (both released from rest) moving under the action of an electric and magnetic field.

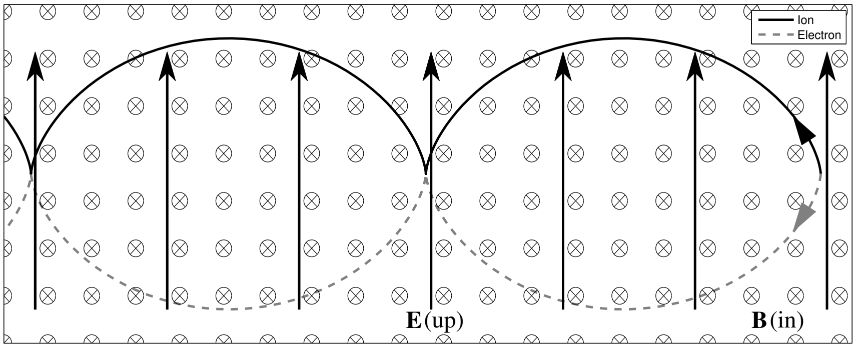

Figure 2.3: The paths of an ion and an electron released from rest (at the far right) in the presence of an electric and magnetic field. The paths are shown to scale and both particles drift to the left.

Now, let us work out the equations of motion to test and extend our understanding. The perpendicular components of Equation 2.13 are

$$\frac{dv_x}{dt} = \frac{q}{m}E_x + \frac{qB}{m}v_y = \frac{q}{m}E_x \pm \omega_c v_y$$

and

$$\frac{dv_y}{dt} = -\frac{qB}{m}v_x = \mp\omega_c v_x$$

where, as before, the top sign applies to ions and the bottom sign applies to electrons. Uncoupling these equations as we did in the case of cyclotron motion gives

$$\ddot{v}_x = -\omega_c^2 v_x \tag{2.24}$$

and

$$\ddot{v}_y = \mp\omega_c \left(\frac{qE_x}{m} \pm \omega_c v_y\right) = -\omega_c^2 \left(\frac{E_x}{B} + v_y\right). \tag{2.25}$$

Notice that these equations are identical to Equations 2.15 and 2.16 obtained for cyclotron motion when $\mathbf{E} = 0$ except for the pesky addition of the $\frac{E_x}{B}$ term in the RHS of Equation 2.25. This term complicates the solution but we can play a nifty mathematical trick that allows us to write the solution down by inspection. The troublesome $\frac{E_x}{B}$ is *constant* (because both E_x and B are constant) and because its time derivatives are zero, we see that

$$\ddot{v}_y = \frac{d^2 v_y}{dt^2} = \frac{d^2}{dt^2}\left(\frac{E_x}{B} + v_y\right)$$

so that

$$\frac{d^2}{dt^2}\left(\frac{E_x}{B} + v_y\right) = -\omega_c^2 \left(\frac{E_x}{B} + v_y\right).$$

Defining $v_y' = \frac{E_x}{B} + v_y$ gives

$$\ddot{v}_y' = -\omega_c^2 v_y' \tag{2.26}$$

The form of Equations 2.24 and 2.26 are identical to those of Equations 2.15 and 2.16 and so the form of the solutions must also be identical.[21] That is, it must be that

$$v_x = v_\perp \exp(i\omega_c t) \tag{2.27}$$

and

$$v_y' = \pm i v_\perp \exp(i\omega_c t)$$

or

$$v_y = \pm i v_\perp \exp(i\omega_c t) - \frac{E_x}{B}. \tag{2.28}$$

Equations 2.27 and 2.28 are those of a gyrating particle *drifting* along the negative $y-$axis with a constant speed of $\frac{E_x}{B}$ as indicated by Figure 2.3.

[21] The effect of the definition $v_y' = \frac{E_x}{B} + v_y$ is to transform the solution into a coordinate system moving with the constant speed $\frac{E_x}{B}$ in the direction of the drift.

This solution is particular to our definitions $\mathbf{E} = E_x\hat{\mathbf{x}} + E_z\hat{\mathbf{z}}$ and $\mathbf{B} = B\hat{\mathbf{z}}$ and it is desirable to obtain a general form for any electric and magnetic fields. To do this, suppose we average the motion over one period of the gyration. The contributions to the motion due to the gyration will cancel while the contributions from the drift will not. If we then define $\mathbf{v}_{gc}$ to be the "guiding center" velocity averaged in this way, it is clear that, because the drift velocity is constant,

$$m\frac{\mathbf{v}_{gc}}{dt} = 0 = q\left(\mathbf{E} + \mathbf{v}_{gc} \times \mathbf{B}\right)$$

or

$$\mathbf{E} + \mathbf{v}_{gc} \times \mathbf{B} = 0.$$

Crossing this last result with the magnetic field gives

$$\mathbf{E} \times \mathbf{B} + \mathbf{v}_{gc} \times \mathbf{B} \times \mathbf{B} = 0$$

so that by vector identities

$$\mathbf{E} \times \mathbf{B} = \mathbf{B} \times \mathbf{v}_{gc} \times \mathbf{B} = \mathbf{v}_{gc}B^2 - \mathbf{B}(\mathbf{v}_{gc} \cdot \mathbf{B}).$$

The parallel component of $\mathbf{v}_{gc}$ is given by Equation 2.23 and the perpendicular component is

$$\boxed{\mathbf{v}_{\perp_{gc}} \equiv \mathbf{v}_E = \frac{\mathbf{E} \times \mathbf{B}}{B^2}} \tag{2.29}$$

which we call the $\mathbf{E} \times \mathbf{B}$ *drift velocity.*

This is a truly remarkable result in that the drift of a charged particle in the presence of any arbitrary $\mathbf{E}$ and $\mathbf{B}$ is independent of anything having to do with the particle (so long as it is charged). In particular, the drift does not depend on the magnitude or sign of the particle's charge, its mass, or its velocity. Thus, both ions and electrons will drift together and will generate zero net current. If one therefore seeks to drive a current in a plasma, it cannot be done with the $\mathbf{E} \times \mathbf{B}$ drift alone. Given a particle's initial conditions, its gyroradius will vary with its mass, with $\mathbf{v}_\perp$, and with the magnitude of its charge. The sense of rotation will vary with the sign of its charge, but so long as the Lorentz force is the only force acting on it, the particle will drift without fail in the direction of $\mathbf{E} \times \mathbf{B}$ at the uniform speed given by the magnitude of $\frac{\mathbf{E}\times\mathbf{B}}{B^2}$.

2.2.3 General Drift Equation

Playing one more mathematical trick will allow us to extend the results we obtained for the $\mathbf{E} \times \mathbf{B}$ drift to the case where the drift is caused by any general force $\mathbf{F}$. The Lorentz force can be written as $\mathbf{F} = \mathbf{F}_E + \mathbf{F}_B$ where $\mathbf{F}_E = q\mathbf{E}$ and $\mathbf{F}_B = q\mathbf{v} \times \mathbf{B}$. Suppose we began the derivation of the $\mathbf{E} \times \mathbf{B}$ drift with $\mathbf{F}$ in place of $\mathbf{F}_E$. The only change would be that, in the end, the electric force $\mathbf{F}_E = q\mathbf{E}$ would be replaced by the general force $\mathbf{F}$. That is, instead of the result

$$\mathbf{v}_E = \frac{q\mathbf{E} \times \mathbf{B}}{qB^2} = \frac{\mathbf{F}_E \times \mathbf{B}}{qB^2}$$

we would instead obtain

$$\boxed{\mathbf{v}_{drift} = \frac{\mathbf{F} \times \mathbf{B}}{qB^2}.} \tag{2.30}$$

This will prove to be a very useful result because it can be used to find the drift resulting from any arbitrary force $\mathbf{F}$. Suppose, for example, that we are interested in the drift resulting from the gravitational force $\mathbf{F}_{grav} = m\mathbf{g}$. We have then

$$\mathbf{v}_{grav} = \frac{m}{q}\frac{\mathbf{g} \times \mathbf{B}}{B^2}$$

which is similar to the $\mathbf{E} \times \mathbf{B}$ drift but different in a few important apsects. This drift is in the direction perpendicular to both the gravitational and magnetic fields and its magnitude depends on both the particle's mass and charge. Thus, ions will drift in one direction at a certain speed while electrons will drift in the opposite direction at a different and a net current will be produced.

2.2.4 Orbit Theory

The cases dealt with above where $\mathbf{E}$ and $\mathbf{B}$ were constant were the "easy" ones and we must now move on to more complicated situations where, promped by the geometries present in Figure 1.1, the magnetic field is not constant in space. We will treat three such cases:

1. There is a gradient in the magnetic field strength in the direction *perpendicular* to the field: $\nabla B \perp \mathbf{B}$. (e.g., as seen in Figure 1.1, on the equatorial plane the field strength increases with decreasing distance from Earth).

2. The magnetic field is curved (e.g., all the field lines shown in Figure 1.1).

3. There is a gradient in the magnetic field strength in the direction *parallel* to the field: $\nabla B \parallel \mathbf{B}$ (e.g., as seen in Figure 1.1, as one follows a field line from the equator to the Earth, the field strength increases).

For these three cases, exact equations of motion are too complicated to derive and we will instead obtain approximate solutions using *orbit theory* that involves averaging quantities over one gyration as we did to obtain the final form of the $\mathbf{E} \times \mathbf{B}$ drift.

Case 1: ∇B Drift

Suppose $\mathbf{E} = 0$ and there is a perpendicular gradient in the field strength (meaning the field strength has a gradient in the direction perpendicular to the field). A charged particle with some initial perpendicular velocity will travel through regions of changing magnetic field strength and we wish to approximate the resulting motion. As we did with the $\mathbf{E} \times \mathbf{B}$ drift, let us first reason through the situation before deriving the result.

Figure 2.4 shows an ion and an electron with initial positions indicated by the large dots and having initial velocities directed up the page. The directions of the magnetic field and its gradient are indicated on the figure. Consider first the ion motion. As it travels up the figure, it will begin to gyrate due to the magnetic force but, as it does so, it will move into regions of higher field strength. Now, this field cannot change the particle energy (*i.e.,* its speed) but it does change the direction. But notice that as it executes cyclotron motion, its gyroradius will change such that it is smaller towards the top of the figure where the field strength is higher and larger towards the bottom of the figure where the field strength is weaker (recall that the gyroradius is $r_L = \frac{mv_\perp}{|q|B}$). As a result, the ion will drift to the *right* in the direction of $\mathbf{B} \times \nabla B$. The electron, on the other hand, with its opposite sense of gyration, will drift to the *left* in the direction of $-\mathbf{B} \times \nabla B$. We therefore expect that, in the presence of a magnetic field with a perpendicular gradient, a drift will result that depends on the sign of the particle's charge. This is a very important difference from what we learned about the $\mathbf{E} \times \mathbf{B}$ drift (which is independent of charge). The ∇B drift, as it is called, apparently can drive a current in the $\mathbf{B} \times \nabla B$ direction.

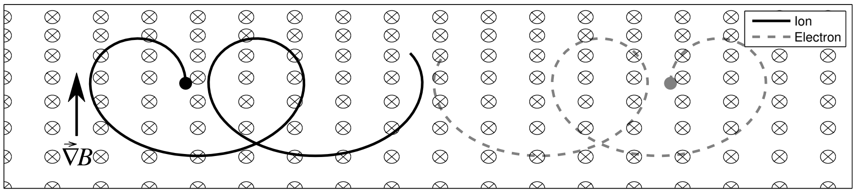

Figure 2.4: The paths of an ion and an electron started with an initial velocity directed toward the top of the figure. The initial position of each particle is indicated by a large dot. The magnetic field is directed into the page and there is a gradient in the field strentgh toward the top of the page. The ion and electron paths are *not* drawn to scale.

To derive an expression for this ∇B drift, let us take $\mathbf{B} = B\hat{\mathbf{z}}$ as before and assume that the gradient in B is along the $y-$direction. The force on a charged particle is

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B}$$

so that

$$F_x = qv_yB \tag{2.31}$$

and

$$F_y = -qv_xB. \tag{2.32}$$

We wish to employ orbit theory wherein the particle motion is averaged over a gyration and our goal will be to obtain an average force that can be inserted into the general drift equation. Quantities averaged over one gyration will be indicated by an overbar as, for example, $\overline{F}_y$.

We see from Equations 2.31 and 2.32 that the forces in the $x-$ and $y-$directions are cross-coupled to the velocity components. Noticing from Figure 2.4 that there is apparently no net motion in the $y-$direction perpendicular to $\mathbf{B}$ but parallel to its gradient, we suppose that the average velocity in that direction will be zero. The average force in the $x-$direction will therefore be zero.

The average force in the $y-$direction is, however, complicated by the presence of the drift that makes $\overline{v}_x \neq 0$. The best we can do here is to approximate $\overline{F}_y$ as follows. Let us assume that B does not change significantly over a gyration. We can then say that $\frac{r_L}{L} << 1$ where L is a parameter that defines the distance over which B changes appreciably. If this condition holds, we can approximate v_x as the $x-$component of the velocity for a particle gyrating in a constant magnetic field as given by the real part of Equation 2.18. That is, we can take

$$F_y = -qv_\perp \cos(\omega_c t)B.$$

Now, we have assumed that B does not change *much* over the gyration, but to be sure it does change *some* (otherwise ∇B would be zero). We can approximate B at any point in the gyration by Taylor expanding about the initial point that we may take to be $x = y = z = 0$. To first order, this gives

$$\mathbf{B} = \mathbf{B}_0 + x\frac{\partial \mathbf{B}}{\partial x} + y\frac{\partial \mathbf{B}}{\partial y} + z\frac{\partial \mathbf{B}}{\partial z}$$

or

$$\mathbf{B} = \mathbf{B}_0 + (\mathbf{r}\cdot\nabla)\,\mathbf{B}$$

where $\mathbf{B}_0$ is the magnetic field at the origin. Because our field has only a $z-$component and the gradient in B is along the y-axis, we have

$$B = B_0 + y\frac{\partial B}{\partial y}.$$

The y term in this expression can be approximated by the real part of Equation 2.21 and the force in the $y-$direction is then

$$F_y = -qv_\perp B_0 \cos(\omega_c t) \mp qv_\perp r_L \frac{\partial B}{\partial y}\cos^2(\omega_c t). \qquad (2.33)$$

We are finally in a position to determine the average force to be substituted into the general drift equation. Averaging the first term on the RHS of Equation 2.33 yields zero and, because the average of $\cos^2(\omega_c t)$ over a gyration is 1/2, we obtain the average force as

$$\overline{F}_y = \mp\frac{1}{2}qv_\perp r_L\frac{\partial B}{\partial y}$$

and the guiding center drift is therefore

$$\mathbf{v}_{gc} = \frac{\mathbf{F}\times\mathbf{B}}{qB^2} = \frac{\overline{F}_y}{qB}\hat{\mathbf{x}} = \mp\frac{1}{2}\frac{v_\perp r_L}{B}\frac{\partial B}{\partial y}\hat{\mathbf{x}}.$$

Our derivation assumed that $\mathbf{B} = B\hat{\mathbf{x}}$ and that the direction of ∇B is along the $y-$axis. Generalizing the solution to our particular case, we can write the result for arbitrary $\mathbf{B}$ and ∇B:

$$\boxed{\mathbf{v}_{\nabla B} = \pm\frac{1}{2}v_\perp r_L \frac{\mathbf{B}\times\nabla B}{B^2}} \tag{2.34}$$

which we call the *grad-B* or ∇B *drift*. There are several important points to note about this drift. First note that no electric field is required to produce it - it results wholly due to the change in B over a gyration. Second, the ± signs indicates that ions and electrons drift in opposite directions. As we noticed before, the ∇B drift will drive a current. Third, the direction of the drift is perpendicular to both the magnetic field and the direction of its gradient. Lastly, do not forget that this result is an approximation based in two instances on the assumption of a circular gyration. The validity of Equation 2.34 is tied to our assumption that $\frac{r_L}{L} << 1$ where L is the length scale of ∇B.

Case 2: Curvature Drift

Consider now the motion of a charged particle in the presence of a constant, curved magnetic field. The particle may have some parallel velocity along a field line and some perpendicular velocity that tends to make it gyrate around the field line. As the particle travels along the field line, it will experience a centripetal acceleration of magnitude $a_R = v_\parallel^2/R_c$ where R_c is the radius of curvature of the field line. The direction of this acceleration will be towards the center of curvature. We may say then that a noninertial centrifugal force

$$\mathbf{F}_R = \frac{mv_\parallel^2}{R_c}\hat{\mathbf{r}} = mv_\parallel^2\frac{\mathbf{R}_c}{R_c^2}$$

is felt by the particle where $\mathbf{R}_c$ is a vector from the center of curvature to the particle's guiding center. The *curvature drift* will then be given by

$$\boxed{\mathbf{v}_R = \frac{\mathbf{F}_R\times\mathbf{B}}{qB^2} = \frac{mv_\parallel^2}{qB^2}\frac{\mathbf{R}_c\times\mathbf{B}}{R_c^2}.} \tag{2.35}$$

Let us put off discussing this result for a moment and realize instead that this result is often incomplete by itself. It is incomplete whenever and whereever the current density is zero because, in such a situation, any curved magnetic field will have $\nabla B \neq 0$: the field strength will get weaker with increasing R_c.

To modify Equation 2.35 in a way that takes this combined effect into account, let us take a cylindrical coordinate system in which $\hat{\mathbf{r}}$ is along $\mathbf{R}_c$, the field line is in the $\hat{\phi}$ direction and the current equals zero. The curl of $\mathbf{B}$ would then be in the $\hat{\mathbf{z}}$ direction and, from Ampere's law with no current, will equal zero. The $z-$component of the curl is then

$$(\nabla \times \mathbf{B})_z = \frac{1}{r}\frac{\partial}{\partial r}\left(rB_\phi\right) = 0$$

and integrating this equation reveals that B_ϕ will be inversely proportional to r. It must be then that $|\mathbf{B}|$ is proportional to the inverse of R_c so that

$$\frac{\nabla B}{B} = -\frac{\mathbf{R}_c}{R_c^2}.$$

Substituting this result into Equation 2.34 gives the ∇B contribution to the total drift:

$$\mathbf{v}_{\nabla B} = \mp\frac{v_\perp r_L}{2B^2}\mathbf{B} \times \left(B\frac{\mathbf{R}_c}{R_c^2}\right) = \frac{m}{2q}v_\perp^2\frac{\mathbf{R}_c \times \mathbf{B}}{R_c^2 B^2} \tag{2.36}$$

and the total drift is then obtained by adding Equations 2.36 and 2.35:

$$\boxed{\mathbf{v}_{R+\nabla B} = \frac{m}{q}\frac{\mathbf{R}_c \times \mathbf{B}}{R_c^2 B^2}\left(v_\parallel^2 + \frac{1}{2}v_\perp^2\right)} \tag{2.37}$$

This curvature plus grad-B drift expression applies to particles drifting in curved magnetic fields where the current density is zero.

We can rewrite Equation 2.37 in a handy way using the equipartition theorem which states that, for a Maxwellian distrubution, each degree of freedom acquires $\frac{1}{2}kT$ of thermal energy. There is one degree of freedom along a magnetic field line and two degrees of freedom perpendicular to it. Thus

$$\frac{1}{2}mv_\parallel^2 = \frac{1}{2}kT$$

and

$$\frac{1}{2}mv_\perp^2 = kT$$

so that

$$\boxed{\mathbf{v}_{R+\nabla B} = \pm\frac{\mathbf{R}_c \times \mathbf{B}}{\omega_c R_c^2 B}\frac{2kT}{m} = \pm\frac{\mathbf{R}_c \times \mathbf{B}}{\omega_c R_c^2 B}v_{th}^2} \tag{2.38}$$

where $v_{th} = \sqrt{2kT/m}$ is the thermal speed.

This $\nabla B + \mathbf{R}_c$ (∇B plus curvature) drift is in the same direction as ∇B and depends on the sign of the charge but, as can be shown, not the particle mass. It will therefore drive a current and, as we will see in a later chapter, will play an important role in determining the motion of charged particles trapped in Earth's magnetic field.

Case 3: $\nabla B \parallel \mathbf{B}$ (Magnetic Bottles)

The final case of single-particle motion we will treat is that of a particle trapped in a magnetic bottle or mirror. The term "magnetic bottle" implies a magnetic field geometry where the field strength increases with position along the field line as illustrated in Figure 2.5.

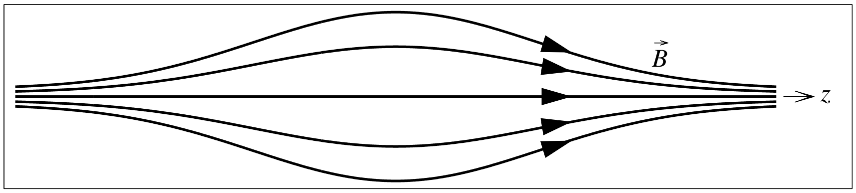

Figure 2.5: A magnetic bottle showing the converging field lines in a geometry where $\nabla B \parallel \mathbf{B}$.

A particle traveling on the central axis of the bottle will have a kinetic energy given by $K = \frac{1}{2}mv_{\|}^2 + \frac{1}{2}mv_{\perp}^2$. That is, there are contributions to the kinetic energy from the velocity in the direction parallel to the field and in the directions perpendicular to the field. The parallel component of the velocity makes the particle travel along the field line and the perpendicular components result in gyration around the field line. As the particle travels to the right from the center of the bottle, it will encounter increasingly stronger magnetic fields and the perpendicular contribution to the kinetic energy will increase as the particle gyrates faster and faster around the field line. However, this increase in the perpendicular contribution to the kinetic energy must come at the expense of the parallel contribution since the total energy cannot change (the magnetic field does no work on the particle). The parallel component of the velocity must then decrease as the particle moves into regions of increasing B. At some point, the parallel component of the velocity will be decreased to zero and the particle will stop translating and simply gyrate around the field line. As we will see below, it turns out that the parallel force responsible for stopping the translation will still be active at the instant translation stops and the particle will be reflected back to regions of decreasing B toward the center of the bottle. The particle will then travel to the left end of the bottle where the situation will repeat. The particle will bounce from end to end, gyrating as it goes and is trapped inside the magnetic bottle. This trapping is not perfect (some particles will escape) but let us defer discussion of the escaping particles until later.

Turning our attention to a derivation of the force responsible for this trapping, we first assume that our bottle is cylindrically symmetric. That is, $B_\phi = 0$ and there is no variation in $\mathbf{B}$ as we circulate around the bottle's central axis. Given the coordinate system shown in Figure 2.5, the trapping force must act along the $z-$axis and we find from Equation 2.13 that this component of force is

$$F_z = q\left(v_r B_\phi - v_\phi B_r\right).$$

B_ϕ is zero by assumption so the first term on the RHS vanishes. As can be seen in Figure 2.5, the field lines converge to the central axis so that $B_r \neq 0$.

To find an approximate expression for B_r, we can impose $\nabla \cdot \mathbf{B} = 0$ on our field in cyclindrical coordinates to find

$$\frac{1}{r}\frac{\partial}{\partial r}\left(rB_r\right) + \frac{1}{r}\frac{\partial B_\phi}{\partial \phi} + \frac{\partial B_z}{\partial z} = 0. \tag{2.39}$$

The second term on the LHS is zero by our assumptions and we are left with the job of using the other two terms to determine B_r and the trapping force.

Suppose we know $\left(\frac{\partial B_z}{\partial z}\right)_{r=0}$ which specifies how the $z-$component of the field varies with position along the bottle and that there is not *too much* variation in this component as we move off the central axis of the bottle (*i.e.,* $\frac{\partial^2 B_z}{\partial r \partial z} \approx 0$). In this case, integrating Equation 2.39 yields

$$rB_r = -\int_0^r r\frac{\partial B_z}{\partial z}dr \approx -\frac{1}{2}r^2\left(\frac{\partial B_z}{\partial z}\right)_{r=0}$$

and

$$B_r = -\frac{1}{2}r\left(\frac{\partial B_z}{\partial z}\right)_{z=0}.$$

The trapping force is then

$$F_z = \frac{1}{2} q v_\phi r \left(\frac{\partial B_z}{\partial z} \right)_{r=0}$$

and can be averaged over one gyration using $\overline{v}_\phi = \mp v_\perp$ (ions will gyrate in the $-\phi-$direction and electrons will gyrate in the $+\phi-$direction) and $\overline{r} = r_L$ to obtain

$$\overline{F}_z \approx \mp q v_\perp r_L \frac{\partial B_z}{\partial z} = \mp \frac{1}{2} \frac{m v_\perp^2}{B} \frac{\partial B_z}{\partial z}.$$

This result can be written in a more convenient and general form if we identify $\nabla_\| B = \frac{\partial B_z}{\partial z}$ as the parallel gradient and realize that the magnitude of the magnetic moment of the gyrating particle is

$$\boxed{\mu = IA = \frac{\frac{1}{2} m v_\perp^2}{B}} \tag{2.40}$$

where I is the current resulting from the gyromotion and A is the area of a circle with radius r_L. With these definitions we may write the $z-$component of the average trapping force as

$$\boxed{\mathbf{F}_\| = -\mu \nabla_\| B.} \tag{2.41}$$

This force is commonly known as the *mirror force* and depends on two quantities: the particle's magnetic moment and the parallel gradient of the magnetic field strength. Note that it is a *restoring force* in the sense that it always points to the center of the bottle, opposite to the direction of increasing field strength. It is this force that is responsible for the bouncing motion of a charged particle trapped in a magnetic bottle.

In §2.3.2 we will discuss magnetic bottles in more detail but let us now take a brief aside and do our first real bit of space physics.

2.3 Periodic Motions in a Dipole Field

As we will see in Chapter 4, Earth's magnetic field can be approximated (although often not very accurately) by that of a magnetic dipole. Figure 2.6 shows a dipole field viewed from a location in Earth's orbital plane and in this section we will consider the periodic motions that result from the interactions of charged particles with such a magnetic field.

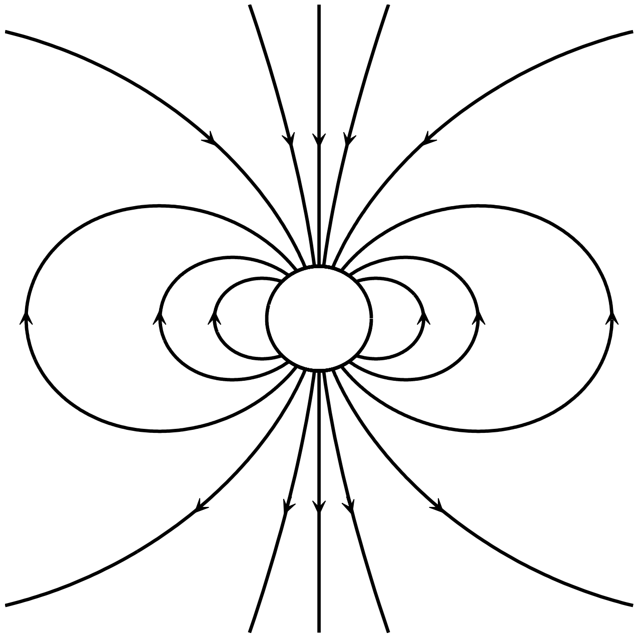

Figure 2.6: A dipole approximation to Earth's magnetic field, viewed from the side.

Suppose a charged particle is located on Earth's equatorial plane on one of the field lines shown in Figure 2.6. The particle will have a velocity given by $\mathbf{v} = \mathbf{v}_\perp + v_\parallel \hat{\mathbf{B}}$. As we have noted before, the perpendicular components of the velocity will cause the particle to gyrate around the field line and the parallel component will cause it to translate along the field line. We can immediately identify the first of what will become three periodic motions that occur: the gyration about a field line.

To identify the second periodic motion, notice that as the particle translates along the field line towards one of Earth's poles, it encounters regions of increasingly strong magnetic field. It is as if the magnetic bottle shown in Figure 2.5 has been bent into the curved shape of a dipole with the center of the bottle at Earth's equator and the two ends at its poles. Just as a charged particle was trapped in the magnetic bottle and bounced from end to end in response to the trapping force, a particle in Earth's dipole field will bounce from pole to pole as it is trapped in the dipole field. This bounce motion is the second periodic motion and the bouncing particles can be at least loosely identified with the famous VanAllen radiation belts.

The third and last periodic motion is easiest to identify if we suppose that $v_\parallel = 0$ when the particle is in the equatorial plane. The particle will still gyrate due to its $v_\perp$ but it will not translate along the field line and will therefore stay in the equatorial plane as it executes cyclotron motion. But notice that there is a gradient to the magnetic field strength in the direction perpendicular to the field. ∇B is directed inward toward Earth and the particle will therefore experience a ∇B drift in the $\pm\mathbf{B} \times \nabla B$ direction depending on the sign of its charge. Ions will drift to the west and electrons will drift to the east while remaining a constant average distance from Earth. A westward current known as the *ring current* will therefore encircle the Earth. This ∇B drift in a circle around Earth is the third periodic motion. Of course, particles with nonzero $v_\parallel$ will also undergo a $\nabla B + \mathbf{R}_c$ drift around Earth as they gyrate and bounce on the curved field lines. In this more realistic case, the ring current will no longer be confined to the equatorial plane and will be distributed over a broad range of latitudes.

> The three periodic motions in a dipole field can be succinctly listed as: gyration, bounce, and drift. Under typical conditions, there are many gyrations per bounce period and many bounces per drift period.

2.3.1 Geomagnetic Storms and the *Dst* Index

Figure 2.7 shows a plot of the *Dst* index from August 1998. The *Dst* or "Disturbance storm time" index is essentially the hourly deviation from average of the low- to mid-latitude northward-pointing component of Earth's magnetic field[22]. In this figure, 7 days of the index are plotted. During geomagnetically quiet conditions, the *Dst* index takes a nearly constant value but, when there is a geomagnetic "storm" caused by enhanced activity on the Sun, the *Dst* index varies in a characteristic way as shown in the figure.

Given what we know about particle trapping and periodic motions in a dipole field, we are in position to understand the most obvious feature seen in Figure 2.7 which is the large and relatively sudden decrease in the northward component of Earth's magnetic field that occurs near the middle of August

[22] The *Dst* index is freely available from the World Data Center for Geomagnetism at Koyoto University, Japan: `http://swdcwww.kugi.kyoto-u.ac.jp/dstdir/index.html`

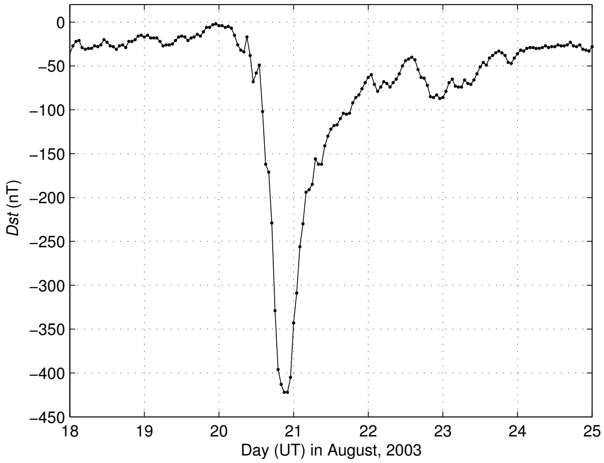

Figure 2.7: Dst index from November 2003 showing the effect of a geomagnetic storm on Earth's magnetic field.

20, 2003. The more gradual recovery back to pre-storm values will soon lead us farther into a discussion of trapping charged particles in a magnetic bottle.

To understand the sudden drop in the Dst index, let us first recall that the ring current is a westward-flowing current that encircles the Earth due to the $\nabla B + \mathbf{R}_c$ drift of ions (to the west) and electrons (to the east). This ring current will produce its own magnetic field that will combine with Earth's field to yield the net or total magnetic field observed at any point. At the equator, Earth's approximately dipole field points due *north* as shown in Figure 2.6 but the field due to the westward-flowing ring current points due *south* by the right-hand rule for currents so that it *weakens* the total observed field. Thus, when the ring current is increased, the observed northward-pointing component of the total field is decreased and the Dst index consequently takes on negative values. So we may associate negative deviations in Dst with a strengthening of the ring current[23].

This increase in the ring current occurs in response to "storms" on the Sun such as solar flares or coronal mass ejections during which the Sun ejects

[23] As it turns out, the Dst index is not influenced by the ring current alone. Other (in fact, *any* other) currents flowing in space will also contribute to the Dst index. However, it can be shown that the ring current generally provides the dominant contribution.

many more charged particles that can become trapped in Earth's magnetic field. These trapped particles gyrate, bounce and drift and significantly inrease the $\nabla B + \mathbf{R}_c$ drift current that produces the ring current. The Dst index decreases in response to this increasing ring current.

We may then ask: once the ring current is increased due to a larger-than-average number of trapped particles drifting in Earth's field, for what reason does it gradually return to pre-storm levels as seen in Figure 2.7? The answer lies in the fact that the trapping is not perfect and ring current particles slowly leak out of the magnetic bottle. To investigate this imperfect trapping, let us introduce the very useful idea of adiabatic invariants.

2.3.2 Adiabatic Invariants

We know from classical mechanics that any periodic motion undergoing adiabatic change[24] has a conserved quantity associated with it. This conserved quantity is the *action integral* defined as $I = \oint pdq$ where p and q are the generalized momentum and coordinate associated with the periodic motion. Such a conserved quantity is called an *adiabatic invariant* and there must be one associated with each of the gyration, bounce and drift motions described above.

1^{st} Adiabatic Invariant, μ

The first of our periodic motions is the cyclotron gyration around a field line due to the Lorentz force and the perpendicular component of a particle's velocity. To compute the action integral and evaluate the adiabatic invariant, we must first identify the approproiate generalized coordinate and momentum. The gyration path is a circle and so the appropriate generalized coordinate is the angular position ϕ so that $dq = d\phi$. The Lagrangian for a charged particle gyrating in a magnetic field is $L = \frac{1}{2}mv_\perp^2 + \frac{1}{2}mv_{||}^2 + q\mathbf{v}\cdot\mathbf{A} = \frac{1}{2}mr_L^2(\dot{\phi})^2 + \frac{1}{2}mv_{||}^2 + qr\dot{\phi}A_\phi$ where $\mathbf{A}$ is the the magnetic vector potential[25].

[24] By *adiabadic change*, we mean that the change in the period of the motion over one cycle is small compared to the period.

[25] Recall that the magnetic vector potential is $\mathbf{B} = \nabla \times \mathbf{A}$. If you have not previously encountered a Lagrangian that includes a magnetic potential energy term and are wondering how it makes its way into the expression (after all, the magnetic force does no work, so how can it have a potential energy?), see any of the good textbook or web references including, *e.g.*, `http://iweb.tntech.edu/murdock/ph4610/magfldlag.pdf`.

The generalized momentum associated with the ϕ coordinate is

$$p_\phi = \frac{\partial L}{\partial \dot{\phi}} = m v_\perp r_L + q r A_\phi.$$

Under adiabatic changes, the action integral is invariant and is given by

$$\oint p_\phi d\phi = \oint m v_\perp r_L d\phi + q \oint r A_\phi d\phi.$$

The RHS may be rewritten as

$$\begin{aligned}
\oint p_\phi d\phi &= \oint m v_\perp r_L d\phi + q \oint \mathbf{A} \cdot dl \\
&= 2\pi m v_\perp r_L + q \int\int (\nabla \times \mathbf{A}) \cdot d\mathbf{S} \\
&= 2\pi m v_\perp r_L + q \int\int \mathbf{B} \cdot d\mathbf{S} \\
&= 2\pi m v_\perp r_L + q B \pi r_L^2 \qquad (2.42)
\end{aligned}$$

where we have assumed that B is constant over the area $\mathbf{S}$ enclosing the path and that $\mathbf{B} \cdot d\mathbf{S}$ is positive. This result can be be manipulated by recalling that $r_L = v_\perp/\omega_c$ and $\omega_c = \pm |q| B/m$ to quickly find that

$$\oint p dq = \pm 6\pi \frac{m}{|q|} \frac{\frac{1}{2} m v_\perp^2}{B}$$

or, recalling Equation 2.40,

$$\oint p dq = \pm 6\pi \frac{m}{|q|} \mu. \qquad (2.43)$$

Now, since the action integral is constant and the particle mass and charge are constant,

> the magnetic moment μ must be constant and we identify it as the *1st adiabatic invariant.*

The magnetic moment is the invariant associated with cyclotron motion but we must not forget that it is an *adiabatic* invariant. That is, μ is conserved only if the quantities on which it depend vary slowly during a gyro-period.

However, it can be shown that μ is a *strong* invariant so that even if $v_\perp$ or B vary with a frequency comparable (but still less than) the gyro-period, μ is still reasonably conserved. But to be sure, if the magnetic field for example is doubled or triped during one quarter of a gyro-period, μ will not be conserved.

Conserved quantities are extrememly useful in determining the motion of particles. Take for example the conservation of energy studied and used extensively in introductory physics courses. The fact that total energy is conserved makes it very easy to find the speed of a particle after it has passed through some change in potential energy. For example, one can find with only a few lines of algebra what will be the speed of a particle at all points as it oscillates on a spring. And, significantly, one can do this without integrating any equations of motion (assuming the potential associated with the spring force is known). Conservation laws give us a very useful alternate approach that we can use here to study in more detail the trapping of charged particles in a magnetic bottle.

Approximately half of the magnetic bottle of Figure 2.5 is shown again in Figure 2.8 but this time with the addition of some markers identifying the locations of the strongest and weakest magnetic fields and the velocity of a particle at the center of the bottle. Let us introduce a notation wherein quantities evaluated at the center of the bottle where the field is weakest are given a subscript of “0” and quantities evaluated at the location in the bottle where the field is strongest are given a subscript of “m”. Thus the particle velocity at the location shown in the figure is $\mathbf{v}_0$ and the field strength there is B_0. This velocity $\mathbf{v}_0$ has components parallel and perpendicular to $\mathbf{B}$ as indicated and we can define the *pitch angle* there as

$$\theta_0 = \tan^{-1}\left(\frac{v_{\perp 0}}{v_{\| 0}}\right). \tag{2.44}$$

A particle with a 90° angle will therefore have no parallel velocity and will not translate along the bottle but will gyrate about a fixed point at the bottle's center. At the other extremes, a particle with a 0° or 180° pitch angle will translate along the length of the bottle[26] but will not gyrate due to its absence of any perpendicular velocity.

As the particle moves within the bottle, bouncing and possibly escaping, its kinetic energy given by $K = \frac{1}{2}mv_\perp^2 + \frac{1}{2}mv_\|^2$ will remain constant since the

[26] By definition, a pitch angle of 0° corresponds to motion in the direction of $\mathbf{B}$ and a pitch angle of 180° corresponds to motion in the direction opposite to $\mathbf{B}$.

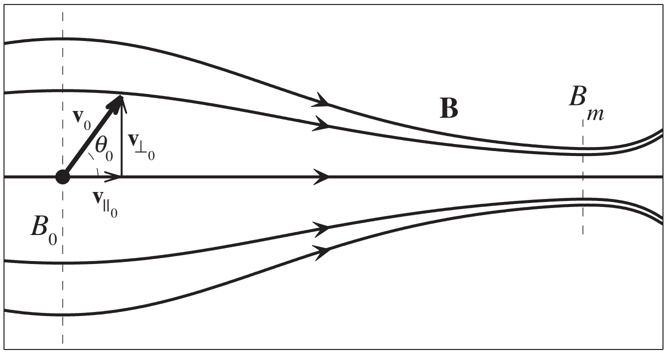

Figure 2.8: Approximately half of a magnetic bottle. The vertical dashed lines indicate locations of the weakest and strongest magnetic fields. A charged particle with some velocity $\mathbf{v}_0$ is located at the center of the bottle where B is weakest.

field does no work on it. Substituting $\frac{1}{2}mv_\perp^2 = \mu B$ and solving the above equation for the kinetic energy, we find that

$$\mu = \frac{K - \frac{1}{2}mv_\|^2}{B}.$$

This equation is interesting because it contains two quantities that are constant as the particle executes its motion within the bottle and two quantities that vary with its motion. Because μ and K are both constant, we see that as the particle moves into regions of increasing magnetic field strength, the parallel component of its velocity must decrease so that the RHS remains constant. On the other hand, as the particle moves into regions of lower magnetic field strength, the parallel component of its velocity must increase so the RHS remains constant. The result is the bouncing, trapped motion we have already discussed but is now viewed in the light of conserved quantities rather than that of the trapping force given by Equation 2.41.

The trapping force equation does, however, provide us with one additional insight. Suppose a particle at the center of the bottle has no perpendicular velocity and is only translating to the right along the central field line. In such a case, its magnetic moment μ and its pitch angle θ_0 will be zero and there will be *no* trapping force. The particle will continue moving unimpeded to the right at a constant speed and will escape the bottle. We may suspect then that even particles with a non-vanishingly small pitch angle may also escape. We suspect the magnetic bottle is leaky and we wish to determine how large the pitch angle may be before the leak is "plugged". To do this, we will return to an approach based on conserved quantities.

Suppose that, as the pitch angle is decreased from 90° and the particles come closer and closer to escaping the bottle, we identify the "last-trapped" particle and label its pitch angle as θ_l. Then for this particle, the turning point must be located where B takes on its largest value. That is, for this last-trapped particle, $v_\| = 0$ where $B = B_m$. Because the kinetic energy of this particle is a constant, we can equate K_0 with K_m. That is, its kinetic energy at the center of the bottle must equal its kinetic energy at the turning point. We have then

$$\frac{1}{2}mv_{\perp_0}^2 + \frac{1}{2}mv_{\|_0}^2 = \frac{1}{2}mv_0^2 = \frac{1}{2}mv_{\perp_m}^2$$

since $v_{\|_m} = 0$. The magnetic moment at the turning point where $B = B_m$ is then

$$\mu = \frac{\frac{1}{2}mv_{\perp_m}^2}{B_m} = \frac{\frac{1}{2}mv_0^2}{B_m}$$

which must equal μ at the center of the bottle by conservation of μ. Thus

$$\mu = \frac{\frac{1}{2}mv_0^2}{B_m} = \frac{\frac{1}{2}mv_{\perp_0}^2}{B_0}$$

and we find for the last-trapped particle that

$$\frac{B_0}{B_m} = \frac{v_{\perp_0}^2}{v_0^2}.$$

But for this particle, $v_{\perp_0} = v_0 \sin\theta_l$ so that

$$\frac{B_0}{B_m} = \sin^2\theta_l.$$

Particles with pitch angles smaller than θ_l will escape the bottle while particles with pitch angles greater than θ_l are trapped in it. If we define a *mirror ratio* as

$$R_m = \frac{B_m}{B_0} \tag{2.45}$$

or the ratio of the strongest to the weakest magnetic field strengths in the bottle, we find that particles can escape the bottle if

$$\sin^2\theta_0 < \frac{1}{R_m}. \tag{2.46}$$

That is, as the ratio of the strongest to the weakest field becomes smaller, the mirror ratio decreases and the bottle becomes more and more "open" and leaky. A very large mirror ratio means the bottle is tightly closed and very few particles will escape. Those particles that do escape are said to be in the *loss cone* formed by rotating the vector $\mathbf{v}_0$ for pitch angle θ_l about the bottle's central field line. Any particle with a velocity vector $\mathbf{v}_0$ lying within this loss cone will escape the bottle and be lost.

Evidence of this loss cone is commonly observed in Earth's dipole magnetic bottle where the location of the strongest field is taken at the point where the charged bouncing particles penetrate far enough along the field lines to encounter the polar atmosphere. Deep in the atmosphere, there is a very high probability that the charged particles will collide with the abundant neutrals, loose their kinetic energy, and "escape" the bottle. Figure 2.9 shows 20 minutes of data from orbit number 9257 of the Fast Auroral SnapshoT (FAST) satellite[27]. The axes show pitch angle versus time and the grayscale indicates the energy flux (approximately proportional to number density) of 0.1-1 keV electrons. Note the near absence of particles with pitch angles near 0°. These particles have travelled nearly along a field line, penetrated deep into the atmosphere due to their small μ and the consequently small trapping force, and been lost.

As another example showing the usefullness of the first adiabatic invariant μ, consider Figure 2.10[28] that shows a diagram of the VASIMR (VAriable Specific Impulse Magnetoplasma Rocket) engine being developed by the Ad Astra Rocket Company and the Johnson Space Center. In this engine, a plasma is generated and confined by magnetic fields but the magnetic field geometry is such that the field diverges and becomes progressively weaker with increasing position along the ejection nozzle.

To accelerate the plasma and provide thrust, the plasma is heated in the perpendicular direction by azimuthal electric fields in resonance with the plasma's ion cyclotron frequency. As a result of this perpendicular heating,

[27] Data from the FAST mission are freely available on the web at: `http://sprg.ssl.berkeley.edu/fast/`.

[28] from: `http://www.daviddarling.info/encyclopedia/V/VASIMR.html`

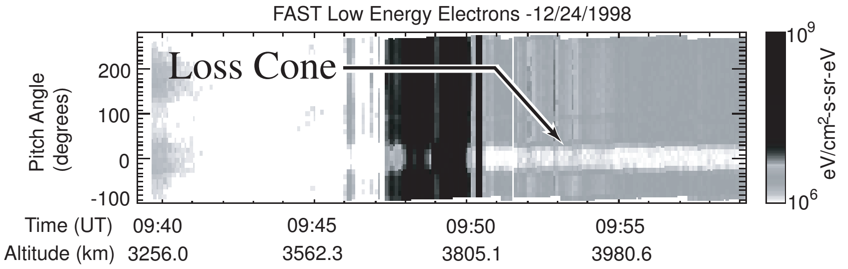

Figure 2.9: Low energy electron data from orbit number 9257 of the FAST satellite showing the presence of a well-defined loss cone.

the ion Larmour radius increases and $\mu = IA = I\pi r_L^2$ tends to increase. But so long as the Larmour radius is changed slowly compared to the ion cyclotron frequency (meaning that the plasma is not heated *too* fast), μ must remain constant. Now, a few lines of algebra will show that μ is proportional to the magnetic flux through a gyro-orbit and thus, as the ions are heated and r_L increases, they are accelerated into the region of increasingly weak magnetic fields in the ejection nozzle. The ejection of these ions at high velocity (and thus high specific impulse) provides the engine with the thrust.

The first adiabatic invariant is the most important and useful of the three since the periodic motion associated with it (cyclotron motion) has periods that are much smaller than those of the bounce and drift periods. We have therefore discussed it at some length and will give a less thorough treatment of the other two invariants.

2$^{\text{nd}}$ Adiabatic Invariant, J

A gyrating particle in Earth's magnetic field will bounce from pole to pole so long as it is not in the loss cone and has some component of its velocity along the field. The second adiabatic invariant is associated with this periodic bouncing and, since the proof is rather lengthy, we omit it here and simply state the result. The second adiabatic invariant is

$$J = \int_{p_1}^{p_2} v_{\|} ds$$

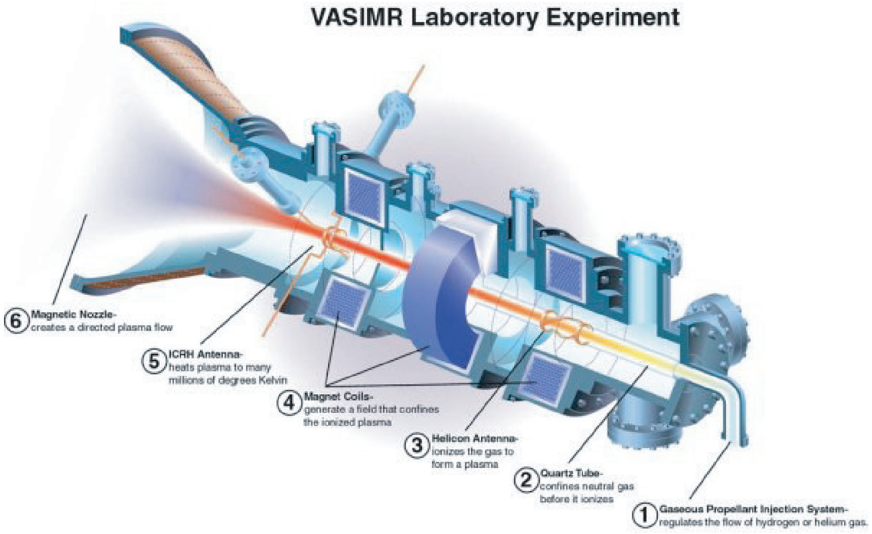
VASIMR Laboratory Experiment
6 Magnetic Nozzle- creates a directed plasma flow
5 ICRH Antenna- heats plasma to many millions of degrees Kelvin
4 Magnet Coils- generate a field that confines the ionized plasma
3 Helicon Antenna- ionizes the gas to form a plasma
2 Quartz Tube- confines neutral gas before it ionizes
1 Gaseous Propellant Injection System- regulates the flow of hydrogen or helium gas.

Figure 2.10: A diagram of the VASIMR engine. Conservation of μ for perpendicularly heated ions provides thrust by accelerating the ions through the ejection nozzle.

where p_1 and p_2 are the turning points at each pole. This result essentially reveals that the length between turning points of a magnetic field line on which a particle is bouncing is a constant.

3$^{\text{rd}}$ Adiabatic Invariant, Φ

Suppose we look down on Earth from above and observe a particle drifting in the circle caused by the $\nabla B + \mathbf{R}_c$ drift which is our third periodic motion. It can be shown that the third adiabatic invariant is simply the flux of magnetic field lines through this drift circle given by

$$\Phi = \oint \mathbf{B} \cdot d\mathbf{s} \tag{2.47}$$

where the area to be integrated over is the area enclosed by the drift circle.

As Exercise 2.11 demonstrates, it can take some days for a particle to complete a full drift circle around Earth. Conditions in space and the field throughout the drift circle often change much more rapidly than this, violating the adiabatic assumption so that Φ is, for most practical uses, not really an invariant at all.

Exercises

2.1: For each of the following cases, calculate the Debye length, plasma parameter and plasma frequency, and use the first two criteria for defining a plasma (assuming the third is automatically satisfied) to state whether or not each case can be considered to be a plasma:

a. Earth's ionosphere: $T_e = 0.08$ eV, $n_e = 1 \times 10^6$ cm^{-3}
(Partial answer: $\lambda_D = 2$ mm, $\Lambda = 10^4$, $\omega_{pe} = 60 \times 10^6$ rad/s (so $f_{pe} = 9$ MHz))

b. Interstellar gas: $T_e = 0.5$ eV, $n_e = 0.1$ cm^{-3}

c. Earth's VanAllen Radiation belts: $T_e = 100$ eV, $n_e = 1 \times 10^3$ cm^{-3}

d. Fusion reactor: $T_e = 2 \times 10^4$ eV, $n_e = 1 \times 10^{14}$ cm^{-3}

e. Typical flame: $T_e = 0.1$ eV, $n_e = 1 \times 10^8$ cm^{-3}

Note: In the field of space physics, you will often encounter "temperatures" given in energy units. When you encounter this, it is to be assumed that Boltzmann's constant k has been absorbed into the "temperature" so that, in this case, T_e is really kT_e. A useful approximation to keep in your head (and with which you may check your conversions) is that $1/40$ eV ≈ 300 K (multiplied by an implicit k).

2.2: In the derivation of the electron plasma frequency, we assumed the ions were stationary because to their mass is much greater than the electron's mass. Derive an expression for the plasma frequency ω_p without assuming stationary ions. Compare this more correct result with the one derived in the text and show they are approximately equal.
(Hint: include the term $n_i = n_0 + \delta n_i$ in Poission's equation and use the ion equations of motion and continuity. You will have five equations and five unknowns but the procedure is identical to that followed in the text.)

2.3: Compute the cyclotron frequency and Larmour radius for the following cases. In each case, take the velocity vector to be perpendicular to the magnetic field. [from Chen, 1983, p.25]

a. A 10-keV electron in Earth's magnetic field of 5×10^{-5} T
(Answer: $\omega_{ce} = 9 \times 10^6$ rad/s, $r_L = 7$ m)

b. A solar wind proton streaming with a speed of 300 km/s in the interplanetary magnetic field of 5×10^{-9} T

c. A 1-keV He^{+} ion in the solar atmosphere near a sunspot where the magnetic field is 5×10^{-2} T

d. A 3.5-MeV He^{++} ash particle in an 8 T fusion reactor

2.4: Beginning with the electron continuity equation, Poisson's equation and the electron momentum equation, show all the steps required to obtain the linearized perturbation Equations 2.9, 2.10 and 2.11.

2.5: Beginning with the linearlized perturbation equations from Exercise 2.4, supply the missing steps required to obtain the plasma frequency given in Equation 2.12.

2.6: In the derivation of the plasma frequency, we assumed oscillating solutions and employed the substitutions

$$\frac{\partial}{\partial t} \to -i\omega$$

$$\nabla \to ik\hat{\mathbf{x}}.$$

Demonstrate the validity of these substitutions.

2.7: Beginning with the two equations immediately preceeding Equation 2.20, use Euler's identify to obtain Equations 2.20 and 2.21.

2.8: Beginning with Equation 2.37, complete the steps required to obtain Equation 2.38.

2.9: Show that, as Equation 2.40 states, $IA = \frac{\frac{1}{2}mv_\perp^2}{B}$ for a particle moving in the presence of a magnetic field. As the text suggests, this can be done by computing the current I due to the charged particle's gyromotion over an area A defined by a circle of radius equal to the Larmour radius r_L.

2.10: For each of the following cases, sketch particle trajectories separately for electrons and protons. Define your coordinate systems and clearly illustrate the direction of the magnetic and electric fields.

a. Assume a static uniform magnetic field oriented along the x-axis with no electric field. Charged particles have initial velocity components of $v_{x_0} = v_{y_0} = 0$ and $v_{z_0} = v_0$.

b. Assume a static uniform magnetic field oriented along the y-axis with a static uniform electric field along the z-axis. Charged particles are initially at rest.

c. Assume a magnetic field along the z-axis that increases in strength with increasing values of z. Charged particles have initial velocity components along the x- and z-axes with $v_{x_0} \approx v_{z_0}$.

2.11: Suppose the magnetic field strength in Earth's magnetic equatorial plane is given by $B = B_0(R_E/r)^3$ where $B_0 = 0.3$ Gauss is the surface equatorial field strength and r is the geocentric distance.

a. Obtain an expression for the drift period (the time it takes a particle to drift around the Earth) of a particle on the equatorial plane with a pitch angle of 90° and energy W.

b. Evaluate this period for both a proton and an electron of 1 keV energy at a distance of $5R_E$ from the center of Earth.

c. Justify the grad-B assumption that $r_L/L << 1$ (where L is a length characterizing the distance over which B changes appreciably) for the cases described in b. above.

In this problem, ignore the curvature drift (using only the grad-B drift). The curvature drift is not actually negligible but, as it turns out, does not seriously impact the numerical answer.

2.12: Complete the missing steps between Eqs. 2.42 and 2.43.

Chapter 3

The Sun and the Solar Wind

3.1 Introduction to the Sun

In 1993, the group *They Might Be Giants* released an album containing the title song, “Why Does the Sun Shine (The Sun Is A Mass of Incandescent Gas)” and the reader who can be persuaded to listen to this song will likely have an amusing and rewarding experience. The Sun is a main sequence star predominantly fueled, as the song lyrically points out, by nuclear fusion of hydrogen into helium. Comparatively speaking, our Sun is an ordinary star but just as ordinary people may effect extraordinary impacts on those closest to them, our Sun is the most significant object in the solar system. This chapter presents an overview of the Sun, its structure and processes, and discusses the solar wind and interplanetary magnetic field (IMF) that fill the space of our solar system and interact with all its celestial objects.

With apologies to those having a keen interest in the worthy field of solar physics, it is admitted at the outset that the focus of this chapter is on those topics that most directly impact our study of the near-Earth space environment. As a result, many interesting and important topics are left untouched and others are merely introduced.

Stars are classified by the Morgan-Keenan system that encodes a star’s temperature (thus its apparent color) and size (an indication of its current point in the life cycle of a star). From the highest to the lowest temperatures, stars are classified as type O,B,A,F,G,K or M. Table 3.1 lists the types and some associated average characteristics relative to values of the Sun (identified by the $\odot$ subscript). Appended to the letter encoding a star’s temperature range is a number that expresses, in tenths, where it lies in that range. For example, the Sun is a type G2 star which means that its approximate surface temperature is two tenths of the way from 6000 K to 5000 K or ~5800 K. Finally, a Roman numeral code from I-V is appended that identifies the approximate size of the star. The largest stars, known as supergiants, are identified as class I and the classification number increases as the star's size decreases. The smallest stars are known as dwarf stars, are on the main sequence and are identified as class V. Our Sun is a spectral type G2V star which means it is a main-sequence star with an approximate surface temperature of 5800 K. Such stars are abundant in our galaxy and Alpha Centauri, the next closest star to us, is also a type G2V star.

Star Type	Approximate Surface Temperature (K)	Mass ($M_\odot$)	Radius ($R_\odot$)	Luminosity ($L_\odot$)
O	>25,000	60	15	1.4×10^6
B	25,000-11,000	18	7	2×10^4
A	11,0000-7500	3.2	2.5	80
F	7500-6000	1.7	1.3	6
G	6000-5000	1.1	1.1	1.2
K	5000-3500	0.8	0.9	0.4
M	<3500	0.3	0.4	0.04

Table 3.1: Average characteristics of stellar spectral types.

Table 3.2 presents a selection of the Sun's "vital statistics"[Beatty et al., 1999, p.25]. To give these numbers perspective, the volume of the Sun is sufficent to contain more than a million Earths and its average density is about one fourth that of Earth and about 40% higher than that of water. The density at the center of Sun is, however, more than 100 times higher than the average. The Sun's radius is about 109 times Earth's radius (R_E) and its mass is 330,000 times that of Earth. The average distance separating the Sun and Earth, defined to be one *astronomical unit* (AU), is ~150 million km which is about 23,500 R_E and photons emitted by the Sun cover this distance in about 8.3 minutes. The Sun's luminosity, which is the total amount of photon energy radiated per second, is a staggering 3.85×10^{26} W and taking humankind's power consumption to be 1.6×10^{13} W[1], we find that in a single second the Sun emits enough photon energy to power humankind for more

[1] see `http://en.wikipedia.org/wiki/World_energy_consumption#Primary_energy`

than 750,000 years. Using Einstein's famous relation $E = mc^2$, we find that the Sun looses more than 4 billion kilograms of mass per second due to its photon radiation (to say nothing of any particles leaving the Sun). It has been loosing mass at essentially this rate for billions of years and will do so for billions more.

> This is worth repeating: for billions of years, the Sun has been loosing billions of kilograms of mass each second, and it will continue doing so for billions of years to come. The Sun is very massive indeed!

Solar Parameter	Value
Age	4.5×10^9 years
Radius	$R_\odot = 6.96 \times 10^8$ m
Mass	$M_\odot = 1.99 \times 10^{30}$ kg
Density	1.4×10^3 kg/m^3 (mean)
	151×10^3 kg/m^3 (center)
Temperature	15.6×10^6 K (center)
	5780 K (photosphere)
	$2 - 3 \times 10^6$ K (corona)
Luminosity	3.85×10^{26} W
Solar Constant	1366 W/m^2
Principal constituents	Hydrogen (92.1%)
	Helium (7.8%)
	All others (0.1%)
Equatorial rotational period	26.24 days[a]
Average Sun-Earth Separation	150×10^9 m

[a] This is the rotational period as observed from Earth. The sidereal period is approximately 24.5 days.

Table 3.2: Vital Statistics of the Sun.

3.1.1 The Sun's Life-Cycle

It appears the Sun was born about 4.5 billion years ago from a gravitationally collapsing cloud of hydrogen-dominated interstellar gas. As this gas cloud collpased it released gravitational potential energy, some of which was released as thermal radiation but much of which was absorbed by the interior, raising the temperature and pressure.[2] At some point during the collapse, temperatures and pressures in the core reached levels sufficient to support the fusion of ionized hydrogen into helium. This lighting of the "nuclear fires" increased the outward pressure gradient force, counteracted the gravitational collpase, and stabilized the Sun's size.

With the beginning of nuclear fusion in its core, the Sun entered the *main sequence* in which this gravitational equlibrium would be expected to persist for some 10 billion years. As the Sun (and in fact as any star) progresses through the main sequence, it gradually brightens due to a slowly increasing rate of fusion in its core. At some time in the future, perhaps 3-4 billion years from now, the energy output of the Sun will be sufficient to doom Earth to a runaway greenhouse effect, raising temperatures to such levels that the oceans will boil [Bennett et al., 2008, pp.495-497].

Once the Sun has exhausted its core nuclear fuel, it will undergo a series of rapid changes over the next few hundred million years whereby it will become a red giant, eventually reach a radius that encompases the Earth, and finally collpase into a dim white dwarf. Earth, if it has not been destroyed by the Sun during its death-throes, will become cold and dark [Bennett et al., 2008, pp.570-572]. But as this eventuality is understood to lie more than 5 billion years in the future and is, in any case, entirely beyond our control, it presents no cause for concern.[3]

[2] The interested student should investigate the *virial theorem* ($\overline{K} = -\frac{1}{2}\overline{U}$ where $\overline{K}$ and $\overline{U}$ are the average kinetic and potential energies) from which many important insights may be gleaned. For example, given the virial theorem and $\overline{E} = \overline{K} + \overline{U}$, it is clear that $\overline{E} = \frac{1}{2}\overline{U}$ so that as the potential energy of a collapsing gas cloud decreases, half of the released energy must leave the system (as, for example, thermal radiation). The virial theorem also reveals (given some assumptions) that the temperature of a gravitationally bound system varies as one over its radius. As a further tease to encourage the study of star formation, notice that the previous point implies that a system under gravitational collapse has a negative heat capacity: As the gas cloud collapses (lowering its energy), its temperature increases!

[3] Is there not a missing punctuation mark in the English language? A '?' indicates that the writer has asked a question; a '!' indicates, among other things, a point of excitement; a '.' indicates, well, just about anything else. Should there not be a mark to indicate that the writer is intending to demonstrate not a question, excitement, or bare statement, but rather a reason for calmness? Let this mark be called the calmness mark and indicate it with the symbol $\ddot{-}$. The sentense preceeding this footnote should be ended with the calmness mark. Rest assured, this footnote will not survive the first real edit of this text $\ddot{-}$

3.1.2 Solar Irradiance

The *solar constant* $F_\odot$ is the amount of photon energy per unit area per second available at 1 AU, at the top of Earth's atmosphere. It can be approximated by assuming that the solar luminosity ($L_\odot$) is radiated equally in all directions so that

$$\boxed{F_\odot \approx \frac{L_\odot}{4\pi(1\ \mathrm{AU})^2} \approx 1366\ \ \mathrm{W/m^2}.}$$

Satellites orbiting above Earth's atmosphere have this amount of power density available, *e.g.*, for conversion to electrical power using solar arrays. Assuming a conversion efficiency of ∼35%, space-based solar arrays require an area of ∼2 m^2 per kilowatt of electrical power produced. Ground-based solar arrays receive significantly less power due to the absorption of solar radiation by the intervening atmosphere. Under ideal conditions, the flux of solar radiation at the surface of Earth is about 1000 W/m^2 and under typical conditions it is far less.

The Sun's *spectral irradiance* defines the emitted photon power per unit area per unit wavelength and its integral over all wavelengths and viewing angles from the Sun yields the luminosity. Figure 3.1 shows the Sun's spectral irradiance at the top of Earth's atmosphere[4]. The numerous "bite-outs" in the spectrum are due to absorption of photons in the Sun's atmosphere and are known as *Fraunhofer lines*[5]. A blackbody fit to Figure 3.1 shows that the Sun has an effective blackbody temperature of ∼5785 K. Note that the photon energy output of the Sun maximizes in the visible portion of spectrum.[6]

[4] Data from `http://rredc.nrel.gov/solar/spectra/am0/`

[5] After the German physicist Joseph von Fraunhofer (1787-1826.)

[6] The Sun appears yellow not because its spectral irradiance peaks in the yellow (see the inset plot in Figure 3.1), but for more complicated reasons. Its irradiance peaks in the blue but has a small variance across the visible spectrum so that it is essentially a source of white light. The Sun appears white in images taken from space. The yellowish color we observe on the ground with our eyes is a result of atmospheric scattering (proportional to $1/\lambda^4$ so that bluer wavelengths are scattered more than redder wavelengths), our eye's spectral sensitivity (we can see yellow colors more easily than we can see blue colors), and the brain's image processing (which makes the Sun appear even yellower because of the sky's blue background).

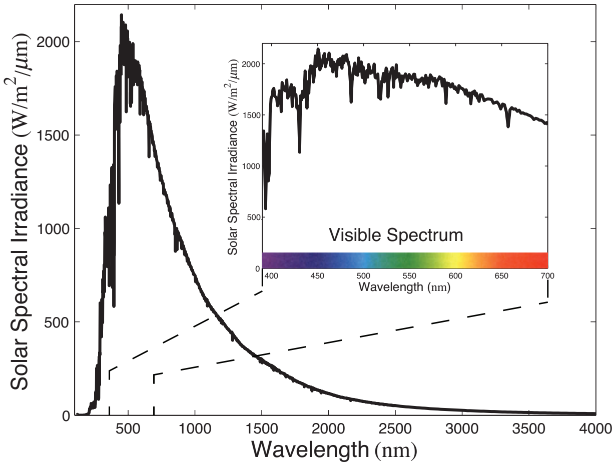

Figure 3.1: Spectral irradiance of the Sun observed at 1 AU.

3.2 Solar Structure

3.2.1 The Sun's Interior

Figure 3.2 shows an illustration of the Sun in cross-section. At the center is the *core* that extends out to about one quarter of the Sun's radius and it is here that the Sun's power is generated through nuclear fusion of hydrogen into helium nuclei. Core temperatures and pressures are such that charged particles collide with energies sufficient to overcome the repulsive Coulomb force. The short-acting strong force then binds, or fuses, the colliding particles together, releasing in the process the energy that powers the Sun. Figure 3.3 illustrates the fusion of hydrogen into helium nuclei through the three steps known as the *proton-proton chain.* Step 1 involves the fusion of two hyrdogen nuclei into deuterium, liberating 0.42 MeV of energy. The release during this step of a positron and a neutrino results in the change of one proton to a neutron. Thus in Step 1,

$$^1_1\mathrm{H} + {}^1_1\mathrm{H} \rightarrow {}^2_1\mathrm{D} + e^+ + \nu + 0.42\ \mathrm{MeV}.$$

The positron indicated in this reaction annihilates very rapidly with an electron, producing two gamma rays with 0.511 MeV of energy each (thus conserving energy) and travelling in opposite directions (thus conserving momentum).

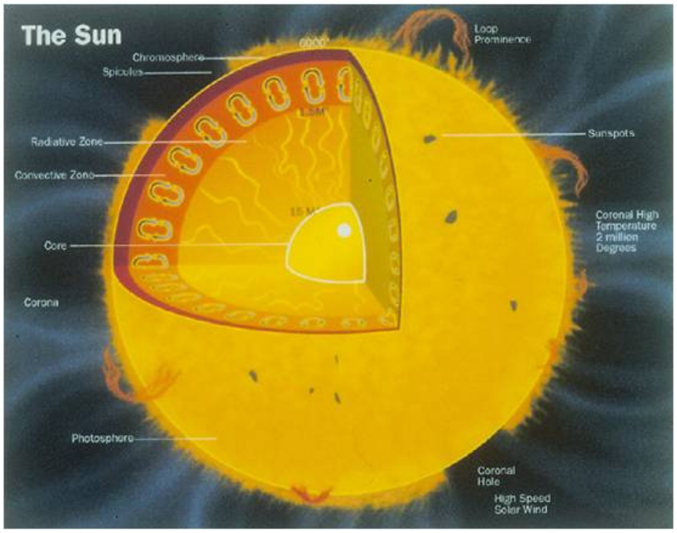

Figure 3.2: The Sun in cross-section, showing its basic structure. Figure from http://ase.tufts.edu/cosmos/view_picture.asp?id=590.

In Step 2 of the proton-proton chain, a deuterium nucleus fuses with another proton to form helium-3, releasing a gamma ray and 5.49 MeV of energy. That is,

$$ {}^{2}_{1}\mathrm{D} + {}^{1}_{1}\mathrm{H} \rightarrow {}^{3}_{2}\mathrm{He} + \gamma + 5.49 \text{ MeV}. $$

The final step of the proton-proton chain proceeds in one of four ways, the most common of which involves the fusion of two helium-3 nuclei into helium-4 with the release of two protons and 12.86 MeV of energy. In this reaction,

$$ {}^{3}_{2}\mathrm{He} +^{3}_{2}\mathrm{He} \rightarrow {}^{4}_{2}\mathrm{He} + 2\,{}^{1}_{1}\mathrm{H} + 12.86 \text{ MeV}. $$

The net change in mass over the three steps of this proton-proton chain is approximately the difference between the mass of a helium-4 atom and the mass of four hydrogen atoms.[7] This mass, converted into energy, amounts to 26.7 MeV and given the Sun's luminosity, it can be shown that the Sun coverts 600 billion kg of hydrogen into helium every second (see Exercise 3.2).

[7] In this approximation, the neutrino mass is neglected.

Another calculation shows that the reactions in the proton-proton chain occur a mind-boggling 10^{38} times per second (see Exercise 3.1).

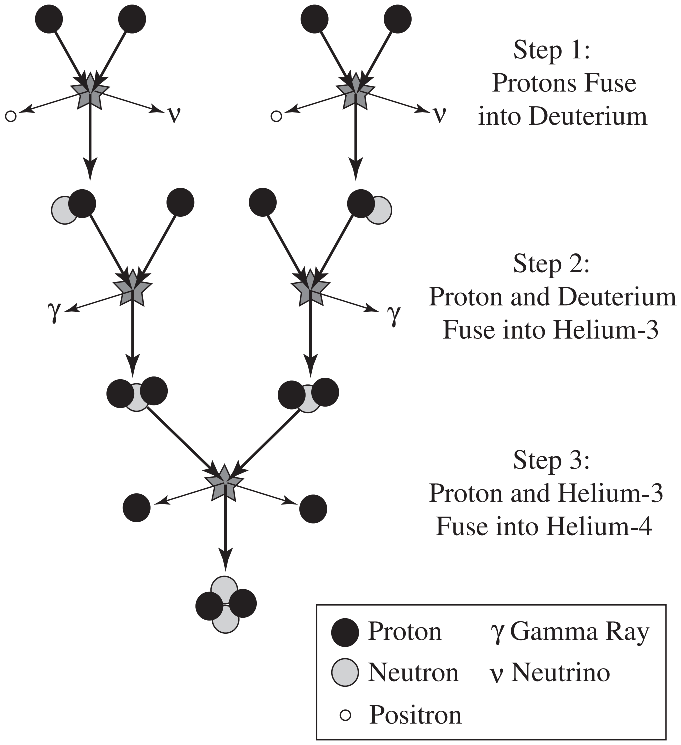

Figure 3.3: Fusion in the Sun's core proceeds along the proton-proton chain of 3 reactions to fuse hydrogen into helium nuclei.

Photons emitted by the core are absorbed and re-emitted innumerable times during their random walk journey to the surface of the Sun and beyond. The size of the Sun and its densities are such that photons emitted by the core require a few hundred thousand years to reach the surface. On Earth, we are bathed in photons that left the outer surface of the Sun some eight minutes ago, but these "same" photons left the core a few hundred thousand years ago!

Proceeding outward away from the core, the *radiative zone* indicated in Figure 3.2 extends to $\sim 0.86R_\odot$ as the temperature steadily drops from its core values. In this radiative zone, outward energy transport is accomplished primarily via photon radiation. Continuing outward, the Sun continues to cool and the next $\sim 0.14R_\odot$ constitutes the *convection zone* where the temperature gradient is large enough that equilibrium cannot be maintained by radiation alone. Here the Sun roils like a pot of water on the stove with areas of hot plasma that rise to the surface being replaced by areas of sinking cooler plasma.

3.2.2 The Sun's Atmosphere

This continual overturning in the convection zone lies just beneath the lowest layer of the Sun's atmosphere, the *photosphere*, which is the visible surface of the Sun and a mere 500 or so km thick. It is here that the blackbody temperature reaches the values of ~5785 K mentioned above. Viewed with the naked eye, the photosphere appears as a smooth surface but viewed with sufficient resolution, its granulation reveals the churning and overturning nature of the underlying convection zone. Sunspots, a topic to be discussed at some length in §3.3.1, sometimes appear on the photosphere. Although sunspots commonly have radii larger than that of Earth, they are not usually visible at Earth without magnification.[8]

Above the photosphere lies the *chromosphere* with a thickness of ~ 2500 km. This middle layer of the Sun's atmosphere has a temperature of about 10,000 K and is the region responsible for most of the Sun's radiated ultraviolet light. The outermost layer of the Sun's atmosphere is known as the *corona* and extends several million kilometers above the photosphere. At least in part due to its very low density and correspondingly low heat capacity, coronal temperatures reach values of over a million Kelvin. Coronal gasses are therefore copious producers of X-rays.

The Sun's corona is spectacularly beautiful but not usually visible with the naked eye due to the overwhelmingly bright photosphere that lies beneath it. Several coronal images recorded from satellite instrumentation will be presented below but here it may be pointed out that rare occurrences on Earth do in fact make it occasionally possible to view the corona with the unaided eye. Specifically, solar eclipses are those rare events when the moon

[8] Naked-eye sunspots are sometimes observed. One of the authors (Hughes) saw one on a cold winter's day in Fairbanks, AK. It was around noon and the Sun was low on the horizon, viewed through a dense ice-fog that dimmed the Sun to the point where it could be looked-at directly. There was a large black spot in the lower left-hand side of the Sun. Upon arriving at work, the observation was confirmed by comparing with the most recent solar image on `spaceweather.com`.

passes between Sun and Earth on a trajectory that blocks at least part of the Sun from view. In one of the solar system's most surprising coincidences, the ratio of our moon's to the Sun's radii almost exactly matches the inverse ratio of their distances from Earth so that, when viewed from Earth, each appear to be very nearly the same size. On rare occasions the moon's trajectory during a solar eclipse entirely blocks the Sun's photosphere which for a few brief moments reveals the corona to the unaided eye. Figure 3.4 shows a photographic image of the corona taken during just such an event, the total solar eclipse of August 11, 1999 as seen from France.[9] This coronal atmosphere is the outermost layer of the Sun and extends into interplanetary space.

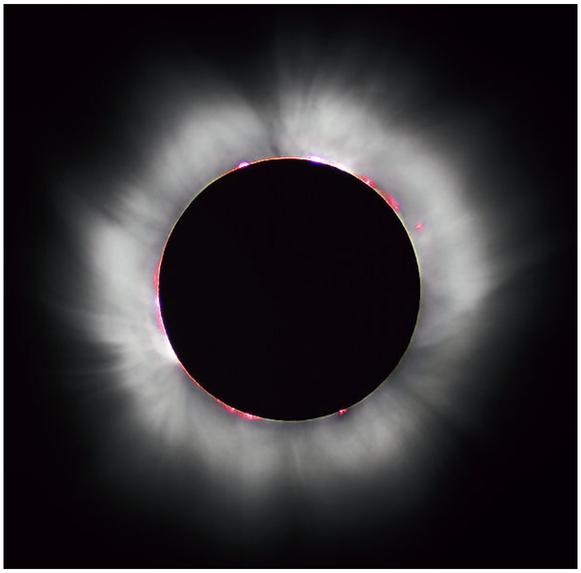

Figure 3.4: The Sun's corona revealed during the total solar eclipse of August 11, 1999. Photo by Luc Viatour.

3.3 Solar Activity

For the vast majority of us who do not live either above the arctic or below the antarctic circles, the Sun rises and sets every day. We set our watches

[9] Image from `http://en.wikipedia.org/wiki/Corona`.

by it and are at least as sure of these two daily occurrences as we are of those two other things we are told are the only guarantees in life[10]. Year after year, decade after decade, and generation after generation we perceive very little, if any, change in either its warmth or its brightness. Our Sun appears to be constant. Careful observers for hundreds if not thousands of years have however noted that the Sun is in fact not constant. Spots appear on it's otherwise uniformly-appearing bright surface. They persist for some time and disappear. More or fewer of these so-called sunspots appear at a given time and their number goes through a relatively ordered cycle. Those living at high latitudes (either northern or southern) note a similar cycle in the occurrence of large auroral displays. The Sun is in fact active and this section describes some of this activity with a particular interest in that which impacts Earth and human technology.

3.3.1 Sunspots

Figure 3.5[11] shows a white-light image of the Sun's photosphere revealing the presence of several large sunspots. Stated most simply, *sunspots* are dark spots on the photosphere. They are dark only in the sense that they are dark*er* than the sorrounding areas and this darkness is indicitave of their lower temperatures. Photospheric temperatures in a sunspot range from 4000-4500 K whereas the average photospheric temperature is ∼5800 K.

Sunspots have been observed and counted for hundreds if not thousands of years and telescopic observations that began around the year 1610 provided the first evidence that the Sun rotated.[12] Observations soon revealed another startling fact regarding sunspots: their numbers and even their locations on

[10] Benajmin Franklin, in a letter to Jean-Baptiste Leroy on November 13, 1789 famously noted, "Our new constitution is now established, and has an appearance that promises permanency; but in this world nothing can be said to be certain, except death and taxes."

[11] From: `http://en.wikipedia.org/wiki/File:Sun_projection_with_spotting--scope.jpg`

[12] It is difficult for us in the modern era, with so much accumulated information and when so much scientific progress is incremental, to appreciate the worldview-changing effects of such fundamental discoveries. All the more so because of this, the student is encouraged to put him or herself in the place of those early investigators and feel the weight of their discoveries. *The Sun rotates!* Students, recall in this context the year in which Galileo was censured by the Roman Inquisition for views expounded in his heliocentric work, Dialog on the Two Chief World Systems, and in which legend records his unpenitent muttering about the Earth: "Nevertheless, it moves." Such were the times.

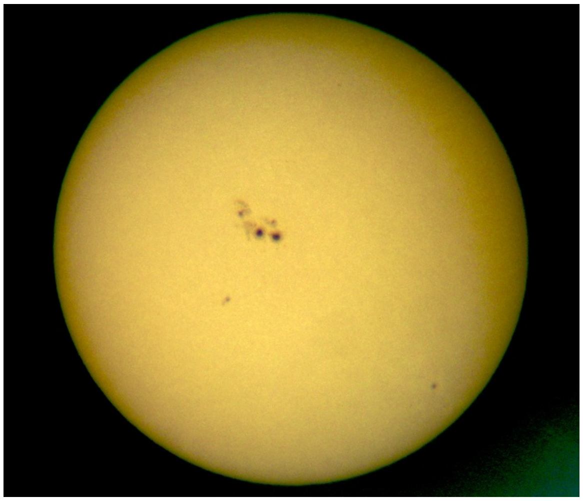

Figure 3.5: A white light image of the Sun's photosphere, showing large sunspots. Photograph by SiriusB.

the Sun varied over time in a quasi-regular pattern.

The *Wolf number*, also known as the International sunspot number (SSN), encodes the number of sunspots and sunspot groups on the photosphere. Observations reveal that, on average, sunspots groups are formed of approximately 10 individual sunspots and so the SSN is calculated as 10 times the number of sunspot groups plus the sum of all individual sunspots not part of the previously counted groups. Historical data and observations from the Royal Greenwich Observatory (RGO), the US Air Force (USAF), and the US National Oceanic and Atmospheric Administration (NOAA) are used to compute daily-, monthly-, and yearly-averaged values of the SSN. Figure 3.6b shows a time series of the monthly-averaged SSN[13] since the year 1874. These data display several interesting features. Perhaps most obviously,

> the SSN varies with a period of about 11 years between successive maxima.

During *solar minimum*, the SSN is nearly zero while at *solar maximum* the number peaks at values that differ significantly from cycle to cycle. Fourier analysis shows that this *solar cycle* has a dominant period of approximately

[13] Data from `http://solarscience.msfc.nasa.gov/greenwch/spot_num.txt`

10.7 years although the variation about this mean is significant. These solar cycles are numbered and solar cycle number one is defined as the cycle that began in March 1755. The solar cycle number is identified in Figure 3.6b as the number near the base of each SSN peak. As of 2010, we have entered solar cycle 24. As you will find in Exercise 3.3.5, several periods of unusually low and high SSN have been identified and these periods are strongly correlated with geophysical observations including Earth's global mean temperature.

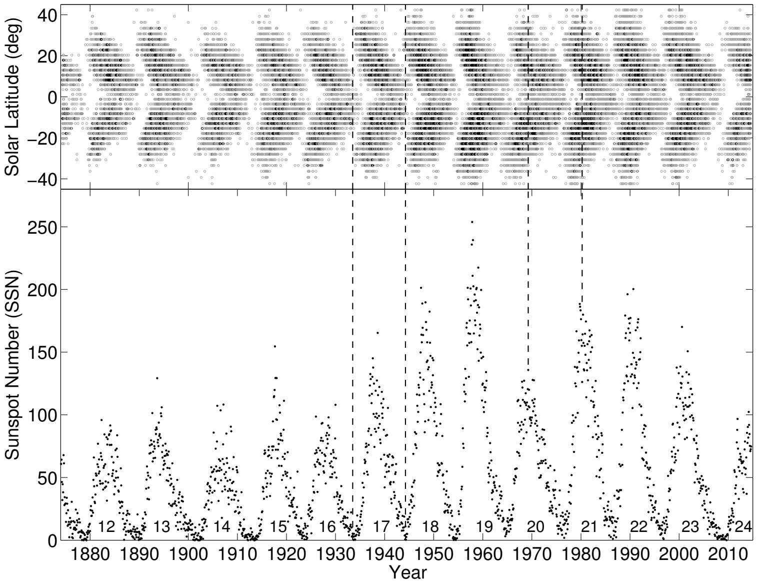

Figure 3.6: a) A so-called butterfly diagram showing the fraction of the solar hemisphere covered by sunspots as a function of both time and solar latitude. Darker colors indicate a larger fraction of coverage with maximum values (shown as black) representing coverage of approximately 2% of the latitudinal strip. b) The monthly-averaged SSN from 1874-2014. The vertical lines identify example times of solar minimum (those passing through SSN minima) and solar maximum (those passing through SSN maxima). The solar cycle number is listed near the base of each SSN peak.

In addition to the counting of sunspots that yields the SSN, observers at the GRO, USAF, and NOAA have also noted the solar latitude at which those sunspots occur. Figure 3.6a shows these data[14] in the form of a so-called *butterfly diagram* that displays the fraction of the solar hemisphere covered by sunspots as a function of both time and solar latitude. In this figure, darker areas indicate a larger fraction of coverage and it can be seen that the vertical lines indicating times of solar maximum pass through the darkest areas on the butterfly diagram. Butterfly diagrams reveal a curious and interesting feature. If we identify the beginning of a solar cycle as the time of solar minimum, we see that as sunspots first begin to appear (that is, as the SSN begins to increase from its local minima near zero), they tend to form at the highest latitudes (typically about 40°) and generally appear at consistently lower latitudes as the 11-year cycle progresses until, just before the beginning of the next cycle, sunspots tend to appear at very low latitudes (typically about 15°).

Much like the Earth and other bodies in the solar system, the Sun generates a magnetic field and this field permeates the Sun itself and, as we will see below, the entire solar system. Further insight into the nature of sunspots and of their importance to the near-Earth space environment comes from observations of the strength and polarity of magnetic fields inside sunspots. The magnetic field strength inside sunspots is several orders of magnitude higher than the photospheric average. Further, it has long been known that sunspots tend to occur in pairs and that the two sunspots forming a pair have opposite magnetic polarity. Figure 3.7 shows an image recorded by the MDI instrument onboard the SOHO satellite of the Sun's photospheric magnetic field polarity. In this figure, gray regions indicate a relatively weak magnetic field and white (black) regions indicate a strong field directed out of (into) the photosphere. For convenience, let us designate the white regions as those of positive polarity and the black regions as those of negative polarity. In addition to the obvious black-white pairings of sunpots in this image, note that in the southern hemisphere, the negative polarity spots tend to lead the positive polarity spots and that the reverse is true in the northern hemisphere.

From one solar cycle to the next, this reversal of polarity ordering from northern to southern hemisphere is consistently found, but amazingly, it is also found that in any given hemisphere, the ordering reverses during each consecutive solar cycle. That is, if in the northern hemisphere of a given solar cycle, positive spots lead negative spots, then two things can be stated.

[14] Data from `http://solarscience.msfc.nasa.gov/greenwch.shtml`

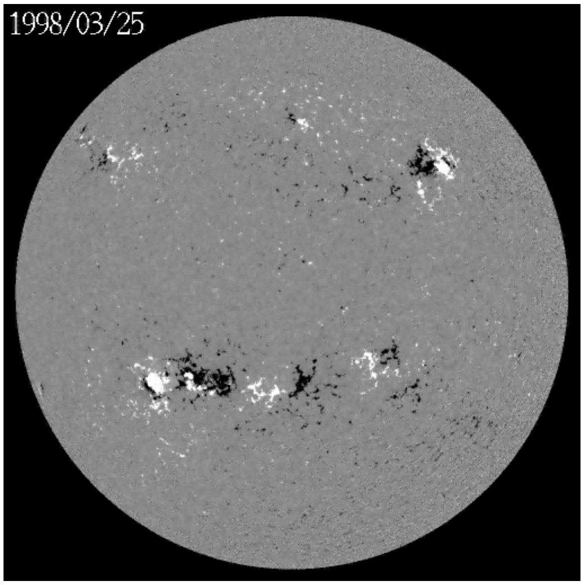

Figure 3.7: Full-disk magnetogram image from the Michelson-Doppler Imager (MDI) instrument onboard the SOlar and Heliospheric Observatory (SOHO) satellite. Gray regions indicate very weak magnetic fields and white (black) regions indicate strong magnetic fields directed out of (into) the solar photosphere. Movie frame from: `http://soi.stanford.edu/press/SSU_2000-/Backside/movies/mag.qt`

First, in the same solar cycle, southern hemisphere spot pairs will be lead by negative spots. Second, in the following (or preceeding) solar cycle, northern hemisphere spot pairs will also be led by negative spots. Thus we may say that in addition to, or superposed onto, the 11-year solar cycle is another *22-year cycle* during which the polarity of sunspot pairs in a given solar hemisphere is repeated. Even more, the Sun has a magnetic field much like Earth does and the polarity of this field reverses every solar cycle so that it repeats every ∼22 years. The student should not allow this stupendous fact to speed by unpondered. Suppose the Sun's magnetic field to be that of a bar magnet located near it's center.[15] This supposed bar magnet reverses polarity every ∼11 years!

The reasons for these polarity reversals (both of the Sun itself and con-

[15] It is not!

sequentially of its sunspot pairs) is an area of current research but we may gain some qualitative insight into the reasons with the addition of two more facts. First is the observational fact that the Sun exhibits differential rotation wherein it rotates fastest at its equator and progressively more slowly towards its poles.[16] The change in rotational period is significant and the Sun at its poles takes at least 33 days to complete a rotation as compared to approximately 25 days near its equator. Second, for reasons we will explore in more detail in §3.6.1, the Sun's plasma and magnetic fields are in a sense locally linked so that they rotate together.

Figure 3.8 presents a time sequence of images of the solar photosphere and it's magnetic field lines. In this figure, the solar magnetic field is represented by the blue lines with white arrows indicating polarity. A helpful model is to think of the Sun as a differentially-rotating ball and each magnetic field line as a Slinky$^{\mathrm{TM}}$ that lies on the surface of the ball. In panel a) at the beginning of the sequence, the magnetic field is directed from south to north and we take this field to be approximately that of the bar magnet mentioned previously. Panel a) corresponds to solar minimum, a state of ordered solar magnetic field and very few if any sunspots. The first effects of the Sun's differential rotation are noticed in panel b) where due to the higher rotation rates near the equator, the solar plasma has begun to distort the magnetic field and, in our model, each Slinky$^{\mathrm{TM}}$ wrapping the surface is becoming longitudinally distored. This distortion continues and becomes more extreme in panels c)-e) as the equatorial regions continue to outstrip the higher latitudes.

Such a collection of wound-up slinkies is a difficult thing to control and at some point first pictured in panel f), one of them has been twisted to such an extent that it begins to kink and break free from the ball's surface. Note the polarity of the magnetic field line on which this "kink" first appears and notice that, at the site of kink, the field will be directed out of the photosphere at the leading edge and into the photosphere at the trailing edge. These two points with their opposite polarities are the sites of the first sunspots in a new solar cycle. The process continues, the Slinkies$^{\mathrm{TM}}$ become more wound-up, more prone to "kinking", and more sunspots appear in both hemispheres. In panels g) and h), field lines in both hemispheres may be traced to see the hemisphere-reversed polarity pattern of sunspot pairs discussed previously. Finally, in panel i) the magnetic field has become highly disordered and there are a large number of sunspot pairs. This final panel

[16] This seems a strange behavior and I cannot explain its cause.

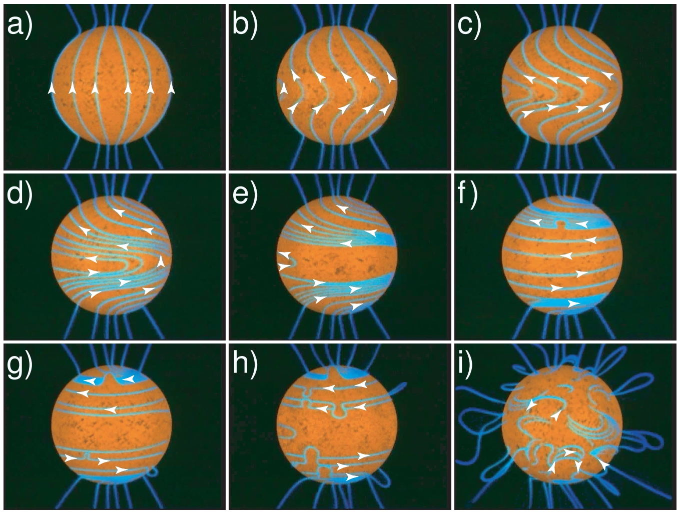

Figure 3.8: A time-sequence of frames beginning with solar minimum and ending with solar maximum. The blue lines indicate the Sun's magnetic field that progresses from ordered to disordered as the solar cycle advances. Adapted from `http://sohowww.nascom.nasa.gov/gallery/-Movies/dynamo/dynamo.mpg`

represents solar maximum and continued progress towards the next solar minimum essentially follows the reverse sequence but with the end result of the next solar minimum polarity being opposite that shown in panel a). Thus in one solar cycle, the Sun's magnetic field has gone from being highly ordered at solar minimum, through a highly disordered state at solar maximum and returned to a highly ordered state with reversed polarity at the next solar minimum.[17]

The student may at this point reasonably ask why so many words have been dedicated to sunspots, their numbers, polarities, and relation to the solar cycle. Two reasons are that the SSN is highly correlated with geomagnetic activity on Earth and that sunspots themselves are directly linked with some causes of this activity. This link is explored in the following section.

[17] The solar cycle, its spatial and time scale, and the processses involved are among the most amazing things the author has ever encountered.

3.3.2 Solar Flares and Coronal Mass Ejections

Figure 3.9 shows an image of the Sun recorded with the SOHO EIT (Extreme ultraviolet Imaging Telescope) instrument at the 304 Å emission line of singly-ionized Helium. Identified in this image are two types of features of interest to us here: solar filaments and prominences. Filaments are dark crack-like features that appear on the solar disk and prominences are bright loop-like structures that extend upward from the photosphere and chromosphere into the corona and are visible near the limb. Filaments and prominences are actually two different manifestations of the same feature in that a prominence is a filament viewed from the side. These features form when relatively dense low-lying plasma at a temperature of ∼80,000 K populates a magnetic loop much like those illustrated in Figure 3.8. In part depending on the wavelength at which they are observed, filaments appear dark for one of a few reasons. As with sunspots, the emissions may be coming from a region that is simply cool*er* and therefore dark*er* than the background against which they are viewed. Also, the number density of the emitting species may be lower in a particular region than in the background, making that region, as before, dark*er* than the background. Third, enhanced absorption at particular wavelengths may cause a region to appear dark.

Prominences, on the other hand, appear bright because they are viewed near the solar limb against the tenuous corona and the dark background of space. Filaments and prominences can form in a day and may persist stably for several months during which time they will have rotated with the Sun into and out of view a number of times. Prominences are truly spectacular features, often spanning distances of many Earth radii and containing plasma with masses exceeding 10^{14} kg. As spectacular as prominences are when they are stable, they are unquestionably more so when they collpase and explode, flooding their path with highly energetic plasma and electromagnetic radiation. These explosions may be loosely called *solar flares* and *coronal mass ejections* (CMEs).

Solar flares and CMEs are likely distinct events triggered by a common mechanism [Kallenrode, 2004, p.188]. Both transfer vast amounts of energy from coronal magnetic fields to the plasma and electromagnetic radiations and for our purposes here we may naively consider solar flares to be smaller, more localized "explosions" of a magnetic loop and CMEs to be large scale eruptions involving a significant fraction of the solar corona. Plasma ejections from these events can attain masses in excess of a billion tonnes travelling

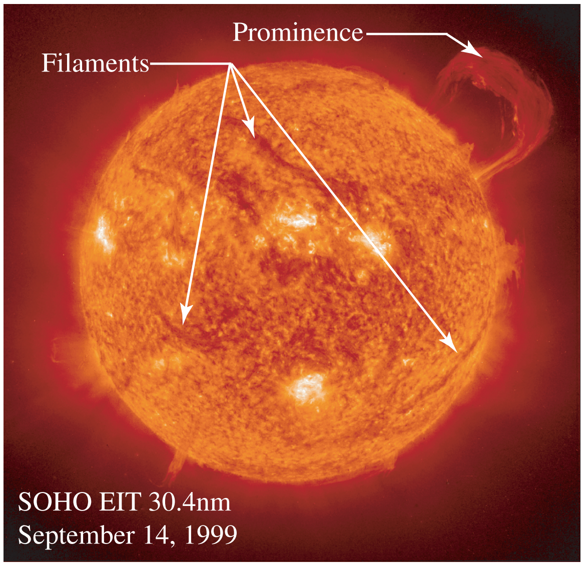

Figure 3.9: September 14, 1999 EUV image of the solar corona from the SOHO EIT instrument at 304Å. Several filaments and a large prominence are visible.

at speeds of over two million miles per hour. The energy output of a large flare can exceed 10^{25} J while that of a large CME can exceed 10^{26} J, more than doubling the Sun's energy output for its duration. Earth at a distance of 1 AU from the Sun is a small target for this energetic ejecta but these events are neither rare nor highly localized in space so that Earth-impacts are a statistical certainty. During solar minimum, approximately one CME is observed per week and this rate increases to several per day during solar maximum. As we will soon see, the ejecta occupies a volume in space that increases as it travels into the solar system so that it may be likened to the pellets of a shotgun blast. With dozens or even hundreds of such blasts occuring per year, our distance from the Sun and Earth's small size offer no statistical protection. Earth's direct interaction with radiations and dense plasma traveling at millions of miles per hour is a certainty. We are protected (at least to some extent) by Earth's own magnetic field and atmosphere, but as this is a topic for future discussions, let us return to the prominences whose collapse is often the trigger for such events.

Figure 3.10 shows frames from an artist's rendered NASA amination illustrating the evolution of a pair of sunspots into a solar flare. The animiation begins in panel a) with a view over the photosphere and through the hot and tenuous corona. A pair of magnetic field loops, which we may identify with the "kinks" illustrated in Figure 3.8, are visible along with their associated sunspots. In panel b) the view has moved from above to below the surface of the photosphere and a newly formed magnetic loop is visible. In panels c)-e) this loop protrudes through the surface of the photosphere and forms a new sunspot pair. If we identify the resulting loop structure visible in panels f) and g), again from above the photosphere, as a prominence, then the edge-on view shown in panel h) would represent a filament. In panels i)-n), some of the energy contained in this magnetic loop is released as a solar flare through a process known as reconnection whereby the magnetic field is reconfigured and plasma is ejected into space. Panels o) and p) illustrate the remnants of the loop. The entire process illustrated in this figure occurs on the time scale of minutes and releases a large amount of energy.

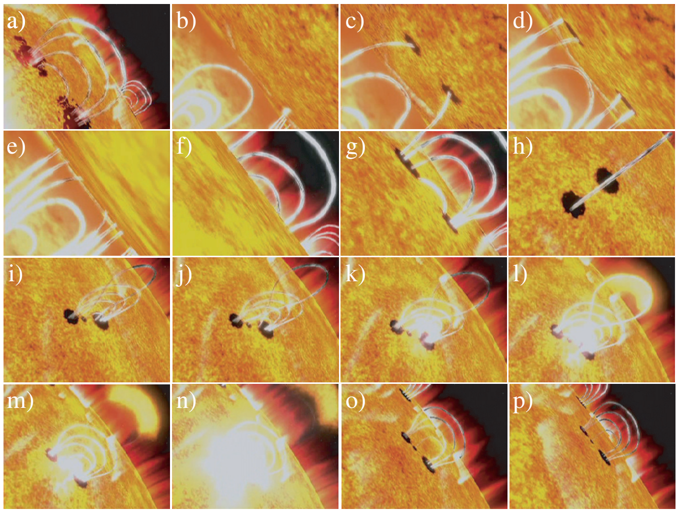

Figure 3.10: An artist's rendering of the evolution of a pair of sunspots into a solar flare. Image frames from `http://sohowww.nascom.nasa.gov/-bestofsoho/Movies/10th/SunspotsForm.mpg`

Solar flares are classified according to the peak energy flux from soft X-rays (1-8 Å) as observed by the Geostationary Operational Environment Spacecraft (GOES) fleet. Table 3.3 shows the classification scheme which is logarithmic with a linear subclassification. To illustrate the scheme, note that the energy flux from a B1 class flare is 10 times higher than from an A1 flare and the energy flux from M7 flare is 7 times higher than from an M1 flare and 700 times higher than from a B1 flare. M and X class flares are powerful enough to cause significant geomagnetic activity and disturbed space weather.

Solar Flare Classification	GOES Peak Energy Flux 1-8 Å(W/m^2)
A(1-9)	$(1-9) \times 10^{-8}$
B(1-9)	$(1-9) \times 10^{-7}$
C(1-9)	$(1-9) \times 10^{-6}$
M(1-9)	$(1-9) \times 10^{-5}$
X(1-ζ)	$(1-\zeta) \times 10^{-4}$

Table 3.3: Soft X-ray solar flare classification scheme

The strongest recorded flare occurred on November 4, 2003 and is officially classified as an X28 although the energy flux saturated the GOES detectors for many minutes and the classification may therefore be only a lower bound. Although this flare was not directed at Earth, its effects caused numerous satellite anomalies and forced astronauts onboard the International Space Station to take shelter in radiation-hardened areas. Solar flares are often associated with CMEs and Figure 3.11 shows a composite image of a large CME that occurred on August 7, 2002 as recorded by the SOHO LASCO (Large Angle Spectrometric COronograph) and EIT instruments and Figure 3.12 shows frames from a SOHO LASCO video of the eruption of large CME on October 28, 2003. Figures 3.11 and 3.12 illustrate the massive scale of these events but the student is strongly encouraged to browse the NASA SOHO archive of images and videos and more fully appreciate the size, violence and regularity of these events.[18]

As visually compelling as these images and videos of our dynamic sun are (to say nothing of the compelling physics involved!), our interest in them is to understand how solar activity impacts the near-Earth space environment and Earth's ground- and space-based technologies. To be sure, the impacts of events such as solar flares and CMEs are significant and many of Earth's

[18] The "Best of SOHO" archive is found at `http://sohowww.nascom.nasa.gov/-gallery/bestofsoho.html`

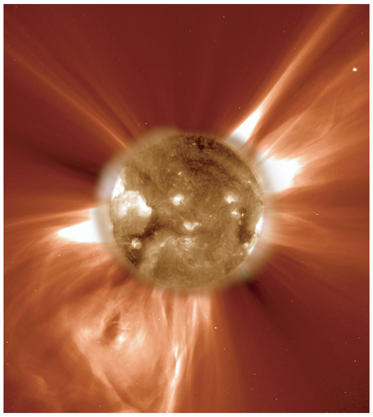

Figure 3.11: SOHO LASCO/EIT composite image of a large CME. From `http://sohowww.nascom.nasa.gov/gallery/images/c2eitcomp.html`

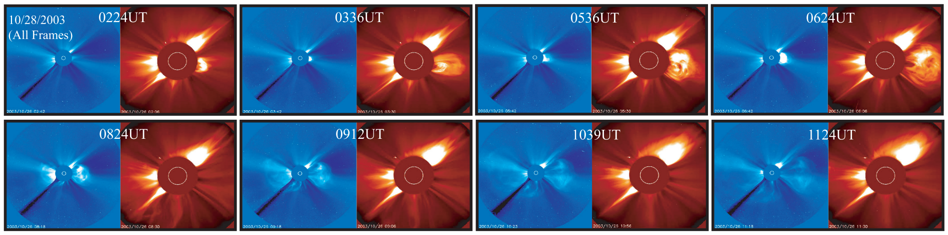

Figure 3.12: SOHO LASCO images showing the eruption of a large CME. The left- and right-hand images in each panel show wide and narrow field-of-view images of the solar corona recorded with two different instruments but closely separated in time. The average time for each set of images is given in the frames. From `http://sohowww.nascom.nasa.gov/gallery/-Movies/flares.html`

processes and phenomena we will encounter throughout this text are associated with them. But let us defer those investigations until later chapters and now move outward from the Sun into the interplanetary medium, remembering that our Sun is active with an ∼11-year solar cycle, that it not infrequently undergoes violent erruptions that spew billions of kilograms of plasma into space at speeds of millions of miles per hour, and that some of this plasma will on occasion be directed towards Earth.

3.4 Introduction to the Solar Wind

"The vacuum of space" is a phrase fairly entrenched in most of our minds - and yet space is most certainly not a vacuum at all. Two items may be mentioned to turn our attention to the possibility of a certain wind from the Sun blowing through the supposed vacuum of interplanetary space. First, recall that the solar corona is the outermost layer of the Sun's atmosphere and that its temperature exceeds a million Kelvin. Could these temperatures be high enough that the coronal plasma "boils off" into space much like steam rising from a pot of water boiling on a stove? That is, could the kinetic energy of coronal plasma exceed the Sun's gravitational potential energy so that the plasma may continuously flow away from the Sun?

Second, it has long been known that comets sometimes have not one, but two (and sometimes even three) distinctly visible tails. Figure 3.13 shows photographs of Comet Hale-Bopp that clearly display two tails. Of these two tails, one appears blue in color and the other appears white or yellow in color. The white or yellow tail is a dust tail composed of small cometary particles weakly pushed by solar photon pressure into a diffuse shape generally following the comet's orbital path. The other (blue) tail is known as the ion tail, is less diffuse, and is swept radially away from the Sun. In 1943 Cuno Hoffmeister suggested that this tail forms when "solar corpuscular radiation" acts on the cometary material [Hoffmeister, 1943][19]. In more modern parlance, this solar corpuscular radiation is called the *solar wind*, the continuous flow of charged particles from the Sun and in which the entire solar system is bathed.[20]

Table 3.4 presents a selection of "vital statistics" for the solar wind and its embedded IMF [Kivelson and Russell, 1995, pp.92-94]. The top part of the table shows observed values, mostly gathered from satellite observations, and the bottom part of the table shows derived properties where n is the (assumed equal) proton and electron number density, k is Boltzmann's constant, T_p and T_e are the proton and electron temperatures respectively, and γ is the ratio of specific heats at constant pressure and constant volume (taken to be $\gamma = \frac{5}{3}$).

[19] The existence of this corpuscular radiation had been suspected and proposed many years prior to Hoffmeister's work in 1943. For example, Kristian Birkeland in 1903 proposed that a stream of electrons coming from the Sun was responsible for the aurora.

[20] Because this tail is observed to be swept radially away from the Sun, we may suspect that solar wind particles have velocities in the same direction.

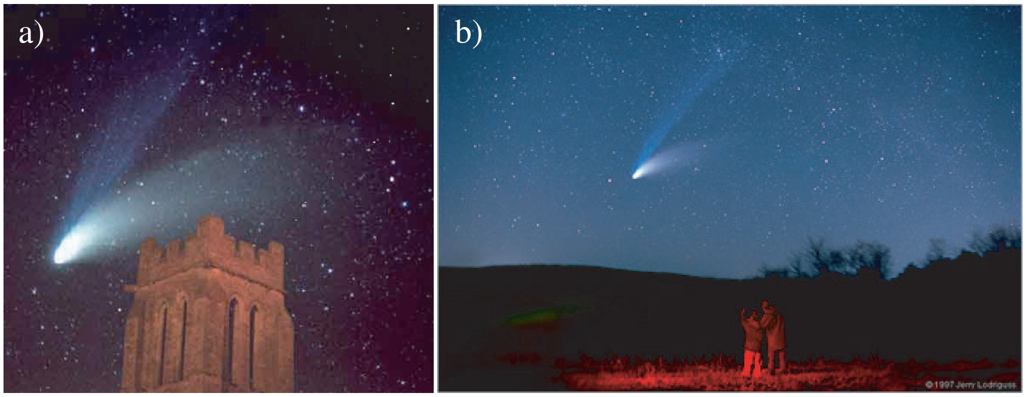

Figure 3.13: Images of Comet Hale-Bopp showing two distinct tails. Photographs a) and b) were taken in 1997 by Malcolm Ellis and Jerry Lodriguss, respectively. Images from a) `http://www.semp.us/publications/-biot_reader.php?BiotID=433` and b) `http://astronomy.swin.edu.au/-cosmos/C/Comet`.

To comment on these values, a few items may be noted. First, notice that contrary to the previous assumption,[21] the number density of electrons is slightly higher than that of the ions; the excess electrons originate mainly from the doubly-ionized Helium ions.[22] The flow speed is variable but the typical number of 450 km/s corresponds to approximately a million miles per hour. Imagine - at 1 AU, approximately a dozen particles in every cubic centimeter of space are constantly flowing towards earth at a million miles per hour! This is the solar wind.

[21] That $n_i = n_e$, which is true for singly ionized ions.

[22] If we assume that protons and doubly-ionized Helium atoms are the only ions in the plasma, it is possible to determine their relative abundances. Taking the rough values as given yields $7.1a_H + 0.25 \times 2a_{He} = 6.6$ and $a_H + a_{He} = 1$ where a_H and a_{He} are the abundances of hydrogen and helium, respectively. From these two equations we find a hydrogen abundance of >92% and a Helium abundance of <8% in the solar wind. Actual abundances vary with solar cycle but are not far from these values.

Solar Wind - Typical Observed Values at 1 AU[a]	
Proton density	6.6 cm^{-3}
Electron density	7.1 cm^{-3}
He^{2+} density	0.25 cm^{-3}
Flow speed[b]	450 km/s
Proton temperature	1.2×10^5 K
Electron temperature	1.4×10^5 K
IMF Magnetic field[c]	7×10^{-9} T
Solar Wind - Typical Derived Properties	
Gas pressure, $p_{gas} = nk(T_p + T_e)$	∼30 pPa
Sound speed, $c_s = \frac{\gamma p}{\rho} = \left(\frac{\gamma k}{m_p+m_e}\left(T_p + T_e\right)\right)^{\frac{1}{2}}$	∼60 km/s
Magnetic pressure, $p_{mag} = \frac{B_{IMF}}{2\mu_0}$	∼ 15 pPa
Proton gyroradius, $r_{L_p} = \frac{v_{\perp p}}{\omega_{cp}}$	∼80 km
Flow time from corona to 1 AU	∼4 days
Proton-proton collision time	$\sim 4 \times 10^6$ s
Electron-electron collision time	$\sim 3 \times 10^5$ s

[a] Real-time solar wind conditions from the ACE satellite located at the L1 Lagrange point can be found at: http://www.swpc.noaa.gov/ace/MAG_SWEPAM_24h.html

[b] This flow is directed nearly radially away from the Sun.

[c] The IMF magnetic field strength is highly variable. Its orientation is nearly parallel to the ecliptic plane and directed approximately 45° from the Sun-earth line.

Table 3.4: Vital Statistics of the Solar Wind and its IMF.

3.5 Parker's Solar Wind

In 1958, Eugene Parker published a much disbelieved but now widely acclaimed paper on the solar wind [Parker, 1958].[23] Motivated by cometary observations that suggested the existence of such a solar wind, Parker investigated the dynamic consequences of those observations and arrived at

[23] For Eugene Parker's telling of the story surrounding this paper (and a good deal of additional advice and insight), see "The Martial Art of Scientific Publication" by E.N. Parker as published in *EOS*, vol. 78, no. 31 on September 16, 1997 - or download the pdf at: `www.hao.ucar.edu/∼travis/seminar/art.pdf`.

two conclusions of particular importance to our study. First, given a set of reasonable assumptions, the Sun's corona cannot be in a state of static hydrodynamic equilibrium. That is, there must be a continuous non-zero flow of coronal mass away from the Sun. Second, he was able to deduce the effect of this flow on the configuration of the IMF. Let us consider these two conclusions in turn.

Having investigated and rejected the possibility that the solar corona could be in static equilibrium,[24] Parker proceeded to develop a theoretical model for the equilibrium state of a steadily expanding corona. His model was spherically symmetric so that all quantities vary only with radial distance from the Sun and it was assumed that the flow was entirely in the radial direction. A further assumption that the expansion was "steady" eliminated all time derivatives.

Parker's model of the corona is a fluid model developed from two fundamental relations: a momentum equation and a mass continuity equation. We encountered a momentum equation in Eq. 2.5 and a continuity equation in Eq. 2.7.[25] The equation of mass continuity in the solar corona is

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0$$

where ρ is the coronal mass density and $\mathbf{u}$ is the fluid flow velocity. Recalling now from Ch. 2 that a momentum equation is essentially Newton's second law applied to some volume of space with an average mass density of ρ, we must idenfity all force densities that act on a volume of the coronal plasma. There are three such forces to mention here. The first is the pressure gradient force that results when the fluid pressure varies with position; the second is the magnetic force resulting from the coronal plasma moving in the presence of a magnetic field; and the third is the gravitational force. Following Parker and neglecting the magnetic force, the momentum equation is

$$\rho \frac{\partial \mathbf{u}}{\partial t} + \rho \left(\mathbf{u} \cdot \nabla \right) \mathbf{u} = -\nabla p - \rho \frac{G M_{\odot}}{r^2} \hat{\mathbf{r}}$$

where p is the coronal pressure, G is the universal gravitational constant and r is the heliocentric distance.

[24] This possibility led to unreasonably high pressures as $r \rightarrow \infty$ and was therefore rejected.

[25] Both of those equations apply to particles. A fluid derivation of their equivalent forms is presented in §5.2.

The two preceeding relations may be simplified by applying Parker's assumptions. Eliminating all time derivatives and enforcing spherical symmetry and radial flow yields a continuity equation given by

$$\frac{1}{r^2}\frac{d}{dr}\left(\rho u r^2\right) = 0 \tag{3.1}$$

and a momentum equation given by

$$\rho u \frac{du}{dr} = -\frac{dp}{dr} - \rho \frac{GM_\odot}{r^2}. \tag{3.2}$$

We seek the solution $u(r)$ of these differential equations.

Parker obtained the solution by employing the equation of state $p = 2nkT$ where n and T are the number density and the assumed constant temperature of protons in the expanding corona[26,27]. The student will notice that substitution of this equation of state into Eq. 3.2 will necessitate an expression for dn/dr that describes the variation in proton number density with heliocentric distance. This expression is obtained from Eq. 3.1 (see Exercise 3.3.8) and, upon substitution, yields a momentum equation with $u(r)$ as the only dependent variable. The momentum equation may then be expressed as

$$\left(u^2 - \frac{2kT_i}{m}\right)\frac{1}{u}\frac{du}{dr} = \frac{4kT_i}{mr} - \frac{GM_\odot}{r^2} \tag{3.3}$$

where m is the combined mass of a proton and an electron and T_i is the (assumed constant) coronal temperature. Figure 3.14 shows six solutions to this equation obtained by varying the boundary condition (the flow speed u at the base of the corona)[28]. Consideration of this interesting figure raises many questions, among which are: Solutions IV and V cross at the so-called critical point (r_c,u_c) - what are the physical significances of r_c and u_c? Which of these solutions represent the actual solar wind? Why may the others be rejected?

[26] The factor of two in the equation of state results from assumed equal contributions to the pressure from two species of particles: protons and electrons.

[27] Rather than including a heat-flow equation to simultaneously solve for $T(r)$, Parker, on the basis of certain solar observations, assumed the corona was isothermal above approximately 1.4 solar radii.

[28] The dashed solutions I,II,III and VI in Figure 3.14 are members of *families* of solutions (an infinite number of solutions exist in each family) and the solid solutions that separate the families are singular in that a single boundary condition results in solution IV and a different single boundary condition results in solution V.

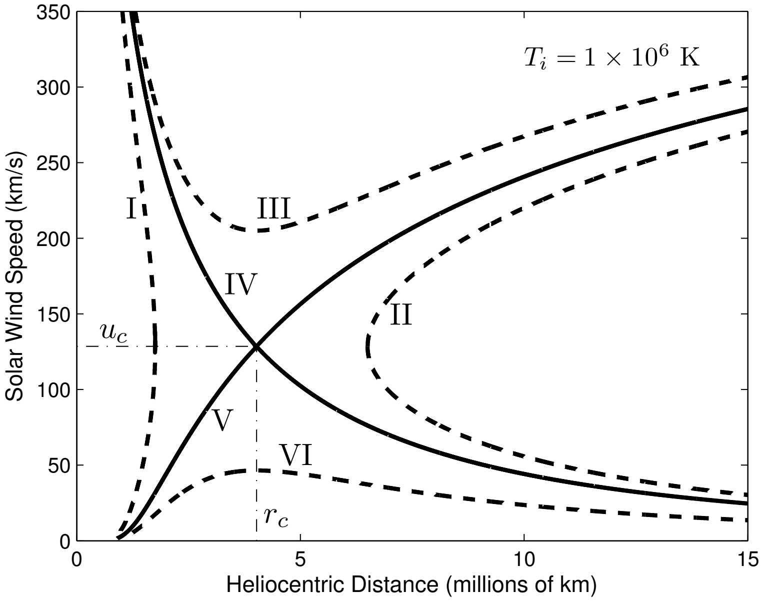

Figure 3.14: Six representative solutions of Equation 3.3 for a million degree solar corona, showing coronal flow speeds as a funciton of heliocentric distance.

To begin answering these questions, let us first identify r_c and u_c. Note that solutions III and VI appear to (and, in fact, *do*) have local extrema where $r = r_c$. Thus, $du/dr\mid_{r_c}= 0$ for these solutions and, setting the RHS of Eq. 3.3 to zero at this same point, we find that

$$r_c = \frac{GM_\odot m}{4kT_i} \approx 5.8R_\odot \qquad (3.4)$$

for a typical coronal temperature of one million Kelvin. Consider now solutions I and II and note that $du/dr \mid_{\substack{u=u_c \\ r\neq r_c}}= \pm\infty$. At $r \neq r_c$ the RHS of Eq. 3.3 is nonzero and finite so that $(u^2 - 2kT_i/m)\mid_{u_c}= 0$ for these solutions and therefore

$$u_c = \sqrt{\frac{2kT_i}{m}}$$

which is the isothermal sound speed in the corona.[29] Thus, for a million Kelvin solar corona, solutions IV and V both pass through the sonic barrier at a heliocentric distance of somewhat less than six solar radii.

We now turn our attention to determining which of the six shown solutions represents the actual solar wind. Solutions I and II may quickly be rejected based on the unphysical double-valued flow speed. It is not physical for the coronal plasma following solution I to flow outward at subsonic speeds from small values of r, reach a maximum distance and then return to small values of r at supersonic speeds. Solution II does not even exist inside the critical radius of $\sim 6R_\odot$. Solution III is entirely supersonic and is rejected on the basis of spectroscopic observations that reveal small Doppler shifts near the base of the corona. Solution IV is supersonic at the base of the corona and is rejected on the same basis. We are then finally left with solutions V and VI as those that may represent the actual solar wind.

Both solutions V and VI have low flow speeds near the base of the corona, in agreement with the previously mentioned observations. They differ in that solution V becomes supersonic beyond the critical radius while solution VI remains subsonic and tends towards zero flow speed as $r \to \infty$. How may we determine which is physical? To foreshadow the answer, let us denote solution V as the solar wind solution and solution VI, due to its lower speeds, as the solar breeze solution. For the solar breeze solution it can be shown that, as the flow speed tends to zero, the density and therefore the pressure tend to an unphysically large constant. Solution VI, the solar breeze solution, is rejected for the same reason as the possibilty of a solar corona in static hydrodynamic equilibrium was rejected. For solution V, the solar wind solution, it can be shown that the density and pressure tend to zero with increasing r and this condition is consistent with the solar wind merging smoothly with the exceedingly low densities of interstellar space.[30] Figure 3.15 shows solar wind speeds (solution Vs) for various coronal temperatures as a function of heliocentric distance. Note that the solar wind at the orbit of Earth for a million degree corona is supersonic with a Mach number of nearly four.

Parker's model of the solar wind includes a number of simlifying assumptions and its validity is open to criticism as a consequence. Signficantly more

[29] For an iosthermal plasma, $\gamma = 1$.

[30] State-of-the-art solar wind models predict that the solar wind does not merge smoothly with the gas of interstellar space (which has a pressure of $\sim 10^{-13}$ Pa). Rather, since the flow is supersonic, it terminates in a shock wave at a boundary known as the heliopause.

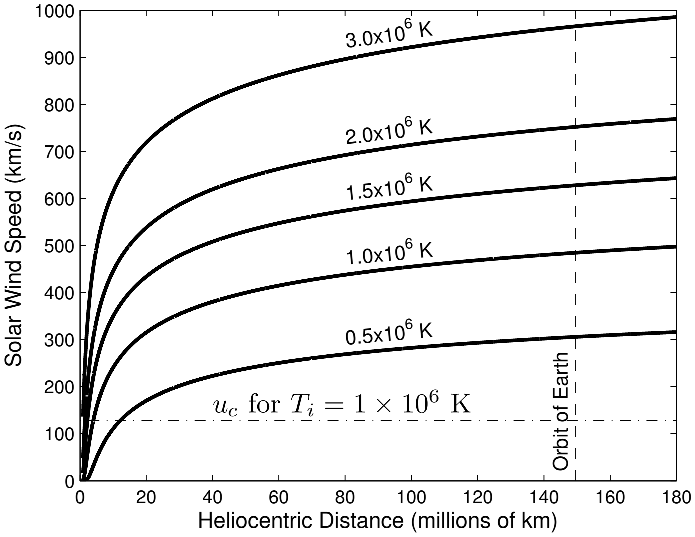

Figure 3.15: Solar wind speed for several coronal temperatures as functions of distance from the Sun

sophisticated models now exist but none call into question the basic nature of the steadily-flowing solar wind and none arrive at significantly different (for our purposes) solar wind speeds at Earth's orbit. Parker's model was a monumental development in the history of space physics and works impressively well.

3.6 Spatial Configuration of the IMF

Having thus conquered the solar wind, aided in part by assumptions that included ignoring the Sun's magnetic field, Parker [1958] turned his attention to the effects of the solar wind on the spatial configuration of that field. In his opening discussion to section V of that paper, Parker noted that no field-free regions are observed on the surface of the Sun so that "each cubic meter of gas flowing outward from the sun is threaded by magnetic lines of force from the main bulk of the sun." The question is: what becomes of these magnetic field lines as they are presumably swept into space with the solar wind to become the IMF? To answer this question, we must first entertain a diversion on the topic of “frozen-in flux”.

3.6.1 Frozen-in Flux

In DC circuit analysis, Ohm’s law takes the form $V = IR$ where V and R are the voltage and resistance across a circuit element and I is the current through the element. The law is an approximation[31] and as the complexity of the circuit increases, the law relating voltage to current also becomes more complex. For example, in an AC circuit containing reactive elements such as capacitors and inductors, the voltage across and the current though a circuit element may not be in phase with each other and a new, more general, form of Ohm’s law is required. This version of Ohm’s law is conveniently expressed in phasor form as $\tilde{V} = \tilde{I}\tilde{Z}$ where the $\tilde{}$ denotes a complex number in phasor form and $\tilde{Z}$ is the impedance. Solving for current, we find that

$$\tilde{I} = \frac{\tilde{V}}{\tilde{Z}} = \tilde{Y}\tilde{V}$$

where $\tilde{Y} = 1/\tilde{Z}$ is the admittance, a measure of how easily and with what phase shift a circuit element allows current to flow in response to an applied voltage $\tilde{V}$.

Currents also flow in plasmas and we desire an Ohm’s law applicable to them. This version of Ohm’s law should be cast in terms of quantities appearing in Maxwell’s equations: instead of a voltage $\tilde{V}$ driving a current $\tilde{I}$ through an element of impedance $\tilde{Z}$ or admittance $\tilde{Y}$, we have the fields $\mathbf{E}$ and $\mathbf{B}$ driving a current density $\mathbf{j}$ through a plasma of conductivity[32] σ. In §7.8, we will derive Ohm’s law from fundamental principles but here it will be abruptly stated in the rest frame of the plasma. In that frame,

$$\mathbf{j}' = \sigma\mathbf{E}'$$

where the $'$ symbols identify quantities in the plasma rest frame. We do not here define σ but suggest that in a collisionless plasma such as the solar wind, the conductivity is so large that it will be insightful to consider it as

[31] Are there any “laws” of physics that are not approximations?

[32] In general, the conductivity of a plasma is a tensor. Here we will treat it as a scalar and take up the more complicated situation in §7.8.

infinite.[33]

Just as the electrons in a perfect conductor almost immediately rearrange themselves in response to an applied static electric field such that the applied field is exactly cancelled, the electric field in the rest frame of an infinitely conductive plasma is also maintained at zero.[34] Most often however, the solar wind is observed from a reference frame other than its rest frame (for example, from the Earth where the solar wind plasma is traveling towards the observer at a million miles per hour) and a transformation is required to account for this difference in reference frames. The required transformation is the Lorentz transformation and, given a typical solar wind speed[35] u, we are justified in taking the non-relativistic limit by ignoring terms in the transformation of order u^2/c^2. The current and electric field then transform as $\mathbf{j}' = \mathbf{j}$ and $\mathbf{E}' = \mathbf{E} + \mathbf{u} \times \mathbf{B}$ where the unprimed quantities are those observed in the non-rest frame and $\mathbf{u}$ is the observed flow speed. Ohm's law then becomes

$$\mathbf{j} = \sigma \left(\mathbf{E} + \mathbf{u} \times \mathbf{B} \right).$$

(As another way of realizing this form of Ohm's law, consider that for a plasma flowing in the presence of a magnetic field $\mathbf{B}$, both terms in the Lorentz force equation ($q\mathbf{E}$ and $q(\mathbf{u} \times \mathbf{B})$) contribute to the plasma motion that results in the current $\mathbf{j}$.)

Given this Ohm's law and an essentially infinite conductivity, the current $\mathbf{j}$ will remain finite only if

$$\boxed{\mathbf{E} + \mathbf{u} \times \mathbf{B} = 0} \tag{3.5}$$

which is an often-made assumption with space plasmas. Here we investigate the consequences of this assumption applied to the interaction between a magnetic field (*e.g.*, the IMF) and a convecting plasma.

Consider a region of space bounded by an open surface S threaded by magnetic flux

$$\Phi = \iint\limits_S \mathbf{B} \cdot d\mathbf{S}.$$

[33] The suggestion is then essentially that collisions limit the rate at which charged particles can move through a plasma. In the absence of collisions, there is nothing to impede the acceleration by $\mathbf{E}'$ of the charged particles, meaning the conductivity is essentially infinite.

[34] Otherwise, the current would be infinite (which is unphysical).

[35] As in the previous section, u is used to denote the speed of a fluid element in the solar wind and is to be distinguished from the speed v of a single particle.

We take this surface to be convecting with the infinitely conductive plasma. It may deform, expand or contract as it convects and we wish to evaluate the change in Φ as it undergoes these motions. That is, we wish to evaluate

$$\frac{d\Phi}{dt} = \frac{d}{dt} \iint\limits_S \mathbf{B} \cdot d\mathbf{S}.$$

Evaluation of this integral is complicated by the fact that the limits of the integral (the surface S) are functions of the differential variable t so that the time derivative may not simply be brought inside the integral.[36] We therefore employ the Leibniz integral rule for three dimensions[37] to obtain

$$\frac{d\Phi}{dt} = \iint\limits_S \left(\frac{\partial \mathbf{B}}{dt} + \cancelto{0}{(\nabla \cdot \mathbf{B})}\mathbf{u} \right) \cdot d\mathbf{S} - \oint_c (\mathbf{u} \times \mathbf{B}) \cdot d\mathbf{c}$$

where $\mathbf{u}$ is the plasma flow velocity with which S convects and $\mathbf{c}$ is a closed contour bounding the surface S. Applying Ampere's law to the first term and Stokes theorem to the last term on the RHS gives

$$\frac{d\Phi}{dt} = -\iint\limits_S (\nabla \times (\mathbf{E} + \mathbf{u} \times \mathbf{B})) \cdot d\mathbf{S} = 0$$

where the last equality must be true given Eq. 3.5. That is, for a perfectly conducting plasma, there is a constant flux through an arbitrary surface convecting with the plasma. We attach to this result the important interpretation that there is no relative motion between magnetic field lines and the

[36] In other words, because S can change shape, $S = S(t)$ and the order of operation cannot be simply exchanged.

[37] This rule is perhaps more well known in one dimension:

$$\frac{d}{dt} \int_{a(t)}^{b(t)} f(x,t)dx = \int_{a(t)}^{b(t)} \frac{\partial}{\partial t} f(x,t)dt + \frac{db(t)}{dt} f(b(t),t) - \frac{da(t))}{dt} f(a(t),t)$$

while the three-dimensional version is:

$$\frac{d}{dt} \iint\limits_{S(t)} \mathbf{F}(\mathbf{r},t) \cdot d\mathbf{S} = \iint\limits_{S(t)} \left(\frac{\partial}{\partial t}\mathbf{F}(\mathbf{r},t) + [\nabla \cdot \mathbf{F}(\mathbf{r},t)]\,\mathbf{u} \right) \cdot d\mathbf{S} - \oint_{c(t)} [\mathbf{u} \times \mathbf{F}(\mathbf{r},t)] \cdot d\mathbf{c}$$

with S, $\mathbf{c}$ and $\mathbf{u}$ defined in the text.

convecting plasma: the magnetic field is *frozen* into the plasma[38,39].

The conclusion of the previous paragraph is too important and too often used in space physics to pass over so quickly. Again, given our assumption of a perfectly conducting plasma, we see that the plasma cannot flow across a magnetic field line (although it can of course flow along it) or, from another reference frame, that a magnetic field cannot diffuse through a perfectly conducting plasma. Although the analogy can be misleading[40] it is often said that

> a perfectly conducting plasma and its magnetic field are frozen together like a bead (the plasma) on a wire or string (the magnetic field line).

We may gain some qualitative insight into the behavior of a magnetized plasma by considering extreme cases for the relative values of the energy densities in the plasma flow and in the magnetic field.

Consider first a case where the plasma flow energy density dominates the magnetic field energy density so that $\rho u^2/2 >> B^2/2\mu_0$. In this case, the plasma flow will dictate the magnetic field geometry. To put it in terms of the bead on a string analogy, a highly energetic bead (the plasma) will easily distort and drag around a nearly tensionless string (the magnetic field)[41]. In the opposite extreme where $B^2/2\mu_0 >> \rho u^2/2$, the magnetic field geometry will dictate the plasma motion. Here it is as if the string is under great tension and the bead is unable to distort it. Of course, between these two extremes the situation is less straightfoward and must be carefully considered.

[38] This interpretation may be formally demonstrated [see *e.g.,* Parks, 2004, pp187-189].

[39] When applying this interpretation, the student must carefully remember that, in general, it is not possible to *uniquely* identify a magnetic field line. In a perfectly conducting plasma, it is true that the magnetic field lines are frozen into the plasma (or, alternately, that an element of the plasma fluid is attached to a given field line) but accepting the frozen-in interpretation implies a field line is defined as that abstract *thing* on which the plasma element is attached. As a hopefully instructive example for the student to think about, consider the drift motion of particles trapped in a dipole magnetic field, paying particular attention to the idea and assumptions of frozen-in flux and the definition of the magnetic field lines.

[40] See the above footnote on interpreting the flux-conserving property of perfectly conducting plasmas.

[41] As we will see in §**??**, the magnetic energy density $B^2/2\mu_0$ manifests itself as both magnetic pressure (in a direction perpendicular to the field line) and as magnetic tension (in a direction parallel to a field line).

3.6.2 The Interplanetary Magnetic Field

Given the conclusions of the previous section, let us accept a solar wind flowing with spherical symmetry radially away from the Sun[42] and the presence of frozen-in solar magnetic field lines carried by it. These magnetic field lines are the IMF and we wish to know something about both its strength and its orientation in space.

To begin, let us consider the locus of points or "path" traced by successive parcels of solar wind plasma flowing radially away from the same source location on the Sun's equator.[43] As each parcel flows outward, the source rotates with the Sun under it so that, as we will see in more detail below, the path takes the shape of an Archimedian sprial. Figure 3.16 shows paths for solar wind parcels launched from four different locations on the source equator. For one of the paths, locations are indicated for six parcels launched from the Sun at equally-spaced time intervals (identified as t_0 through t_5) with the most recent parcel near the source surface and the most distant parcel at 1 AU.[44]

Assuming the solar wind speed is constant, the radial and azimuthal (or longitudinal) positions of a parcel are given by

$$r(t) = r_0 + ut \tag{3.6}$$

$$\phi(t) = -\omega_\odot t + \phi_0 \tag{3.7}$$

where r_0 is a constant equal to the source surface radius, u is the solar wind speed, $\omega_\odot$ is the Sun's rotational angular velocity, and ϕ_0 is a constant specifying the longitude on the Sun's equator from which the parcel was launched. To identify the shape of the path through space (rather than the variation of each component with time as given in Eqs. 3.6 and 3.7), it is necessary to eliminate time from these two equations. Performing the

[42] More modern work including observations has revealed that the solar wind is not spherically symmetric. See, for example, the amazing data collected by the Ulysses satellite that orbits the Sun at high inclination and revealed the presence of a high-speed polar wind.

[43] To be more precise, it should be noted that these parcels are not actually launched from the Sun's "surface" (at $r = R_\odot$); they are launched from a source surface some distance from the Sun. A reasonable approximation is to take this source surface to be at the critial radius given by Equation 3.4.

[44] I am obliged to remark here, as I think everyone is who ever discusses this topic, that the effect is similar to that observed with a rotating garden sprinkler.

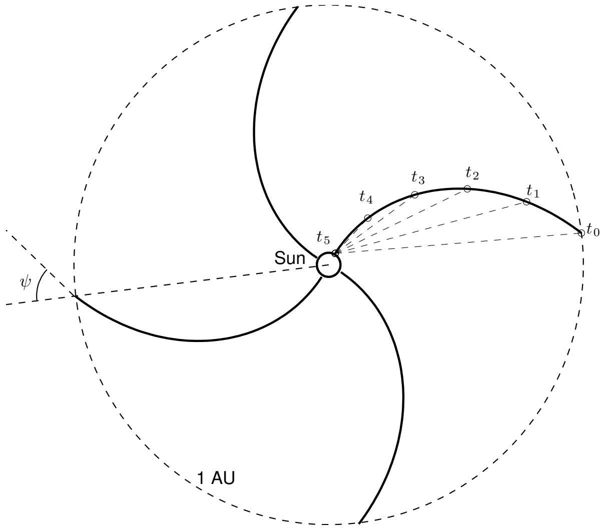

Figure 3.16: Spiral paths of solar wind parcels launched from four equally-spaced locations on the Sun's equator, taking the solar wind speed to be a constant $u = 450$ km/s. For one of the paths, the locations of six successively-launched parcels are indicated. The solar radius has been exaggerated by a factor of 10 (and therefore nearly equal to the source radius) for convenient display and Earth's orbit at 1 AU is indicated by the large dashed circle.

required substitution yields

$$r = r_0 - \frac{u}{\omega_\odot}(\phi - \phi_0)$$

which is the equation of an Archimedian spiral.

Because the Sun's magnetic field is frozen into these plasma parcels, we may expect that the IMF will take the same shape. That is, as each parcel of plasma distorts its local field and carries it in the radial direction, the base of the fieldline rotates with the Sun under it so that the field takes on the same sprial shape discussed above. Continuing this conceptual model, the magnetic field must be parallel to the spiral that, given Equations 3.6 and

3.7, has an unscaled tangent vector[45] **s** given by

$$\mathbf{s} = u\hat{\mathbf{r}} - r\omega_{\odot}\hat{\phi}. \tag{3.8}$$

If ψ is then the sprial angle between an IMF field line (assumed tangent to the path) and the Sun-Earth line as shown in Figure 3.16, Equation 3.8 shows that

$$\tan\psi = \frac{B_\phi}{B_r} = \frac{-r\omega_{\odot}}{u}. \tag{3.9}$$

At the Sun's surface, $r = 0$ and $\psi_{\odot} = 0$ or 180° (the IMF is radial or anti-radial at the Sun)[46] and, as you can show by substituting the appropriate values,[47]

$\psi_{\oplus} \approx 45°$ or $135°$ (the equatorial IMF at Earth makes an angle of approximately 45° with the Sun-Earth line).

While the above discussion reveals the very interesting result that the equatorial IMF is wound up in the shape of an Archimedian sprial, it does not tell us the strength of that field. Both this IMF and the plasma carrying it will interact with Earth's own magnetic field and, in addition, form part of the space environment in which satellites may operate. We therefore wish to estimate the strength of the IMF as a function of radial distance from the Sun.

The equatorial IMF has both radial and azimuthal components and we may find the field strength by solving for each component separately. The IMF must satisfy $\nabla \cdot \mathbf{B} = 0$ which, for spherical coordinates, yields

$$\nabla \cdot \mathbf{B} = \frac{1}{r^2}\frac{d\left(r^2 B_r\right)}{dr} = 0$$

where it has been assumed that $\partial B_\phi/\partial\phi = 0$. Integrating this result gives

$$B_r(r) = B_0\left(\frac{r_0}{r}\right)^2 \tag{3.10}$$

[45] Imagine a particle moving along the path defined by these two equations. Because velocity is always tangent to the path, the tangent vector is parallel to $\mathbf{v} = \frac{dr(t)}{dt}\hat{\mathbf{r}} + r\frac{d\phi(t)}{dt}\hat{\phi}$.

[46] See footnote 50 for an explanation of why there are two possible values.

[47] See Exercise 3.3.9.

where the terms with the '0' subscript are taken at any reference radius. Thus we see that the radial component of the spiral field falls off as $1/r^2$. To find the azimuthal component, we may use Equations 3.9 and 3.10 to find

$$B_\phi(r) = \frac{-r\omega_\odot}{u} B_r = -B_0 \frac{r_0^2 \omega_\odot}{ur}.$$

which shows that the azimuthal component of the IMF fall off as $1/r$. The strength of the IMF is then

$$\boxed{B(r) = B_0 \left(\frac{r_0}{r}\right)^2 \sqrt{1 + \left(\frac{\omega_\odot r}{u}\right)^2}.} \tag{3.11}$$

Taking typical solar magnetic field strengths at $r_0 = r_c$, it can be shown that the strength of the IMF at Earth is consistent with the 7 nT given in Table 3.4 (see Exercise 3.11).

3.6.3 The Heliospheric Current Sheet

So far we have been considering the IMF in the solar "equatorial" plane. While the general results are relevant, we have left a few important points unconsidered. First, the Sun's magnetic moment is not usually aligned with its rotational axis, resulting in an offset between its rotational and magnetic equators[48] that depends on solar longitude. Second, the magnetic equator may have a complicated, or at least a rippled, structure. It is to the consequences of these points that we now turn our attention.

Figure 3.17a illustrates the structure of the so-called helispheric current sheet within the inner solar system where the spiral angle is small (that is, at small distances from the Sun where the IMF is nearly radial). Note that the IMF in the magnetic equatorial plane is directed away from the Sun above the plane and toward the Sun below it. There is therefore a curl to the IMF that, from $\nabla \times \mathbf{B} = \mu_0 \mathbf{j}$ (ignoring the displacement current), results in a clockwise current when viewed from above.[49] This current is in the form of a "thin" sheet, known as the *heliospheric current sheet* with a

[48] As an aid in visualizing the concept of a magnetic equator, consider a dipolar magnetic field that has a magnetic equator along the locus of points where the magnetic field has no radial component.

[49] During the succeeding solar cycle when the Sun's magnetic polarity has reversed, the direction of this current will also have reversed.

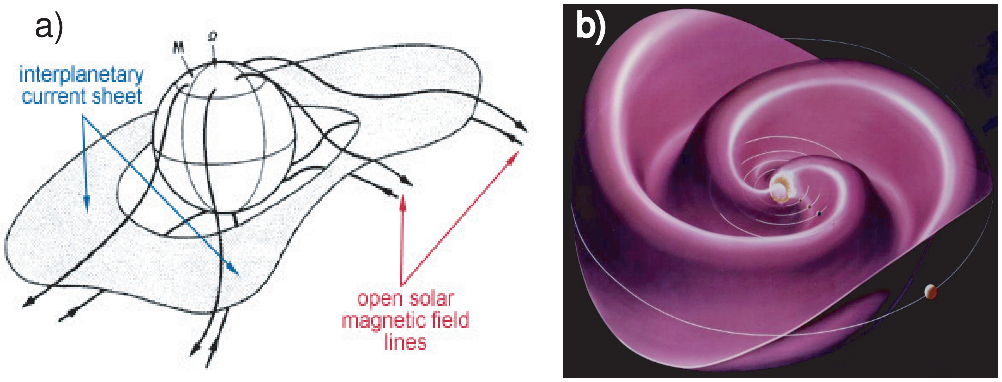

Figure 3.17: a) The heliospheric current sheet (adapted from Smith et al. [1978]). b) The ballerina model (figure from the public domain).

typical thickness of $\sim 5R_E$ flowing around the Sun's magnetic equator. If the magnetic equator is tilted with respect to the Sun's rotational axis, the heliospheric current sheet will also be rotated so that Earth, and the other planets, will be sometimes above and sometimes below the plane. Given the magnetic polarity shown in Figure 3.17a, when a planet is above the current sheet, it will experience an IMF with a positive radial component; when it is below the current sheet, it will experience an IMF with a negative radial component.[50] Thus for a tilted, planar magnetic equator, there are two so-called magnetic sectors. If the magnetic equator is not planar but rippled or having some more complicated structure (as if often the case near solar maximum), there will be four or more magnetic sectors.

Figure 3.17b shows an artist's rendition of the heliospheric current sheet for distances extending to the orbit of Jupiter. Here the spiral angle becomes significant and in addition to the sector structure imposed by a tilted, rippled magnetic equator, the current sheet forms a twisted pattern often described as resembling that of a twirling ballerina's skirt.

[50] This change in sign of the radial component is responsible for the two values of the spiral angle ψ given previously.

3.7 Summary

Our Sun is a main-sequence star with a surface temperature of ∼5800 K. It's energy output is powered by nuclear fusion in its core resulting, at least in part, in a luminosity of 3.85×10^{26} W. Embedded in the Sun and its atmosphere is its magnetic field that, due to the Sun's faster rotation about its equator than its poles, leads to the formation of sunspots, the 11- and 22-year solar cycles and the generation of solar storms that can impact technology near Earth.

While the surface of the Sun is relatively cool at ∼5800 K, its coronal atmosphere reaches temperatures into the millions of Kelvin. This atmosphere continually "boils off" forming the solar wind and carrying away on the order of a billion kilograms of plasma every second. Embedded in this plasma and carried along with it is a remnant of its magnetic field known as the IMF (Interplanetary Magnetic Field).

Both in its quiescent state and during storms, solar photons, plasma, and the IMF fill the solar system and interact with its cellestrial objects. At Earth the solar luminosity has an irradiance of $\sim$1340 W/m^2, the solar wind has a proton density of $\sim$7 cm^{-3} with a nearly equal number of electrons, and speeds of ∼450 km/s. The IMF at Earth has a typical but highly variable magnitude of 7 nT and is oriented nearly parallel to the ecliptic plane but rotated to an angle of ∼45° or 135° relative to the Sun-Earth line.

As the Sun is the main driver of dynamics at Earth, the following chapters examine the impact these solar photons, plasmas, and magnetic fields have on our planet and the technology we have deployed around it.

Exercises

3.1: How many times does the proton-proton chain of fusion reactions occur each second in the Sun?

3.2: What mass of hydrogen is converted into helium each second in the Sun?

3.3: Estimate the required dimensions for a ground-based solar array capable of meeting the United States' electrical power demands.

3.4: Obtain the Sun's zero air mass spectral irradiance data from the web site listed in the text, perform a blackbody fit to those data, and verify the given temperature of ~5785 K.

3.5: Research the historical SSN data and investigate and comment on the Maunder minimum, the Dalton minimum, the Spörer minimum, and the Modern maximum.

3.6: Given the values listed in Table 3.4 (p.77), estimate the Sun's mass loss rate due to particle outflow. Compare the value with the mass loss rate due to its luminosity.

3.7: Complete the missing steps that lead to Equations 3.1 and 3.2.

3.8: Taking I to be the mass per second passing through concentric spherical shells, show from Equation 3.1 that

$$\frac{dn}{dr} = -\frac{I}{4\pi m}\left(\frac{2}{ur^3} + \frac{1}{u^2r^2}\frac{du}{dr}\right).$$

3.9: Use Equation 3.9 to calculate the spiral angle of the IMF at each planet.

3.10: Verify Equation 3.11 using the given expressions for B_r and B_ϕ.

3.11: Use Equation 3.11 to calcualate the strength of the IMF at Earth and compare with the value given in Table 3.4. To do this, take, for example, the reference radius to be $r_0 = r_c$ and research the solar magnetic field to obtain B_0 from sources outside this text.

3.12: Estimate the magnitude of the current flowing in the heliospheric current sheet. Assume the current sheet extends to a distance of 100 AU from the Sun and that the IMF has an average value of 0.05 nT. (Hint: Use Ampere's Law. Answer: $I \approx 1 \times 10^9$ A.)

Chapter 4

Earth's Magnetosphere

4.1 History

Since many hundred years before Christ, it has been known that certain mineral-laden rocks attract iron. Around the year 1000, Chinese investigators discovered that some such rocks, known as *lodestones*[1], always pointed in the same direction when free to pivot in the plane parallel to Earth's surface. Imagine the suprise and delight that followed when it was appreciated that this direction was essentially along the lines of longitude. The magnetic compass had been invented and uses in navigation and timekeeping quickly followed [Merrill and McElhinny, 1983].

Most historians of science would agree that "modern science" began around the year 1600 with the towering works of Copernicus, Kepler and Galileo. Among these giants stands William Gilbert, personal physician to Queen Elizabeth I and, of more immediate interest to us here, an early physicist who struggled to understand the nature of magnetism. His crowning achievement was *De Magnete*, written in Latin (as was the custom in those days) and published in the year 1600.[2] In this work, Gilbert hypothesized that loadstones always pointed in the same direction because Earth itself was a lodestone and that their opposite poles attracted each other. Gilbert may have been incorrect in the details but this was a marvelous step in the history of science: *the Earth is a magnet*. We may now ask two questions.

[1] The word lodestone comes from Middle English and could be interpreted as 'course stone', indicating their usefullness in navigation.

[2] It has since been translated and reprinted in English.

First, where and what is the source of this magnetism and second, what is the nature and extent of the field? We will concentrate on the latter question and give but slight treatment to the former.

4.2 The Source of the Geomagnetic Field

In the middle 1800s, Carl Gauss and Wilhelm Weber organized a "Magnetic Union" to establish a world-wide network of magnetic observatories. For the most part, the observatories were located in a chain across Siberia and at a great number of locations in both the northern and southern hemispheres set up by the British empire. Armed with data from this network, Gauss was able to determine not only the general location of the source of Earth's magnetic field (inside the Earth) but also its strength (more on this later).

The structure of Earth's interior can be divided into three regions: its crust, mantle and core. The crust and mantle are generally nonmagnetic and therefore, because the source of geomagnetic field is inside the Earth, that source must lie within the core. The core is divided into two parts: a solid inner core and a liquid outer core. There are very good reasons to belive that the geomagnetic field does not originate in the solid inner core[3] and so we are left, by elimination if for no other reasons, with the liquid outer core as the source region.

The detailed mechanism by which the geomagnetic field is generated is a topic of current research but it can be safely stated that most geophysicists agree that Earth's magnetic field is generated in the liquid outer core by a magnetohydrodynamic self-exciting dynamo. Were we to leave the reader with such obfuscated terminology, we would be guilty of something very similar to deceit and so a brief, mainly qualitative explanation follows.

Magnetohydrodynamics (or MHD) theory is discussed more fully in §5.2 but for now, let us just say that this is a theoretical framework that describes magnetized, electrically conductive fluids such as plasmas or, as in this case, molten metal. A dynamo is a generator of electric current and thus of magnetic fields and a self-exciting dynamo is one in which the operation of the dynamo strengthens the initial magnetic field through positive feedback. Figure 4.1 illustrates such a dynamo. In this system, a conducting disk is

[3] Two of these are the high temperature of the inner core (that greatly exceeds any expected Curie temperatures) and the fact that the geomagnetic field undergoes occasional reversals.

made to rotate on a conducting axis in the presence of a magnetic field. With the sense of rotation as shown in the figure, the Lorentz $q(\mathbf{v} \times \mathbf{B})$ force acts to produce an excess of positive charge at the outside edge of the disk. This excess charge is made to flow as a current through a solenoid oriented so that its magnetic field enhances the original field responsible for the current. Thus, the dynamo is said to be "self-exciting". Of course, energy must be supplied to maintain the rotation.

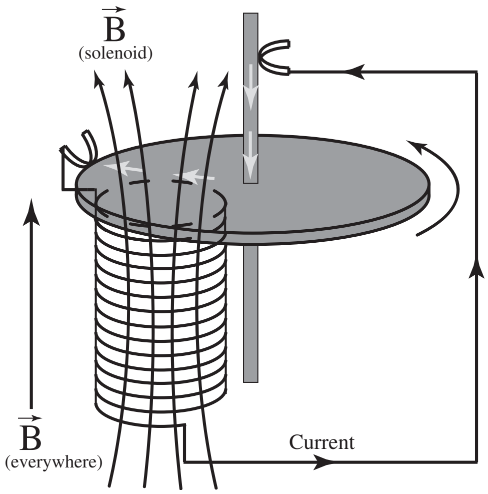

Figure 4.1: A self-exciting dynamo. The Lorentz force acting on charged particles in the conducting disk drives a current through the solenoid that provides positive feedback to the system.

There are no conducting disks in the outer fluid core but the core does support the flow of currents and a great deal of research suggests that a dynamo is responsible for the geomagnetic field. This research is consistent, at least qualitatively, with many observations including polarity reversals and other variations in the field [see, *e.g.,* Glatzmaier and Roberts, 1995, Kono and Roberts, 2002].

4.3 Introduction to the Main Field

4.3.1 The Dipole Approximation

The magnetic field surrounding Earth is the resultant of fields from many different sources. The dynamo-generated internal field discussed above constitutes over 90% of the net field and for this reason it is often referred to as the *main field.* Near Earth, the main field is nearly dipolar where the dipole axis is tilted approximately 11° from the rotational axis and is offset slightly from the center of the planet. The north magnetic pole is located near the south geographic pole and the south magnetic pole is located near the north geographic pole. That is, if we conceptualize the main field as that of a short bar magnet located inside Earth, the bar magnet would be upside down.

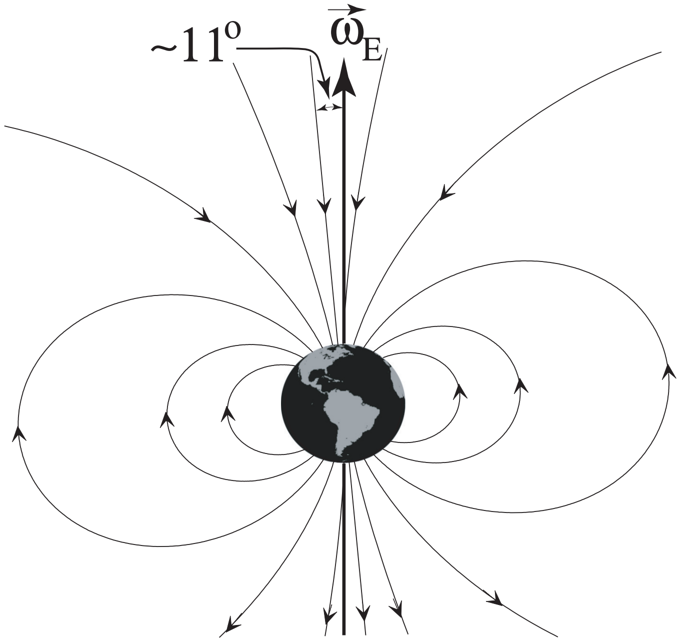

Figure 4.2: An illustration of Earth's main field, apporoximated as a dipole tilted by $\sim 11°$ relative to the spin axis. Note that the north magnetic pole is near the south geographic pole.

The dipole poles are the two points on Earth's surface where the dipole field is perpendicular to the surface. These occur at approximately 78° north and south geographic latitude near Thule, Greenland and Vostok Station, Antarctica. Each of these locations are about 800 miles from the geographic poles. A purely dipole field has no azimuthal component and is expressed in spherical coordinates by

$$\mathbf{B} = B_r\hat{\mathbf{r}} + B_\theta\hat{\theta} \tag{4.1}$$

with

$$B_r = \frac{2M}{r^3}\cos\theta \quad \text{and} \quad B_\theta = \frac{M}{r^3}\sin\theta \tag{4.2}$$

so that

$$\boxed{B = \frac{M}{r^3}\left(1+3\cos^2\theta\right)^{\frac{1}{2}}} \tag{4.3}$$

where r is the radial distance from the center of the dipole, θ is the polar angle measured from the dipole axis and M is the dipole moment.

$$\boxed{\text{Currently, Earth's dipole moment is } M \approx 7.9\times10^{15}\ \ \mathrm{T}\cdot\mathrm{m}^3.}$$

It is useful to obtain an equation that describes the shape of these dipole field lines and, given the tilt and offset of the dipole, this is most easily done in spherical coordinates referenced to the dipole axis rather than to Earth's spin axis. A field line is, by definition, everywhere tangent to the field and so a set of similar triangles reveals that

$$\frac{rd\theta}{dr} = \frac{B_\theta}{B_r} \tag{4.4}$$

which can be integrated (see Exercise 4.1) to obtain the equation of a dipole field line given by

$$r = r_0\sin^2\theta \tag{4.5}$$

where r_0 is a constant equal to the distance from the dipole origin (approximately the center of Earth) to the field line on the dipole "equator" where $\theta = 90°$. Equation 4.5 can be made more useful and intuitive by introducing a few minor changes. Long habit has accustomed us to think of position from the pole in terms of latitude instead of the polar angle θ and so we introduce the quantity $\Phi = \frac{\pi}{2} - \theta$ (or $\Phi = 90° - \theta$ in degrees) that is the *dipole latitude*[4] so that

$$r = r_0\cos^2\Phi. \tag{4.6}$$

[4] Note that dipole latitude Φ is entirely unrelated to the 3rd adiabatic invarient (Equation 2.47) which is regrettably identified by the same symbol.

The final change is motivated by the fact that distances in the near-Earth space environment are often given in terms of Earth radii. For example, the number $6.6R_E$ is familiar to us as the radius of a geosynchronous orbit. And so we wish to rescale Equation 4.6 so that the unit of length is R_E. This is done simply by redefining the constant r_0 as L, the distance in Earth radii from the dipole center to the point where the field line crosses the dipole equator.[5] Our final equation for the dipole field line is then

$$\boxed{r = L\cos^2\Phi.} \tag{4.7}$$

Due to azimuthal symmetry about the dipole axis, the locus of points on field lines with a fixed L value map out a surface known as an $L-$*shell* that can be visualized by revolving a given field line around the dipole axis. Every field line on a given $L-$shell reaches the same maximum distance from the Earth[6] and penetrates Earth's surface at the same dipole latitude. This latitude at which a field line on a given $L-$shell penetrates the surface is known as the *invariant latitude* Λ. An expression for Λ can be obtained from Equation 4.7 by solving, as a function of L, for the latitude at which $r = 1$ (that is, by finding the latitude at which the field line is $1R_E$ from the center of Earth). Thus,

$$\boxed{\Lambda = \cos^{-1}\left(\sqrt{\frac{1}{L}}\right)} \tag{4.8}$$

uniquely relates the dipole latitude at which a field line penetrates the Earth to the maximum distance it reaches from the center of the Earth. Figure 4.3 illustrates field lines and invariant latitudes for L=1,2,3,4,5 and 6.

4.3.2 Magnetic Elements

There are seven commonly used *magnetic elements* that describe the geomagnetic field at any location with respect to a geographic coordinate system. They are

1. Z: The vertical (radial) component, measured positive down.

[5] More properly, we divide Equation 4.6 by R_E so that both r and r_0 are measured in units of R_E.

[6] Here, as we will do in general, we ignore the slight offset between the dipole origin and the center of the Earth.

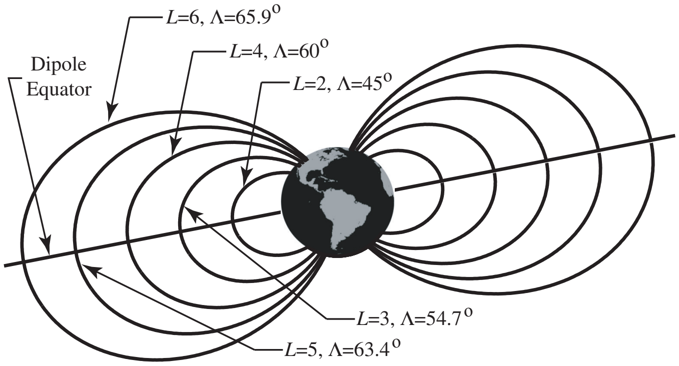

Figure 4.3: Dipole field lines and invariant latitudes for L=1,2,3,4,5 and 6.

2. H: The horizontal (tangential) component, measured along the magnetic meridian.

3. I: The magnetic inclination or dip angle given by $\tan I = \frac{Z}{H} = -2\tan\Phi$. Inclination is reckoned positive downward.

4. B (or F): The total field strength given by $B = \sqrt{H^2 + Z^2}$.

5. X: The geographic northward component of H given by $X = H\cos D$.

6. Y: The geographic eastward component of H given by $Y = H\sin D$.

7. D: The magnetic (or compass) declination given by $\tan D = \frac{Y}{X}$. Magnetic declination is measured positive eastward.

Obviously, these are not all independent. Figure 4.4 illustrates their interrelationships.

4.3.3 Main Field Models

The dipole approximation to the main field is nicely described by Equations 4.1-4.8 but the actual main field is not so straightforward. Complications include the "dipole" not being exactly centered on the center of Earth and the fact that the main field is not purely dipolar. Thus, accurate representation of the main field requires a more detailed description than that of a pure dipole and empirically-constrained spherical-harmonic models are the

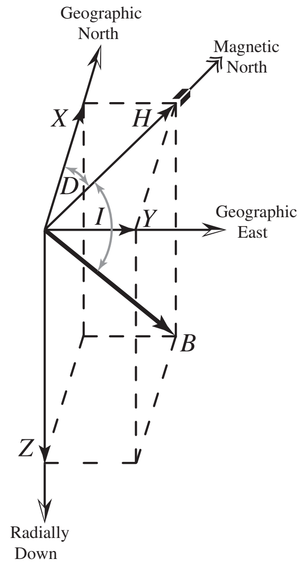

Figure 4.4: The seven magnetic elements, referenced to both a geographic and geomagnetic coordinate system. The dashed lines form a rectange with top and bottom surfaces parallel to the Earth's local surface.

tools of choice. The term "empirically-constrained" here is meant to convey the notion that such models are mathematical fits to recorded sets of observational data. Further complicating the matter, Earth's main field is not constant but is changing with time and so the mathematical fits must account for secular and harmonic trends in the field.

The two most widely-used main field models are the World Magnetic Model (*WMM*) [McLean et al., 2004] and the International Geomagnetic Reference Field (*IGRF*) that, when combined with its set of so-called definitive coefficients, is known as the Definitive Geomagnetic Reference Field (*DGRF/IGRF*) [International Association of Geomagnetism and Aeronomy (IAGA), Division V, Working Group 8, 2003]. The WMM is produced by the National Geophysical Data Center (NGDC) and the British Geological Survey (BGS) and is the standard model for the US Department of Defense, the UK Ministry of Defence, the North Atlantic Treaty Organization (NATO) and the World Hydrographic Office (WHO) navigation and attitude referencing systems. The DGRF/IGRF, on the other hand, tends to be used more extensively by magnetic field modelers and space physicists.[7]

Both of these models employ spherical-harmonic analysis to represent the main field. While it is beyond the scope of this book to describe in detail the methods used to arrive at the models, the student should be familiarized with the process. In essence, the job that must be accomplished is this: given a set of accepted magnetic field measurements distributed over the surface and near Earth at a given time, determine the coefficients on a complete set of orthogonal polynomials so that the polynomial fit represents the observed field as accurately as possible. This must be done for data sets taken at different times so that interpolation can be performed to model the field at times and locations for which data are not available. Both the WMM and the DGRF/IGRF do essentially this and are freely available on the internet where the source code can be downloaded and web-based interfaces are available.[8]

Figure 4.5a shows DGRF/IGRF isomagnetic contours for the year 2000 of the dipole component of the surface main field. The contours are labelled in units of Gauss where $\boxed{1 \text{ G} = 10^{-4} \text{ T}}$ so we see that the dipole field has strengths of approximately 0.3 G near the equator and 0.6 G near the dipole poles. As would be expected for a dipole field (although it is tilted and offset with respect to Earth's rotational axis), the contours are 'orderly' with well-defined poles which happen to be located near Thule, Greenland and Vostok Station, Antarctica[9].

The full DGRF/IGRF model reveals that the surface main field shown in Figure 4.5b is actually quite non-dipolar. The geomagnetic field is much more spatially variable than the dipole approximation would suggest and Thule and Vostok Station, the dipole poles, occupy positions of apparently no special importance when the full field is considered. It is important to note, however, that because the non-dipolar contributions to the main field fall off more rapidly than the dipole contribution, the main field at altitudes spanning some tens of kilometers to a few Earth radii is actually more dipolar

[7] I (Hughes) don't know why one set of organizations uses the *WMM* while the other tends to use the *DGRF/IGRF*. Perhaps you know (or will find out) and be kind enough to tell me.

[8] See, for example: `http://www.ngdc.noaa.gov/geomag/WMM/DoDWMM.shtml` and `http://omniweb.gsfc.nasa.gov/vitmo/igrf_vitmo.html`.

[9] Vostok Station, Antarctica is said to be the location of the of the lowest reliably measured temperature on Earth (-128.6 °F). Although I don't know much about it's history, I do know that it was established in 1957 during the enormously productive *International Geophysical Year* (*IGY)*). Thule, Greenland is home to the USAF's northernmost base.

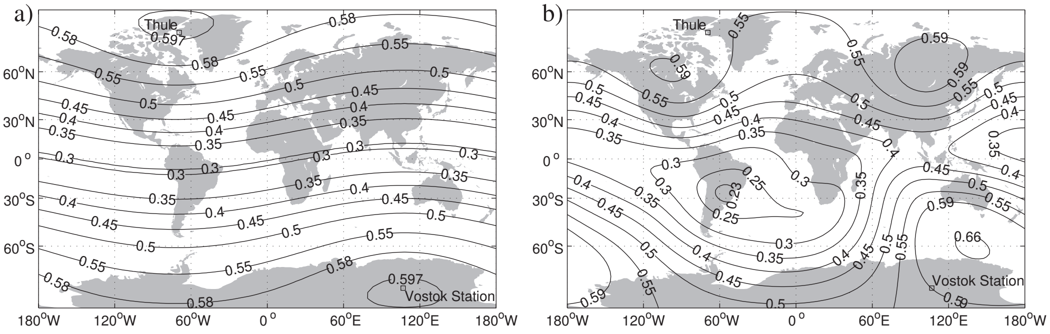

Figure 4.5: DGRF/IGRF isomagnetic contours of Earth's b) main field and a) its dipole component. The contours are labelled in units of Gauss.

than the surface field.[10] For this reason, magnetic phenomena in space (*e.g.*, the aurora, which will be discussed later) are often organized nearly around the dipole poles.

Figure 4.5b shows a relatively deep minimum in the main field near Brazil. This region of weak geomagnetic field is known as the *South Atlantic Anomaly* (SAA) and is largely due to the offset in the dipole field that is centered several hundred kilometers from the center of Earth in the direction away from the SAA, thereby weakening the surface field at the more distant locations. Recall from §2.3 that particles trapped in Earth's "magnetic bottle" execute three periodic motions: gyration, bounce and drift. The relatively weak fields over the SAA lower the mirror ratio (Equation 2.45) and open the magnetic bottle for particles bouncing on field lines in the SAA's range of $L-$shells and longitudes. Significantly, this range of $L-$shells includes the VanAllen radiation belts that, as we will see in §4.4.3, contain highly energetic electrons and protons that pose a serious threat to orbiting spacecraft and astronauts. For example, the Hubble Space Telescope passes through the SAA many times each day and several high-voltage instruments are powered down during each pass to avoid radiation damage. Astronauts passing through the SAA report a higher-than-usual indicence of "shooting stars" in their visual fields as a result of radiation interacting with the optic nerves.

[10] The n^{th} term in a spherical harmonic expansion is weighted by a factor proportional to $r^{-(n+1)}$ where r is the geocentric distance. It is because of this dependence that lower-order terms dominate at larger distances from Earth.

4.3.4 Magnetic Poles

There is a good deal of interest in the notion of Earth's magnetic poles. However, given the surface field shown in Figure 4.5b, it can easily be appreciated that the definition of a "pole" for such a field configuration is not entirely straightforward. Two definitions, each defining a different pole, are often encountered in practice.

1. *Geomagnetic poles* are the two locations on the surface where the best-fit dipole field is vertical.

2. *Dip poles*, also known as the *magnetic poles* are the surface positions where the geomagnetic field is vertical.

Note that the dip and geomagnetic poles differ in location because the geomagnetic field is not purely dipolar. The dip poles can be (and in fact are) identified using direct measurements by locating the positions where the surface field is vertical. It is not possible to locate the geomagnetic poles using direct measurements because there is no unique feature of the geomagnetic field that indicates the locations of the best-fit dipole field poles. The locations of the geomagnetic and dip poles can be approximated using a geomagnetic field model (such as the IGRF) but modeled dip poles are found to differ somewhat from the results of direct measurement.

Figure 4.6 shows locations from the IGRF model of the north and south dip and geomagnetic poles for the years 1900-2010 [Data from British Geological Survey, 2009][11]. The positions of the dip poles change rapidly in response to solar wind and IMF conditions and both the dip and geomagnetic poles have secular trends that are clearly visible in Figure 4.6.

At least during the past 100 years or so, the dip poles have been moving at a much higher average speed than the geomagnetic poles.[12] Over the past $\sim$100 years, the northern hemisphere dip pole has moved with an average speed of about 15 km/yr. The southern hemisphere dip pole as moved about half as far over the same time. Interestingly, it is also fairly clear from Figure 4.6 that the position of the northern hemisphere (southern magnetic) dip pole is accelerating. These changes in the geomagnetic field are presumably related to the amazing phenomena of magnetic reverals.

[11] Also see: `http://www.geomag.bgs.ac.uk/education/poles.html`.

[12] This would indicate that the geomagnetic field's higher order harmonics are changing at a more rapid rate than the dipole contribution.

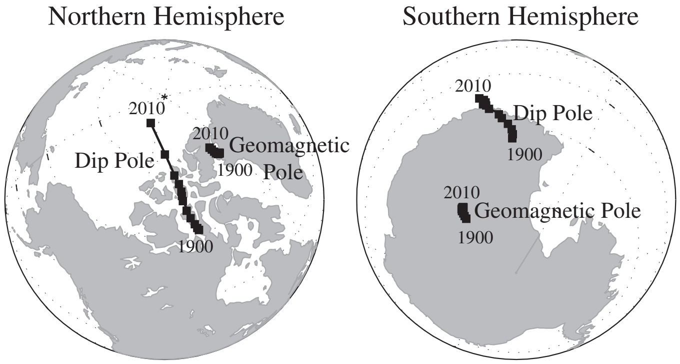

Figure 4.6: Locations of the IGRF model geomagnetic and dip poles for the northern and southern hemispheres for the years 1900-2010.

4.3.5 Magnetic Reversals

The polarity of the solar magnetic field reverses about every 11 years near the peak of solar max. Evidence suggesting a corresponding phenomenon on Earth was discovered as early as 1904 when lava flows were observed to contain remnant magnetizations roughly opposite to that of the present geomagnetic field [David, 1904, Brunches, 1906]. It is by no means a trivial task to verify that such reversed magnetizations actually indicate geomagnetic field reversals rather than the results of processes internal or local to the rocks being studied. However, the task has been undertaken by generations of researchers and the overwhelming body of evidence indicates that Earth's magnetic field does indeed undergo polarity reversals [Jacobs, 1994].

The paleomagnetic evidence indicates that the polarity of the geomagnetic field has reversed many hundreds of times and that reversals occur, on average, about every 250,000 years with the reversals being being accompanied by dramatic changes in geomagnetic field intensity and, of course, direction. Typically, reversals are accomplished over the course of a thousand or more years [Jacobs, 1994]. There is currently no uncontested evidence that links geomagnetic polarity reversals with biological extinctions.

The geomagnetic field is currently in an unusually long period of constant polarity with the last reversal occurring some 780,000 years ago. Possibly suggestive of an imminent (on geologic time scales) reversal, the intensity of the current geomagnetic field is decreasing at an accelerating rate and is now

$\sim$35% lower than its maximum values of about 2000 years ago.[13]

4.4 The Magnetosphere

The geomagnetic field is the net magnetic field around Earth resulting from the vector superposition of all naturally-occurring magnetic fields including the main field, other crustal and mantle fields, and fields produced by currents flowing in the near-Earth space environment. It was appreciated as early as 1930 [Chapman and Ferraro, 1930] that the geomagnetic field would serve as an obstacle to the oncoming solar wind plasma,[14] around which it would have to flow in a manner somewhat analogous to water in a stream flowing around a rock. In this section, we will investigate this solar wind-geomagnetic field interaction to understand the structure of the resulting *magnetosphere* and the currents that support it.

4.4.1 Formation of the Magnetosphere

Consider first the geomagnetic field to be that of a dipole and assume there is no solar wind. The dipole will extend into space as the field strength falls off as $1/r^3$, as we have previously seen. Our goal here will be to add the solar wind, allow it to interact with the geomagnetic field, and understand qualitatively why the magnetosphere results.

Solar wind and magnetospheric plasmas are highly conductive so that the frozen-in flux theorem discussed in §3.6.1 holds. This theorem can often be taken to imply that once a charged particle begins gyrating around a particular field line, it must always remain on that field line.[15] Many R_E in front of Earth in the direction of the Sun, the solar wind ram pressure greatly exceeds (typically by a few orders of magnitude) both the solar wind thermal pressure and the magnetic pressure of the geomagnetic field. Thus there is more energy density in the particle flows than in the field and, as with the case of the solar wind's interaction with the solar atmosphere, particle

[13] This very interesting topic is worthy of more discussion than is given here.

[14] Recall that strong observational evidence for this plama was reported by Cuno Hoffmeister (involving comet tails) in 1943.

[15] But recall that they *don't* do this. Frozen-in-flux was obtained from a *fluid* description. Nevertheless, the analogy is both tempting and helpful (unless it leads to incorrect physics (which it can), in which case it is both tempting and harmful.)

motions control the magnetic field topology. Figure 4.7 illustrates how this results in the formation of the magnetosphere.

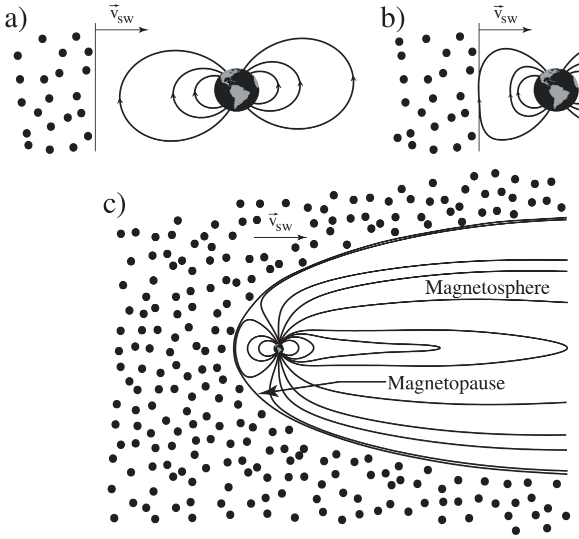

Figure 4.7: An illustration of the formation of the magnetospheric cavity by the solar wind. a) A dipole approximation to the geomagnetic field in vacuum, before an oncoming front of solar wind plasma interacts with the shown field lines. b) The dayside magnetic field lines are compressed as by the solar wind. c) The solar wind flows flows around the magnetosphere, dragging some field lines with it into a long tail.

Figure 4.7a illustrates a situation where, after being "switched off" for some time, an oncoming front of solar wind plasma is about to interact with the approximately dipolar vacuum geomagnetic field. In panel b), the solar wind plasma has reached the geomagnetic field and, because it cannot flow across the field lines because of frozen-in flux, it exerts a ram pressure on them that compresses the field on the dayside. As this dayside field is being compressed, the field strength and magnetic pressure increase until the point is reached where the magnetic pressure equals the ram pressure and the particles no longer control the magnetic field topology. The solar wind plasma must then flow around the geomagnetic field as illustrated in panel c). As the flow is diverted, the dipole field geometry results in some field lines being "dragged" with the flow into a long "tail". The final result is a magnetosphere that is compressed on the dayside and elongated on the nightside. The magnetosphere is separated from the solar wind by a pressure-balance boundary known as the *magnetopause.*

The task of determining the exact shape of the magnetopause boundary is difficult and, as pointed out by Baumjohann and Treumann [1997, pp187-188], requires the numerical solution of a second-order three-dimensional non-linear partial differential equation. We will not engage this problem here because, as we will very shortly see, there are excellent and accessible models for the magnetospheric field lines that will elucidate details of the magnetosphere's structure.

4.4.2 Size of the Magnetosphere

Although the task of determining the exact shape of the magnetopause is beyond us, two obtainable quantities go a long way towards illustrating the size of the magnetosphere: the so-called magnetopause standoff distance and the limiting width of the tail.

The magnetopause standoff distance is the distance from the center of Earth to the front, or *nose* of the magnetopause and a derivation of an expression for this quantity is useful and instructive. Considering the nose of the magnetopause shown in Figure 4.7c to be in equilibrium, we may equate the external with the internal pressures to obtain an expression for the equilibrium position. There are three external pressures: the magnetic pressure of the IMF, the solar wind thermal plasma pressure, and the solar wind plasma ram pressure. The internal pressures are the magnetospheric magnetic and plasma thermal pressures. Pressure balance at the magnetopause nose requires

$$p_{ext} = p_{int}$$

where p_{ext} and p_{int} are the total external and internal pressures, respectively. Substituting the pressures mentioned above yields

$$\epsilon_{B_{IMF}} + \epsilon_{sw_{thermal}} + \epsilon_{ram} = \epsilon_{B_{mp}} + \epsilon_{mp_{thermal}} \tag{4.9}$$

where ϵ is an energy density that, given the geometry under consideration, is proportional to pressure.[16]

[16] "Energy density is proportional to pressure" is an interesting phrase that you should investigate. Under certain circumstances it is true. Under others, it is not.

Our task is now to evaluate each term in Equation 4.9 and use the result to determine the equilibrium position. Let us first turn our attention to the LHS terms. As we saw in Table 3.4, under typical solar wind conditions at 1 AU, $\epsilon_{sw_{thermal}} \approx nk(T_p + T_e) \approx 30$ pPa and $\epsilon_{B_{IMF}} = B_{IMF}^2/2\mu_0 \approx$ 15 pPa.[17] The remaining LHS term, the solar wind ram pressure, is the momentum flux density of the bulk solar wind plasma flow absorbed by the magnetopause nose. That is, the ram pressure is just the flow momentum absorbed by the magnetopause nose per unit area per unit time. As shown below, $\epsilon_{ram} = K\rho_{sw}v_{sw}^2 \approx 3$ nPa where $K \approx 0.9$ is an efficiency factor and ρ_{sw} is the solar wind mass density. This ram pressure, while also very small, is approximately two order of magnitude larger than the solar wind thermal and IMF magnetic pressures and these two latter pressures can therefore be neglected in evaluating Equation 4.9.

An expression for solar wind ram energy density (or pressure) is obtained by evaluating the quantity $\epsilon_{ram} = \Delta P_{sw}/(A\Delta t)$ where ΔP_{sw} is the solar wind momentum absorbed by an area A of the magnetopause in time Δt. Figure 4.8 illustrates a constant flux of solar wind with speed[18] u_{sw} and mass density ρ_{sw} on the magnetopause nose. In a time interval Δt, all solar wind particles contained in the shaded box will impact the magnetopause nose and, if they stagnate there, all of their momentum will be absorbed. In reality, the flow does not stagnate and only the momentum fraction K is absorbed with the remainder flowing around the magnetopause nose rather than being absorbed by it. Thus,

$$\begin{aligned}\Delta P_{sw} &= KP_{sw} = Km_{box}u_{sw} = K\rho_{sw}V_{box}u_{sw} \\ &= K\rho_{sw}A(u_{sw}\Delta t)u_{sw} = K\rho_{sw}u_{sw}^2 A\Delta t\end{aligned}$$

where m_{box} is the mass of solar wind particles contained in the shaded box and V_{box} is the volume of the shaded box. The solar wind ram pressure is then

$$\epsilon_{ram} = \frac{K\Delta P_{sw}}{A\Delta t} = K\rho_{sw}u_{sw}^2.$$

Using numbers typical of the solar wind, one finds that $\epsilon_{ram} \approx 3$ nPa.

On the magnetospheric side of the magnetopause (the RHS of Equation 4.9), it is assumed that $\epsilon_{B_{mp}} >> \epsilon_{mp_{thermal}}$ so that the thermal pressure may

[17] These are very tiny pressures indeed!

[18] Assuming this value (along with others in this derivation) is constant.

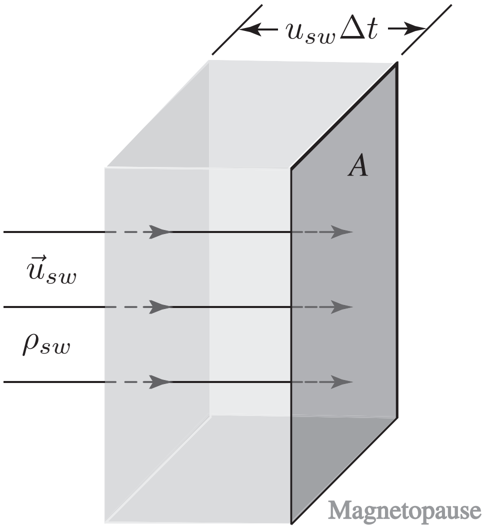

Figure 4.8: Solar wind ram pressure equals an efficiency factor times the momentum of all solar wind particles in the shaded box per unit area per unit time.

be neglected.[19] Equation 4.9 then reduces to

$$K\rho_{sw}v_{sw}^2 = \frac{B_{mp}^2}{2\mu_0}. \tag{4.10}$$

In the final tally, the boundary between the solar wind and the magnetosphere is a pressure balance surface between the solar wind ram pressure and the magnetopause magnetic field pressure.

We are now in position to inquire about B_{mp}, the strength of the geomagnetic field at the magnetopause nose. If the field were dipolar and knowing the field strength at the surface of Earth, we could simply apply a $1/r^3$ falloff from the center of Earth to obtain a general expression for B_{mp}. However, the dayside field is not dipolar but is compressed by the solar wind ram pressure. It is not feasible to derive an exact expression for the compressed field strength and we account for it by simply introducing a compression factor a by which to multiply a dipolar field. The constant a will certainly be larger than 1 and, for the time being, 2 is a reasonable estimate[20] that we will make

[19] Spacecraft observations justify this assumption.

[20] A value of $a = 2$ would be obtained if the field were compressed by an image dipole.

more precise later.

Equation 4.3 shows that at the magnetopause nose, which we assume lies on the dipole equator,

$$B_{mp} = aB_{dipole} = \frac{aM_{dipole}}{r^3} = \frac{aB_0}{L_{mp}^3}$$

where L_{mp} is the magnetopause nose $L-$shell and B_0 is the surface field strength. Equation 4.10 then gives

$$K\rho_{sw}u_{sw}^2 = \left(\frac{aB_0}{L_{mp}^3}\right)^2 \frac{1}{2\mu_0} = \frac{(aB_0)^2}{2\mu_0 L_{mp}^6}$$

that may be solved for the quantity we are interested in:

$$\boxed{L_{mp} = \left(\frac{(aB_0)^2}{2\mu_0 K\rho_{sw}u_{sw}^2}\right)^{\frac{1}{6}}}$$

which is the dimensionless distance (in R_E) from the center of Earth to the magnetopause nose. Combining all the constants in the above expression, taking $a = 2.44$ which measurements indicate is typical, and changing to the system of units most often used when reporting solar wind conditions, this expression reduces to

$$\boxed{L_{mp} \approx 107.4\left(n_{sw}u_{sw}^2\right)^{-\frac{1}{6}}} \tag{4.11}$$

where n_{sw} is the effective solar wind proton number density [21] in cm^{-3} and u_{sw} is the solar wind speed in km/s. For typical solar wind conditions, we then find that

$$\boxed{L_{mp} \approx 10.}$$

The above result indicates that the nose of the magnetopause has a typical standoff distance of approximately $10R_E$. This is an important number and perhaps most important because it exceeds the $6.6R_E$ orbital radius of the geostationary satellite fleet. Under normal circumstances, these satellites are

[21] The "effective" proton number density is somewhat higher than the actual proton number density to account for the presence of some He^{++} in the solar wind with its mass factor of 4.

within the magnetosphere and shielded by it from many of the radiations of space.

The limiting width of the magnetosphere tail may be estimated in a way very similar to what was done above. As the solar wind flows down the tail, it exerts essentially no ram pressure on the tail and only those terms due to the solar wind thermal pressure and IMF magnetic pressures need to be considered. Equating these pressures with typical tail magnetic field pressures, one finds that $R_T \approx 1.8 L_{mp}$ [Baumjohann and Treumann, 1997, p.189] where R_T is the perpendicular distance from the center of Earth to the sides of the magnetopause. That is, under typical conditions, the limiting width of the magnetotail is about $2 \times 18 R_E$. We can therefore approximate the size of the magnetosphere as illustrated in Figure 4.9.

The magnetosphere extends $\sim 10 R_E$ into space towards the Sun and reaches a width of $\sim 2 \times 18 R_E$ across its flanks. The downstream tail of the magnetosphere reaches to some $1000 R_E$ in the antisunward direction.

All of Earth's geomagnetic field lines are contained within this volume.[22]

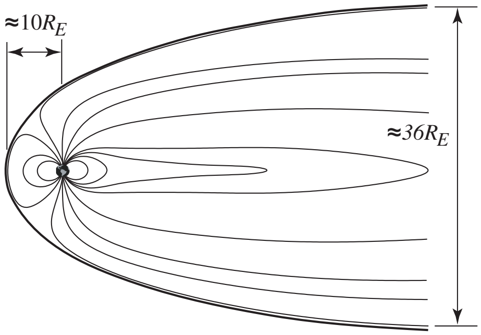

Figure 4.9: Approximate size parameters for Earth's magnetosphere.

[22] Earth's magnetosphere encompasses a huge volume (but Jupiter's encompasses much more!). Our moon orbits Earth at a typical radius of $60 R_E$ so that it is sometimes inside and sometimes outside the magnetosphere.

4.4.3 Structure of the Magnetosphere

The magnetosphere is a complex system consisting of many regions having differing magnetic field geometries and plasma properties. In order to identify and discuss these different regions, we must first obtain an accurate representation of the magnetic field throughout the magnetosphere. This is done most conveniently with a magnetospheric magnetic field model and, of those that have been developed and are in use, among the most popular are the series of models by Nikolai Tsyganenko [*e.g., Tsyganenko, 1989, 1995, Tsyganenko and Sitnov, 2005]*. The Tysganenko models are semi-empirical best-fit representations of the magnetospheric magnetic field based on satellite observations and solutions of Maxwell's equations. The codes for these models are available online[23] and Figure 4.10 illustrates several field lines obtained with a Tsyganenko model that illustrate the large-scale structure of the magnetosphere.

Bow Shock

The solar wind approaches the magnetosphere with a typical speed of 450 km/s. If this number is not shockingly large, perhaps it will be after conversion into the more familiar units of miles per hour: The solar wind approaches the magnetosphere with a typical speed of a million miles per hour!

It is tempting at this point to say something like, "Well, such a high speed is clearly supersonic,[24] so just as an airplane traveling supersonically moves with a shock wave in front of it, the magnetsophere must have a shock wave in front of it." Tempting as such an analogy may be, it fails on a critical point. An airplane travels through the air, a collisional gas that supports sound waves (thus the term super*sonic*). The solar wind at 1 AU is essentially collisionless and therefore does not support sound waves. Where there are no sound waves to travel sonically there is no consequence to things traveling supersonically. Nevertheless, there is in fact a type of shock wave standing in front of the magnetopause (called the *bow shock* and shown in Figure 4.10) and its effects and reasons for forming there are interesting. To motivate and guide our thinking about why there could or should be a shock in the collisionless solar wind plasma, let us first think about shocks in an

[23] See, e.g., `http://geo.phys.spbu.ru/ tsyganenko/modeling.html` or `http://www.igpp.ucla.edu/public/tpoiii/Pliny/MATLAB/Tsyg/`.

[24] Recall the solar wind sound speed of ~ 60 km/s given in Table 3.4.

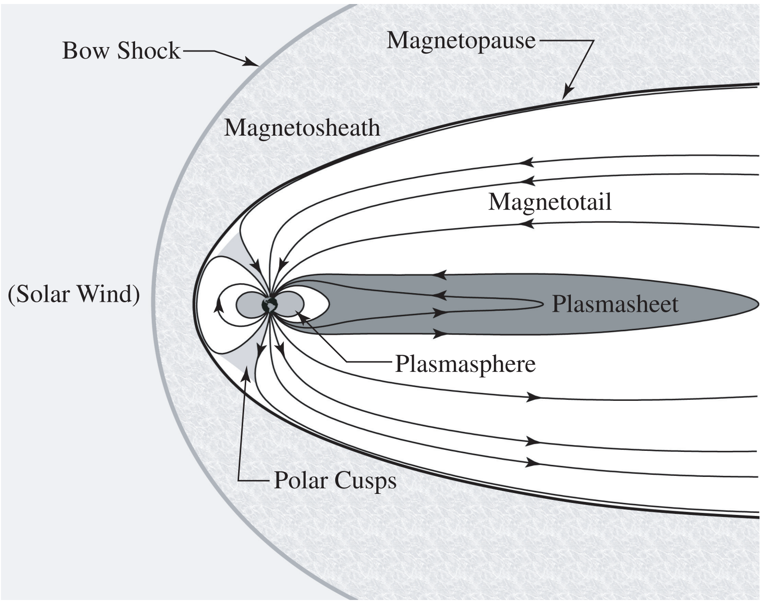

Figure 4.10: Cross section of the magnetosphere identifying regions as determined by magnetic field geometries and properties of the local plasma populations. Earth's radiation belts are not pictured.

abstract and qualitative sense as framed by the following analogy.

Suppose you plan a dinnertime visit to a friend's home across town. You compose a letter stating your plan, put it in the mailbox, and immediately hop into your car and drive to his home. Of course you reach there before the letter does and your friend is awkwardly surprised to find you expecting dinner. The awkwardness occurred because you travelled faster than the letter supplying the information required for a more typical reception.

Suppose further that the laws of human interaction allow for no awkwardness.[25] How may the awkwardness of your sudden arrival be avoided? There are many ways and the supposed law against awkwardness would require you to bring at least one into effect. Perhaps as you hop into your car after mailing the letter, you also send a text or email containing the equivalent information. Your friend would then receive it in time to avoid the awk-

[25] Snicker, snicker. But the physical side of this analogy is that the laws of classical physics tend to provide for smooth transitions between states (*i.e.,* no "awkwardness").

wardness. On the physical side of this analogy, the mailed letter corresponds to usual physical processes that convey information (*e.g.*, sound waves) and the text/email corresponds to an unusual way to convey information (*e.g.*, a shock wave).

Consider now an object moving through some medium. The medium must[26] flow around the object to avoid adding its negative relative momentum to the object through direct collision and eventually stopping its motion.[27] If this flowing-around-rather-than-colliding is to happen, information about the object's impending arrival must travel upstream ahead of the object itself so that the medium has time to respond appropriately. In a medium that supports sound waves, this information may be carried by a compressional sound wave so that a pressure pulse causes the required flowing-around effect through the laws of, for example, fluid mechanics and thermodynamics.

But what if the object moves faster than the speed at which information is typically transmitted? In this case, nature avoids the 'awkwardness' of discontinuous transitions and the relevant physical processes result in a faster way of transmitting the required information. The information is carried by a *shock wave* that travels ahead of the object. The effect of the shock wave is to, among other things, slow down the flow speed so 'ordinary' processes can provide for the required flow around the object. We may then ask what these 'ordinary' processes are and with what speed they propagate.

Although the collisionless solar wind does not support sound waves, it does support many other types of waves. In particular, three waves known as Alfvén waves and the fast and slow magnetosonic waves exist in the solar wind's highly conductive magnetized plasma. A detailed development of these waves is beyond the scope of this section and we simply state without justification that our magnetosphere's bow shock is a fast magnetosonic shock with a typical fast magnetosonic Mach number of $\boxed{M_{ms} \approx 8}$ (from which you can calculate a typical wave speed of ~ 60 km/s).

The bow shock is an irreversible (entropy-increasing) wave that causes a transition from super-magnetosonic to sub-magnetosonic flow. Upon interacting with the shock wave, the plasma flow speed decreases but mass,

[26] Of course, *must* is too strong of a word. What is about to be described tends (approximately) to happen but we are not free to dictate that nature *must* work this way.

[27] That is, the medium must get out of the oncoming object's path to avoid a collision (and a possible pileup of mass) that slows down the object. To avoid creating a trailing vacuum with the same end result, the medium must also flow into the vacancy left behind by the passing object.

energy and momentum are conserved across the shock boundary.[28] So-called *Rankine-Hugoniot relations* based on these conserved quantities can be derived that relate upstream to downstream plasma parameters. For example, conservation of mass flux across the boundary requires that

$$\frac{\partial(\rho_m u_n)}{\partial n} = 0$$

where n is the shock-normal direction, ρ_m is the mass density and u_n is the velocity component normal to the shock. Thus

$$(\rho_m u_n)\,|_{upstream} = (\rho_m u_n)\,|_{downstream}$$

or, to use a common notation,

$$[\rho_m u_n] = 0$$

where the square backets indicate the difference between upstream and downstream quantities. We see that because the downstream speed is lower than the upstream speed, the downstream mass density (of the shocked solar wind) must be higher than the upstream mass density. To conserve mass flux across the bow shock, the plasma is more dense on the downstream side than on the upstream side. Other conservation relations are developed by Kivelson and Russell [1995, ch. 5] and the reader is highly recommended to this interesting discussion.

The geocentric distance to the nose of the bow shock is about 30% greater than to the magnetopause nose and it thus has a standoff distance of approximately 13 R_E under typical solar wind conditions [Baumjohann and Treumann, 1997, pp.192-193].

Magnetosheath

As illustrated in Figure 4.10, the region between the bow shock and the magnetopause is known as the *magnetosheath* and is, for the most part, populated with the shocked solar wind plasma and IMF[29]. The ratio of particle

[28] If mass were not conserved across the boundary, the boundary would become more (or less) massive over time. This is neither expected nor observed. Energy is conserved across the boundary because physicists believe in energy conservation more than almost anything else (although energy may of course change form from, for example, directed flow kinetic energy to the random motions of thermal energy). Momentum is conserved across the boundary because the boundary exerts no net force on the plasma.

[29] A small fraction of the magnetosheath plasma comes from the magnetosphere.

densities in the sheath to those in the solar wind range from approximately 4 near the bow shock nose to around 2 at locations far from the nose. Plasma flows in the magnetosheath are highly turbulent with average speeds that vary strongly with position and, most importantly, with distance from the nose. As the solar wind is shocked upon entry into the magnetosheath, much of its kinetic energy (which in the solar wind was largely contained in the directed flows) is transferred into thermal energy and the temperature of magnetosheath plasmas can be as much as 20 times higher than in the solar wind. Thus, as compared to the upstream solar wind, the magnetosheath is populated with relatively dense, turbulent, slow-moving, hot plasma.

Magnetopause

The magnetopause was discussed in §4.4.2. It is the pressure-balance boundary between the solar wind and the magnetosphere, inside of which all of the geomagnetic field lines are contained.

Polar Cusps

Turning our attention to regions inside the magnetopause, inspection of Figure 4.10 revels several distinct magnetic field geometries. Note first the *polar cusps* which are the high latitude regions in the northern and southern hemispheres where magnetic field lines transition from dayside field lines[30] to those that are swept past Earth into the tail. The two cusps are somewhat elongated in latitude and longitude and their magnetic field geometry suggests they link the magnetopause to the inner magnetosphere and the underlying atmosphere. That is, shocked magnetosheath plasma can flow along cusp field lines and have direct-entry access to the low altitude regions of Earth. We might then expect a continual flow of plasma down these field lines into the lower atmosphere, a topic we will take up again in §**??**.

Magnetotail

Magnetospheric field lines are classified as either "open" or "closed". Closed field lines emerge from the southern hemisphere and return to the northern,

[30] The term 'dayside field lines' refers to those field lines that cross the geomagnetic equator on the dayside. Such field lines may, for example, exist on the nightside (in the dark) at 1 R_E but by definition must cross the geomgnetic equator at sunlit locations.

penetrating the surface at a so-called *conjugate* location. All field lines in a dipole approximation are closed. Open field lines, on the other hand, do not link one hemisphere with the other. They 'leave' or 'enter' Earth's surface from one hemisphere[31] and do not connect to the other. Instead, they gradually and approximately align with the Sun-Earth direction and *merge* or *reconnect* with the IMF some hundreds of R_E downstream from Earth. The interesting subject of magnetic reconnection will be taken up in §5.2.1.

The magnetotail is the region of the magnetosphere having open field lines. Because the field lines are open and plasma is free to stream along the field lines, down the tail, and to eventually merge again with the solar wind, the magnetotail is characterized by plasma of low electron density (0.1 cm^{-3} or less) and energy (typically less than 0.5 keV). In the near-Earth tail, magnetic field strengths of around 20 nT are typical and the field strength gradually weakens until it approaches that of the IMF in the distant tail.

Magnetotail field lines converge in the high-latitude regions into what is known as the *polar cap.* Solar wind electrons streaming along these field lines with energies of a few hundred eV precipitate into the polar cap and form a spatially homogenous *polar rain.* The magnetotail is approximately circular in cross-section and is separated into two *lobes*, the northern and southern lobes, by a region of relatively dense plasma known as the plasmasheet.

Plasmasheet

The region of closed field lines near the center of the tail is known as the plasmasheet due to the closed field line geometry and the consequent higher electron densities of $\sim$ 0.5 cm^{-3}. The solar wind is the main source of electrons and protons in the plasmasheet and these particles must first pass through the tail before becoming part of the plasmasheet population. As they do so, the particles are energized in the tail to typically $\sim$ 0.6 keV for electrons and $\sim$ 5 keV for protons. The plasmasheet is typically 2-6 R_E thick and is a very important source of the auroral particles we will discuss in §**??**.

[31] Given the polarity of Earth's magnetic field, field lines 'leave' from the southern hemisphere where the field has a positive $\hat{\mathbf{r}}$ component and 'enter' the northern hemisphere where the same component is negative.

Plasmasphere

The final magnetospheric region identified in Figure 4.10 is the *plasmasphere* which may be thought of as the magnetospheric extension of the ionosphere to be discussed in Chapter 7. The plasmasphere is a region of closed field lines forming a torus of cold (~ 1 eV), dense ($10^2 - 10^3$ electrons cm^{-3}) plasma that maps to "subauroral" field lines, or field lines that map to Earth at latitudes equatorward of where the aurora is observed. In contrast to the rest of the magnetosphere, which is generally aligned along the Sun-Earth direction, the plasmasphere almost corotates with Earth, taking on average 27 hours to complete a full rotation.

Radiation Belts

Explorer I was the first successful American satellite. It was launched from Cape Canaveral on January 31, 1958 at 10:48 PM EST and carried a scientific payload centered around James Van Allen's cosmic ray[32] detection instrumentation. Van Allen's telling of why and how his instrumentation came to be onboard Explorer I is interesting and insightful [see, *e.g.,* Gillmor and Spreiter, 1997, pp.235-251]. As is well known, data from Exlorer I revealed that "My God, space is radioactive!"[33] Van Allen and his team had discovered what would come to be called the Van Allen Radiation Belts.

Particles trapped on Earth's closed magnetic field lines will gyrate, bounce, and drift. Some of these particles are highly energetic and it is these particles that form the radiation belts. Figure 4.11 shows a pleasantly old-school image of the radiation belts. Note that there are two belts: an inner belt composed of energetic protons and electrons, and an outer belt composed of mainly energetic electrons. Between these two belts is the so-called slot region.

The inner belt is temporally stable, extends to about $1R_E$ above Earth's surface, peaking at an altitude of ~ 3000 km, and results from the β-decay of neutrons triggered by cosmic ray collisions with nuclei in the upper atmo-

[32] Cosmic rays are, in fact, not "rays" at all, but are highly energetic particles of cosmic origin. The most energetic cosmic rays have energies in excess of 10^{20} J. Approximately 99% of cosmic rays are atomic nuclei with the remainder being solitary electrons. About 90% of the nuclei are protons. Although supernovae are one source of cosmic rays, other sources and energization mechanisms are being actively investigated. Cosmic rays are a continual source of error in both ground- and space-based electronics.

[33] Said by Ernest Ray, a colleague of Van Allen at the Univ. of Iowa.

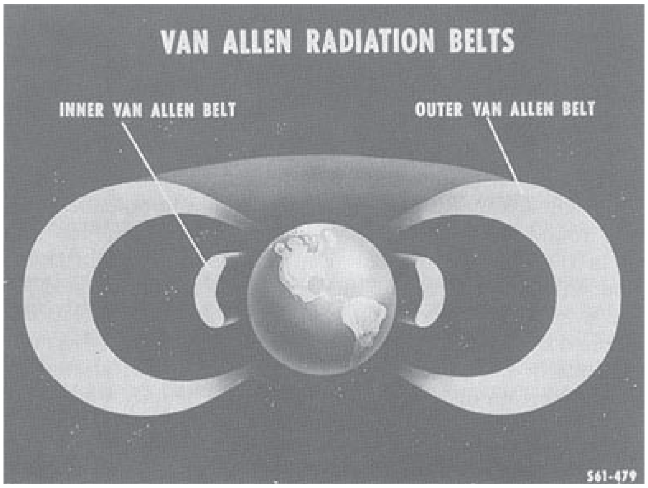

Figure 4.11: The Van Allen Radiation Belts. Image from NASA.

sphere. Proton energies in the inner belt extend to a few hundreds of MeV while the electrons have typical energies in the range 1-5 MeV. Inner belt protons are an important source of single-event upset events in space-based electronics. In particular, protons in the South Atlantic Anomaly, discussed above, provide the most intense radiation source in low Earth orbit.

In contrast to the inner belt, the outer electron belt is temporally variable,[34] extends from about 3-10R_E above Earth's surface, peaking at an altitude of $\sim 4-5R_E$, and is populated by electrons with typical energies in the range 0.1-10 MeV.

4.4.4 Currents in the Magnetosphere

Previously in this chapter we have presented the magnetosphere as being shaped by the solar wind pressure distorting an approximately dipolar geomagnetic field into a blunted (on the dayside), elongated (on the nightside) cavity. The resulting magnetic field configuration is complex, dividing naturally into the several regions discussed above, each embedded in magnetic and possibly electric fields. As we saw in Chapter 2, plasma single-particle

[34] Energetic particle fluxes in the outer belt change rapidly with timescales on the order of hours in response to geomagnetic storms (recall, for example, the discussion surrounding §2.3.1 about trapped particles and the Dst index).

motions in certain electric and magnetic field configurations result in currents and, furthermore, *any* configuration of electric and magnetic fields must satisfy Maxwell's equations. Note in particular Ampere's law,

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}$$

by which we see that whereever there is a curl to the magnetic field, there must be a current (either conduction or displacement). In space plasmas considered on large spatial and long time scales, displacements currents can almost always be ignored and it is conduction currents that support the magnetic field geometries.

Certainly there are regions in the magnetosphere where there is a curl to the magnetic field and certainly there are regions where the magnetic field geometry is such that single-particle motions result in ions and electrons drifting in different directions. Thus we realize there must be currents in the magnetosphere that work self-consistently with our previous discussions to support the structure of the magnetosphere. In this section we will investigate the currents that support the stationary, average structure of the magnetosphere. Other currents associated with magnetospheric dynamics will be discussed in Chapter 5.

Figure 4.12 illustrates the major magnetospheric current systems. All these currents must flow in closed loops and they are not independent, but rather are interconnected in such a way that the current is everywhere divergenceless and that the magnetic field produced results in the geometry illustrated in Figure 4.10.[35] The four major magnetospheric current systems are:

1. the magnetopause current,
2. the neutral sheet and tail current,
3. the ring current and
4. the field-aligned (or Birkeland) currents.

The Magnetopause Current

The direction of the geomagnetic field at the magnetopause nose is approximately normal to the ecliptic plane and has a strength of ∼70 nT. Moving

[35] A divergenceless current that flows in a closed loop is required by charge continuity: $\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0$, where ρ is the charge. For stationary systems, the time derivative is zero (showing that the current is everywhere divergenceless) and an application of the divergence theorem shows that the current must flow in a closed loop.

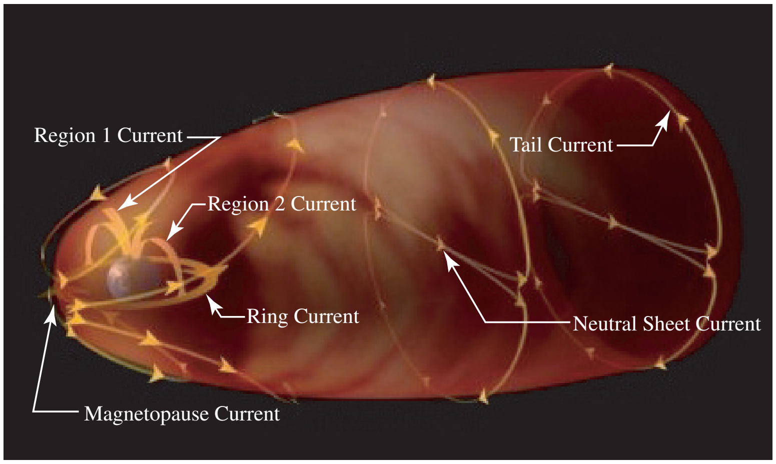

Figure 4.12: An illustration of the major magnetospheric current systems. Figure adapted from `http://www.meted.ucar.edu/hao/aurora/-txt/x_m_3_1.php` and ©The COMET Program.

some small distance towards the sun, just across the magnetopause and into the solar wind, the magnetic field is the IMF with variable orientation and typical strengths of only 6 nT. Such a field configuration has a nonzero curl and from Ampere's law we find that a current must flow on the surface of the magnetopause to support the magnetic field geometry on both sides of the boundary. Visualizing this geometry at the magnetopause nose, we find that the magnetopause current[36] must flow from west to east (dawn to dusk) so that it strengthens the magnetic field on the earthward side of the magnetopause and weakens it on the sunward side.

The magnetopause is a three dimensional structure and the magnetopause current is not limited in spatial extent to the region near the nose. As shown in Figure 4.12, this current flows over a large portion of the magnetopause, generally circulating around the cusps where the geomagnetic field has zero strength.

The physical mechanism by which this current is formed is, in the wonderful way in which nature works, self-consistent with the pressure balance requirement previously discussed for the magnetopause position. As the so-

[36] The magnetopause current is also known as the *Chapman-Ferraro current*, after the investigators who in 1931 first proposed its existence.

lar wind plasma encounters the magnetopause, the flow is generally diverted around the flanks of the magnetospheric cavity, temporarily penetrating the magnetopause before returning to the solar wind due to the Larmour radius of the gyration orbits. The rotational directions of gyrations for ions and electrons are opposite, resulting in the dawn to dusk current at the magnetopause nose and the overall current pattern illustrated in Figure 4.12. The reader is referred to Parks [1991, pp.254-256] for further details.

The Neutral Sheet and Tail Current

Inspection of Figure 4.10 reveals another region of the magnetosphere having magnetic fields with a definite curl, and thus another region where current must flow in accordance with Ampere's law. Notice the magnetic field geometry in the tail and, in particular, in the center of the plasmasheet. Here, as one goes from the southern to the northern tail lobe, the field direction changes from begin nearly away from the Earth to nearly towards the Earth. The strong curl to this field is supported by a *neutral sheet current* that, as illustrated in Figure 4.12, flows from the dawnside to the duskside through the plasmasheet and then, because it cannot simply terminate but must close on itself, flows around the magnetopause flanks as the *tail current* where it merges with the magnetopause current and closes with the neutral sheet current.

Given Ampere's law and typical values for the magnetotail field strengths (B_T), and the length of the tail (L_T), one can estimate the tail current (I_T) required to support the field configuration. Applying Ampere's law to the neutral sheet current gives[37]

$$\begin{aligned}
\oint \mathbf{B}_T \cdot d\mathbf{l} &= \mu_0 I_{NS} \\
2B_T L_T &= \mu_0 (2I_T) \\
J_T &= \frac{I_T}{L_T} = \frac{B_T}{\mu_0}.
\end{aligned}$$

where I_{NS} is the neutral sheet current and the factor of 2 on the RHS of the middle equation accounts for the converging currents from the northern

[37] The path of integration on the LHS of Ampere's law is a rectangle oriented perpendicular to the ecliptic plane, parallel to the Sun-Earth line, and situated to enclose the tail current.

and southern tail lobes. Using typical values of $B_T = 20$ nT and $L_T = 100R_E$ gives $J_T = 0.02$ A/m and $\boxed{I_T = 1 \times 10^7 \text{ A}}$. (Ten million amps is a large current!) Measurements indicate this current flows across a cross-tail potential difference of ∼60 kV that indicates a massive power of $P = IV = 1 \times 10^{12}$ W is required to support this current. This power is extracted from the solar wind as it is diverted to form and flow around the magnetopause.

The Ring Current

The *ring current* was first introduced in §2.3 where it was noted that charged particles trapped in Earth's "magnetic bottle" gyrate, bounce and drift around Earth such that an east-to-west current is produced. The current is carried by trapped ions (which drift westward) and electrons (which drift eastward) within the inner part of the plasmasheet at approximately $4-6R_E$, with ions in the energy range of 10-100 keV carrying most of the current. Contrary to the illustration in Figure 4.12 and to the name itself, the ring current does not simply consist of a ring of current around the equator but rather is carried by a sort of torus of drifting particles spanning a large range of latitudes.

The ring current produces a magnetic field that weakens the geomagnetic field at the surface and its magnitude can therefore be estimated using low latitude ground-based magnetometers.[38] During geomagnetic storms, the density and energy of ring current particles increases and the diamagnetic field they produce decreases the surface field strength by levels often exceeding 100 nT.[39]

Birkeland Currents

The ring current encircles Earth but is stronger on the nightside than on the dayside. Current continuity then requires that, as the ring current flows from the nightside to the dayside, part of it must be diverted into another current system. For the same reason, as the ring current flows from the dayside to the nightside, the fact that its strength increases means that it must be augmented by contributions from another current system. Figure

[38] But recall the warning in Chapter 2 footnote 23: the low-latitude geomagnetic field is not influenced by the ring current alone.

[39] During extreme events, the surface magnetic field strength can be reduced by as much as a few percent.

4.12 illustrates this interplay between the ring current and what is known as the *region 2* current system. Region 2 currents are field-aligned or *Birkeland currents* which are named after Kristian Birkeland who first proposed their existence. They flow along field lines towards Earth on the duskside and away from Earth on the dawnside.

Another Birkeland current system known as the *region 1* system flows along field lines at higher latitudes than those of the region 2 system and partially closes the magnetopause current system. The polarity of the region 1 currents is generally opposite that of the region 2 system, flowing away from Earth on the duskside and towards Earth on the dawnside. These region 1 and region 2 current systems are key factors in driving the patterns of ionospheric convection described in §7.9.

4.5 Summary

To be written.

Exercises

4.1: Illustrate the similar triangles leading to the ratio given in Eq. 4.4 and perform the required integration to obtain Eq. 4.5.

4.2: Use the DGRF/IGRF magnetic field model to show that, out to at least a few Earth radii, the geomagnetic equatorial field falls off as $1/r^3$.

4.3: Verify the substitutions leading to Eq. 4.11 and insert typical solar wind values to show that $L_{mp} \approx 10$.

4.4: By approximately what factor is the solar wind density increased on the downstream side of the bow shock relative to the upstream side?

4.5: Typical solar wind values are given in Table 3.4 (p.77). For a typical solar wind speed, determine the density required to move the magnetopause inside geosynchronous orbit. For a typical solar wind density, determine the solar wind speed required to move the magnetopause inside geosynchronous orbit. Investigate solar wind data and comment on how often these required values are realized.

Chapter 5

Magnetospheric Dynamics

5.1 Introduction

The previous chapter presented a *static* picture of Earth's magnetosphere. In that picture, the incoming solar wind and IMF interact with Earth's geomagnetic field to produce the magnetopause and other magnetospheric regions as well as the magnetospheric currents that flow self-consistently to support the structure. There was little mention or even hinting that this structure and the currents that support it evolve over time. But in fact the magnetosphere is quite dynamic and it could be argued that these dynamics are the more important and compelling aspects of the magnetosphere. If Chapter 4 presented a "snapshot in time" of the magnetosphere, this chapter is intended to develop and display the movie.

Those fortunate enough to live or travel at high latitudes need not wait long nor look far for strong suggestions that the magnetosphere is impressively, if not amazingly, dynamic. As the Norwegian explorer Fridtjof Nansen put it [Nansen, 1897, p.253],

> "Presently the aurora borealis shakes over the vault of heavens its veil of glittering silver - changing now to yellow, now to green, now to red. It spreads, it contracts again in restless change; next it breaks into waving many-folded bands of shining silver, over which shoot billows of glittering rays, and then the glory vanishes. Presently it shimmers in tongues of flame over the very zenith, and then it shoots a bright ray right up from the horizon, until the whole melts away in the moonlight, and it is as though

one heard the sigh of a departing spirit."

Nansen was neither the first nor the last to be awestruck by the dynamic, and sometimes violently dynamic, aurora. This wonderful phenomenon is perhaps the most dramatic evidence that exciting things are at work in the magnetosphere. Before understanding what these things are and what effects they have, it is first necessary to introduce some additional physics.

5.2 Magnetohydrodynamics

5.2.1 Introduction to Magnetic Reconnection

Much of magnetospheric dynamics is driven by the process of *magnetic reconnection*. The theory of this process is difficult, largely beyond the scope of this text and, to some extent, still an area of active research. Even so, the basic idea as illustrated in Figure 5.1 is reasonably straightfoward. When oppositely directed magnetic fields are forced together, they can "merge", or "reconnect", ejecting the local plasma in the process. In this figure, the black magnetic field lines at the top are directed opposite to the light gray field lines at the bottom. Both sets of field lines are carried into the reconnection region by an inflow of plasma that carries the field lines with the flow in consequence of the frozen-in flux condition. Near the horizontal neutral line, a strong current sheet is present to support the magnetic field configuration. At the center of the figure is an *X-point*, a point where two oppositely directed field lines reconnect to become one (darker gray) field line that preserves the original field polarities but is highly "kinked". Highly kinked magnetic field lines exhibit a sort of "magnetic tension", discussed in some detail in §**??**, that is analogous to the tension in a kinked spring. As a result of this tension, the field lines, in an attempt to become unkinked, are ejected from the reconnection region and carry ejected plasma with them.

A key observation from Figure 5.1 will lead us into some theory: note from this figure than at locations where this reconnection happens, plasma from one field line becomes intermixed with plasma from another field line. Such mixing violates the frozen-in flux assumption and we may be able to understand why, or at least where, reconnection happens by investigating where in magnetosphere the frozen-flux assumption is invalid.

Let us begin our investigation of the frozen-in flux condition and reconnection by introducing the theory of *magnetohydrodynamics* or *MHD* for short.

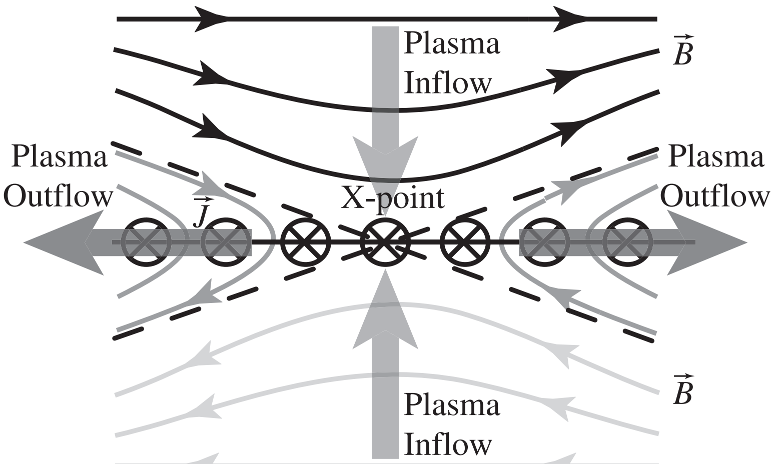

Figure 5.1: Cross-sectional view through a reconnection region showing oppositely directed and reconnected magnetic fields. Vertical plasma flows carry the magnetic field into the reconnection region where, at the X-point, they reconnect and eject the plasma. A current density **J**, directed into the page, supports the magnetic field geometry.

Rather than modeling a plasma as a collection of single *particles* as we did in §2.2, MHD models plasmas as a *fluid* or as a collection of fluids. In particular, MHD models the time evolution of magnetized fluid plasmas, from which the term *magneto-hydro-dynamics* is obtained. As with any fluid description, quantities in MHD are locally averaged. For example, the density at some point is the average density over some volume that is small compared to the system being modeled but large enough to contain a statistically significant number of particles. MHD also assumes that locally averaged quantities vary on time scales that are long compared to microscopic motions (such as plasma oscillations and gyroperiods), and on spatial scales that are large compared to the Debye length and the thermal gyroradius. That is, MHD deals with average, bulk plasma motions and describes the large-scale, slow evolution of a magnetized plasma. It cannot be used to study the "microphysics" or single-particle motions involved in a process but it is an extrordinarily valuable tool for understanding the large-scale picture.

5.2.2 The MHD Equations

A complete MHD description requires the simultaneous solution of many equations and we may determine the number of required equations by counting the number of unknowns that must be modeled. In a fluid plasma, the fluid velocity, mass density and pressure may all vary in space and time and, taking the velocity to be a three dimensional vector, this yields five unknowns. But MHD models magnetized fluids so we must also account for the electric and magnetic fields and currents flowing the in plasma. Each of these are vector quantities and together they introduce nine more unknowns for a total of 14. A full MHD description therefore requires 14 independent equations. Let us begin a presentation of these equation by introducing the fluid equations.

In Chapter 2, we developed the continuity equation (Equation 2.7 which is an expression of charge conservation. Here we may write an analogous equation expressing conservation of mass for a given fluid species (*e.g.,* electron, ions or neutrals). The continuity equation for a species s is

$$m_s \frac{\partial n_s}{\partial t} + m_s \nabla \cdot (n_s \mathbf{v}_s) = m_s (S_s - L_s) \tag{5.1}$$

where m_s and n_s are the mass and number density of a species and S_s and L_s are the source and loss rates for the species in number per second. Assuming charge neutraity and singly-ionized ions (so that $n_e = \sum_i n_i$) and a fully ionized plasma (so that there are no neutral species), equation 5.1 may be written for each species being modeled and summed over species to obtain a continuity equation for the system as a whole. Further assuming there are no sources of losses, we obtain

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \mathbf{v} = 0 \tag{5.2}$$

where ρ is the total mass density and $\mathbf{v}$ is the center of mass velocity. This continuity equation is the first of the 14 required equations.

We encountered a momentum equation during our derivation of the plasma frequency where it was noted that momentum equations are essentially expressions of Newton's 2$^{\text{nd}}$ law. Here, a given species s may be exposed to a variety of forces including those from pressure gradients, electric and magnetic fields, gravity and collisions. A typical momentum equation for species

s may then be

$$\rho_s\left[\frac{\partial \mathbf{v}_s}{\partial t}+\left(\mathbf{v}_s\cdot\nabla\right)\mathbf{v_s}\right]+m_s\mathbf{v}_s\left(S_s-L_s\right)=-\nabla p_s+\rho_{qs}\mathbf{E}+\mathbf{J}_s\times\mathbf{B}+\frac{\rho_s\mathbf{F}_g}{m_s}+\mathbf{K}_s \tag{5.3}$$

where $\rho_{qs}=q_sn_s$ is the charge density, p_s is the species partial pressure, $\mathbf{J}_s=q_sn_s\mathbf{v}_s$ is the current density associated with species s, $\mathbf{K}_s$ is the rate of momentum change in species s due to collisions and $\mathbf{F}_g$ is the force of gravity on species s. Taking the same assumptions as before, we may add the electron and ion momentum equations to obtain

$$\rho_s\left[\frac{\partial \mathbf{v}}{\partial t}+\left(\mathbf{v}\cdot\nabla\right)\mathbf{v}\right]=-\nabla p+\mathbf{J}\times\mathbf{B}+\frac{\rho\mathbf{F}_g}{m_p} \tag{5.4}$$

since there can be no sources or losses of charged particles (there being no neutrals to ionize), the electric forces are oppositely directed on each species and therefore cancel and the net momentum is conserved. Furthermore, the fact that ion masses greatly exceed electron masses makes it possible to neglect the electron gravitational term in Equation 5.4. This vector equation accounts for the next three of our required equations (bringing the current total to four).

If we assume an ideal gas law for the equation of state, the total scalar pressure is

$$p=\sum_s n_skT_s \tag{5.5}$$

which, for a fully ionized plasma, reduces to $p=\sum_i n_ikT_i+n_ekT_e$. We cannot count this equation among the required 14 since doing so would introduce more unknowns (the temperatures).

We may obtain more independent equations by examining the fields. In the MHD limit of slow changes over large spatial distances, the displacement current in Ampere's law may be neglected, leading to

$$\nabla\times\mathbf{B}=\mu_0\mathbf{J}. \tag{5.6}$$

Faraday's law is

$$\nabla\times\mathbf{E}=-\frac{\partial\mathbf{B}}{\partial t}. \tag{5.7}$$

Equations 5.6 and 5.7 bring the total number of independent equations to 10 and we may suppose that 2 more may be obtained from the two scalar

Maxwell equations. But note that $\nabla \cdot \mathbf{B} = 0$ and $\nabla \cdot \mathbf{E} = \rho_q/\epsilon_0$ (where ρ_q is the charge density that, assuming charge neutrality, is equal to zero) are not dynamical equations in the sense that they do not describe the time variations of **B** and **E**. Rather, they are more like initial conditions that are also satisfied at any time. So having now 10 equations, we need four more.

One of the required four may be obtained by imposing conservation of either energy or entropy on the plasma. Conservation of energy is the more problematic of these two choices as it introduces the heat flux vector into the system which requires either further approximation or the addition of three more independent equations to specify the heat flux. Conservation of entropy, however, simply requres that the pressure and mass density are related by

$$p\rho^{-\gamma} = \text{constant}$$

so that

$$\frac{\partial p}{\partial t} + \mathbf{v} \cdot \nabla p = c_s^2 \left(\frac{\partial p}{\partial t} + \mathbf{v} \cdot \nabla \rho \right) \tag{5.8}$$

where c_s is the plasma sound speed defined by $c_s = \gamma p/\rho$ and γ is the ratio of specific heats [Kivelson and Russell, 1995, p47].

One more vector equation (to bring our total to 14 and close the system) may be obtained from a generalized Ohm's law that is a relationship between the current density **J** and the fields. This relation may be derived by subtracting the ion and electron momentum equations and performing some algebra. The derivation would require a significant detour at this point and we instead refer the reader to Baumjohann and Treumann [1997, pp141-142] and Kivelson and Russell [1995, p.48] for details that result in

$$\mathbf{j} = \sigma \left[(\mathbf{E} + \mathbf{v} \times \mathbf{B}) + \frac{1}{n_0 e} \nabla p_e - \frac{1}{n_0 e} \mathbf{j} \times \mathbf{B} - \frac{m_e}{n_0 e^2} \left(\frac{\partial \mathbf{j}}{\partial t} + \nabla \cdot (\mathbf{j}\mathbf{v}) \right) \right]. \tag{5.9}$$

where σ is the plasma *conductivity.* The conductivity of a plasma will be discussed in detail in §7.8 where we will find that σ, in general, is a tensor that relates the magnitude and direction of current flow in response to applied forces.

Often the last terms on the RHS of Equation 5.9 are dropped, resulting in

$$\mathbf{J} = \sigma(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{5.10}$$

and, in *ideal* MHD, the conductivity is assumed to be infinite so that

$$\mathbf{E} + \mathbf{v} \times \mathbf{B} = 0. \tag{5.11}$$

So we have succeeded in accumulating 14 MHD equations that can be used to describe the slowly-varying time evolution of large-scale plasma fluids. It was a nontrivial task to assemble the equations and the reader may appreciate that the simultaneous solution of the set is, in general, quite a formidable task. Formidable though they may be, the MHD equations are routinely solved numerically and such models of the magnetosphere provide rich insights into its dynamics and response to changing conditions in space.

Chapter 6

Earth's Neutral Atmosphere

6.1 Motivation

Earth's neutral atmosphere[1] is a spherical layer of gas surrounding our planet and, compared to the scale of regions we have investigated up to this point, it is an *exceedingly thin* spherical shell. In fact, half of the atmospheric mass lies beneath 6 km in altitude and only something on the order of a ten thousandth of a percent of it lies above 100 km in altitude where space is considered to begin. Most of Earth's atmosphere is therefore beneath the near-Earth space environment and omission from this book could be argued on this point alone. But we should appreciate, for example, that atmospheric drag is a primary consideration for satellites in low-Earth orbit, that the neutral atmosphere is central to the formation of Earth's ionosphere, that the neutral atmosphere protects us from various radiations from space, that auroras result largely from the interaction of magnetospheric plasma with the neutral atmosphere, and that the dynamics of the atmosphere have definite impacts on space-based technology. Given the importance of topics on this list and the context and perspective that a basic understanding of the neutral atmosphere can provide, this chapter attempts to walk a middle ground between the typically plasma-dominated field of space physics and the chemistry- and fluid dynamics-dominated field of aeronomy. This middle ground regrettably ignores most of the compelling dynamics of the atmosphere and focuses on

[1] Translated from Greek, the word *atmosphere* literally means 'vapour ball'.

general, static properties.[2]

6.2 Introduction

Water, earth, air and fire are the four classical 'elements'[3] and we may consider air to be equivalent to the neutral atmosphere that we breathe. The first evidence that air was not an element, but rather a mixture of various gasses, was provided around the year 1630 by Jan Baptista van Helmont (1577-1644) who identified carbon dioxide (called by him *gas sylvestre*) released by burning charcoal.[4] The next consitituent of our neutral atmosphere to be discovered was molecular oxygen, first around 1772 by Carl Wilhelm Scheele (1742-1786) and then independently by Joseph Priestly (1733-1804) in 1774.[5] In addition to his discovery of oxygen in 1774, Scheele simultaneously also discovered molecular nitrogen. He labelled O_2 and N_2 as 'fire air' and 'foul air' respectively to indicate that the first supports combustion while the latter suppresses it.

N_2 and O_2 are the two most common constituents of the atmosphere at ground level and together they account for around 99% of its atoms and molecules. As we will see below, CO_2 is the fourth most common constituent (and yet it was the first to be discovered). The missing element in the top four constituents is argon: an inert, colorless, odorless, and tastless noble gas. No doubt these properties added to the delay in its being identified as a relatively common constituent. Following the discoveries of CO_2, O_2

[2] So we will largely ignore meteorology, weather, climate change, global circulation, gravity waves, tides, *etc.* To briefly indicate the complexity and importance of these topics, note that among them are some of the most mature, developed, challenging, and important problems related to the study of our planet. Humans have tried for millenia to predict the weather and even today's state-of-the-art models run on the fastest supercomputers cannot determine with certainty whether or not it will rain two days in the future at a given location. Additionally, questions concerning global climate change must be considered among the most important facing mankind.

[3] Often the fifth element or quintessence is also listed. This quintessence is the aether that lies beyond the other four.

[4] van Helmont claims to have have coined the term 'gas', by which he meant vapors different from the ordinary air we breathe. *Gas Sylvestre* may be loosely translated from latin as "gas from wood".

[5] Credit for the discovery of O_2 is usually given to Priestly because he was the first to publish his findings.

and N_2, more than a century passed before Lord Rayleigh[6] (1842-1919) and Sir William Ramsay (1852-1916) isolated argon from air in 1894. Rayleigh and Ramsay were awarded the 1904 Nobel prizes in physics and chemistry, respectively, for their discovery. Given these discoveries and others that followed, we now know that the neutral atmosphere is a gaseous mixture of atoms and molecules that may contain in suspension, at certain locations and times, small amounts of solids and liquids including dust and water droplets.

Table 6.1 summarizes the atmospheric composition at ground level. The abundances of nitrogen, oxygen and the noble gasses do not vary significantly with location or time. The other consitituents, significantly including water vapor which is not included in the table, may vary by several percent in relative abundance due to local conditions.[7]

Ground-level Atmospheric Constituent	Chemical Formula	Relative Abundance (%)	Number Density (m^{-3})
Nitrogen (molecular)	N_2	78.1	$\sim 2 \times 10^{25}$
Oxygen (molecular)	O_2	20.9	$\sim 5.6 \times 10^{24}$
Argon	Ar	0.93	$\sim 2.5 \times 10^{17}$
Carbon Dioxide	CO_2	0.035	
Neon	Ne	0.0018	
Helium	He	0.00052	
Methane	CH_4	0.00015	
Krypton	Kr	0.00011	
Hydrogen (molecular)	H_2	0.00005	
Xenon	Xe	0.00001	

Table 6.1: Composition of Earth's neutral atmosphere at ground level.

Thus the air we breathe is mainly nitrogen (~78%), oxygen (~21%), and argon (~1%) with a density and temperature such that the average sea-level pressure is 101.325 kPa, defined to be one standard atmosphere (or atm)[8].

[6] Wikipedia relates that he was: John William Strutt, 3rd Baron Rayleigh, OM (Order of Merit), PRS (President of the Royal Society).

[7] For example, water vapor near the surface varies from a relative abundance near zero over deserts to about 4% over oceans and methane (CH_4) is relatively more abundant over cow pastures than deserts.

In many ways, Earth's atmosphere is surprising. Our nearest planetary neighbors have atmospheres dominated by the oxidized CO_2 molecule while Earth's atmosphere is oxidizing, alone in the solar system in its relatively high (and far from thermodynamic equilibrium) abundances of life-sustaining O_2.[9]

Local deviations from the average sea-level pressure of 1 atm are of course strong drivers of atmospheric dynamics but here we will assume the atmsophere is in static equilibrium. Given this assumption, we may investigate how atmospheric pressure varies with altitude.

6.3 The Hydrostatic Approximation

The pervasive force of gravity exerts a constant downward force on the atmosphere's constituents and we may therefore expect the number density (and pressure) of this gas to be highest at the lowest altitudes and to decrease with increasing altitude. To obtain an expression for this variation, we recognize that static equilibrium requires a balance of vertical forces acting on a parcel of air. Figure 6.1 shows these forces where p_0 is the pressure at the bottom of the parcel, A is the parcel's cross-sectional area, δz is the parcel's height, δm is the airmass in the parcel, and δp is the assumed change in pressure over the parcel height. If ρ and δV are the parcel's mass density and volume, respectively, the parcel's weight is given by $w = (\delta m)\, g = (\rho\, \delta V)\, g = (\rho A\, \delta z)\, g = \rho A g\, \delta z$.

In static equilibrium, $\sum F_z = 0$ so from Figure 6.1,

$$\sum F_z = -\left(p_0 + \delta p\right) A - \rho A g\, \delta z + p_0 A = 0.$$

Doing a line or so of algebra and taking the limit $\delta z \to 0$, we find that

$$\frac{\partial p}{\partial z} = -\rho g \tag{6.1}$$

which is known as the *hydrostatic equation.* The pressure p and mass density ρ in this equation are not independent and are related to each other through

[8] Note that this pressure is the magnitude of weight per unit area of a column of air extending from sea level to the "top" of the atmosphere and that 1 atm = 101.325 kPa = 14.696 psi = 760.0 mmHg = 29.92 inHg.

[9] The primary source of this excess O_2 is of course biological activity (primarily photosynthesis) fueled by radiation from the Sun. In the absence of biological activity, the surface concentration of O_2 would be approximately 10^{13} times smaller than observed[Wayne, 1991, p.5].

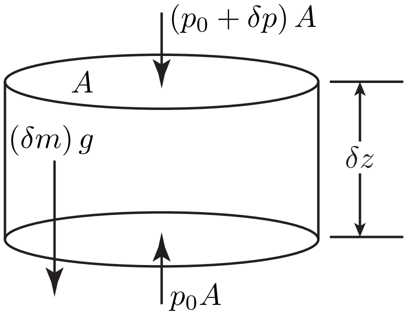

Figure 6.1: The vertical forces on an air parcel in static equilibrium.

an equation of state. Let us take the ideal gas law, $p = nkT$, where n is the number density of the gas, k is Boltzmann's constant and T is the gas temperature. The ideal gas law may be written as

$$p = \rho kT/m_p \tag{6.2}$$

where m_p is the average constituent mass[10] and $n = \rho/m_p$. Solving Equation 6.2 for the mass density ρ, substituting into Equation 6.1, and separating variables yields

$$\frac{\partial p}{p} = -\frac{m_p g}{kT}\partial z \tag{6.3}$$

that may be integrated to find

$$p = p_0 \exp\left(-\int_{z_0}^{z} \frac{m_p g}{kT} dz'\right) = p_0 \exp\left(-\int_{z_0}^{z} dz'/H\right)$$

where

$$H \equiv \frac{kT}{m_p g} \tag{6.4}$$

is the *scale height*.[11] It can be shown that the scale height at ground level is approximately 8 km (see Exercise 6.2).

[10] For air near ground level, the average constituent mass is $m_p \approx 0.78(m_{N_2}) + 0.21(m_{O_2}) + 0.01(m_{Ar}) \approx 0.78(28) + 0.21(32) + 0.01(40)$ amu ≈ 29 amu $\approx 4.8 \times 10^{-26}$ kg.

[11] Note that this result may be obtained directly (neglecting variations with altitude) from the Boltzmann distribution we encountered in Chapter 2. In this case, the force of gravity is directed straight down and its potential energy (mgz) 'competes' with the random thermal energy given by kT so that we must have $p = p_0 e^{-\frac{mgz}{kT}}$.

The scale height varies with neutral temperature, the average constituent mass, and the gravitational acceleration g. To develop intuition about the physical meaning of the scale height, consider an atmosphere for which H is constant over altitude. If we further assume for this purpose that neither g nor m_p vary with altitude, the temperature must also be constant (*i.e.*, the atmosphere is isothermal). Equation 6.3 can then be integrated and combined with the ideal gas law to find

$$p = p_0 e^{-(z-z_0)/H} \quad \text{and} \quad n = n_0 e^{-(z-z_0)/H}$$

which show that atmospheric pressure and number density decrease exponentially. An increase in altitude of ζ scale heights results in a factor of e^{ζ} decrease in the atmospheric pressure and number density. Note carefully that an increase in temperature results in a larger scale height and less severe falloff with increasing altitude (and *vice versa*).[12]

Taking the ground-level scale height to be 8 km as mentioned above, it can be seen that:

half the atmosphere lies below an altitude of 6 km (~20,000 ft)[13]; 90% lies below 18 km (~11 miles) and 99.9997% lies below 100 km (~62 miles).

These numbers and the extreme thinness of our atmosphere should not pass by unappreciated. Fewer than 4 miles above our heads, less than half of the atmosphere remains.

To give another physical interpretation of the scale height H, note that it is the thickness of a constant-value layer. That is, if the exponentially decaying atmosphere were compressed from above until the pressure and number density were constant over altitude and equal to the reference (*e.g.*, ground level) value, the thickness of this compressed layer would be exactly one scale height. This interpretation follows directly from the integrals

$$\int_{z_0}^{\infty} p dz = p_0 H \quad \text{and} \quad \int_{z_0}^{\infty} n dz = n_0 H.$$

[12] This point will be recalled in §**??** in the context of atmospheric drag. As the atmosphere warms, density and satellite drag at a given altitude must increase.

[13] This is approximately the altitude of Denali, *aka* Mt. McKinley. Mt. Everest has an altitude of over 8.8 km and the atmospheric pressure at its summit is approximately 1/3 of that at sea level; 2/3 of Earth's atmosphere lies below the summit of Mt. Everest!

In general the scale height H is not constant with altitude. As we will see below, the atmospheric temperature varies significantly with altitude and, at least for the upper atmosphere, so too does the average constituent mass. Of course, the acceleration of gravity decreases with increasing altitude as $1/(R_E + z)^2$.

6.4 Atmospheric Temperature

Having found the atmospheric pressure and number density to fall off exponentially with scale height H as altitude increases, let us turn our attention to the temperature of Earth's atmosphere. It is no mean feat to quantitativey evaluate from first principles the atmospheric temperature at ground level, to say nothing about the additional complications required to determine its variations with altitude. Here we will employ general approximations and an empirical model to illuminate the problem and hopefully provide useful understanding.

6.4.1 The Surface Temperature

To a good level of approximation, the atmosphere is heated by two sources of energy: the Sun and the Earth. Recall from §3.1.2 that the Sun provides a power density of $F_\odot \approx 1366$ W/m^2 from above and the Earth, being continually warmed by the Sun,[14] radiates as a blackbody in the far infared and provides energy from below. We may crudely estimate Earth's surface temperature by equating the solar energy it absorbs to the blackbody radiation energy it emits.[15] The projected area of Earth intercepting the Sun's energy is πR_E^2 and we assume, because the Earth rotates, that this energy is on average distribued across the whole surface of Earth. Some fraction of this incident radiation is relfected from the surface and the reflection coefficient is known as the *albedo*. Although the albedo varies with wavelength and surface conditions, a globally-averaged value of 0.30 is generally accepted.[16] That is, 70% of the incident solar energy is absorbed by Earth's surface and

[14] Here we ignore other sources of energy within the Earth including those from radioactivity and other geothermal sources.

[15] That is, we assume Earth is in radiative equilibrium with its environment.

[16] For visible light, the albedo of: forrested areas is ∼0.15, desert sand is ∼0.40, and fresh snow is ∼0.90.

30% is reflected. We may then use the Stefan-Boltzmann law to equate

$$\pi R_E^2 \left(1 - a_E\right) F_\odot = 4\pi R_E^2 \sigma T_E^4 \tag{6.5}$$

where a_E is the globally-averaged albedo, σ is the Stefan-Boltzmann constant[17] and T_E is Earth's surface temperature.

Evaluating Equation 6.5 yields $T_E = 255$ K for the surface temperature of Earth (see Exercise 6.5). Note that this is the temperature of Earth's surface, not the temperature of the atmosphere at ground-level. The actual average temperature on Earth's surface is some 30 K warmer than this and the discrepancy is largely due to complicating effects attributable to Earth's atmosphere. To crudely suggest the basis of these effects, it must be appreciated that Earth's atmosphere acts as a "blanket" wrapping Earth and absorbing the vast majority of its emitted infared (blackbody) radiation. Some of this absorbed infared radiation is lost to space while some is directed back towards Earth, in effect (but not in fact) reducing its albedo and raising the surface temperature. Allowing for conduction and convection, we may then take the atmospheric temperature at ground level to be equal to the observed globally-averaged surface temperature of ~287 K.

6.4.2 The Temperature Profile

Figure 6.2a shows perhaps the simplest model of what might be the expected temperature profile for Earth's atmosphere as heated by the two sources discussed above. At the highest altitudes, the temperature is determined almost entirely by the energy input from the Sun. As some fraction of this energy is absorbed, less is available for heating at lower altitudes so the energy input decreases with decreasing altitude. Earth's reradiated infrared blackbody radiation is the dominant heat source at low altitudes and this contribution decreases rapidly with increasing altitude.[18] The resulting temperature profile should be approximately proportional to the sum of these two energy inputs at each altitude and we therefore expect relatively high values

[17] σ may helpfully be remembered as the "5-6-7-8 constant" because $\sigma = 5.67 \times 10^{-8}$ W/m^2/K^4 to three significant figures.

[18] Water vapor is the dominant absorber of Earth's infared near the ground. It's density decreases with a scale height of ~2 km which is much shorter than those for the dominant consituents due to condensation and precipitation as the air temperature decreases with increasing height. In addition to water vapor, CO_2 and O_3 near the surface also contribute to infrared absorption.

at the ground and at the highest altitudes with a minimum value falling at intermediate altitudes.

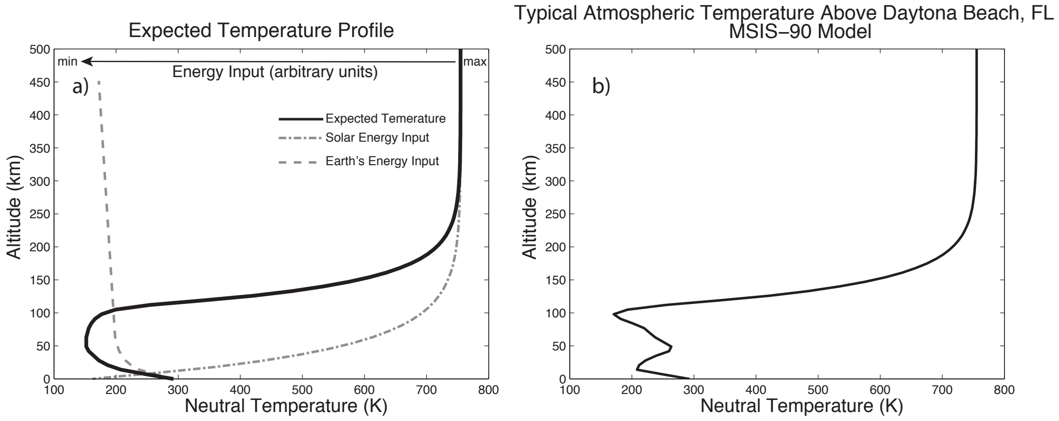

Figure 6.2: a) The expected temperature profile for Earth's atmosphere based on two energy sources: the Sun (heating from above) and the Earth (heating from below); b) a typical temperature profile from the MSIS-E-90 model.

Figure 6.2b shows a typical atmospheric temperature profile over Daytona Beach, FL, obtained from the MSIS-E-90 atmospheric model.[19] This profile looks generally similar to the expected profile in that temperatures are highest near the ground and at the highest altitudes with lower values at intermediate altitudes. The local maximum near 50 km is however unexpected and indicates an additional local heating source at that level. As it turns out, this local source is due to absorption of solar EUV by the *ozone layer*. Ozone is an extremely efficient absorber of EUV photons, particularily in the so-called Hartley bands near 260 nm. These wavelengths are harzardous to life due to their mutating effects on DNA and their absorption above ground level is therefore of critical importance. Not considering so-called smog ozone near ground level, ozone is naturally produced in the atmosphere by photodissociation of O_2 and recombination to form O_3. An equilibrium balance between this source and loss mechanisms catalyzed by natural OH and NO and man-made chlorofluorocarbons results in an ozone layer with peak densities near 25 km but with sufficient densities at higher

[19] Data from the MSIS-E-90 empirical model are available from: http://omniweb.gsfc.-nasa.gov/vitmo/msis_vitmo.html

altitudes to result in the unexpected temperature enhancements shown in Figure 6.2b.

The temperatures at the highest altitudes shown in Figure 6.2 vary significantly in reponse to the flux of UV and higher energy photons incident on the atmosphere. At times of very high solar activity, the limiting temperature may reach 2000 K and may be as low as ∼450 K during very low solar activity. We may then expect a strong correlation between the limiting temperature and the solar cycle (see Exercise 6.6).

Figure 6.3 shows the same temperature profile as in Figure 6.2b but with a logarithmic altitude scale that accentuates the rapid fluctuates in temperature near the ground. Note that the sign of dT/dz changes several times as altitude increases. For the first 12 km or so above the ground, the temperature decreases with increasing altitude and this region is known as the *troposphere*.[20] The slope then changes sign and temperature increases with increasing altitude throughout the *stratosphere* until the level near 50 km. In the *mesosphere* the slope reverses sign and the temperature reaches its minimum value near 100 km. Above this level, the temperature in the *thermosphere* increases with altitude until reaching its limiting value near 750 K.

The naming scheme described above organizes the atmosphere into regions based on whether the temperature increases or decreases with increasing altitude. The boundaries between adjacent layers, indicated by the horizontal lines in Figure 6.3, are denoted the *tropopause*, *stratopause*, and *mesopause*.[21]

6.5 Atmospheric Composition

We saw in Table 6.1 that Earth's atmosphere at ground level consists of molecular nitrogen (∼78%) and oxygen (∼21%) with relatively small additions of argon and a number of other trace or highly variable constituents. It is natural to wonder how this composition varies with altitude. Certainly,

[20] The vertical extent of the troposphere varies with latitude and season. Its mean vertical extent is about 18 km at the equator and 8 km at the poles.

[21] In the general field of "space physics", a boundary between two adjacent regions is typically called a *pause*, with the addition of a prefix to identify the underlying layer. Thus, the magnetopause is the boundary between the magnetosphere and interplanetary space, the tropopause is the boundary between the troposphere and the stratosphere, etc.

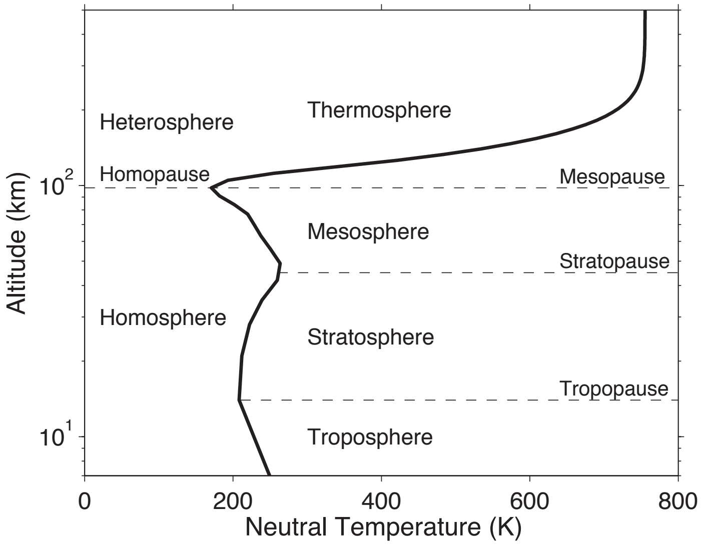

Figure 6.3: The same MSIS-E-90 temperature profile shown in Figure 6.2b but with a logarithmic altitude scale. Atmospheric regions are identified as classified by: the sign of dT/dz (near the center), and by composition (near the left side of the plot).

as we have previously seen, the total number density decreases exponentially with altitude at a rate set by the scale height, but here we wonder whether or not, at higher altitudes, the relative concentrations remain the same as at ground level. The answer is perhaps unexpected: it is at first "yes" and then, above a certain altitude, becomes "no".

Figure 6.4 shows MSIS-E-90 composition data from the same date, time and location as in Figure 6.3. There are several points of interest to note from this figure. The horizontal axis is logarithmic so the exponential falloff in number density with altitude is clearly visible below ~100 km. Above ~80 km, atomic species begin to appear that did not exist at lower altitudes and that are certainly not normally present at ground level. One may well wonder why they suddenly appear in the middle atmosphere and why their relative abundances do not correlate with those of associated (and the presumed source) molecules. That is, given the dominance of N_2 at almost all altitudes shown here, why are the concentrations of both O and H orders of magnitude higher than that of N?

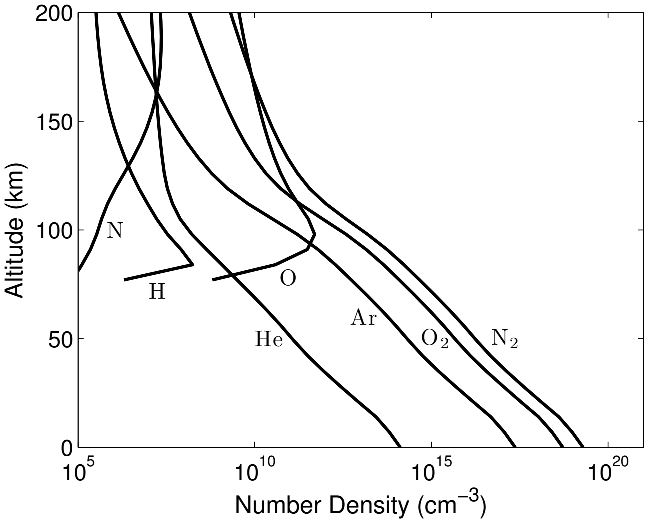

Figure 6.4: MSIS-E-90 number densities of the most dominant atmospheric constituents up to 200 km altitude.

Either a short or any more complete discussion of these points leads farther into topics of atmospheric chemistry than required for our purposes here. Let us simply note that these atomic species are produced through photodissociation by solar EUV of O_2, N_2, and H_2O to yield O, N, and H respectively. The bond energies involved are approximately 5.13 eV for O_2, 9.79 eV for N_2 and 4.7 eV for H_2O, requiring photodissocating EUV photons of wavelengths less than 242 nm, 127 nm, and 260 nm respectively. We may now appreciate the relative concentrations of these three atoms as shown in Figure 6.4: the concentrations of O_2 and N_2 in the mesosphere and lower thermosphere differ by less than an order of magnitude but require solar EUV photons of vastly different energies to yield the associated atomic species. In fact, in a typical solar spectrum, photon fluxes energetic enough to dissociate O_2 may be 10^4 times higher than those with energies sufficient to dissociate N_2. This difference in energies and EUV fluxes is the main reason why densities of O are far higher than those of N and why N is never a dominant constituent at any altitude even though N_2 is the most dominant constituent at lower altitudes.

Hydrogen densities fall in between the two extremes discussed above because the required photon fluxes are relatively high in the upper atmsophere but the densities of H_2O, from which H is obtained, are very low. As it turns out, O_2, O, and N_2 are efficient absorbers of photons in the 1-100 nm range and essentially all solar photons within this range are absorbed above 80 km and are not significantly present near the ground where concentrations of H_2O are highest.

Figure 6.5 again shows the previous MSIS-E-90 composition data but in a format intended to more fully illustrate typical variations with altitude. The horizontal axis shows relative number density (N_i/N_{total}) of a given species where N_i is the number density of the i^{th} species and $N_{total} = \sum_i N_i$. Note that below to about 100 km, the relative concentrations of N_2, O_2, and Ar are apparently constant and equal to their ground level values, indicating that all species are decaying at the same exponential rate set by a common scale height as defined in Equation 6.4. This region below about 100 km is therefore known as the *homosphere* because the relative concentrations here do not vary with altitude. This density variation of different species with a common scale height is perhaps a bit unexpected: if a gas composed of several different species is left undisturbed, heavier species will "sink" beneath the lighter ones and the density of each species will vary with a unique scale height that depends on its mass. The absence of this gravitational separation indicates that the homosphere is well mixed. Something is stirring the pot, so to speak, fast enough that gravitational separation is not effective. This issue of mixing in the lower atmosphere will be addressed in §6.6.

Figure 6.5 shows very different features above ~100 than in the underlying homosphere. At these higher altitudes it appears that individual species may (and in fact do) gravitationally separate and vary independently of each other with a species-dependent scale height given by

$$H_i = \frac{kT}{m_i g} \tag{6.6}$$

where the i subscript represents the i^{th} species and it is assumed that all species have the same temperature. As m_i increases, the scale height decreases, meaning that the exponential falloff occurs over a shorter altitude scale relative to less massive species. Thus, as expected, lighter elements will "float" on top of the more massive ones. This region of the atmosphere where gravitational separation leads to varying relative concentrations based on the mass of individual species is known as the *heterosphere*. The boundary

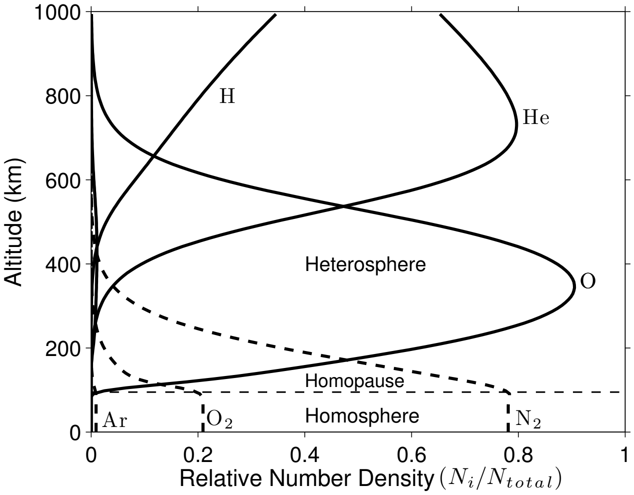

Figure 6.5: MSIS-E-90 relative number densities of the most dominant atmospheric constituents up to 1000 km altitude. Atmopsheric regions classified by composition are listed near the plot center.

between the well-mixed homosphere and the gravitationally-separating heterosphere is, as you may well expect, known as the *homopause*. This graviational separation leads to atomic oxygen being the dominant constituent from ∼180-550 km, reaching a maximum relative number density of ∼90%. At higher altitutdes, the lighter elements begin to dominate and He has the highest relative concentration from ∼550-2500 km. At altitudes above 2500 km, H, the lightest element, dominates and is nearly fully ionized by solar EUV. This uppermost region is sometimes known as the *protonosphere*.

6.6 Atmospheric Stability

6.6.1 Overview

We saw in the previous section that the neutral atmosphere is "well-mixed" below the homopause at ∼100 km and not mixed (and therefore graviationally separating) above this level. The difference between these two regimes has to do with atmospheric stability. In essence, pacels of air are constantly being displaced vertically.[22] In some regions, parcels of air that are displaced upwards are more dense than the surrounding ambient atmosphere and consquently "sink" back toward their original altitude. These parcels do not significantly contribute to vertical mixing. The atmopshere in these regions is stable to vertical perturbations. In other regions, the upwardly displaced parcels are less dense than the ambient atmosphere and continue to rise (or "float" higher) in an unstable atmosphere and provide significant mixing between altitudes.[23] We now turn out attention to what makes the atmosphere stable or unstable at various altitudes.

Figure 6.6 is reproduced for convenience from Figure 6.3 and provides a clue to the cause of atmospheric stability. Note in this figure that the homopause occurs at approximately the same altitude as the mesopause and that above this altitude, temperature monotonically increases with altitude. Below the homopause the temperature either increases or decreases (depending on the layer) with altitude. This obvious contrast suggests the explanation for why the homosphere is well-mixed and the heterosphere is not may be related to the variation in temperature with altitude.

To investigate the stability a little more quantatively, let us consider the adiabatic expansion of an air parcel as it rises in altitude as a result of some vertical perturbation. Two regimes need to be considered in turn: first, a regime where the ambient temperature increases with increasing altitude (as in the thermosphere) and then where it decreases with increasing altitude (as in the tropsphere). In both regimes, the upwardly displaced parcel maintains pressure balance with its surroundings[24] so that

$$\begin{aligned} p_d &= p_a \\ n_d k T_d &= n_a k T_a \\ n_d &= \frac{T_a}{T_d} n_a \end{aligned}$$

where the d and a subscripts refer to values of the displaced and ambient values respectively.

In the first regime where temperature increases with altitude, T_d must be less than T_a because the parcel starts at a lower altitude (where temperatures

[22] By, for example, winds over mountain ranges or convection in thunderstorms.

[23] Analogous comments could be made for air parcels that are displaced downwards.

[24] That is, we assume the parcel is displaced at a speed lower than the sound speed but, to maintain the adiabatic assumption, fast enough that no thermal energy is transferred to or from the parcel.

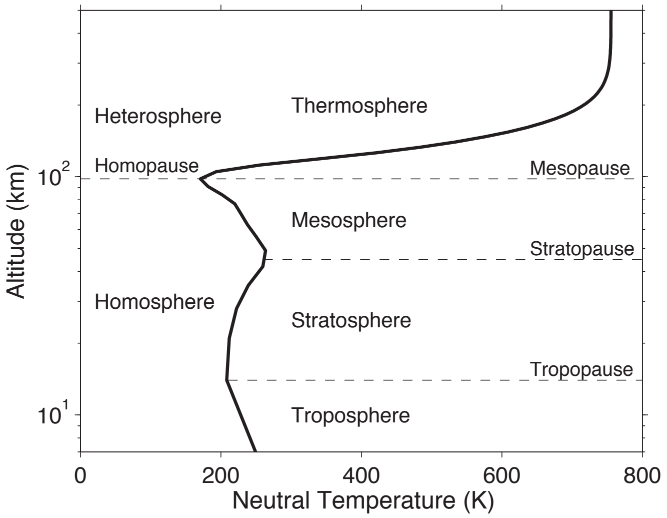

Figure 6.6: The same MSIS-E-90 temperature profile shown in Figure 6.2b but with a logarithmic altitude scale. Atmospheric regions are identified as classified by: the sign of dT/dz (near the center), and by composition (near the left side of the plot). (Reproduced from Figure 6.3.)

are lower) and the parcel's temperature decreases further as it adiabatically expands. The parcel's density is then

$$n_d = \frac{T_a}{T_d} n_a > n_a \tag{6.7}$$

so that it "sinks" back toward its equilibrium position and is stable to vertical perturbations. A similar agument for a parcel displaced vertically downward shows that $n_d < n_a$ and the parcel "floats" back toward its equilibrium position and is stable to vertical perturbations. In general,

> regions of the atmosphere where temperature increases with altitude (*e.g.*, the thermosphere) are stable to vertical perturbations. In consequence, these regions are not generally well-mixed by atmosperic motions and tend to gravitationally separate according to species-dependent masses and temperatures.

In the second regime where temperature decreases with altitude, the situation is more complicated. As the displaced parcel rises, it will maintain pressure balance, adiabatically expand, and cool as before. Opposite from the previous case however, the ambient temperature also decreases and it is not now clear how T_d will compare with T_a. From Equation 6.7, if $T_d < T_a$, the parcel will be stable to vertical perturbations as discussed above. If, on the other hand, $T_d > T_a$, the vertically displaced parcel will be less dense than the ambient atmosphere, will "float" even higher, and will not return to its equilibrium position. In this case, the parcel is unstable to vertical perturbations and, in consequence, the region will be well-mixed by atmospheric motions and will not gravitationally separate according to species-dependent masses and temperatures.

> In regions of the atmosphere where temperature decreases with altitude (*e.g.*, the troposphere), atmospheric stability is determined by the rate of ambient temperature decreases with altitude relative to the rate of adiabatic temperature decrease with altitude.

The rate of adiabatic temperature decrease with altitude is known as the *adiabatic lapse rate*. This important quantity is discussed below.

6.6.2 Adiabatic Lapse Rate

The[25] first law of thermodynamics is a statement of energy conservation. It can be stated as: the change in a system's internal energy[26] equals the heat[27] added minus the work accomplished by the system.[28] Rearranging terms as

[25] First, a note on notation for this section. The thermodynamic variable p is always shown in lowercase. Quantities given as *per unit mass* are shown as lowercase. For example, a system's heat capacity at constant volume is C_v whereas the specific heat at constant volume is $c_v \equiv C_v/M$ where M is the system mass. The subscript m is used to indicate a *molar* quantity (that is, a quantity per unit mole of the substance). The molar specific heat per at constant volume is then $c_{v,m} \equiv C_v/M/N_m$ where N_m is the number of moles in the system.

[26] That is, the total energy of a system with a stationary center of mass.

[27] "Heat" is energy *transferred* between thermodynamic systems by kinetic interaction.

[28] In mechanics, the "work" term in energy conservations laws is usually written as the work done *on* the system. For whatever reason, it is more common in thermodynamics to deal with the work done *by* the system. One is of course the negative of the other and this accounts for the "minus the work accomplished by the system" term in the first law of thermodynamics. Consistency with mechanics would suggest that "plus the work done on the system" is a more reasonable expression. However, the accompanying change of sign would be a pesky thing until one got used to it.

convenient, the first law of thermodynamics is

$$dQ = dU + p\,dV$$

where dQ is the heat added, dU is the change in internal energy and $p\,dV$ is the work accomplished by the system. For an ideal gas, the internal energy is an explicit function of temperature alone so that $dU = \left(\frac{\partial U}{\partial T}\right)_V dT \equiv C_v dT$ where C_v is the heat capacity at constant volume and

$$dQ = C_v dT + p\,dV.$$

Dividing this expression by the number of moles in the system, N_m, and the system mass M, we obtain

$$dq_m = c_{v,m} dT + \frac{p}{M} dV_m \tag{6.8}$$

where $c_{v,m}$ is the molar specific heat at constant volume and V_m is the molar volume given by $V_m = V/N_m$.

Equation 6.8 can yield the adiabatic lapse rate by assuming an adiabatic process ($dq_m = 0$), an equation of state (the ideal gas law) and an expression for the variation in pressure with altitude (the hydrostatic equation). The approach taken below is essentially to relate variations in molar volume to variations in pressure using the ideal gas law and finally to variations in altitude using the hydrostatic equation. The first law of thermodynamics can then be used to evaluate the adiabatic variation in temperature with altitude.

For an ideal gas,

$$pV = (nV)kT = N_m RT \tag{6.9}$$

where $R = 8.31441$ J/mol/K is the universal molar gas constant. Dividing this expression by the number of moles N_m in the system and expanding the differentials yields

$$p\,dV_m + V_m dp = R\,dT.$$

Setting $dq_m = 0$ (the adiabatic condition) in Equation 6.8 and substituting the result for $p\,dV_m$ into the previous expression gives

$$Mc_{v,m} dT + R\,dT - V_m dp = 0$$

or

$$C_{v,m} dT + R\,dT - V_m dp = 0.$$

Now, Mayer's relation states that $R = C_{p,m} - C_{v,m}$ for an ideal gas[29] so that

$$\begin{aligned} C_{v,m}dT + (C_{p,m} - C_{v,m})\, dT - V_m dp &= 0 \\ C_{p,m}dT - V_m dp &= 0. \end{aligned}$$

Substituting the hydrostatic relation from Equation 6.1 for dp and rearranging, we find that

$$\Gamma_d = -\frac{dT}{dz} = \frac{g}{c_p} \tag{6.10}$$

where Γ_d is the dry air adiabatic lapse rate, the rate at which temperature falls with increasing altitude for a adiabatically rising parcel of dry air in an exponential atmosphere. This last result is obtained using the relation $c_p = N_m C_{p,m}/M$.

The molar heat capacity of dry air at constant pressure is approximately given by $c_p = 1003.5$ J/kg/K so that $\Gamma_d \approx 9.8$ K/km. That is, a parcel of dry air lifted vertically in Earth's atmosphere will adiabatically cool at the rate of ~9.8 K/km.

If the ambient air temperature is falling at a rate lower than $\Gamma_d \approx 9.8$ K/km (or increasing) with increasing altitude, the displaced parcel will be cooler and more dense than the ambient air and the parcel is stable to perturbations as discussed above.[30] The parcel will be unstable to vertical perturbations in the opposite condition, in which case it will continue to rise.

[29] Note that $c_{p,m} > c_{v,p}$. Energy added at constant volume is entirely used to increase the internal energy (because no work is done on or by the gas when the volume is constant). However, when heat is added at constant pressure, some of the added energy is used to increase the volume (doing work) and only the remainder is available to increase the internal energy (and therefore the temperature). Thus it takes more energy transferred as heat to raise the temperature at constant pressure than at constant volume. Mayer's relation formalizes this and evaluates the difference for an ideal gas. It is most simply derived under isobaric assumptions for which the first law of thermodynamics becomes

$$\begin{aligned} dU &= dQ - pdV \\ C_v dT &= C_p dT - pdV \\ C_v dT &= C_p dT - N_m R dT. \end{aligned}$$

Mayer's relation follows by solving for the universal gas constant R.

[30] Recall the pressure balance argument: $p_d = p_a$ (by assumption) so that $T_d < T_a$ results in $n_d > n_a$ and the displaced parcel is stable. The case of reversed inequalities leads to the displaced parcel being unstable to perturbations.

Tropospheric air that is saturated with water adiabatically cools at a rate significantly lower than Γ_d due to the release of latent heat of condensation as the air rises.[31] This, in effect, makes warm, moist air more unstable to verical perturbations and is an important consideration in, as just one example, the formation of thunderstorms.

The tropsphere and mesosphere, where temperatures decrease with altitude, are at least marginally unstable. Vertical perturbations from a wide variety of sources[32] contribute to keeping these layers well-mixed so the relative concentrations of species remain constant with altitude. The thermosphere, where temperature increases with altitude, is absolutely stable and not well-mixed, leading to gravitational separation and a resulting diffusive equilibrium. The stratosphere is sandwiched between the troposphere and the mesosphere and, as its positive temperature gradient would indicate, is stable to vertical perturbations and horizontally stratified.[33] It is also well-mixed as Figure 6.5 clearly indicates and, in light of its stability, this is perhaps a bit unexpected. The stratosphere is mixed from the combined actions of several effects including, for example, global circulation of the atmosphere and turbulent eddies. Gravitational separation is not effective in the stratosphere due to its high number density (and the resulting high rate of collisions which impede any tendency toward gravitational separation) and the small mass differences among the dominant constituents (which are mainly molecular nitrogen and oxygen).

6.7 Summary

To be written.

[31] Γ for saturated air at ground level is approximately half the value of Γ_d although an exact treatment is beyond the scope of this development.

[32] These sources include, but are not limied to: convection and turbulence in the tropsphere due to conductive heating of ground-level air, and to the effects of breaking gravity waves in the mesosphere.

[33] Thus its name.

Exercises

6.1: In footnote 8, it was asserted that atmospheric pressure at sea level is the weight per unit area of a column of air extending from sea level to the top of the atmosphere. Use the hydrostatic equation to show that the following generalization of that statement is true: The atmospheric pressure at any altitude is the weight per unit area of the overlying column of air.

6.2: Show that the scale height at ground level is $\sim$8 km.

6.3: Show that the total mass of Earth's atmosphere is $\sim 5 \times 10^{18}$ kg. (Hint: Begin with atmospheric pressure at ground level and calculate the weight of the overlying air.)

6.4: Plot the magnitude of acceleration due to gravity, g, from ground level up to an altitude of 1000 km. Comment on what seems important about its variation with altitude.

6.5: Evalute Equation 6.5 to verify the stated value of $T_E = 255$ K.

6.6: Use the MSIS-E-90 Atmosphere Model to illustrate the variation in the atmosphere's limiting temperature with F10.7. (Note that F10.7 is a common index measuring the noise level generated by the Sun at a wavelength of 10.7 cm at Earth's orbit. It is strongly correlated with the sunspot number and the solar cycle.)

6.7: Verify the wavelengths given in §6.5 required to photodissociate O_2, N_2, and H_2O.

6.8: Demonstrate the equivalence $pV = (nV)kT = N_mRT$ used in Equation 6.9.

6.9: Complete the steps required to obtain Equation 6.10.

Chapter 7

Earth's Ionosphere

In Chapter 6 on the neutral atmosphere, we encountered various layers classified according to physically significant properties. For example, variations in the temperature profile gave rise to the labels troposphere, stratosphere, mesosphere and thermosphere. Variations in composition, on the other hand, gave rise to the labels homosphere and heterosphere. Layers named according to differing physical properties may overlap and we found, for example, that the troposphere, stratosphere and mesosphere occupy the same range of altitudes as does the homosphere. In this chapter we introduce and consider new atmospheric layers classified according to the density of free electrons. These layers constitute Earth's *ionosphere*, the partially ionized plasma region that coexists with the neutral atmosphere.

7.1 Introduction

The discovery of Earth's ionosphere has its roots in the history of radio and it was, in fact, an experiment conducted by Guglielmo Marconi in 1901 that prompted widespread interest in its possible existence. Marconi was able to demonstrate that longer than line-of-sight radio transmissions were possible.[1] The explanation of this unexpected observation came in 1902

[1] It would be difficult in our modern world of wireless communication to overstate the importance of this experimental result. Marconi is sometimes represented as an amateur tinkerer. He had no formal college or university education but if that makes him an amateur, he was brilliant one. This monumental experiment occurred during December, 1901 and resulted in the letter 'S' being transmitted, via Morse Code, from a station in Cornwall, England to a receiving station in Signal Hill, Newfoundland. The story of

independently from Arthur Kennelly and Oliver Heaviside who suggested that a conducting layer above the Earth would reflect transmitted signals back to ground level, allowing for the achieved propagation distances. The proposed layer was named the "Kennelly-Heaviside layer" and more direct evidence for its existence was obtained in 1925 by Edward Appleton and Miles Barnett using instruments very similar to what are today known as ionosondes. Over time, the Kennelly-Heaviside layer became known simply as the ionosphere.[2] Appleton was later awarded a Nobel prize for his work in the field.[3]

7.2 Static Electron Density Profile

There are two necessary ingredients for a planet to have an ionosphere:

1. something to ionize (*i.e.,* a neutral atmosphere)
2. something to do the ionizing (*e.g.,* solar EUV radiation)

and of course Earth satisfies these requirements. An assumed equilibrium between the sources and losses of ionization at each altitude then determines the vertical structure of the resulting ionosphere. This vertical structure is in fact not static but rather is continually changing in response to differing conditions that alter the sources and losses of ionization. Despite these variations, we can consider the two requirements above as they apply to Earth and make an educated guess as to what form the vertical structure of the ionosphere should take.

7.2.1 A Qualitative Argument

Figure 7.1 shows the general dependence with altitude of the neutral atmospheric density and the intensity of solar EUV radiation. The neutral density

Marconi and this expiment is well documented and worth reading.

[2] This evolution of the name should be expected. In space physics, the usual geometry of things leads us to call most things a (insert prefix here)sphere.

[3] Edward V. Appleton was awarded the 1947 Nobel Prize in Physics "for his investigations of the physics of the upper atmosphere especially for the discovery of the so-called Appleton layer." In the modern nomenclature presented below, the Kennelly-Heaviside layer is known as the E-region and the Appleton layer is known as the F-region of Earth's ionosphere.

is highest at ground level and generally decays exponentially. The intensity of EUV is highest at the highest altitudes and decreases with decreasing altitude as an ever-increasing fraction of the incident EUV is absorbed by the exponentially increasing atmosphere. If the ionosphere is formed by solar EUV ionizing the neutral atmosphere,[4] it would be reasonable to expect that

> the ionization profile should be approximately proportional to the product of the atmsopheric density and the solar EUV intensity.

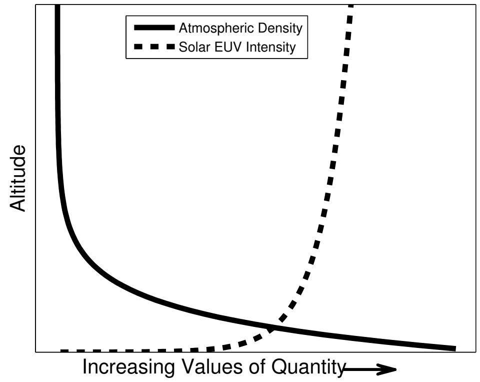

Figure 7.1: Representative vertical profiles of atmosphereic density and solar EUV intensity.

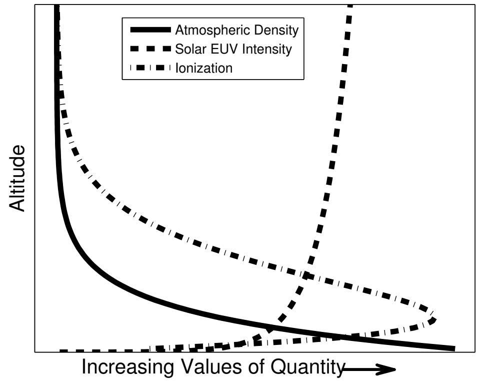

Figure 7.2: Representative vertical profiles of atmospheric density, solar EUV intensity and ionization.

Figure 7.2 shows this product and reveals several interesting features that are generally true of the ionosphere. Note that:

1. At low altitudes, no ionization is expected due to the low intensity of the ionizing solar EUV. (There is nothing to do the ionizing.)
2. At the highest altitudes, no ionization is expected due to the low neutral atmospheric density. (There is nothing to ionize.)
3. At intermediate altitudes the ionization reaches a maximum, increasing sharply from below and decaying relatively slowly as altitude increases from the peak.

[4] We will see below that solar EUV is a primary (but certainly not the only) contributor to the ionization of our neutral atmosphere.

The International Reference Ionosphere (IRI) is an empirical standard model of the ionosphere compiled from a wide variety of available data sources.[5] The model will output, among many other useful things, ionization profiles for any date, latitude, and longitude. For comparison with the profile shown above, Figure 7.3 shows an ionospheric profile from the IRI above Daytona Beach, FL.[6] This profile is similar to the one shown in Figure 7.2 and suggests that our general considerations were correct.

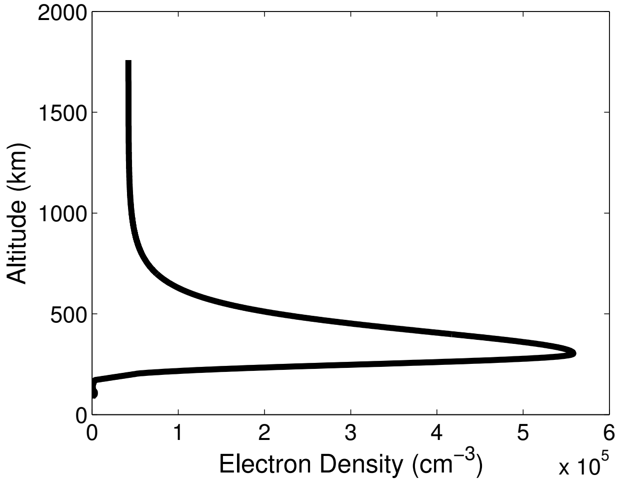

Figure 7.3: Typical IRI model output of the electron density profile above Daytona Beach, FL.

There are two ways to obtain more detailed knowledge of the ionospheric profile: theory and experiment. The IRI model is empirical and we will return to it and other experimental observations for more insights, but let us first turn to theory.

[5] The IRI model is sponsored by the Committee on Space Research (COSPAR) and the International Union of Radio Science (URSI) and is freely available through the web at: `http://modelweb.gsfc.nasa.gov/ionos/iri.html` or `http://modelweb.gsfc.nasa.gov/models/iri.html`.

[6] The geographic latitude of Daytona Beach, FL is 29.2°N and so this location serves as a representative mid-latitude site.

7.2.2 The Chapman $\alpha-$Layer

Sydney Chapman was one of the most accomplished physicists of the 20th century. This highly accomplished and excellent man published more than 450 scientific papers and served as Advisory Scientific Director of the Geophysical Institute in Fairbanks, AK from 1951-1970, during which time he maintained a dual appointment at the National Center for Atmospheric Research (NCAR) in Boulder, CO where he spent the bulk of each year. Dr. Chapman's contributions spanned the fields of kinetic theory of gasses, meteorology, geomagnetism, and ionospheric physics. His derivation of the ionospheric profile is classic. Given below is a derivation based on Rishbeth and Garriott [1969, pp.89-94] of the so-called *Chapman $\alpha-$layer*.

The Chapman $\alpha-$layer is a steady-state model of the ionospheric electron density profile obtained by balancing, as a function of altitude and solar zenith angle, the production rate of photoelectrons with losses due to recombination. We will assume that photoelectrons are produced at a given altitude by incoming solar EUV and that recombination at the same altitude is proportional to the product of electron and ion densities. Taking the constant of proportionality to be α, the Chapman $\alpha-$layer results.

The derivation begins with a number of simplifying assumptions. It is assumed that a single wavelength of solar EUV is responsible for any ionization, that the atmosphere consists of a single species of gas, that the atmosphere is plane, horizontally stratified, and that the scale height is constant. While it is of course true that none of these assumptions are strictly valid, the derived result is useful and interesting nonetheless.

Figure 7.4 shows the coordinate system we will use in the derivation. Altitude h is measured, as always, from the ground up, χ is the solar zenith angle (the angle between the local vertical and the Sun), and s is distance measured along the direction of solar EUV propagation. The intensity of solar EUV at the top of the atmosphere is denonted as I_∞.

As the radiation penetrates the neutral atmosphere, some of it is absorbed and the intensity varies according to Beer's law[7]: $\frac{dI}{ds} = -I\sigma n$ where σ is the absorption cross-section and n is the neutral number density. That is, the decrease in intensity with distance along the path is proportional to the number of neutral particles per unit volume and the cross-section (the effective

[7] Beer's law is also known as the Beer-Lambert law, the Lambert-Beer law, or the Beer-Lambert-Bougeur law. Apparently the law was discovered by Pierre Bouguer (1698-1758) before 1729.

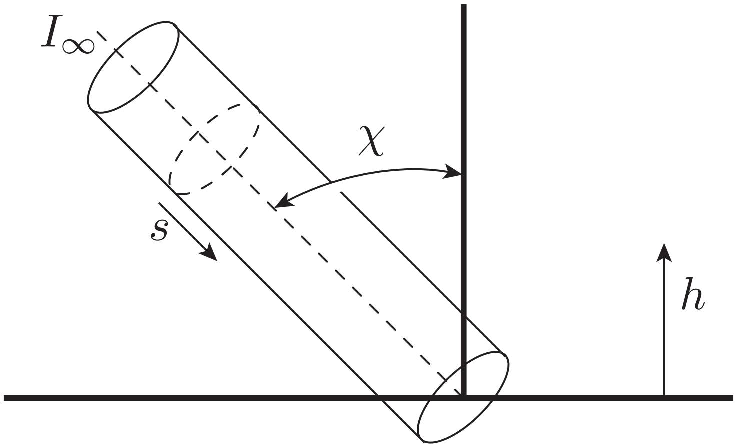

Figure 7.4: The coordinate system used to derive the Chapman α-layer.

absorbing area) of each neutral. If we define an "ionization efficiency" η to be the number of photoelectrons produced per unit energy absorbed in this way, it is clear that the rate of ionization is given by[8]

$$q = -\eta \frac{dI}{ds} = \eta I \sigma n \tag{7.1}$$

which yields the number of electrons freed per unit volume per unit time. As one would expect, this rate is proportional to the energy absorbed. We will assume that η is constant.

The following bit of notation will prove useful. For an element ds of the path of radiation, let us define an increment of optical depth as $d\tau = -\frac{dI}{I} = \sigma n ds$. Integrating this equation from the top of the atmosphere to the point of unit optical depth gives:

$$\int_{I_\infty}^{I} \frac{dI'}{I'} = -\int_0^\tau d\tau'$$

which yields

$$I = I_\infty e^{-\tau}. \tag{7.2}$$

Thus,

[8] As a sure sign that we have run out of letters, the q in the following equation is the ionization rate and is not a particle's electric charge.

for every unit of optical depth travelled along the path s, the intensity decreases by a factor of e.

Remember that our goal is to derive the steady-state electron density profile. To do so, we must evaluate the production function q, balance it with losses due to recombination and evaluate the result as a function of altitude. Presently, we know that the production function depends on the derivative of intensity which varies with optical depth. So let us obtain an expression for the optical depth as a function of altitude. Certainly

$$\tau(s) = \int_0^s \sigma n \, ds$$

but we require $\tau(h)$, not $\tau(s)$. The change of variable is accomplished by relating altitude (h) to the ionizing radiation's path (s). From the geometry of Figure 7.4, we have $dh = -\cos\chi \, ds$ so that $ds = -\sec\chi \, dh$ and

$$\tau(h, \chi) = -\int_\infty^h \sigma n \sec\chi \, dh = \sigma \sec\chi \int_h^\infty n \, dh.$$

This integral can be simplified. Assuming a constant scale height, we have $n(h) = n_0 e^{-\frac{h}{H}}$ so that

$$\int_h^\infty n \, dh = \int_h^\infty n_0 e^{-\frac{h}{H}} \, dh = n_0 H e^{-\frac{h}{H}} = n(h)H.$$

This interesting result reveals that if the neutral atmosphere above a certain altitude h were to be compressed to the same pressure or density as that present at h, the thickness of the resulting layer would be exactly one scale height H. Combining this result with our previous expression for $\tau(h, \chi)$ gives

$$\tau(h, \chi) = \sigma n(h) H \sec\chi. \tag{7.3}$$

We are now in a position to evaluate the production function q as function of altitude and solar zenith angle.

$$q = \eta I \sigma n = \eta I_\infty e^{-\tau(h,\chi)} \sigma n_0 e^{\frac{h}{H}} = \eta I_\infty \sigma n(h) e^{-\tau(h,\chi)}. \tag{7.4}$$

Knowing the production function, we can determine the location in the ionosphere where it takes it maximum value. That is, we can determine the location where the most photoelectrons are produced per second per unit volume. To do this, we must take the derivative of q with respect to some spatial variable and, since $\frac{dI}{ds}$ is already known from Beer's law, let us find the maximum of q in terms of the path s. From Equation 7.4, q maximizes where

$$\frac{dq}{ds} = \frac{d\left(\eta\sigma In\right)}{ds} = \eta\sigma\left(n\frac{dI}{ds} + I\frac{dn}{ds}\right)_m = 0$$

where the m subscript indicates that the given quantity is to be evaluated at the peak of q. The terms in parentheses summing to zero, we can separate variables and evaluate each individually:

$$\frac{1}{n}\frac{dn}{ds} = \frac{1}{n}\frac{dn}{dh}\frac{dh}{ds} = \frac{1}{n}\left(\frac{-n_0}{H}e^{-\frac{h}{H}}\right)(-\cos\chi) = \frac{\cos\chi}{H}$$

and Beer's law gives

$$\frac{1}{I}\frac{dI}{ds} = -\sigma n_m.$$

Summing these two terms to zero and rearranging yields an important insight that we will make use of several times:

$$\sigma n_m H \sec\chi = \tau(s_m, \chi) = 1. \tag{7.5}$$

That is, assuming the scale height H is constant, the production of photoelectrons is maximized at "optical depth unity" where $\frac{I}{I_\infty} = \frac{1}{e}$ or where the intensity of solar EUV is $\approx \frac{1}{3}$ of its value at the top of the atmosphere.

Let us use this insight to find the maximum production rate when the Sun is directly overhead ($\chi = 0$). Affixing the subscript 0 to indicate a quantity for which $\chi = 0$, we have from Equation 7.4

$$q_{m_0} = I_\infty \eta\sigma n_m e^{-\tau(s_m,0)} = \frac{I_\infty \tau(s_m,0)}{H}e^{-\tau(s_m,0)}$$

but $\tau(s_m, \chi) = 1$ so that

$$q_{m_0} = \frac{\eta I_\infty}{eH}.$$

Using Equation 7.5 and a bit of algebra, the normalized production rate at all heights and solar zenith angles can be expressed as

$$\frac{q}{q_{m_0}} = \frac{I}{I_\infty}\frac{ne}{n_{m_0}} = \frac{I}{I_\infty}e^{\frac{1-(h-h_{m_0})}{H}}.$$

If we define zH to be the "reduced height" and again make use of Equation 7.5 to realize that

$$\frac{I}{I_\infty} = e^{-\tau} = e^{-\frac{n\sigma H n_{m_0} \sec\chi}{n_{m_0}}} = e^{e^{-z}\sec\chi}$$

then we find that the production function is given by

$$q(z,\chi) = q_{m_0} e^{1-z-e^{-z}\sec\chi}. \tag{7.6}$$

Figure 7.5a shows a plot of this production function, normalized by q_{m_0}, as a function of reduced height. Note that, as should be expected, the production of photoelectrons maximizes for $\chi = 0$ (*i.e.,* when the Sun is directly overhead) and that, as the Sun sets, the maximum production rate of photoelectrons both decreases in magnitude and shifts to higher altitudes. We will return to this point after balancing this source of photoelectrons with the losses to determine the steady-state electron density profile.

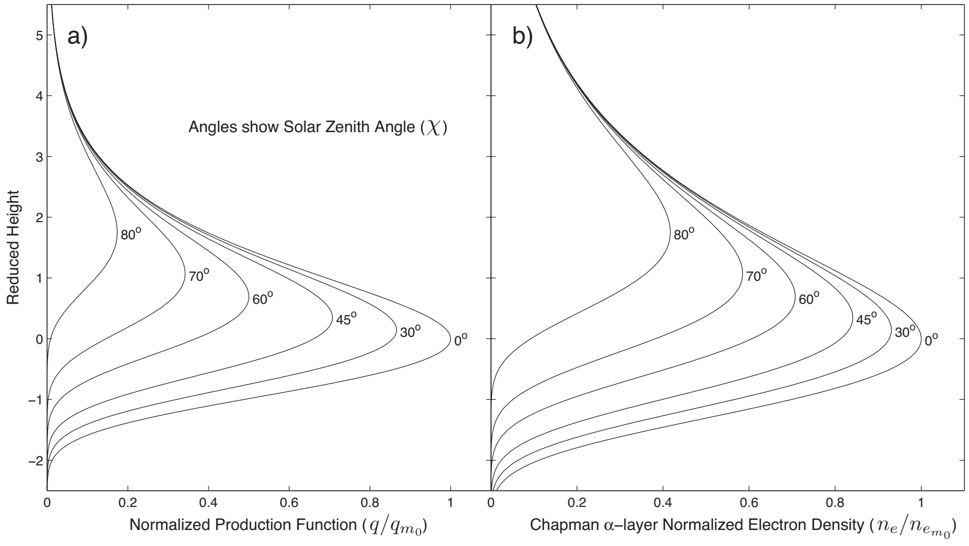

Figure 7.5: Chapman α-layer production function and electron density as a function of altitude and solar zenith angle.

The only loss mechanism we will consider is recombination whereby an electron and ion recombine to yield a neutral. This loss rate should be proportional to the product of the ion and electron densities. Taking the rate constant to be α and enforcing neutrality gives the loss rate $L = \alpha n_e^2$ where n_e is the electron density. In equilibrium, the rate of photoelectron production will equal the rate of loss so that $q = L = \alpha n_e^2$. Substituting the result from Equation 7.6 and solving for n_e gives the electron density profile as

$$n_e(z,\chi) = n_{e_{m_0}} e^{\frac{1}{2}\left(1-z-e^{-z}\sec\chi\right)} \tag{7.7}$$

which is the Chapman $\alpha-$layer.

Figure 7.5b shows plots of this layer as a function of solar zenith angle and altitude. Note that the electron density profile displays the same trends with solar zenith angle as were observed with the production function:

as the Sun sets and the solar zenith angle increases, the ionosphere decays and lifts.

This decaying and lifting are both due to the fact that, as the Sun sets, the incoming EUV passes through more of the atmosphere at higher altitudes (thus, *lifting*) and the condition of unity optical depth where the production is maximized is reached at higher altitudes where the density of neutrals available to ionize is lower (thus, *decaying*).

This simple theory of the Chapman $\alpha-$layer is useful for the insights it provides but, as it turns out, is too simple to predict many of the ionosphere's observed features. Complications that break each of Chapman's assumptions are more or less important at different altitudes and, as a result, the ionosphere rarely closely resembles an actual Chapman layer. Let us therefore turn to observations for further insights.

7.3 Ionospheric Layers

Notice again Figure 7.3 (p. 162) that shows a typical electron density profile above Daytona Beach, FL obtained from the IRI model and note two things in particular. First, the electron density in the ionosphere apparently varies by more than five orders of magnitude. Second, while the large peak at approximately 300 km is the most obvious feature, there is a slight hint of a smaller peak near 100 km. These items suggest that the linear scale used to plot Figure 7.3 may be concealing interesting structure and that a logarithmic scale might be more revealing.

Figure 7.6 shows the same electron density profile as Figure 7.3 but, in this case, the horizontal axis is logarithmic. We do in fact see more structure in this figure and, in a manner similar to what was used to identify atmospheric layers classified according to the neutral temperature profile, we identify and name three ionospheric regions (or layers):

1. The $D-$region is the lowest-lying ionospheric layer and exits from about 50-90 km.
2. The $E-$region spans ∼90-140 km and peaks near 110 km.
3. The $F-$region begins near 140 km and extends upwards in altitude until it merges with the magnetosphere.

As seen in Figure 7.6, there is no real peak to the $D-$region and it appears as a mere "shoulder" on the $E-$region that lies above it. The $F-$region is broken into two subregions: the F_1-region and the F_2-region with the F_1-region appearing as a shoulder on the above-lying F_2-region.

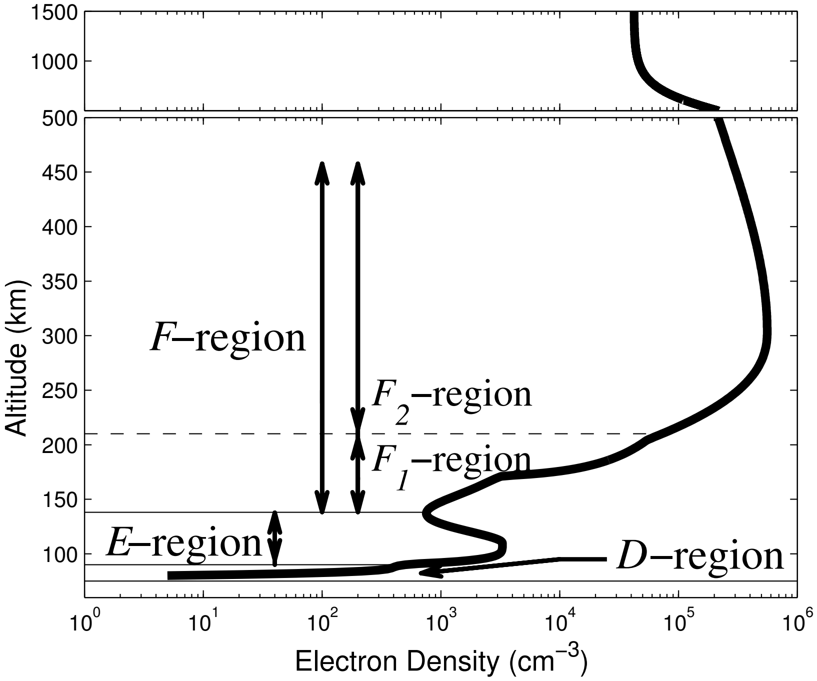

Figure 7.6: Typical IRI model output of the electron density profile above Daytona Beach, FL. The logarithmic scale reveals several ionospheric layers.

Given the prediction from Chapman theory of a single ionospheric layer, we may well ask why there are these several ionospheric layers rather than a single one. The answer is essentially that nature is more complicated than Chapman theory and its several simplifying assumptions. Indeed, nature is *so* complicated that an entire course (or lifetime!) could be dedicated to the physics involved in the generation of the several layers. Here we will give a brief overview of the variations that are observed in the layers and of the physical processes responsible for their formations. We do this first for the quiet-time[9] ionosphere and then for the disturbed ionosphere.

7.4 Variations in The Quiet-Time Ionosphere

There are many sources and sinks of ionization in the quiet-time ionosphere and all of them vary with location on Earth. It should then not be suprising that the quiet-time ionosphere exhibits many systematic variations. This section will introduce several of the most important or dominant, which include diurnal, latitudinal, seasonal, and solar cycle variations, and will introduce some important chemical and transport processes.

Diurnal Variations

Figure 7.7 shows IRI electron density profiles over Daytona Beach at local midnight and noon on March 15, 2002 and Figure 7.8 shows the maximum electron densities in the $E-$ and $F-$regions for each hour on the same day. Several important features can be noted:

1. Ionospheric electron densities are much higher during the day than during the night (Figures 7.7 and 7.8).

2. While the $F-$region decays relatively slowly after sunset at approximately 1900 LT, the $E-$region decays much more rapidly and has a larger fractional change (Figure 7.8).

3. While the $D-$region is not observed at midnight, both the $E-$ and $F-$regions are clearly present throughout the night (Figure 7.7).

[9] By *quiet-time*, we mean during average conditions in the absence of any enhanced solar, magnetospheric, or auroral activity.

4. Both the $E-$ and the $F-$regions reappear suddenly at sunrise at approximately 0500 LT (Figure 7.8).

5. The F_1-region, while present at noon as a shoulder below the F_2-region, has disappeared by midnight (Figure 7.7).

6. The altitude of the $F-$region maximum is somewhat lower in altitude at noon compared to its altitude at midnight (Figure 7.7).

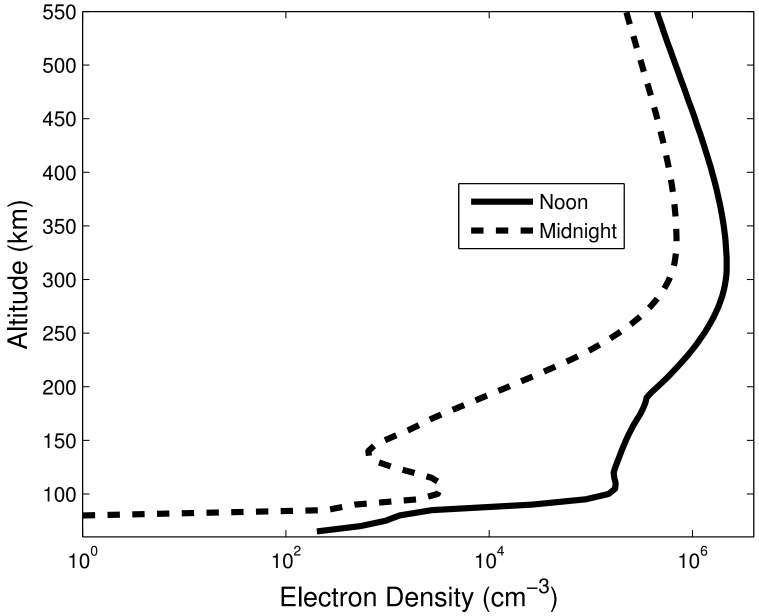

Figure 7.7: IRI electron density profiles above Daytona Beach, FL at local noon and midnight on March 15, 2002.

The first and last of these items are consistent with Chapman theory from which we learned that, as the Sun sets and the solar zenith angle increases, the ionosphere both decays and lifts. Item number four suggests that solar EUV is a dominant daytime source of ionization but item number 3 suggests that there may be other sources active at night. In fact there are a great number of ionoziation sources active during both daytime and nighttime and Table 7.1 lists the most dominant.

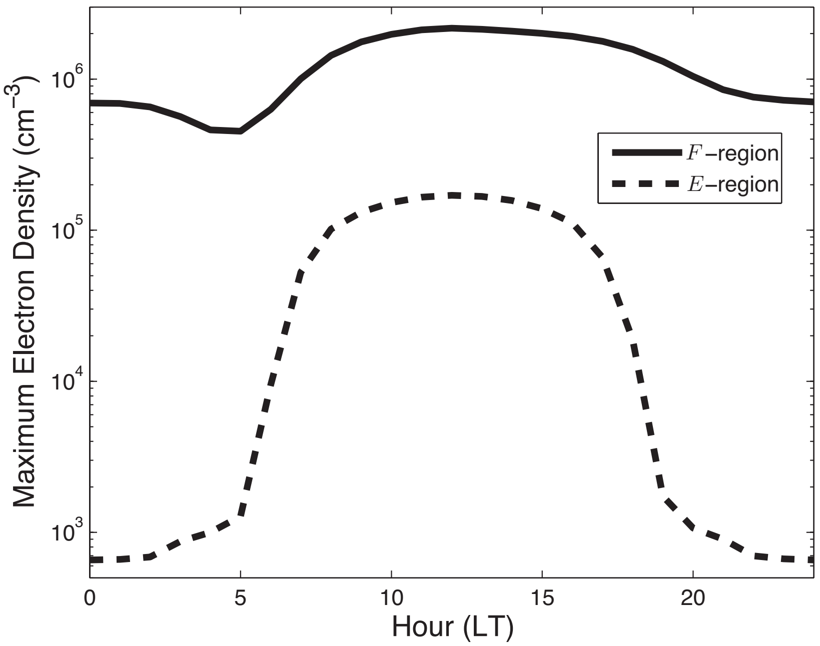

Figure 7.8: The maximum electron densities in the $E-$ and $F-$regions versus local time on March 15, 2002 (from IRI).

Region	Day	Night
D	Solar Ly-α (1216Å) Galactic X-rays (1-10Å) Glactic cosmic rays	Scattered Ly-α (Geocorona) Galactic cosmic rays
E	Solar Ly-β (1027Å) Solar Ly-α (1216Å) Solar EUV (911-1027Å) Galactic X-rays (10-170Å)	Scattered Ly-β (Geocorona) Scattered Ly-α (Geocorona)
F	Solar He$^+$ (304Å), He (584Å) Solar EUV (170-911Å)	Scattered He (Geocorona) Conjugate photoelectrons

Table 7.1: Dominant day and nigh quiet-time ionospheric ionization sources. The sources are listed in decreasing order of dominance.

With the single exception of conjugate photoelectrons, each of the listed sources ionizes neutral constituents and produces a photoelectron. As we have mentioned before, an equilibrium balance between the sources and losses due primarily to recombination results in the electron density profile. Speaking generally, we can say that differences in ionization sources, atmospheric constituents, densities, and chemistry account for the formation of the different ionospheric layers.

Ionization in the quiet-time ionosphere is generally accomplished through photoionization, the process whereby a photon (EUV, X-ray, gamma ray, etc) imparts enough energy to an atom or molecule to free an electron. Photoionization processes have the general form

$$X + h\nu(\lambda < 100\text{nm}) \rightarrow X^+ + e^-$$

where X is a neutral atom or molecule. This reaction indicates than a photon with sufficient energy may ionize the neutral, resulting in an ion and a free electron.

Once the ion and free electron have been created, a variety of chemical processes take place that determine, in the end, which ion species are the most numerous. The most important of these chemical processes are charge exchanges wherein colliding neutral and ionized species exchange charge, possibly dissociating in the process. Examples of charge exchange reactions are

$$N_2^+ + O_2 \rightarrow N_2 + O_2^+ \quad (\text{non} - \text{dissociative})$$

and

$$O^+ + N_2 \rightarrow NO^+ + N \quad (\text{dissociative}).$$

Given the variety of atomic and molecular species present in Earth's atmosphere, the student should not be suprised to read that many dozens of possible reactions could be identified and that each of the reactions are more or less likely to occur as quantified by an associated reaction rate.

The dominant loss mechanism in the ionosphere is recombination that generally takes two forms:

1. Radiative recombination wherein an ion and electron combine to yield a neutral and a photon.

2. Dissociative recombination wherein a molecular ion and an electron combine to yield two (possibly excited) neutral consitutents.

For example,

$$NO^+ + e^- \rightarrow N + O.$$

As it turns out, typical reaction rates for dissociative recombination reactions are very much larger than those for radiative recombination reactions. Thus, loss mechanisms are much more efficient in the lower ionosphere ($D-$region) where there are significant molecular densities than in the $F-$region that is dominated by less massive atomic species. It is this difference in loss mechanisms that is principally responsible for the $F-$region existing throughout the night while the $D-$region essentially disappears at sunset.

Figure 7.9 shows the principle charged constituents in the IRI ionosphere over Daytona Beach on March 15, 2002. Note that ion densities in the lower parts of the ionosphere are dominated by heavy molecular species while the $F-$region is dominated by lighter atomic species. As altitude continues to increase, lighter atoms become more predominant until, above approximately 1000 km, ionized hydrogen is the principle ion. The ionosphere above approximately 1000 km is therefore often referred to as the *protonosphere*.

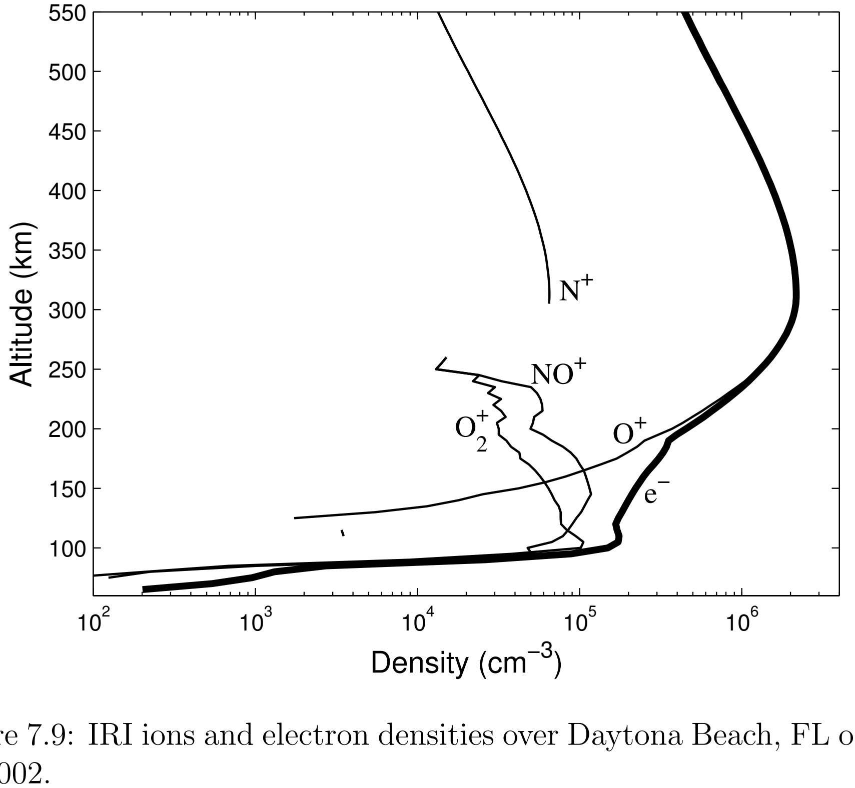

Figure 7.9: IRI ions and electron densities over Daytona Beach, FL on March 15, 2002.

The goecorona and conjugate photoelectron sources listed in Table 7.1 deserve more explanation. Geocorona is the "glow" of scattered light (mainly solar far ultraviolet) that surrounds Earth to a distance of at least $15R_E$.[10] Some of this scattered light reaches Earth's nightside where it can serve as an ionization source. Conjugate photoelectrons arrive due to the tilt of Earth's magnetic field and the high conductivity along a magnetic field line. As illustrated in Figure 7.10, photoelectrons produced in the sunlit hemisphere can easily travel along a field line where they contribute to the electron density in the nighttime hemisphere.

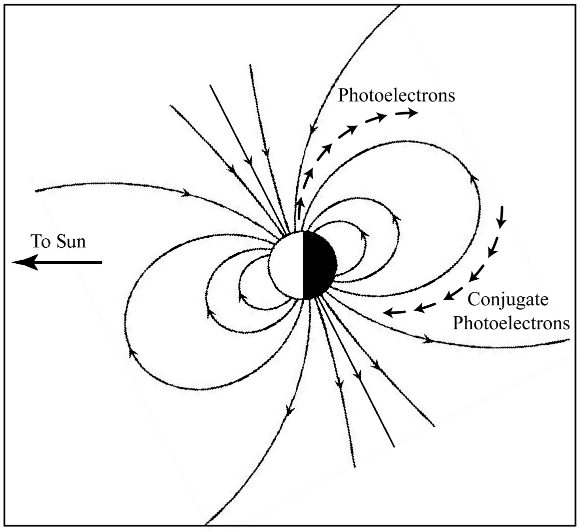

Figure 7.10: An illustration showing how photoelectrons produced in the sunlit hemisphere appear in the nighttime hemisphere as conjugate photoelectrons. The tilt of Earth's magnetic field is exaggerated for effect.

Latitudinal Variations

There is at least one reason to suspect variations in ionospheric electron density with latitude: on a given day at a given local time, the solar zenith angle varies with latitude. Thus we expect a dominant effect to be that electron densities will be highest at latitudes where the Sun is overhead at

[10] A quick web search will result in many impressive pictures of the geocorona.

noon. For example, at the equinoxes, the Sun is overhead at noon on the equator and it is there that we expect to find the highest noon-time electron densities. Equinoctal electron densities at noon should generally decrease with increasing distance from the equator.

Figure 7.11 shows the maximum IRI $F-$region electron densities on March 15, 2002 as a function of geomagnetic latitude. The general effect mentioned above is clearly evident with electron densities generally increasing towards the equator. But there are two major departures from this general trend. First, notice the *equatorial anomaly* which is a *decrease* in the electron density at the magnetic equator and an apparent *increase* on either side extending to about $\sim \pm 20°$ magnetic latitude. Second, a *mid-latitude trough* appears at night as a region of depressed electron density centered at just under 60° magnetic latitude.

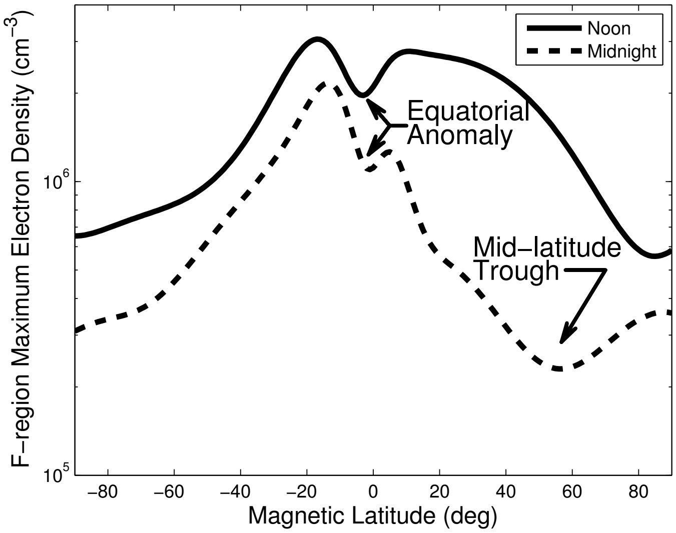

Figure 7.11: The noon and midnight maximum $F-$region electron densities versus geomagnetic latitude on March 15, 2002 (from IRI).

The equatorial anomaly results from the so-called *fountain effect* that essentially acts to eject $F-$region ionization from the magnetic equator and deposit it at $\sim \pm 20°$ magnetic latitude. A detailed explanation of the fountain effect is beyond the scope of this text but it may be noted here that it is due to the $\mathbf{E} \times \mathbf{B}$ drift of equatorial plasma where the electric field is an $F-$region eastward-directed field[11] and the magnetic field is the horizontal, northward-pointing equatorial geomagnetic field. Figure 7.12 illustrates the resulting drift. Once the ionization has been ejected, it settles along magnetic field lines to approximately ±20° magnetic latitude under the influence of gravity and pressure gradient forces.

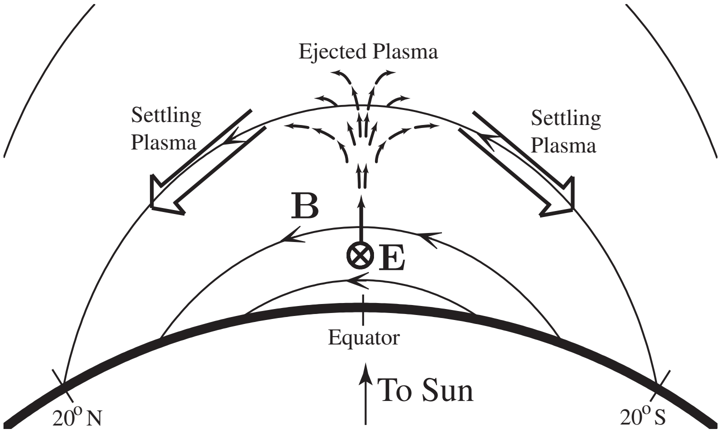

Figure 7.12: The equatorial anomaly and plasma fountain.

A *mid-latitude trough* is commonly observed on the nightside at subauroral latitudes on magnetic field lines mapping to the plasmapause that marks the boundary between the plasmasphere and the plasmasheet. It is due to stagnation of convection in the high-latitude ionosphere (disucssed in §7.9). This convection provides a source of ionization to the high-latitude nightside ionosphere but does not extend to plasmaspheric latitudes. The termination of of this additional nightside ionization source results in a decrease (or trough) in electron density.

[11] This electric field results from collisional coupling of atmospheric winds with the $E-$region plasma. A short and not-very-precise explanation is as follows: High solar heating at the equator tends to drive $-\nabla p$ winds away from the equator. These north/south winds drag the local $E-$region plasma with them that, given the northern- and southern-hemisphere magnetic field geometry, result in an eastward-directed electric field (from Equation 3.5). This electric field maps up in altitude along the nearly equipotential magnetic field lines where a eastward $F-$region electric field acts with the local geomagnetic field to result in an upward $\mathbf{E} \times \mathbf{B}$ drift.

Solar Cycle Variations

As we have seen in Chapter 3, solar activity varies on an 11-year cycle. While the sunspot number may be the most familiar parameter that varies over the cycle, many other quantities vary as well. In particular, the flux of ionizing X-rays and EUV vary in concert with the sunspot number. As a result, Earth's ionospheric densities also vary with the solar cycle and both daytime and nighttime electron densities tend to be higher during solar max than during solar min. Figure 7.13 shows the sunspot number[12] and maximum IRI ionospheric $F-$region electron densities ($\mathrm{n_mF}$) over Daytona Beach for the past three solar cycles. The high degree of correlation is obvious. Although the maximum $D-$ and $E-$region densities are not shown in this figure, they are also correlated with sunspot number although the magnitude of their relative fluctuations are less than for the $F-$layer.

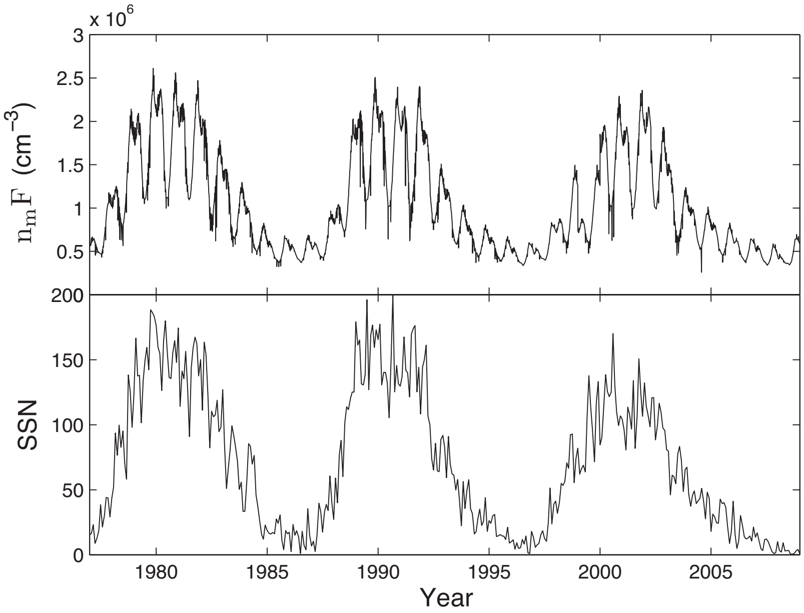

Figure 7.13: Average monthly sunspot numbers (bottom panel) and maximum IRI $F-$region electron densities (top panel) for the past three solar cycles.

[12] From: http://www.ngdc.noaa.gov/stp/SOLAR/ftpsunspotnumber.html#international

Figure 7.14 shows IRI electron density profiles over Daytona Beach at noon and midnight during solar max (March 15, 2002) and solar min (March 15, 2008) conditions. Note that the maximum $F-$region electron densities at solar max are increased by nearly a factor of 10 over their values at solar min. In fact, for some altitudes the midnight solar max electron densities in the $F-$region exceed those at noon during solar min! It can also be seen in this plot that, as was mentioned in the previous paragraph, relative variations in the $D-$ and $E-$ regions over the solar cycle are much less significant.

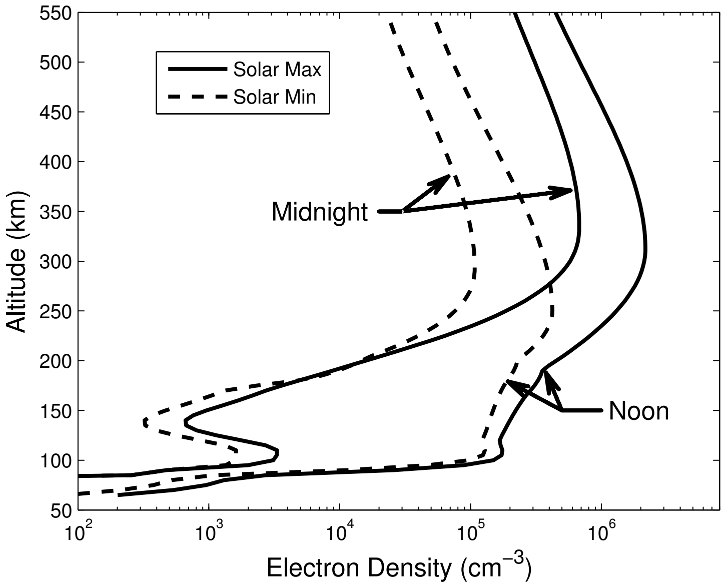

Figure 7.14: IRI electron density profiles over Daytona Beach, FL at noon and midnight during solar max (March 15, 2002) and solar min (March 15, 2008) conditions.

Another feature in the electron density data shown in Figure 7.13 can be detected by the discerning eye. Superposed on the solar-cycle variation is a seasonal variation in $\mathrm{n_mF}$ known as the *seasonal anomaly* wherein electron densities are higher in the winter than in the summer.[13] Main causes of this anomaly are related to seasonal variations in the neutral atmsopheric chemical composition and temperature. As the chemical composition varies

[13] By a factor of 2 or so.

with season and as reaction rates vary with temperature, the ionospheric density varies in response.

7.5 Vertical Sounding of the Ionosphere

Instruments known as *ionosondes* and the data they produce are nearly ubiquitous in ionospheric studies and it is no exaggeration to say that they have contributed significantly to a large fraction of our understanding of ionospheric structure and variability. In their most basic form, ionosondes are essentially radars that transmit short-duration pulses over a broad range of frequencies and measure the time delay required for each pulse to return to the instrument. These time delays are then processed to obtain a profile of the ionospheric electron density.

Before discussing this instrument and the data it produces, it will be helpful to introduce a major result from magnetoionic theory, the theory that describes how electromagnetic waves propagate through a magnetized plasma. This result is the index of refraction for a wave of frequency ω. Let us first review and introduce some new notation and terminology. The electron plasma frequency given in Equation 2.12 defines an unmagnetized plasma's natural oscillation frequency when perturbed from equilibrium and the electron gyrofrequency given by $\omega_{ce} = \frac{eB}{m_e}$ defines the frequency with which an electron gyrates around a magnetic field line. If we then let the electron/neutral collision frequency be ν_{en}, we may define the following dimensionless quantities:

1. $X = \frac{\omega_{pe}^2}{\omega^2}$

2. $Y = \frac{\omega_{ce}}{\omega}$

3. $Z = \frac{\nu_{en}}{w}$

where ω is the electromagnetic wave frequency.

The reasons for making these definitions is that the index of refraction we require is a complicated expression and these definitions are used to simplify its appearance somewhat. First obtained by E. V. Appleton and known as the *Appleton-Hartree dispersion relation*, the index of refraction of an electromagnetic wave in a magnetized plasma is

$$n^2 = (\mu - i\chi)^2 = 1 - \frac{X}{1 - iZ - \left(\frac{Y_T^2}{2(1-X-iZ)}\right) \pm \left(\frac{Y_T^4}{4(1-X-iZ)^2} + Y_L^2\right)^{\frac{1}{2}}} \tag{7.8}$$

where μ and χ[14] are the real and imaginary parts of n, respectively, $Y_L = Y\cos\theta$, $Y_T = Y\sin\theta$, θ is the angle between the wave normal and the magnetic field and the ± allows for different wave polarizations. For perpendicular propagation (that is, for propagation in a direction perpendicular to the magnetic field), the '+' sign represents the *ordinary* mode in which the polarization is linear and along the magnetic field and the '-' sign represents the *extraordinary* mode in which the polarization is linear and perpendicular to the magnetic field. For parallel propagation, the '+' and '-' signs refer to left- and right-hand circularly polarized signals, respectively. The ordinary mode is so-named because waves with this polarization propagate through the plasma as if it were unmagnetized since electrons are accelerated along the magnetic field and are therefore unaffected by its presence. In the extraordinary mode, however, electrons are accelerated in the direction perpendicular to the magnetic field and are therefore affected by both the electric and magnetic terms in the Lorentz force equation.

Equation 7.8 is complex, complicated and unwieldy and we seek simplifications to aid in our understanding of wave propagation through a plasma. Ignoring electron/neutral collisions (by setting $Z = 0$) and any effects due to the magnetic field (by setting $Y = 0$), we obtain a much simplified result:

$$\mu^2 = 1 - X = 1 - \left(\frac{\omega_{pe}}{\omega}\right)^2 \tag{7.9}$$

and $\chi = \Im(n) = 0$ where $\Im(n)$ deontes the imaginary part of n.

There are several points of interest here. First, note that μ^2 is always less than 1 and must be real for the wave to propagate. Since the phase speed v_p is given by $v_p = c/\Re(n) = c/\mu$ where c is the speed of light in a vacuum, we find that the phase speed of an electromagnetic wave in a plasma is *always* greater than c. This is not a violation of special relativity since information does not travel at the phase speed but at the group speed v_g that, for the

[14] Here we are using standard notation that is often, and unfortunately, redundant. It should be carefully noted that μ and χ, when used in this context, bear no relation to the magnetic moment, first adiabatic invariant or the solar zenith angle.

cases we will treat, is given by $v_g = \Re(n)c = \mu c$. We see then that v_g is always less than c. Next, note that μ in this approximation depends only on the ratio of the plasma frequency to the wave frequency. That is, the index of refraction depends only on the wave frequency and the plasma's electron density. Third, note that μ can be real or imaginary. For $\omega > \omega_{pe}$, μ is real and the wave propagates undamped through the plasma, but if $\omega < \omega_{pe}$, μ is imaginary. Under these conditions, the wave is evanescent, its amplitude decreases exponentially and the wave packet will be reflected.

Let us now consider the propagation of an electromagnetic wave pulse launched from a transmitter (specifically, an ionosonde) vertically into the ionosphere. The ionosphere does not extend to ground level where the ionosonde is located, so below the ionosphere we have $n_e = 0$ and, given Equation 7.9, $\mu = 1$. The wave's phase and group speeds equal the speed of light in vacuum. As the pulse enters the bottomside of the ionosphere and the electron density begins to increase, ω_{pe} increases and both μ and the group speed decrease - the wave pulse slows down. As the wave continues to propagate upward into the increasingly more dense plasma of the ionosphere, the pulse will continue to slow down as μ decreases and one of two possibilities will eventually result. On one hand, the pulse may eventually reach an altitude where $\omega = \omega_{pe}$ at which the group speed goes to zero. At this point the wave becomes evanescent and the pulse will be reflected. The trip down will be the revese of the trip up in that the group speed will increase as the wave descends into regions of lower electron density until it finally exits the bottomside of the ionosphere from where it will propagate at speed c to the ionosonde for reception. On the other hand, the wave frequency may be higher than the maximum value of ω_{pe} in the ionosphere in which case the group speed will never equal zero and wave pulse will not be reflected but will penetrate the ionosphere and be lost.

Given this understanding, a few points may be noted. First, it is clear that if we could measure the time between the launch of a pulse at some frequency ω and its return, we could obtain a measure of the altitude at which $\omega = \omega_{pe}$. Second, transmissions at increasingly higher frequencies will penetrate increasingly farther into the ionosphere (and thus take longer to return) until the frequency matches the maximum ionospheric plasma frequency. All higher frequencies will penetrate the ionosphere and will not return. There will be no reflections from the topside[15] of the ionosphere.

[15] The term *topside* refers to altitudes above the peak of the F-region.

Last, sweeping the wave frequency and measuring the time-of-flight for each pulse would then provide a profile of ω_{pe} and from it, a profile of n_e as a function of altitude.

Figure 7.15 illustrates this basic operating principle of the ionosonde. This figure shows the same IRI data as Figure 7.6 but the horizontal axis includes both the electron density and the corresponding electron plasma frequency $f_{pe} = \omega_{pe}/2\pi$. Ionosonde transmissions at varying frequencies are indicated as either reflecting from altitudes where the wave frequency equals the plasma frequency or as penetrating the ionosphere for wave frequencies greater than the maximum plasma frequency.

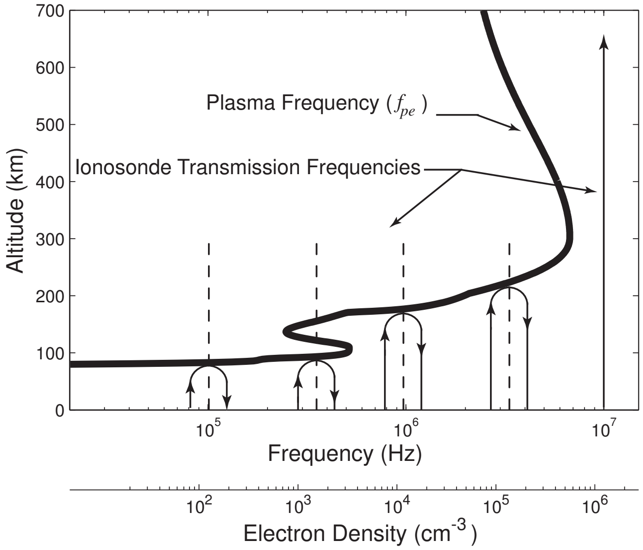

Figure 7.15: An ionospheric profile plotted as functions of electron density and electron plasma frequency. Ionosonde transmissions at frequencies that both reflect and penetrate the ionosphere are shown.

Suppose the time delay between transmission and reception of each transmitted ionosonde pulse shown in Figure 7.15 was measured. The *virtual height* of each corresponding reflection point can then be defined as half the distance travelled at speed c during the delay (since the pulse is travelling vertically upward during only half of the delay). These virtual heights can then be plotted as a function of transmitted frequency to obtain a so-called *ionogram*. Figure 7.16 shows a simulated ionogram obtained by ray-tracing an ordinary mode signal into the same IRI density profile shown in Figure 7.15. In essence, the ray-tracing code evaluates the integral

$$\Delta t = 2\int_0^{h_r} \frac{dh}{v_g(h)}$$

to obatain the delay time and virtual height of the reflection point for each transmitted frequency where, in the above expression, Δt is the time delay between transmission and reception and h_r is the reflection altitude.

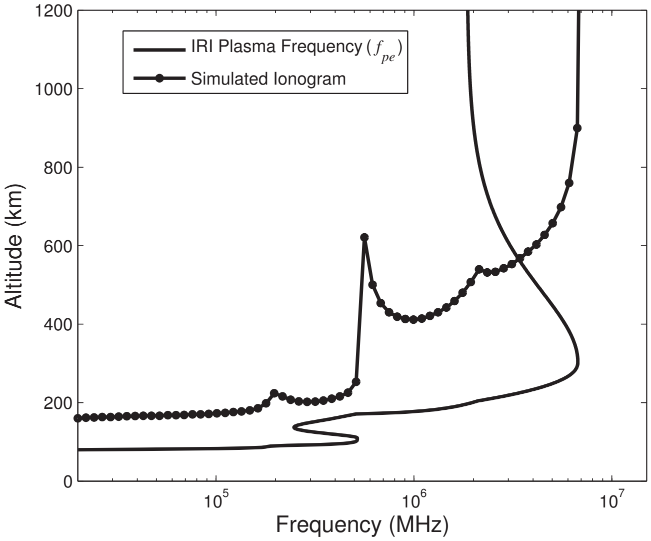

Figure 7.16: A simulated ionogram obtained by ray-tracing an ordinary mode wave into a model ionosphere.

Notice that the virtial heights are all higher than the actual reflection altitudes (or *true heights*) that occur where $\omega = \omega_{pe}$ or, equivalently, where $f = f_{pe}$. The reason for this is that the pulse does not travel at c as assumed in the calculation of virtial height but rather at the group speed that is always less than c in the ionosphere. Thus, the true heights will always be less than the virtual heights. While it is a relatively straightforward task to create a simulated ionogram from an ionospheric electron density profile, in practice the reverse operation must be performed: Ionsondes obtain ionograms from which the true height ionospheric profile must be obtained. This process of obtaining ionospheric profiles from virtual height ionograms is known as *inverting* an ionogram and can be achieved by evaluating for each transmitted frequency the true height given by

$$h_{true} = \frac{1}{2}\int_0^{\Delta t} v_g dt = \frac{1}{2}\int_0^{\Delta t} n(t)cdt.$$

Ionogram iversion is routinely performed by software in real-time and reduces the virtual heights to the true heights that represent the ionospheric profile.

One other feature of the simulated ionogram shown in Figure 7.16 deserves discussion at this point. At frequencies corresponding to the $E-$ and $F-$region peaks, the ionogram displays vertical asymptotes in the virtual heights. These asymptotes appear due to the presence of local maxima in electron density (and thus in plasma frequency) at those altitudes. Transmissions at frequencies equal to the local maximum must travel through a range of altitudes just below the peak where the plasma frequency is very close to, but just less than, the transmitted frequency. Thus, over that range of altitudes the index of refraction and group speed will be very small and the pulse will take a very long time to reach the reflection altitude. The corresponding virtual height will asympotically approach infinity.

There are a large number of ionosondes (perhaps many dozens) operating around the word for the purpose of obtaining ionospheric profiles and monitoring ionospheric conditions and variability and many of the produced ionograms are available in on-line databases. Such databases make available a wealth of information that can be used to diagnose and understand ionospheric contitions.

7.6 The Disturbed Ionosphere

We discussed in §7.4 several systematic vaiations in ionospheric parameters with latitude, local time, season and solar cycle. In addition to these variations, the ionosphere also responds to conditions in the Earth and space environments that vary on much shorter time scales. As a single example among a great many possibilities, the arrival at Earth of a solar flare may suddenly and dramatically alter the structure of the ionosphere and so we now turn our attention to the disturbed ionosphere. That is, we will investigate and consider the consequences of several important departures from quiet-time conditions. We will begin at the lowest altitudes (in the $D-$region) and proceed upward to the $E-$ and $F-$regions.

7.6.1 $D-$region absorption and PCA Events

Broadly categorized, radiowaves are that portion of the electromagnetic spectrum from 3 kHz to 300 GHz and transmissions at these frequencies are often used for communications. Indeed, the GPS, AM, FM, television, aviation and marine bands all lie within the radio band and any ionospheric impacts on these transmission are of significant technological importance.

In our ionosonde discussion, the Appleton-Hartree dispersion relation was introduced and we may recall that the index of refraction of an electromagnetic wave propagating through a plasma such as the ionosphere depends on three quantities, each relative to the wave frequency: the electron plasma frequency ratio X, the electron gyrofrequency ratio Y, and the electron/neutral collision frequency ratio Z. To simplify the full expression, we assumed there was no magnetic field so that $Y = 0$ and no collisions so that $Z = 0$. Let us now consider the effect of the collisional term Z that, as can be seen from Equation 7.8, makes the index of refraction n complex.

Suppose the wave (specifically the radiowave) under consideration is a plane wave with with an electric field amplitude given by $E = E_m e^{i(\omega t - kz)}$ where k is the wavevector. Now,

$$k = \frac{2\pi}{\lambda} = 2\pi\frac{f}{v} = n\frac{\omega}{c}$$

so that if n is real $(Z = 0)$ we have

$$E = E_m e^{i(\omega t - \frac{\mu\omega}{c}z)}$$

and the wave propagates with an undamped amplitude. However, if n is complex $(Z \neq 0)$ then

$$E = E_m e^{i(\omega t - \frac{(\mu - i\chi)\omega}{c}z)} = E_m e^{-\frac{\chi\omega}{c}z} e^{i(\omega t - \frac{\mu\omega}{c}z)}$$

and the wave propagates with the same speed and amplitude as before but its amplitude is damped exponentially. To gain physical insight into the reason for this damping, consider the result of a collision between a neutral and an electron or ion that has been accelerated by the wave's electric field. In accelerating the charged particle, the wave has done work and transferred some of its energy to it. If the charged particle then suffers a collision with a neutral, at least some of this energy will be immediately lost or, at best, transferred to the neutral. In either case, the energy will not be transferred back to the wave and it therefore suffers damping. We see then that radiowaves propagating through the ionospheric plasma that is embedded in the neutral atmosphere will suffer damping to an extent dependent on the number density of charged and neutral particles.

Neutral densities in the $D-$region are several orders of magnitude higher than at the peak of the $E-$region and many order of magnitude higher than in the $F-$region. We should therefore expect that a significant amount of $D-$region ionization would lead to radiowave absorption, and indeed it does. Viewed in this light, we realize it is quite convenient that the $D-$region normally has very low electron densities. Tyically, there are very few electrons in the $D-$region to be accelerated and to loose their energy through electron/neutral collisions. This is, however, not always the case. Auroral substorms or solar storms such as flares and coronal mass ejections may temporarily elevate $D-$region charged particle densities to several times their normal levels and during such events, radiowave propagation is very significantly degraded.

At the onset of an auroral substorm, large numbers of electrons precipitate into the lower ionosphere and radiowave absorption increases in response. Figure 7.17 shows approximately one hour of radiowave data[16] collected by a broadband ground-based radio receiver stationed at Arviat in northern Canada's Nunavut Territory. In this figure, radiowave intensity is plotted as a function of frequency and time with darker pixels indicating higher intentities. Many shortwave fixed-frequency transmissions are clearly visible as is the AM broadcast band from $\sim 500 - 1600$ kHz. On this day, an auroral substorm occured at about 0652 UT, and the intensity of radiowaves over much of the plot is greatly diminished due to ionospheric absorption. This absorption is most evident in the near disappearance of the AM broadcast band. It is interesting and the subject of much research that the substorm onset was accompanied by the generation of auroral radio waves (in addition to visible auroral emissions), three examples of which can be seen here. Auroral roar occurs at ~ 3000 kHz, begins before onset and continues through the expansion phase. MF burst and auroral hiss are the broadband emissions first visible at substorm onset.

More dramatic and long-lasting than auroral absorption, *Polar Cap Absorption* (PCA) events result when intense solar flares produce large fluxes of energetic protons (10 MeV) that are guided by Earth's geomagnetic field

[16] Courtesy of Dr. J. LaBelle, Dartmouth College

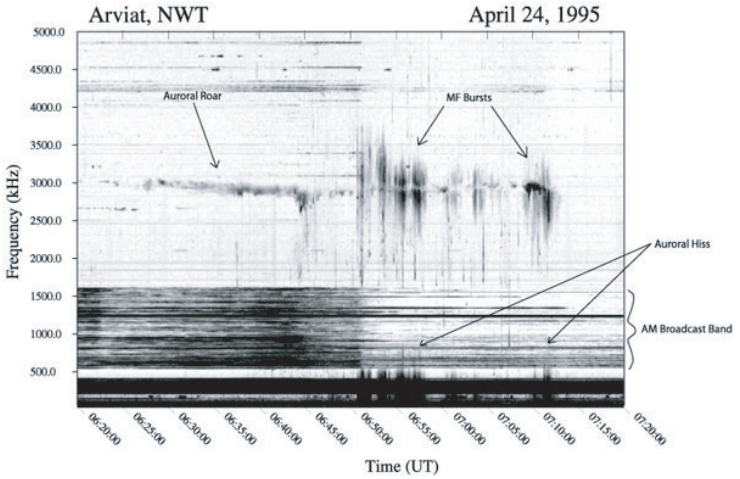

Figure 7.17: Intensity of radiowave signals received at Arviat, Nunavut on April 24, 1995.

lines directly into the polar cap. These highly energetic protons penetrate to $D-$region altitudes and the associated increases in ionization and radiowave absorption last for hours or even days. These dramatic events are detected with, among other instruments, riomenters (discussed below), ionosondes (which show an absence of reflected signals during the event) and radio receiers.

Figure 7.18 shows radiowave data[17] recorded during a 22-day period during which two solar flares occured. The top panel shows radiowave intensity in the same format as Figure 7.17 where the blank (white) intervals represent times during which the receiver was turned off for housekeeping. The bottom panel shows the radiowave intensity integrated over the band relative to the maximum value observed during the period. The times of two solar flares are indicated and the resultant PCA with its decrease in radiowave intensity is obvious and dramatic, reaching a minimum value of more than 30 dB below pre-event levels. This unusually long-lasting PCA event lasted for nearly 2 weeks and, at it height, increased radiowave absorption by more than a factor of 1000.

It is a true saying that "One man's noise is another man's signal" and here we may change the wording a bit to read, "One instrument's absorption

[17] Courtesy of Dr. J. LaBelle, Dartmouth College

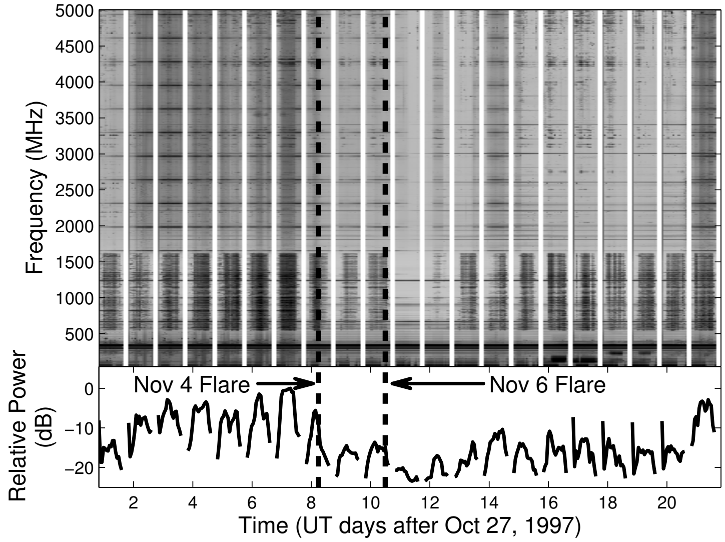

Figure 7.18: Intensity of radiowave signals received at Arviat, Nunavut over a 22 day period containing two solar flares and an intense PCA event. The top panel shows the broadband spectrum and the bottom panel shows the radiowave power integrated over the band relative to the maximum value observed during the period.

is another instrument's signal." Relative Ionospheric Opacity meters, or *riometers* are essentially very sensitive ground-based radio receivers tuned to a portion of the radio band (typically near 30-40 MHz) unused by man but populated with broadband signals from stellar and galactic sources. The intensity of these signals is known very precisely and variations from expected levels can be attributed to absorption by the intervening ionosphere. Riometer data are central to a great number of space environment studies and can reveal, for example, times and loations where energetic charged particles are precipitating into the lower ionosphere.

7.6.2 Sporadic-E

Phenomena known as *sporadic-E*, or E_s for short, are observed on ionograms as $E-$layer echoes that extend to higher frequencies than usual. They are thus layers of higher than usual $E - layer$ charge densities and both rocket and radar observations reveal that they can be very thin, sometimes perhaps less than a kilometer thick. There are actually many different phenomena that, in the end, are classified as sporadic-E and there are accordingly many different causes for the effect. In this section, we will introduce three causes for E_s, one each for low, middle, and high latitudes.

E_s is very common at low latitudes during daytime hours when it can be present more than 90% of the time near the geomagnetic equator. A principle cause of low latitude E_s is instabilities in the equatorial electrojet but a detailed discussion of this point is beyond the scope of this text. The interested reader is referred to Kelley [1989, pp.154ff] for more information.

At middle latitudes, the combined effects of a continuous flux of meteoric metallic ions into the ionosphere and wind shears driven by gravity waves and tides result in E_s that occurs near 110 km in altitude and is most prevalent in summer during the daytime. To understand how the meteoric ions becomes concentrated into E_s, consider their response to a gravity wave or tide with a wavelength such that a strong zonal shear in the wind field is produced. The meteoric ions, which tend to drift with the wind field, will be subjected to a force $\mathbf{F} = q(\mathbf{U} \times \mathbf{B})$ where $\mathbf{U}$ is the wind velocity and $\mathbf{B}$ is the geomagnetic field. At certain altitudes, this force will tend to make the ions converge from above and below, increasing the charge density. $E-$region electrons, on the other hand, are strongly tied to the geomagnetic field lines due to their higher gyrofrequencies and do not respond significantly to the wind shear. The electrons will, however, move up or down a field line into regions of higher ion density so that charge neutrality is approximately maintained. Thus, in regions of high wind shear, significant increases in plasma density are possible.

At high latitudes, prepipitation of auroral electrons with energies in the range of 1-10 keV also leads to the production of E_s. In this case, a precipitating high-energy electron may ionize many neutral atmospheric constituents along its path before its energy is low enough to produce the excitation responsible for visible emissions. Thus, precipitating electrons can produce a flood of ionization over a range of altitudes and result in either thin or thick layers of E_s. Auroral E_s tends to occur, as one would expect, during the nighttime hours and is correlated with geomagnetic activity.

7.6.3 Spread-F

7.7 TEC

7.8 Currents and Conductivities in the Ionosphere

7.8.1 Qualitative Introduction

In an ordinary electric circuit consisting of, for example, a voltage source, some lengths of wire, capacitors, resistors, and other circuit elements, molecular-scale collisions between charge carriers (electrons) and the bulk material impede the flow of current and result in resistance. Except when thinking of actual resistors, we are typically conditioned to think badly of these collisions as they add resistance to otherwise ideal circuit elements. Simply put, in an ordinary electric circuit, it would be ideal if there were no unintended collisions between electrons and the bulk material as current is flowing through a circuit.

As we will see in this section, we must disabuse ourselves of this conditioning if we are to appreciate the processes by which the ionosphere supports the flows of its many important currents. In the ionosphere, collisions between current-carriers (electrons and ions) and the "bulk material" (the neutral atmosphere) are, perhaps surprisingly, often absolutely essential to its ability to conduct current. The reason for this is tied to the forces that drive those currents. Electric and magnetic fields are the dominant current drivers in the ionospheric plasma, far outweighing the influence of, for example, pressure gradient and gravitational forces. To glimpse the importance of collisions to ionospheric conductivity, let us first consider the ionosphere's response to these current drivers in the absence of collisions.

The ionosphere is embedded in Earth's magnetic field and, when acted upon by only this magnetic field and an electric field, we recall from §2.2.2 that all charged particles (electrons and ions) will drift at the $\mathbf{E} \times \mathbf{B}$ velocity in the same direction at the same speed. Thus, an applied electric field will produce no current if the $\mathbf{E} \times \mathbf{B}$ drift is the only effect. In order to drive a current, it must somehow be arranged that electrons and ions do not drift together at the same speed in the same direction. As it will turn out, this arrangement can be effected by electron/neutral and ion/neutral collisions.

Before deriving the ionospheric conductivity and discussing the altitudes over which the ionosphere can most readily support the flow of current, let us consider qualitatively the physics behind the process. As our system, we will take crossed electric and magnetic fields embedded in the ionospheric plasma and the neutral atmosphere. The electric and magnetic fields will accelerate the charged particles and the neutral atmosphere will produce collisions between charged and neutral particles.

Recall that electrons and ions $\mathbf{E} \times \mathbf{B}$ drifing in the absence of collisions are executing two superposed motions: gyration around magnetic field lines and drift of the "guiding center" at the $\mathbf{E} \times \mathbf{B}$ velocity. For a charged particle starting from rest, the $\mathbf{E} \times \mathbf{B}$ drift takes effect in two stages: first the electric force acts to accelerate the particle in the direction of $q\mathbf{E}$ and then the full Lorentz force causes gyration and drift in the $\mathbf{E} \times \mathbf{B}$ direction. The introduction of collisions with neutrals at a frequency ν_c complicates this process and we consider two extremes. First, ν_c may be much smaller than the gyrofrequency. That is, a charged particle may execute many gyrations between collisions. In this case we expect the particle's average motion to be in the same direction as the $\mathbf{E} \times \mathbf{B}$ drift, but perhaps at a somewhat lower speed due to the impeding collisions. Second, ν_c may be larger than the gyrofrequency. That is, there may be several or many collisions during a gyroperiod. In this case, if we assume the charged particle is brought to rest after each collision, we see that the particle will spend more time moving in the direction of $q\mathbf{E}$ than gyrating since its velocity is often zero.

It seems clear then that the ratio ω_c/ν_c where ω_c is the gyrofrequency will play an important role in determining the direction a charged particle moves in response to an applied electric field. If $\omega_c/\nu_c >> 1$, we expect the particle to generally move in the $\mathbf{E} \times \mathbf{B}$ direction but if $\omega_c/\nu_c << 1$ we expect it to move in the direction of $q\mathbf{E}$. Gyrofrequencies vary relatively little in the ionosphere but the collision frequency varies over many orders of magnitude since it is essentially proportional to the neutral density and we will find that each of the two extremes (and, of course, intermdediate values) hold for electrons and/or ions over some range of altitudes.

Our job at this point is twofold. First, we must develop Ohm's law for a magnetized plasma[18]. Second, we will examine the ionosphere's conductivity

[18] While we are perhaps most used to thinking of Ohm's law as $V = IR$, a relation between voltage and current, physicists often use the form $\mathbf{j} = \tilde{\sigma} \cdot \mathbf{E}$ where $\mathbf{j}$ is the current density and $\tilde{\sigma}$ is the conductivity tensor. See §3.6.1 for an introduction to Ohm's law.

profile to determine at which altitudes currents may be expected to flow. The first job will be done using the momentum equation and the second will be done by incorporating observational data.

A general momentum equation accounting for electric and magnetic fields and collisions between charged particles of species s and neutrals is

$$m_s \frac{d\mathbf{v}_s}{dt} = q_s \left(\mathbf{E} + \mathbf{v}_s \times \mathbf{B}\right) - m_s \nu_c \left(\mathbf{v}_s - \mathbf{u}\right) \tag{7.10}$$

where the last term quantifies the momentum loss per second due to collisions of the charged particle with a neutral moving with velocity $\mathbf{u}$.

7.8.2 Conductivity of an Unmagnetized Plasma

As a starting point, we first consider electron current under a number of simplifying assumptions: the electrons are in steady state (so that $\frac{d\mathbf{v}_e}{dt} = 0$), $\mathbf{B} = 0$, both the ions and neutrals are stationary ($\mathbf{u} = 0$) and Coulomb collisions are ignored. Equation 7.10 then reduces to

$$\mathbf{E} = -\frac{m_e \nu_e}{e} \mathbf{v}_e \tag{7.11}$$

where ν_e is the electron/neutral collision frequency and we wish to use this relation to obtain Ohm's law. Since, in this case, there is no magnetic field, all charged particles will move in the direction of $q\mathbf{E}$ and Ohm's law will be given by $\mathbf{j} = \sigma \mathbf{E}$ or, alternatively, by $\mathbf{E} = \eta \mathbf{j}$ where $\eta = 1/\sigma$ is the *resistivity* and σ is the *conductivity.*

The electron current density is

$$\mathbf{j}_e = -e n_e \mathbf{v}_e$$

and we obtain Ohm's law for electrons by solving Equation 7.11 for the electron velocity and substituting into the above relation to obtain

$$\mathbf{j}_e = \frac{n_e e^2}{m_e \nu_e} \mathbf{E} = \sigma \mathbf{E}. \tag{7.12}$$

Ions, of course, also contribute to the current and, since there are typically several ion species in the ionosphere (see, for example, Figure 7.9) their contribution must be summed over species. This can be done by applying Equation 7.10 to each species, solving for the velocity and substituting the result into Ohm's law. Adding the current density contributions from each species (electrons and ions) results in the total current density

$$\mathbf{j} = \sum_s q_s n_s \mathbf{v}_s = \mathbf{j}_e + \sum_i \mathbf{j}_i = \left(\frac{n_e e^2}{m_e \nu_e} + \sum_i \frac{n_i e^2}{m_i \nu_i} \right) \mathbf{E} = \sigma_0 \mathbf{E}$$

where σ_0 is the unmagnetized conductivity and m_i and ν_i are the ion mass and collision frequency, respectively.

Note that the conductivity

$$\sigma_0 = \frac{n_e e^2}{m_e \nu_e} + \sum_i \frac{n_i e^2}{m_i \nu_i} = \frac{\epsilon_0 \omega_{pe}^2}{\nu_e} + \sum_i \frac{\epsilon_0 \omega_{pi}^2}{\nu_i} \tag{7.13}$$

is generally dominated by the electron term since the more massive ions contribute very little to the current due to their comparatively small velocities. A useful approximation is then

$$\sigma_0 \approx \sigma_{0_e} = \frac{n_e e^2}{m_e \nu_e} = \frac{\epsilon_0 \omega_{pe}^2}{\nu_e}. \tag{7.14}$$

The conductivity of an unmagnetized plasma varies as we might expect with electron density and collision frequency. For a fixed electric field, the conductivity and thus the current density decreases with increasing collision frequency and increases with increasing electron density. As in the case of an ordinary circuit with current flowing through wires, collisions in an unmagnetized plasma impede the motion of the current-carriers and decrease the conductivity. Space plasmas are, however, generally magnetized and to obtain the conductivity in this case, we must use the full Lorentz force in the momentum equation.

7.8.3 Conductivity of a Magnetized Plasma

As with the previous case, we begin by assuming the electrons are moving in steady state so that $\frac{d\mathbf{v}_e}{dt} = 0$, that both the ions are neutrals are stationary, and that there are no Coulomb collisions. Equation 7.10 then reduces to

$$\mathbf{E} + \mathbf{v}_e \times \mathbf{B} = -\frac{m_e \nu_e}{e} \mathbf{v}_e \tag{7.15}$$

which must be solved for velocity to obtain the electron current density and the conductivity in Ohm's law. In this magnetized case, it is convenient at first to adopt a coordinate system aligned with the magnetic field so that $\mathbf{B} = B\hat{\mathbf{z}}$ and separately solve for each component of velocity and current density in Ohm's law.

The field-aligned component of Equation 7.15 is

$$E_z + v_{e_x}\cancelto{0}{B_y} - v_{e_y}\cancelto{0}{B_x} = -\frac{m_e\nu_e}{e}v_{e_z}$$

and can be directly solved for the parallel component of velocity and substituted into the z-component of Ohm's law to find that

$$j_{e_z} = \sigma_{\|_e}E_z$$

where the parallel conductivity $\sigma_{\|_e}$ is given by

$$\sigma_{\|_e} = \frac{n_e e^2}{m_e\nu_e} = \frac{\epsilon_0\omega_{pe}^2}{\nu_e}$$

and is identical under our assumptions to the electron conductivity of an unmagnetized plasma. This is to be expected since, if the electric field has a component E_z along the direction of the magnetic field, the motion of charged particles in that direction is not at all affected by the magnetic field. Charged particles flow along the magnetic field as accelerated by the parallel component of $\mathbf{E}$ and impeded only by collisions. Including ion motion as in the unmagnetized case results in

$$j_z = \sigma_{\|}E_z \tag{7.16}$$

where $\sigma_{\|}$ is identical to σ_0 given by Equation 7.13. The parallel current in a magnetized plasma is driven by the parallel component of the electric field with conductivity $\sigma_{\|} = \sigma_0 \approx \sigma_{0_e}$.

The $x-$component of Equation 7.10 is

$$E_x + v_{e_y}B_z - v_{e_z}\cancelto{0}{B_y} = -\frac{m_e\nu_e}{e}v_{e_x}$$

that, following the procedure we have used in the previous cases, can be solved for the velocity to find that

$$v_{e_x} = -\frac{e}{m_e\nu_e}E_x - \frac{eB_z}{m_e\nu_e}v_{e_y}$$

and

$$j_{e_x} = -en_e v_{e_x} = \frac{e^2 n_e}{m_e \nu_e} E_x + \frac{eB_z}{m_e \nu_e} en_e v_{e_y}.$$

Using $\sigma_{\|_e} = \frac{n_e e^2}{m_e \nu_e}$ and $\omega_{ce} = \frac{eB}{m_e}$ this relation can be rewritten as

$$j_{e_x} = \sigma_{\|_e} E_x - \frac{\omega_{ce}}{\nu_e} j_{e_y} \tag{7.17}$$

and in the very same way the $y-$component of the current is

$$j_{e_y} = \sigma_{\|_e} E_y + \frac{\omega_{ce}}{\nu_e} j_{e_x}. \tag{7.18}$$

Notice that the current in one of the perpendicular directions (x or y) depends on the current in the other perpendicular direction. The current densities are coupled and since we wish to express the current $\mathbf{j}_e$ as a conductivity multiplied by the $\mathbf{E}$, it must be that the conductivity is a tensor so that $\mathbf{j} = \tilde{\sigma} \cdot \mathbf{E}$. To see this, let us decouple Equations 7.17 and 7.18 in the usual way of solving two equations for two unknowns. Substituting Equation 7.18 into Equation 7.17 and solving for j_{e_x} and then j_{e_y} yields

$$j_{e_x} = \frac{\nu_e^2}{\nu_e^2 + \omega_{ce}^2} \sigma_{\|_e} E_x - \frac{\nu_e \omega_{ce}}{\nu_e^2 + \omega_{ce}^2} \sigma_{\|_e} E_y$$

and

$$j_{e_y} = \frac{\nu_e^2}{\nu_e^2 + \omega_{ce}^2} \sigma_{\|_e} E_y + \frac{\nu_e \omega_{ce}}{\nu_e^2 + \omega_{ce}^2} \sigma_{\|_e} E_x.$$

If we now make the definitions

$$\sigma_{P_e} = \frac{\nu_e^2}{\nu_e^2 + \omega_{ce}^2} \sigma_{\|_e} \quad \text{and} \quad \sigma_{H_e} = \frac{\nu_e \omega_{ce}}{\nu_e^2 + \omega_{ce}^2} \sigma_{\|_e}$$

the current densities become

$$j_{e_x} = \sigma_{P_e} E_x - \sigma_{H_e} E_y \quad \text{and} \quad j_{e_y} = \sigma_{H_e} E_x + \sigma_{P_e} E_y$$

where σ_{P_e} and σ_{H_e} are the Pedersen and Hall electron conductivites, respectively, associated with the Pedersen and Hall electron currents. The electron current density is then $\mathbf{j}_e = \tilde{\sigma}_e \cdot \mathbf{E}$ where

$$\tilde{\sigma}_e = \begin{pmatrix} \sigma_{P_e} & -\sigma_{H_e} & 0 \\ \sigma_{H_e} & \sigma_{P_e} & 0 \\ 0 & 0 & \sigma_{\|_e} \end{pmatrix}.$$

To gain insight into the current and these conductivities, let us consider the magnitude and direction of current that flows in response to electric fields directed along each of the three axes. First, let us take $\mathbf{E} = E\hat{\mathbf{z}}$ which is parallel to the magnetic field. In this case, the current $\mathbf{j}_e = \tilde{\sigma}_e \cdot E\hat{\mathbf{z}} = \sigma_{\|_e} E\hat{\mathbf{z}}$ is along the magnetic field and has a magnitude determined by the electron parallel conductivity. If $\mathbf{E} = E\hat{\mathbf{x}}$ then $\mathbf{j}_e = \sigma_{P_e} E\hat{\mathbf{x}} + \sigma_{H_e} E\hat{\mathbf{y}}$ and has components along $\mathbf{E}$ with a magnitude determined by the Pedersen conductivity (thus, the Pedersen current) and perpendicular to $\mathbf{E}$ with a magnitude determined by the Hall conductivity (thus, the Hall current). Note that the Pedersen current is parallel to $\mathbf{E}$ but perpendicular to $\mathbf{B}$ while the Hall current is perpendicular to both $\mathbf{E}$ and $\mathbf{B}$. Finally, if $\mathbf{E} = E\hat{\mathbf{y}}$, the electron current is $\mathbf{j}_e = -\sigma_{H_e} E\hat{\mathbf{x}} + \sigma_{P_e} E\hat{\mathbf{y}}$ and we again see that the Pedersen current is parallel to $\mathbf{E}$ but perpendicular to $\mathbf{B}$ while the Hall current is perpendicular to both $\mathbf{E}$ and $\mathbf{B}$.

In this derivation we have assumed the ions are stationary and we have therefore ignored their contribution to the current density. This contribution can be included to find the total current density $\mathbf{j} = \mathbf{j}_e + \mathbf{j}_i$ by performing the same algebra as above but with ions instead of electrons to obtain

$$j_{i_x} = \sum_i \left(\frac{\nu_i^2}{\nu_i^2 + \omega_{ci}^2} \sigma_{\|_i} E_x - \frac{\nu_i \omega_{ci}}{\nu_i^2 + \omega_{ci}^2} \sigma_{\|_i} E_y \right)$$

and

$$j_{i_y} = \sum_i \left(\frac{\nu_i^2}{\nu_i^2 + \omega_{ci}^2} \sigma_{\|_i} E_y + \frac{\nu_i \omega_{ci}}{\nu_i^2 + \omega_{ci}^2} \sigma_{\|_i} E_x \right)$$

where, as before, the summation is over ion species so that

$$\begin{aligned} j_x &= j_{e_x} + j_{i_x} \\ &= \left(\frac{\nu_e^2}{\nu_e^2 + \omega_{ce}^2} \sigma_{\|_e} + \sum_i \frac{\nu_i^2}{\nu_i^2 + \omega_{ci}^2} \sigma_{\|_i} \right) E_x - \\ &\quad \left(\frac{\nu_e \omega_{ce}}{\nu_e^2 + \omega_{ce}^2} \sigma_{\|_e} + \sum_i \frac{\nu_i \omega_{ci}}{\nu_i^2 + \omega_{ci}^2} \sigma_{\|_i} \right) E_y \end{aligned}$$

and

$$\begin{aligned} j_y &= j_{e_y} + j_{i_y} \\ &= \left(\frac{\nu_e^2}{\nu_e^2 + \omega_{ce}^2} \sigma_{\|_e} + \sum_i \frac{\nu_i^2}{\nu_i^2 + \omega_{ci}^2} \sigma_{\|_i} \right) E_y + \\ & \left(\frac{\nu_e \omega_{ce}}{\nu_e^2 + \omega_{ce}^2} \sigma_{\|_e} + \frac{\nu_i \omega_{ci}}{\nu_i^2 + \omega_{ci}^2} \sigma_{\|_i} \right) E_x. \end{aligned}$$

Defining Pedersen and Hall conductivities as

$$\sigma_P = \frac{\nu_e^2}{\nu_e^2 + \omega_{ce}^2} \sigma_{\|_e} + \sum_i \frac{\nu_i^2}{\nu_i^2 + \omega_{ci}^2} \sigma_{\|_i} \tag{7.19}$$

and

$$\sigma_H = \frac{\nu_e \omega_{ce}}{\nu_e^2 + \omega_{ce}^2} \sigma_{\|_e} - \sum_i \frac{\nu_i \omega_{ci}}{\nu_i^2 + \omega_{ci}^2} \sigma_{\|_i} \tag{7.20}$$

the total current density may be written as

$$\mathbf{j} = \tilde{\sigma} \cdot \mathbf{E} \tag{7.21}$$

where the conductivity tensor of a magnetized plasma is

$$\tilde{\sigma} = \begin{pmatrix} \sigma_P & -\sigma_H & 0 \\ \sigma_H & \sigma_P & 0 \\ 0 & 0 & \sigma_\| \end{pmatrix}. \tag{7.22}$$

As with the electron currents, the Pedersen conductivity is associated with the Pedersen current that flows in a direction parallel to **E** but perpendicular to **B** and the Hall conductivity is accosicated with the Hall current that flows in a direction perpendicular to both **E** and **B**.

Given this understanding of the parallel, Pedersen and Hall conductivities and currents, we can easily generalize Equation 7.21 for an arbitrary magnetic field **B** not necessarily aligned with the $z-$axis. The result is

$$\mathbf{j} = \sigma_\| \mathbf{E}_\| + \sigma_P \mathbf{E}_\perp - \sigma_H \frac{\mathbf{E}_\perp \times \mathbf{B}}{B}$$

where $\mathbf{E}_\|$ is the component of **E** along **B** given by $\mathbf{E}_\| = (\mathbf{E} \cdot \mathbf{B})\mathbf{B}/B^2$ and $\mathbf{E}_\perp$ is the component of **E** perpendicular to **B** given by $\mathbf{E}_\perp = \mathbf{E} \times \mathbf{B}/B$. Parallel currents flow along the direction of $\mathbf{E}_\|$, Pedersen currents flow along the direction of $\mathbf{E}_\perp$ and Hall currents flow along the direction of $-\mathbf{E}_\perp \times \mathbf{B}$.

7.8.4 Ionospheric Conductivity Profile

Let us now investigate the altitude ranges in Earth's ionosphere over which these currents may be expected to flow. To do this, notice first that the Pedersen and Hall conductivities are proportional to the parallel conductivity that is a function of electron density, electron and ion masses, and the collision frequencies. Figure 7.19 shows the contributions to the Pedersen and Hall conductivities from a single species (either an electrons or an ion species) normalized by the corresponding parallel conductivity. The horizontal axis shows the ratio of the gyrofrequency to the collision frequency so that for $\omega_c/\nu_c >> 1$ the particle experiences many gyrations per collision while for $\omega_c/\nu_c << 1$ the particle experiences many collisions per gyration. In agreement with our initial qualitative discussion, we see that when there are many gyrations per collision, the Hall conductivity dominates so that the current flows in the $q\mathbf{E} \times \mathbf{B}$ direction while the Pedersen conductivity dominates when the particle experiences many collisions per gyration, in which case the current flows in the direction of $\mathbf{E}_\perp$.

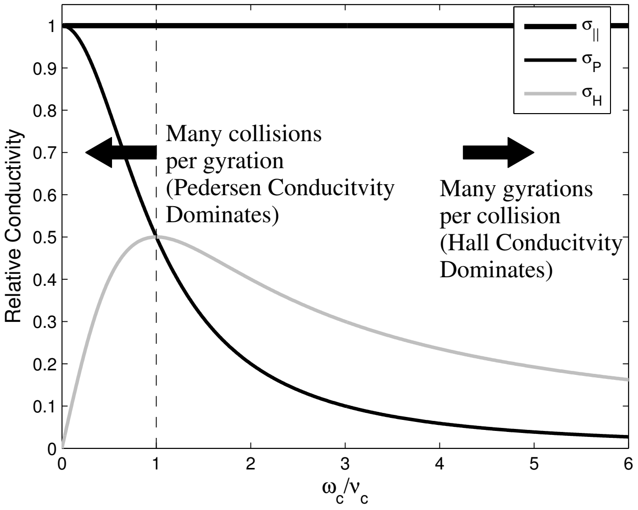

Figure 7.19: Pedersen and Hall conductivities of a single species relative to the corresponding parallel conductivity.

To obtain conductivity profiles, we must first determine the cyclotron, plasma and collision frequency profiles. Electron and ion cyclotron profiles can be obtained for a species s using $\omega_{cs} = |q_s|B/m_s$ where q_s is the species charge (assumed to be equal to the electron charge for both ions and electrons), B is the magnitude of the geomagnetic field that can be obtained from the DGRF/IGRF magnetic field model, and m_s is the species mass. For ions, m_s can be taken to be the average ion mass obtained from the IRI model. Plasma frequency profiles can be obtained in a similar way. The collision frequencies are more complicated.

The ion collision frequency is the frequency with which ions collide with neutrals and this, as derived by Chapman [1956] is given by

$$\nu_i = \nu_{in} = (2.6 \times 10^{-9}(n_n + n_i)A^{-1/2} \quad \mathrm{s}^{-1}$$

where n_n and n_i are the neutral and ion densities per cubic centimeter and A is the mean molecular weight of the neutrals and ions. Following Kelley [1989, p.460], we take these two weights to be equal to the mean molecular weight and obtain the values from the MSIS atmospheric model. Coulomb collisions between ions and electrons are not included in this expression for ν_i since they impart a neglibible change to the massive ion's momentum. Coulomb collisions do, however, significantly alter the electron's momentum and their effect must be included in the electron collision frequency that becomes $\nu_e = \nu_{en} + \nu_{ei}$. Where ν_{en} is the electron/neutral collision frequency and ν_{ei} is the Coulomb collision frequency. Electron/neutral collisions will dominate at lower altitudes where the neutral density is lowest and electron density is highest but Coulomb collision will make a significant contribution in the $F-$region where neutral densities are relativily low and electron densities are relatively high. Nicolet [1953] gives the collision frequencies as

$$\nu_e = \nu_{en} + \nu_{ei} = (5.4 \times 10^{-10})n_n T_e^{1/2} + [34 + 4.18 \ln(T_e^3/n_e)]n_e T_e^{-3/2} \quad \mathrm{s}^{-1}$$

where T_e is the electron temperature which can be obtained from the IRI model.

With profiles of the plasma, gyro, and colision frequencies in hand, the parallel, Pedersen and Hall conductivities given in Equations 7.13, 7.19 and 7.20 can be determined. Figure 7.20a) shows the electron and ion gyro and collision frequency profiles obtained using the MSIS, IRI and DGRF/IGRF models for March 15, 2002 at 00UT. Note that three distinct regions roughly corresponding to the $D-$, $E-$ and $F-$regions may be identified as deliniated by the two horizontal lines. The lower horizontal line indicates the altitude for which the electron cyclotron frequency equals the electron collision frequency and the upper horizontal line indicates the altitude for which the ion cyclotron frequency equals the ion collision frequency.

In region 1 below about 90 km (roughly corresponding to the $D-$region), the electron and ion collision frequencies both exceed the corresponding gyrofrequencies and we expect all conductivities in this region to be low due to the combined effect of the high collision frequencies and low $D-$region electron densities. In region 1, there are not many charged particles available to carry current and the motion of the these particles is severely impeded by collisions.

In region 3 (roughly corresponding to the $F-$region), the neutral density is low enough that both the electron and ion collision frequencies are smaller than the respective gyrofrequencies and both electrons and ions gyrate many times between collisions. We therefore expect that the Hall conductivity will be low in this region since both electrons and ions will be predominantly $\mathbf{E}\times\mathbf{B}$ drifting together. The parallel conductivity in region 3 is expected to be large since electron and ion densities are highest in the $F-$region (*i.e.,* there are many charged particles to carry current) and the collision frequencies are low due to the small neutral densities. Any perpendicular currents in this region should be Pedersen currents carried by ions since $\omega_{ci}/\nu_i << \omega_{ce}/\nu_e$ here. That is, since the ions gyrate fewer times per collision than do electrons, the ion contribution to the Pedersen conductivity should exceed the electron contribution.

In region 2 (roughly corresponding to the $E-$region, the electron cyclotron frequency exceeds the collision frequency and we therfore expect that electrons will predominantly move in the direction of $\mathbf{E}\times\mathbf{B}$ and significantly contribute to the Hall conductivity. Ions on the other hand, suffer many collisions per gyration in this region since the collision frequency exceeds the gyrofrequency and we expect that they will contribute negligibly to the Hall but significantly to the Pedersen conductivities. In region 2 then, we expect electron Hall currents and ion Pedersen currents. The Hall conductivity typically exceeds the Pedersen conductivity and so important electrojet currents that flow in the $E-$region are typically Hall currents carried by electrons.

In agreement with these expectations, Figure 7.20b) shows parallel, Pedersen and Hall conductivity profiles. The main plot shows the combined conductivities given by Equations 7.13, 7.19 and 7.20 while the inset plots show electron and ion contributions to each conductivity.

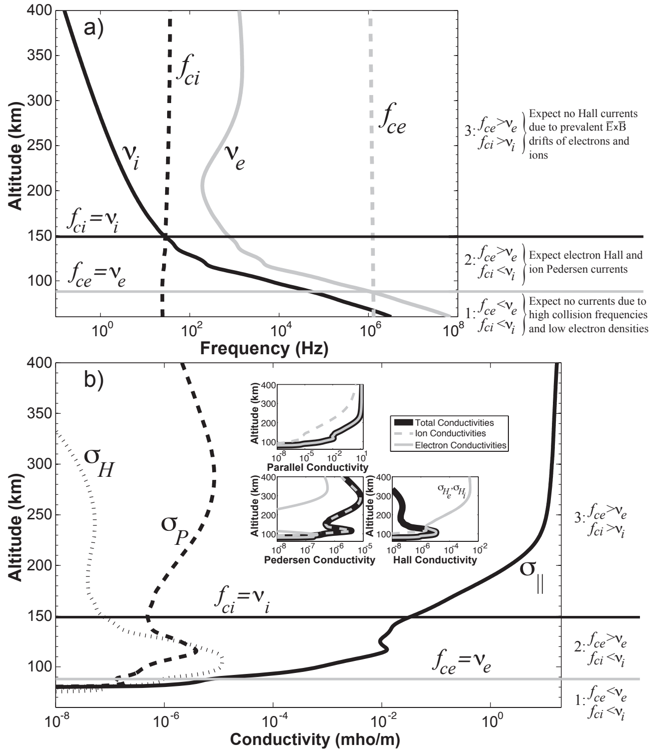

Figure 7.20: a) Electron and Ion collision- and gyro-frequencies. b) Parallel, Pedersen and Hall conductivity profiles. The inset plots show contributions to the total conductivities from electrons and ions. Panels a) and b) represent nighttime solar-max conditions.

7.9 Ionospheric Convection

In this section we consider the ionospheric counterpart to magnetospheric convection that, as we saw in §?? for the case of southward IMF, is the process wherein magnetic flux from the magnetopause is transported, after merging with the IMF, over the polar cap into the magnetotail. This depletion of dayside flux is replenished by the sunward flow of plasma (with its frozen-in magnetic field) from the magnetotail around the magnetosphere at a rate that is dependent on the existing conditions and state of the magnetosphere-ionosphere system. The high-latitude ionosphere acts as a sort of film upon which the global-scale process of magnetospheric convection is imaged. Observations of ionospheric convection are therefore of significant importance since they can be used to determine the strength of magnetospheric conection and the amount of solar wind energy and momentum being coupled into the magnetosphere.

Figure 7.21 illustrates the high-latitude electric fields that are responsible for driving ionospheric convection. In Figure 7.21a, the Sun is located into the page and the solar wind flows out of the page across the open polar cap magnetic field lines. As it flows past these open field lines, an electric field $\mathbf{E}_{pc} = -\mathbf{v}_{sw} \times \mathbf{B}$ is present as required by the generalized Ohm's Law (Equation **??**) to maintain a finite current density in the highly conductive solar wind. This electric field maps down the essentially equipotential polar cap field lines into the ionosphere where the $\mathbf{E}\times\mathbf{B}$ drift drives the ionospheric plasma in the antisunward direction. This antisunward drift of ionospheric plasma is the ionospheric image of that portion of magnetospheric convection resulting in the movement of magnetopause field lines from the dayside, over the polar cap and into the magnetotail. The strength of this polar cap electric field $\mathbf{E}_{pc}$ and its associated potential (which often exceeds 60 kV) is therefore related to the strength of magnetospheric convection and thus to the rate at which magnetopause field lines are being loaded into the tail.

The viscous interaction of the solar wind with the magnetopause makes an An alternate and self-consistent view of this process can be gained by considering the region 1 currents shown in Figure 7.21b. Recall from §4.4.4 the region 1 currents that flow near the poleward edge of the auroral oval and close part of the dayside magnetopause current by flowing into the ionosphere in the morning sector and out of the ionosphere in the evening sector. This flow of current results in regions of excess positive and negative charge as indicted by the plus and minus signs in Figure 7.21b. The resulting polar cap electric field is the same field $\mathbf{E}_{pc}$ discussed in the previous paragraph.

The region 2 currents that partially close the ring current flow near the equatorward edge of the auroral oval and are, in general, of opposite polarity to the region 1 currents. These currents also result in regions of excess positive and negative charges as indicated in the same figure and result in an auroral zone electric field that, in the dawn and dusk sectors, is directed opposite to polar cap electric field. The $\mathbf{E} \times \mathbf{B}$ drift driven by this auroral zone electric field is in the sunward direction and is the ionospheric image of the magnetospheric sunward flow of plasma from the tail to the dayside. A typical 2-cell pattern of ionospheric convection typically results (during southward IMF) as indicted in Figure 7.21b. The region of highly sheared

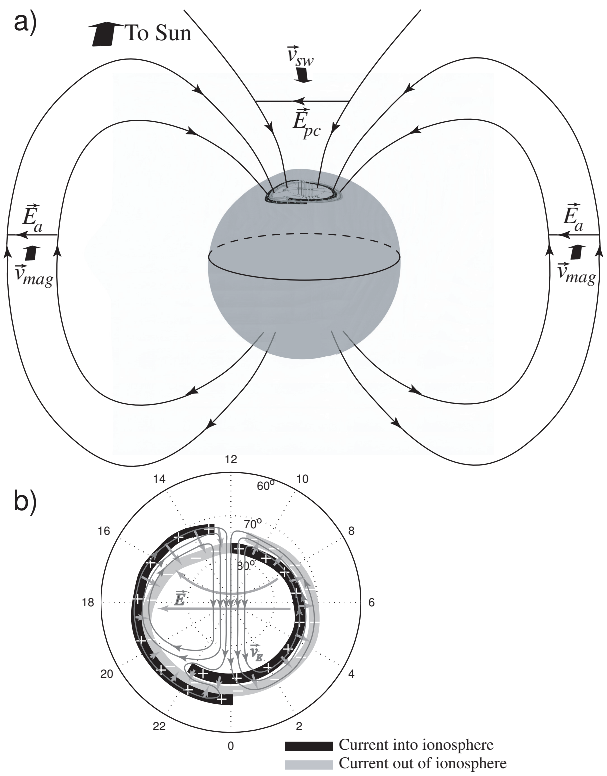

Figure 7.21: a) Three-dimensional representation of the high-latitude electric fields that drive ionospheric convection. b) A representation of the high-latitude Birkeland currents, electric fields and resulting plasma flows.

flow in the premidnight sector is known as the Harang discontinuity that results from the overlapping of the region 1 and 2 currents as shown.

Returning to a Figure 7.21a, an alternate and self-consistent view of the auroral zone electric field can be seen. As the process of magnetospheric convection returns tail plasma to the dayside, it flows across the closed auroral zone magnetic field lines where the auroral zone electric field $\mathbf{E}_a = -\mathbf{v}_{mag} \times \mathbf{B}$ is present, again as required by the generalized Ohm's Law. This electric field maps along the equipotential magnetic field lines and presents itself in the auroral zone as the electric field shown in Figure 7.21b.

additional, small contribution to the polar cap potential and the entire ionospheric convection pattern is observed to vary significantly on the time scale of minutes in response to changing conditions including the solar wind density and velocity, the orientation of the IMF and the state of the magnetosphere. The two-cell pattern described above is most common during southward IMF but distorted patterns are often observed, as are, for example, four-cell patterns during periods of northward IMF.

Appendices

Appendix A

Electrodynamics Review

A.1 Some fundamental constants

https://www.youtube.com/watch?v=Fo3DudOzV4k

Physical constants appear frequently in electromagnetic theory:

$$\begin{aligned}
\mu_0 &= 4\pi \times 10^{-7} \quad \text{H/m} \quad \text{(permeability of free space)} \\
\epsilon_0 &\approx 8.854 \times 10^{-12} \quad \text{F/m} \quad \text{(permittivity of free space)} \\
c &= \frac{1}{\mu_0\epsilon_0} \approx 3 \times 10^{8} \quad \text{m/s} \quad \text{(speed of light in free space)} \\
m_e &\approx 9.109 \times 10^{-31} \quad \text{kg} \quad \text{(electron mass)} \\
m_p &\approx 1.672 \times 10^{-27} \quad \text{kg} \quad \text{(proton mass)} \\
e &\approx 1.6 \times 10^{-19} \quad \text{C} \quad \text{(elementary charge)} \\
q_e &= -e \quad \text{(electron charge)} \\
q_p &= e \quad \text{(proton charge)}
\end{aligned}$$

A.2 Some useful results from basic electricity and magnetism

Electromagnetic theory is structured as a field theory in which electromagnetic fields, produced by *source* charges and currents, exert forces on other nearby *test* charges and currents. The force law which encodes the interaction of the fields with test charges is the Lorentz force law:

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \tag{A.1}$$

where $\mathbf{F}$ is the force (both electric and magnetic) experienced by test charge q moving at velocity $\mathbf{v}$ through a region of space containing electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$.

Electromagnetic fields are generated by source charges and currents. The complete set of equations describing the connection between sources and fields is the Maxwell equations, reviewed below in a form similar to that used by most introductory textbooks.

Gauss's Law for E

Gauss's law is a quantitative statement linking the electric fields to one of their sources - electric charges:

$$\oint_{\partial V} \mathbf{E} \cdot d\mathbf{a} = \frac{q_{enc}}{\epsilon_0}. \tag{A.2}$$

In this equation ∂V represents a closed surface on the boundary of volume V and q_{enc} is the charge enclosed inside that surface. The enclosed charge is often represented in terms of a charge density (ρ_c, units C/m^3) by integration of this density over the enclosed volume V so that

$$q_{enc} = \int_V \rho_c dV. \tag{A.3}$$

Gauss's Law for B

A type of Gauss's law also exists for the magnetic field:

$$\oint_{\partial V} \mathbf{B} \cdot d\mathbf{a} = 0. \tag{A.4}$$

This formula indicates that the total amount of magnetic field piercing any closed surface ∂V is always zero. This gemoetric interpretation means that magnetic field lines always form closed loops, which, in turn, implies there are no point sources of magnetic field.[1]

[1] The point source of magnetic field, or *magnetic monopole*, has been predicted by theory but has never been observed and, for this reason, the RHS of Gauss's law for $\mathbf{B}$ equals zero. If magnetic monopoles are ever observed, Gauss's law for $\mathbf{B}$ will become $\oint_{\partial V} \mathbf{B} \cdot d\mathbf{a} = \mu_0 q_{mag,enc}$ where $q_{mag,enc}$ is the magnetic charge enclosed by volume V.

Faraday's Law

Point charges are not the only source of electric field. Time-varying magnetic fields also produce electric fields, a behavior accounted for by Faraday's law of induction:

$$\oint_{\partial S} \mathbf{E} \cdot d\boldsymbol{\ell} = -\frac{d\Phi_B}{dt}. \tag{A.5}$$

The contour of integration on the left hand side of this equation is around the closed loop ∂S which forms the boundary of surface S. Φ_B is the magnetic flux through surface S, which must be an open surface (one that does not divide space into two disconnected regions) for this equation to be non-degenerate. In terms of the magnetic field, this flux is defined by the integral

$$\Phi_B = \int_S \mathbf{B} \cdot d\mathbf{a}. \tag{A.6}$$

Ampere's Law

The final Maxwell equation is Ampere's law, which outlines the connection between electric currents and the magnetic fields that they produce:

$$\oint_{\partial S} \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 I_{enc}. \tag{A.7}$$

The current enclosed by the loop ∂S comprises two contributions: the conduction current I_c formed by moving charges and the displacement current I_d which is related to *local* electric field fluctuations (produced by *remote* charge density fluctuations). The (enclosed) conduction current is often written in terms of the current density $\mathbf{J}$ (units of A/m^2) as

$$I_c = \int_S \mathbf{J} \cdot d\mathbf{a}. \tag{A.8}$$

The displacement current belies another source of magnetic field other than moving charges - namely time-varying electric fields, i.e.:

$$I_d = \epsilon_0 \frac{d\Phi_E}{dt}. \tag{A.9}$$

Φ_E is the electric flux defined in a manner analogous to magnetic flux (equation A.6) above:

$$\Phi_E = \int_S \mathbf{E} \cdot d\mathbf{a}. \tag{A.10}$$

A.3 Integral forms of Maxwell's equations

The four Maxwell equations (the two Gauss's laws, Faraday's law, and Ampere's law) outlined above provide a means to calculate the electric and magnetic fields from given charge and current distributions. These equations may be written out in full integral form by substituting the field expressions for all fluxes and the integral expressions for enclosed current and charge:

$$\oint_{\partial V} \epsilon_0 \mathbf{E} \cdot d\mathbf{a} = \int_V \rho_c dV \tag{A.11}$$

$$\oint_{\partial S} \mathbf{E} \cdot d\ell = -\frac{d}{dt}\left[\int_S \mathbf{B} \cdot d\mathbf{a}\right] \tag{A.12}$$

$$\oint_{\partial V} \mathbf{B} \cdot d\mathbf{a} = 0 \tag{A.13}$$

$$\oint_{\partial S} \frac{\mathbf{B}}{\mu_0} \cdot d\ell = \int_S \mathbf{J} \cdot d\mathbf{a} + \frac{d}{dt}\left[\int_S \epsilon_0 \mathbf{E} \cdot d\mathbf{a}\right] \tag{A.14}$$

These forms of Maxwell's equations, known as the integral forms, are usually not the most convenient way of calculating electromagnetic fields. Instead they are a useful means for introducing electromagnetic theory and providing a conceptual tool for understanding the physical meaning behind Maxwell's equations. However, certain, very specific, types of problems (those involving discontinuities or a high degree of symmetry) may be most easily analyzed through these forms (cf. Griffiths, Cheng). All of the familiar results (e.g. Coulomb's Law, Biot-Savart Law, etc.) from electromagnetic theory may be derived from these equations or equivalent forms. See Cheng's book for a particularly well-organized presentation of this idea and derivations linking most electromagnetic formulae to Maxwell's equations.

A.4 Differential forms of Maxwell's equations

Two results from vector analysis, which are useful for manipulating the Maxwell equations above, are the Divergence theorem and Stokes' theorem. These theorems may be written, for a vector field **A**, as:

$$\oint_{\partial V} \mathbf{A} \cdot d\mathbf{a} = \int_V (\nabla \cdot \mathbf{A})\, dV \tag{A.15}$$

$$\oint_{\partial S} \mathbf{A} \cdot d\ell = \int_S (\nabla \times \mathbf{A}) \cdot d\mathbf{a}, \tag{A.16}$$

respectively. In these equations the notation ∂V indicates a boundary surface for volume V and ∂S indicates a bounding contour for surface S. Surface ∂V and contour ∂S are both *closed* in the sense that ∂V separates space into unconnected 'inside' and 'outside' volumes and ∂S is a closed loop. The left hand sides of these theorems bear obvious resemblance to the left hand sides of the integral Maxwell equations. These theorems can be used to convert the integral forms of Maxwell's equations to the, generally more useful, differential forms.

Conversion of the integral forms of Maxwell's equations to differential forms is illustrated with Gauss's Law and Faraday's law. The remaining conversions follow from identical logic and mathematical steps. The left hand side of Gauss's law may be rewritten using the divergence theorem, which yields:

$$\int_V \nabla \cdot (\epsilon_0 \mathbf{E})\, dV = \int_V \rho_c dV. \tag{A.17}$$

Consolidating all quantities onto the left hand side of the equation, we find:

$$\int_V \left[\nabla \cdot (\epsilon_0 \mathbf{E}) - \rho_c\right] dV = 0. \tag{A.18}$$

Note that the two, previously distinct, integrals have been combined since they were over the same volume V. It may not have been obvious in the previous discussion, but V has the additional property that it is arbitrary in the sense that the integral Maxwell equation is valid for any choice of V. Because of this, for equation A.18 to be valid for all possible integration volumes V, the left hand side integrand must be identically zero, hence

$$\nabla \cdot (\epsilon_0 \mathbf{E}) - \rho_c = 0. \tag{A.19}$$

This result is the differential form of Gauss's law. As with the rest of the Maxwell equations it is most often written with the causative field source (charge, in this case) on the right hand side and the resulting field on the left hand side:

$$\nabla \cdot (\epsilon_0 \mathbf{E}) = \rho_c. \tag{A.20}$$

Conversion of Faraday's law, proceeds by first transforming the left hand side of equation A.12 with Stokes' theorem:

$$\oint_S (\nabla \times \mathbf{E}) \cdot d\mathbf{a} = -\frac{d}{dt}\left[\int_S \mathbf{B} \cdot d\mathbf{a}\right]. \tag{A.21}$$

Similar to the previous derivation, we will attempt to combine the two integrals, which are over the same surface. In order to do this we must first pull the time derivative through the surface integral operation. At first glance, this appears straightforward because time (which we differentiate with respect to) and space (which we integrate over) are independent. However, one must take care since the surface integral limits could, in principle, change with time (as with a surface that is 'elastic' and freely distorts with time).[2] Thus, in order to bring the time derivative under the integral without introducing an additional complicating term, we must stipulate that the surface S does not change with time. Bringing the derivative under the integral, then causes it to directly operate on the magnetic field, which is a function of both space and time. Hence, the time derivative may be interpreted as a partial derivative with respect to time (provided that S also does not move). This renders the following form of Faraday's law:

$$\oint_S \left[\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} \right] \cdot d\mathbf{a}. \tag{A.22}$$

As with Gauss's law, this integral equation is valid for all choices of the surface S. For the above relation to be true, it must be the case that the integrand is zero. Hence, the differential form of Faraday's law:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \tag{A.23}$$

Transforming the remaining equations (generalized Ampere's law and Gauss's law for magnetic fields) involves similar manipulations and ideas to the derivations for Gauss's law and Faraday's law. The resulting set of differential equations (the differential forms of Maxwell's equations) is listed below:

$$\begin{aligned}
\nabla \cdot (\epsilon_0 \mathbf{E}) &= \rho_c & \text{(A.24)}\\
\nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} & \text{(A.25)}\\
\nabla \cdot \mathbf{B} &= 0 & \text{(A.26)}\\
\nabla \times \left(\frac{\mathbf{B}}{\mu_0} \right) &= \mathbf{J} + \frac{\partial}{\partial t} (\epsilon_0 \mathbf{E}) & \text{(A.27)}
\end{aligned}$$

These are the primary electromagnetic laws (along with the Lorentz force) used in this text.

[2] See the discussion surrounding footnote 37 on p. 85.

A.5 Potentials in electromagnetic theory

Maxwell's equations may be reformulated in terms of the scalar potential Φ and the vector potential $\mathbf{A}$. For many (perhaps most) problems it is easier to solve the potential forms of these equations instead of the set A.24 - A.27. That this reformulation is possible may be seen by combining Gauss's law for the magnetic field with the vector identity $\nabla \cdot (\nabla \times \mathbf{A}) = 0$ (valid for any vector field $\mathbf{A}$). These two facts together imply that we may to introduce a magnetic vector potential $\mathbf{A}$ which is related to the magnetic field through the relation:

$$\mathbf{B} = \nabla \times \mathbf{A}. \tag{A.28}$$

To invoke this definition does not violate Gauss's law (for mag. field), and so, it is permitted.

The connection between the electric field and the potentials Φ and $\mathbf{A}$ is derived from Faraday's law by substituting in the magnetic vector potential.

$$\nabla \times \mathbf{E} = -\frac{\partial}{\partial t} (\nabla \times \mathbf{A}) \tag{A.29}$$

Interchanging the order of the time and space derivatives and combining all terms under a single curl operator on the left hand side gives:

$$\nabla \times \left(\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} \right) = 0. \tag{A.30}$$

This equation may be combined with the vector identity $\nabla \times \nabla \Phi = 0$ (valid for all scalar fields Φ). Together these two pieces of information imply that we may introduce a scalar potential Φ related to the electric field and vector potential by:

$$\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} = \pm \nabla \Phi. \tag{A.31}$$

This mathematical relation is valid for either a positive or negative sign on the right-hand side. The negative sign is chosen for the potential convention, which, it turns out, preserves the interpretation of potential at some field point as being work per unit charge required to bring a test charge from very far away to that point. Hence we find the electric field is:

$$\mathbf{E} = -\nabla \Phi - \frac{\partial \mathbf{A}}{\partial t}. \tag{A.32}$$

The development above has shown consistency of our potential definitions, equations A.28 and A.32, with Gauss's law for mag. fields and Faraday's law. It does not provide a means to solve for the potentials. Fortunately, when equations A.28 and A.32 are substituted for the electric and magnetic fields in the remaining Maxwell equations, Gauss's Law and Ampere's law, a set of equations which fully specify the potentials, and hence the fields, results. This process involves quite a bit of detailed algebra, along with some assumptions about $\nabla \cdot \mathbf{A}$ (which one can actually specify arbitrarily - see Griffiths text). In the end it produces a set of inhomogeneous wave equations. These, however are not directly used in this book, so we relegate further discussion on this point to a full course in electrodynamics.

A.6 Simplified forms of the Maxwell equations

One of the most common simplifications of the Maxwell's equations is the so called electrostatic approximation. This approximation is obtained by neglecting fluctuations in the magnetic field, i.e. $\partial \mathbf{B}/\partial t = 0$. From this assumption and Faraday's law, it is seen that the electric field is curl free; hence it is the gradient of some scalar potential function:

$$\mathbf{E} = -\nabla \Phi. \tag{A.33}$$

The inverted form of this equation (Φ in terms of $\mathbf{E}$, assuming $\Phi(\mathbf{r} \to \infty) = 0$) can be derived from the fundamental theorem of gradients (see Griffiths text) and is often useful:

$$\Phi(\mathbf{r}) = -\int_{\infty}^{\mathbf{r}} \mathbf{E} \cdot d\boldsymbol{\ell}. \tag{A.34}$$

If we combine equation A.33 with Gauss's law, an equation specifying the electrostatic potential results:

$$\begin{aligned} -\nabla \cdot (\epsilon_0 (\nabla \Phi)) &= \rho_c \\ \nabla^2 \Phi &= -\frac{\rho_c}{\epsilon_0}. \end{aligned} \tag{A.35}$$

This equation is known as the Poisson equation. The solution for finite distributions of source charge in unbounded space with the boundary conditions

$\Phi(\mathbf{r} \to \infty) = 0$ can be calculated by evaluating the particular integral:

$$\Phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho_c(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} dV', \tag{A.36}$$

where the integral is assumed to be over some volume containing all of the charge distribution of interest $\rho_c(\mathbf{r}')$. In the equations describing the potential solutions (e.g. equation A.36), the primed variables represent source locations and the unprimed variable represent field points.

Another approximate form of Maxwell's equations, very often used in magnetospheric and solar physics, is the quasistatic form. This simplification results from neglecting the displacement current from the generalized Ampere's law (equation A.27):

$$\nabla \times \left(\frac{\mathbf{B}}{\mu_0} \right) = \mathbf{J}. \tag{A.37}$$

The system of Maxwell's equations with this simplification can be reduced into one equation if we invoke the vector potential and substitute it into equation A.37:

$$\nabla \times (\nabla \times \mathbf{A}) = \mu_0 \mathbf{J}. \tag{A.38}$$

Invoking a vector identity to expand the double curl operation yields

$$\nabla (\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} = \mu_0 \mathbf{J}. \tag{A.39}$$

Finally we take $\nabla \cdot \mathbf{A} = 0$ (gauge freedom, see Griffiths book for details), which reduces the equation to a vector Poisson equation:

$$\nabla^2 \mathbf{A} = -\mu_0 \mathbf{J}. \tag{A.40}$$

By analogy with the electrostatic Poisson equation, we may simply write down the integral solution (compare with equation A.36):

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} dV'. \tag{A.41}$$

Note, again, that this type of solution is valid for finite current distributions and $\mathbf{A}(\mathbf{r} \to \infty) = 0$.

A.7 Dipole approximations

The integral solutions in equations A.36 and A.41 for vector potentials find wide use in solving static problems. In general, they are very difficult to evaluate exactly, except in circumstances of (perhaps unreasonably) simple geometry. Fortunately, there exist systematic simplifications for these integrals based off of series expansions of the inverse distance $|\mathbf{r}-\mathbf{r}'|^{-1}$ part of the integrand. The most commonly invoked expansion is the so-called multipole expansion in which the contributions to the total potential are represented as a superposition of monopole, dipole, quadrupole, etc. terms. The utility of this approach is that, if the field point $\mathbf{r}$ is somewhat far from the source distribution of charge or current which creates the potential, then the potential near that field point can be represented accurately with only a few series terms. More precisely, the higher order contributions (quadrupole, etc.) to the potentials at large distances from their sources are 'usually' negligible. Here we present, without derivation, the results for the potential of a pure electric or magnetic dipole for later use (e.g. describing the Earth's magnetic field in the inner magnetosphere):

$$\Phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\frac{\mathbf{p}\cdot(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} \tag{A.42}$$

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\frac{\mathbf{m}\times(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}. \tag{A.43}$$

The parameters $\mathbf{p}$ and $\mathbf{m}$ are the electric and magnetic dipole moments, respectively, which have fixed values for a particular charge/current distrubtion. These formulas are also good approximations for non-ideal sources, provided that we are 'far enough' away from the source.

A.8 Wave solutions and conventions

The full set of Maxwell's equations (i.e. that including the displacement current) admit wave solutions, and, indeed, wave solutions, in general, are extraordinarily common in other branches of physics (e.g. hydrodynamics, solid mechanics, etc.). The easiest way to prove the existence of electromagnetic waves from Maxwell's equations is to neglect the field sources and examine the interplay between induction as a source of the electric field and the displacement current as a source of magnetic field. Omitting charge and current density from the Maxwell equations leaves:

$$\begin{aligned} \nabla \cdot (\epsilon_0 \mathbf{E}) &= 0 && \text{(A.44)} \\ \nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} && \text{(A.45)} \\ \nabla \cdot \mathbf{B} &= 0 && \text{(A.46)} \\ \nabla \times \left(\frac{\mathbf{B}}{\mu_0}\right) &= \frac{\partial}{\partial t}(\epsilon_0 \mathbf{E}). && \text{(A.47)} \end{aligned}$$

If we take the curl of Faraday's law from this set we can generate a recognizable wave equation in just a few steps:

$$\begin{aligned} \nabla \times (\nabla \times \mathbf{E}) &= -\nabla \times \left(\frac{\partial \mathbf{B}}{\partial t}\right) \\ &= -\frac{\partial}{\partial t}(\nabla \times \mathbf{B}). \end{aligned} \quad \text{(A.48)}$$

Next, Ampere's law is used to rewrite the $\nabla \times \mathbf{B}$ terms in the far right expression in the above statement:

$$\begin{aligned} \nabla \times (\nabla \times \mathbf{E}) &= -\frac{\partial}{\partial t}\left(\mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}\right) \\ &= -\frac{1}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2}. \end{aligned} \quad \text{(A.49)}$$

A vector identity may be used to simplify the double curl term on the left hand side of the above expression, namely:

$$\nabla \times (\nabla \times \mathbf{E}) = \nabla (\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}. \quad \text{(A.50)}$$

Combining this identity with equation A.49 gives

$$\nabla (\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E} = -\frac{1}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2}. \quad \text{(A.51)}$$

Invoking Gauss's law (recall that we have assumed zero charge density) allows us to remove the $\nabla \cdot \mathbf{E}$ term. Rearranging the remainder then gives a homogeneous vector wave eqauation:

$$\nabla^2 \mathbf{E} - \frac{1}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2} = 0. \quad \text{(A.52)}$$

An identical equation for the magnetic field may be derived by apply the same procedure above to the source-free Ampere's law. Note that these equations do not tell us about how the waves were generated (the radiation process), merely how they behave as they propagate away from their sources.

A completely general treatment of the process of electromagnetic radiation and wave solutions to Maxwell's equations is well beyond the scope of this course. However, it is useful to point out a certain type of solution that satisfies equation A.52 - the uniform plane wave, mathematically represented in the following form:

$$\mathbf{E}(\mathbf{r},t) = \tilde{\mathbf{E}}e^{i\mathbf{k}\cdot\mathbf{r}-i\omega t}. \tag{A.53}$$

In this formula, $\tilde{\mathbf{E}}$ is a complex vector constant, $\mathbf{k}$ is the wavenumber, $i = \sqrt{-1}$, and ω is the angular frequency. For our purposes, when discussing this type of wave, all quantities in equations A.53 and A.52 (derivative and position vectors) should be represented in Cartesian coordinates, e.g. $\mathbf{k}\cdot\mathbf{r} = k_x x + k_y y + k_z z$. The usefulness of representing wave solutions in the form of equation A.53 is that they can be almost easily differentiated, i.e.

$$\begin{aligned} \frac{\partial \mathbf{E}}{\partial t} &= -i\omega\mathbf{E} \\ \nabla\cdot\mathbf{E} &= i\mathbf{k}\cdot\mathbf{E} \\ \nabla\times\mathbf{E} &= i\mathbf{k}\times\mathbf{E} \\ \nabla\Phi &= i\mathbf{k}\Phi, \end{aligned} \tag{A.54}$$

when $\mathbf{E}$ and Φ are plane waves. These properties are helpful when we look for characteristics of wave solutions in other systems of equations. Specifically, a useful approach to systems of equations is often to guess that plane wave solutions exist and then to substitute in the general form of the plane waves (equation A.53) to this system to derive the properties of those waves, e.g. wavenumber frequency relation, phase and group speeds, etc. This process is equivalent to a Fourier decomposition of the solutions to whatever system of equations is being analyzed.

A.9 Electromagnetic field energy and momentum

One of the key features of electromagnetic fields is that they are able to store and transport both energy and momentum. This result comes from electromagnetic conservation laws derived from Maxwell's equations and the Lorentz force law. The algebra is involved (see Griffiths), and we merely document the results here for later use:

$$\begin{aligned} \frac{\partial w}{\partial t} + \nabla \cdot \mathbf{S} &= -\mathbf{J} \cdot \mathbf{E} && \text{(A.55)} \\ \frac{\partial \mathbf{g}}{\partial t} - \nabla \cdot \mathsf{T} &= -\left(\rho_c \mathbf{E} + \mathbf{J} \times \mathbf{B}\right). && \text{(A.56)} \end{aligned}$$

Equation A.55 is known as the Poynting theorem and is statement of conservation of energy for the electromagnetic fields. Equation A.56 is a statement of conservation of field momentum. The symbol w in equation A.55 represents the energy density (units $\mathrm{J/m^3}$) stored in the fields and can be separated into electric and magnetic contributions,

$$w = w_E + w_B, \tag{A.57}$$

where w_E and w_B represent the electric field energy and the magnetic energy, respectively. These are defined, in a vacuum, by:

$$\begin{aligned} w_E &= \frac{1}{2}\epsilon_0 E^2 \\ w_B &= \frac{B^2}{2\mu_0}. \end{aligned} \tag{A.58}$$

Transport of the field energy is quantified by the Poynting flux:

$$\mathbf{S} = \frac{1}{\mu_0}\mathbf{E} \times \mathbf{B}, \tag{A.59}$$

which measures amount of energy passing through unit surface area in unit time, i.e. $\mathbf{S} \cdot d\mathbf{a}dt$ is the total energy (in joules) passing through area $d\mathbf{a}$ in time interval dt.

In equation A.56, the quantity $\mathbf{g}$ is the field momentum density, defined by:

$$\mathbf{g} = \frac{\mathbf{E} \times \mathbf{B}}{\mu_0 c^2} = \frac{\mathbf{S}}{c^2}. \tag{A.60}$$

The Maxwell stress, T represents the transport of electromagnetic (linear) momentum, and is defined by:

$$\mathsf{T} = \epsilon_0 \left[\mathbf{E}\mathbf{E} + c^2\mathbf{B}\mathbf{B} - \frac{1}{2}\left(E^2 + c^2 B^2\right)\right]. \tag{A.61}$$

The Maxwell stress is a way of quantifying the amount of momentum passing through unit surface area in unit time, i.e. the quantity $\mathsf{T} \cdot d\mathbf{a}dt$ is the total momentum (in kg m/s) passing through area $d\mathbf{a}$ in time interval dt.

With the definitions and concepts of field energy, Poynting flux, field momentum, and Maxwell stress one can construct an intuitive picture of the conservation laws of equations A.55 and A.56. The left hand sides of these equations both have two terms: one representing intrinsic change in energy or momentum and one representing flow of energy or momentum. These two terms describing energy or momentum variation are balanced against local source of energy or momentum on the right hand sides of the equations. In the Poynting theorem, $-\mathbf{J} \cdot \mathbf{E}$ represents the energy density transferred from charges to the fields - a local (i.e. existing at each point in space) source of field energy. Likewise in equation A.56 $-(\rho_c \mathbf{E} + \mathbf{J} \times \mathbf{B})$ is recognizeable as a force (density, compare to the Lorentz force law) - which is a source of momentum (density) for the fields (being the opposite the force density exerted on the particles)

Those familiar with fluid mechanics will recognize these equations as being similar to fluid conservation laws like the Euler equations. While we do not directly use these equations in this course, they are still important as a way to succintly define, from a classical standpoint, the concept of field energy, energy transport, momentum, and momentum transport.

Exercises

A.1: Discuss the physical interpretation of Maxwell's equations. For each law try to write down a single sentence that outlines the meaning of that law. Take care to not fall into the logical trap of simply naming and explaining the mathematical symbols - instead you should endeavor to explain what these laws say about the relationship between electromagnetic fields and their sources.

A.2: Derive the differential forms of Gauss's law for the magnetic field and Ampere's law from their integral forms in equations A.13 and A.14.

A.3: Derive an equation for the magnetic field of a pure dipole by directly taking the curl of equation A.43. You are encouraged to follow these steps: (a) set your system up so that the dipole axis is in the z-direction, i.e. $\mathbf{m} = m\hat{\mathbf{e}}_z$, the dipole is located at $\mathbf{r}' = 0$. (b) express the field point in spherical coords as $r\hat{\mathbf{e}}_r$ and evaluate the cross product in equation A.43 (c) Convert/retain the result in spherical coordinates and take the curl. ANSWER:

$$\mathbf{B} = \frac{\mu_0 m}{4\pi r^3} \left(2\cos\theta\hat{\mathbf{e}}_r + \sin\theta\hat{\mathbf{e}}_\theta\right) \tag{A.62}$$

A.4: Derive a wave equation for the magnetic field by starting from the source-free Maxwell equations and invoking a procedue similar to that used to derive the wave electric field equation.

A.5: Verify equations A.54 by directly operating on the uniform plane expression with each derivative operator (grad, div, curl, time-derivative).

Bibliography

W. Baumjohann and R. Treumann. *Basic Space Plasma Physics*. Imperial College Press, London, England, 1997.

J. Beatty, C. Petersen, and A. Chaikin, editors. *The New Solar System*. Sky Publishing Corp., Cambridge, MA 02138, 4th edition, 1999.

J. Bennett, M. Donahue, N. Schneider, and M. Voit. *The Cosmic Perspective*. Pearson Addison-Wesley, San Francisco, CA 94111, 5th edition, 2008.

The British Geological Survey. Magnetic poles, June 2009. `http://www.geomag.bgs.ac.uk/poles.html`.

B. Brunches. Recherches dur le direction d'aimantation des roches volcaniques. *J. Phys.*, 5:705, 1906.

S. Chapman. The electrical conductivity of the ionosphere: A review. *Nuovo Cimento*, 5(Supplemental):1385–1412, 1956.

S. Chapman and V. Ferraro. A new theory of magnetic storms. *Nature*, 126: 129, 1930.

F. F. Chen. *Introduction to Plasma Physics and Controlled Fusion*, volume 1: Plasma Physics. Plenum Press, New York, NY 10013, 2nd edition, 1983.

P. David. Sur la Stabilité de la direction d'aimantation dans quelques roches volcaniques. *C.R. Acad. Sci. Paris*, 138:41, 1904.

C. S. Gillmor and R. Spreiter, editors. *Discovery of the Magnetosphere*. American Geophysical Union, Washington, D.C. 20009, 1997.

G. A. Glatzmaier and P. H. Roberts. A three-dimensional self-consistent computer simulation of a geomagnetic field reversal. *Nature*, 377:203–209, 1995.

C. Hoffmeister. Physikalische untersuchungen auf kometen ii, die bewegung der schweifmaterie und die repulsivkraft der sonne beim kometen. *Z. Astrophys.*, 22:265, 1943.

International Association of Geomagnetism and Aeronomy (IAGA), Division V, Working Group 8. The 9th Generation International Geomagnetic Reference Field. *Phys. Earth Planet Int*, 140:253–254, 2003.

J. A. Jacobs. *Reversals of the Earth's Magnetic Field.* Cambridge University Press, New York, NY 10011-4211, 1994.

M.-B. Kallenrode. *Space Physics: An introduction to plasmas and particles in the heliosphere and magnetospheres.* Springer, Berlin, Germany, third edition, 2004.

M. C. Kelley. *The Earth's Ionosphere.* Academic Press, Inc., San Diego, CA 92101, 1989.

M. Kivelson and C. Russell, editors. *Introduction to Space Physics.* Cambridge University Press, New York, NY, 10011, 1995.

M. Kono and P. H. Roberts. Recent geodynamo simulations and observations of the geomagnetic field. *Rev. Geophys.*, 40(1013):4, 2002. doi: 10.1029/2000RG000102.

S. McLean, S. Macmillan, S. Maus, V. Lesur, A. Thomson, and D. Dater. The US/UK World Magnetic Model for 2005-2010. *NOAA Technical Report*, NESDIS/NGDC-1, 2004.

R. T. Merrill and M. W. McElhinny. *The Earth's Magnetic Field.* Academic Press, Inc., London, England, 1983.

F. Nansen. *Farthest North.* Harper and Brothers Publisher, New York, NY, 1897.

M. Nicolet. The collision frequency of electrons in the ionosphere. *J. Atmos. Terr. Phys.*, 3:200–211, 1953.

E. Parker. Dynamics of the Interplanetary Gas and Magnetic Fields. *The Astrophysical Journal*, 128:664–676, 1958.

G. K. Parks. *Physics of Space Plasmas.* Addison-Wesley Publishing Company, Redwood City, CA 94065, 1991.

G. K. Parks. *Physics of Space Plasmas.* Westview Press, Boulder, CO, 80301-2877, 2nd edition, 2004.

H. Rishbeth and O. K. Garriott. *Introduction to Ionospheric Physics.* Academic Press, Inc., New York, NY 10003, 1969.

E. J. Smith, B. T. Tsurutani, and R. L. Rosenberg. Observations of the Interplanetary Sector Structure up to Heliographic Latitudes of 16°: Pioneer 11. *J. Geophys. Res*, 83(A2):717–724, 1978.

N. A. Tsyganenko. A magnetospheric magnetic field model with a warped tail current sheet. *Planet. Space Sci.*, 37:5–20, 1989.

N. A. Tsyganenko. Modeling the earth's magnetospheric magnetic field confined within a realistic magnetopause. *J. Geophys. Res.*, 100(A4):5599–5612, 1995.

N. A. Tsyganenko and M. I. Sitnov. Modeling the dynamics of the inner magnetosphere during strong magnetic storms. *J. Geophys. Res.*, 110(A3), 2005. doi: 10.1029/2004JA010798.

R. P. Wayne. *Chemistry of Atmospheres.* Oxford University Press, New York, NY, 2nd ed. edition, 1991.

Index

Physical Constants

Name	**Symbol**	**Value**	**Unit**
Elementary charge	e	$1.602176565(35) \times 10^{-19}$	C (measured)
Speed of light in vacuum	c	2.99792458×10^{8}	m/s (by def)
Permeability of the vacuum	μ_0	$4\pi \times 10^{-7}$	H/m (by def)
Permittivity of the vacuum	$\epsilon_0 \equiv \frac{1}{\mu_0 c^2}$	$8.854187817... \times 10^{-12}$	F/m
Molar gas constant	R	8.31441	$\mathrm{J{\cdot}mol^{-1}{\cdot}K^{-1}}$
Avogadro's constant	N_A	6.0221367×10^{23}	$\mathrm{mol^{-1}}$
Boltzmann's constant	$k = R/N_\mathrm{A}$	1.380658×10^{-23}	J/K
Stefan-Boltzmann constant	σ	$5.670373(21) \times 10^{-8}$	$\mathrm{W/m^2/K^4}$
Electron mass	m_e	$9.1093897 \times 10^{-31}$	kg
Proton mass	m_p	$1.6726231 \times 10^{-27}$	kg
Neutron mass	m_n	1.674954×10^{-27}	kg
Atomic mass unit	amu	$1.66053892 \times 10^{-27}$	kg
Radius of the Sun	$R_\odot$	696×10^{6}	m
Mass of the Sun	$M_\odot$	1.989×10^{30}	kg
Radius of Earth	R_E	6.378×10^{6}	m
Mass of Earth	M_E	5.976×10^{24}	kg
Astronomical unit	AU	$1.4959787066 \times 10^{11}$	m