Challenges to Understanding the Earth’s Ionosphere and Thermosphere

R. A. Heelis and A. Maute

1University of Texas at Dallas, Dallas, TX, USA, 2High Altitude Observatory, National Center for Atmospheric Research, Boulder, CO, USA

Abstract

We discuss, in a limited way, some of the challenges to advancing our understanding and description of the coupled plasma and neutral gas that make up the ionosphere and thermosphere (I-T). The I-T is strongly influenced by wave motions of the neutral atmosphere from the lower atmosphere and is coupled to the magnetosphere, which supplies energetic particle precipitation and field-aligned currents at high latitudes. The resulting plasma dynamics are associated with currents generated by solar heating and upward propagating waves, by heating from energetic particles and electromagnetic energy from the magnetosphere and by the closure of the field-aligned currents applied at high latitudes. These three contributors to the current are functions of position, magnetic activity, and other variables that must be unraveled to understand how the I-T responds to coupling from the surrounding regions of geospace. We have captured the challenges to this understanding in four major themes associated with coupling to the lower atmosphere, the generation and flow of currents within the I-T region, the coupling to the magnetosphere, and the response of the I-T region reflected in the neutral and plasma density changes. Addressing these challenges requires advances in observing the neutral density, composition, and velocity and simultaneous observations of the plasma density and motions as well as the particles and field-aligned current describing the magnetospheric energy inputs. Additionally, our modeling capability must advance to include better descriptions of the processes affecting the I-T region and incorporate coupling to below and above at smaller spatial and temporal scales.

Plain Language Summary

The ionosphere is the region of Earth’s upper atmosphere made up of a mixture of charged and neutral gases between approximately 50 and 1,000 miles (80–1,600 km) above the Earth’s surface. Sandwiched between the lower atmosphere and the magnetosphere, the ionosphere reacts to weather and climate near the Earth’s surface and to eruptions and sunspot activity on the Sun. The ionosphere absorbs the harmful radiation from the Sun and determines the fidelity of all radio communication, navigation, and surveillance transmissions through it. It is part of a complex, coupled system that changes on scales from meters to the planetary radius, and from seconds to decades. Understanding how the behavior of this region is controlled, by internal interactions and by the external regions to which it is coupled, is the preeminent challenge for the next generation of scientists. These challenges in understanding Earth’s ionosphere are associated with deciphering the many changes in neutral and plasma density and their relationships to the coupling with the Earth’s lower atmosphere, the generation and flow of currents within the region, and the coupling to the magnetosphere. Addressing these challenges requires advances in observing the composition and dynamics of the neutral particles and simultaneous observations of the charged particles, as well as the particles and field-aligned current describing the coupling of the ionosphere to the magnetosphere. Additionally, our modeling capability must advance to include better descriptions of the processes affecting the ionosphere and thermosphere region and to incorporate coupling with the regions below and above at smaller spatial and temporal scales.

1. Introduction

About 80 km altitude above the Earth’s surface, the absorption of solar radiation at extreme ultraviolet (EUV) and X-ray wavelengths by the neutral atmosphere produces a mixture of neutral gas, positively charged ions and electrons, collectively referred to as the ionospheric plasma. The Earth’s approximate dipole magnetic field threads the entire ionosphere-thermosphere (I-T) region, connecting the two hemispheres at middle and low latitudes along geomagnetic field lines (e.g., Thébault et al., 2015). At high
latitudes the geomagnetic field passes through the I-T and extends to great distances into the inner magnetosphere, the outer magnetosphere, and the magnetopause (e.g., Dungey, 1965; Tsyganenko, 1995).

At upper mesospheric and lower thermospheric altitudes, the electrons are magnetized and strongly bound to the Earth's magnetic field. However, the neutral number density is so much larger than the plasma number density that collisions between the ions and the neutral gas allow currents to flow perpendicular to the magnetic field. Thus, the region becomes an anisotropic electric conductor.

Above 500 km, collisions between the neutral gas and the ions are so low that the charged gas behaves as a magnetized collisionless plasma. However, in the region from 100 to 500 km altitude, the interactions between the charged and neutral species are so important that it is not possible to discuss or understand the behavior of one of the species without consideration of the other (e.g., Prölls, 2012). This has led to the term ionosphere-thermosphere or I-T to describe this region, highlighting the importance of considering both the neutral gas and the plasma. We will use the term I-T for this purpose throughout this article.

The I-T region of our space environment is host to a multitude of space assets varying in size from the International Space Station to small satellites with dimensions of tens of centimeters. Each is subjected to a highly variable drag due to the sensitivity of the neutral atmosphere to energy sources from the Sun and the magnetosphere, making orbit tracking and re-entry predictions particularly challenging. Many resources use communication and navigation signals that must pass through the I-T, which can produce signal fading and signal loss due to the presence of plasma structure. Finally, the existence of an electrical conducting layer in space above the conducting Earth itself can lead to induced currents in the ground that can be particularly hazardous to distribution systems that utilize highly conductive elements such as electrical grids and pipelines (e.g., McIntosh, 2019; Moldwin, 2008). To protect space-based assets and minimize the impact of space weather on society with reliable specification and forecast, it is critical that we understand how to specify its state and predict its evolution by advancing our holistic understanding of the I-T system.

Overviews of the fundamental processes that govern the behavior of the I-T appear in a number of key reference books (e.g., Hargreaves, 1992; Kelley, 2009; Rishbeth & Garriott, 1969; Schunk & Nagy, 2009). In addition, there exist insightful reviews and papers collected in monographs of specific areas relevant to the I-T behavior. We will refer to this material during the course of the discussion. However, this article is written to highlight the major challenges to advancing our understanding of the I-T system and its connections to the space environment from above and below. Observations of the I-T have been made for many years and are readily accessible from one of the World Data Centers (e.g., https://spdf.gsfc.nasa.gov), and mission data are permanently archived at the NASA National Space Science Data Center. These databases together with numerical and empirical models, many of which are accessible from the Community Coordinated Modeling Center (https://ccmc.gsfc.nasa.gov), represent the legacy upon which future advances will be made.

In the I-T, the neutral atmosphere transitions from the well-mixed gases that exist near the surface of the Earth to diffusively separated constituents that are distributed by altitude in the thermosphere according to their mass. This condition of diffusive equilibrium, in which the partial pressure gradient in the gas is balanced by the gravitational force, is commonly maintained in the upper neutral atmosphere. Within the I-T, absorption of far ultraviolet (FUV) and EUV radiation from the Sun heats the neutral gas, dissociates molecular oxygen to produce atomic oxygen, and ionizes the neutral gas to create the ionosphere during the daytime. The neutral gas density decreases with increasing altitude while the plasma density has local maxima in altitude termed the E region (approximately 90–150 km), associated with the balance between the ion production and chemical losses, and the F region (approximately 150–500 km) associated with the balance between ionization sources, chemical losses, vertical diffusion, and transport associated with winds and electric fields.

Figure 1 shows for local noon and local midnight the fractional contribution to the total neutral mass density (the mass mixing ratio) and the number density of atomic oxygen, O, and molecular nitrogen, N₂, as a function of altitude in the thermosphere. These results, from the output of the Thermosphere-Ionosphere-Electrodynamical General Circulation Model (TIEGCM) for solar minimum conditions (Maute, 2017), show little variation in the mixing ratio between noon and midnight. At the lowest altitudes, the contribution from molecular oxygen is significant (see Figure 1, right panel), but O and N₂
dominate above about 150 km, and $O$ above approximately 250 km. In this case the neutral mass densities at noon and midnight are $5.4 \times 10^{-14}$ g/cm$^3$ and $4.7 \times 10^{-14}$ g/cm$^3$ around 250 km, respectively, but changes driven by solar activity can be as large as a factor of 3.

Figure 2 shows the distribution of total plasma density as a function of altitude and latitude for typical noon and midnight conditions corresponding to the TIEGCM simulation in Figure 1 with solar radio flux $F_{10.7}=68$ sfu and geomagnetic Kp index Kp = 1$-$. The dipole magnetic field of the Earth is overlaid to emphasize the organization of the plasma by the magnetic field. The results are shown for 0° geographic longitude where the magnetic equator, indicated by the dashed line in Figure 2, lies north of the geographic equator. At low and middle latitudes, the plasma number density has altitude peaks during the daytime identifying the $E$ region in a narrow area near 110 km and the $F$ region with a peak near 250 km at middle latitudes and near 350 km at the equator. While the $E$ region plasma density is determined by chemical equilibrium conditions, the $F$ region plasma density is strongly influenced by the daily plasma dynamics that lifts the plasma during the day and forms the so-called equatorial ionization anomaly (EIA) that is weakly preserved throughout the night and appears as ionization peaks displaced from the magnetic equator. At low and middle latitudes, the $E$ region is largely absent at night due to the lack of a solar photoionization and the dominant chemical loss processes. However, at high latitudes energetic particle precipitation from the magnetosphere can maintain the $E$ and $F$ region plasma density as can be seen by the appearance of $E$ and $F$ region density peaks poleward of 60° in both hemispheres at midnight. Note that at midday, the southern polar region is in darkness and without appreciable precipitation leading to low densities at the highest latitudes in the Southern Hemisphere polar cap.

The conductivity, which varies significantly with height, determines how currents are generated and closed in the I-T. The conductivity tensor is typically specified in terms of the direct conductivity, which captures the relative mobility of the ions and electrons in response to an equivalent electric force in the direction parallel to the magnetic field and the Pedersen and Hall conductivities, which specify the relative mobility of the ions and electrons in response to an electric force perpendicular to the magnetic field and parallel and perpendicular to the electric force, respectively. Expressions for the conductivity components with varying degrees of specificity can be found in books by Kelley (2009) and Schunk and Nagy (2009).

It is important to recognize that the direct conductivity is very large everywhere in the I-T, allowing the Earth’s magnetic field lines to be considered as electric equipotentials for all spatial scales above 10 km (Farley, 1959; Heelis & Vickrey, 1990). By contrast, the Pedersen and Hall conductivities are strong.
functions of altitude and local time (see Figure 3) and largely determine the most significant currents that flow perpendicular to the magnetic field. Figure 3 illustrates the Hall and Pedersen conductivities associated with the plasma density variation for solar minimum conditions in Figure 2. Note that the color scale differs by a factor of 100 between the daytime and nighttime conductivity distributions. The daytime Hall conductivity resides in a narrow layer near 110 km. This peak corresponds to the relative plasma density maximum in the $E$ region. By contrast, the daytime Pedersen conductivity is distributed over a wider altitude range reflecting a dependence on the plasma density in the $E$ and $F$ regions. During the nighttime the $F$ region Pedersen conductivity decays rather slowly because the chemical loss time of the plasma is quite slow (approximately 1 hr) and is partly compensated by downward plasma drifts perpendicular to the magnetic field and downward diffusion parallel to the magnetic field. Thus, $E$ and $F$ region Pedersen conductivities become comparable for the illustrated solar minimum conditions, but for solar maximum conditions (not illustrated), the $F$ region Pedersen conductivity dominates. The nighttime Hall conductivity remains in a narrow layer near 110 km but is significantly less than the daytime value by a factor of 100. At high latitude during nighttime, the effects of energetic particle precipitation on the Hall and Pedersen conductivities are seen by the elevated values at midnight. Almost no effects are seen at midday since the contributions of cusp precipitation are not included in this model.

Figure 2. Electron density log$_{10}$[#/cm$^3$] from TIEGCM simulation conditions for doy 181, 2009; 12 LT; and 0 LT. The magnetic field geometry and the magnetic equator are indicated by the dotted and dashed lines, respectively. See Figure 1 for details about the simulation.

Figure 3. Pedersen and Hall conductivity log$_{10}$ [S/m] at 12 LT (left panels) and 0 LT (right panels) from TIEGCM simulation conditions for doy 181, 2009. See Figure 1 for details about the simulation.
In the $F$ region, the plasma distribution is in large part determined by the plasma motion (Pfaff, 2012). Parallel to the magnetic field, the plasma responds to the collisional forces with the neutral gas and the gravitational force and plasma pressure gradient force components. Neutral winds move plasma along the magnetic field in the $F$ region producing hemispheric asymmetries in the plasma density as seen in Figure 2. Field-aligned plasma motions allow solar-produced ionization to fill the outer reaches of the magnetic flux tubes at low and middle latitudes where charge exchange between $O^+$ and $H$ produces a reservoir of cold $H^+$ in the plasmasphere (see, e.g., Goldstein, 2006, for a review). At high latitudes, parallel plasma motions result from frictional and particle heating and provide an upward flux of ions (Coley et al., 2003), some of which escape into the magnetosphere (Engwall et al., 2009). Perpendicular to the magnetic field, the plasma motion transports plasma horizontally and vertically and combined with diffusive motions parallel to the magnetic field they produce plasma density gradients that define large- and small-scale features throughout the I-T. Key to describing these plasma motions is a determination of the currents, and many of the significant challenges arise from identifying the source of those currents and how they evolve in time.

Currents perpendicular to the magnetic field arise within the I-T through collisions between the neutrals and charged particles. Currents are also produced by the plasma pressure gradient (diamagnetic currents) and by the gravitational force component perpendicular to the magnetic field (e.g., Alken et al., 2017). In addition, the connection of the I-T to the magnetosphere results in a current systems of magnetospheric origin that close through the I-T (e.g., Siscoe & Maynard, 1991). All these currents close in various paths through the I-T aided by the highly conducting magnetic field and controlled by the spatial distribution of the conductivity. Thus, the currents and the associated plasma motion can be distributed over great distances, making it a challenge to identify the sources.

The daily variation of absorption of solar radiation within the I-T and in the lower atmosphere provides heat sources that lead to periodic large-scale oscillatory motions in the neutral gas that move (migrate) with the apparent motion of the Sun around Earth. Additional heat sources from features fixed with respect to the Earth lead to large-scale nonmigrating oscillatory motions. These tidal oscillations are usually designated by letters that describe the period (diurnal, semidiurnal, etc.), the propagation direction (eastward or westward), and the wave number in longitude. There also exist global scale oscillations with periods of several days, which are referred to as planetary waves (PWs). Other energy sources from the ground and from the lower atmosphere may also produce waves with smaller spatial scales that propagate into the I-T region. These features are collectively referred to as gravity waves. The waves interact with each other and with any average prevailing flow as they propagate through the atmosphere. The winds that reside in the $E$ region and the $F$ region produce the currents associated with plasma motion at low and middle latitudes (e.g., Richmond, 1979; Richmond & Thayer, 2013). In the $F$ region, the wind is affected by collisions with the plasma and the so-called ion drag force, and thus, the forces producing the plasma motion and the plasma motion itself are coupled to the neutrals. The relative motion between the plasma and the neutral gas not only provides the ion-drag force but also heats the plasma and the neutral gas through dissipation of kinetic energy. This so-called frictional heating is most dramatic at high latitudes where the plasma-neutral velocities are highest but may also be effective at lower latitudes.

At high latitudes the horizontal currents and associated plasma motion in the I-T are dominated by field-aligned currents imposed from the magnetosphere. The closure paths for these currents are dependent on the spatial distribution of the field-aligned currents and the distribution of ionospheric conductivity, both of which evolve in time. The most fundamental of the high-latitude current systems are the so-called Regions 1 and 2 currents that are consistent with the familiar two-cell circulation in the plasma at high latitudes (see Figure 10, black dashed lines). This circulation of the plasma is imprinted on the neutral gas through collisions, and the circulations of the plasma and the neutral gas change the plasma number density and the conductivity. Thus, the dynamo current associated with the wind is changed, as is the distribution of the field-aligned current. Therefore, the I-T is not a passive receiver of the field-aligned current from the magnetosphere but is an active element that evolves in time to influence the flow of electromagnetic energy between the I-T and the magnetosphere. In addition, the closure paths for the field-aligned currents imposed at high latitudes are not confined to high latitudes but may extend to the magnetic equator over short time periods. Thus, the plasma dynamics of the entire I-T is influenced by its connections to the magnetosphere, and the magnetospheric dynamics are influenced by the evolving conductivity and winds of the I-T.
Finally, we note that the time-evolving motion of the plasma throughout the I-T produces large-scale variations in the plasma density that are identified by significant spatial gradients. The existence of currents perpendicular to the magnetic field, and spatial gradients in the plasma density and the plasma velocity, provides sources of free energy for the formation of plasma instabilities. Many of these instabilities in the $E$ and the $F$ regions produce the scattering of radio signals that in some cases are used to probe the region and in other cases are detrimental to the propagation of critical communication and navigation signals.

Our goal is to understand the internal workings of the I-T well enough to describe how it interacts with the forcing region below and the magnetosphere above. While the physical processes at work can be listed separately, they interact nonlinearly at different rates and different scale sizes. Those interactions are both influenced by the drivers and also modify the drivers. Thus, the effectiveness of different processes will depend on initial conditions and a response time, which are functions of spatial scale. The initial conditions for the I-T response encompass variations on time scales of years, to seasonal and day to day, and in themselves, there are questions about I-T longer-term variations that will not be the focus of this report. This coupled I-T system presents a challenging environment for which carefully planned observations and detailed advanced modeling are required to discover the optimal way to specify and predict its behavior. It should be recognized that challenges emerge as models are developed and observations are interpreted. We have not therefore attempted to “list” all the challenges that come to mind. Rather, we discuss those that are most apparent to the authors in four overarching themes associated with wind systems in the upper atmosphere, current flow and closure in the I-T region, coupling of the I-T region to the magnetosphere, and plasma and neutral structures that interact with each other over different spatial and temporal scales.

2. Major Challenges in the Ionosphere-Thermosphere System

2.1. To Understand the Wind Variations in the Upper Atmosphere

Neutral winds serve to redistribute mass and momentum in the neutral gas, affecting its chemistry and heat balance. In addition, collisions between charged and neutral species generate currents and associated plasma motions that also affect the distribution, chemistry, and heat balance of the I-T. A broad understanding of the neutral wind variation in the I-T is therefore essential to an overall appreciation of the behavior of the neutral gas and plasma in the I-T.

Wind measurements in the upper thermosphere are rather sparse and poorly distributed around the globe. Most available measurements are incorporated into a global model that is widely used to determine the effects of ion-neutral coupling on the plasma (Drob et al., 2013). However, such model-derived winds often depart from local measurements in the $F$ region (e.g., Meriwether et al., 2011) and from the extensive structure seen at lower altitudes from chemical release experiments (Larsen, 2002). As more wind measurements become available (e.g., Makela et al., 2013; Visser et al., 2013), we might expect improved specifications, but presently, our understanding is mostly based on models of the coupled ionosphere and thermosphere. An example of our general understanding of the wind variation is presented in Figure 4. The winds shown here represent those for low solar activity and moderately low magnetic activity, where the influences of high-latitude energy inputs, discussed in more detail later, are small but can nevertheless be seen in zonal and meridional wind enhancements at the highest latitudes. The signature of upward propagating waves and tides with the reversal of wind direction with height can be identified in the $E$ region and lower $F$ region at middle and low latitudes, while in the upper thermosphere in situ heating becomes more dominant. This displayed smooth variation is in strong contrast to the limited number of observations indicating larger wind variability mentioned earlier. Presently, the sparsely observed neutral wind system limits our ability to fully understand the following:

1. the spectrum of wave and tidal variations propagating from the lower atmosphere into the upper atmosphere;
2. the critical processes at the transition to space near 100 km altitude;
3. all the processes affecting the in situ wind in the upper atmosphere; and
4. the effect of high-latitude energy inputs on the neutral wind systems at different spatial scales.

In the following, we give examples highlighting specific issues in each of these four areas.
2.1.1. Wave Spectrum From the Lower Atmosphere

Tides associated with the thermal balance of the atmosphere vary significantly from one day to the next with amplitudes doubling or tripling (e.g., She et al., 2004), and numerical studies demonstrate that variability from the lower atmosphere is a source of the ionospheric day-to-day variability (Liu et al., 2013). Several effects contribute to the day-to-day tidal variability including variability of the tidal sources, variability of the propagation conditions, variability of nonlinear interactions and modulations of tides, and perhaps the chaotic nature of the system (Lorenz, 1969). Discovering the most prominent causes of tidal variability is an essential step in the specification of the spectrum of wind variability driving the I-T system from below.

The background condition through which tides propagate influences the tidal characteristics seen in the I-T. Numerical simulations indicate that the autocorrelation in time of the longitudinally average zonal wind decreases from the stratosphere (10–30 days) to the mesosphere and lower thermosphere (MLT) region (<5 days). A possible explanation for the short correlation time in the thermosphere is the increasing influence of wave dissipation. The correlation time of some tides, for example, the diurnal migrating tide, also decreases with altitude, suggesting an influence of the mean atmosphere through which they propagate. However, other tidal modes (e.g., semidiurnal migrating tide) do not strictly follow this behavior, indicating that other processes influence the tidal variability (e.g., Liu, 2016).

Compared to the day-to-day variability, tidal variability on longer time scales (approximately 10 days and longer) in the lower thermosphere is better correlated to the tidal variations in the troposphere and stratosphere (e.g., Liu, 2016). Thus, over time scales of weeks to years, meteorological events such as El Niño and Southern Oscillation (ENSO), Quasi-Biennial Oscillation (QBO), sudden stratospheric warming (SSW), and PW activity provide opportunities to study the coupling between the lower atmosphere and the I-T system from the observational and modeling perspective.

ENSO periods are linked to an approximately 10–30% increase in variability in the lower thermosphere, in the neutral winds, and in the $F$ region ionosphere (e.g., Pedatella & Liu, 2013). ENSO and QBO can also

---

![Figure 4](https://example.com/figure4.png)

**Figure 4.** Zonal and meridional wind [m/s] at 12 LT (left panels) and 0 LT (right panels) from TIEGCM simulation conditions for doy 181, 2009. See Figure 1 for details about the simulation.
cause thermospheric neutral density variations of the order of 1−3% (Liu et al., 2017). QBO-like signatures are seen in the ionospheric current system, suggesting an impact on winds in the $E$ region (e.g., Yamazaki et al., 2017), where a correlation with $E$ region scintillation (sporadic $E$) is also suggested (e.g., Chang et al., 2018). SSW periods can be associated with 100% changes in plasma density in the $F$ region (e.g., Chau et al., 2012; Goncharenko et al., 2010) and approximately 5–30% in neutral density (e.g., Liu et al., 2017).

These are significant changes in the I-T system, but the importance of different mechanisms is still not understood to the level of prediction. In general, the long-term events, for example, ENSO, QBO, and SSW, have in common that the observed tidal spectrum in the lower thermosphere region is significantly modified (e.g., Fuller-Rowell et al., 2010; Lieberman et al., 2007; Wu et al., 2008). Several mechanisms are suggested to contribute to these changes, such as modified zonal mean state through which tides propagate (e.g., Sassi et al., 2004), modified nonlinear interaction between wave-tide and tide-tide, and modified tidal excitation by latent heat, radiative heat, and ozone (e.g., Yiğit et al., 2016). In addition, the lunar gravitational tidal signal can be significantly enhanced during SSW periods (e.g., Forbes et al., 2013). Finally, on time scales of several days, PW signals are observed in the I-T (e.g., Gu et al., 2014; Pancheva et al., 2009) even though PWs cannot easily propagate above the mesopause.

In addition to variability associated with large-scale meteorological events and tidal oscillations, the I-T also experiences neutral atmospheric motions on much smaller spatial and temporal scales. It is becoming increasingly clear that gravity waves can have a major influence on the I-T system; however, getting a global picture of the gravity wave spectrum from observations is a challenge (Alexander et al., 2010). Gravity wave amplitudes, although small in the lower atmosphere, grow exponentially with decreasing atmospheric density. Breaking gravity waves can then deposit momentum and change the heat flux, which modifies the background conditions through which the tides propagate (e.g., Liu & Hagan, 1998; Orland & Alexander, 2006). Gravity waves may also interact directly with propagating tidal oscillations to modify the tidal wave amplitude (e.g., Thayaparan et al., 1995; Walterscheid, 1981) and can also contribute directly to small-scale turbulent mixing and transport of mass and heat.

An overview of gravity wave measurements, the gravity wave distribution, and their momentum flux is provided by Geller et al. (2013) and Alexander et al. (2010). Gravity wave sources are variable and complex, including flow over mountain ridges, shear along wind jets, earthquakes and associated tsunamis, tropical convections, and frontal systems (Fritts & Alexander, 2003). Several possible pathways for gravity waves to reach the thermosphere and ionosphere are identified such as direct propagation and generation of secondary and tertiary waves in the mesosphere and thermosphere (e.g., Becker & Vadas, 2018; Vadas & Becker, 2019; Vadas & Liu, 2009), but all are sensitive to the background wind and temperature conditions. While most of these pathways are revealed in high resolution and/or regional models (e.g., Becker & Vadas, 2018; Heale et al., 2017; Miyoshi et al., 2014), their more global effects on the I-T are not directly included in models that do not have the necessary spatial resolution. Rather, these models parametrize in ad hoc ways, for example, the gravity wave effects of momentum deposition and eddy diffusion at different spatial scales. In addition, the direct effects of gravity waves, such as their impact on the plasma, cannot be captured using parametrizations.

Prominent examples of gravity wave effects in the I-T system are medium-scale traveling ionospheric disturbances (MSTID), which have been visualized in maps of total electron content (TEC) generated from ground-based Global Navigation Satellite System (GNSS) receivers (Tsugawa et al., 2007) and have been associated with earthquakes (e.g., Makela et al., 2011; Tsugawa et al., 2011), tornados (e.g., Nishioka et al., 2013), and tropical convection (e.g., Fukushima et al., 2012). Gravity waves propagate away from the source with amplitudes that increase with height (Hickey et al., 2009) forming traveling atmospheric disturbances (TADs) and largely horizontal winds that move plasma parallel and perpendicular to the magnetic field, directly through collisions and indirectly through dynamo forces, respectively (Huba et al., 2015; Shiokawa et al., 2005). The effect of gravity waves and associated TADs on the current and plasma distribution will be discussed in section 2.2 and on the small-scale plasma structure and irregularity development in section 2.4.

More comprehensive observations of PWs, tides and gravity waves in the I-T system are required to quantitatively connect I-T variability to the lower atmospheric variations and the propagation conditions, and to advance our model descriptions to the level required to capture I-T effects of waves with different scale sizes.
2.1.2. Critical Processes at the Transition to Space

Much of the variability in the winds in the I-T is produced by waves that originate below the turbopause and propagate upward through the region where turbulent mixing is prominent, into the regions dominated by molecular diffusion at higher altitude. In the transition region to space around 100 km turbulent mixing can transport species from high- to low-concentration regions, radically changing the temperature and composition structure of the atmosphere and the propagation characteristics of the waves.

While turbulent heating can be locally important in the MLT region (e.g., Lübken, 1997), gravity wave dissipation can significantly modify the energetics and composition nonlocally since they result in heating and introduce vertical fluxes of heat and atmospheric constituents (Garcia et al., 2007; Liu, 2000; Walterscheid, 1981). Small-scale waves can also contribute significantly to large wind shear generation in the MLT (Fritts et al., 2004; Liu, 2017), which then can affect sporadic E layer generation (Haldoupis, 2011), described in section 2.4.

Correctly capturing the effects of transport and mixing by dissipative waves and turbulence on the atmosphere is therefore critical to describing the effects of waves and reconciling the observed behavior of the I-T with the conditions at the lower boundary. Presently, such effects are captured in models in general by an eddy diffusion coefficient with a magnitude that depends on the scale sizes being considered (e.g., Liu, 2016; Siskind et al., 2014). The eddy diffusion is anisotropic on large scales with more efficient eddy transport in the horizontal than vertical direction. In most numerical models (e.g., Thermosphere-Mesosphere-Ionosphere-Electrodynamic GCM and Whole Atmosphere Community Climate Model-eXtended) only vertical, and not horizontal, eddy diffusion is considered. However, quantifying the eddy diffusion coefficient is a challenge, and a brief discussion highlighting the complexities of observational efforts to determine this coefficient is provided by, for example, Lübken (1997) and Vlasov and Kelley (2014).

In numerical models the eddy diffusion coefficient is poorly quantified and often adjusted ad hoc to improve agreement with some I-T observations, for example, annual/semiannual variation in thermospheric neutral density (Qian et al., 2009), but at the same time, it is changing other parameters in the neutral gas and plasma as well (e.g., Siskind et al., 2014).

*Improvements in the methodologies to represent the effects of eddy diffusion and the physical descriptions of the processes are required to more confidently specify the response of the I-T to wave forcing on various scales from below.*

2.1.3. Capturing Processes in the Upper Thermosphere

In the I-T system, the neutral wind is strongly influenced by the coupling to the lower atmosphere. But the upper thermosphere itself is a source of winds that are dependent on internal processes and forcing through magnetosphere-ionosphere (M-I) coupling. The winds generated by these sources interact with the plasma and feed back through the ion drag force. The coupling and feedback is highly dependent on location and geophysical conditions, and the neutral flow can modify the plasma distribution and the neutral density, as well as transport momentum and energy globally. Thus, localized processes have effects beyond their source region. Understanding how each of these influences evolves in space and time is necessary to describe the behavior of the I-T.

In the low- and middle-latitude region, the neutral wind in the upper thermosphere is primarily influenced by solar EUV heating and infrared radiation (IR) cooling. But the wind itself also affects the thermal balance of atmosphere, producing important feedback that must also be considered. During solar maximum, ion drag forces become important and change the wind such that the pressure gradient is increased and the thermal balance is adjusted to maintain a strong day-night energy difference (Hsu et al., 2016). During solar minimum, viscous drag becomes more important leading to a reduced day-night difference through divergent daytime and convergent nighttime winds and associated adiabatic cooling and heating, respectively. Above 300 km the neutral wind varies very little with altitude but is strongly influenced by solar radiation and by ion drag. For example, at night when the neutral wind tends to be eastward due to the afternoon peak in thermospheric pressure, the F region plasma moves also eastward reducing the ion drag force on the neutral gas and allowing an increased wind to the east. Thus, the daily averaged wind is eastward, which describes the atmospheric superrotation (Liu et al., 2006; Rishbeth, 1972; Wharton et al., 1984). However, below 300 km the neutral wind varies greatly as the drag forces and the influence of upward propagating tides and their associated daily variability increase. Below 250 km, for example, the zonal wind can...
change direction, as seen in Figure 4, with important impacts on the current and the stability of the ionosphere and the thermal structure and composition of the neutral atmosphere.

While computational models can verify that the height variation of the wind has profound effects on the I-T system, they must be combined with observations of the plasma and neutral density and their motions to validate and, subsequently, to unravel the roles of lower atmosphere forcing and internal processes.

2.1.4. Processes at High Latitude on Different Scales
Conceptually, we often separate the low- and middle-latitude from the high-latitude region, even though the regions are strongly coupled. Energy input into the high-latitude region can be highly variable and initiate responses on different spatial scales. The I-T response is not constrained to the region of energy input, but it can spread in space and persist longer than the initial energy input. At middle and high latitudes the neutral dynamics combines the effect of tides and waves with the effect from electromagnetic and particle energy inputs from the magnetosphere. The magnetospheric energy inputs remain imprinted on the neutral atmosphere well after the magnetospheric forcing has ceased or changed. Thus, separating the effects from the lower atmosphere and the magnetosphere is a significant challenge.

The electromagnetic energy inputs from the magnetosphere, which set the ionospheric plasma in motion, also provide the momentum forcing that creates a large-scale circulation of the neutral atmosphere at altitudes above 300 km. In this regime the observed similarities between the neutral circulation and the plasma convection in a long-term, large-scale average sense, including dependencies on the interplanetary magnetic field (IMF) orientation, are a comforting confirmation that physical processes are understood (Dhadly et al., 2017). However, the large-scale driver of the momentum forcing, the plasma drift, can change rapidly and be variable over a wide range of spatial scales. The momentum exchange between neutral gas and plasma can be surprisingly fast, ranging from a few minutes to more than an hour (e.g., Conde et al., 2018; Kosch et al., 2010). At lower altitudes the time scales are much longer, but in all cases it should be recognized that the neutral dynamics reacts generally on time scales longer than those associated with the reconfiguration of the global plasma convection pattern, on the order of 13 min (Ridley et al., 1998).

Recently a comprehensive observational network over Alaska has allowed the study of winds in the $F$ region in connection with ion flow and particle precipitation with unprecedented spatial and temporal resolution (Conde et al., 2018). These observations indicate that the $F$ region wind (approximately at 240 km altitude) responds to changes in the drivers within 15 min on spatial scales of 100 km or smaller. Good spatial coverage also reveals stronger $F$ region vorticity flow than divergent flow, consistent with momentum forcing from the plasma.

Observations suggest that the response time is reduced in regions of enhanced plasma density associated with auroral arcs where the neutrals would be more effectively accelerated by the plasma flow (Kiene et al., 2018; Walterscheid & Lyons, 1992). However, increased momentum forcing also changes the neutral pressure, and gravity waves, generated by unstable neutral flow in aurora arcs, can distribute momentum meridionally to decrease the wind shear (Larsen & Conde, 2011). At high latitudes the spatial extent and temporal persistence of the neutral circulation being different from those of the plasma lead to significant differences between simultaneous colocated neutral wind and plasma velocity measurements (Killeen et al., 1984), which are more pronounced at lower altitudes where the ion-neutral coupling times are longer. Where the plasma and neutral velocities differ, the I-T is subject to frictional heating or Joule heating (Strangeway, 2012) that subsequently modifies the neutral wind and density patterns over a range of spatial scales that are different at different altitudes (Walterscheid & Brinkman, 2008; Walterscheid et al., 1985). The effectiveness of these energy dissipation processes is also highly dependent on altitude (Thayer, 1998). Below 120 km momentum, forcing and heating tend to be equally important in driving the neutral flow (Kwake & Richmond, 2017). However, the neutral gas density is so large that the effects on the neutral temperature and dynamics are quite small.

At high latitudes the spatial extent and temporal persistence of the neutral circulation being different from those of the plasma lead to significant differences between simultaneous colocated neutral wind and plasma velocity measurements (Killeen et al., 1984), which are more pronounced at lower altitudes where the ion-neutral coupling times are longer. Where the plasma and neutral velocities differ, the I-T is subject to frictional heating or Joule heating (Strangeway, 2012) that subsequently modifies the neutral wind and density patterns over a range of spatial scales that are different at different altitudes (Walterscheid & Brinkman, 2008; Walterscheid et al., 1985). The effectiveness of these energy dissipation processes is also highly dependent on altitude (Thayer, 1998). Below 120 km momentum, forcing and heating tend to be equally important in driving the neutral flow (Kwake & Richmond, 2017). However, the neutral gas density is so large that the effects on the neutral temperature and dynamics are quite small.

At the highest altitudes the Joule heating rate per unit volume is much smaller, but the heating rate per unit mass is large and can significantly affect the relatively low neutral gas density. In the intermediate region between 150 and 250 km, the effects of Joule heating are most dramatically reflected in the temperature and dynamics of the neutral atmosphere (Huang et al., 2012). The heating effects depend strongly on the solar wind forcing and increase more rapidly than the momentum forcing (Kwake & Richmond, 2017).
Precipitating energetic particles represent an additional energy input that on large scales tends to be spatially organized in a similar way to the electromagnetic energy input. When the IMF has a southward component, particle precipitation appears in the well-known configuration of the auroral oval (Akasofu, 1963). However, within this region there are significant variations that occur over smaller spatial scales and shorter time scales such as are seen in the cusp region (Onsager et al., 1993) and during magnetically disturbed periods across the nightside (Wing et al., 2013). Selected features will be discussed in section 2.4, but here we note that the neutral wind variations associated with these features have not been systematically described.

The magnetospheric energy inputs to the I-T not only modify the neutral winds at high latitudes but also affect the winds at middle and low latitudes via a modified pressure distribution. Particle and frictional heating at high latitudes also produces a vertical circulation of the neutral atmosphere that is coupled to the horizontal motions associated with the disturbance dynamo (see section 2.2). During times of high magnetic activity, these so-called disturbance winds can dominate observations at middle latitudes (e.g., Blanc & Richmond, 1980). However, modeling the winds is highly dependent on quantifying the magnetospheric energy inputs, which are poorly specified depending on the spatial and temporal scales (see section 2.3). Impulsive forms of magnetospheric energy input lead to TADs in the neutral atmosphere, which move away from the source region equatorward and poleward and are associated with large-scale ionospheric disturbances (LSTID) discussed in section 2.4. The plasma and neutral gas response to the electromagnetic and particle energy dissipation strongly depend on the spatial scale, the duration over which the energy is deposited, and the evolution of the energy input. Variations in the energy input characteristics can produce both large-scale reconfigurations and impulsive propagating disturbances in the I-T system.

To capture the neutral and plasma response to magnetospheric energy input, we need to understand the spatial and temporal scales over which the drivers are most effective. This significant challenge is exacerbated by the lack of comprehensive observations.

2.2. To Connect the Winds to the Currents in the I-T System

At high latitudes and during times of moderate to strong magnetic activity, the dynamics of the I-T are strongly influenced by current systems from the magnetosphere. There are unique challenges associated with this interaction between the I-T and magnetosphere that will be discussed in section 2.3. However, the energy transferred to the I-T system by magnetosphere-ionosphere coupling and more directly by solar radiation, produces global wind systems that in turn influence the plasma motion. This interaction between the charged and neutral particles in a magnetized planetary atmosphere determine the underlying daily, seasonal, and longer-term baseline behavior of the I-T. The physical processes determining the relationships between the neutral gas motion and the plasma motion are time dependent and discussed in detail by Vasyliunas (2012). Here, we confine our discussion to quasi-steady state conditions for which it should be understood that no causative links between the plasma motions, the currents, and the electric field can be established. Rather, what is referred to as a dynamo current or wind-related current represents the magnetic stresses that are distributed along the magnetic field and are required to balance the collisional force between the plasma and the neutral gas. In the F region, the behavior of electric field E and the plasma drift V may be viewed equivalently through the expression \( \mathbf{E} = -\mathbf{V} \times \mathbf{B} \).

This baseline behavior results from a complex feedback loop with the following high-level features:

1. The neutral winds are associated with currents perpendicular to the magnetic field that are dependent on the ionospheric conductivity.
2. The plasma motion and associated current change the spatial and temporal distribution of plasma and the neutral gas throughout the I-T.
3. The spatial and temporal distribution of plasma and the neutral gas influences the conductivity directly and the wind system indirectly, through the ion drag force.

In order to understand the connections between the winds and the plasma motions, we must be able to specify the spatial and temporal evolution of the winds, the ionospheric conductivity distribution, and the distribution of the current. Presently, there are no contemporaneous measurements of these key parameters that allow us to describe these variables.
2.2.1. Ionospheric Conductivity

The ionospheric conductivity is the key tensor variable determined by the mobility of charged particles. As outlined in section 1, it is dependent on the plasma and neutral number densities as well as the charged particle gyrofrequencies and their collision frequencies with neutrals. At low and middle latitudes the plasma is produced in the daytime by solar EUV radiation, and the conductivity perpendicular to the magnetic field is largest in the F region, where there is a simple balance between the photoproduction and the chemical losses. The F region also provides a significant contribution to the conductivity, but plasma transport is now important. The major processes operating in the daytime are quite well understood and can be described in models of the ionosphere and thermosphere that quite reliably predict the distribution of ionospheric conductivity (e.g., Jin et al., 2011; Qian et al., 2014; Ridley et al., 2006; Wang et al., 2011, 2012) as shown in the left panels of Figure 3 for the TIEGCM. Such models are additionally aided by efforts to provide a long-term specification of the spectral content of the solar radiation (Lean et al., 2011) that can be related to the routinely measured solar radio flux at 10.7 cm \( F_{10.7} \) and the MgII line emission with the latter shown to be a better proxy for short term (<81 days) EUV variations (Viereck et al., 2001). In addition, the dramatic changes in the EUV irradiance that accompany solar flares can now be accounted for to first order (Chamberlin et al., 2008) and be included in specifications of ionospheric and thermospheric behavior (e.g., Pawlowski & Ridley, 2008; Qian et al., 2019) during the daytime.

The nighttime conductivity in Figure 3 (right panels) is distributed over a much wider altitude range than during the daytime, although the nighttime magnitudes are smaller. This results in quite different current closure paths, which occupy the E and the F regions for currents associated with the wind systems (Richmond, 1995) and by the externally imposed field-aligned currents from the magnetosphere (Wiltberger et al., 2017). Finally, plasma transport is an increasingly more important factor in determining the conductivity at higher altitudes at night, leading to a tighter coupling between the winds, the plasma motion, and the plasma density (Coley et al., 1994).

The growth of plasma irregularities at low and middle latitudes is also dependent on the underlying ionospheric conductivity variations across the terminator (e.g., Krall et al., 2013; Sultan, 1996) where the coupling between the winds and the plasma motions are changing. During the nighttime, specification of relatively small ionization sources such as scattered solar radiation, starlight and moonlight, and weak X-ray emissions (Titheridge, 2001) becomes more important in determining the plasma density below 200 km altitude. In addition to the specification of ionization sources, the nighttime E region plasma density becomes strongly dependent on the lifetime of the plasma. Thus, the molecular ion species produced by solar ionization during the day decay rapidly, while metallic ion species deposited by meteor ablation remain for many days and may be spatially localized due to neutral wind shears in the region (Haldoupis, 2012). These ionization layers called sporadic E layers have significant effects on the dynamics and stability of the local plasma on magnetic flux lines that pass through them and produce a feedback to the neutral and plasma dynamics that is poorly described due to the lack of information about the spatial extent and dynamics of these layers (Chu et al., 2014). An insightful review of the evolution of sporadic layers and their coupling to the F region is given by Tsunoda (2013).

*With no reliable way to specify the ionization sources at middle and low latitudes and no systematic way to measure the relatively low plasma number density in the nighttime E region, either from the ground or in situ, a significant challenge remains to accurately specify the nighttime ionospheric conductivity.*

In addition to the variable ionization produced from solar radiation, the ionospheric conductivity at high latitudes is strongly dependent on, and at night dominated by, the precipitation of energetic particles from the magnetosphere and the solar wind. This ionization input is deposited over a wide region encompassing the auroral zone and the polar cap. It is difficult to accurately specify since the height varying ionization rate is dependent on the incident particle energy and pitch angle. With particle energies ranging from a few eV to MeV, the impacts are distributed over altitudes from 60 to 250 km. Using a model of the neutral atmosphere, the ionospheric conductivity distribution can be derived as a function of altitude from sophisticated electron transport models and computational schemes to degrade the initial specification in energy and pitch angle (e.g., Lummerzheim & Lilensten, 1994) and from faster methods specifying the ionization rates for energetic electrons (McGranaghan et al., 2015) and protons (Fang et al., 2013).
In many instances the height integrated auroral conductances provide significant insights into the spatial and temporal variability driven by precipitating particles. These parameters can be derived from knowledge of the mean energy of the precipitating particles and the associated total energy flux (e.g., Robinson et al., 1987), and global specification of these parameters can be derived from UV imagery of the auroral emissions (e.g., Coumans et al., 2004). Coincident specifications of auroral conductivity from auroral emissions, incident precipitating particle spectra and more direct measurement of the plasma number density from incoherent scatter radar, have allowed verification of these techniques to derive the conductivity distribution from energetic particle precipitation (Doe et al., 1997; Knight et al., 2018; Lanchester et al., 2009).

However, a methodology to provide a global specification of the altitude distribution of auroral conductivity remains an observational and computational challenge requiring more systematic observations in both hemispheres.

### 2.2.2. Current Flow Paths

The distribution of ionospheric current perpendicular to the magnetic field indicates where the electromagnetic, collisional, and other forces are applied to the plasma and how the electromagnetic energy is dissipated in the I-T. Perpendicular currents are associated internally with the wind forcing and the gravitational and plasma pressure gradient forces (Richmond, 2016), and externally with field-aligned currents applied to the I-T at high latitudes (Siscoe & Maynard, 1991; Siscoe et al., 2013). All these currents flow through the I-T according to the spatial distribution of the source current and the conductivity. The currents flow in closed circuits and must be locally continuous, leading to complex current paths (e.g., Harper, 1977) that are a challenge to describe and visualize, especially when spatial gradients in the conductivity exist. It is extremely difficult to directly measure the current density, and yet a description of the current systems in the I-T is essential to understanding the bulk plasma motion and how it is affected by internally generated and externally imposed currents.

In principle, the current distribution in the I-T can be extracted from coupled models of the ionosphere, thermosphere, and magnetosphere such as the Space Weather Modeling Framework (SWMF) (Tóth et al., 2005), the Open Geospace General Circulation Model (OpenGGCM) (Raeder et al., 2001), and the Coupled Magnetosphere-Ionosphere-Thermosphere (CMIT) model (Wiltberger et al., 2004). In addition, details of the altitude distribution of the current and the associated dissipation rates can be derived from incoherent scatter radar measurements (Thayer, 1998). However, observations of the larger-scale spatial distributions of the winds, the plasma motions, and the conductivity, which are required to identify the important current sources and determine how the current is flowing, are not yet available to validate or constrain such models.

While it is convenient to consider the conductivity of the low and middle latitudes in two layers in the E and F regions, respectively, this convenience should not hide a visual picture that employs field-aligned currents to distribute the current closure paths continuously according to the spatial distribution of the conductivity. For example, wind systems with short vertical wavelengths may be associated with local regional current loops with no influence on the larger scale and more global plasma motion. Conversely, the net current associated with tidal winds in the E region is additionally distributed in the F region by field-aligned currents that connect the two regions. Likewise, the F region wind-driven current is distributed through the E and F regions. In addition, at low and middle latitudes the geomagnetic field links the currents at different geographic locations in each hemisphere. Thus, field-aligned currents flow continuously to ensure that the total current is divergence free everywhere.

In the E region, the currents associated with the winds are dependent on the conductivity distribution, with the challenges associated with its description mentioned above. The tidal wind systems and the large daytime E region conductivity are related to a large-scale current called the Solar-quiet (Sq) current system that circulates anticyclonically in the Northern Hemisphere and cyclonically in the Southern Hemisphere. Near the equator, the horizontal magnetic field and the altitude distribution of the E region conductivity severely limit the vertical current density, and the eastward zonal current is increased in a so-called Cowling channel called the equatorial electrojet (e.g., Yamazaki & Maute, 2016, for a review). The eastward equatorial electrojet current and counter electrojets are easily detected on the ground (e.g., Forbes, 1981; Reddy, 1989). However, in determining the EEJ and CEJ magnitude, there is some ambiguity due to the off equatorial current, which depends on the height variation of the neutral wind and is not well known (e.g., Aiken et al., 2008; Fang et al., 2008).
The left panel of Figure 5 illustrates the daytime horizontal zonal current density of the wind dynamo forcing near 110 km derived from the horizontal wind model (HWM) (Drob et al., 2008) and conductivity distributions based on International Reference Ionosphere (IRI) (Bilitza et al., 2011) and Mass Spectrometer and Incoherent Scatter radar model (MSIS) (Hedin, 1991). The wind dynamo forcing is $\sigma U \times B$ with the conductivity tensor $\sigma$, the neutral wind vector $U$, and the geomagnetic field $B$. Local noon corresponds to approximately 72° magnetic longitude where the magnetic declination is small.

While the eastward dynamo current (red contours) is prominent in the equatorial region, it may also be seen that a westward current flows at some longitudes, representing the imprint of different tidal modes in the wind at this altitude. It should be appreciated that this distribution will change daily according to the tidal and wave variability as described previously. The current distribution also changes with altitude, as shown in the right panel of Figure 5. Here the zonal current is shown as a function of magnetic latitude and altitude to emphasize that the current is not symmetric about the magnetic equator and the current flows both eastward and westward in the altitude between 90 and 120 km. These features demonstrate that local current loops can be driven internally in the dynamo region and that the wind-related currents can be hemispherically asymmetric.

It is important to recognize that the plasma motion perpendicular to the magnetic field is related to the integrated effects of currents that can differ in magnitude and direction at different altitudes along the magnetic field. This is dramatically illustrated in Figure 6 showing the horizontal current, due to the wind dynamo, that is required to maintain a divergence-free system, as a function of magnetic latitude and magnetic longitude at 110 and 150 km. Note that local noon is approximately at 72° magnetic longitude. Here the center of each arrow designates the position, and the most obvious differences appear near 20° magnetic latitude where the direction of the current can be quite different. Note that vectors are not shown within 8° of the magnetic equator since they disproportionately show the equatorial electrojet in the upper panel and field-aligned currents in the lower panel.

Unraveling the contribution of the wind systems to the net current that influences the plasma motion at remote locations remains a significant challenge requiring comprehensive measurements of the plasma and neutral dynamics to be interpreted in a model that self-consistently couples the high- and the low-latitude drivers in both hemispheres.

As described in section 1, during the daytime, the $E$ and $F$ regions may be regarded as parallel resistive paths, and thus, the generally eastward daytime current associated with $E$ region winds in the equatorial region is distributed between these two regions and in the $F$ region drives an upward and westward drift of the plasma (Fejer et al., 2005, 2008). Figure 7 shows the equivalent ionospheric $S_q$ current driven by quiescent tidal wind systems in the $E$ region. This and the following Figures 8 and 9 describe the current patterns derived for a particular universal time. It is well recognized that at other universal times, the same large-scale features will be displayed but with some spatial distortions depending on the offset between...
the geographic and geomagnetic equator. At the solar terminators the spatial gradients in the $E$ region conductivity and the large vertical shear in the horizontal neutral wind produce large local gradients in the horizontal current that are accompanied by field-aligned currents that further change the distribution of the current between the $E$ and $F$ regions. The first-order effect is a rotation in the $Sq$ current so that at dusk, it flows poleward and subsequently westward at higher latitudes. However, the contribution of current driven by winds in the $F$ region becomes equally important in the evening. Near the equator the magnetic field is almost horizontal, and only the zonal wind with respect to the plasma drives a current perpendicular to the magnetic field in the magnetic meridian. During the daytime these currents close meridionally through the highly conducting $E$ region. However, at night and across the terminators, the current flow and plasma flow are adjusted to accommodate the spatial gradients in the conductivity. Eccles et al. (2015) describe different flow paths for this current at the dusk terminator near the equator and their relationship to the so-called prereversal enhancement (PRE) in the vertical drift of the $F$ region plasma (Fejer et al., 1991). Fesen et al. (2000) describe the dramatic changes in the PRE of the vertical plasma drift in the equatorial ionosphere at dusk, which are dependent on changes in the $E$ and $F$ regions conductance across the dusk terminator. Richmond and Fang (2015) identified that the neutral wind in the EIA region contributes most to the PRE magnitude. Rodrigues et al. (2012) and Heelis et al. (2012)

Figure 6. Total daytime horizontal current density $J_{f,\text{hor}}$ [$\mu A/m^2$] at 110 km (top panel) and 150 km (bottom panel) with neutral winds from HWM07 (Drob et al., 2008) for day 181 with $F10.7 = 130$ sfu. Note that the vectors within $\pm 8^\circ$ magnetic latitude are not illustrated since they are dominated by equatorial electrojet and field-aligned current contributions.
have attempted to describe the current distribution for specific configurations of the wind and the conductivity. However, the lack of appropriate measurements of the wind and the plasma motions have limited the advancements that can be made in understanding this important phenomenon. A similar drift phenomena appears at the dawn terminator (Zhang et al., 2015) but has been given less attention to date.

Figure 7. Equivalent current function due to undisturbed wind dynamo (Sq). Simulation with the TIEGCM for 21 March 2010 and $F_{10.7} = 70$ sfu with tidal climatology from Global Scale Wave Model (GSWM) (Hagan et al., 1999) and no high-latitude forcing. Modified reproduction of Figure 49 by Yamazaki and Maute (2016) distribution under the terms of Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

Figure 8. Equivalent current function due to disturbance dynamo. The equivalent current due to the disturbance dynamo is obtained from the difference of two simulations, both with conditions as described in Figure 7, but one with and the other without high-latitude forcing representative of $K_p = 5^-$. Modified reproduction of Figure 49 (Yamazaki & Maute, 2016) distribution under the terms of Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).
At middle latitudes, the steadily inclined magnetic field line in the I-T produces a more easily visualized network of wind-driven current generators in the E and F regions. Again, the influence of the winds in the E and F regions are weighted by the ionospheric conductivity distribution. However, frictional and particle heating at high latitudes adds significant variability to the midlatitude wind system dependent on magnetic activity (Fuller-Rowell et al., 1994). The additional disturbance winds (see section 2.1) blow equatorward and westward at middle latitudes creating a counterclockwise circulation and a clockwise equivalent current as shown in Figure 8 (Yamazaki & Maute, 2016). It is evident that across the dayside, the Sq and disturbance dynamo currents are in opposition. However, the center of circulation of these two systems is different, and the magnitude of each contribution is dependent on the prevailing wind systems and the spatial extent they occupy. Thus, each can dominate at different locations and under different conditions.

While the methodology for incorporating the disturbance dynamo into a description of the plasma dynamics is in place, a description of the spatial distribution of the wind system is dependent on a poorly specified magneto-spheric energy distribution.

The middle and low latitudes may also be influenced by the closure of current systems applied to the I-T at high latitudes. These currents associated with the penetration electric fields do not originate from wind systems in the I-T and are discussed further in the following section. However, for completeness we include a description of three current drivers that influence the plasma dynamics at low and middle latitudes with magnitudes that change in location and time and are controlled by different external factors.

Under quasi-steady state conditions, the neighboring high-latitude Regions 1 and 2 currents close through the auroral zone and the polar cap, effectively shielding the middle and low latitudes from magnetospheric influences. However, temporal variations almost always result in imbalances in these currents with a subsequent closure path through the midlatitude ionosphere associated with so-called prompt penetration plasma drifts (Huang et al., 2007). Figure 9 shows the equivalent current contours that result from this source. Here, it is apparent that while most of the current circulates through the auroral zone at high latitudes, the daytime low latitude I-T experiences a current that circulates in a manner similar to that seen in the Sq current in Figure 7. The current associated with this high-latitude driver can be substantial, and the associated plasma motions can significantly influence the spatial distribution of the plasma.
plasma density (Balan et al., 2009). On the dayside, significant upward drifts are observed, and large east-west gradients in the vertical plasma drift are associated with the requirement for the current flow across the terminators to be continuous (Basu et al., 2007).

Near the equator empirical models of the prompt penetration drifts and those attributable to the disturbance dynamo have been assembled (Fejer & Scherliess, 1997). However, at middle latitudes, the plasma drift patterns often overlap with similar influences from the disturbance winds and prompt penetration effect. These two influences have quite different temporal development with respect to storm-time epoch, but each may evolve over time scales of a few hours, making their contributions at any given time difficult to identify.

TADs represent an additional contributor to the current systems and plasma motions that are fundamental to understanding the stability of the I-T. As described in section 2.1, TADs originate from impulsive energy sources applied internally and externally to the I-T. The resulting change in the atmospheric pressure produces waves in the neutral wind that propagate away from the source and are modified by the prevailing larger scale wind system. The waves that are present in the conducting region of the ionosphere drive currents perpendicular to the magnetic field and plasma motions parallel to the magnetic field.

The associated changes in the plasma density appear as traveling ionospheric disturbances (TIDs). The observation of TIDs that appear at magnetically conjugate locations (Otsuka et al., 2004) suggests that many such signatures are associated with electrodynamic plasma motion perpendicular to the magnetic field produced by the dynamo action of a TAD. It is also possible that TAD disturbances can propagate into the lower $F$ region where a plasma density variation may be directly driven by motion parallel to the magnetic field. Alternatively, at middle latitudes, the atmospheric wave can produce field-aligned motions through ion-neutral collisions and subsequently change the flux-tube integrated Pedersen conductivity that renders the region unstable to the currents driven by the larger scale winds (Perkins, 1973; Zhou & Mathew, 2006). The relationships between TAD and TID are further complicated by TAD-induced changes in the ionospheric density that once produced may be preserved during the night in the absence of a neutral perturbation.

Figure 10. Schematic of field-aligned and perpendicular current to the magnetic field in the I-T system with the perimeter latitude at the equator. Cusp, Region 1, Region 2, and substorm currents originate in the magnetosphere and drive the high-latitude plasma convection shown by the dashed black curves. A penetration current, shown in yellow, arises when the Regions 1 and 2 currents are regionally unbalanced. The disturbance current, shown in blue driven by winds produced by high-latitude energy inputs, is differentiated from the Sq current, shown in red, driven primarily by winds in the $E$ region. Question marks indicate regions where the current path is poorly specified being dependent on the wind and conductivity distribution.
2.3. To Understand the Impacts of Coupling to the Magnetosphere

At high latitudes the I-T is connected to the magnetosphere through a system of field-aligned currents (Green et al., 2009) and Alfvén waves (Keiling et al., 2003) that exchange electromagnetic energy through precipitating energetic particles carrying some of the field-aligned current and providing heat and ionization to the I-T (Akasofu, 1981) and through the transport of plasma between the ionosphere and the magnetosphere affecting the dynamics and the plasma distributions in each region (Welling et al., 2015).

Figure 10 schematically illustrates the major current systems that influence the dynamics of the I-T (Anderson et al., 2008; Carter et al., 2016; Iijima & Potemra, 1978). On the dayside across local noon, so-called cusp currents, named for their source at the magnetospheric cusps, provide the driver for ionospheric convection that is dependent on the orientation of the IMF and reflects the temporal variability with which closed magnetic flux is converted to open magnetic flux at merging sites on the dayside magnetopause (Stauning et al., 2010). These currents also feed the so-called northward Bz (NBZ) current system when the IMF is northward (Iijima & Shibaji, 1987). So-called Region 1 currents originate near the magnetopause and reflect the drivers for antisunward convection across the polar cap (Siscoe et al., 2013). Region 2 currents originate inside the magnetosphere in response to the changing pressure in the sunward flow of plasma (Tanaka, 1995). Finally, the usually quasi-uniform current that flows across the tail of the magnetosphere can be diverted through the ionosphere during a substorm (Kepko et al., 2015) producing enhanced flows and a reconfiguration of the convection pattern on the nightside. The field-aligned currents shown here are largely closed across the auroral zone and the polar cap where they are associated with the ionospheric convection pattern shown by the black dashed curves. However, the Regions 1 and 2 currents may be regionally imbalanced, and in this case, a closure path, primarily through the dayside plasmasphere, constitutes the so-called penetration field.

We have seen that currents are driven by neutral winds due to tides and waves in the I-T, which are labeled Sq. Additionally, the wind system in the I-T is also modified by high-latitude frictional heating and particle heating, which can modify the current in the low and middle latitudes, and is labeled the disturbance dynamo. Figure 10 also shows regions marked by question marks that emphasize the uncertainties in how much penetration current flows across the nightside, the local time extent of the disturbance field, and the latitude extensions of the disturbance and penetration fields that determine their combined effects.

Typically, numerical models of the I-T response to magnetospheric drivers utilize empirical models to specify both particle precipitation and plasma convection representative of a field-aligned current distribution (e.g., Codrescu et al., 2012; Richmond & Maute, 2013). Some work has been done in which the field-aligned current density is specified rather than the plasma motions (e.g., Marsal et al., 2012). Work toward a self-consistent coupling between the I-T, the magnetosphere, and the solar wind is also advancing. In the following, we discuss the challenges in these areas.

2.3.1. Describing the Magnetospheric Drivers

Solid advances in our understanding of the I-T response to magnetospheric drivers have been obtained by utilizing models of the precipitating particle distribution at high latitudes. In many cases the precipitating particle distribution is characterized by an average energy and energy flux (Fuller-Rowell & Evans, 1987), and in other cases, a more sophisticated division of the spectral characteristics is made (Mitchell et al., 2013; Newell et al., 2010). By reconstructing the distribution function from these descriptions, the ionization and heating rates in the atmosphere can be derived as described earlier. While these models limit the description of the particle spectral distribution functions, there is no reason in principle that a model could not describe the input particle distribution functions in more detail (McIntosh & Anderson, 2014). For example, advances have been made by describing separately the characteristics of the discrete, diffuse, and cusp precipitation, which have been used in regional high-resolution models to more completely examine the I-T response to M-I coupling (e.g., Brinkman et al., 2016). In a similar way the response of the I-T to storm time and substorm energy inputs has also been conducted (Clausen et al., 2014). Such representations of auroral precipitation are also shown to be well captured in auroral emissions (Coumans et al., 2002; Oyama et al., 2017). However, individual observations of auroral precipitation and auroral emissions from the ground and from space show significant spatial and temporal variations (e.g., Gillies et al., 2014; Kataoka et al., 2015; Zhang & Paxton, 2008) that are not captured in the presently existing empirical models or specified by a parametric description.
A continuing challenge remains to describe the spatial structure, the temporal variability, and the energy spectrum of auroral precipitation to properly describe its effects in the I-T. This is particularly important at smaller scales and during times of geomagnetic activity, when higher particle fluxes are more dynamically distributed across the auroral zone (Akasofu, 2017). It is also a particular challenge during periods of northward IMF conditions when the aurora forms at high latitudes and migrates across the region traditionally referred to as the polar cap (Cumnock, 2005; Milan et al., 2010).

Similar challenges exist for the specification of the plasma drift in the high-latitude I-T. Foundational empirical models still receive wide use (e.g., Weimer, 2005) and have been extended with evermore data available from ground-based radars (Cousins & Shepherd, 2010). The challenges are now to more fully distinguish hemispheric differences in the patterns and the factors that determine the deviations from a large-scale average model specification (e.g., Grocott et al., 2017). During times of southward IMF, the spatial asymmetry in the convection pattern produced by the \( B_y \) component of the IMF is opposite in each hemisphere. However, the effects of dipole tilt, ionospheric conductivity gradients, and the magnitude of the geomagnetic field itself are all contributing to the hemispheric differences in the ionospheric convection. Differences in the magnetic field strength may also produce differences in the configuration of the plasma flows in each hemisphere (Förster & Haaland, 2015). Differences in the maximum potential drop across the polar regions (Pettigrew et al., 2010) emphasize the importance of considering field-aligned potential drops that moderate the current delivered to the ionosphere from the magnetosphere.

Recently, simulations of the solar wind interaction with the magnetosphere (Peng et al., 2010) and observations of the field-aligned current systems at high latitude (Laundal et al., 2018) suggest that the IMF \( B_y \) component also influences the orientation of plasma flow features and the convection reversal boundary. Apart from well-documented and modeled seasonal variations (Shue et al., 2005), it is also likely that emerging features in the plasma convection and field-aligned currents that are hemispherically asymmetric also appear in the particle precipitation patterns. Early observations by Stenbaek-Nielsen et al. (1973), for example, suggest that hemispheric differences in the magnetic field strength are responsible for the different particle precipitation intensity at conjugate locations.

Presently, there are no large-scale models of the plasma convection, field-aligned current density distribution, or particle precipitation that are derived from a contemporaneous data set that ensures their self-consistency. Thus, there are often physical inconsistencies between the models of ion convection and particle precipitation produced largely by the separate observations and/or parametrization used to construct them (Sheng et al., 2019). Overcoming this observational challenge will represent a first-order advance in our ability to model the behavior of the I-T at high latitudes in a consistent manner.

Next generation advances in constructing models of precipitation, convection, and field-aligned currents should also take into account the dynamics of the boundaries (Landry & Anderson, 2018; Sotirelis & Newell, 2000) and use simultaneous observations of precipitation and convection to ensure that the relative locations of systematic convection and precipitation features are preserved. Increasingly available are plasma drift measurements from ground-based radars (Bristow & Spaleta, 2013), ground-based magnetometer measurements (Waters et al., 2015) at high latitudes, and observations of magnetic field perturbations from satellite constellations (Coxon et al., 2016; He et al., 2012). All available data sources permit assimilation of energetic auroral particles, electric field, and field-aligned currents to be employed to derive instantaneous distributions of the electric potential and auroral particle precipitation at high latitudes with a temporal cadence of a few minutes (Lu, 2017). Convection patterns derived in this way are able to describe the quite complex distributions of the electric potential and precipitation but are so far only available for case studies. Therefore, empirical models can be expected to be widely used until more effective global specifications of the drivers are available.

**Advancement of empirical models of high-latitude plasma convection, particle precipitation, and field-aligned current density should utilize contemporaneous observations of the key parameters with the best available spatial coverage in both hemispheres taken over long enough periods to expose dependencies on magnetospheric and interplanetary parameters.**

### 2.3.2. Temporal Variations

At the largest spatial scales the temporal evolution in models of the high-latitude drivers of the I-T will be specified by the temporal variation of geomagnetic or solar wind variables such as the auroral electrojet...
index, the interplanetary electric field, and the interplanetary magnetic field orientation. These variables will determine the location and intensity of the energetic particle flux, the field-aligned current, and major plasma convection features such as the convection reversal boundary.

It is well recognized that the solar wind drivers operate over a wide range of temporal scales ranging from minutes to many hours. For scales greater than about 1 min (Song et al., 2009) the influence of these temporal variations on the large-scale behavior of the I-T is determined by imposing a time sequence of distributions for the electrostatic potential and the energetic particle precipitation characteristics. However, whether derived from models or from data assimilation, this approach may impose unrealistic conditions on the I-T. For example, time successive realizations of the potential distribution from an empirical model can relocate the convection reversal boundary to lower latitudes at dusk, placing a location that was previously on closed magnetic flux lines in the auroral zone, on open magnetic flux lines in the polar cap. However, physically, we expect the plasma previously in the auroral zone to convect through the ionospheric footprint of a reconnection region on the dayside to appear in the polar cap.

The expanding/contracting polar cap model (Cowley & Lockwood, 1992) has been put forward to address this issue for temporal changes associated with southward oriented IMF conditions. However, the methodology by which such a model can be reconciled with observations, or utilized in data assimilation procedures, is quite primitive (Lockwood & Morley, 2004) and not extensively employed in numerical models of the I-T behavior. Work in this area suggests that, during periods of southward IMF, the field-aligned current distribution imposed on the ionosphere may be divided into dayside and nightside distributions that are each controlled by different conditions (Grocott et al., 2010; Milan, 2013; Sotirelis et al., 2017). We might expect that the increasing spatial density and temporal cadence over which measurements are made in the future will lead to higher fidelity with which evolving features on the dayside and the nightside can be described.

Applying temporally evolving magnetospheric drivers to the I-T in a physically realistic way requires further advances in data assimilation and in the coupling of the solar wind-magnetosphere interaction to the ionosphere and thermosphere.

### 2.3.3. Northward IMF

Generally regarded as the quiet time state of the I-T, the behavior of the I-T at high latitudes during periods of dominantly northward IMF is not as well studied as the behavior during moderate and disturbed conditions, which prevail when the IMF is directed southward and the I-T experiences large energy inputs associated with magnetic storms and substorms. Nevertheless, periods of dominantly northward IMF result in configurations of high-latitude energy inputs from the magnetosphere that are significantly different from those seen during weakly northward or southward IMF, and thus, it is important to understand the response of the I-T during such periods.

The most reproducible characteristic of the convection is a region of sunward flow at highest latitudes (Bhattarai et al., 2012; Cumnock et al., 1995; Huang et al., 2000) that is most frequently associated with the so-called northward Bz (NBZ) field-aligned current system (Vennerstrom et al., 2002). Additionally, a region of sunward flow persists at the lowest latitudes as it does when the IMF is directed southward. Observations of plasma motions and equivalent electric field at high latitudes (Cumnock et al., 1995; Heppner, 1977) have also noted the presence of highly structured flows on the nightside and particularly in the dark hemisphere when the IMF is northward. At highest latitudes filamentary field-aligned currents are pervasive (Lühr et al., 2016), and energetic particle precipitation features such as Sun-aligned arcs occupy the region normally devoid of such features during southward IMF (Frey, 2007). Generally, the spatial variability in the high-latitude convection features during periods of northward IMF does not allow the features to be well represented in large-scale averages of the electric potential (Weimer, 2005), and while a conceptual model for the temporal evolution of convection features exists for a southward IMF orientation, this is not the case for northward IMF. Similarly model representations of the auroral precipitation or hemispheric power do not correctly describe the spatial distribution of the input. Thus, at this time the large-scale electrical connections between the I-T and the magnetosphere during times of northward IMF are poorly understood, and the methodologies by which the momentum transfer and heating rates to the I-T should be specified are not established.

While the magnetospheric forcing of the I-T system is less intense and confined to higher latitudes during periods of northward IMF, it is necessary to describe the temporally and spatially evolving patterns of plasma motions...
and particle precipitation and to develop an approach to modeling it in order to understand the response of the I-T to this condition that is present almost half the time.

2.3.4. Multiple Scales

In large-scale global models of the coupled ionosphere and thermosphere at high latitudes, the specification of large-scale patterns of convection and precipitation usually underestimates the observed temperature in the thermosphere (Fesen et al., 1997). This has led to the consideration of the impacts of smaller-scale features in the plasma drift and the particle precipitation that are always present in the observations. Various descriptions of the plasma velocity structure have been given (Cousins & Shepherd, 2012; Johnson & Heelis, 2005) and models of the spatial distribution of the plasma velocity structure and its intensity have been developed (Matsuo & Richmond, 2008). The associated structure in the particle precipitation has received less attention, but Deng et al. (2009) have modeled the effects of velocity variability in a global description of auroral precipitation to demonstrate the significant effects on the thermospheric density and temperature.

Within a large-scale model, the effects of small-scale structures that appear at the subgrid level are accounted for in parametric increases in the frictional heating rate and the particle heating rate. However, advances in computational capabilities allow adjustable time steps and adaptable grids in space to incorporate smaller scale convection features more explicitly (Zhu et al., 2018). Advances in defining multiscale features in the ionospheric plasma motions and the field-aligned currents associated with them are now being made (Chen & Heelis, 2018; Gabrielse et al., 2018; McGranaghan et al., 2017). At the present time our descriptions of the spatial and temporal variations in the observed plasma velocity structure are rather primitive, and observations of auroral emissions that are associated with spatial structures with scales of 100 to 500 km show that many times, these features are moving (e.g., Kim et al., 2017; Zou et al., 2014, and references therein). Thus, a more advanced assessment of the impact of velocity and precipitation structures on the I-T will require the spatial and temporal ambiguity to be systematically resolved for different scales.

Ultimately, a self-consistent picture of the evolving precipitating particle distribution, the field-aligned current, and the plasma convection that results from the coupled magnetosphere and I-T is required. These self-consistent patterns have to be validated with observations over a range of spatial scales to establish how the global I-T response is dependent on the duration of drivers at different scale sizes.

2.3.5. Magnetosphere-Ionosphere interactions

As we examine the linkages between the I-T behavior and the magnetospheric drivers, we discover that the neutral wind and the ionospheric conductivity are also key active elements influencing the flow of electromagnetic and particle energy between the magnetosphere and the I-T. This active feedback between the magnetosphere and the I-T is established by the following connections:

1. between the field-aligned current and the particle precipitation that is associated with it,
2. between the particle precipitation and the ionospheric density and conductivity that it produces,
3. between the currents perpendicular to the magnetic field and the plasma and neutral gas motions, and
4. between the field-aligned currents induced by the changes in the neutral wind and ionospheric conductivity.

There exist significant challenges to understand each of these interactions.

As mentioned in section 1, field-aligned currents at high latitudes provide the channels for electromagnetic energy flow between the ionosphere and the magnetosphere. Precipitating energetic electrons are frequently the carriers of upward field-aligned currents, leading to the observation of adjacent regions of energetic particle inputs and electromagnetic energy inputs to the I-T (Evans et al., 1977). However, there also exist regions where the spatial separation of particle and electromagnetic energy inputs is not well preserved (Deng et al., 2015). This suggests that the magnetospheric mechanism that produces the field-aligned current and the precipitating particle flux requires further investigation (Borovsky & Valdiana, 2018) despite our ability to establish relationships between the electron energy and the field-aligned current intensity in some regions (Cran-McGreehin & Wright, 2005; Knight, 1973). In the I-T, energetic particle precipitation changes the ionospheric conductance, which is further modified by the transport of plasma associated with the closure of the field-aligned current. In regions of upward field-aligned current, carried by precipitating electrons, the ionospheric conductance is enhanced (Johnson & Wing, 2015), while in downward field-aligned current regions, the current carriers in the ionosphere are assumed to be thermal...
The electromagnetic energy flux imposed by the field-aligned currents is dissipated as momentum transfer to the neutral gas (e.g., Thayer et al., 1995) and frictional heating (Strangeway, 2012). The resulting plasma motions redistribute the plasma horizontally and vertically over a wide range of altitudes (Fuller-Rowell et al., 1994), thus changing the ionospheric conductivity. At any location in the I-T, the gradient in the horizontal current, determined principally by the conductivity and the relative velocity of the plasma and the neutral gas, must be consistent with the field-aligned current, which delivers the electromagnetic energy flux and organizes the particle precipitation.

**However, the role of the I-T in controlling the partitioning of particle and electromagnetic energy inputs through changes in ionospheric conductivity is not well understood.**

Finally, we emphasize the large range of time scales over which the plasma and the neutral gas are affected by coupling to the magnetosphere. While the plasma motion may respond rapidly to changes in the electromagnetic energy inputs, the bulk flow of the neutral gas responds much more slowly and tends to integrate the momentum inputs from the plasma. Thus, the neutral atmosphere can act as a “fly-wheel” to provide an underlying modification to the current associated with the neutral wind and thus the field-aligned currents that deliver the electromagnetic energy (Deng et al., 1993). The extent to which changes in the neutral wind influence the field-aligned current distribution at high latitudes cannot be evaluated without simultaneous observations of the temporally evolving plasma and neutral density, plasma motion, and neutral circulation.

At the shortest time scales of a few minutes or less, electromagnetic energy is transferred between the ionosphere and the magnetosphere through propagating Alfvén waves (Lysak & Song, 2006; Pakhotin et al., 2018). This energy input provides structured electric fields that can be of similar magnitude to those electrostatic features described earlier (Keiling et al., 2019). They do not affect the bulk circulation of the plasma in the ionosphere but selectively heat the I-T in the upper region (Lühr et al., 2004) due to the resonant trapping of Alfvén waves by the plasma density gradient in the topside ionosphere (e.g., Lotko & Zhang, 2018; Verkhoglyadova et al., 2018). Alfvénic energy sources need to be identified in a systematic way to be incorporated into models of the I-T behavior since they represent heat sources that redistribute the plasma and the neutral gas horizontally and vertically.

The spatially and temporally evolving state of the I-T presents a changing conductivity and current distribution that is an active element in the electrodynamic coupling between the ionosphere and the magnetosphere. In outer magnetosphere modeling, these changes are reflected in changes in the ionospheric conductance and changes in the lower-latitude electric potential distribution that constitute a boundary condition (Merkin & Lyon, 2010). Changes in these conditions are seen to affect the cross polar-cap potential by as much as 10% and change the pressure distribution in the magnetotail. In the inner magnetosphere, changes in the ring-current plasma pressure determine the Region 2 current distribution (Tofollo et al., 2003). The ionospheric conductivity then determines the distribution of the Regions 1 and 2 currents and the resulting plasma motions in the I-T and the magnetosphere.

If the Region 2 current is not closed through the auroral zone, where the ionospheric conductivity is maintained by particle precipitation, then the current may flow partially through the I-T in a loop at lower latitudes as shown in Figure 10. The concept of penetration electric fields has been well studied and observed near the equator, where enhanced vertical plasma drifts associated with a zonal current appear during magnetically active periods (e.g., Huang et al., 2010) and at subauroral latitudes in the dusk sector, where so-called subauroral polarization streams (SAPs) describe the zonal plasma drift associated with the meridional current (Foster & Burke, 2002; Foster & Vo, 2002). The temporal development of the flow paths for the currents originating in the outer and inner magnetosphere does not occur in isolation from those generated by the disturbance dynamo and the tidal oscillations in the I-T, since they too can modify the ionospheric conductivity. Thus, the development of coupled models of the ionosphere and the magnetosphere that consider all these sources has led to many further insights into how the ionospheric conductivity and the plasma and neutral motions in the I-T modify the Region 2 current distribution originating in the inner magnetosphere. Large-scale penetration of currents inside the plasmasphere have been demonstrated (Huba & Szeykin, 2014), while smaller-scale features such as subauroral polarization streams are sensitive to the initial conditions (Lin et al., 2019; Raeder et al., 2016). But all the models are dependent on a specification...
of the electric potential around the polar cap boundary and on the methodology by which the auroral conductance is specified.

In addition to controlling the current flow paths perpendicular to the magnetic field, the motion of the ionospheric plasma parallel to the magnetic field by virtue of the frictional heating and particle heating adds another element of active feedback between the I-T and the magnetosphere. It is well established that the cycle of frictional and particle heating, imposed on the plasma as it circulates at high latitudes, produces a net outward flux of ionospheric plasma to the magnetosphere (Yau et al., 1988). The effects in the magnetosphere appear in global simulations (Wiltberger, 2015) and are reflected in changes in the cross-polar cap potential (Winglee et al., 2002) and the intensity of flow bursts (Moore et al., 2013). These connections between the I-T and the magnetosphere, which change the distribution of field-aligned currents, will have significant temporal delays associated with the circulation of ionospheric ions through the magnetosphere that will further complicate identification of the relative roles of frictional heating, particle precipitation, and plasma transport on the ionospheric conductance and the field-aligned current. Nevertheless, modeling initiatives have produced interesting insights at the largest scales (Wang et al., 2008), and the existence of field-aligned potential drops, which allow some independence between the dynamics of the outer magnetosphere and the I-T, is now being considered (Connor et al., 2016).

The connections to the magnetosphere and the corresponding response of the I-T are clearly influenced by electromagnetic energy flow, precipitating particles, and thermal plasma outflow. The challenge of incorporating these processes into computational models is matched by the acquisition of observations in the I-T that specify how these connecting properties vary in space and time within the I-T.

2.4. To Understand How Structures in the Plasma and Neutral Density Are Formed and Evolve

Neutral density variations are influenced by the plasma density distribution through ion-neutral coupling and by compositional changes caused by varying neutral dynamics and chemistry. The plasma density itself depends among others on the composition and dynamics of the neutral gas, making the I-T a strongly interconnected system. It is important to recognize that coupling effects are present over a wide range of spatial and temporal scales at every location and that variations at specific temporal and spatial scales cannot be properly understood without consideration of the entire spectrum of variations.

To describe and understand the structures in the neutrals and the plasma requires the simultaneous consideration of temporal variations from days to minutes and spatial variations from the planetary radius to a few kilometers.

The schematic in Figure 11 illustrates how plasma and neutral structures on various scales are embedded in the I-T system and coupled to different regions. Both large- and small-scale features exist throughout the I-T. At high latitudes for example, the large-scale convection pattern produces large-scale storm-enhanced densities (SEDs) and depletions (midlatitude trough) in the plasma density bounded by regions of auroral precipitation that are particularly apparent during magnetically active periods. Spatial and temporal variations in the plasma flows and discrete auroral emissions produce smaller-scale variations in the neutral density and plasma patches that have unstable gradients at their edges. At low latitudes, large-scale plasma motions produce the EIA, and accompanying ion-neutral interactions are reflected in a corresponding equatorial density anomaly (EDA). Smaller-scale plasma features arise at night due to the unstable configuration of the entire F layer. At middle latitudes the propagation of atmospheric disturbances is visible in medium- and large-scale disturbances in the atmosphere and the ionosphere, and the plasma disturbances are themselves unstable producing smaller-scale variations.

In the following, we will describe observed phenomena spanning these different scales to highlight the gaps in our understanding and the challenges to closing them.

2.4.1. Equatorial Plasma Structures

Plasma irregularities resulting from instabilities occur over a wide range of spatial scales from centimeters to hundreds of kilometers (Basu et al., 1978; Huang, 2017; Stolle et al., 2006; Su et al., 2001; Xiong et al., 2012). At the largest scales, near the magnetic equator, an interchange instability, analogous to the Rayleigh-Taylor fluid instability, can operate in the bottomside F region where the plasma density increases rapidly with altitude. In the evening, the F region is uplifted, and the E region quickly diminishes leading to large vertical plasma density gradients and the generation of equatorial plasma bubbles (EPBs) that rise into the F
region with the density 1–3 orders of magnitude smaller than the ambient density (see reviews by Abdu & Kherani, 2011; Balan et al., 2018). EPBs are plasma-depleted flux tubes with a large range of spatial scales from less than 10 km to more than 1,000 km (Eastes et al., 2019; Hei et al., 2005) that evolve as depletion wedges from original bottomside perturbations (seeds) in ways that are not fully understood. Hysell and Kudeki (2004) describe a collisional shear instability in the bottomside $F$ region that rapidly produces irregularities with scale sizes near 100 km. Small- and large-scale features may be initiated by gravity waves (Abdu et al., 2009) and by medium scale waves, respectively, in the neutral atmosphere (Tsunoda, 2010). The subsequent growth of equatorial plasma perturbations from the bottomside $F$ region through the $F$ region peak and into the topside ionosphere is controlled by the growth conditions (Krall et al., 2010a), which are dependent on the ion-neutral collision frequency, the neutral wind, and the vertical plasma motion. Observations suggest an increased appearance of plasma structure in the topside equatorial ionosphere when the magnetic meridian is aligned with the solar terminator (Tsunoda, 1985) and when the PRE vertical drift is most intense (Fejer et al., 1999; Huang & Hairston, 2015). However, the appearance of plasma structures is highly dependent on longitude (McClure et al., 1998; Rodriguez-Zuluaga et al., 2019) and is not precisely organized by magnetic declination, suggesting that the seeds are themselves longitudinally and seasonally dependent.

During times of moderate to high solar activity, irregularities are seen soon after sunset and decay uniformly throughout the night, consistent with the appearance of a PRE in the vertical plasma drift. The influence of PWs on the PRE (Abdu et al., 2006; Bertoni et al., 2011; Jose et al., 2017) may also be revealed in the increased occurrence of equatorial plasma structures during periods of strong PW activity such as SSWs (e.g., Smith et al., 2016). However, at low solar activity the vertical plasma motion (Stoneback et al., 2011) and the appearance of equatorial plasma structures (Smith & Heelis, 2017) do not conform to these usual expectations. Under these conditions, irregularities are distributed more uniformly throughout the night (Zhan et al., 2018), leading to questions concerning when a bottomside perturbation is produced and the relative contributions of gravity, neutral winds, and plasma motions to the growth rate (Huba & Krall, 2013; Zhan & Rodrigues, 2018).

It is a significant challenge to identify the seeds for development of equatorial plasma perturbations and reveal the importance of various contributions to the instability growth rate of equatorial irregularities. To make advances, simultaneous measurements of the height variation in plasma and neutral density and velocity in the bottomside $F$ region through the lifetime of events are required.
The presence of large-scale plasma density gradients, which define the edges of plasma bubbles or depletions, provides the seed for smaller-scale features that are ultimately the cause of radio scintillation (Basu & Aarons, 1977). These smaller-scale irregularities are produced by kinetic types of plasma wave instabilities (e.g., Gary, 1984; Ossakow, 1981). In addition to the small-scale kinetic instabilities in the $F$ region, the equatorial electrojet, described earlier, also provides a source of free energy for a family of plasma irregularities described by Farley (2009), with spatial scales that vary from a few meters to a few kilometers. They have been studied extensively from the coherent radar backscatter that they produce. While the prevailing conditions that produce the structures are reasonably well understood, the combination of mechanisms that produce the observed radar backscatter intensities remains to be fully investigated (Young et al., 2017).

### 2.4.2. Midlatitude Plasma Structures

The midlatitude ionosphere is largely devoid of structured ionization sources and of large or structured electric fields from magnetospheric sources. However, contrary to expectations, there exists significant structure in the plasma density particularly during times of magnetic activity. In some instances, structures originating in the bottomside $F$ region near the equator can propagate to high apex altitudes to be seen in the topside ionosphere at midlatitudes (Earle et al., 2006; Pfaff et al., 2008). Most plasma structures at middle latitudes are largely due to the presence of large- and medium-scale perturbations that propagate through the region and are produced by neutral winds in the lower thermosphere that transport plasma parallel and perpendicular to the magnetic field.

So-called TIDs are observed at $F$ region heights with horizontal wavelengths of hundreds to thousands of km during the daytime and nighttime. TIDs are a major source of geolocation error, which warrants understanding their generation and evolution. In general, TIDs are distinguished by the source generating the TADs. Large-scale TIDs (LSTIDs) are due to aurora activity (Jonah et al., 2018), while medium-scale TIDs (MSTIDs) are associated with lower atmospheric disturbances (e.g., Crowley & Azeem, 2018) (see also section 2.1). The TADs propagate away from the source with the limiting speed increasing with height (Bruinsma & Forbes, 2009; Hickey et al., 2009) and largely horizontal neutral motions move plasma parallel and perpendicular to the magnetic field directly through collisions and indirectly through dynamo generated currents, respectively (Huba et al., 2015; Shiokawa et al., 2005) (see also section 2.2).

Once formed, the plasma structures described above are themselves unstable and also provide the seed for plasma irregularities with smaller-scale sizes. Informative reviews of the processes at work are given in an introduction by Hysell et al. (2016) and by Shalimov (2014). A better understanding of TID generation, morphology, and evolution is needed in order to assess their role in the generation of plasma density gradients, plasma irregularity development, and associated impacts on global navigation systems.

LSTIDs are generated by enhanced aurora activity, but we do not understand the generation mechanism to the level of forecasting. Hunsucker (1982) suggested that the momentum forcing and frictional heating from the plasma generate these large-scale neutral waves that can be important in the I-T system to transfer energy from the high to lower latitudes. LSTID occurrence increases during geomagnetic storms (e.g., Tsugawa et al., 2004), but LSTIDs are also seen without substorm occurrence (e.g., Lyons et al., 2019). In simulations turbulent electron heating associated with the microscopic Farley-Buneman plasma instability (FBI) (Buneman, 1963; Farley, 1963) is associated with modified TADs through increased Joule heating (Wiltberger et al., 2017; Liu et al., 2016, 2018). There are still major gaps in our understanding of MSTID generation. Observations at conjugate points support the notion that the plasma moves up and down under the influence of electrodynamic forces (e.g., Otsuka et al., 2004; Saito et al., 1995). It is also suggested that the angled TID wavefront with respect to the field line (e.g., Martinis et al., 2010) results from filtering of the waves as they propagate from high latitudes (Kelly, 2011). The action of waves in the $E$ region producing sporadic layers (Cosgrove & Tsunoda, 2003; Yokoyama & Hysell, 2010) has also been suggested to increase the otherwise low instability growth rate (e.g., Makela & Otsuka, 2012). A variety of observational techniques are employed to detect MSTIDs (e.g., Kotake et al., 2007; Martinis et al., 2010; Park et al., 2014; Shiokawa et al., 2003) with variations in occurrence climatologies (Hernández-Pajares et al., 2006; Otsuka et al., 2013) that depend on season and solar activity (Arras et al., 2008; Kotake et al., 2006).

Often, observations are localized, leading to the identification of isolated plasma density enhancements called blobs. These features may occur on the same magnetic flux line as a plasma depletion (e.g., Huang et al., 2014; Park et al., 2008; Yokoyama et al., 2007) consistent with results from models with the
appropriate plasma dynamics (Krall, Huba, Joyce, et al., 2010; Krall et al., 2010b). However, blobs are observed without the occurrence of bubbles especially during solar minimum, and a connection between MSTIDs and blobs is suggested by observations with similar occurrence distributions (e.g., Choi et al., 2012; Haaser et al., 2012; Kil & Paxton, 2017; Miller et al., 2014).

We need an improved general understanding of the generation mechanisms for TIDs, bubbles and blobs and their evolution, morphology, interconnections, and association with plasma irregularity generation. This will require comprehensive collocated observations of the plasma and neutral dynamics that include the wave vector identification.

2.4.3. Plasma Density Enhancements

During geomagnetic active periods, an increase in electron density, named SED (Foster, 1993), is observed at middle latitudes in the subauroral region where the plasma density distribution is most strongly influenced by the residence time in sunlight of a convecting volume of plasma. As mentioned previously, the convective paths of the plasma with respect to the Sun are determined by the combination of plasma drifts resulting from neutral winds and magnetospheric currents and atmospheric corotation. Thus, at certain longitudes, local times, and seasons, the plasma may dwell for longer time in sunlight than in neighboring regions (Sojka et al., 2012). Foster (1993) suggested that the expanded high-latitude ion convection can transport the high-density solar-produced plasma poleward. In addition to the zonal transport, vertical flow, due to poleward convection that produces regions of reduced loss but continued production, can contribute to the SED (e.g., David et al., 2011; Deng & Ridley, 2006; Heelis et al., 2009). However, penetration electric fields, equatorward neutral winds, and changes in neutral composition may also play a role (Lin et al., 2009; Liu et al., 2015; Lu et al., 2012; Wang et al., 2010), and simultaneous measurements of the plasma and neutral motions and composition are required to discover the relative importance of these different influences on the generation and evolution of SEDs.

Once produced, the enhanced density region may be erratically convected into darkness at higher latitudes, where the evolving convection pattern produces tongues of ionization (TOIs) and polar cap patches that move toward the nightside (David et al., 2011). While the convection plays an important role in the development of TOI from SEDs (e.g., Thomas et al., 2013), TOIs also form during geomagnetic moderate times (Liu et al., 2015), indicating that other factors such as season and longitude may play important roles.

In the polar cap, the distribution of plasma in the $F$ region is highly dependent on the orientation of the IMF, which determines the direction, magnitude, and variability of the plasma transport, and on the solar zenith angle, which determines the local photoionization rate. When the IMF is directed southward, plasma flows antisunward across the polar cap from the dayside cusp to the nightside. Electron precipitation with characteristic magnetosheath energies of a few 100 eV is present in the cusp with fluxes that vary temporally in accord with periods of enhanced reconnection at the dayside magnetopause (Hosokawa et al., 2016; Skjoveeland et al., 2011). Thus, the plasma flowing antisunward from the cusp is transported from the region of SED, through the region of variable electron precipitation, to form TOIs that are present in the polar cap during times of moderate to strong geomagnetic activity (Liu et al., 2015; Thomas et al., 2013). The temporal variation in the convection speed and cusp ionization rate in the dayside has been suggested as one possible cause for the highly variable plasma density that emerges from the cusp region in so-called plasma patches (Carlson, 2012).

Ionospheric plasma patches are most easily detected when the plasma enters the cusp from the equatorward, sunlit side and exits the cusp on the poleward, dark side. These conditions allow the plasma entering the cusp to have the highest density while retaining any structure that was created in the cusp region itself (David et al., 2016; Sojka et al., 1994). Less-intense enhancements can be caused by soft particle precipitation in the cusp and polar cap (Goodwin et al., 2015; Rodger et al., 1994; Oksavik et al., 2006).

Polar cap patch occurrence and evolution depend on many factors such as convection, universal time, and the offset between the geographic and geomagnetic poles, producing longitudinal and seasonal dependencies with interhemispheric differences (Chartier et al., 2018; Coley & Heelis, 1998; Noja et al., 2013; Spicher et al., 2017). While many studies have found that polar cap patches occur preferentially in the winter hemisphere, it has been pointed out that the definition of a polar patch (at least a factor of 2 increase from the ambient density) may prevent the identification of the same formation mechanisms operating in the summer hemisphere (e.g., Chartier et al., 2018; Noja et al., 2013). Polar cap patches are also associated with


neutral density enhancements (approximately 10% during solar minimum), increased heating (approximately 10% neutral temperature increase), and neutral wind changes that depend on the ambient conditions (e.g., Ma & Schunk, 1997; Schunk & Zhu, 2008; Schunk et al., 2005).

Precipitation also adds to the structure in the polar cap with Sun-aligned arcs reaching from noon to midnight across the polar cap. Sun-aligned arcs are only a few hundreds of km wide, can last for several hours, and are associated with a few keV or less electron precipitation (Robinson & Mende, 1990). These smaller arcs are connected to a flow channel between the upward and downward current region (Maggiolo et al., 2012). Some arcs follow the direction of IMF Bz (Valladares et al., 1994), while others move poleward independent of Bz (Hosokawa et al., 2011). Sun-aligned arcs may directly produce plasma density enhancements that subsequently remain in regions where the solar photoionization rates are low, and where they may be distorted and reoriented by the changing plasma convection pattern (Hosokawa et al., 2011).

The spatial gradients in the plasma density that define the edges of these structures can provide the seed for smaller-scale structure that may be directly produced by externally imposed electric field structure (Kivanç & Heelis, 1999) or may develop from the growth of the gradient drift instability or the velocity shear instability (Basu et al., 1990). The roles of these different mechanisms are not well assessed and may change as the plasma enhancement moves through regions with different convection speeds and directions during a transit from the dayside to the nightside. All these issues and the relative importance of variations in convection and precipitation in producing plasma structure seen in the polar cap remain significant challenges of great importance since they control when and where the smaller-scale plasma structure, producing disruptions and degradation of radio signals, will appear (Clausen et al., 2016; Wang et al., 2016).

An understanding of the plasma and neutral density structures, which include identification of the sources for plasma density enhancements and the subsequent ion-neutral interactions that determine their evolution and reconfiguration at high latitudes, will require multipoint observations of the plasma and neutral parameters with sufficient spatial and temporal resolution to follow the bulk transport properties as well as the formation of smaller scale structures.

2.4.4. Neutral Density Variations by Large-Scale Processes

Large-scale variations in the neutral density result from the overall thermal balance of the planet, from the magnetospheric energy inputs, and from the dynamic coupling with the plasma. Qian and Solomon (2012) and Emmert (2015) provide comprehensive reviews of neutral density variations and the state of the field. Specific reviews describing neutral density variations driven by the solar wind (Prölss, 2011) and other space weather phenomena (Liu et al., 2017) expose the breadth of issues to be considered. In the following, we mention a few areas highlighting the need for better understanding of the coupled I-T system for accurate specification of the neutral density distribution.

Neutral density variations depend strongly on forcing from solar irradiance, the solar wind-magnetosphere interaction, and the coupling to the lower atmosphere. Forecasting neutral density variations therefore strongly depends on the ability to forecast these forcings. Data assimilation methods focusing on neutral density exist but are challenged by the limited amount of in situ neutral density data at altitudes between 200 and 500 km. Several different approaches, employed for quiescent and disturbed times, were applied recently to overcome the data sparseness (e.g., Matsuo et al., 2012, 2013; Mehta & Linares, 2018; Sutton, 2018).

While it is convenient to discuss the neutral wind behavior and the neutral density behavior as separate phenomena, they cannot be understood in isolation. Thus, the previous discussions in sections 2.1 and 2.3 are immediately applicable with the neutral wind variability being among the most important parameters to be considered. At high latitudes the winds are strongly influenced by ion drag and frictional heating from the plasma convective flow, while at lower latitudes the energy and momentum deposited by upward propagating waves from lower altitudes tend to dominate the behavior of the neutral atmosphere. Thus, understanding large-scale variations in the neutral atmosphere density requires that the effect of systematic variations in the wave spectrum from the lower atmosphere be distinguished from similar variations induced by magnetic activity and associated high latitude drivers.

In addition to variations in the neutral density produced by wave motions, it is also important to consider density changes induced by changes in the neutral temperature. In this regard we have described the
effects of eddy diffusion that must be correctly modeled to account for temperature changes at the turbopause. But it has also been shown that the NO has a powerful influence on the thermospheric temperature and can be radically changed by particle precipitation during solar and geomagnetic storms and subsequently transported to lower latitudes (e.g., Maeda et al., 1992; Mlynczak et al., 2003). During the recovery from geomagnetic storms, the physical processes leading to an overcooling of the atmosphere are well understood. Increases in precipitating particle flux enhance the production of NO, which cools the thermosphere by radiative cooling that may be enhanced for hours to days (e.g., Knipp et al., 2017; Lei et al., 2012). This important cooling has been related to indices of the Joule heating input into the I-T system (e.g., Lu et al., 2010), but this provides only a global NO cooling estimate. Global, large-scale patterns of NO can be derived to understand, in a statistical sense, the spatial variability in the cooling (e.g., Flynn et al., 2018). However, specification of the NO cooling distribution during individual events, as is required to improve specification of the neutral density, must take into account the spatial distribution of the particle sources, the transport of the neutral gases, and the spatial distribution of the cooling rate (Mlynczak et al., 2018).

In addition to the propagation of variations from below, there are local processes that change the neutral density distribution in the thermosphere. The daytime equatorial neutral density is enhanced in regions aligned with the EIA, and accordingly, this increase is termed the equatorial mass anomaly (EMA) (e.g., Liu et al., 2005, 2007). Similar enhancements are observed in the neutral temperature called the equatorial temperature anomaly (ETA) (e.g., Raghavarao et al., 1991). These features serve to illustrate the wide range of local processes that influence the neutral density (Emmert, 2015; Liu et al., 2017). These large-scale structures in the ionosphere and thermosphere have significant impacts on satellite drag (e.g., Zhang et al., 2018) and on radio scintillation. Many have been studied in detail, such as field-aligned ion drag (Maruyama et al., 2003; Miyoshi et al., 2011), chemical heating by charge exchange (Fuller-Rowell et al., 1997), and heat transport due to the background wind (Hedin & Mayr, 1973) and tidal propagation (Miyoshi et al., 2011). A few modeling studies have examined the mechanism generating the ETA finding field-aligned ion drag (Hsu et al., 2014; Lei et al., 2012, 2014) and energy transfer from electrons/ions to neutrals (Lei et al., 2014) to be dominant mechanisms. However, any prediction of their magnitude and occurrence will require the identification of the various mechanisms and the conditions under which they are most effective. This can only be accomplished with simultaneous measurements of plasma and neutral density, drift, and temperature taken systematically over all seasons.

### 2.4.5. Neutral Density Enhancements in the Polar Cusp

In the dayside polar cap significant neutral density enhancements are observed regularly (40% of the times) and are of sufficient magnitude to affect satellite orbit determinations (e.g., Huang et al., 2017; Liu et al., 2010; Lühr et al., 2004; Rentz & Lühr, 2008). The enhancements scale with geomagnetic activity but are also observed during geomagnetic quiescent time (e.g., Huang et al., 2017). Recent studies suggest that these enhancements are associated with direct energy input into the polar cap region (e.g., Knipp et al., 2011). The neutral density enhancements can also be accompanied by lower-altitude density reductions (e.g., Brinkman et al., 2016; Clemmons et al., 2008) suggesting an energy input, such as frictional heating or monoenergetic particle precipitation, that is localized in altitude.

There exist some observational evidence, to suggest that some neutral density enhancements are associated with small-scale field-aligned current and Alfvénic waves causing ohmic heat in the plasma that is subsequently transferred to the neutral gas (Liu et al., 2010; Lühr et al., 2004). Lotko and Zhang (2018) examined Alfvénic heating effects in the cusp finding an increase in the $F$ region Joule heating rate, which can be effective in heating the $F$ region neutrals. Cusp neutral density enhancements associated with increased electron temperature, small-scale FAC, and vertical plasma motion are consistent with increased electron precipitation (e.g., Kervaveli & Lühr, 2013; Lühr & Marker, 2013). Other studies suggest that poleward TADs associated with atmospheric gravity waves can lead to neutral density enhancements (e.g., Bruinsma & Forbes, 2007; Lin et al., 2018; Lu et al., 2016). More recently enhanced frictional heating in antisunward moving plasma patches has also been suggested as a production mechanism for neutral density enhancements in the $F$ region (Lin et al., 2018).

To identify the most influential mechanisms in producing neutral density changes on local to global scales, the drivers from the lower atmosphere and from the magnetosphere have to be specified at different spatial scales and integrated in numerical models.
2.4.6. Compositional Changes

Knowing the thermospheric composition is crucial to understand neutral and plasma density variations in the I-T system. Changes in the composition can easily connect different latitude and altitude regions, leading to global effects on time scales from minutes to seasons. A challenge in capturing the composition is that we need to better quantify the energy input into the I-T system and the dynamical and chemical coupling to the mesosphere. An overview of the established and suggested mechanisms affecting neutral and plasma density is given in the introduction by Jones et al. (2018).

The ratio of atomic oxygen to molecular nitrogen is very important to understand the neutral and plasma density variations. An increase in atomic oxygen compared to molecular nitrogen increases the mass density at a particular height through changing the scale height and can also lead to an increase in plasma density through increased ion production. The ratio can be changed by the thermospheric spoon effect (Fuller-Rowell, 1998), which describes the large-scale meridional-vertical circulation associated with the increased solar heating in the summer hemisphere leading to upwelling of N₂ rich air in the polar region, which is transported into the other hemisphere, and downwelling of O rich air in the winter hemisphere. During the transport into the other hemisphere, O and N₂ experience increased diffusive separation when moving equatorward and downward, leading to an increase of the O/N₂ ratio (e.g., Rishbeth et al., 2000). This meridional circulation can be significantly modified by M-I coupling and the associated high-latitude heating (e.g., Burns et al., 1995; Zhang et al., 2004). The changes at a fixed altitude in the neutral composition directly influence the photoproduction and chemical loss of major F region ions. While these are simplified pathways to change composition and modify neutral and plasma density, the exact variations strongly depend on the distribution of solar and auroral heating and the neutral wind, which in general are poorly quantified.

In addition to the large-scale circulation effects, it is known that upward propagating waves and tides from the lower atmosphere can modify the composition and therefore the neutral and plasma density. Longitudinal variations at fixed local times are found in GUVI O/N₂ observations indicative of tidal origin, but there are questions about the mechanism since the eastward movement of major tidal signals is not coherent with the O/N₂ variation (e.g., He et al., 2010; Kil & Paxton, 2011). Similarly, PW signals can be present in the composition and affect the neutral and plasma density (e.g., Forbes, Zhang, et al., 2018; Forbes, Maute, et al., 2018; Sassi et al., 2016).

On smaller spatial scales, Qian et al. (2009) argued that gravity wave breaking introduced seasonally varying eddy mixing, which they captured by introducing eddy diffusion parameters in a model to reproduce the observed neutral density changes. As discussed in section 2.1 and shown by the multitude of studies with different eddy diffusion parameterization (e.g., Pilinski & Crowley, 2015; Salinas et al., 2016; Siskind et al., 2014), it is a major challenge to understand and quantify the effect of eddy diffusion. Lately, it has been suggested that the upper mesosphere chemistry coupled to the thermospheric spoon effect might be important for the neutral and plasma density distribution (e.g., Jones et al., 2018). Similarly, it was demonstrated that the MLT composition can significantly contribute to neutral density error in models (Mehta et al., 2019).

The composition highlights the coupling of the I-T system by being influenced by the lower atmosphere dynamics, the mesospheric chemistry, solar and highlatitude heating, and winds. Constellation observations are needed to quantify the different potential drivers and to quantify the importance of the different mechanisms.

3. Future Directions

It should be apparent from the foregoing discussion that the presented challenges to advance our understanding of the I-T require significant efforts to improve the scope of observations and the capabilities of computational models.

It is clear that comprehensive measurements of the neutral wind are an essential component of the advances to be made and, presently, such measurements are sparsely distributed in time and space. Various methodologies for remote sensing from space (Killeen et al., 1982; Shepherd et al., 2012) and from the ground (Makela et al., 2013) have provided valuable insights into the global and regional behavior of the winds in the I-T with a relatively small volume of measurements being made in situ (Wharton et al., 1984). These data are the basis of empirical models for the wind (e.g., Droeb et al., 2015) that are still used in ionospheric models to
investigate the effects of dynamo-driven currents and ionospheric motions parallel to the magnetic field. Refined techniques for remotely sensing the wind below 200 km are emerging (Wu et al., 2016). However, many challenges described here require the wind and the ion drift to be specified in the same volume. Overcoming this challenge requires innovative data gathering schemes to connect two regions along the magnetic field (Immel et al., 2018) and knowledge of the ionospheric conductivity distribution in altitude, or in situ observations of the wind and the ion drift, which can be accomplished from an orbital platform or from ground-based facilities.

Addressing the problems of sparseness in data coverage requires a wider distribution of ground-based observations and the deployment of satellite constellations. The challenges to data distribution from the ground arise principally from the availability of appropriate sites with good meteorological and sunlight conditions. The challenges to data distribution from satellite constellations arise from the access to the critical range of altitudes above 100 km and the availability of robust instrumentation that can provide the required measurements of neutral and plasma density and velocity.

The good news is that all these challenges can be overcome, and we may expect that effective observation programs using ground- and space-based assets will be undertaken in the next 5 to 10 years. The advent of small spacecraft with small propulsion units will allow the deployment of constellations to all altitudes above 150 km. It is important to emphasize, however, that the transition from larger, fully instrumented measurement platforms to smaller platforms with more targeted measurement capabilities is not straightforward. It will require innovations and validation activities to demonstrate the capabilities required to advance our understanding. Nevertheless, the availability of multipoint measurements at different locations will allow a first-order separation of space and time. The use of propulsion also allows orbit parameters to be tailored to the prevailing neutral pressure, which determines the ion-neutral coupling efficiency as was emphasized during recent periods of very low solar activity (Heelis et al., 2009; Solomon et al., 2013).

One of the most prominent targets for future observations and complementary modeling is directed toward a better understanding of the wave sources at low altitudes and the factors affecting the propagation of those waves into the I-T. As has been discussed, there are many sources of variability, and the degree to which the variability can be predicted will depend on isolating the sources and better understanding of the drivers. Improving our specification of the propagation characteristics requires further descriptions of the heat sources and sinks that affect the altitude distribution of the neutral density, the composition, and the prevailing bulk flow.

Within the I-T itself the prominent target is to understand how the coupling between the plasma and the neutral gas affects the response of the system to external energy inputs. These inputs come from the wave motions discussed above and from the interactions of the I-T with the magnetosphere, which are applied at high latitudes. We have emphasized here that winds and currents in the I-T effectively distribute a localized energy input throughout the region. Thus, localized behaviors cannot be studied in isolation from the more global state of the I-T. However, specific mechanisms for energy deposition are localized in space and time, and thus, observational and modeling targets can be restricted. For example, particle and frictional heating are largely restricted to high latitudes, and dedicated observational initiatives are required to separate the spatial and temporal variations in these energy inputs and the local response of the I-T. While the nature of the energy inputs and the local I-T response can be described with appropriately distributed measurements at high latitudes, it should also be recognized that the I-T response to these inputs will extend to middle and low latitudes, where the magnetic field has a different orientation and another set of spatially distributed measurements will be required to describe it.

Observations alone are insufficient to advance our understanding or to provide a legacy on which future developments will depend. Rather, an advancement will be solidly established when physics-based models are able to explain observed phenomenology and establish the drivers that are responsible for it. In the foregoing discussion, we have seen that the challenges to our physical models lie in specification of the model drivers, descriptions of multiscale processes, and accounting for interactions with neighboring regions. Each of these challenges is attached to observational challenges to constrain the model and to verify the results that are obtained. In principle, many of the model challenges are overcome by the use of adaptive grids to account for smaller scales and nonlinear processes and by the use of a framework that connects neighboring regions of geospace at the appropriate spatial and temporal scales. However, appropriate
physics for the smaller scales and for the coupling of different geospace regions have to be employed. Significant progress is being made in this direction (e.g., Connor et al., 2016) at the largest spatial scales, and we may expect that increases in computational capability will allow processes operating on smaller spatial and temporal scales to be considered.

At the present time the most consequential observational challenge is to discover the most important spatial scales that are responsible for the deposition of energy into the I-T and the time scales over which they are applied. The corresponding modeling challenge is to reconcile the observed state of the I-T with the spatial and temporal scales over which the I-T responds. The spatial density over which observations are available is likely to remain a significant constraint. Thus, the development of physics-based assimilation procedures and machine learning, to confidently create a volume or surface distribution for a particular parameter given a sparse distribution of samples at any given time, will be of increasing value to the research endeavor.

At the highest level our challenge is to describe the I-T with sufficient clarity to identify the most important processes that control its role in the interaction of our planet with the Sun and the interplanetary medium. In so doing, we will discover the important features that distinguish our planet from its neighbors. We will also improve predictions of the distribution of neutral density, charged particle density, and the currents that flow in the geospace environment, all of which affect space- and ground-based systems that we increasingly rely upon.

Acknowledgments

The data to produce Figures 1–6 are available at the UCAR/NCAR Digital Access Service Hub repository (https://doi.org/10.5065/eled-h060). R. A. H. was supported by NASA Grant NNG12FA44C and AFOSR Grant FA9550-16-1-0364. A.M. was supported by NASA through HSR Grant NNX16AG64G and Explorer Grant NNX14AF03G and by AFOSR through Award FA9550-17-1-0248. ICON is supported by NASA’s Explorers Program through Contracts NNG12FA44C and NNG12FA42I. This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977. The authors would like to thank Hanli Liu and Art Richmond for valuable comments on an earlier draft and the two reviewers for their time and constructive comments, which helped improve the manuscript.

References


