**Imaging the D-region Ionosphere with Subionospheric VLF Signals**

by

**Forrest Gasdia**

B.S., Embry-Riddle Aeronautical University, 2014

M.S., Embry-Riddle Aeronautical University, 2016

M.S., University of Colorado Boulder, 2019

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Aerospace Engineering Sciences

2021

Committee Members:

Robert Marshall, Chair

Tomoko Matsuo

Delores Knipp

Cora Randall

Jeffrey Thayer

Gasdia, Forrest (Ph.D., Aerospace Engineering Sciences)

Imaging the D-region Ionosphere with Subionospheric VLF Signals

Thesis directed by Prof. Robert Marshall

Long radio waves in the very low frequency (VLF) radio band propagate great distances in the space between Earth's surface and the base of the ionosphere. This characteristic has led to the use of VLF radio for resilient strategic communication and positioning, navigation, and timing (PNT) systems. Unlike the propagation of high frequency (HF) radio waves, which refract high in the ionosphere, VLF waves are guided by the surface and D-region boundaries of the Earth-ionosphere waveguide (EIWG). Propagation is affected by the electromagnetic characteristics of both boundaries, and although ground conductivity is relatively stable, the ionosphere is constantly changing. Understanding how the lower ionosphere influences the propagation of low frequency signals is essential for the operation of reliable VLF systems. Observing subionospheric VLF signals also provides an opportunity to remotely sense the conductivity profile of the D-region ionosphere. The D-region responds to a large number of phenomena, yet is difficult to measure because of its very low electron density and its location above aircraft and below satellites. Snapshots of the D-region produced from subionospheric VLF signals over a large area would not only improve knowledge of typical conditions in the lower ionosphere, but may also provide insight into external perturbing events.

This dissertation examines VLF propagation in the Earth-ionosphere waveguide through realistic models of the ionosphere. It then uses simulated observation experiments to demonstrate two methods for combining observations across an array of VLF receivers to "image" the D-region. The problem of estimating an ionosphere perturbed by energetic particle precipitation (EPP) is also analyzed. The results indicate that two-dimensional estimates of the D-region may be possible with real VLF data in the near future.

## Dedication

In memory of my grandfather, William F. Hough.

## Acknowledgements

First I would like to acknowledge my advisor, Robert Marshall. Bob has been supportive of all of my research paths, thoughtful in his advice, and been an amazing example of how to grow and run a research group. When I arrived at CU, the LAIR group was just Bob, André Antunes de Sá, and me, and today there are more than a dozen members. André paved the way, making it easier for me to navigate each step of the way through my defense. I would like to thank the entire LAIR group for listening to my (sometimes long) presentations and especially Wei Xu for his technical feedback and collaboration.

This dissertation would not have been possible without the professors I have had at CU and throughout my education. In particular, Prof. Tomoko Matsuo introduced me to data assimilation and many of the methods I applied to the VLF imaging problem. I would like to thank my entire defense committee for posing thoughtful questions and showing interest in VLF.

I would like to acknowledge my family and friends who have supported me throughout this journey. Thank you for being patient with me as I changed and cancelled plans so that I could complete this dissertation on time. A big thank you to my Aunt Susan and Matt who generously welcomed me into their home in the weeks leading up to the defense.

This work was financially supported by a fellowship award through the National Defense Science and Engineering Graduate (NDSEG) Fellowship Program, sponsored by the Air Force Office of Scientific Research (AFOSR), the Army Research Office (ARO), and the Office of Naval Research (ONR).

## Contents

**Chapter**

**Appendix**

# Tables

**Table**

## Figures

**Figure**

# Chapter 1

# Introduction

## 1.1 VLF propagation and the D-region ionosphere

It has been recognized from the early days of practical radio experimentation that long radio waves can efficiently propagate for thousands of kilometers over Earth's surface (P. Hansen, 2016). Commercial broadcasters, ships at sea, aviators, and militaries are just some of those who have utilized extremely low frequency (ELF), very low frequency (VLF), and low frequency (LF) radio systems since the beginning of the 20$^{th}$ century (Watt, 1967; Gebhard, 1979). Except for a few select systems, modern applications are limited by large antennas, very low data rates, and limited spectrum. One example of current VLF use is communication with submarines by global navies, but there is also renewed interest in using VLF for positioning, navigation, and timing (PNT) to augment global satellite navigation systems (GNSS) (Burke, 2019; Navigation Innovation and Support Programme (NAVISP), 2020; Naval Sea Systems Command (NAVSEA), 2021).

VLF radio waves propagate efficiently over great distances because they reflect off the ionosphere—the charged region of Earth's upper atmosphere. The VLF radio band covers wavelengths of 10 to 100 km, so the base of the ionosphere is only a small number of wavelengths distance above Earth's surface; together they form the boundaries of the "Earth-ionosphere waveguide" (EIWG). The propagation of long radio waves through this naturally occurring waveguide is affected by the electromagnetic characteristics of both the ground and ionosphere. Although ground conductivity is relatively stable, the lower ionosphere is constantly changing.

Understanding how the ionosphere influences low frequency signals propagating through the EIWG is essential for operating accurate and reliable PNT systems and planning low frequency communication links.

Despite indirectly interacting with the lower ionosphere through longwave radio transmissions for a century, our understanding of the D-region ionosphere, where the wave reflection occurs, is limited by the difficulty in measuring it directly. The D-region is between 50 to 90 km altitude, so it is above aircraft and high-altitude balloons, but below satellites. A limited number of sounding rockets have made measurements in the D-region, but these have been sparse in space and time. The electron density is extremely low so it only reflects long radio waves; higher frequencies are absorbed or pass through. We lack a comprehensive "snapshot" of the D-region to understand its spatial characteristics as it responds to a great number of external factors.

### 1.2 D-region interactions

The D-region ionosphere is an important interaction region between the neutral atmosphere and ionospheric plasma above. It responds to a large number of external influences which makes remote sensing and understanding the D-region both challenging and rewarding. In some cases, observations of the D-region are actually a diagnostic for understanding external perturbing phenomenon. For example, improved knowledge of the D-region has consequences for our understanding of ozone production and models of atomic oxygen (Verronen et al., 2005; Siskind et al., 2013), initiation of atmospheric electrical phenomena (Marshall et al., 2006; Qin et al., 2014; NaitAmor et al., 2017), monitoring of climate change (Laštovička et al., 2006; Clilverd et al., 2017a), and forecasting absorption along medium and high frequency (MF/HF) communication links (Siskind et al., 2017). Although the D-region is primarily driven by the apparent motion of the sun, it responds to a large number of phenomena: solar X-ray flares (Šulić et al., 2016), solar eclipses (Clilverd et al., 2001; Guha et al., 2010), astronomical gamma-ray bursts (Tanaka et al., 2010), atmospheric waves (Marshall & Snively, 2014), earth-

quakes (Phanikumar et al., 2018), direct heating with radio transmitters (Inan, 1990), nuclear detonations (Zmuda et al., 1964), lightning (Marshall, 2012), energetic particle precipitation from the radiation belts as a result of solar storms (Clilverd et al., 2009) and induced by radio transmitters (Imhof et al., 1983; Cohen et al., 2012), and meteorology (Rozhnoi et al., 2014). Simões et al. (2012) and Silber and Price (2017) describe an even greater number of the D-region's interactions with the neutral atmosphere.

As VLF waves propagate subionospherically from a transmitter to a receiver, the otherwise stable amplitude and phase of the signal will change—sometimes rapidly, sometimes slowly—when one of the above phenomena occur along the propagation path. Because the distance between a transmitter and receiver is often thousands of kilometers, different conditions usually exist at different locations along the propagation path. Interpreting the received signal is difficult because of the number of phenomenon that can affect the observations and, ironically, because VLF is sensitive to very small changes in electron density. For example, Bainbridge and Inan (2003) report sensitivity to a change from 55 $e^-/cm^3$ to 79 $e^-/cm^3$ at 80 km altitude.

### 1.3 Inverting VLF observations

In any waveguide, the reflectivity of the upper and lower boundaries and the distance between them determines the field distribution of propagating waves in the free space of the waveguide. Therefore, the average ionosphere height and reflectivity over a propagation path can be estimated by monitoring the amplitude and phase of a VLF transmitter signal at a fixed receiver. The relationship between the EIWG parameters and the signal amplitude and phase is highly nonlinear; a computer model is required to compute the ionosphere reflectivity and reproduce observations for known ground and ionosphere conductivity profiles. Thomson (1993), McRae and Thomson (2000), and many others have used receivers to monitor the amplitude and/or phase of VLF signals of opportunity in order to retrieve the path-average conductivity profile by iteratively fitting the data to computer-modeled observations.

Estimating the electron density profile of the D-region with subionospheric VLF is an

ill-posed, nonlinear inverse problem. Estimates of the electron density are non-unique for a given observation, particularly if the profile is assumed to vary along the propagation path. The problem is underdetermined because the electron density must be estimated in more locations than there are observations. Complicating the solution is the nonlinear relationship between the states being estimated and the observations, and the fact that the observations are non-local. Each observation contains integrated information on the ionosphere and ground along the entire propagation path and the weighting function of the integration has a nonlinear relationship with the estimated states. That is, the response of an observation to a local perturbation in the ionosphere depends on the location of the perturbation, the intensity of the perturbation, and the unperturbed background ionosphere along the rest of the propagation path—all of which are unknown in the general estimation problem.

### 1.4 Contributions

The great majority of D-region estimates using subionospheric VLF are path-average estimates or investigations into localized perturbations using relative change in amplitude or phase. There is recent work on simultaneous estimation of the D-region using networks of receiver observations (McCormick & Cohen, 2018; Gross & Cohen, 2020), but the community still lacks maps of D-region electron density like those that exist for the upper ionosphere, e.g. in the form of GNSS-derived total electron content (TEC). This work aims to expand the capability of subionospheric VLF receiver arrays by exploring techniques to generate geographically varying electron density profiles. Improved estimates of typical D-region ionospheres benefit both scientific modeling and practical applications of longwave radio. Techniques capable of estimating spatially varying D-region electron density profiles are particularly useful for understanding the spatial and temporal extents of D-region disturbances and the phenomena that cause them. This work has also resulted in the creation of a new mode-theory computer program to model longwave propagation in the EIWG. The contributions can be summarized as:

- A simpler and more robust computer code has been created for modeling the propagation of VLF waves in the Earth-ionosphere waveguide;
- A sensitivity analysis of the VLF amplitude and phase response to different EIWG conditions has been performed.
- An ensemble Kalman filter technique for estimating the spatially-varying D-region has been developed and assessed.
- An iterative cost-function minimization approach for estimating the spatially-varying D-region with poor *a priori* information has been developed and assessed.
- The feasibility of imaging an energetic particle precipitation (EPP) patch has been investigated.

## 1.5 Organization

This dissertation contains seven chapters and two appendices. The contents of each of the following chapters is described below.

Chapter 2 provides background for the rest of the dissertation. It begins by characterizing the D-region ionosphere and introduces models used to describe typical electron density profiles. The chapter then discusses various methods for observing the D-region and compares propagation models that have been used by previous researchers as the forward model in estimates of the D-region.

Chapter 3 is a mathematical overview of the mode theory of longwave propagation in the Earth-ionosphere waveguide. A high level understanding of this theory helps explain the complexity of the propagation and inverse problem, as well as the restrictions it places on the choice of estimation technique. Appendix A provides a more extensive discussion of the theory suitable for those interested in a detailed understanding of mode theory propagation codes.

Chapter 4 introduces the Longwave Mode Propagator (LMP) model developed for this dissertation. It discusses the motivation for creating this model and how the model is different from the Long-Wavelength Propagation Capability (LWPC). Accuracy is verified by comparing the model results against LWPC and a finite difference propagation model. A collection of model results for a variety of Earth-ionosphere waveguide conditions are then presented that indicate the sensitivity of subionospheric VLF observations. Appendix B presents the results of tests that were used to determine default parameters used by LongwaveModePropagator.jl.

Chapter 5 formally introduces the D-region estimation problem, presents the common receiver array scheme used to generate the estimate maps, and discusses the mathematics of the estimation techniques used in this dissertation. The results of simulated observation experiments demonstrating the effectiveness of these estimation techniques, including in the presence of model errors, are presented in Chapter 6. Chapter 6 also discusses challenges with imaging an ionosphere perturbed by an energetic particle precipitation (EPP) event.

Finally, Chapter 7 summarizes the conclusions made in the previous chapters and suggests areas for improvement and future work.

# Chapter 2

# Background

## 2.1 The D-region ionosphere

Solar radiation ionizes gasses in Earth's upper atmosphere, forming a region of weakly ionized plasma known as the ionosphere. This region extends from about 50 km to over 800 km above Earth's surface. The intensity of solar radiation increases exponentially with height while the neutral atmospheric density decreases exponentially with height. The resulting charge density profile is a simple model ionosphere attributed to Chapman (1931). Details of the photo-chemistry, however, produce a more complicated distribution of free electrons in the real ionosphere. There are three primary layers of locally higher electron density at night (F, E, and D) and four at day ($F_2$, $F_1$, E, and D), listed in order of decreasing altitude. Figure 2.1 displays these layers for typical conditions according to the International Reference Ionosphere (IRI) model (Bilitza et al., 2017). The F layer (150 to 800 km) contains the bulk of the ionization and is relatively well understood by scientists. The E layer (100 to 150 km) is best known for sporadic and significant increases in electron density, a phenomenon that remains under study. At the base of the ionosphere is the D-region (50 to 100 km), characterized by free electron densities of only 1 to 1000 $e^-/cm^3$ (Mitra, 1978).

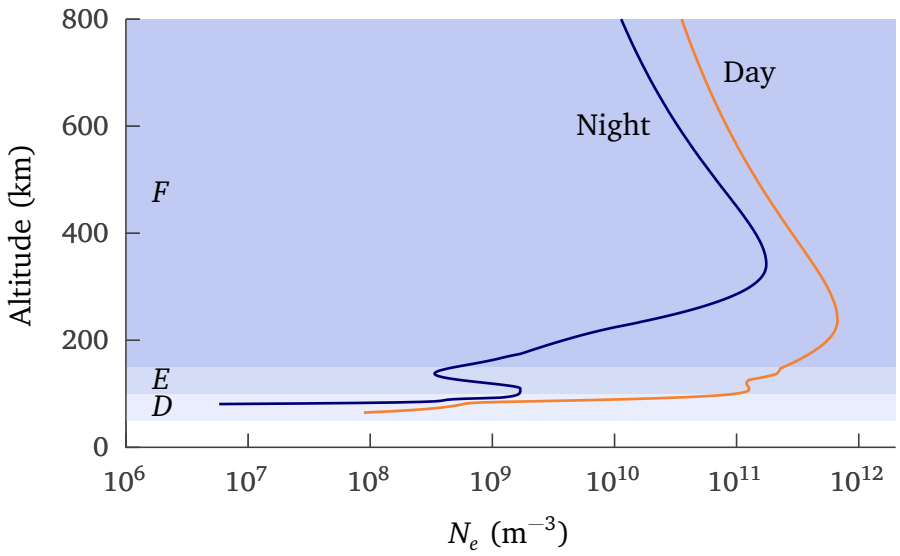


Figure 2.1: Electron density profiles for typical daytime and nighttime ionospheres for 4 April 2020 over Boulder, CO, as generated by the International Reference Ionosphere (Bilitza et al., 2017).

### 2.1.1 Characteristics

The electron density of the D-region is so low that it can be safely ignored for many ionospheric studies and applications. The plasma frequency[1] often does not exceed 300 kHz in this region, so propagation is not affected for the great majority of modern radio systems. However, users of medium frequency (MF, 0.3 to 3 MHz) and high frequency (HF, 3 to 30 MHz) systems, including military, aviation, maritime communications, and shortwave broadcasters, can experience signal attenuation due to the D-region. This occurs because the electron-ion collision frequency is comparable to the wave frequency. Sudden ionospheric disturbances (SIDs) or polar cap absorption (PCA) events can enhance the D-region electron density to the point of causing HF radio blackouts because the attenuation is so great (Smirnova et al., 1988). In fact, the only regularly published D-region forecasts are the D-region absorption predictions

[1] Electron plasma frequency is the frequency of electrostatic oscillation of electrons if slightly displaced from the background collection of ions in a plasma. It is $f_p = (N_e e^2/4\pi^2\epsilon_0 m)^{1/2}$ for electron density $N_e$, charge $e$, and mass $m$. Radio waves of lower frequency than the plasma frequency are reflected from the plasma.

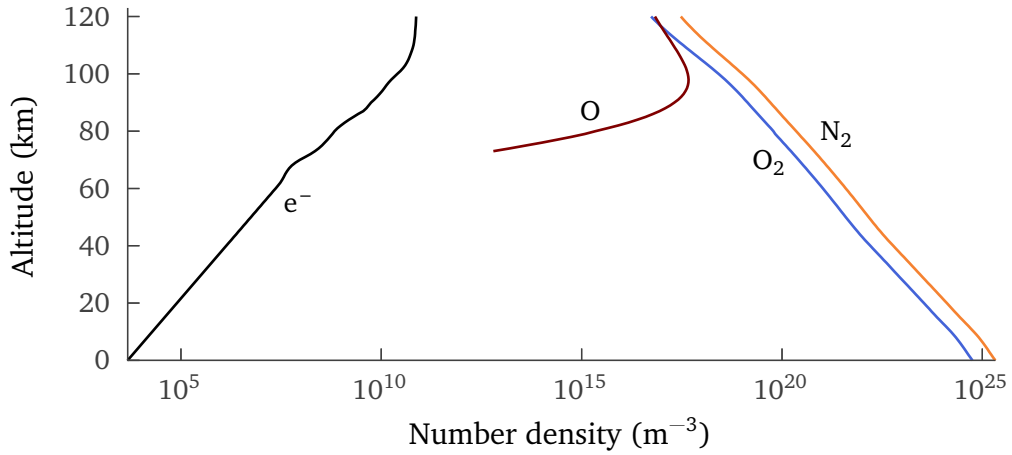


Figure 2.2: MSIS density profiles for the major neutral constituents at D-region altitudes pictured alongside the FIRI electron density (Picone et al., 2002; Friedrich et al., 2018). The profiles are for daytime, 2020-03-01, at 60° N.

(D-RAP) to explain degraded HF propagation conditions (Space Weather Prediction Center, 2020).

The primary ionization source for the daytime D-region is solar radiation. Short wavelength X-rays, Lyman-$\alpha$, and ultraviolet light ionize the atmospheric constituents $N_2$, $O_2$, and smaller amounts of NO, $H_2O$, O, $O_3$, Na, and Ca (Nicolet & Aikin, 1960; Mitra, 1978). Because of its relatively low altitude, D-region chemistry is particularly complicated and the minor constituents play an important role (Friedrich & Rapp, 2009; Siskind et al., 2017). The high density of neutrals at 50 to 90 km altitude, depicted in Fig. 2.2, results in frequent ion-neutral collisions and recombination is rapid. At night the D-region might be expected to disappear without solar radiation to continuously ionize the region. In fact, many scientists do not consider the D-region to be present at nighttime, e.g. (Kelley, 2009), but as we will later see, long radio waves continue to reflect from heights below the typical altitude of the E-region at night. Scattered Lyman-$\alpha$, galactic cosmic rays, meteors, and even starlight may each be partly responsible for maintaining some ionization in this region throughout the night (Nicolet & Aikin, 1960; Torkar & Friedrich, 1983).

Beyond the complicated chemistry and variety of ionization sources that produce a tenuous

and difficult-to-predict D-region density profile, there are several external influences: neutral atmosphere temperature variations and acoustic and gravity waves may carry localized variations in the electron density profile (Lay & Shao, 2011; Marshall & Snively, 2014; Pal & Hobara, 2016); changes in D-region conductivity as a result of atmospheric electrical phenomena, including lightning and sprites, has been observed (Inan et al., 1995; Cheng & Cummer, 2005; Marshall et al., 2006; Inan et al., 2010; Qin et al., 2014); after solar storms, precipitation of energetic electrons from Earth's radiation belts enhances high-latitude regions of the D-region (Imhof et al., 1983); and anthropomorphic phenomena such as atmospheric nuclear detonations and the operation of high power transmitters also influence the region (Crain & Tamarkin, 1961; Koons et al., 1980; Graf et al., 2013).

### 2.1.2 Models

A number of models, varying in complexity and spanning from purely empirical to detailed photo-chemistry, have been developed specifically to describe the D-region ionosphere. The major models of the upper atmosphere and ionosphere, including the International Reference Ionosphere (IRI), Whole Atmosphere Community Climate Model (WACCM), the Global Ionosphere-Thermosphere Model (GITM), and Sami2 is Another Model of the Ionosphere (SAMI2) do not provide reliable coverage into the D-region (Huba et al., 2000; Ridley et al., 2006; Bilitza et al., 2011; Marsh et al., 2013). WACCM-D extends WACCM by including 307 reactions of 20 positive ions and 21 negative ions relevant to the D-region (up from five ions and 15 reactions in WACCM) (Verronen et al., 2016). Similarly, the Faraday-International Reference Ionosphere (FIRI), discussed in detail below, extends the IRI model into the D-region by blending sounding rocket measurements with a simple ion-chemical model of the lower ionosphere (Friedrich & Torkar, 2001).

Models have also been developed that exclusively cover the D-region. The Originally Austrian Study of the Ionosphere (OASIS) is a steady-state ion-chemical model with about 55 reactions that simplifies the negative ion chemistry into either $O_2^-$, $O^-$, or all other negative

ions $X^-$ (Torkar & Friedrich, 1983; Siskind et al., 2015). The state of the art is the Sodankylä Ion and Neutral Chemistry (SIC) model. It is a comprehensive one-dimensional chemistry model including more than 400 reactions between 41 positive ions, 29 negative ions, and 34 neutral species and takes into account external forcing from solar radiation, particle precipitation, and galactic cosmic rays (Verronen et al., 2005; Turunen et al., 2009; Verronen et al., 2016).

A few simple models are used in this dissertation or are otherwise particularly relevant. A discussion of each follows.

#### 2.1.2.1 GPI

Glukhov et al. (1992) introduces a basic chemical model consisting of only four kinds of charged particles: electrons, negative ions (e.g. $O_2{}^-$, $CO_3{}^-$, $NO_2{}^-$), light positive ions (e.g. $O_2{}^+$, $NO^+$), and heavy positive ion clusters (e.g. $H^+(H_2O)_n$), where the number densities of each are referred to as $N_e$, $N^-$, $N^+$, and $N_x^+$, respectively. This model is referred to as the Glukhov, Pasko, and Inan (GPI) model. Accuracy at heights below ~50 km was improved in an update by separating negative ions into light negative ions, e.g. $O_2{}^-$, and heavy negative ions, e.g. $NO_3{}^-$, denoted by number densities $N^-$ and $N_x^-$ (Lehtinen & Inan, 2007).

The evolution of the species in the GPI model is given by

$$\mathrm{d}N_e/\mathrm{d}t = Q - \beta N_e + \gamma N^- + \gamma_x N_x^- - (\alpha_d N^+ + \alpha_d^c N_x^+) N_e \tag{2.1}$$

$$\mathrm{d}N^-/\mathrm{d}t = \beta N_e - \gamma N^- - \alpha_i (N^+ + N_x^+) N^- - A N^- \tag{2.2}$$

$$\mathrm{d}N_x^-/\mathrm{d}t = -\gamma_x N_x^- - \alpha_i (N^+ + N_x^+) N_x^- + A N^- \tag{2.3}$$

$$\mathrm{d}N^+/\mathrm{d}t = Q - \alpha_d N_e N^+ - \alpha_i (N^- + N_x^-) N^+ - B N^+ \tag{2.4}$$

$$\mathrm{d}N_x^+/\mathrm{d}t = -\alpha_d^c N_e N_x^+ - \alpha_i (N^- + N_x^-) N_x^+ + B N^+ \tag{2.5}$$

but because of the neutrality condition $N^+ + N_x^+ = N_e + N^- + N_x^-$, one of the equations can be eliminated. The ionization source $Q$ is calculated as the source necessary to produce a steady state solution of the equations from initial background density profiles of $e^-$, O, $O_2$, and $N_2$, which are provided from other models e.g. MSIS (Picone et al., 2002). The coefficients $\alpha$,

$\beta$, and $\gamma$ describe neutralization and recombination, attachment rate, and detachment rate, respectively, which are provided by a variety of sources (Lehtinen & Inan, 2007). Although it requires far fewer computations, the five-species GPI model results in electron density profiles that agree to within about 40 % of the SIC model and can also be used to model disturbed D-region profiles by providing an ionization source in addition to the equilibrium source profile $Q$ (Marshall & Cully, 2020). Figure 2.3 shows the charge profiles resulting from GPI given neutral species and temperature profiles from MSIS and electron density from FIRI at daytime in the northern hemisphere.

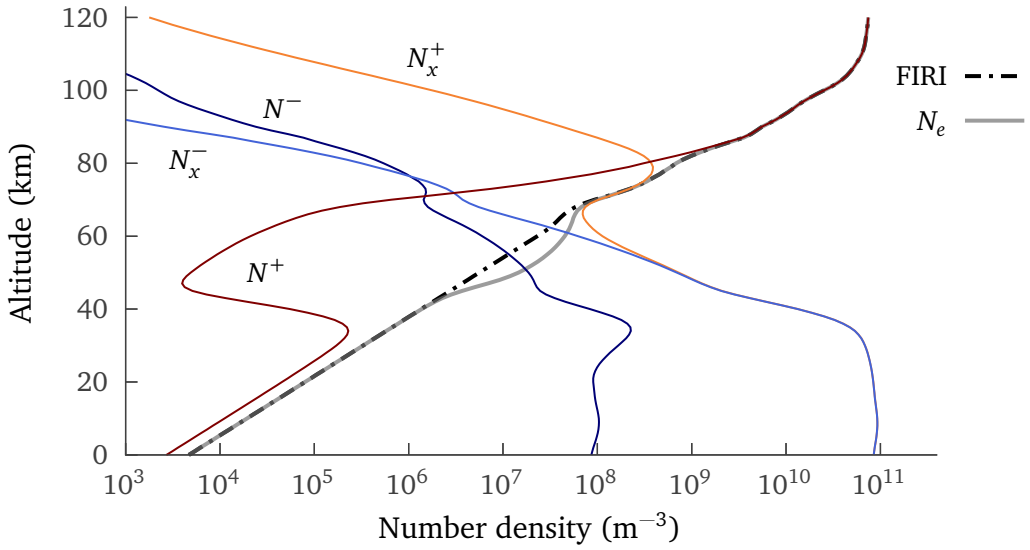


Figure 2.3: GPI density profiles for electrons and light and heavy negative and positive ions. The input neutral density and temperature profiles are from MSIS for 2020-03-01 2000 UTC and 60° N, 102° W. The original electron density profile is an FIRI-2018 profile interpolated for the appropriate month, solar zenith angle, latitude, and F10.7 index, and exponentially extrapolated to the ground.

A custom implementation of GPI has been programmed for this dissertation using default values of the coefficients. Equations (2.1) to (2.3) and (2.5) are integrated using the `Rodas4` 4th-order stiffly stable Rosenbrock solver from DifferentialEquations.jl and $N^+$ is solved from the neutrality condition (Rackauckas & Nie, 2017). As in version 5.4 of the implementation from Lehtinen and Inan (2007), charge equilibrium is established by performing a binary search for

the ionization source $Q$ that satisfies steady-state equilibrium of the equations at each height of the profile. An additional ionization source, such as energetic particle precipitation, can be added to the equilibrium source and the equations solved at each height once more.

#### 2.1.2.2 FIRI

The Faraday International Reference Ionosphere (FIRI) is a semi-empirical model of the undisturbed ionosphere based on an ion-chemical model adjusted to satisfy in situ sounding rocket profiles (Torkar & Friedrich, 1983; Friedrich & Torkar, 2001; Friedrich et al., 2018). The model was generated using a total of 120 profiles using the Faraday rotation technique and a further 207 profiles from Langmuir probe measurements—all collected from the literature. The published FIRI-2018 model is in tabular form and consists of 1980 electron density profiles for 11 solar zenith angles between 0° and 130°, latitudes 0°, 15°, 30°, 45°, and 60° N, and three solar activity levels for the middle of each month of the year. The model is claimed to be valid for altitudes above 60 km and densities over $10^6\,\mathrm{e}^-/\mathrm{m}^3$.

To apply the model to scenarios requiring profiles at arbitrary solar zenith angle and latitude, a tool for interpolating across the published tabular profiles has been developed for this dissertation.[2] Additionally, the tabular profiles can either be averaged across the unspecified fields or quantiles can be obtained. Although the model is only expected to be valid for altitudes down to about 60 km, an exponential extrapolation can be performed to lower altitudes.[3] Examples of the FIRI profiles interpolated across latitude and solar zenith angle and extrapolated down to 40 km are shown in Fig. 2.4.

#### 2.1.2.3 Exponential parameterization

One of the oldest yet most common ways to describe the electron density of the lower ionosphere derives from Wait and Spies (1964) and is commonly referred to as the “Wait and

[2] github.com/fgasdia/FIRITools.jl

[3] There is some evidence that the electron density at and below 60 km altitude is sufficient to affect the propagation of long radio waves (Siskind et al., 2018; Xu et al., 2019).

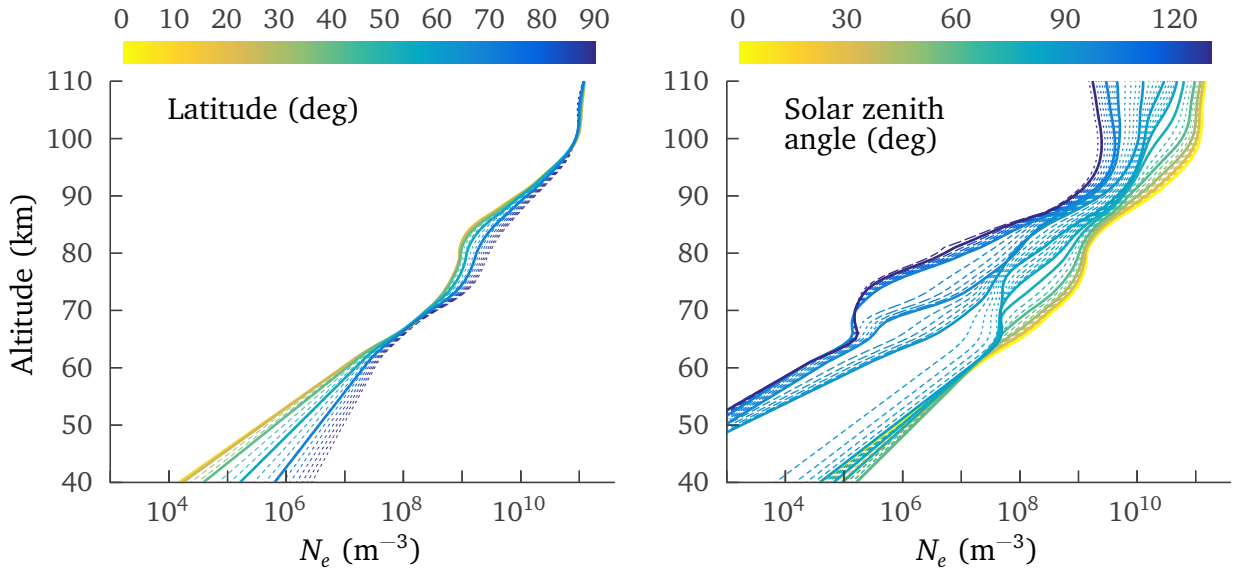


Figure 2.4: Left: Published FIRI-2018 electron density profiles (solid lines) averaged over all 12 months and F10.7 indices at a solar zenith angle of 45° and interpolated at latitudes every 5° (thin dashed lines). Right: Published average FIRI-2018 profiles (solid lines) at 45° N and interpolated between 0° and 130° solar zenith angle (thin dashed lines). Interpolation between solar zenith angles of 80° and 100° are every 1° to better resolve the terminator. All other angles are interpolated in 5° increments.

Spies" or "Wait" profile. To characterize the propagation of VLF radio waves in the Earth-ionosphere waveguide, Wait and Spies found it convenient to define a "conductivity parameter"

$$\omega_r = \omega_p^2 / \nu \tag{2.6}$$

where $\omega_p$ is the angular plasma frequency and $\nu$ is the effective electron-neutral collision frequency. Laboratory data from Phelps and Pack (1959) compared favorably to partial reflection and sounding rocket observations of collision frequency in the lower ionosphere and was used to fit the curve

$$\nu(z) = 5 \times 10^6 \exp\big(-0.15(z-70)\big) \quad (\mathrm{s}^{-1}) \tag{2.7}$$

for height $z$ in kilometers.

Wait and Spies (1964) fits the profile

$$\omega_r = (2.5 \times 10^5) \exp\big(\beta(z-h')\big) \quad (\mathrm{rad\,s}^{-1}) \tag{2.8}$$

parameterized by a reference height $h'$ (km) and constant $\beta$ (km$^{-1}$) to VLF propagation data assuming the collision frequency profile in Eq. (2.7). To obtain an expression for the electron density profile, we substitute the expressions for $\omega_p^2$ and $\nu$ into Eq. (2.8) and solve for $N_e$, resulting in

$$N_e(z) \approx 1.43 \times 10^{13} \exp\left(-0.15h'\right) \exp\left((\beta - 0.15)(z - h')\right) \tag{2.9}$$

or some equivalent form. It can be shown that, neglecting Earth's magnetic field, reflection occurs approximately at the height at which $\omega_p^2/\omega^2 = \nu/\omega$ (Ratcliffe, 1959).[4] This is an equivalent statement to $\omega_r = \omega$. Therefore, noting that $h'$ is the height at which $\omega_r = 2.5 \times 10^5$ rad s$^{-1}$, $h'$ is the reflection height for a 40 kHz radio wave.

It is extremely common in the VLF community to describe the D-region electron density profile by the $h'$ and $\beta$ parameters which best fit observations. It is important to remember that the Wait profile inherently assumes the collision frequency profile Eq. (2.7) and was derived using fits to VLF propagation data. Others have used similar exponential profiles. Jacobson et al. (2009) uses the exponential parameterization from Volland (1995), which uses a different reference density for $N_e$ than Wait and Spies (1964). Bainbridge and Inan (2003) found the best fit to their phase observations was achieved with an interpolation between the profiles of Wait and Spies (1964) and the non-exponential profile from Reagan et al. (1981). Others have used multiple Wait profiles as a function of height to better match observations (Xu et al., 2019).

As we will see later, VLF has limited sensitivity above the height where most of the reflection occurs. Therefore, Wait's exponential profile does not directly describe a true electron density profile throughout the D-region, rather it is an *effective* profile for VLF applications. Figure 2.5 compares the best-fit Wait profiles for day and night midlatitude GPI profiles. Also plotted are several sounding rocket profiles from Siskind et al. (2018) which contain a knee captured by GPI but not the exponential Wait model.

[4] The reflection condition with no magnetic field is $X = Z$ using the URSI symbols for magnetoionic theory.

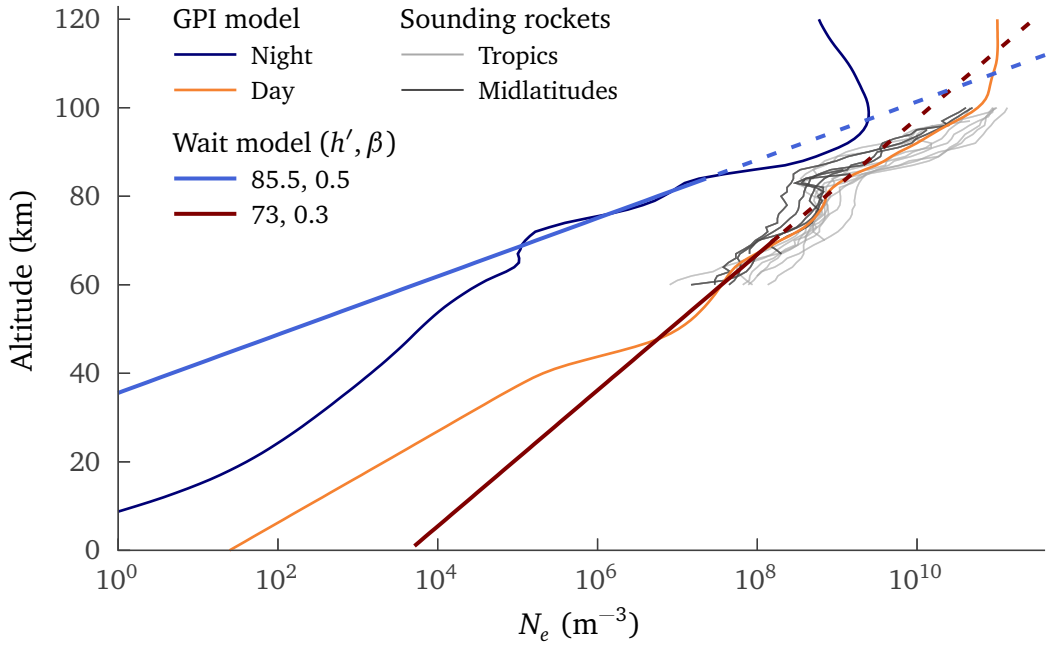


Figure 2.5: The GPI model does a good job capturing the daytime knee of the electron density profile measured by sounding rockets (Siskind et al., 2018). The $h'$ and $\beta$ of Wait's model on this plot were adjusted to fit the VLF amplitude resulting from the GPI profiles along a 2000 km path from a 24 kHz transmitter. Wait's exponential ionosphere inherently lacks the ability to capture features in the more realistic GPI profile. The Wait profiles are dashed above the approximate reflection height at which $\omega_r = \omega$.

**Ferguson ionosphere**

Ferguson (1980) fit exponential Wait profiles to 570 electron density profiles from the literature. Ferguson then performed a multiple linear regression analysis over solar zenith angle $\chi$, geographic latitude $\phi$, Zurich smoothed sunspot number SSN, month number $m$, and quiet (0) or disturbed (1) conditions $M$. The resulting expressions for $h'$ and $\beta$ are:

$$h' = 74.37 - 8.097\cos\chi + 5.779\cos\phi - 1.213\cos\big(2\pi(m-1/2)/12\big) - 0.044\,\mathrm{SSN} - 6.038M \tag{2.10}$$

$$\beta = 0.3849 - 0.1658\cos\chi - 0.08584\cos\phi + 0.1296\cos\big(2\pi(m-1/2)/12\big) \tag{2.11}$$

Ferguson (1980) also analyzed data from 28 aircraft flights that recorded radial profiles of the field strength amplitude and found differences from his model in the nighttime field strength

| | Day | | Night | |
|---|---|---|---|---|
| Frequency (kHz) | $h'$ (km) | $\beta$ (km$^{-1}$) | $h'$ (km) | $\beta$ (km$^{-1}$) |
| 10 | 74 | 0.3 | 87 | 0.3 |
| 60 | 74 | 0.3 | 87 | 0.8 |

Table 2.1: D-region electron density exponential profile parameters used in the Long-Wavelength Propagation Capability (LWPC). A linear fit is assumed in frequency for nighttime $\beta$ such that $\beta = 0.4\,\mathrm{km}^{-1}$ for a 20 kHz signal. REPRODUCED FROM Ferguson (1995).

suggesting errors of 2 to 4 km. The model does not contain the level of variability observed in the flight data, and there was a significant difference in $h'$ between the data and model at high latitudes. Additionally, the Ferguson model predicts a much smoother variation with solar zenith angle than likely exists. It does a particularly poor job of capturing the terminator (Marshall et al., 2017). Nonetheless, the model is trivial to compute and only requires knowledge of the current time if the sunspot number and conditions flag are ignored, as they often are.

**LWPC ionosphere**

Data from Morfitt (1977) and Ferguson (1980) were used to build a subjectively determined best fit Wait ionosphere for the U.S. Navy's Long-Wavelength Propagation Capability (LWPC), described in detail below. The $h'$ and $\beta$ parameters, shown in Table 2.1, were reanalyzed in Ferguson (1995) against the original data and found to generally agree to ∼6 dB in amplitude. The LWPC model simply has daytime (solar zenith angle $\chi < 90°$) and nighttime ($\chi > 99°$) because there was insufficient data to accurately model the influence of solar zenith angle. A linear step model is used for the transition between day and night. At nighttime, the geomagnetic dip angle dictates another linear step transition that adjusts $h'$ between the midlatitude and polar nighttime ionosphere (Ferguson, 1998).

## 2.2 Observing the D-region

The D-region is notoriously difficult to measure. Its very low electron density means HF ionosondes are not sensitive to the region. In situ measurements are not possible using aircraft,

balloons, or satellites because none operate in the D-region's altitude range. Sounding rockets pass through the region, but most sounding rockets do not deploy their instruments as low as the D-region. Nonetheless, several methods of directly or indirectly sensing the D-region ionosphere have been used over the past 100 years. In this section we will briefly review those measurement techniques and place extra emphasis on previous work indirectly observing the D-region using long radio waves.

### 2.2.1 Measurement techniques

Several methods somewhat similar to traditional ionosondes have been used in past studies of the D-region. Appleton and Piggott (1954) present a study of lower ionosphere absorption measurements over a period of 18 years. The studies were conducted by launching a 4 MHz radio wave pulse into the ionosphere and measuring the effective reflection coefficient $\rho$ by comparing the amplitude of the series of echoes from sequential reflections of the wave between the ground and ionosphere; absorption is then $|\log \rho|$. Appleton and Piggott (1954) were able to study seasonal variations, relationship to the sunspot number, variation with frequency,[5] the degree of day-to-day variability, and sudden ionospheric disturbances, but care had to be taken in the analysis to separate D-region effects from the E- and F-regions. Riometers continue to be used today to measure HF absorption due to the D-region in order to improve absorption prediction models (Rogers & Honary, 2014).

Belrose and Burke (1964) discuss a similar technique using partial reflection to more directly measure electron density and collision frequency in the D-region. Partial reflection uses measurements of the reflected amplitude of both ordinary and extraordinary waves pulsed at two lower HF frequencies to infer the absorption and reflection profiles and hence collision frequency and electron density profiles. Because the D-region reflections are weak compared to reflections that occur off the F-region, it is necessary to perform the experiment at a low-noise site with large antenna arrays and use very high transmitter power levels; Belrose and

[5] Measurements were carried out later in the study at frequencies other than 4 MHz.

Burke (1964) used a 1 MW transmitter at 2.66 MHz and 100 kW transmitter at 6.275 MHz. This technique can observe reflections from altitudes below 50 km, but the equipment requirements are those of a major experiment. It may be possible using modern digital receivers to reduce the necessary transmitter power levels, but I am not aware of these types of measurements being made in the past three decades.

More recently, measurements of the upper D-region have been obtained using incoherent scatter radar (ISR). At least one observation with the ALTAIR radar was conducted coincident with a rocket-born Faraday rotation experiment and the two were in good agreement (Friedrich et al., 2006). Unfortunately, the sensitivity of ISR to the low electron densities in the D-region limits their applicability to daytime altitudes of above 85 km except in cases of very strong disturbances. During storms, the electron density is enhanced to lower altitudes; during a recent pulsating aurora event the EISCAT radar observed down to 68 km (Miyoshi et al., 2015).

The most direct means of measuring the D-region is by sounding rocket. A relatively small number have been launched by the United States specifically to study the D-region. Langmuir probes can be used to make in situ measurements of the electron density (Smith, 1965), but the aerodynamic flow around the rocket body can alter the plasma density from the undisturbed ionosphere, even when the probes are deployed on booms (Friedrich et al., 2018). An alternative approach is to place a radio transmitter on the rocket and measure the Faraday rotation resulting from propagation through the ionosphere with a fixed receiver on the ground (Jacobsen & Friedrich, 1979). Both of these techniques were used to build the FIRI D-region model (Section 2.1.2.2). Unfortunately, sounding rocket flights for the D-region are sparse in space and time, limiting our understanding of the region under varying conditions.

### 2.2.2 Longwave observations and estimation

Long radio waves in the ELF through LF bands (0.3 to 100 kHz) are sensitive to the very low electron densities in the D-region. The electron density profile can be estimated from these signals as a result of the fact that the radio field measured at a fixed point responds to changes

in the electromagnetic properties of a waveguide's boundaries. Therefore, the propagating electromagnetic field measured at a distant receiver can be used to infer properties of the Earth-ionosphere waveguide as they change over the course of a day or in response to external sources. We are interested in the ionosphere so we assume we have perfect knowledge of the ground conductivity,[6] although in practice this can be a source of error.

Long radio waves can be generated naturally by lightning or by man-made transmitters.[7] For the most part, longwave estimation techniques observe the D-region indirectly. It is necessary to invert the measurements through a nonlinear process to estimate the D-region conductivity profile that reproduces the measurements. Below we discuss some techniques first using lightning-generated longwave signals and then using narrowband man-made transmitters.

#### 2.2.2.1 Lightning atmospherics

Lightning discharges emit most of their radio frequency energy in the ELF and VLF radio bands (Cummer et al., 1998). This energy is "guided" by the ground and D-region ionosphere which act as reflecting boundaries in a naturally occurring waveguide for the efficient propagation of long radio waves over thousands of kilometers, a scenario depicted in Fig. 2.6. Cummer (1997) outlines in detail how lightning atmospherics ("sferics") can be used to infer the average D-region electron density along the propagation path to a receiver. The major difference between lightning and man-made transmitter signals is that lightning contains rich frequency information; it is a wideband signal as opposed to narrowband. The receiver amplitude need not be absolutely calibrated. Instead, the observed frequency spectra over tens of kilohertz is compared to the spectrum obtained by a forward propagation model. A table of different exponential $h'$ and $\beta$ parameters are modeled and the closest matching spectrum determines the ionosphere. One difficulty with using lightning sferics as a radio source is that

[6] VLF is also used to remotely sense electrical properties of the ground, e.g. Pedersen et al. (2009) or Mogensen et al. (2014).

[7] VLF waves have even been generated directly in the D-region by modulated HF heating using a high power steerable transmitter (Cohen et al., 2010a).

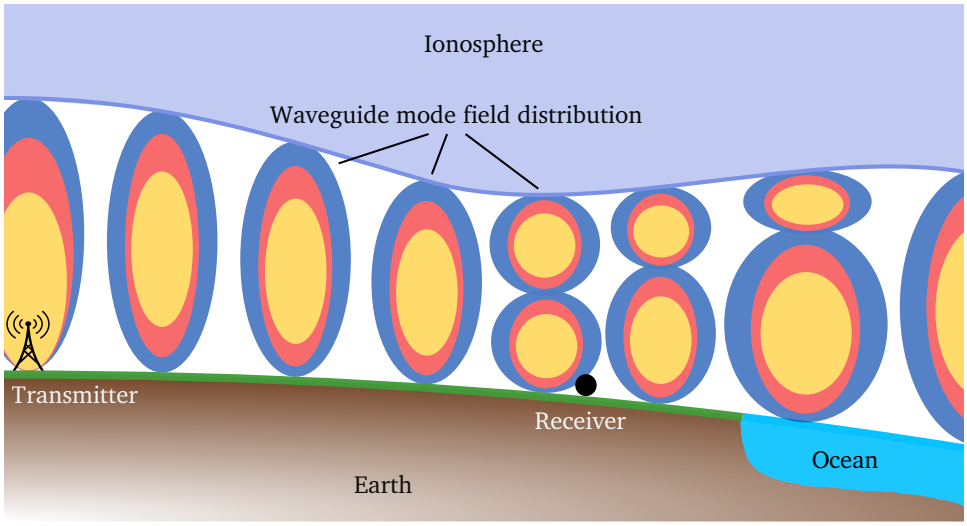


Figure 2.6: A notional diagram of the Earth-ionosphere waveguide. Long radio waves generated by lightning discharge or man-made transmitters propagate between the conductive ground and ionosphere. The energy is partitioned into resonant modes that are subject to the electromagnetic boundary conditions set up by the conductivity profile of the ionosphere and ground.

the radio source location must be approximately known for comparison of a real observation with a forward model. Cummer et al. (1998) obtained the source location using the National Lightning Detection Network (NLDN). Another issue is that the lightning or thunderstorm might perturb the ionosphere, affecting the electron density estimate (Cheng & Cummer, 2005; Marshall et al., 2008).

Lay et al. (2014) present a technique using lightning sferics to *locally* observe the D-region rather than obtain the average electron density profile along the propagation path from lightning to receiver. The technique compares time domain lightning waveforms between 1 to 160 kHz from direct ground wave propagation to first-hop ionospheric reflections. The results are compared to a propagation model and are sensitive to the ionosphere at the midpoint of the signal-receiver path. There were several differences between the exponential $h'$ and $\beta$ parameters obtained by Lay et al. (2014) and typical guided wave estimates of the D-region, even when the midpoint was located away from a thunderstorm. In particular, quiet nighttime

profiles were found to have very high $\beta$ values of 2.8 $km^{-1}$ compared to 0.5 to 0.6 $km^{-1}$ for path-average techniques.

#### 2.2.2.2 Narrowband signals

Radio transmitters in the early 20$^{th}$ century operated at low frequencies and played a significant role in the development of early models of the lower ionosphere (Watson & Nicholson, 1918; 1919). Although radio research and applications quickly moved to higher frequencies for practical reasons, interest in long waves returned after World War II at the beginning of the Cold War when high power VLF transmitters were built for near-global strategic communication and navigation. Prior to the Global Positioning System (GPS), the United States operated the Omega navigation system which relied on a global network of phase-synchronized transmitters which could be used by ships and aircraft to determine their position through multiranging or hyperbolic navigation modes (Morris et al., 1994). Russia continues to operate their "Alpha" navigation system today. Several navies operate VLF communication transmitters because of their resiliency and ability to communicate with submerged submarines due to the relatively great skin-depth of sea water at long radio wavelengths. Although the bandwidth of VLF transmitters is too narrow for voice communication, one-way encoded data is transmitted nearly continuously from high power VLF transmitters around the globe.

The U.S. Navy operates several transmitters across United States territory using minimum-shift keying (MSK) modulation at 200 baud. These are signals of opportunity for estimating the D-region ionosphere. Because MSK is phase-continuous, the phase trellis can be estimated and unwrapped to obtain an effectively continuous-wave (CW) signal (Cohen et al., 2010b). Although the instantaneous transmitted signal amplitude and phase is unknown, the receiver can be calibrated to receive absolute amplitudes and the phase can be referenced to a stable local oscillator coordinated with GPS. Three days of phase data from the NML transmitter in North Dakota observed at a receiver in Colorado is shown in Fig. 2.7. Some processing difficulties result from these transmitters being used for military communication rather than

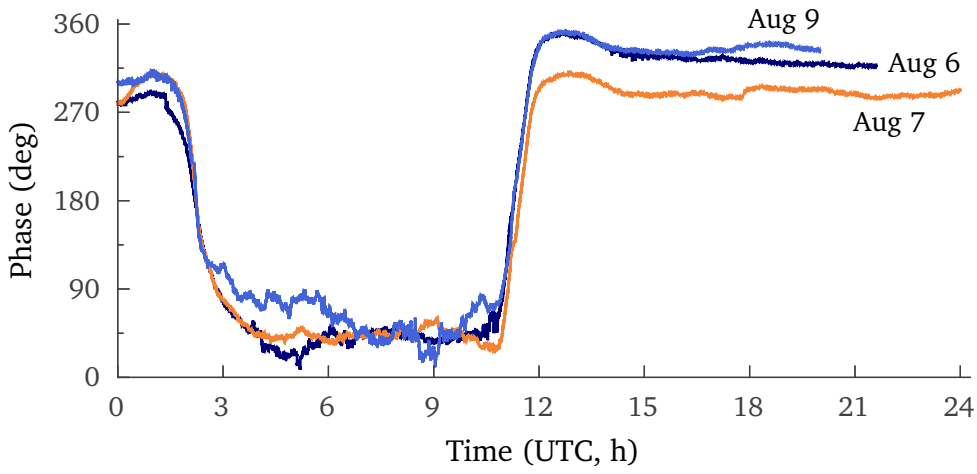


Figure 2.7: Phase of the NML transmitter measured at Table Mountain in Colorado for three summer days in 2017. The relatively smooth phase occurs during daytime and the highly variable segment is at night. Each phase curve has been shifted so all three days correspond.

for safety-of-life navigation. They have unscheduled maintenance time off and lose phase continuity during these periods, and although some sources consider the transmitters highly phase stable (McRae & Thomson, 2000), carrier phase instabilities have been observed (Gross et al., 2018).

The typical VLF D-region estimation experiment includes several magnetic loop antennae and radio receivers placed across a region of several hundred to several thousand kilometers. A computer continuously records the observed amplitude and/or phase of signals at known transmitter frequencies. Transmitter-receiver paths are examined one at a time and candidate ionospheres are run through a forward propagation model.[8] The modeled amplitude and/or phase is compared to the observed amplitude and/or phase and the candidate ionosphere is adjusted until the difference between the modeled and observed signals match to within some tolerance (Thomson, 1993; McRae & Thomson, 2000; Thomson, 2010). Early experiments benefited from knowledge of the Omega transmitter's predictable signals, but later experiments using Navy transmitter signals require a reference to (relatively stable) midday phase or multiple

[8] The complicated nonlinear relationship between the ground, Earth's magnetic field, and the ionosphere's collision frequency and charge density profiles on the observed field requires the use of computer codes (usually LWPC) for the forward model.

simultaneous receiver measurements across different paths (Thomson et al., 2017). Careful processing of the signal using the polarization ellipse can be used to circumvent the problem of unknown transmitter phase (Gross et al., 2018). Although most published work has taken a fairly brute force approach to estimating $h'$ and $\beta$, Gross and Cohen (2020) applied an artificial neural network (ANN) to amplitude and carrier phase data from across an array of receivers to simultaneously estimate the average parameters along each path.

Narrowband VLF observations have also been used in other ways for more targeted experiments. Rather than using the absolute values, *changes* in amplitude and/or phase have been directly correlated to changes in the ionosphere conductivity profile, frequently for studying solar eclipses (Clilverd et al., 2001; Guha et al., 2010; Cohen et al., 2018; Xu et al., 2019), but older experiments applied this technique to monitor atmospheric nuclear detonations (Zmuda et al., 1963). Temporal spacing of amplitude minima observed by a receiver is used by Marshall and Snively (2014) to study atmospheric gravity waves and by Samanes et al. (2015) to study modal interference across the terminator. Rapid changes in the VLF signal, known as "early/fast" events, have been correlated with sprites (Marshall et al., 2006). Bainbridge and Inan (2003) used a relatively dense linear array of phase-coherent receivers to decompose the VLF signal into its constituent waveguide modes to infer a non-exponential electron density profile.

Although VLF receivers are usually stationary on the ground, the U.S. Navy conducted several experiments with receivers placed on high-speed aircraft with the goal of producing improved models of the ionosphere. By placing receivers on aircraft, an amplitude curve could be observed over a great distance in a period of time for which the ionosphere is relatively constant. These experiments also used oblique multi-frequency VLF sounders located on the ground to provided additional information about the ionosphere through the multiple wave frequencies (Pappert & Morfitt, 1975; Morfitt, 1976; 1977). The sounder system appears to have operated in the 1960s and 1970s, and no doubt a similar system would provide deep insight into the D-region today. Unfortunately, technical and licensing issues related to VLF spectrum usage would need to be overcome to operate such a transmitter.

## 2.3 Longwave propagation models

To estimate the D-region using subionospheric field measurements of narrowband VLF transmitters, it is necessary to use a forward model that mathematically propagates the transmitted signal through the Earth-ionosphere waveguide to a distant receiver. As alluded to in the previous section, the model observation is iteratively compared to the real observation, adjusting the candidate ionosphere with each step to produce a better match between the modeled and real observations. The forward model is thus an integral part of the nonlinear estimation process.

The relationship between the state of the ionosphere along a propagation path and the observed amplitude and phase of a transmitted field is complicated and highly nonlinear.[9] A large number of variables contribute to the observed field: the conductivity profile of the ionosphere, the ground conductivity, and Earth's magnetic field all vary spatially within the Earth-ionosphere waveguide, and the transmitter and receiver antennae orientations and positions must also be considered. To capture all of these influences, the forward model for this problem is expressed as a computer program.

Several computer codes have been developed to model the propagation of long waves in the Earth-ionosphere waveguide using ray, modal, or finite difference approaches. Less than a wavelength ($\sim 5\,\mathrm{km}$) from the transmitter, the field is strongly reactive and best calculated as a direct radiation field. Beyond that distance, both a direct ground wave and ionosphere-reflected sky wave can be distinguished and a ray model is usually applied. Far from the transmitter ($\gtrsim 1000\,\mathrm{km}$) the field is best represented as the Earth-ionosphere guided field and is commonly modeled using mode theory. Finite difference approaches do not favor one regime over the other except for the fact that the compute time increases with the size of the model domain.

[9] The estimation problem is nonlinear in the sense that the relationship between the states being estimated (parameterizations of the ionosphere's conductivity profile) and the observations are nonlinear. The propagation problem can be formulated as the solution of Maxwell's linear differential equations with particular boundary conditions. The system is complicated by the ionosphere being a spatially varying, dispersive, and anisotropic medium, and nonlinear phenomena such as transmitter heating could also be modeled.

Regardless of the model, "full wave" solutions using fairly accurate representations of Maxwell's equations must be applied because VLF waves are partially-reflected, partially-transmitted, and partially-absorbed in the lower ionosphere. Whereas geometric optics approaches can be applied to the propagation of high frequency (HF) radio waves under the assumption that the ionosphere changes relatively slowly compared to the wavelength (10 to 100 m), this is violated at VLF (10 to 30 km).

Ray theory is favored over short distances from the transmitter because the field can be calculated from just a few ray "hops" between the ground and ionosphere. More and more hops must be considered to accurately compute the field at greater distances. Ray approaches are relatively uncommon, but have been compared to mode theory (Volland, 1961) and have recently been applied to D-region estimation using lightning sferic sources (Jacobson et al., 2009). To reiterate, these are not ray optics approaches. Jacobson et al. (2009) solves Maxwell's equations for a plane wave by integrating four simultaneous complex differential equations through the anisotropic, collisional, horizontally stratified ionosphere using the method of Pitteway (1965).[10] Wavefront curvature is approximated using a summation over an angular spectrum of plane waves. Although all angles contribute to the field at the receiver, those angles near the specular angle of ray optics have the greatest influence.

At great distances from the transmitter it is more efficient to represent the field with a sum over a small number of waveguide modes rather than a large number of ray hops. Mode theory is the most commonly used forward model for D-region estimation, and it is the method implemented in the Long-Wavelength Propagation Capability (LWPC), the *de facto* standard for modeling VLF propagation. LWPC was the culmination of several decades of effort by the U.S. Naval Electronics Laboratory Center (NELC, now Naval Information Warfare Systems Command—NAVWAR)[11] to develop longwave propagation models for signal coverage analysis

[10] In the Longwave Mode Propagator developed for this dissertation (see Chapter 4), the method of Pitteway (1965) is used to integrate wavefields for mode conversion.

[11] The U.S. Naval Electronics Laboratory Center (NELC) merged into the Naval Ocean Systems Center (NOSC) and then the Space and Naval Warfare Systems Command (SPAWAR) during development of the program that became LWPC. In 2019 SPAWAR was renamed to Naval Information Warfare Systems Command (NAVWAR).

(Sheddy et al., 1968; Pappert et al., 1970; Morfitt & Shellman, 1976; Ferguson & Snyder, 1980; 1987; Ferguson, 1998). Application of mode theory to the Earth-ionosphere waveguide was solved in a way amenable to computer solution by Budden (1955) and Budden (1962) and is detailed in Chapter 3. In mode theory, the Earth-ionosphere system is modeled explicitly as a waveguide consisting of a series of horizontally homogeneous segments along the propagation path. Reflection coefficients are computed for the ground and ionosphere in each segment and resonant eigenangles corresponding to waveguide modes are identified. The fields radiated from the transmit antenna are summed over the eigenangles and carried into the next waveguide segment through a mode conversion algorithm. Mode theory models can compute the field at a distant receiver in just seconds on a modern computer.

The final category is finite difference propagation models. The simplest forms include finite-difference frequency-domain (FDFD) and finite-difference time-domain (FDTD) models, which directly solve a discretized form of Maxwell's equations on a spatial grid (Chevalier & Inan, 2006). Both two- and three-dimensional variants of FDTD have been applied to propagation in the Earth-ionosphere waveguide (Simpson & Taflove, 2007; Marshall et al., 2017). These models compute the electric and magnetic fields at high resolution across the model grid and are capable of incorporating complicated structures in the EIWG (Marshall, 2012). Two slightly different models are the recursive finite element model known as the Stanford Full-wave Method (FWM) (Lehtinen & Inan, 2008; Lehtinen & Inan, 2009) and the method from Nagano et al. (2003) which applies a matrix method to compute the full wave solution between thin layers of the horizontally stratified ionosphere. All four of these methods require considerable computational resources. FDFD requires large matrix inversions, FDTD has a numerical stability requirement to resolve the wave propagation in space and time over great distances (Courant et al., 1928), and FWM has a step size requirement that scales its number of integrations with the distance of interest from the transmitter squared. The Nagano method requires computing full wave solutions for over one billion plane waves. Although some of these have been used as forward models for D-region estimation, e.g. (Han et al., 2011; Xu et al., 2019), they are

prohibitively expensive for the imaging problem that is the focus of this dissertation. Therefore, we focus on mode theory models because of the large number of forward model runs required to perform the nonlinear estimate of the D-region from observations of VLF signals.

# Chapter 3

## Mode theory of propagation

The structure formed by Earth's conductive surface and the ionosphere above acts as a naturally occurring waveguide for long radio waves in the ELF through LF radio bands. The "height" of the D-region ionosphere is approximately 75 km, so the distance between ground and the ionosphere is only on the order of 1 to 10 free space wavelengths. This chapter describes how the Earth-ionosphere waveguide (EIWG) can be mathematically modeled using waveguide mode theory to determine the electric field strength of a dipole transmitter signal measured at a distant receiver in the waveguide. The resulting model is then used as the forward propagation model in the D-region estimation problem.

Explanations of longwave propagation assumed a concentric conducting shell around Earth at a time when formal waveguide theory was still in its infancy, but it did not take long for the theory of guided waves to appear as a plausible explanation for phenomena being observed on radio signals (Watson & Nicholson, 1919; Hollingworth et al., 1926). By 1936, observations of interference patterns as a function of distance from the transmitter were well known and already being used to approximate the height of the ionosphere (Best et al., 1936). Much of the theory presented here, and employed by both LWPC and the Longwave Mode Propagator (LMP, Chapter 4), was developed by K. G. Budden in the 1950s and 60s. Budden (1955) presented an approach to compute the reflection coefficient of the anisotropic ionosphere as though it were a sharp boundary, and in 1962 he presented the mode theory necessary to calculate the electromagnetic fields from a known exciter in the waveguide. Others, especially James R. Wait,

contributed significantly to our understanding of longwave propagation (Wait, 1957; 1958; 1963), but it was Budden's method that was best suited to numerical solution of general EIWG scenarios with a computer. By 1968, Sheddy et al. published results from a FORTRAN 63 computer program developed at the U.S. Naval Electronics Laboratory Center using Budden's model (Sheddy et al., 1968). This group would continue developing the methods and computer program that would become LWPC over the next few decades. Ultimately, the model that will be discussed here allows for:

- multiple species in the ionosphere (electrons, positive and negative ions)
- general (continuous) number density profiles $N(z)$ as a function of height $z$ for each species
- general (continuous) species-neutral collision frequency profiles $\nu(z)$ as a function of height
- the curvature of Earth
- the effect of Earth's magnetic field
- the effect of Earth's ground conductivity and permittivity
- varying of the magnetic field, ground, and ionosphere profiles as a sequence of horizontally homogeneous slabs along the propagation path.

The system is modeled with Earth's surface and the ionosphere forming two walls of a waveguide with free space in between. A radio source placed in the waveguide will excite discrete modes that propagate a great distance. The total field observed by a receiver in the waveguide is the sum of the fields corresponding to each propagating mode. The propagation model can therefore be broken down into three main problems: 1) identify resonant modes of a homogeneous waveguide segment, 2) calculate the electromagnetic fields associated with each mode, and, if necessary, 3) propagate the fields from one waveguide segment into the next waveguide segment. An overview of the process is shown in Fig. 3.2. Step 3 refers to mode

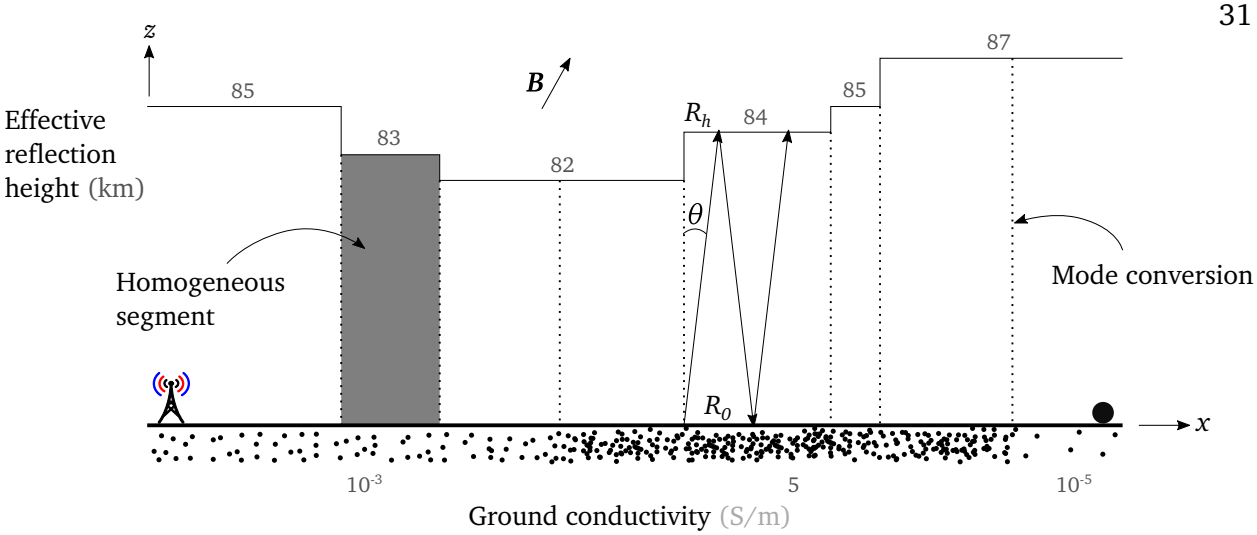


Figure 3.1: The Earth-ionosphere waveguide as modeled with mode theory. There is a ground reflection coefficient $R_0$, ionosphere reflection coefficient $R_h$, and free space in between.

conversion and is usually required to model realistic scenarios over long propagation paths where it is inevitable that at least one of the ionosphere, ground, or magnetic field will change significantly along the path. This segmented waveguide model is depicted in Fig. 3.1.

Readers seeking additional explanation of each of the steps presented in this section are strongly encouraged to read Appendix A. No single book captures all of the background and methods necessary to explain EIWG mode theory. Budden (1961b) presents his work on general waveguide modes. Budden (1988) is a treatise on electromagnetic propagation through the ionosphere, but has relatively little to say about longwave propagation in the EIWG. Watt (1967) discusses many of the practical aspects of transmitting and receiving VLF radio signals. Much of the published theory and discussion of modeling methods is spread across several decades of U.S. Navy technical reports written during the development of LWPC (Morfitt & Shellman, 1976; Ferguson & Morfitt, 1981; Pappert & Ferguson, 1986; Shellman, 1986). This chapter and the appendices of this dissertation make an attempt to explain much of EIWG mode theory in one place at a level appropriate for first year graduate students. At the same time, gaps

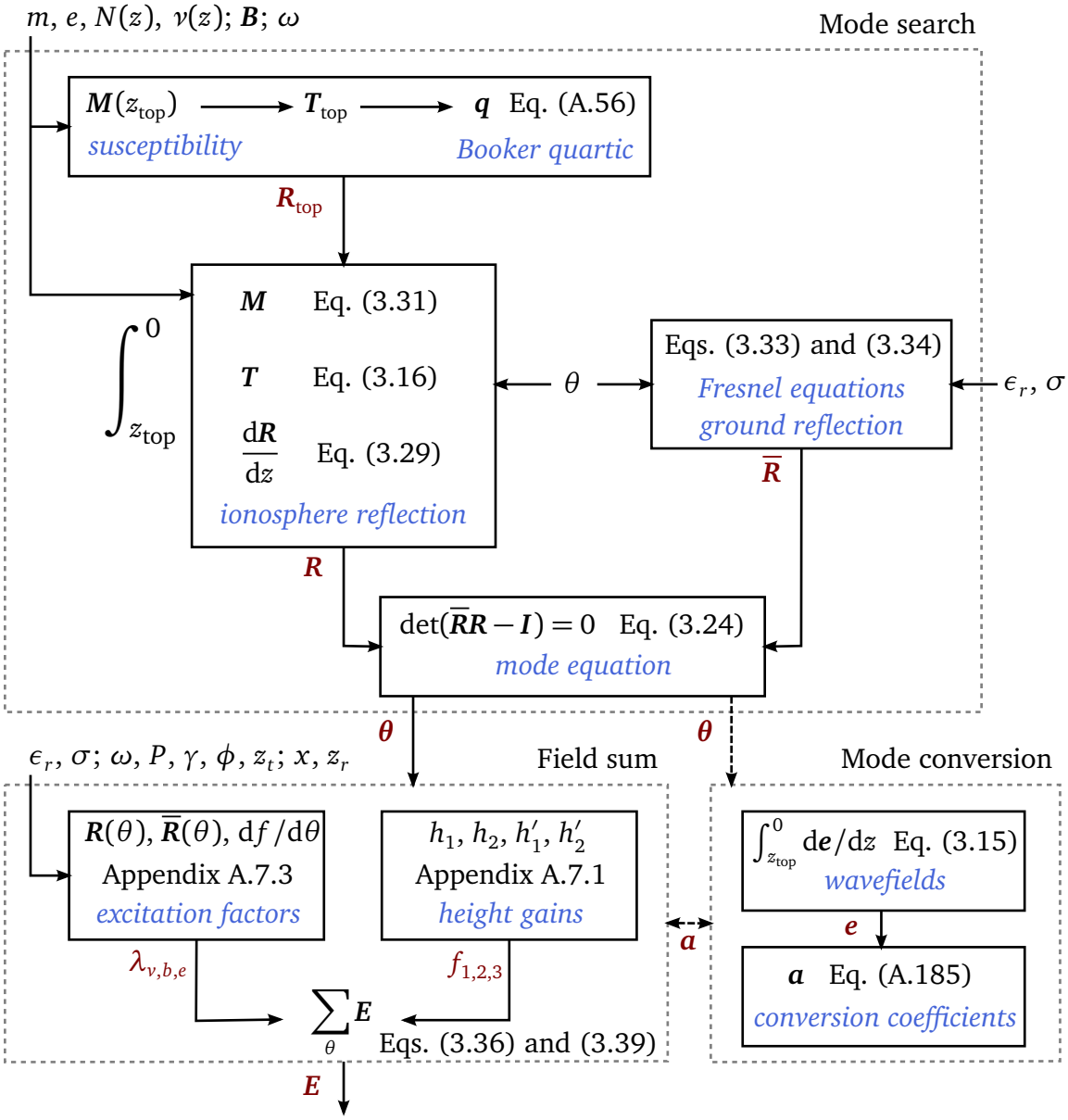


Figure 3.2: Steps used by mode theory to compute the electric field $\boldsymbol{E}$ observed by a receiver in the Earth-ionosphere waveguide. The diagram is read primarily from top to bottom. It begins with a search for the resonant modes in a homogeneous segment of the waveguide. This requires solving the mode equation for the eigenangle associated with each mode, which in turn requires computation of the ground and ionosphere reflection coefficients. After the dominant waveguide modes are identified for every waveguide segment along the propagation path, the wavefield profile can be computed in each segment and used to compute the mode conversion coefficients at each transition between segments. The fields at the transmitter are propagated through each segment until they're eventually computed at the receiver. Variables are defined in the text.

remain in a thorough derivation of waveguide excitation, to which the reader is pointed largely to Budden (1961b) and the results of Pappert in e.g. Pappert et al. (1967) and Pappert et al. (1970).

Before beginning, we assume that all fields vary harmonically with time $t$ so that e.g. $\boldsymbol{E}$, $\boldsymbol{H}$, $\boldsymbol{D}$, $\boldsymbol{P}$ oscillate with angular frequency $\omega$. If some real (measurable) field component is $F(t)$, then

$$F(t) = A\cos(\omega t + \phi) = \mathrm{Re}\big(A\exp(i\phi)\exp(i\omega t)\big). \tag{3.1}$$

The $\exp(i\omega t)$ will be implicitly multiplied onto every field—not explicitly written. For example, the electric field vector $\boldsymbol{E}(t) = \mathrm{Re}(\boldsymbol{E}_0\exp(i\omega t))$ will be written as $\boldsymbol{E}$. Conveniently, the time derivative $\partial/\partial t$ of a field is equivalent to multiplication by $i\omega$.

The Earth-ionosphere waveguide coordinate system, shown in Fig. 3.3, has $z$ directed vertically upward into the ionosphere, $x$ is horizontal along the waveguide propagation path, and $y$ completes the right-handed coordinate system. This is a two-dimensional model with invariance in $y$. To be concise, we will frequently use the notation $S = \sin\theta$ and $C = \cos\theta$.

## 3.1 Maxwell's equations in the ionosphere

In this section we derive a special form of Maxwell's equations for radio wavefields in the cold, magnetized, collisional ionosphere. As these are fundamental equations, it should not be surprising that they are needed both for the derivation of the ionosphere reflection coefficient and for calculation of the mode conversion coefficients.

A radio wave propagating in the ionosphere generates a current density $\boldsymbol{J}$ which arises from the motion of charges. If $\boldsymbol{r}$ is the average vector displacement of an electron from the position it would have occupied if there were no radio field, then the average electron velocity is $\partial\boldsymbol{r}/\partial t$ and the current density generated by a volume of electrons with density $N$ is

$$\boldsymbol{J} = Ne\ \partial\boldsymbol{r}/\partial t\,. \tag{3.2}$$

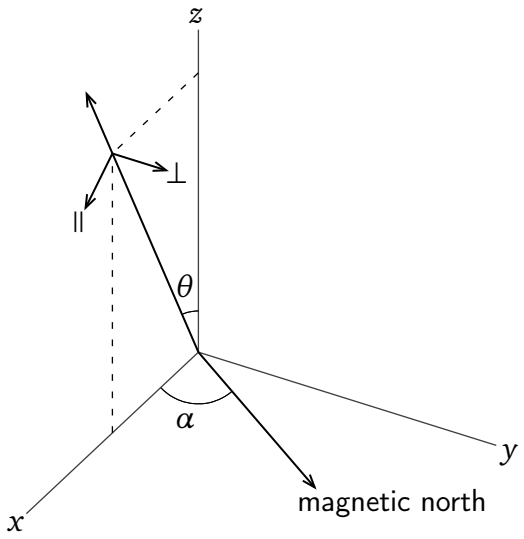


Figure 3.3: The wave propagates within the $x$–$z$ plane. The long vector represents the wave normal of an upgoing wave. The vectors marked || and ⊥ represent electric field components parallel and perpendicular to the plane of incidence, respectively. A variety of coordinate systems are used throughout the literature and for specific derivations, although this is the most common and simplifying.

The electric polarization can be defined

$$\boldsymbol{P} = Ne\boldsymbol{r} \tag{3.3}$$

so that

$$\boldsymbol{J} = \partial \boldsymbol{P}/\partial t\,. \tag{3.4}$$

Maxwell's equations for an electromagnetic field are given by

$$\begin{aligned} \operatorname{div}\boldsymbol{D} &= 0 \\ \operatorname{div}\boldsymbol{B} &= 0 \\ \operatorname{curl}\boldsymbol{E} &= -\mu_0\,\partial \boldsymbol{H}/\partial t = -i\omega\mu_0\boldsymbol{H} \\ \operatorname{curl}\boldsymbol{H} &= \partial \boldsymbol{D}/\partial t = i\omega\boldsymbol{D} \end{aligned} \tag{3.5}$$

where we make use of the time derivative of the implicit $\exp(i\omega t)$ time variation.

To simplify the equations it is common to define

$$\mathscr{H} = Z_0\boldsymbol{H} \tag{3.6}$$

where $Z_0$ is the impedance of free space $(\mu_0/\epsilon_0)^{1/2}$. $\mathscr{H}$ measures the magnetic field in terms of the electric field that would be associated with it in a progressive plane wave in free space (Budden, 1988, p. 31). $\boldsymbol{E}$ and $\mathscr{H}$ have the same units.

The latter two of Maxwell's equations can also be written[1]

$$\operatorname{curl}\boldsymbol{E} = -ik\mathscr{H}, \qquad \operatorname{curl}\mathscr{H} = ik\epsilon_0^{-1}\boldsymbol{D}. \tag{3.7}$$

Expanding the curls, these are

$$\begin{aligned}
\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} &= -ik\mathscr{H}_x, & \frac{\partial \mathscr{H}_z}{\partial y} - \frac{\partial \mathscr{H}_y}{\partial z} &= ik\epsilon_0^{-1}D_x \\
\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} &= -ik\mathscr{H}_y, & \frac{\partial \mathscr{H}_x}{\partial z} - \frac{\partial \mathscr{H}_z}{\partial x} &= ik\epsilon_0^{-1}D_y \\
\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} &= -ik\mathscr{H}_z, & \frac{\partial \mathscr{H}_y}{\partial x} - \frac{\partial \mathscr{H}_x}{\partial y} &= ik\epsilon_0^{-1}D_z
\end{aligned} \tag{3.8}$$

We assume that the ionosphere is stratified along the $z$-axis only. A plane wave of generally elliptical polarization is incident from below the ionosphere in the $x$–$z$ plane at an oblique angle $\theta$ to the vertical. $x$ and $y$ dependence of all field components is through $\exp(-ikSx)$ (see Appendix A.2.4, Clemmow and Heading, 1954, p. 323, or Budden, 1988, p. 181). Consequently,

$$\partial/\partial x = -ikS, \qquad \partial/\partial y = 0. \tag{3.9}$$

Substituting these directly into Maxwell's equations:

$$\frac{\partial E_x}{\partial z} = -ik(\mathscr{H}_y + SE_z), \qquad \frac{\partial E_y}{\partial z} = ik\mathscr{H}_x, \qquad SE_y = \mathscr{H}_z \tag{3.10}$$

$$\frac{\partial \mathscr{H}_x}{\partial z} = ik(\epsilon_0^{-1}D_y - S\mathscr{H}_z), \qquad \frac{\partial \mathscr{H}_y}{\partial z} = -ik\epsilon_0^{-1}D_x, \qquad S\mathscr{H}_y = \epsilon_0^{-1}D_z \tag{3.11}$$

We are interested in $\boldsymbol{E}$ and $\mathscr{H}$, so it is necessary to determine the electric displacement vector $\boldsymbol{D}$ in terms of $\boldsymbol{E}$. Banerjea and Saha (1947), summarizing the work of previous authors, show that for propagation in the ionosphere we can use the polarization vector $\boldsymbol{P}$ obtained from the susceptibility tensor $\boldsymbol{M}$ as

$$\boldsymbol{P} = \epsilon_0 \boldsymbol{M}\boldsymbol{E}. \tag{3.12}$$

[1] Work backwards from Eq. (3.7) with the definition of $k$ and $Z_0$ to arrive at the equivalent forms in Eq. (3.5).

The electric displacement contains the polarization vector through

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}. \tag{3.13}$$

Section 3.2.4 discusses the construction of $\boldsymbol{M}$, but for now it is only necessary to know that the $3 \times 3$ susceptibility tensor of the ionosphere can be calculated. These steps are also summarized in Clemmow and Heading (1954), who were the first to present the next step of the derivation using our coordinate system.

The only components that appear in derivatives in Eq. (3.10) or Eq. (3.11) are $E_x$, $E_y$, $\mathscr{H}_x$, and $\mathscr{H}_y$. We may use the last equations in (3.10) and (3.11) to eliminate $E_z$ and $\mathscr{H}_z$, leaving us with four first-order differential equations, consistent with the existence of four characteristic waves in the ionosphere. Let $\boldsymbol{e}$ be a column vector with the four components[2] :

$$\boldsymbol{e} = \left(E_x, -E_y, \mathscr{H}_x, \mathscr{H}_y\right)^\top. \tag{3.14}$$

Then we can write the differential equations for the ionosphere, Eqs. (3.10) and (3.11), in matrix-vector form as

$$\frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}z} = -ik\boldsymbol{T}\boldsymbol{e} \tag{3.15}$$

where (Clemmow and Heading, 1954, p. 323; Budden, 1961a, ch. 18)[3]

$$\boldsymbol{T} = \begin{pmatrix} -\dfrac{SM_{31}}{1+M_{33}} & \dfrac{SM_{32}}{1+M_{33}} & 0 & \dfrac{C^2+M_{33}}{1+M_{33}} \\ 0 & 0 & 1 & 0 \\ \dfrac{M_{23}M_{31}}{1+M_{33}} - M_{21} & C^2 + M_{22} - \dfrac{M_{23}M_{32}}{1+M_{33}} & 0 & \dfrac{SM_{23}}{1+M_{33}} \\ 1 + M_{11} - \dfrac{M_{13}M_{31}}{1+M_{33}} & \dfrac{M_{32}M_{13}}{1+M_{33}} - M_{12} & 0 & -\dfrac{SM_{13}}{1+M_{33}} \end{pmatrix}. \tag{3.16}$$

[2] The minus sign on $-E_y$ results in nice properties for the matrix $\boldsymbol{T}$ (Budden, 1988, sec. 7.14).

[3] Budden (1988, p. 182) notes that Eq. (3.16) can be made even more general by replacing $\boldsymbol{M}$ with the electric permittivity tensor $\boldsymbol{\epsilon}$ (Budden, 1988, eq. 3.55).

Equation (3.15) is the key result of this section and allows us to compute the wavefields of a radio wave in the ionosphere at any height as long as we know the fields at some other height. The integration of Eq. (3.15) is required to compute the mode conversion coefficients in Section 3.4. Equation (3.15) is also used to derive the differential equation for the ionosphere reflection coefficient (Appendix A.4.1) and for computing the Booker quartic, which is used to determine starting values for the wavefields and reflection coefficients.

## 3.2 Mode search

The first task of a mode theory propagation model is to identify resonant modes of a horizontally homogeneous segment of the Earth-ionosphere waveguide. By horizontally homogeneous we mean the ground conductivity, background magnetic field, and density and collision frequency profiles of the ionosphere are constant in the $x$ direction along the propagation path. The number density and collision frequency profiles still vary with height along $z$.

In this section we introduce: 1) the fundamental equation of mode theory, 2) calculation of the reflection coefficient for the ionosphere, and 3) calculation of the ground reflection coefficient. The fundamental equation of mode theory describes the condition for modes to propagate in the waveguide and depends on the reflection coefficients of ground and the ionosphere, which are themselves functions of the wave angle of incidence. Each angle that satisfies the fundamental equation represents a waveguide mode for which we can later compute the electric field in the waveguide.

### 3.2.1 The fundamental equation of mode theory

We begin by considering a waveguide bounded by reflecting planes at $z = 0$ and $z = h$ with reflection coefficients $R_0(\theta)$ and $R_h(\theta)$, respectively. A plane wave is incident on the boundary at $z = h$ with its normal at an angle $\theta$ from the $z$-axis. For detailed background,

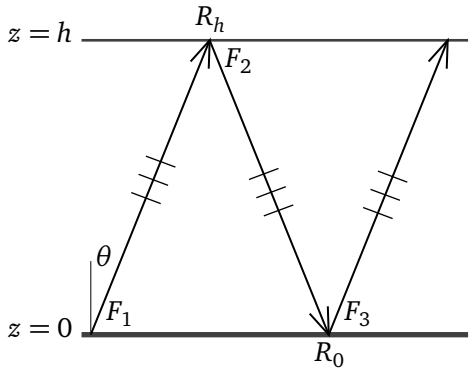


Figure 3.4: For a field to propagate down the waveguide, the twice reflected wave $F_3$ must be identical to the original upgoing wave $F_1$.

read Appendix A.2 for an explanation of reflection coefficients and Appendix A.3 for a brief introduction to waveguides before continuing with this section.

Let some field component $F$ of the incident wave be given by

$$F_1 = F_0 \exp\big(-ik(Sx + Cz)\big) \tag{3.17}$$

where $S = \sin\theta$ and $C = \cos\theta$. When the wave meets the boundary at $z = h$, it generates a reflected wave in which the field is

$$F_2 = R_h(\theta) F_0 \exp\big(-ik(Sx - Cz)\big) \exp(-2ikCh). \tag{3.18}$$

The second exponential ensures that the ratio $F_2/F_1 = R_h$ at $z = h$, as in Eq. (A.13). This second wave travels downwards and is reflected from the boundary at $z = 0$ and gives a third wave in which $F$ is

$$F_3 = R_0(\theta) R_h(\theta) F_0 \exp\big(-ik(Sx + Cz)\big) \exp(-2ikCh). \tag{3.19}$$

This scenario is depicted in Fig. 3.4.

In order for the two crossing waves $F_1$ and $F_2$ to propagate down the waveguide, $F_3$ must be identical with $F_1$. This requires that

$$R_0(\theta) R_h(\theta) \exp(-2ikCh) = 1 \tag{3.20}$$

which is the fundamental equation of mode theory.[4] Each discrete $\theta$ for which this criteria is met is known as an *eigenangle* and represents a waveguide mode. To satisfy the equation when the boundaries are not perfect reflectors, $\theta$ must be complex-valued. Equation (3.20) applies for any boundaries of reflection with a known $R$, even if the boundary is a stratified medium like the ionosphere. We can choose any plane $z = h$ as the boundary and use $R(\theta) = F_2/F_1$ at this plane. This seemingly toy concept is critical for the development of much of the mode theory of the propagation of long waves in the Earth-ionosphere waveguide.

### 3.2.2 Mode equation with change in polarization

As discussed in Appendix A.2.3, in an anisotropic ionosphere the reflection coefficient is actually a complex matrix $\boldsymbol{R}(\theta)$. The mode equation, Eq. (3.20), was derived from the condition that a plane wave which makes a double passage from ground to the ionosphere and back, undergoing two reflections, gives rise to a wave which is in phase with the original wave. If the wave also undergoes a change in polarization on reflection, we add the condition that the twice reflected wave must also match the original wave in polarization (Budden, 1961b, p. 150).

If the column vector $\boldsymbol{e}$ contains the electric field components parallel and perpendicular to the plane of incidence, then

$$\overline{\boldsymbol{R}}(\theta)\boldsymbol{R}(\theta)\boldsymbol{e} = \boldsymbol{I}\boldsymbol{e}, \tag{3.21}$$

where $\overline{\boldsymbol{R}}$ is the reflection matrix looking down at the ground, $\boldsymbol{R}$ is the reflection matrix looking up at the ionosphere, and $\boldsymbol{I}$ is the $2 \times 2$ identity matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, is the condition that the twice reflected wave be identical to the original upgoing wave. This is equivalent to

$$\left(\overline{\boldsymbol{R}}(\theta)\boldsymbol{R}(\theta) - \boldsymbol{I}\right)\boldsymbol{e} = \boldsymbol{0}. \tag{3.22}$$

For a nontrivial eigenvector $\boldsymbol{e}$ to exist, the square matrix $\overline{\boldsymbol{R}}\boldsymbol{R} - \boldsymbol{I}$ must be singular (Budden, 1962, p. 541; Martin, 1965, p. 998). Thus, reintroducing the explicit height term, the modal

[4] Eq. (3.20) satisfies Eq. (A.36) when the boundaries are perfectly reflecting: $\exp(-2ikh\cos\theta) = \exp(-2in\pi) = \cos(2n\pi) - i\sin(2n\pi) = 1$.

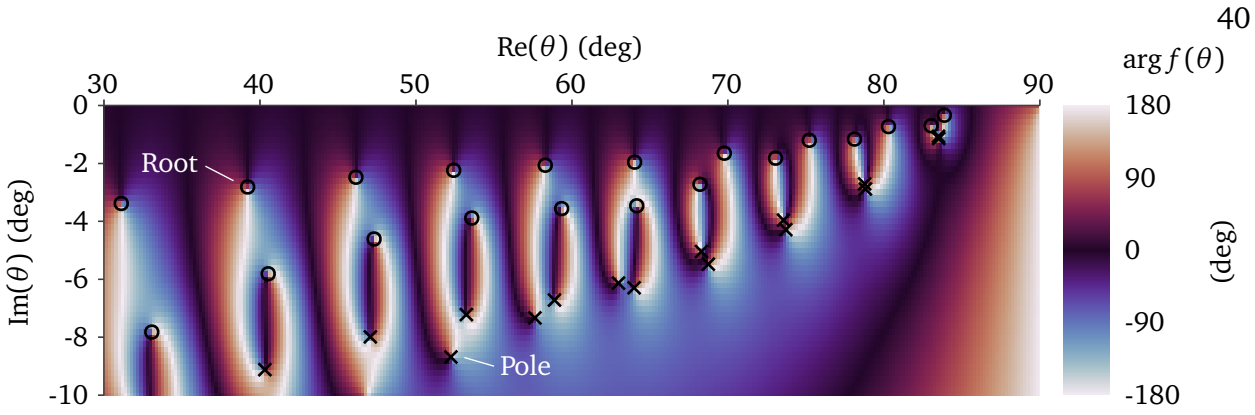


Figure 3.5: Complex phase map of the mode equation $f(\theta)$ (Eq. (3.24)) for the ionosphere profiles shown in Fig. 3.6 and ocean-like ground conductivity. Roots and poles of the function are marked with ∘ and ×, respectively.

equation is

$$\det\left(\overline{\boldsymbol{R}}(\theta)\boldsymbol{R}_h(\theta)\exp(-2ikCh)-\boldsymbol{I}\right)=0 \tag{3.23}$$

where $\boldsymbol{R}_h$ is the ionosphere reflection matrix at height $h$. The effect of the exponential multiplied onto the ionosphere reflection matrix is to reference it to the ground (Budden, 1962, p. 541). If both reflection matrices are already referenced to the ground ($z=0$), then the matrix mode equation is equal to

$$\det\left(\overline{\boldsymbol{R}}(\theta)\boldsymbol{R}(\theta)-\boldsymbol{I}\right)=0. \tag{3.24}$$

A phase map of the left side of Eq. (3.24) is shown in Fig. 3.5. The mode finder must identify each of the eigenangles $\theta$ that are complex roots (zeros) of the left side of Eq. (3.24). Eigenangles with large real components and small imaginary components are the lowest order modes and have the most significant influence on the waveguide fields.

### 3.2.3 Ionosphere reflection coefficients

Budden (1955) presents two methods for obtaining numerical solutions of the differential equations which describe the reflection of long radio waves from the ionosphere. Both methods require integration of differential equations from an initial solution of $\boldsymbol{R}$ at the top of the ionosphere downwards to the space below the ionosphere where the reflection coefficients are

directly available. We proceed with Budden's second method, described by Budden as "greatly superior" to the first because it reduces by half the number of required integrations and reduces the risk of floating point overflow (Budden, 1955, pp. 517, 528). A detailed derivation of the differential equations for $\boldsymbol{R}$ with respect to height is provided in Appendix A.4.1.

The components of the ionosphere reflection coefficient matrix $\boldsymbol{R}$ are

$$\boldsymbol{R} = \begin{pmatrix} {}_{\parallel}R_{\parallel} & {}_{\perp}R_{\parallel} \\ {}_{\parallel}R_{\perp} & {}_{\perp}R_{\perp} \end{pmatrix} \tag{3.25}$$

where the notation ${}_{a}R_{b}$ means subscript $a$ is the incident component and subscript $b$ is the reflected component. The symbol "$\parallel$" indicates the electric field is parallel to the plane of incidence and "$\perp$" indicates the electric field is perpendicular to the plane of incidence. Additionally, in free space

$${}_{\parallel}R_{\parallel} = \mathscr{H}_y^R/\mathscr{H}_y^I, \quad {}_{\parallel}R_{\perp} = E_y^R/\mathscr{H}_y^I \tag{3.26}$$

$${}_{\perp}R_{\parallel} = \mathscr{H}_y^R/E_y^I, \quad {}_{\perp}R_{\perp} = E_y^R/E_y^I \tag{3.27}$$

where superscript $R$ and $I$ represent the reflected and incident fields, respectively.

The differential equations $\mathrm{d}\boldsymbol{R}/\mathrm{d}z$ are derived using the definition of a reflection coefficient and the matrix differential equation for the wave fields, Eq. (3.15). The four elements of $\boldsymbol{e}$ fully describe the total electromagnetic field of the radio wave in the horizontally stratified ionosphere. Budden assumes that there is a very thin slice of vacuum (thin enough so it doesn't make a difference to the fields) in the ionosphere which allows us to describe the wave fields as if they are in free space. They are resolved into upgoing and downgoing elliptical waves, each of which are further resolved into plane wave components aligned to the $x, y, z$ coordinate frame. If $\boldsymbol{S}$ is a transformation matrix to resolve $\boldsymbol{e}$ into these component waves $\boldsymbol{f}$, such that $\boldsymbol{e} = \boldsymbol{S}\boldsymbol{f}$, then we can rewrite Eq. (3.15) as

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}z} = -i\boldsymbol{S}^{-1}\boldsymbol{T}\boldsymbol{S}\boldsymbol{f} = -\frac{1}{2}i\boldsymbol{W}\boldsymbol{f}. \tag{3.28}$$

Budden (1955) then makes use of the definition of the reflection coefficient matrix through $\boldsymbol{d} = \boldsymbol{R}\boldsymbol{u}$ for downgoing wavefields $\boldsymbol{d} = (f_3, f_4)^\top$ and upgoing wavefields $\boldsymbol{u} = (f_1, f_2)^\top$ to derive

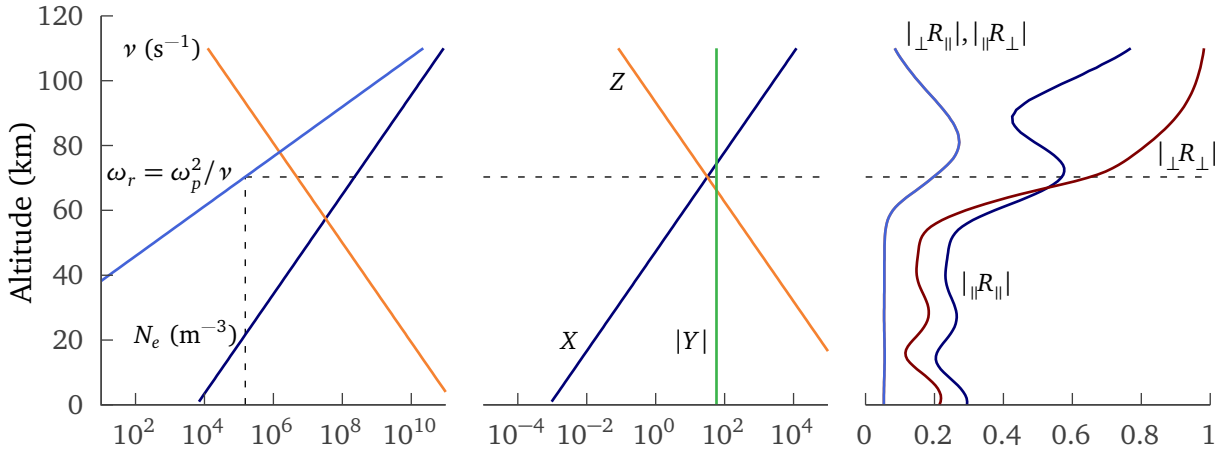


Figure 3.6: Magnetoionic parameters $X$, $Y$, and $Z$ directly track with $N_e$, $B$, and $\nu$. Wait's conductivity parameter $\omega_r$ is also shown (Wait & Spies, 1964). The profiles are for Wait's exponential density profile parameters $h' = 72\,\text{km}$, $\beta = 0.3\,\text{km}^{-1}$, vertical $|\boldsymbol{B}| = 50\,000\,\text{nT}$, and a 24 kHz radio wave at real angle of incidence $\theta = 75°$ from the vertical. The off-diagonal components of the ionosphere reflection coefficient matrix are identical because of the vertical magnetic field.

the nonlinear matrix differential equation

$$\frac{2i}{k}\frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}z} = \boldsymbol{W}_{21} + \boldsymbol{W}_{22}\boldsymbol{R} - \boldsymbol{R}\boldsymbol{W}_{11} - \boldsymbol{R}\boldsymbol{W}_{12}\boldsymbol{R} \tag{3.29}$$

for the four $2 \times 2$ submatrices of $\boldsymbol{W}$.

An initial $\boldsymbol{R}$ can be computed at a point high in the ionosphere from the Booker quartic (see Appendix A.4.2 and Appendix A.4.3), then Eq. (3.29) can be integrated downwards through the ionosphere into free space below. At any height where the integration stops, $\boldsymbol{R}$ at that height is the reflection coefficient of the ionosphere as though it were sharply bounded by free space below. Profiles of the components of $\boldsymbol{R}$ as it is integrated downwards through an ionosphere with exponential density and collision frequency profiles are shown in Fig. 3.6.

### 3.2.4 Constitutive relations and susceptibility

So far we have assumed the existence of the ionosphere susceptibility tensor $\boldsymbol{M}$, but in this section we will briefly explain what the susceptibility tensor physically describes. Within

the whole of EIWG mode theory, nearly all of the magnetoionic physics appears through the susceptibility tensor.

Looking back to Maxwell's equations, Eq. (3.8), we would like to express the displacement $\boldsymbol{D}$ and, by extension, the electric polarization $\boldsymbol{P}$, in terms of the electric intensity $\boldsymbol{E}$. The resulting expressions are known as the constitutive relations where the relation between $\boldsymbol{P}$ and $\boldsymbol{E}$ is through the susceptibility matrix $\boldsymbol{M}$:

$$\boldsymbol{P} = \epsilon_0 \boldsymbol{M} \boldsymbol{E}. \tag{3.30}$$

In Appendix A.5 we build up the net electric polarization resulting from the influence of an imposed electric field, electron-neutral collisions, and a background magnetic field. We assume electrons move over a background grid of infinitely massive positively charged particles such that the overall collection of particles is homogeneous and neutrally charged. It is straightforward to derive the equations of motion of a single electron subject to each of the above three external forces and extend the results to a collection of electrons.

The resulting susceptibility tensor is:

$$\boldsymbol{M} = -\frac{X}{U(U^2 - Y^2)} \begin{pmatrix} U^2 - l_x^2 Y^2 & -il_z YU - l_x l_y Y^2 & il_y YU - l_x l_z Y^2 \\ il_z YU - l_x l_y Y^2 & U^2 - l_y^2 Y^2 & -il_x YU - l_y l_z Y^2 \\ -il_y YU - l_x l_z Y^2 & il_x YU - l_y l_z Y^2 & U^2 - l_z^2 Y^2 \end{pmatrix} \tag{3.31}$$

where $X = Ne^2/\epsilon_0 m\omega^2 = \omega_n^2/\omega^2$, $Y = |e\boldsymbol{B}/m\omega| = \omega_H/\omega$, $Z = \nu/\omega$, and $U = 1 - iZ$ for angular wave frequency $\omega$, angular plasma frequency $\omega_n$, and angular gyrofrequency $\omega_H$. $l_x$, $l_y$, and $l_z$ are direction cosines of the magnetic field vector with respect to the EIWG coordinate system. Thus, the susceptibility is determined by the species density, collision frequency, and background magnetic field.

The influence of multiple ionospheric species, e.g. electrons and negative and positive ions of arbitrary mass and charge, can be considered by computing a susceptibility tensor for each species individually and summing them (Budden, 1988). In practice, only the magnetoionic parameters $X$, $Y$, and $Z$ need to be computed for each species before forming the total $\boldsymbol{M}$. The

derivation of $\boldsymbol{M}$ assumed charge neutrality and that should be maintained when using multiple species, i.e. $\sum N_{e,-i} = \sum N_{+i}$ for ions charged $\pm i$.

There are some limitations to the derived susceptibility tensor. First, we neglect the effect of the radio wave's magnetic field on the motion of electrons. Budden (1988, ch. 3.7) examines this scenario for typical radio transmitters and finds the relative influence of the magnetic field is usually less than $10^{-4}$ times the influence of the electric field and is reasonable to neglect. Nonetheless, ionospheric heating experiments and lightning may generate significant magnetic fields that influence the propagation of the radio wave through its interaction with the ionosphere. Another simplifying assumption used to derive Eq. (3.31) is that the electron-neutral collision frequency $\nu$ is independent of the electron velocity $v$. The collision frequency above is an averaged effective collision frequency, but collision frequency is actually a function of velocity, which in turn is a function of temperature (Budden, 1988, ch. 3.12). Finally, although electron-electron collisions are unlikely to have a meaningful impact on the propagation of radio waves in the ionosphere, there may be higher order effects from electron-ion collisions that occur for radio frequencies near the ion gyro-frequency (Budden, 1988, p. 62). Each of these are neglected in the susceptibility tensor above, but have minimal influence on the propagation of typical communication or navigation radio signals.

Although not directly related to the ionosphere, one way to include the effect of Earth's curvature is by means of a fictitious medium in the free space of the Earth-ionosphere waveguide. This medium has an index of refraction $\mathfrak{n}$ that varies as a function of height such that it mimics the path of a plane wave propagating through free space over a surface curving out from under it. This correction can be applied directly as a modification to the diagonal elements of the susceptibility tensor, which is related to $\mathfrak{n}^2$. For details, see Appendix A.5.4.

At this point, all of the steps necessary to compute the ionosphere reflection coefficient $\boldsymbol{R}$ over a curved Earth have been introduced.

### 3.2.5 Ground reflection coefficients

The second half of the fundamental equation of mode theory requires computation of the ground reflection coefficient matrix. We consider the ground to be a simple isotropic boundary with a reflection coefficient $\overline{\boldsymbol{R}}$ that can be described by the Fresnel reflection equations,[5] derived in Appendix A.6.

The resulting ground reflection coefficient matrix is

$$\overline{\boldsymbol{R}} = \begin{pmatrix} {}_{\parallel}\overline{R}_{\parallel} & 0 \\ 0 & {}_{\perp}\overline{R}_{\perp} \end{pmatrix} \tag{3.32}$$

where

$${}_{\parallel}\overline{R}_{\parallel} = \frac{C\mathfrak{n}_g^2 - \left(\mathfrak{n}_g^2 - S^2\right)}{C\mathfrak{n}_g^2 + \left(\mathfrak{n}_g^2 - S^2\right)} \tag{3.33}$$

$${}_{\perp}\overline{R}_{\perp} = \frac{C - \left(\mathfrak{n}_g^2 - S^2\right)^{1/2}}{C + \left(\mathfrak{n}_g^2 - S^2\right)^{1/2}} \tag{3.34}$$

and $\mathfrak{n}_g^2$ is the squared refractive index of the ground (see Appendix A.6.3)

$$\mathfrak{n}_g^2 = \epsilon_r - i\frac{\sigma}{\omega\epsilon_0}. \tag{3.35}$$

## 3.3 Waveguide excitation and field strengths

This section is arguably both the most important and mathematically complicated. Up to this point we have established much of the formalism for reflection coefficients, waveguide propagation, and magnetoionic physics. In this section we answer the question we are ultimately interested in: *for a particular VLF radio source located in the waveguide, what electromagnetic fields would we measure at a point elsewhere in the waveguide*? The theory used here is largely credited to Budden (1962) and his use of complex residue theory to solve some of the integrals associated with fields radiated from the source. Wait (1970a, ch. 6 and 7) develops the theory

[5] Galejs (1972) examines the effect of a more realistic multi-layer, anisotropic ground.

analytically in cylindrical and spherical coordinates. Readers are encouraged to seek out these sources, as well as Budden (1961b), for details on the derivation of the theory; familiarity with complex analysis as well as some antenna theory is necessary for understanding the details. The interested reader should see Appendix A.7 for some of the intermediate mathematics used to describe EIWG mode theory.

Qualitatively, it is easy to imagine that a line quadrupole source placed in free space will emit fields which can be represented as an angular spectrum of plane waves. If the source is located in a waveguide, the field measured at a distant point in the waveguide is the result of summing the up- and downgoing waves as a result of reflecting off the lower and upper waveguide boundaries, integrated over the complex angular spectrum emitted by the source. The waveguide boundaries will filter out a discrete set of these waves that satisfy the mode equation. Budden (1962) rigorously works through this process and finds that complex poles crossed by the contour of integration of the field sum are the waveguide modes and the residues are excitation factors that describe how a given field component is excited by this particular waveguide. Geometric spreading due to Earth's spherical shape is considered, as are height gain functions, describing the vertical field distribution, to model elevated transmitters or receivers.

Over time, the expression for the total waveguide field of a dipolar radio source has been distilled into a practical form. An arbitrarily oriented transmitting antenna is modeled as the geometrically-weighted sum of fields emitted from a vertical, end-on, and broadside oriented dipole. The total $j$ component of the $E$-field in the waveguide is then:

$$\begin{aligned} E_j(x) = \frac{Q}{\big(\sin(x/R_e)\big)^{1/2}} \sum_n \Big( & \lambda_{v,n} \cos(\gamma) f_{1,n}(z_t) \\ & + \lambda_{b,n} \sin(\gamma) \sin(\phi) f_{3,n}(z_t) \\ & + \lambda_{e,n} \sin(\gamma) \cos(\phi) f_{2,n}(z_t) \Big) f_{j,n}(z_r) \exp\big(-ik(S_n - 1)x\big) \end{aligned} \tag{3.36}$$

| | | |
|---|---|---|
| mode | $n$ | |
| electric field component | $j$ | $x$ |
| | | $y$ |
| | | $z$ |
| excitation factor | $\lambda$ | $v$ vertical source |
| | | $b$ broadside |
| | | $e$ end-on |
| source orientation | $\gamma$ | dipole moment angle from vertical |
| for | $\phi$ | angle from $x$ |
| height gain function | $f$ | $1 = x$ component |
| | | $2 = y$ |
| | | $3 = z$ |
| height | $z$ | $r$ receiver |
| | | $t$ transmitter |
| great circle distance from $t$ to $r$ | $x$ | |
| sine of *ground*-referenced eigenangle | $S$ | |
| source amplitude scalar | $Q$ | |

The factor to the left of the summation corrects for the radiated power of the transmitter and spreading due to the spherical shape of the Earth. The terms grouped by the large parentheses collectively model the transmitter excitation of the waveguide. The height gain function evaluated at $z_r$ models the receiver by selecting a specified field component at a particular height in the waveguide. Finally, the exponential is a propagation factor that describes the propagation of the mode in the horizontal direction along the waveguide.

This formulation can be attributed to Pappert, who over a series of papers calculated the height gains and excitation factors for the $E_x$, $E_y$, and $E_z$ fields and arbitrarily positioned and oriented antennas in the guide (Pappert, 1968; 1970; Pappert & Bickel, 1970; Pappert & Shockey, 1971; Pappert & Shockey, 1976). Multiple formulations have been used in the waveguide propagation codes developed by the U.S. Navy. It is important that the height gain and excitation factors used are self-consistent. In addition, most formulations, including this one, assume the reflection coefficient is referenced to the ground. The excitation factors $\lambda$ and height gain functions $f$ are presented in Appendix A.7.

## 3.4 Mode conversion

Up to this point we have assumed the waveguide is homogeneous—that the electromagnetic properties of the ground and ionosphere are constant in the horizontal axis of the waveguide. In reality, both the ground and ionosphere change over a propagation path; for the ground, this might be a sharp transition from soil to ocean water, and for the ionosphere, this might be a day/night transition. The terminator transition is relatively slow compared to the soil/ocean transition; nonetheless, a stepped model with several homogeneous segments of waveguide can be used to represent the true smooth transition.

Crombie (1964) observed periodic fading of VLF signals received over paths during sunset and sunrise. To explain his observations, he suggests that a mode propagating with a phase velocity determined by the conditions on the day side are converted into two waveguide modes supported by the night side waveguide. This is largely considered to be an accurate description of the process (Walker, 1965; Clilverd et al., 1999; Chand & Kumar, 2017). Therefore, to model VLF propagation in inhomogeneous waveguides, we model conversion of energy scattered from each mode into other modes at every sharp transition in the waveguide, as shown in Fig. 3.1.

Assume a mode of unit amplitude is incident upon a discontinuity in the waveguide. This mode can be described by its height gain function $\overrightarrow{f}_{j,1}(z)$, which is a vector of electromagnetic fields at each height in the waveguide. Subscript $j$ denotes the $j^{\text{th}}$ mode and subscript 1 denotes it is in waveguide segment 1. The overarrow denotes the direction of propagation. The sharp impedance discontinuity in the waveguide generates both reflected, back-propagating modes, as well as transmitted, forward-propagating modes. Discontinuities that might exist further down the guide generate additional backward-propagating modes. Continuity yields

$$\overrightarrow{f}_{j,1} + \sum_m R_{mj} \overleftarrow{f}_{m,1} = \sum_n T_{nj} \overrightarrow{f}_{n,2} + \sum_l R_{lj} \overleftarrow{f}_{l,2} \tag{3.37}$$

where every height gain function is a function of $z$. $R_{ab}$ and $T_{ab}$ are reflection and transmission coefficients describing conversion of mode $b$ into mode $a$ for modes $j$, $l$, and $m$. Note that

$R$ here is not the reflection coefficient of the ionosphere, but a reflection coefficient between waveguide segments.

One problem with using Eq. (3.37) to implement mode conversion is the requirement to know the future propagation history—the term with $R_{lj}$. By assuming that no significant reflection occurs, then we can set $R_{lj}$ and $R_{mj}$ to 0. This is known as Kirchoff's approximation and is used throughout the literature (Wait, 1970b). Equation (3.37) then simplifies to

$$\vec{f}_{j,1} = \sum_n T_{nj} \vec{f}_{n,2}. \tag{3.38}$$

Clearly, if we can solve for $T_{nj}$, then we can describe the scattering of energy from each incident mode $j$ into each transmitted mode $n$. The method we follow is known as FULLMC for "full mode conversion" because it explicitly integrates the wavefields in the neighboring segments in order to establish the mode conversion coefficients (Pappert & Smith, 1972). Appendix A.8 explains the process to compute the conversion coefficients in greater detail. Ultimately, the mode sum equation for an inhomogeneous waveguide is

$$\begin{aligned} E_{j,p}(x) = \frac{Q}{\big(\sin(x/R_e)\big)^{1/2}} \sum_m \sum_n a_{m,n,p} \Big( & \lambda_{v,n} \cos(\gamma) f_{1,n}(z_t) \\ & + \lambda_{b,n} \sin(\gamma) \sin(\phi) f_{3,n}(z_t) \\ & + \lambda_{e,n} \sin(\gamma) \cos(\phi) f_{2,n}(z_t) \Big) f_{j,m}(z_r) \\ & \cdot \exp\big(-ik\big(S_{n,\mathrm{tx}} x_{\mathrm{tx}+1} + S_{m,p}(x - x_p)\big)\big). \end{aligned} \tag{3.39}$$

where $a_{m,n,p}$ is the mode conversion coefficient from mode $m$ to $n$ in the $p^{\text{th}}$ segment. The homogeneous mode sum, Eq. (3.36), should be used within the transmitter slab. Equation (3.39) should be used in waveguide segment $p$ where $p$ occurs after the transmitter segment. The excitation factors making up the transmitter term are all defined within the transmitter segment and the other terms of Eq. (3.39) carry the fields into segment $p$.

## 3.5 Amplitude and phase of the field

The amplitude and phase of the electric field is usually specified rather than the complex field. The amplitude is defined as

$$A = 20\log_{10}(|E|) \tag{3.40}$$

and the phase, relative to free space, is

$$\phi = \arg(E). \tag{3.41}$$

The phase-distance curve is unwrapped in such a way that it accumulates the remainder of $\phi_i - \phi_{i-1}$ after division by $2\pi$. It is not unusual for phase profiles to cycle through $2\pi$ rad.

# Chapter 4

## Implementation and validation of the Longwave Mode Propagator

The mode theory model presented in Chapter 3 is an analytical solution for the electric field strength in the Earth-ionosphere waveguide aided by numerical integration of differential equations and iterative root finders. For several decades, the standard propagation model in the VLF community has been the mode theory model called the Long-Wavelength Propagation Capability (LWPC). LWPC was developed in the early days of modern computers and significant effort was devoted to maximizing compute efficiency and minimizing runtime. From then to now, computer processing power has increased by several orders of magnitude (Schaller, 1997). Sheddy et al. (1968) states that a predecessor code to LWPC took 10 minutes to find one solution of the mode equation—a process that takes less than 1 second on a personal computer today.

The speed enabled by mode theory to simulate longwave propagation makes it an obvious choice as forward model for estimation of the ionosphere or for communication or navigation link planning over a large geographic region. Unfortunately, LWPC has several shortcomings. Its mode finder, described in Morfitt and Shellman (1976) and Shellman (1986), interpolates $\boldsymbol{R}$ across the complex plane in what is described as the "weakest link" in the model. The root finding algorithm assumes there are no poles in the region of the complex plane being searched for roots. Therefore, to improve the likelihood that this criteria is met, the mode equation and equations for $\boldsymbol{R}$ and $\overline{\boldsymbol{R}}$ are modified. Numerical difficulties can occur in the LF band (Pappert, 1981) and I have observed numerical problems or program failure for ionosphere profiles with

low Wait $\beta$ parameter. Additional downsides to LWPC are mentioned in Gasdia and Marshall (2021).

A new implementation of EIWG mode theory called LongwaveModePropagator.jl[1] (LMP) has been constructed for this dissertation in the Julia programming language (Bezanson et al., 2017; Gasdia & Marshall, 2021). The most significant difference between LWPC and LMP is the use of the global complex roots and poles finding (GRPF) algorithm to identify the eigenangles (Kowalczyk, 2018). GRPF allows LMP to directly solve the physical mode equations shown in the previous chapter because it differentiates between complex roots and poles. LWPC uses modified forms of the fundamental mode equation, ground reflection, and differential equation for the ionosphere reflection coefficients. Using the physical equations and identifying the complex poles greatly simplifies the code base of LMP, as does the use of external libraries and Julia's native linear algebra capabilities. Mode finding with GRPF is also significantly more robust than LWPC's mode finder; $\boldsymbol{R}$ is computed explicitly rather than interpolating across the complex plane. Although this helps ensure accuracy, it is more computationally expensive than interpolation.

This chapter discusses the implementation of mode theory in LongwaveModePropagator.jl, validates the model against LWPC and a two-dimensional finite difference time domain (FDTD) code, and presents examples of longwave propagation in different EIWG scenarios.

## 4.1 Implementation

There are three steps involved in simulating the propagation of longwaves in the EIWG with LongwaveModePropagator.jl: 1) the user defines the scenario, 2) waveguide modes are identified for all homogeneous segments in the scenario, and 3) the desired field is summed over each mode, including mode conversion between homogeneous segments if the receiver is not located in the same segment as the transmitter.

[1] github.com/fgasdia/LongwaveModePropagator.jl

Every EIWG scenario in LMP consists of a `Transmitter`, `Receiver`, and `Waveguide`.[2] The `Transmitter` is defined by a dipole antenna with specified orientation, transmit frequency, and radiated power, and the `Receiver` specifies a field component and is located some distance from the transmitter. Two `Waveguide` types are currently supported. A `HomogeneousWaveguide` defines the `BField` magnitude and direction, conductivity and relative permittivity of the `Ground`, and one or more `Species` that make up the ionosphere. Each `Species` has a defined charge, mass, number density profile, and collision frequency profile. A `SegmentedWaveguide` simply consists of a series of `HomogeneousWaveguides`, each beginning a specified distance from the transmitter.

The `Transmitter`, `Receiver`, and `Waveguide` are passed to LMP's `propagate` function which returns the complex electric field, amplitude, and phase that would be observed by the `Receiver`. A brief outline of the mode search and field sum computation performed internally by `propagate` follows.

### 4.1.1 Mode search

A theoretical discussion of the waveguide mode equation has already been presented in Chapter 3 with further detail in Appendix A. Because of the global complex root and pole finding (GRPF) algorithm, LMP's mode search follows very closely to theory (Gasdia & Marshall, 2021). The GRPF algorithm is implemented in `RootsAndPoles.jl` (Gasdia, 2020). The algorithm works by sampling the mode function (the left side of Eq. (3.24)) at the nodes of a triangular mesh grid. The complex phase of the function is analyzed and a discretized version of Cauchy's argument principle is applied to the phase map to detect the presence of roots and/or poles in regions of the mesh. The argument principle relates the difference between the number of roots and poles of a function $f$ to the contour integral

$$q = \frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z)} \, \mathrm{d}z \tag{4.1}$$

[2] See the LongwaveModePropagator.jl documentation at https://fgasdia.github.io/LongwaveModePropagator.jl/stable/ for similar types with special characteristics.

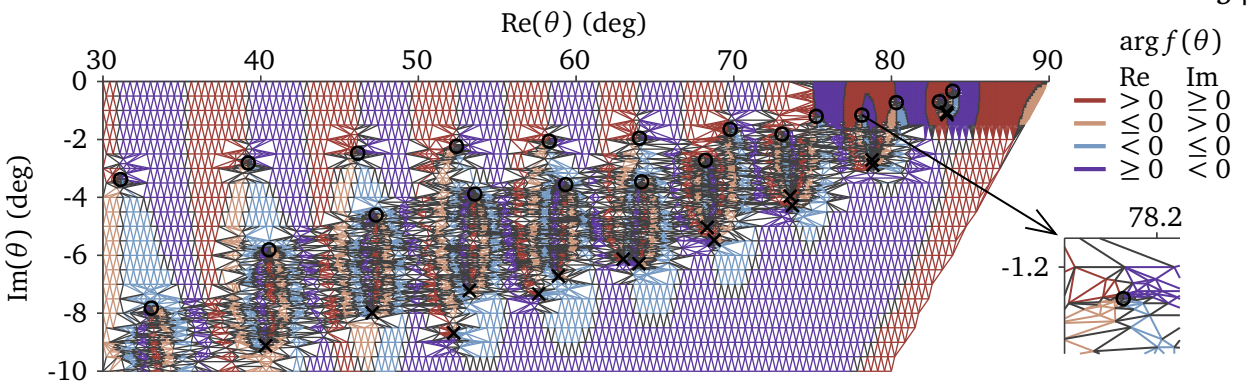


Figure 4.1: The final mesh grid used by the global complex root and pole finding (GRPF) algorithm applied to an Earth-ionosphere waveguide scenario. Roots of the mode equation are identified with ∘ and poles with ×. This figure is the GRPF counterpart to Fig. 3.5, which shows the mode equation phase map on a fine dense grid.

where $q$ is the number of roots minus the number of poles contained within the contour.[3] Kowalczyk (2018) discretizes Eq. (4.1) to use the difference in quadrant of the complex phase along a contour of mesh nodes. If $q$ is zero, there are either no roots or poles contained in the contour or there are an equal number of roots and poles, in which case the initial mesh grid was too coarsely sampled for the problem. However, for non-zero $q$, the mesh grid is automatically refined by Delaunay triangulation to obtain a better estimate of the location of the root or pole. This refinement may split a pair of a closely spaced root and pole. The sign of $q$ identifies whether a root or a pole has been located. LongwaveModePropagator.jl assigns each root as an `EigenAngle` and simply ignores the poles that have been identified.

An example of GRPF applied to the same waveguide as Fig. 3.5 is shown in Fig. 4.1. The top right quadrant of the search region uses a denser initial mesh to identify closely spaced root and pole pairs that often appear in this region of the complex plane. These are the lowest order modes, so it is particularly important that they be identified. The lower right triangular region is not included in the initial mesh because no roots occur here.

The mode equation itself is a single line of code; the ground reflection coefficient $\overline{\boldsymbol{R}}$ is also

[3] Technically, $q$ represents the difference of the sum of multiplicities of the roots and of the poles. See any textbook on complex analysis for further details.

trivial to compute. However, computing the ionosphere reflection coefficient is extremely costly because each evaluation requires an integration of four differential equations (one for each entry of the matrix). Because $\mathrm{d}\boldsymbol{R}/\mathrm{d}z$ is complex valued, it is as if we are actually integrating *eight* differential equations to compute $\boldsymbol{R}$ for a single complex angle $\theta$. During the mode search, $\boldsymbol{R}$ is calculated at every mesh node—usually tens of thousands of $\theta$s are searched for every homogeneous segment of ionosphere. The time required for this search represents much more than 90 % of the total LongwaveModePropagator.jl runtime. Reducing the runtime of the mode search without sacrificing robustness of the current implementation remains one of the most lucrative areas for improvement.

The starting solution for the ionosphere reflection coefficient $\boldsymbol{R}$ is computed at 110 km altitude from the upgoing Booker wavefields. The Booker quartic is solved from the susceptibility tensor at 110 km using the root finder package PolynomialRoots.jl (Skowron & Gould, 2012). $\mathrm{d}\boldsymbol{R}/\mathrm{d}z$ is integrated down to the ground ($z = 0$) using the `Vern7` 7/6 Runge-Kutta method from OrdinaryDiffEq.jl (Verner, 2010; Rackauckas & Nie, 2017). Comparisons of different solvers and tolerances are shown in Appendix B.1. Unlike LWPC, LongwaveModePropagator.jl always integrates to the ground; there is no need to determine an optimal height below the ionosphere at which to reference the reflection coefficients to minimize the presence of complex poles, and there is little benefit to stop the integration early and use modified Hankel functions to re-reference the reflection coefficients (as is done by LWPC) because the solver is very efficient when the electron density is low.

### 4.1.2 Mode field sum

Unlike the mode search, the mode field sum in LongwaveModePropagator.jl is very similar to LWPC because they both use Pappert's derivations of excitation factor and height gain functions. Due to the presence of $\mathrm{d}f/\mathrm{d}\theta$ in the excitation factor, where $f$ is the modal equation, LongwaveModePropagator.jl also includes explicit analytical derivatives with respect to $\theta$ for many of the functions involved in computing the mode equation. These were largely

computed using a computer algebra system and are checked against finite difference solutions in LongwaveModePropagator.jl's test suite. The excitation factors and height gains are formulated such that there is a transmitter term and a receiver term. This makes extension to inhomogeneous `SegmentedWaveguides` simpler. As shown mathematically in Appendix A.7 and Appendix A.8, the transmitted $\mathscr{H}_y$ field is simply the transmitter term in the transmitter slab, and it is carried into the following slabs by applying the mode conversion coefficients. The receiver term is the height gain function for the receiver multiplied by the appropriate correction factor from $\mathscr{H}_y$ to the field component measured by the receiver.

Mode conversion coefficients for segmented waveguides are calculated using the FULLMC theory (Pappert & Smith, 1972). For each homogeneous segment, waveguide modes are first identified and then the wavefields and adjoint wavefields[4] are integrated using the orthonormalization method suggested by Pitteway (1965). This integration occurs downwards from a great height in the ionosphere where there are only upgoing waves. Equation (3.15) is integrated simultaneously for wavefield vectors $\boldsymbol{e}$ for both upgoing waves. Unfortunately, naïve integration of these fields results in numerical issues as the values representing the wavefields increase with the downwards integration. Both an evanescent and travelling wave are present, but the evanescent wave grows by so many orders of magnitude more than the travelling wave as to exceed the machine precision of IEEE 64-bit floating point operations. This is known as "numerical swamping" of the travelling wave solution. To overcome this, Pitteway (1965) repeatedly constrains the second wavefield to be orthogonal to the first. After completing the integration, each orthonormalizing correction must be undone to reveal the correct wavefields. An example of the wavefields scaled and with scaling removed is shown in Fig. 4.2.

The `Tsit5` 4/5 Runge Kutta solver is used to integrate $\mathrm{d}\boldsymbol{e}/\mathrm{d}z$ and the fields are scaled to the waveguide boundary conditions identically to LWPC. Various differential equation solvers were compared for integration of the wavefields and are shown in Appendix B.2.

[4] The adjoint waveguide is identical to the original waveguide except for a change of sign of the $x$ component of the magnetic field.

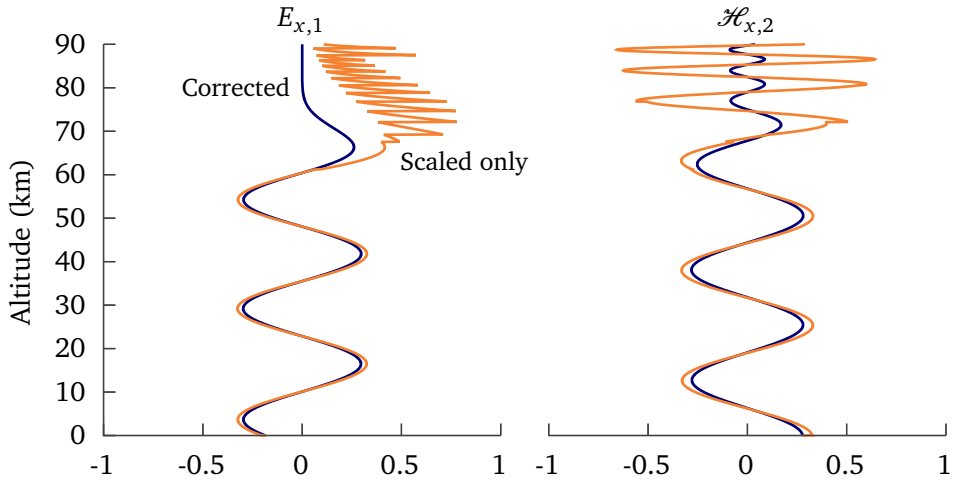


Figure 4.2: A sample of one of each of the upgoing wavefields for the ionosphere shown in Fig. 3.6, but at a real angle of incidence of 60°. The orthonormalization procedure from Pitteway (1965) is applied to the "Scaled only" curve during downwards integration through the ionosphere and then "Corrected" on the way back up from the ground.

## 4.2 Validation

LongwaveModePropagator.jl was validated against LWPC and the EMP2D finite difference time domain (FDTD) model described in Marshall et al. (2017). Results are shown in Gasdia and Marshall (2021), but some highlights are presented here. First, it is important to note that these are comparisons and not necessarily a measure of correctness of the LongwaveModePropagator.jl model. There are limitations that manifest as errors in all three models.

The mean absolute deviation (MAD) of $E_z$ amplitude between each combination of the three models are shown in Fig. 4.3 over a path between 400 to 3000 km from the transmitter for different combinations of Wait ionosphere profiles. The scenarios all use homogeneous waveguides having an ocean-like ground of $\epsilon_r = 4$, $\sigma = 81\,\mathrm{S\,m^{-1}}$ and a vertical magnetic field of 50 000 nT. The transmit frequency is 24 kHz. The source used in EMP2D was not calibrated to produce the same radiated power as LWPC and LMP, so was shifted to the level where the MAD is minimized to LWPC or LMP in their respective plots. Thus, EMP2D is useful for characterizing

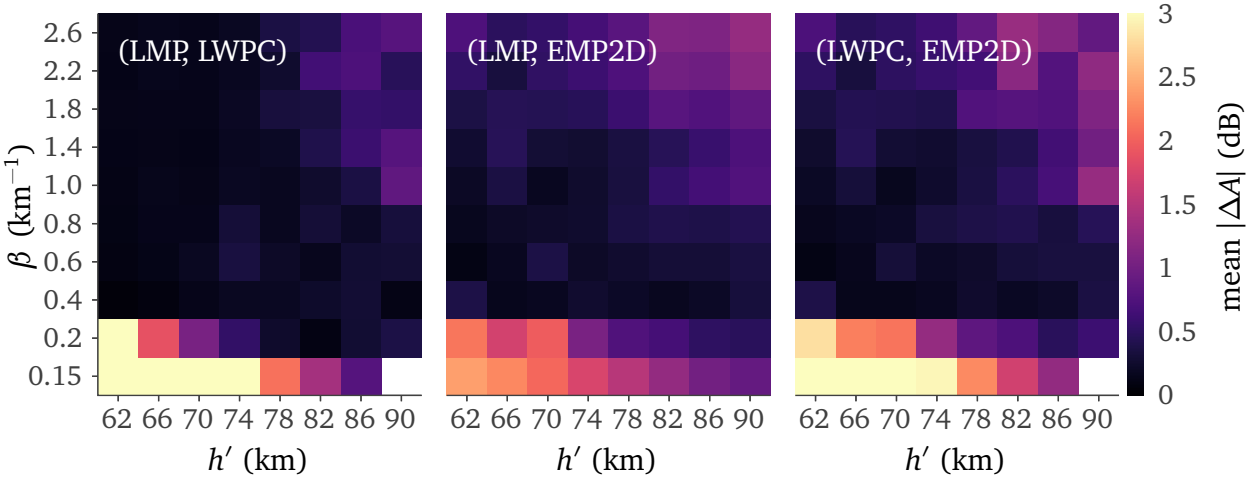


Figure 4.3: Mean absolute deviation of amplitude along a propagation path 400 to 3000 km from the transmitter for Wait ionospheres between each combination of pairs of the LMP, LWPC, and EMP2D models. Note that the $\beta$ values are not evenly distributed. LWPC failed for one of the scenarios with $\beta = 0.15\,\mathrm{km}^{-1}$.

relative changes in amplitude or phase along the propagation path, but not for establishing the true field amplitude.

All three models are in good agreement for "typical" Wait ionospheres (Section 2.1.2.3) between $\beta = 0.4$ to $0.8\,\mathrm{km}^{-1}$ and the entire $h'$ range plotted, from 62 to 90 km. The mode theory models are a better match to EMP2D at lower $h'$ than higher $h'$ for high $\beta$ values. The greatest discrepancies occur for low $\beta$ ionospheres, but in general LMP is a better match to EMP2D than LWPC is. This can be seen in Fig. 4.4, where the slope of the amplitude decay in LMP is closer to the slope of the EMP2D curve despite missing a significant null at 500 km. An expanded mode search region and finer initial mesh grid were not able to identify any "missing" modes from the plotted results.

Another interesting model comparison is for a segmented ionosphere that coarsely models a transition from a daytime to nighttime ionosphere using Wait's parameters. The $E_z$ amplitude curves are shown in Fig. 4.5. All three models are in excellent agreement along the path. This not only validates the mode conversion implementation of LMP, but demonstrates that neglecting backwards reflecting waves (an assumption made by FULLMC) is a valid assumption

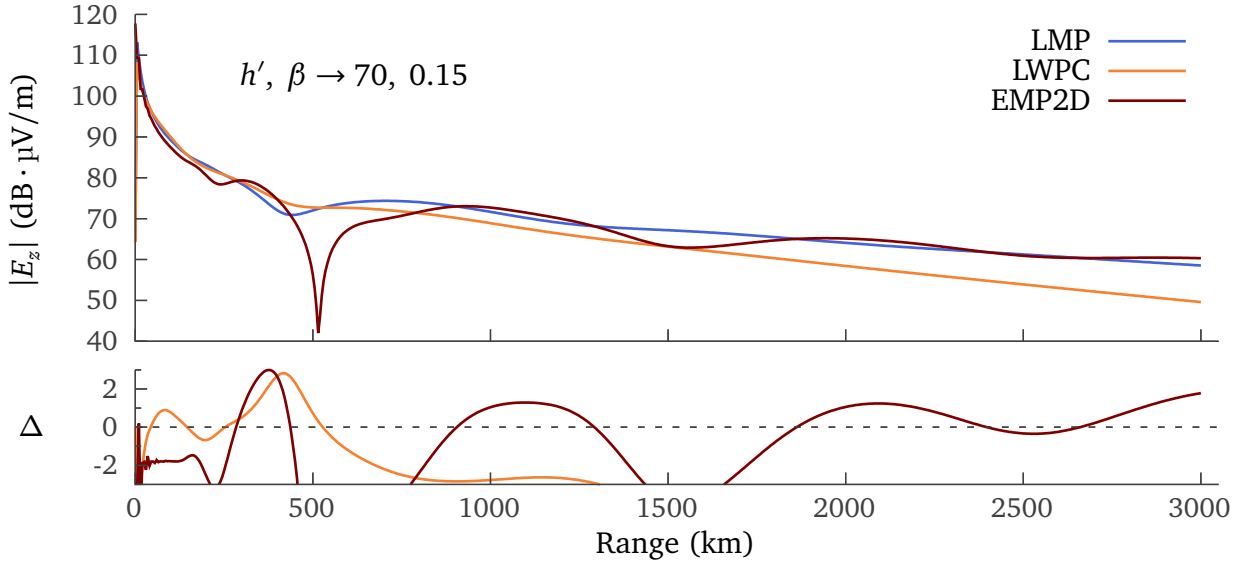


Figure 4.4: Top: $E_z$ amplitude curves for the mode theory and EMP2D FDTD models for a Wait ionosphere with $h' = 70\,\text{km}$ and $\beta = 0.15\,\text{km}^{-1}$. The EMP2D model is shifted to minimize its mean absolute deviation from the LMP curve. Even if it was shifted to minimize the distance to the LWPC curve, the slope of the LMP curve is a better match to EMP2D. Bottom: The amplitude difference curves between LWPC and EMP2D compared to LMP. MODIFIED FROM GASDIA AND MARSHALL (2021), ©IEEE 2021.

for reasonable changes in $h'$ and $\beta$ along the path. EMP2D includes the effect of backward propagating waves automatically through the FDTD solution of Maxwell's equations.

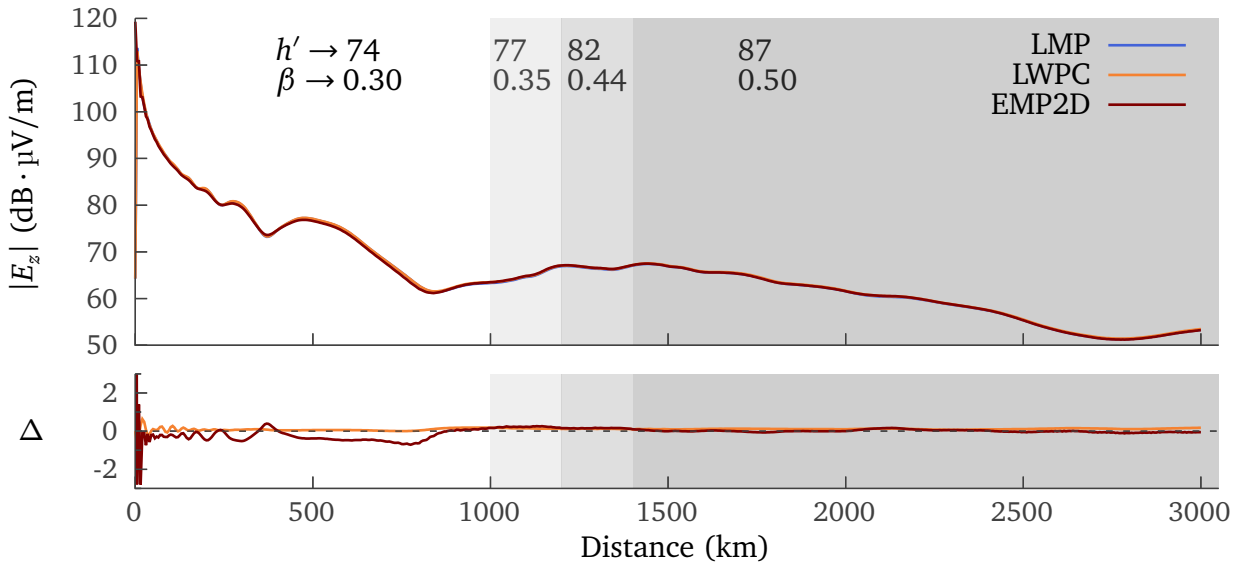


Figure 4.5: Top: $E_z$ amplitude along the ground for a day-to-night segmented ionosphere over an ocean-like ground. The $h'$ and $\beta$ parameters for each segment are marked along the figure with four segments shaded in increasingly darker gray for the transition from day to night. Bottom: The amplitude difference from LWPC and EMP2D to LMP. MODIFIED FROM GASDIA AND MARSHALL (2021), ©IEEE 2021.

## 4.3 Propagation examples

It can be instructive to examine how each parameter of the EIWG scenario influences propagation along the path. In this section we present a collection of amplitude and phase curves for changes in $h'$, $\beta$, transmitter frequency, collision frequency profile, magnetic field direction, ground conductivity, the presence of multiple ionospheric species, and density perturbations vertically and along the propagation path. True ionospheres include combinations of these effects and vary along the path, but the following plots indicate the sensitivity of the measurements to each parameter and illustrate the nonlinear nature of some of these effects.

Each example modifies the same underlying homogeneous exponential daytime scenario with the parameters:

- 24 kHz transmitter
- vertical 50 μT magnetic field
- ground with $\sigma = 0.001\,\mathrm{S\,m^{-1}}$, $\epsilon_r = 15$
- $h' = 75\,\mathrm{km}$
- $\beta = 0.35\,\mathrm{km^{-1}}$

**Change in $h'$**

Using Wait's exponential ionosphere profiles, $h'$ correlates closely with the effective height of the waveguide (the dominant reflection height of the wave). Figure 4.6 shows that as $h'$ increases, the field strength pattern moves outward from the transmitter. A change in $h'$ of just $\sim 2\,\text{km}$ changes the phase by 45° across much of the plotted range. The typical phase measurement noise is 1°.

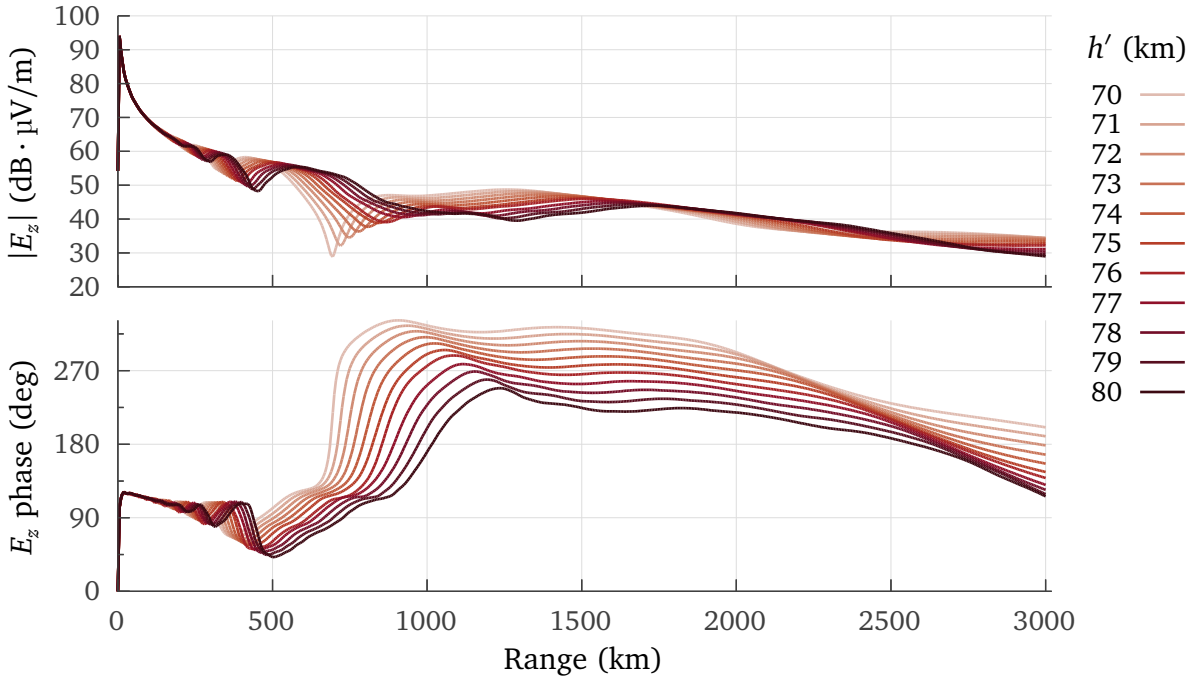


Figure 4.6: Amplitude and phase along the ground using the standard parameters ($\beta = 0.35\,\text{km}^{-1}$) and adjusting $h'$ in steps from 70 km to 80 km.

**Change in $\beta$**

Changes in $\beta$, shown in Fig. 4.7, are more difficult to describe. This plot covers the entire typical day to night range of $\beta$ values, with day from 0.2 to 0.4 km$^{-1}$ and night from 0.5 to 0.7 km$^{-1}$ (McRae & Thomson, 2000; Thomson et al., 2007). In general we see that as $\beta$ increases, the influence of $\beta$ on amplitude and phase decreases. We also see that the amplitude is higher on average for high $\beta$ ionospheres, which corresponds with a "sharper" waveguide boundary. There is very little sensitivity of subionospheric VLF to $\beta$ above 1.0 km$^{-1}$. Higher $\beta$ also tends to result in deeper nulls.

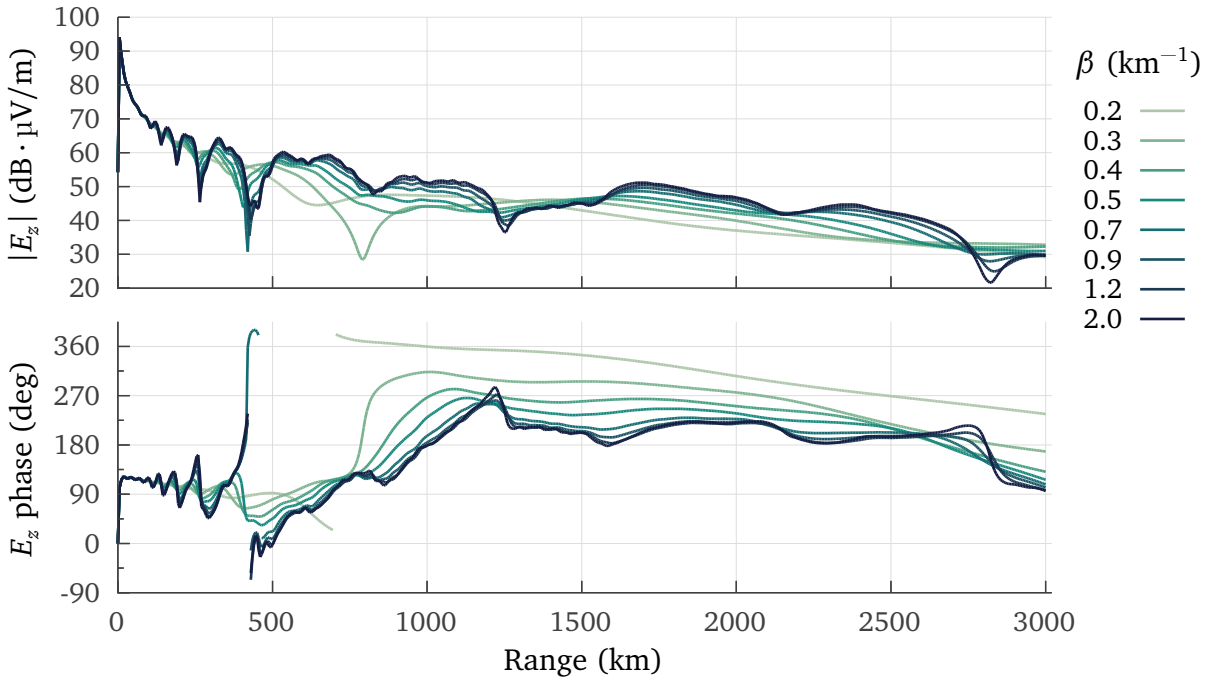


Figure 4.7: Amplitude and phase along the ground using the standard parameters ($h' = 75$ km) and adjusting $\beta$ in unequal steps from 0.2 km$^{-1}$ to 2 km$^{-1}$.

### Change in frequency

The effect of transmitter frequency, shown in Fig. 4.8, is very similar to the effect of $h'$ on amplitude and phase. This is logical given that the physics of waveguides is largely determined by the guide height relative to the wavelength. Increasing frequency (decreasing wavelength) is similar to increasing $h'$.

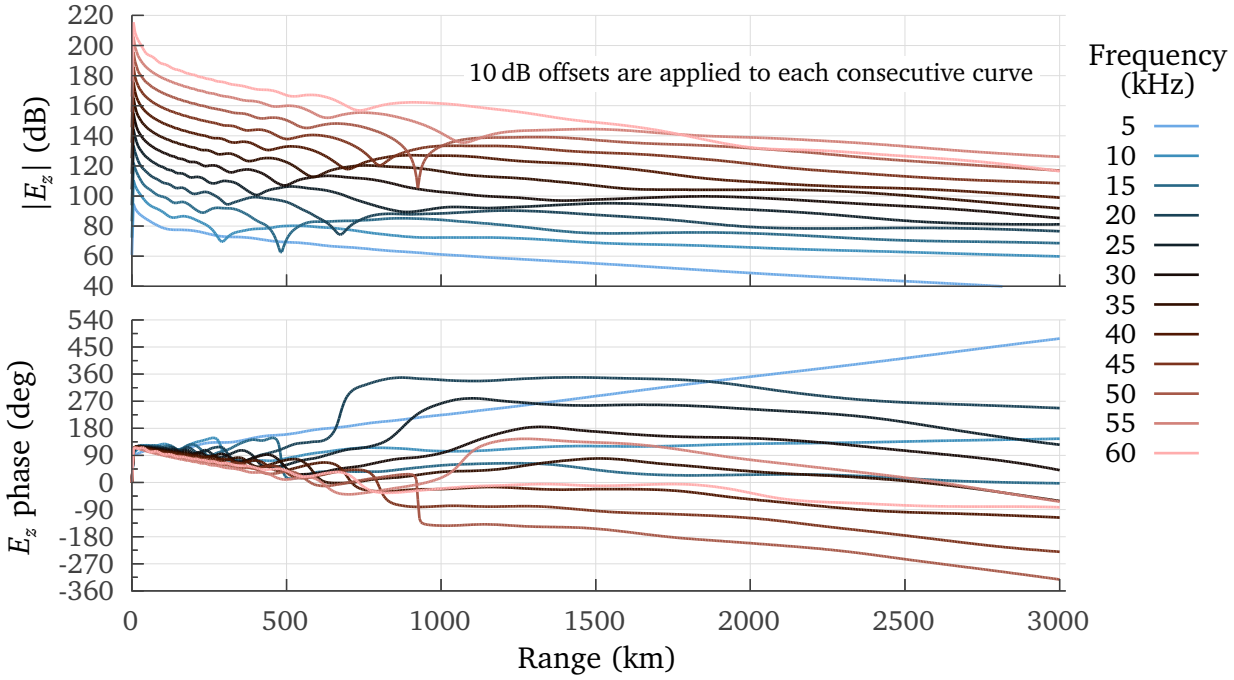


Figure 4.8: Amplitude and phase along the ground for transmitter frequencies between 5 kHz and 60 kHz. Each amplitude curve is shifted by 10 dB so the field strength pattern is easier to see.

**Change in collision frequency**

Wait's number density profiles were derived assuming a particular exponential collision frequency profile (Section 2.1.2.3). The amplitude and phase curves resulting from scaling the collision frequency up or down by a factor of 5 and changing the slope of the exponential profile by ±0.01 are shown in Fig. 4.9. The collision frequency profile is assumed to be fairly stable, but may vary in the presence of storms. Even small differences in the true profile will result in recovering different $h'$ and $\beta$ values when estimating the ionosphere using subionospheric VLF signals. Increasing the collision frequency has an effect similar to increasing $h'$. This is expected given the relationship between $h'$ and collision frequency through Wait's conductivity parameter.

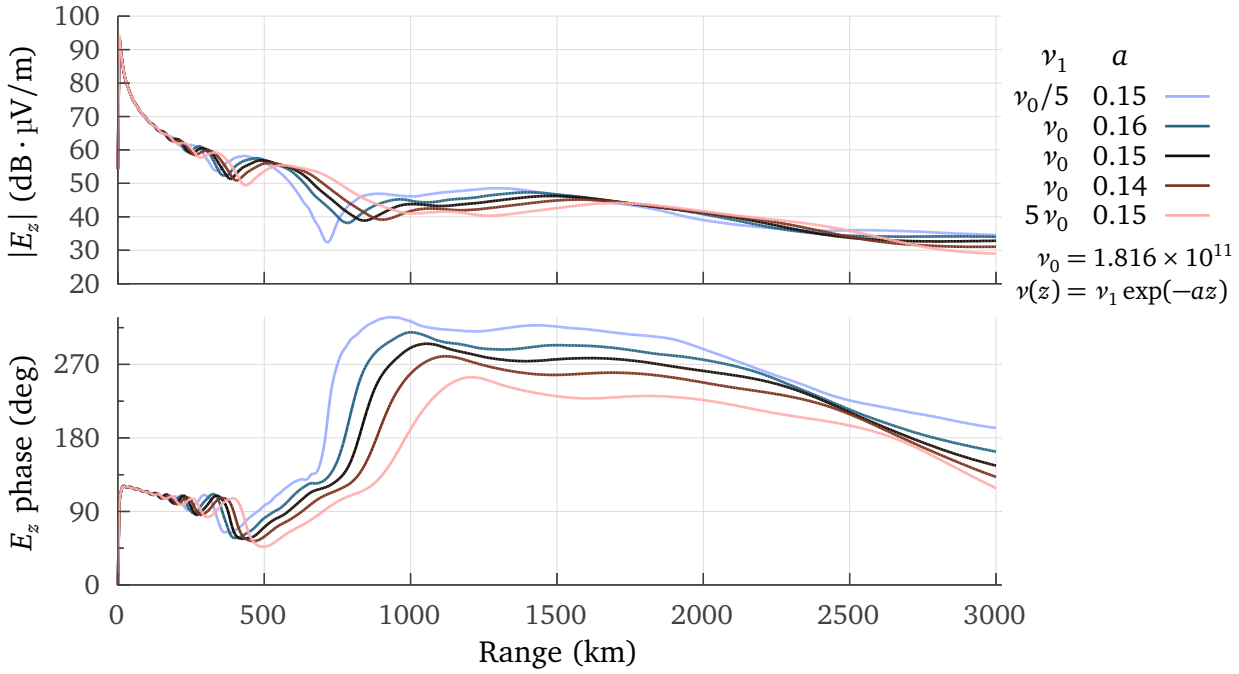


Figure 4.9: Amplitude and phase along the ground for several different exponential collision frequency profiles. The Wait collision frequency profile is $\nu_0 \exp(-0.15z)$.

**Change in magnetic field**

The influence of Earth's magnetic field on the propagation of VLF waves has been observed from the early days of longwave radio experiments (Crombie, 1958). The effect manifests as a difference in propagation between the east and west directions. At day there is very little difference, but at night when the reflection occurs at higher altitudes, the influence of the magnetic field is more pronounced. Both are shown in Fig. 4.10. There is no simple way to generalize the effect of the magnetic field on the amplitude and phase, but the difference between eastward and westward propagation is easily measured. This demonstrates why it is important to consider the magnetic field direction in propagation models.

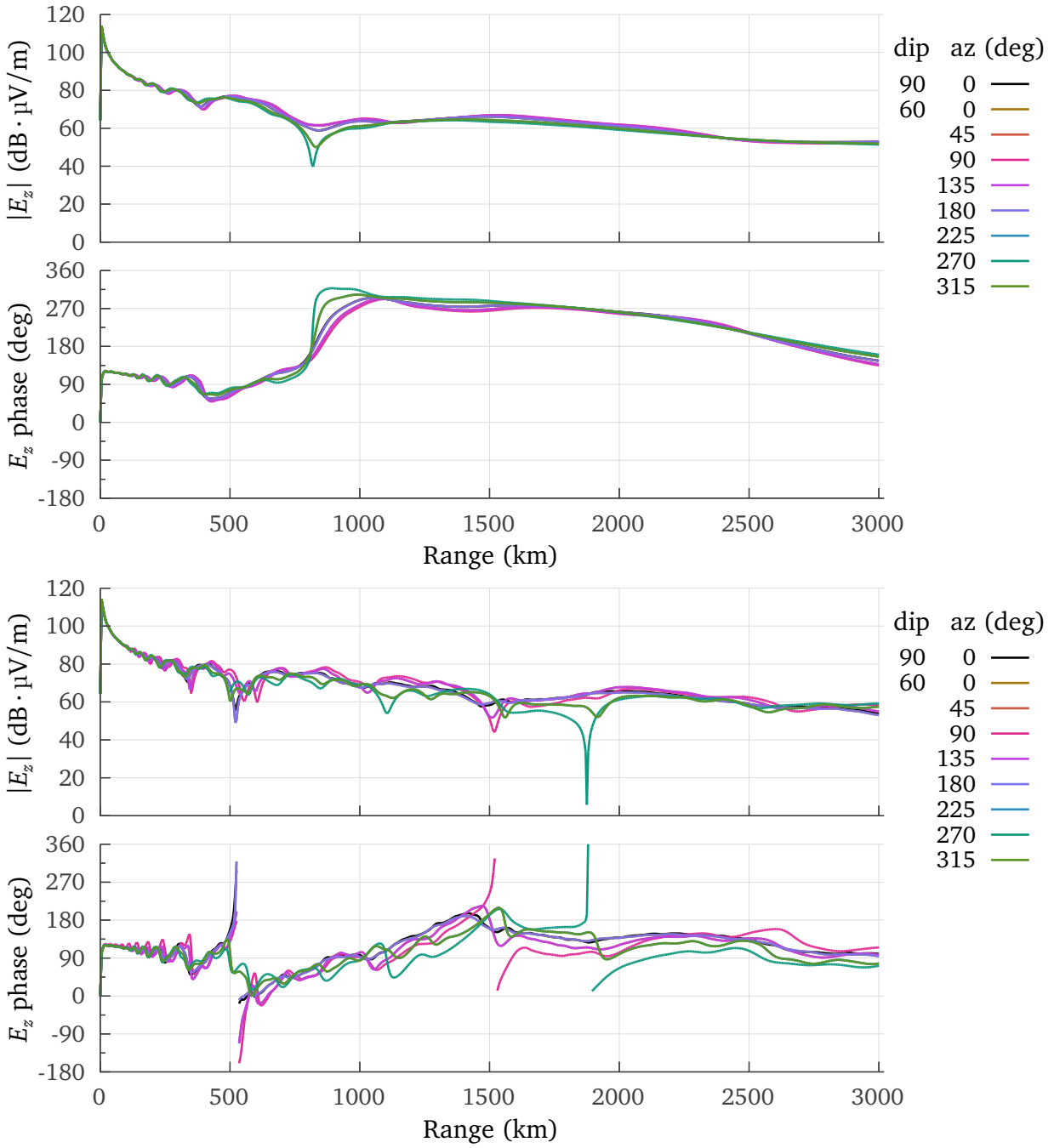


Figure 4.10: Top: With the daytime ionosphere $h' = 75\,\text{km}$, $\beta = 0.35\,\text{km}^{-1}$, magnetic field direction has a relatively small but measurable influence on the $E_z$ field along the ground. Bottom: Magnetic field has a stronger effect on nighttime propagation; here the ionosphere has $h' = 82\,\text{km}$, $\beta = 0.6\,\text{km}^{-1}$.

**Change in ground**

The electrical parameters of the ground have a significant affect on the attenuation of longwaves in the Earth-ionosphere waveguide. Figure 4.11 shows that the field strength over very low conductivity icy polar regions and dry soil are roughly 30 dB down from high conductivity sea water over much of the path. Although ground conductivity is assumed known from VLF ground conductivity maps for D-region estimation, the maps are several decades old and do not capture changes in urban areas or soil moisture (Fan & van den Dool, 2004). It is unknown to what extent this might affect VLF measurements. Some researchers attempt to use all-sea paths to reduce ionosphere estimation error from errors in the assumed ground conductivity (Thomson et al., 2007).

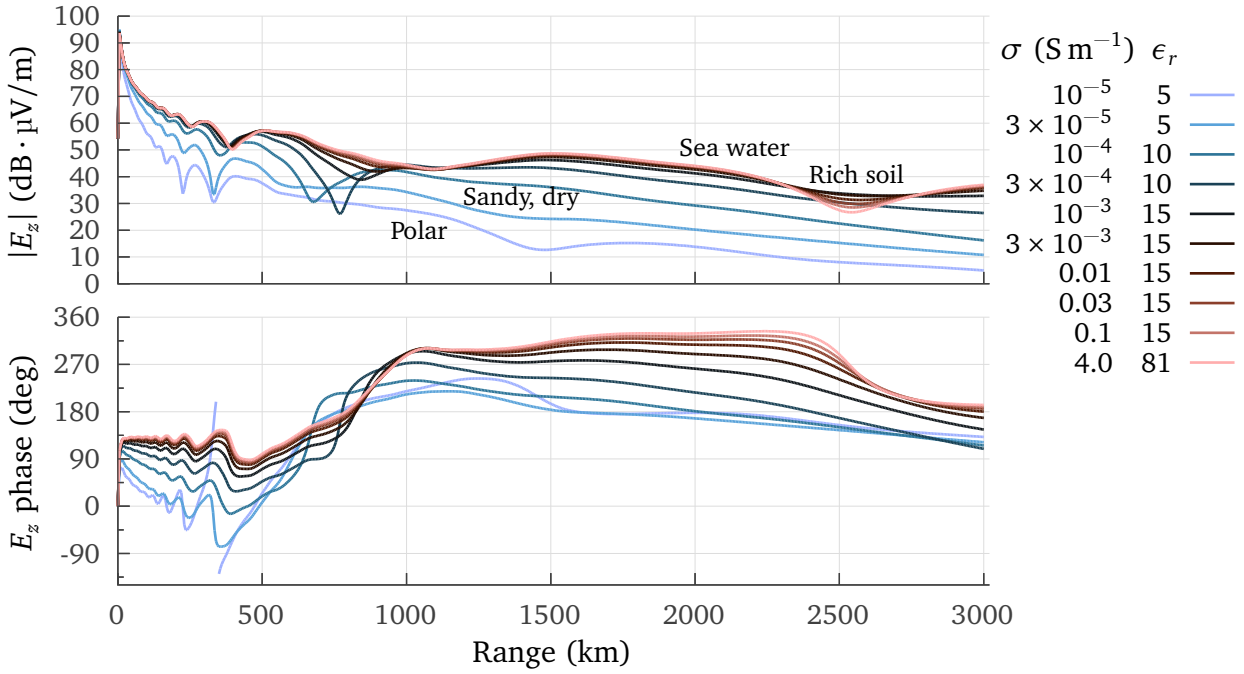


Figure 4.11: Amplitude and phase curves for the 10 standard grounds used in LWPC and the Morgan (1968) VLF ground conductivity maps.

**Multiple ionosphere species**

LongwaveModePropagator.jl allows any number of species to be included in the ionosphere. LWPC has limited support for a positive and negative ion to be included with electrons. Figure 4.13 shows amplitude and phase curves for an electrons-only ionosphere and an ionosphere with electrons, light positive and negative ions, and heavy positive and negative ions with masses of 58 000 $m_e$ and 120 000 $m_e$, respectively. The number density profiles for all of these species are shown in Fig. 4.12. These are profiles generated using MSIS and GPI. The presence of ions in the ionosphere has very little influence on the amplitude and phase. After a null near 750 km from the transmitter there is a constant offset of about 0.3 dB and 0.3°. The difference in amplitude is above the typical VLF receiver noise floor and could effectively bias the ionosphere estimate during an inversion process that uses an electrons-only ionosphere, but it would be difficult to measure the influence of the ions directly because uncertainty is dominated by the electrons.

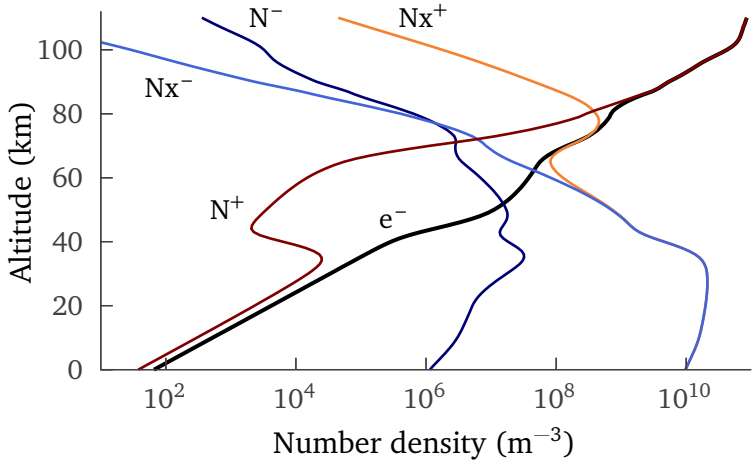


Figure 4.12: Density profiles of the five ionospheric species used to produce the amplitude and phase curves in Fig. 4.13. The curves are computed using the GPI model with neutral species from MSIS and an electron density profile from FIRI.

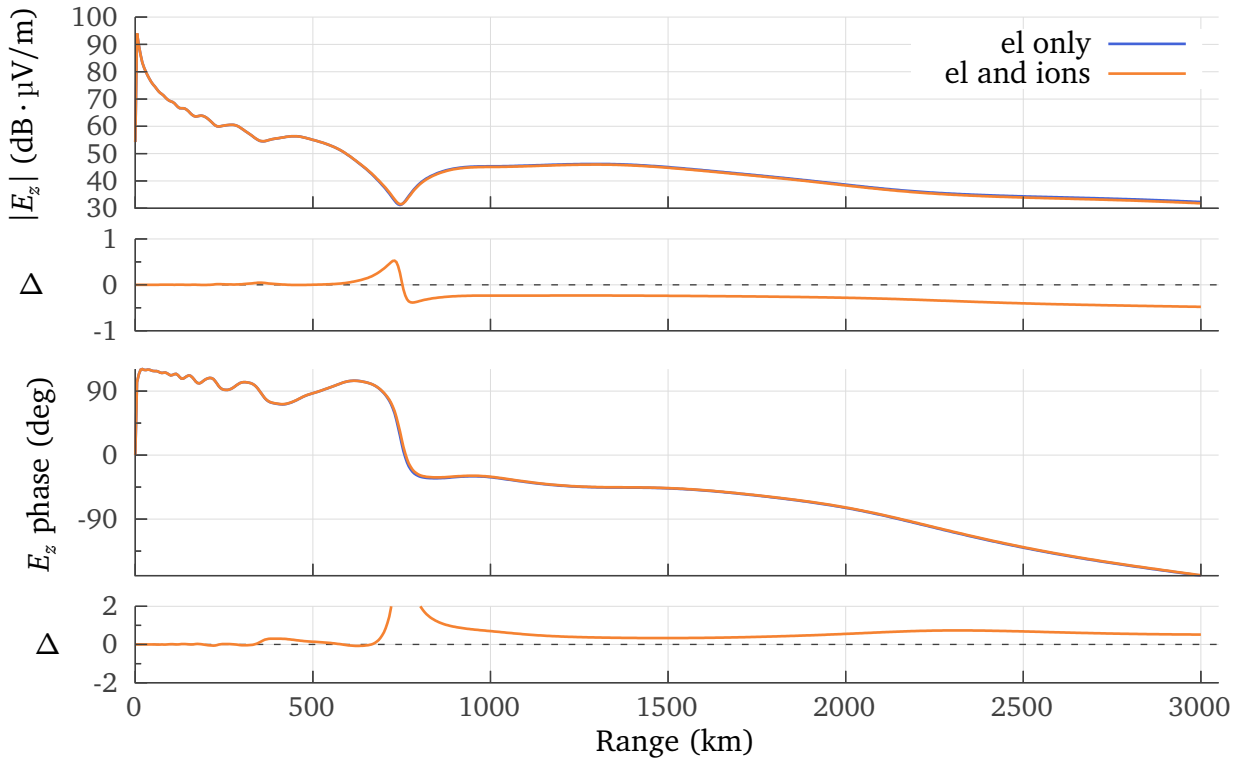


Figure 4.13: Amplitude and phase curves for an electrons-only and five-species ionosphere using the profiles in Fig. 4.12. There is a small, nearly constant bias in amplitude and phase over much of the path between the two ionospheres.

**Localized perturbation**

So far, all of the ionospheres examined have been homogeneous. In reality, most ionosphere perturbations occur in a localized region along the propagation path. Here, a Gaussian perturbation having a peak amplitude of 100 times the background electron density and a half width of 5 km is placed on a background Wait profile at different locations in the waveguide. Profiles are plotted in Fig. 4.14. The perturbation is located at 40, 50, 60, 70, or 80 km altitude and 0 through 2500 km along the propagation path in 500 km increments. The disturbed waveguide segment is 300 km long and the surrounding ionosphere is homogeneous daytime with $h' = 75$ km, $\beta = 0.35\,\text{km}^{-1}$ or homogeneous nighttime with $h' = 82$ km, $\beta = 0.6\,\text{km}^{-1}$.

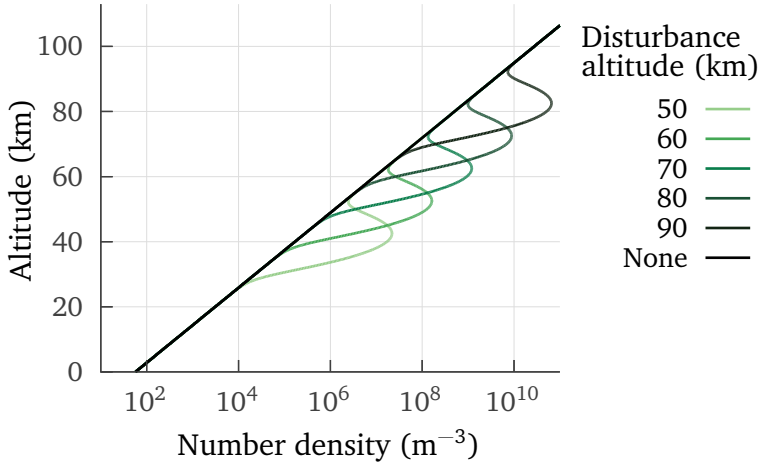


Figure 4.14: Gaussian disturbance profiles with a peak amplitude of 100 times the electron density at the center height and a half width of 5 km.

Figure 4.15 shows the amplitude curves for the different disturbance positions with background daytime and nighttime ionospheres. The dark blue colors are the curves associated with disturbances that occur for the first 300 km and are located sequentially at 40 through 80 km altitude and the reddish curves are for the disturbances that occur furthest from the transmitter. The amplitude response to a given disturbance is complicated—there are both short- and long-range affects. Although this may not be true in general, long-range effects for these perturbations are less for the nighttime ionosphere.

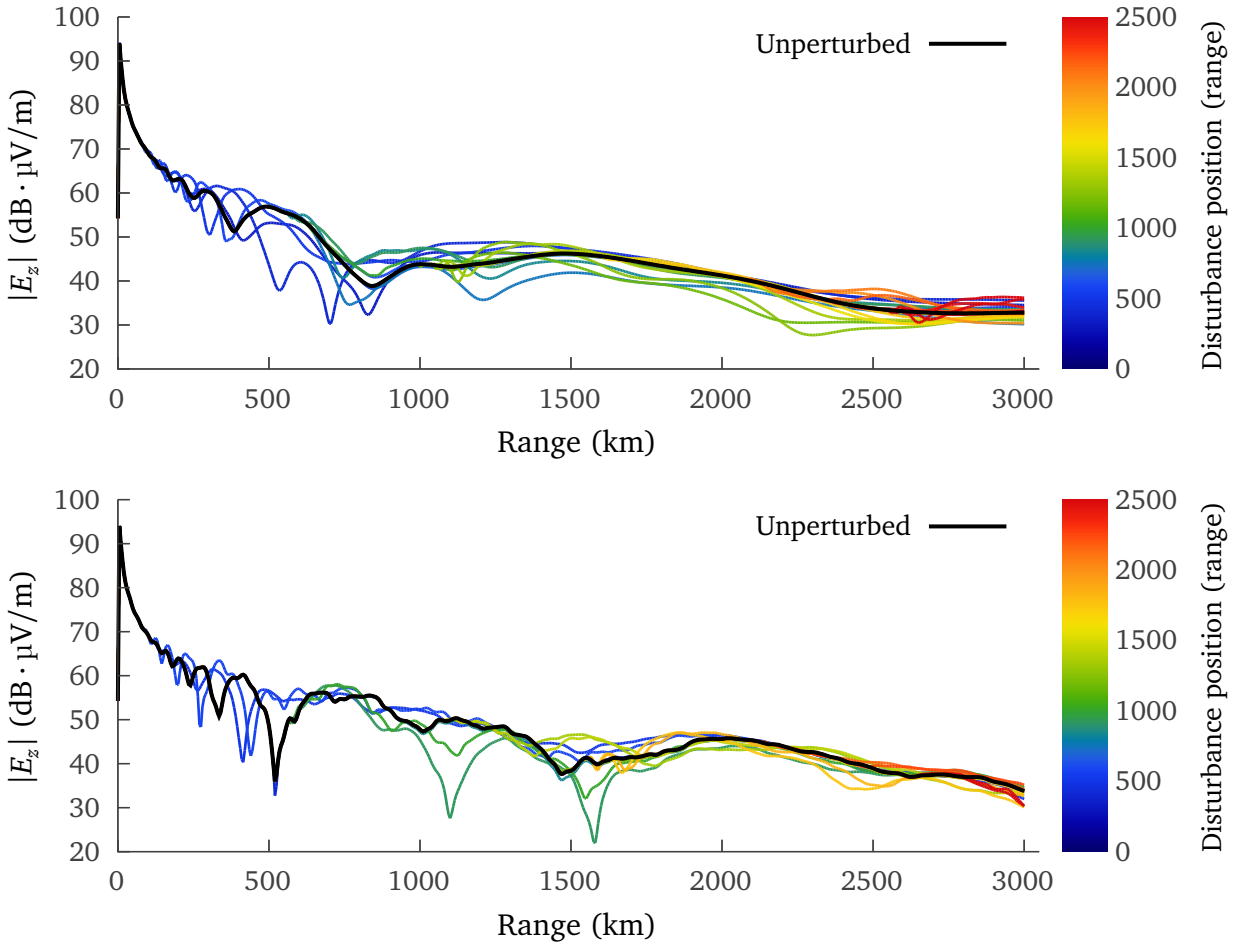


Figure 4.15: Amplitude curves along the ground for disturbed ionospheres with a Gaussian perturbation that is shifted up in altitude and out in distance from the transmitter for a background daytime ionosphere (top) and nighttime ionosphere (bottom).

Figure 4.16 shows cross sections of the amplitude difference between the disturbed and undisturbed ionospheres at 2000 km, 2500 km, and 3000 km from the transmitter. The variability suggests that for an observed change in amplitude at a single receiver, it is difficult to know where the perturbation causing the deviation is located along the propagation path without additional information. The electron density of the background nighttime ionosphere is sufficiently low that the perturbation only begins to affect the amplitude at 60 km altitude, but there are small differences for perturbations as low as 40 km in the daytime. Although not presented here, the magnitude of the perturbation also influences the measurement.

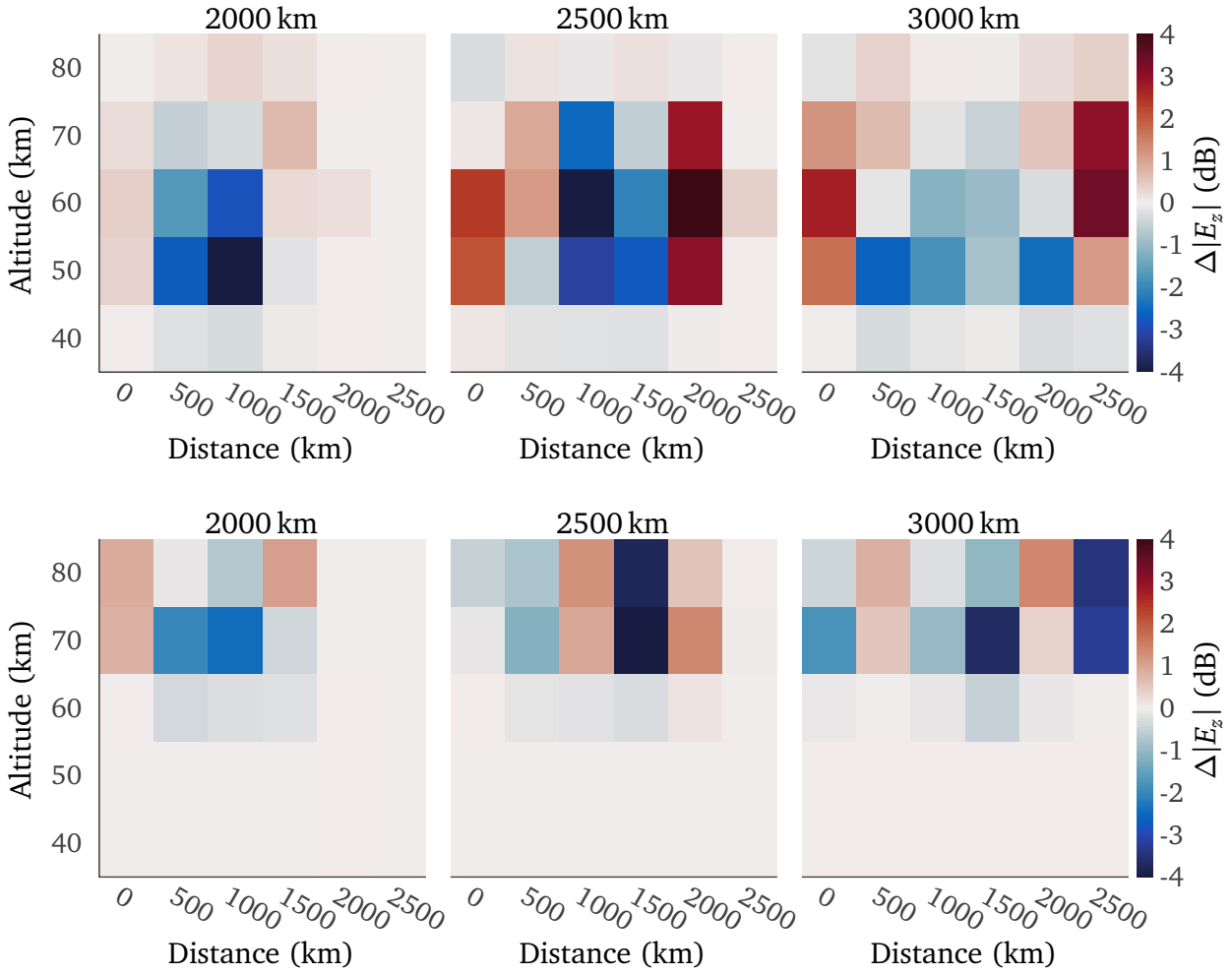


Figure 4.16: Cross-sections of the difference in amplitude between the disturbed and undisturbed daytime (top) and nighttime (bottom) ionospheres at 2000 km, 2500 km, and 3000 km from the transmitter. The heatmap axes specify the altitude of the perturbation and its distance from the transmitter.

# Chapter 5

# D-region imaging techniques

## 5.1 Problem description

Although constant path-average estimates of the D-region have been produced using subionospheric VLF for decades, the goal of this work is to leverage arrays of VLF receivers to produce a spatially varying estimate of electron density in the D-region. This estimate can be represented as a geographic map or "image" of electron density at a set altitude. Imaging the D-region is a nonlinear, underdetermined, and ill-posed inverse problem. It is underdetermined because the number of states $\boldsymbol{x}$ being estimated (electron density at each point in space) greatly exceeds the number of observations $\boldsymbol{y}$ (usually fewer than 100). In reality, the electron density varies smoothly across space, so it is necessary to discretize the state representation. The problem is ill-posed because there are potentially a near-infinite number of combinations of electron density along a path that produce the same observation at a receiver. Conventional D-region estimation techniques (Section 2.2.2) circumvent this problem by discretizing the electron density so there is only one density profile used for the entire path—this is the "path average" estimate.

Localization of the observations is also complicated. Theoretically, receivers are sensitive to conductivity at any height in the waveguide, but the sensitivity of the observation varies nonlinearly based on the relative positions of the transmitter, local ionosphere, receiver, and the surrounding waveguide conductivity. The sensitivity of an observation to electron density perturbations at different locations in the waveguide was demonstrated in Section 4.3.

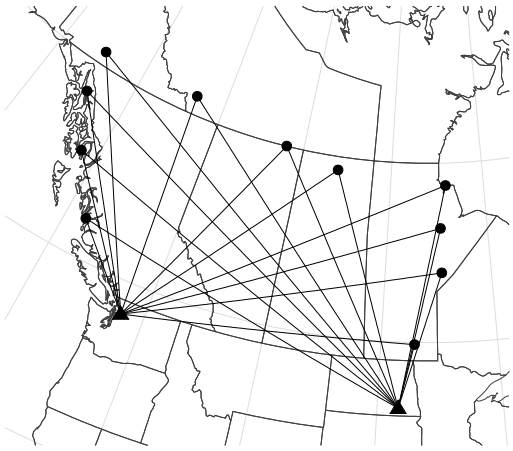

Figure 5.1: Propagation paths from the NML and NLK VLF transmitters (▲) in North Dakota and Washington to receivers (•) across Canada. The great circle propagation paths (black lines) appear as straight lines because the map is drawn using a Lambert conformal conic projection.

In the horizontal plane (in latitude/longitude) the receivers observe the states sparsely. Assuming stationary transmitters and receivers, as shown in Fig. 5.1, the observations are only sensitive to electron density along or immediately near the propagation path.[1] Therefore, the solution method must not only constrain the estimate along each path, but fill in the gaps between paths.

## 5.2 Solution scheme

The basic approach to solving this nonlinear estimation problems is to: 1) begin with a best-guess ionosphere, 2) apply a forward propagation model to the initial estimate, 3) compare the model and real observations, 4) adjust the ionosphere estimate, and 5) repeat steps 2)–4) until the difference between the forward modeled and real observations is within some threshold. In particular, once the residuals between the model and real ionosphere are comparable to the measurement noise, no further information can be extracted from the observations and the iterations can stop. Given the non-uniqueness of the problem, it is not necessarily true

[1] Density perturbations in the ionosphere can cause the radio wave to scatter so that a measurement at the receiver may be affected by the ionosphere off the great circle propagation path (Poulsen, 1991; Burns et al., 2021).

that the final estimate is the "correct" ionosphere or even the global minimum in some sense of the estimate fit. To paraphrase the geophysicist Albert Tarantola, "observations can only reject models, not suggest them" (Tarantola, 2006). The solution we choose should not only satisfy the observations, but have physical characteristics that we encode in some way based on our preexisting knowledge of the ionosphere. Practically, this means the estimated ionosphere should have a spatial correlation that reasonably represents the D-region, and we may want it to be similar to a prior estimate of the ionosphere as well. It is not hard to produce an estimate of the D-region with acceptable measurement residuals but completely implausible variations of the ionosphere. In this section we will define some common characteristics of our solution scheme and briefly discuss several estimation and optimization techniques before presenting details of two methods we applied to this problem.

### 5.2.1 State representation

To reduce the dimensionality of the problem we follow the usual practice of parameterizing the electron density in the vertical direction by assuming Wait's exponential profile (Eq. (2.9)). Therefore, the final ionosphere estimate can be represented as a pair of geographic maps: one for $h'$ and one for $\beta$. The state vector $\boldsymbol{x}$ is then just a stacked vector of $h'$ and $\beta$ values for each geographic location. As usual with Wait's profile, we inherently assume the collision frequency profile in Eq. (2.7) and additionally we assume an electrons-only ionosphere. Because of this parameterization in the vertical direction, although it's technically three-dimensional, we often describe our solution as a two-dimensional estimate, as opposed to the "one-dimensional" path average estimates that are produced when assuming a constant ionosphere profile along the propagation path.

Two different approaches have been used to reduce dimensionality in the horizontal plane. In Gasdia and Marshall (2019), $h'$ and $\beta$ were estimated on a latitude/longitude grid. A spherical spline interpolation (Dierckx, 1993) was used to interpolate from the grid onto latitude/longitude points every 120 km along each propagation path to determine the $h'$ and $\beta$

values for each waveguide segment in the propagation model. At higher latitudes this scheme resulted in northern regions of the grid having significantly denser estimates than the southern portion of the estimation region.

A new scheme performs the estimation on a grid defined on a projected plane. The North America Equidistant Conic (ESRI:102010) projection is used for the simulated observation experiments in this dissertation[2] and is compared to lat/lon coordinates in Fig. 5.2. As an equidistant projection, distances along meridians are proportionally correct and distances along the 20° N and 60° N latitudes are also correct (Snyder, 1987).[3] Latitude/longitude points every 100 km are extracted along each propagation path, projected into the plane, and then $h'$ and $\beta$ is interpolated from the grid control points onto the propagation path. This scheme means the ionosphere is estimated with equal spatial density over the estimation region, and because the interpolation occurs on a plane grid, a greater number of interpolation methods are available. The error from applying a spherical interpolation method to coordinates defined on the WGS84 ellipsoid is also eliminated.

### 5.2.2 Observations

The receiver observations and forward model outputs have been thoroughly discussed earlier in this dissertation. For the simulated observation experiments in Chapter 6, both vertical electric field $E_z$ amplitude ($\mathrm{dB}\cdot\mu\mathrm{V/m}$) and phase (deg) are measured at every receiver so $\boldsymbol{y}$ is a stacked vector of amplitudes and phases for each propagation path. In Gasdia and Marshall (2019) we used only phase measurements, and although we have found that using both amplitude and phase measurements outperforms either alone, it is not usually a necessity to use both. The typical amplitude measurement noise is $\sigma_A = 0.1\,\mathrm{dB}$ and phase noise is $\sigma_\phi = 1°$. VLF receivers are capable of sampling tens of times per second, much faster than the

[2] In PROJ notation (Evenden, 1990) this is: `+proj=eqdc +lat_0=0 +lon_0=0 +lat_1=20 +lat_2=60 +x_0=0 +y_0=0 +ellps=GRS80 +datum=NAD83 +units=m +no_defs`.

[3] Distances on the equidistant conic projection are proportionally correct with respect to the WGS84 ellipsoid, but are not equivalent to the great circle distances between points.

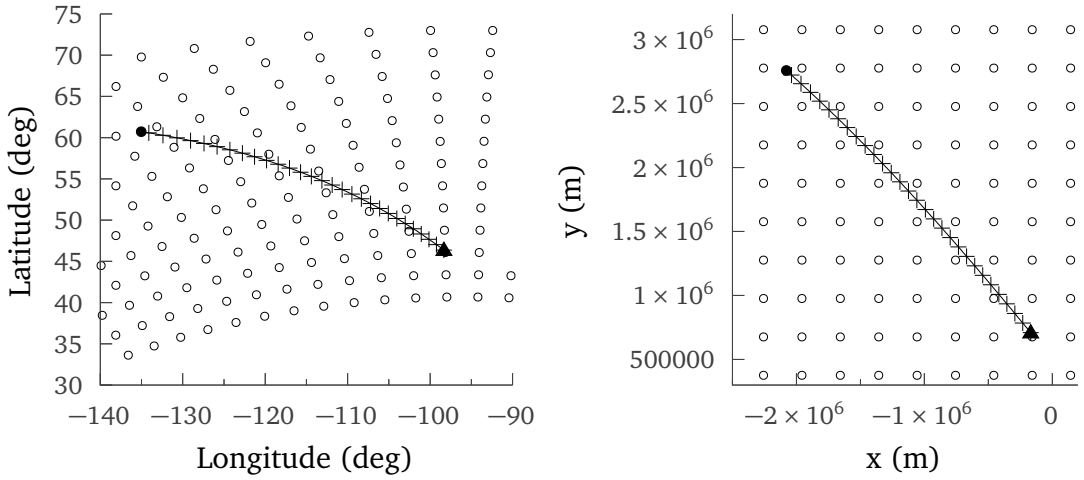


Figure 5.2: The great circle propagation path from the NML transmitter (▲) to a receiver (•) in Whitehorse, Yukon. The ionosphere is segmented every 100 km along the path (+) for the forward model. Estimate control points (∘) are also plotted. The plot on the left is in latitude/longitude and the plot on the right uses the ESRI:102010 projection.

time it takes to make the estimate, so the effective measurement noise can be reduced further by averaging samples over several seconds. However, the measurement noise is already very low compared to other uncertainties in the problem. Unlike most previous D-region estimates using subionospheric VLF, we assume that the observations are calibrated to the transmitter, i.e. we assume the radiated power is known and that phase measurements are coherent with the transmitter. This is a significant assumption because we have little insight into the operation of the U.S. Navy VLF communication transmitters. One way to achieve an instantaneous reference field for the transmitter is to place a receiver in its immediate vicinity, but further research into this technique is required. It would also be beneficial to study the ability of the estimation methods to cope with a bias between the real and modeled observations.

Several research groups operate networks of VLF receivers (Clilverd et al., 2009; Samanes et al., 2015; Pal & Hobara, 2016; Gross & Cohen, 2020), and it is typical that some propagation paths between transmitters and receivers will criss-cross (pass through the same geographic point). Example paths were shown in Fig. 5.1. This constrains the ionosphere estimate because the estimate at the crossing point must be consistent with both path measurements. However,

if we assume that the ionosphere is spatially correlated, then the ionosphere estimate *off* of any propagation path is statistically related to the estimate *along* a propagation path. This means that measurements from paths that are close together but not crossing can both contribute information to the estimate in the region near both paths. We expect the correlation between the ionosphere from one point to another to decrease with distance until the ionosphere at the second point can no longer be meaningfully predicted from the first. Intuition tells us that a receiver in Florida measuring the signal from the NAA transmitter in Maine should probably not be used to estimate the ionosphere over California. This is because the spatial correlation length of the ionosphere is shorter than the distance from the east coast to the west coast of the United States.

Each observation from the receiver array needs to be localized so that it is only used to update a subset of states across the map. Several localization schemes have been tested, all based on the great circle distance from the path to the estimate control points. This is complicated by the indirect relationship between the ionosphere and receiver: the measurement is nonlinearly sensitive to the ionosphere along the propagation path, it is not clear how off-path control points should be localized, and the two-dimensional forward models do not consider the off-path estimate. Other indirect observations like those produced from satellite-borne microwave radiometers use weighting functions to localize the observations (Fertig et al., 2007), but for our problem we do not have a fast way to establish what the weights should be for a given estimate. In Gasdia and Marshall (2019) we use a localization region around each propagation path that circles the transmitter, spreads outwards toward the midpoint of the propagation path, then converges inwards until circling the receiver. Any estimation points within this region are given a weight of 1 and any points outside of this region have a weight of 0. In Chapter 6, any estimate control point within ground range $r$ from a propagation path is given a weight of 1 and all others a weight of 0. Localization for one path is shown in Fig. 5.3. Additional points are removed from around the edges of the estimation region to reduce the number of points which must be estimated.

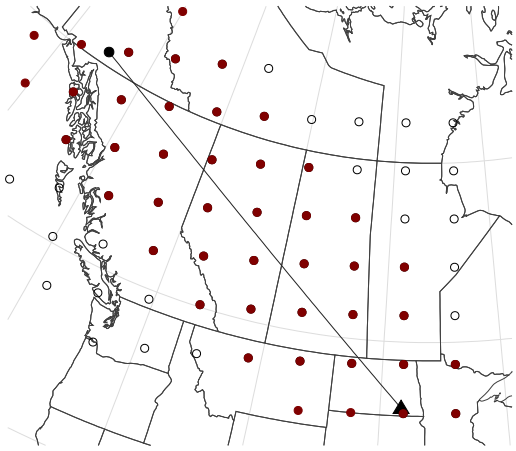

Figure 5.3: Localization of the estimate control points (•) for the propagation path from the NML transmitter to a receiver in the Yukon. The control points not localized to this path are labeled with ◦. Additional filtering of points by a latitude/longitude boundary box occurs in the simulated observation experiments.

The width of the localization for the path shown in Fig. 5.3 may appear very large; it includes control points up to 600 km from the propagation path. Choosing an appropriate localization region is a balance between a physically plausible region of measurement sensitivity and the estimate grid spacing. Too small a localization region around each path and no information is shared between the paths. Depending on the estimation method, there may even be holes in the map.

### 5.2.3 Methods for optimization or estimation

Numerous approaches have been developed to solve nonlinear estimation or optimization problems, so we will briefly discuss a few of them before detailing the methods we have applied. One way to characterize these methods is by whether or not they converge to a local or global solution. Local methods usually begin from an initial estimate and use a method-specific formula to generate a new estimate that reduces the error compared to the previous estimate. This is repeated until any change of the estimate increases the error. Local methods are susceptible to being trapped with an estimate that is locally good, but not necessarily the best estimate (or the

correct estimate). Global methods explicitly sample a wide range of states so that they do not get trapped with the solution nearest to the starting estimate. In the extreme, every possible combination of states could be tried until the combination with the lowest error is identified. Unfortunately, this is often prohibitively expensive when the forward model is slow or there are a large number of states. Instead, information about the probability distribution of the error surface can be utilized so time is not wasted sampling states that are not likely to contain the minimum error. We can build this information using techniques like Markov chain Monte Carlo (MCMC) or Hamiltonian Monte Carlo (HMC) (Mosegaard & Tarantola, 1995; Tromp, 2020). When less interested in characterizing the probability distribution than locating the best global fit, variants like simulated annealing (SA) can be applied (Ingber, 1989). For very expensive forward models, Bayesian optimization assumes an underlying distribution for the objective function, e.g. a Gaussian process, which is used to determine what estimate should be tried next in order to provide useful information about the global best estimate (Galuzzi et al., 2018). Many generic global optimization methods do not easily fit into our grid interpolation scheme because most interpolators become ill-conditioned or behave poorly when values vary wildly from grid point to grid point.

When a reasonable initial estimate is available, the most efficient methods are local methods that leverage the model gradient (and sometimes Hessian) to rapidly converge to the solution. This is a technique commonly applied to cost functions for imaging problems using optimization methods like steepest descent or L-BFGS (Modrak & Tromp, 2016). Computing the gradient of the objective function with a forward finite difference method requires $2n$ forward model runs for $n$ states. Not only can this be extremely expensive, but often the finite difference is a poor approximation to the true gradient and additional iterations of the optimization method are required to compensate. We have found for the D-region imaging problem that the error can be so bad that convergence is not possible. Optimization methods have been

developed to approximate the gradient with $n$ runs (Spall, 1998), but applied to our problem these struggled to produce valid $\beta$ estimates; again, this is likely because poor "gradients" were obtained.

Other geophysical estimation problems using the gradient for optimization rely heavily on the adjoint method. An adjoint model linearizes the relationship between perturbations in the observations and states (Plessix, 2006). For complicated nonlinear systems the adjoint model is usually a computer code (Naumann, 2019), but the implementation of mode theory is ill-suited to the development of an adjoint model. The primary difficulty is developing an adjoint for the complex-valued complex-argument root finder which is responsible for the majority of the model compute time. The root finder is also an issue for automatic differentiation tools (Bartholomew-Biggs et al., 2000). Although less efficient, it is necessary to treat the D-region imaging problem with derivative-free optimization methods. The best-known of these is Nelder-Mead (Nelder & Mead, 1965), which "crawls" a simplex of $n+1$ estimates towards a local solution. The method has been successfully applied to many problems despite the lack of rigorous theory supporting it (Powell, 2007).

## 5.3 Ensemble Kalman filter

The Kalman filter is an estimation method originally derived for linear dynamic systems with Gaussian noise processes (Kalman, 1960). If these assumptions are met, then the filter generates the minimum mean square error estimate from a sequence of noisy measurements. Because of the Gaussian process assumption, the state estimate is represented by a mean $\boldsymbol{x}$ and variance, or covariance matrix for multiple states, $\boldsymbol{P}$. Two steps are iterated as measurements become available. First, the state estimate and covariance are propagated using a forecast model from the time of the last estimate to the current time. Second, the measurement residual is used to update the state estimate and covariance. There is no forecast model for the D-region and we are interested in large spatial scales that change slowly relative to the measurement

rate, so we focus on the measurement update step. The update equations are

$$\boldsymbol{K} = \boldsymbol{P}_b \boldsymbol{H}^\top (\boldsymbol{H} \boldsymbol{P}_b \boldsymbol{H}^\top + \boldsymbol{R})^{-1} \tag{5.1}$$

$$\boldsymbol{x}_a = \boldsymbol{x}_b + \boldsymbol{K}(\boldsymbol{y}_o - \boldsymbol{H}\boldsymbol{x}_b) \tag{5.2}$$

$$\boldsymbol{P}_a = (\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H})\boldsymbol{P}_b \tag{5.3}$$

where subscript $b$ indicates "background" representing the previous (prior) values and $a$ indicates "analysis" representing the new values. The process is depicted at a high level in Fig. 5.4. These equations are written assuming a linearized forward model $\boldsymbol{H}$ such that the noisy observations can be modeled by $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \varepsilon$ with a Gaussian measurement noise realization $\varepsilon$ sampled from the covariance matrix $\boldsymbol{R}$. $\boldsymbol{y}_o$ are real observations; $\boldsymbol{y}$ without subscript $o$ indicates forward model "observations". Equation (5.2) clearly describes the basic procedure of updating the previous state estimate $\boldsymbol{x}_b$ based on how close the forward model observation is to the real observation. The Kalman gain $\boldsymbol{K}$ provides the optimal weighting between the previous estimate and the measurements based on the confidence in the previous estimate $\boldsymbol{P}_b$ and the measurement noise $\boldsymbol{R}$.

### 5.3.1 LETKF

Several variations on the Kalman filter have been developed so that it can be applied to problems with a nonlinear forward model $\mathcal{H}(\boldsymbol{x}_b)$. Although these are no longer optimal filters, they have been applied successfully to numerous problems (Budhiraja et al., 2007). One way to incorporate a nonlinear forward model is to represent the state mean and covariance by an ensemble population of states—the sample ensemble mean is the state mean and the sample covariance is the state covariance. The forward model can be applied to each member of the ensemble individually. There are several ensemble Kalman filter (EnKF) algorithms (Evensen, 2003), but we have implemented the measurement update step of the deterministic local ensemble transform Kalman filter (LETKF) because it has a simple implementation while

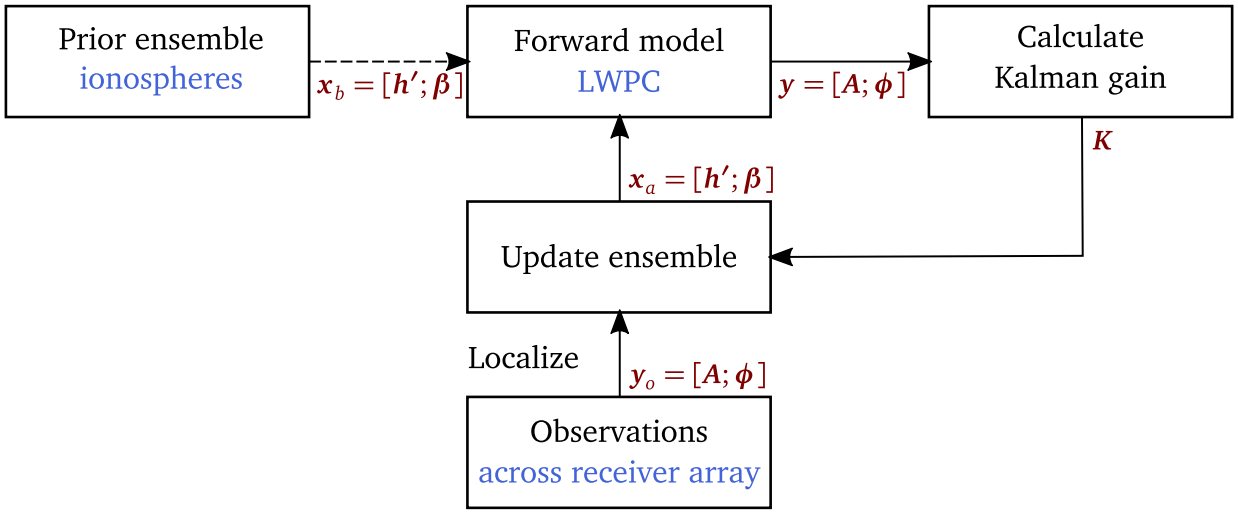


Figure 5.4: Schematic of the Kalman filter measurement update as applied to the D-region estimation problem. The Kalman gain is computed independent of the observations because a fixed measurement noise is assumed. The cycle repeats for a sequence of array observations over time.

supporting measurement localization and covariance inflation (Hunt et al., 2007).[4] In the LETKF, Eqs. (5.1) to (5.3) become

$$\widetilde{\boldsymbol{P}}_a = \left((k-1)\boldsymbol{I} + \boldsymbol{Y}_b^\top \boldsymbol{R}^{-1} \boldsymbol{Y}_b\right)^{-1} \tag{5.4}$$

$$\boldsymbol{x}_a = \boldsymbol{x}_b + \boldsymbol{X}_b \widetilde{\boldsymbol{P}}_a \boldsymbol{Y}_b^\top \boldsymbol{R}^{-1} \left(\boldsymbol{y}_o - \mathcal{H}(\boldsymbol{x}_b)\right) \tag{5.5}$$

$$\boldsymbol{P}_a = \boldsymbol{X}_b \widetilde{\boldsymbol{P}}_a \boldsymbol{X}_b^\top \tag{5.6}$$

where $\widetilde{\boldsymbol{P}}_a$ is the representation of $\boldsymbol{P}_a$ in the space spanned by the ensemble perturbations and capital $\boldsymbol{X}_b$ and $\boldsymbol{Y}_b$ are the zero-mean vectors $\boldsymbol{x}_b - \overline{\boldsymbol{x}_b}$ and $\boldsymbol{y}_b - \overline{\boldsymbol{y}_b}$ over the ensemble (Fertig et al., 2007). Gasdia and Marshall (2019) list the implementation steps for this problem. Where we say "LETKF" throughout the rest of this dissertation we are referring only to the measurement update step or a few iterations of the measurement update. The full LETKF algorithm includes a forecast step that is commonly used in other data assimilation applications.

[4] The LETKF was designed to be computationally efficient by performing most operations in the ensemble space, but the size of the D-region estimation problem is small compared to most applications of ensemble Kalman filters.

### 5.3.2 Prior state ensemble

The Kalman filter needs an initial state estimate and covariance to get started. This is called the prior distribution and contains all available information about the states before a measurement has been made. Solving our problem with an ensemble Kalman filter, we construct $k$ initial ionospheres that together have a mean that we believe to be close to the truth, a variance that represents our uncertainty in the prior, and that individually have a physically realistic spatial correlation. This is accomplished using a multivariate Gaussian distribution. Each of the scenarios in Chapter 6 use different prior means, and in general they could come from previously published "typical" $h'$ and $\beta$ values for day or night, a simple model like Ferguson Eqs. (2.10) and (2.11), or even an exponential fit to a more complicated chemical model. We then use equation (4.10) from Gaspari and Cohn (1999) to compute a covariance matrix between estimate control points with Gaussian-like spatial correlation:

$$C_0(z,1/2,c) = \begin{cases} -\frac{1}{4}(|z|/c)^5 + \frac{1}{2}(z/c)^4 + \frac{5}{8}(|z|/c)^3 - \frac{5}{3}(z/c)^2 + 1, & 0 \le |z| \le c \\ \frac{1}{12}(|z|/c)^5 - \frac{1}{2}(z/c)^4 + \frac{5}{8}(|z|/c)^3 + \frac{5}{3}(z/c)^2 - 5(|z|/c) + 4 - \frac{2}{3}c/|z|, & c \le |z| \le 2c \\ 0, & 2c \le |z| \end{cases} \tag{5.7}$$

We assume that the D-region has a spatial length scale of $\sigma = 1200\,\text{km}$, which is converted to length scale $c$ for Gaspari and Cohn equation (4.10). As the length scale decreases, the LETKF performance in the simulated observation experiments degrades. One result of applying two-dimensional estimate to the D-region would be improved knowledge of the real spatial correlation length.

The covariance matrix is multiplied by constants for the variance in $h'$ and $\beta$ to scale each to an appropriate uncertainty. In Gasdia and Marshall (2019) we found that performance plateaus after an ensemble size of about $k = 100$ ionospheres. Three ensemble members are shown in Fig. 5.5. Note that $h'$ and $\beta$ are independent (not correlated), although we have also experimented with them partially correlated. Correlating may improve the filter convergence

as long as the correlation realistically models the true ionosphere. This is not always easy to achieve because although typical ionospheres have a positive correlation between $h'$ and $\beta$, perturbations may have a negative correlation between $h'$ and $\beta$ (as $h'$ decreases, $\beta$ increases).

Because each prior ensemble ionosphere is smooth and the filter tends to retain relatively smooth estimates, we are able to use a stiff thin plate spline to interpolate directly between the estimate control points onto the propagation paths for the forward model waveguide segments.

### 5.3.3 Comments

The LETKF appears to be a relatively efficient estimator, requiring $k \times p \times t$ forward model runs for ensemble size $k$, $p$ propagation paths, and $t$ iterations. We iterate fewer than 10 times and the filter tends to converge in just a few iterations. The filter also produces an estimate of uncertainty computed from the variance of the ensemble. Unfortunately, because the D-region estimation problem is ill-conditioned, it is possible given a poor prior that the filter converges to the wrong ionosphere with high confidence. The requirement for the prior mean to be not more than about two standard deviations from the true ionosphere (and ideally closer) significantly limits the ability of the LETKF to estimate perturbed ionospheres when the perturbation is not known *a priori*. The EnKF is essentially a local (as opposed to global) estimation method. The standard deviation of the prior cannot be simply increased to accommodate high uncertainty because the problem is sufficiently nonlinear that the distribution of residuals will be highly non-Gaussian and the ensemble will likely diverge (Gasdia & Marshall, 2019). A second downside to the technique is the need to explicitly define the localization of each path measurement to the states: this is currently ad hoc, as described in Section 5.2.2.

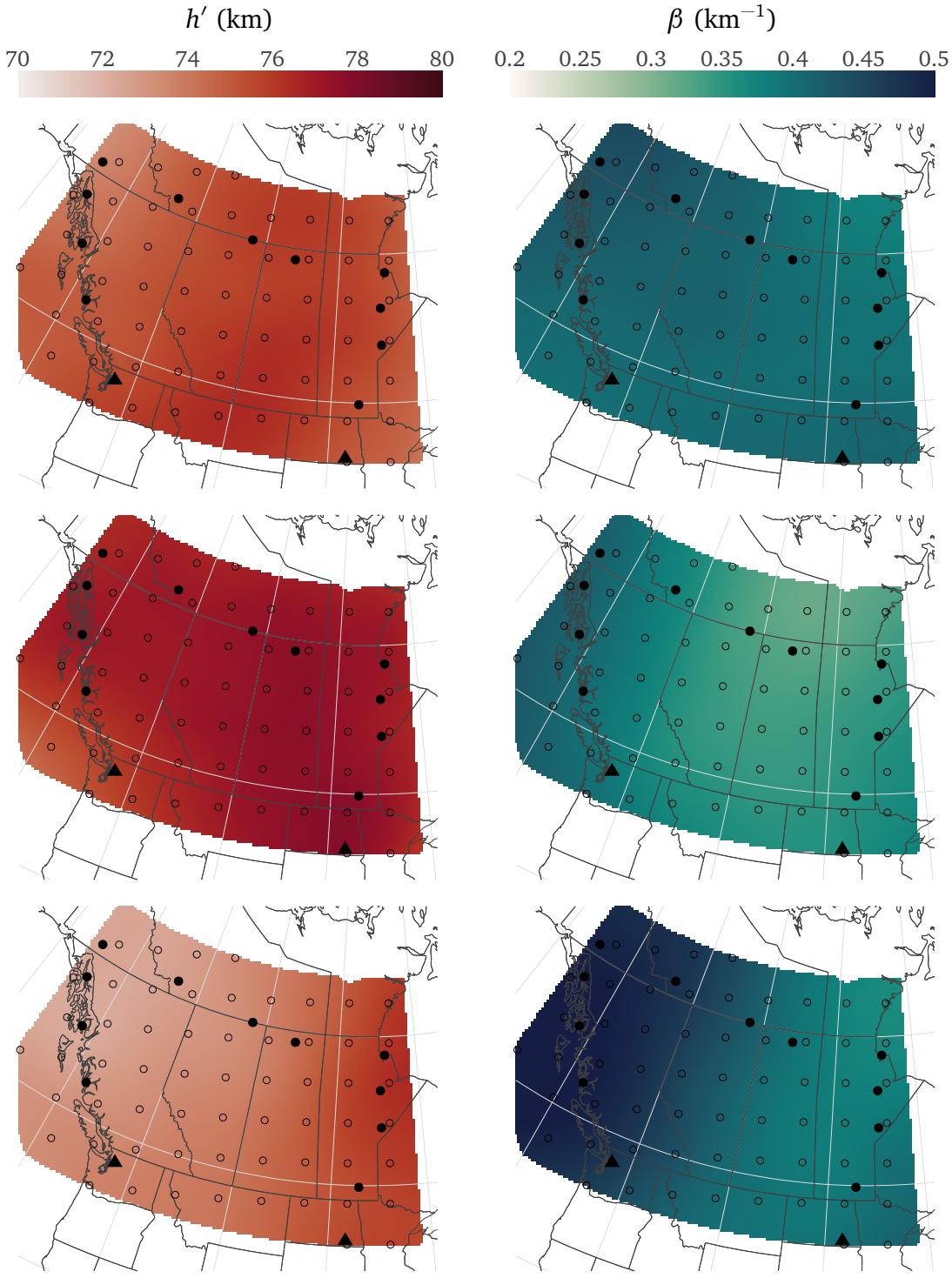


Figure 5.5: Three prior ionospheres selected from the ensemble for the daytime scenario of Chapter 6. The ensemble mean has $h' = 75\,\text{km}$ and $\beta = 0.4\,\text{km}^{-1}$ and standard deviation $\sigma_{h'} = 1.8\,\text{km}$ and $\sigma_\beta = 0.04\,\text{km}^{-1}$. $h'$ and $\beta$ are independent, although we could have chosen them to be correlated.

## 5.4 Nonlinear optimization

Approaching D-region imaging as a bounded optimization problem, we can define a cost function $J$ to be minimized such that

$$\boldsymbol{x}_{\text{est}} = \underset{x\in(x_{\min},x_{\max})}{\arg\min}\ J(\boldsymbol{x}) \tag{5.8}$$

where the subscript $a$ notation for the analysis has been dropped in favor of "est" for the estimate because it is a point estimate rather than statistical distribution. The "prior" for this method is similarly an initial vector $\boldsymbol{x}$ rather than a distribution. The cost function is

$$J(\boldsymbol{x}) = \Phi_d + \lambda\Phi_m \tag{5.9}$$

$$= \|\boldsymbol{y}_o - \mathcal{H}(\boldsymbol{x})\|_H + \lambda\|\nabla\boldsymbol{x}\|_{\text{TV}} \tag{5.10}$$

which has been partitioned into data cost $\Phi_d$ and model (estimate) cost $\Phi_m$. The model cost regularizes the solution—it penalizes estimates that do not have the properties we want in the ionosphere. Because the problem is ill-conditioned, minimizing $\Phi_d$ alone may result in an ionosphere estimate that happens to produce small measurement residuals, but is unphysical.

### 5.4.1 Data cost

We compute data cost using the pseudo-Huber norm $\|\cdot\|_H$ of the measurement residuals. The Huber norm is an $l_2$-norm near 0 and an $l_1$-norm outside of ± some value $\epsilon$. Although the $l_2$-norm has nice properties for optimization, large residuals produce very high cost and thus the $l_2$-norm is not robust to outliers. The $l_1$-norm is the sum of absolute values of the residuals, so it's robust, but it is not differentiable at zero. The Huber norm provides the quadratic properties of $l_2$ near the solution and the robustness of $l_1$ when far from it (Guitton & Symes, 2003). The pseudo-Huber loss

$$L_H(d) = \epsilon^2\left(\sqrt{1+(d/\epsilon)^2}-1\right) \tag{5.11}$$

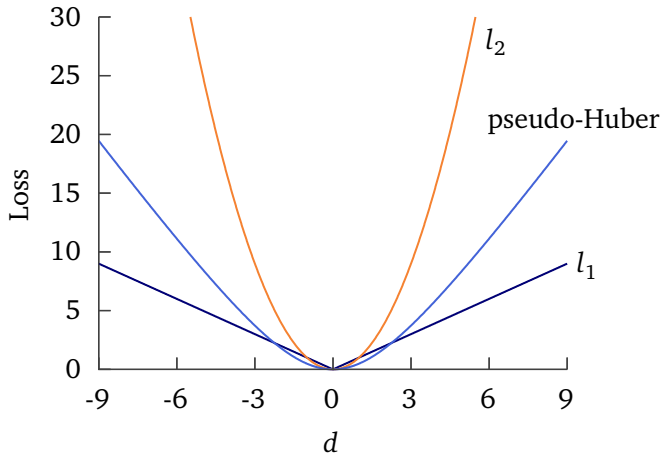


Figure 5.6: Loss functions of the $l_1$, $l_2$, and pseudo-Huber norms. The pseudo-Huber loss is approximately quadratic within ±3 and approximately linear outside of ±3.

is a smooth approximation of the Huber norm loss (Hartley & Zisserman, 2004) for value $d$. Figure 5.6 compares the loss functions for these three norms. The norm is the sum of this loss over each element of vector $\boldsymbol{d}$.

Before applying Eq. (5.11) to the measurement residuals, we divide them by the typical measurement noise figures:

$$\boldsymbol{d} = \frac{\boldsymbol{y}_o - \mathcal{H}(\boldsymbol{x})}{\sigma} \tag{5.12}$$

where $\sigma = 0.1\,\text{dB}$ for amplitude and $\sigma = 1°$ for phase. When both amplitude and phase are observed, separate $\boldsymbol{d}$s are computed for each and then stacked into a single vector. Dividing by the measurement noise effectively normalizes the measurements so that amplitude and phase are approximately equally weighted in the data cost. We set $\epsilon = 3$ so that outside of $\pm 3\sigma$ the residuals are penalized with an $l_1$-norm and inside they are penalized with an $l_2$-norm.

### 5.4.2 Model cost

The model cost is the total variation (TV) of the spatial gradient of the ionosphere estimate. Unlike Tikhonov regularization, which imposes a quadratic penalty and results in very smooth model estimates, TV imposes an $l_1$ penalty on the model gradient. This allows TV-regularized

estimates to retain sharp gradients that may exist in real perturbed ionospheres (Vogel & Oman, 1998). We compute the TV regularization separately for $h'$ and $\beta$. The entire scaled model cost is

$$\lambda_{h'} \sum_{i,j} \sqrt{\Delta u^2_{h':i,j} + \Delta v^2_{h':i,j} + \alpha^2_{h'}} + \lambda_{\beta} \sum_{i,j} \sqrt{\Delta u^2_{\beta:i,j} + \Delta v^2_{\beta:i,j} + \alpha^2_{\beta}} \tag{5.13}$$

where $\Delta u$ and $\Delta v$ represent the horizontal and vertical gradients, respectively, across the $h'$ and $\beta$ maps in the plane projection.[5] The gradients are discrete approximations using the Prewitt operator (Burger & Burge, 2016). $\alpha$ is a small term that ensures that the expression is differentiable at zero. As $\alpha$ gets larger, the estimated model becomes smoother (Anagaw & Sacchi, 2012). The square root expression is summed over every estimate control point in the grid.

$\lambda$ is sometimes called the regularization parameter. It scales the relative influence of the data cost and model cost. Unlike the Kalman gain, there is no optimal formula to compute what value it should have. For uncorrelated zero-mean Gaussian measurement noise of standard deviation $\sigma$, the squared $l_2$-norm data cost is a $\chi^2$ random variable with an expectation equal to the number of observations $N$. The discrepancy principle states that the regularization parameter should result in an estimate for which the final data misfit $\Phi_d \approx N$ (Constable et al., 1987). This is a formal way of saying that the estimate should proceed until the measurement residuals have decreased to the measurement noise. In practice the optimization is often performed for several values of $\lambda$ (Farquharson & Oldenburg, 2004). A plot of the final data cost against a Tikhonov regularization model cost for several $\lambda$ has the general appearance of an "L" and is referred to as an L-curve. It is often the case that the $\lambda$ value corresponding to the maximum curvature point of the L-curve represents a good solution (P. C. Hansen, 1992; P. C. Hansen & O'Leary, 1993). However, there is no guarantee with total variation (TV) regularization that the curve will be L-shaped with a corner containing the optimal regularization parameter (Mead, 2020). The "L-curve" for the terminator scenario of Chapter 6, shown on the left of Fig. 5.7, is

[5] Although it does not immediately appear similar to the $l_1$-norm, the absolute value of a gradient is the square root of the sum of squares of the individual gradients (in $u$ and $v$).

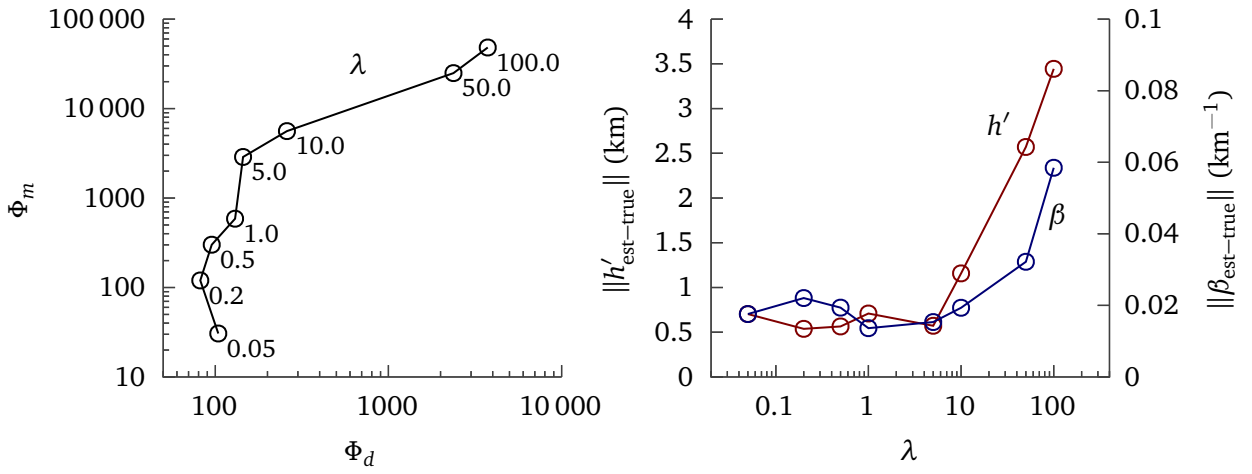


Figure 5.7: Left: "L-curve" plot of model $\Phi_m$ against data $\Phi_d$ cost for different regularization parameter values $\lambda$. Right: Mean absolute error of $h'$ and $\beta$ at the final estimate control points.

constructed using the final estimate and does not look like the typical L-shape. On the right of Fig. 5.7, the mean absolute deviation of the $h'$ and $\beta$ estimates from the true values at the control points is plotted as a function of $\lambda$. The error essentially plateaus for $\lambda \lesssim 1$, which indicates the estimate is limited by the data cost $\Phi_d$.

I found that $\lambda = 1$ worked well for our simulated observation experiments, although 5 or 10 would have also been reasonable. Several different $\lambda$ values should be tried when applying the method to real data. It's also not unreasonable to use different $\lambda$ values for different ionosphere conditions. To give appropriate weighting to the $\beta$ model cost, we scaled $\lambda_{h'} = \lambda$ to $\lambda_\beta = 10\lambda$ because $\beta$ has values roughly an order of magnitude less than $h'$.

### 5.4.3 COBYLA

Although a local and/or global method could be applied to this problem, we have had good results using a local method on unperturbed ionospheres. It is difficult to predict the performance of gradient-free nonlinear optimization methods on a particular problem without trying them. After applying very fast simulated annealing (VFSA, Ingber, 1989), covariance matrix adaptation evolution strategy (CMA-ES, N. Hansen and Ostermeier, 2001), differential

evolution (Storn and Price, 1997), locally-biased dividing rectangles (DIRECT-L, Gablonsky and Kelley, 2001), controlled random search (CRS, Kaelo and Ali, 2006), constrained optimization by linear approximations (COBYLA, Powell, 1994), bound optimization by quadratic approximation (BOBYQA, Powell, 2009), a variant of Subplex (Sbplx, Rowan, 1990), and others as my own implementations and across several libraries including Feldt (2018) and Johnson (2021), the most consistently effective method was COBYLA. Given the number of variables in designing the solution scheme and tuning each algorithm, it is possible that another algorithm might perform more efficiently under the right conditions, but COBYLA has been very robust for our problem.

The algorithm is similar to Nelder-Mead except it constructs linear polynomial approximations to the cost function by interpolation at the $n + 1$ vertices of the simplex. At each iteration, the worst simplex vertex is replaced by one that minimizes the linear polynomial within an automatically calculated trust region. The algorithm also attempts to keep the simplex in a regular shape. Because it incrementally "crawls" its way to the solution with step size limited by a trust region, the estimate tends to remain relatively smooth, which is easy on the grid interpolator. Although it supports constrained minimization, we had greater success applying our algorithm directly to the cost function, Eq. (5.9), without constraints except for bounds on $h'$ and $\beta$. Compared to the LETKF, the algorithm has a greater number of steps and includes several logic branches, but it has already been implemented in several software libraries including NLopt (Johnson, 2021), which was used for the results in Chapter 6. Unfortunately, like Nelder-Mead, Powell (1994) remarks that it is difficult to formulate meaningful theoretical bounds on the efficiency of COBYLA. Depending on the shape of the cost function surface it could converge more quickly because of the linear approximation of the cost function, or it could converge more slowly if it is difficult for the algorithm to maintain a regularly shaped simplex (SAS Institute Inc., 2021).

### 5.4.4 Multiscale interpolation

As the problem size (number of estimated states) increases, it becomes increasingly difficult for any method to find the cost function minimum. COBYLA uses $n+1$ vertices in a simplex and only changes a single vertex per iteration. Therefore, $n$ cost function evaluations ($n \times p$ paths total propagation model runs) must occur to change each state once. For a dense grid of estimate control points this would be very inefficient. Instead, we apply COBYLA several times from large to small spatial scales. First, we use a sparse estimate grid with $n \approx 10$ control points. COBYLA is able to make relatively large changes to the initial estimate in order to capture major trends in the ionosphere. We iterate COBYLA until the forward model has been run approximately 600 times ($600p$ propagation model runs). Then, we interpolate the resulting ionosphere onto a new, slightly denser grid, and run COBYLA for another 600 iterations. Our current scheme uses four spatial scales with grid spacings of 900 km, 600 km, 300 km, and 100 km. $n$ increases at each scale and COBYLA is not able to move the simplex as far within 600 iterations, but at each step the estimate does not need to move as far to approach the minimum.

The ionospheres defined on the coarse grids were often not suitable for rigid spline interpolation. Instead, we apply a locally weighted regression (LWR, Cleveland, 1979) to interpolate from the control points to the propagation paths and control points at the next spatial scale. LWR is also known as loess when applied to smoothing points on a scatter plot. To interpolate across our two-dimensional estimation grid, we apply a Gaussian weighting function

$$w = \exp\left(-\frac{r^2}{2\tau^2}\right) \tag{5.14}$$

for distance $r$ and "bandwidth" $\tau$. We use the implementation from Hoffimann (2018) which scales distances relative to the dimensions of the grid. We found that $\tau = l \times 10^{-7}$ for length scale $l$ was able to produce reasonably accurate interpolations of the truth. As $\tau$ increases, the interpolation becomes smoother and sharp transitions in the real ionosphere cannot be accurately represented by even the 100 km grid. Table 5.1 shows $l$ for each grid spacing used

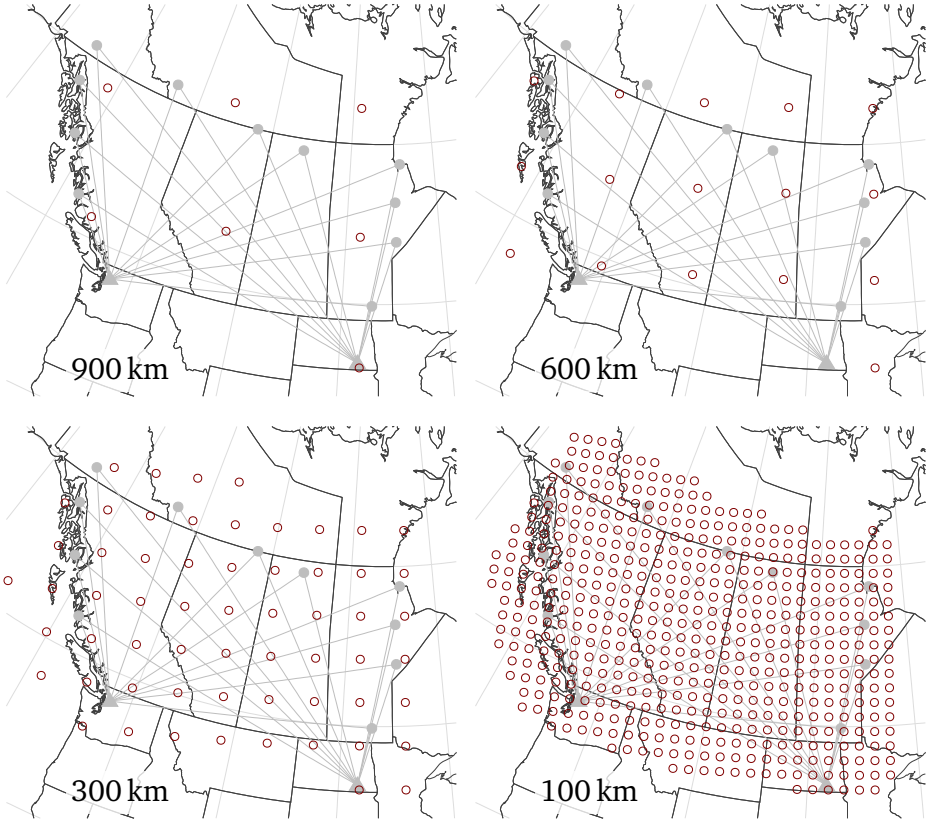


Figure 5.8: Estimate control points (∘) at the four spatial scales using projected spacings of 900, 600, 300, and 100 km. The estimate begins at 900 km and is then interpolated onto each finder scale after 600 iterations of COBYLA.

in the simulated observation experiments of Chapter 6. A statistically rigorous interpolator like kriging could encode the physical correlation length of the ionosphere into the interpolation, but most kriging implementations are not monotonic. The kriging interpolation has significant unphysical oscillations at transitions in the ionosphere as expected at the terminator and at perturbations.

### 5.4.5 Comments

This multiscale estimation technique is less efficient than the LETKF, requiring $2400p$ forward model runs compared to $1000p$ for ten iterations of the LETKF and $p$ propagation paths. However, it is robust to poor initial estimates and does not require explicit encoding

| Grid spacing (km) | Length-scale $l$ (km) |
|---|---|
| 900 | 1000 |
| 600 | 800 |
| 300 | 600 |
| 100 | 400 |

Table 5.1: Estimation grid spacing in projected kilometers and the length scale used for interpolation bandwidth and localization in WGS84 kilometers. The projected distances are proportional to the WGS84 distances, but they are not exactly the same. However, over much of the map they are within $\sim 1\,\%$.

of localization—it happens implicitly through the LWR interpolator. It may also be possible to increase the efficiency of this method by adjusting the 600 iterations to a smaller value, stopping early if some condition on the model or data cost is met, or using a different number of iterations for the different spatial scales. We found 600 to be a reasonable compromise between estimate accuracy and run time, but maximizing efficiency of the method was not an explicit goal.

Although this method does not provide a measure of estimate confidence, the robustness of the estimate can be tested by restarting the method with a different initial ionosphere. The method should converge to approximately the same ionosphere. The measurement residuals indicate if the estimate is valid, but the estimate itself is subjective because of the ill-conditioned nature of the problem.

# Chapter 6

## Simulated observation experiments

Simulated observation experiments use artificial observations to test the estimation methods against known "truth" ionospheres. To decrease dependence on the forward propagation model, the estimation methods use LWPC as the forward model and the simulated observations are generated using LMP with independent Gaussian noise added to each observation. Although Chapter 4 showed that the two models are in good agreement, in some scenarios the difference between the models can build to a measurable level for ionospheres with a large number of segments (every 100 km along the path). Both models use identical ground conductivity maps and magnetic field vectors from the International Geomagnetic Reference Field (IGRF)-13 (Thébault et al., 2015).

The truth ionospheres are defined as continuous functions of latitude/longitude, not confined to the estimate grid. This places a lower bound on the estimate accuracy that can be obtained due to error in the interpolation scheme from the grid points, i.e. even if the ionosphere is perfectly estimated at all of the control points, the interpolated surface differs somewhat from the continuous truth ionosphere.

Because the estimate maps are produced from interpolating or extrapolating from the non-rectangular grid of control points, it is necessary to define a region where we believe the estimate is reasonably confident. The LETKF variance at each control point can be interpolated across the map, but instead we explicitly use the 100 km path segments to construct a map of kriging variance. Kriging is an interpolation method commonly used in geophysics that uses

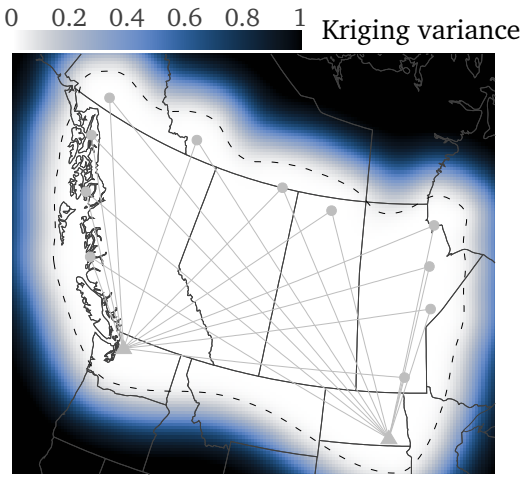


Figure 6.1: Kriging variance assuming "measurements" every 100 km along each propagation path with a Gaussian variogram having range 600 km and a sill of 1. The dashed line represents the $\sigma^2 = 0.2^2$ threshold used to mask the interpolation boundary in the following maps.

Gaussian processes to make the best linear unbiased prediction of a sampled field at a new point (Krige, 1951; Matheron, 1963). Like the Kalman filter, it also provides an estimate of the variance at that point. If we assume our "measurements" really occur at the 100 km intervals over which the interpolated $h'$ and $\beta$ surfaces are sampled for the forward model, then those regions of the map with dense paths (and ionosphere segments) have lower variance than those that are sparsely sampled. We set up an ordinary kriging (OK) problem with a variogram range of 600 km (equal to the localization range) and sill of 1. Whereas the previously mentioned scale lengths correspond to standard deviations, the variogram range indicates the maximum range at which the field at one point is correlated to another. The "sill" represents the variance of the field beyond the variogram range. We mask the regions of the estimate maps where the kriging variance is greater than $\sigma^2 = 0.2^2$, shown in Fig. 6.1, although we could choose a different value depending on what we consider an acceptable level of uncertainty.

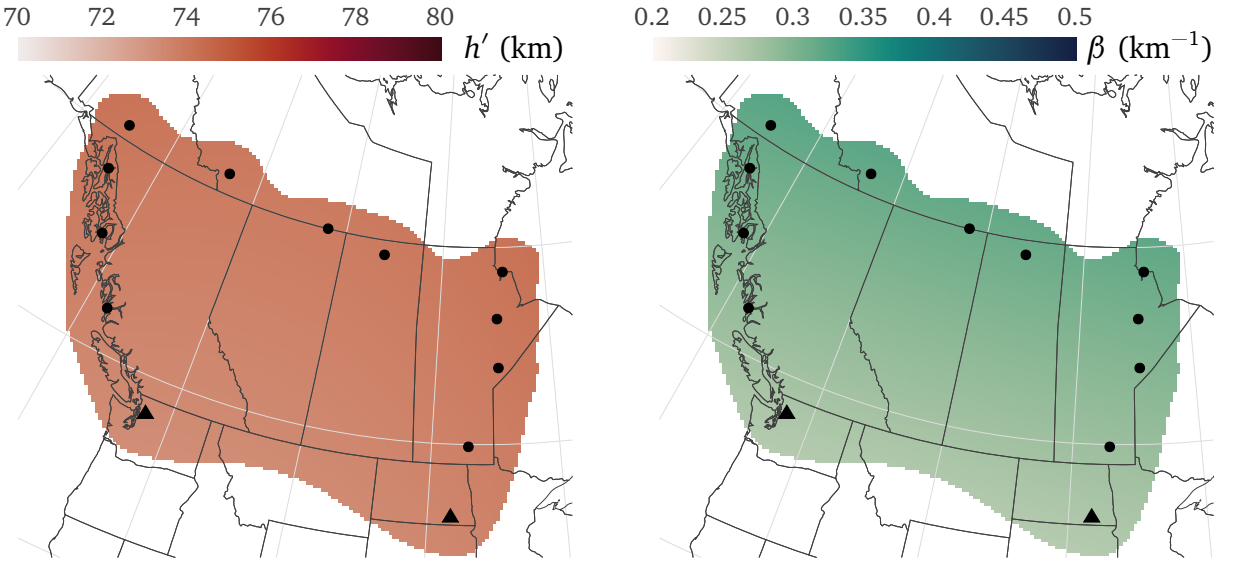


Figure 6.2: Daytime truth ionosphere defined using the Ferguson model at 2020-03-01 2000 UTC. Transmitters at the south of the map are marked with ▲ and receivers are marked with •. Although experiments in real life may require receivers placed near each of the transmitters, no such receivers were used in these simulated observation experiments.

## 6.1 Daytime scenario

### 6.1.1 Truth model

This section shows the results of a simulated observation experiment for a "typical" daytime ionosphere defined using the Ferguson model (Section 2.1.2.3) at 2020-03-01 2000 UTC. The true $h'$ and $\beta$ maps (without interpolation) are shown in Fig. 6.2. The ionosphere is very slowly varying—it's nearly homogeneous.

### 6.1.2 Prior model

With real data, Ferguson could be used as the prior mean (for LETKF) or initial estimate (for the cost function minimization), but here we use a constant ionosphere with $h' = 75\,\text{km}$ and $\beta = 0.4\,\text{km}^{-1}$. The standard deviation of the LETKF prior is $\sigma_{h'} = 1.8\,\text{km}$ and $\sigma_\beta = 0.04\,\text{km}^{-1}$. As previously mentioned, the LETKF uses a Gaussian-like spatial covariance with a length scale

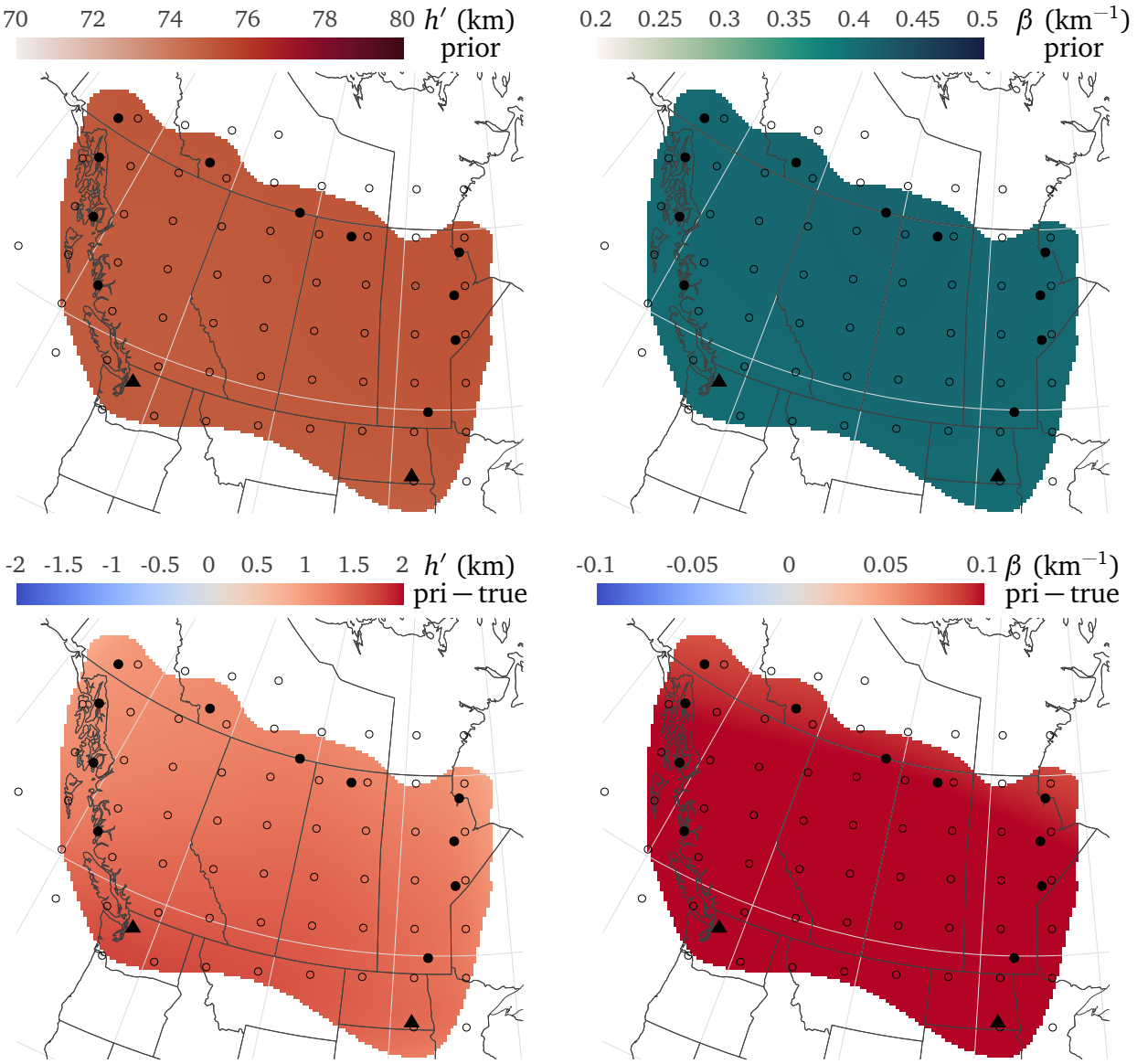


Figure 6.3: Top: Initial ionosphere for the day scenario (mean of the ensemble prior for the LETKF). Bottom: Difference between the prior mean and true ionosphere.

of 1200 km to generate the prior (Fig. 5.5).

For easier comparison to the estimates, the prior $h'$ and $\beta$ (ensemble mean for LETKF) and error between the prior and true ionosphere is shown in Fig. 6.3. Both $h'$ and $\beta$ are biased high compared to the truth.

### 6.1.3 Estimates

To avoid having pages and pages of plots, I will often show sequences of plots for $h'$ only. $\beta$ usually follows a similar trend.

Figure 6.4 shows the $h'$ error over six iterations of the LETKF. Note that the result of the first iteration is a large change from the prior error. We often observe that the first iteration makes the largest change in the states and removes much of the bias from the prior ionosphere model. The next five iterations have relatively little change.

Figure 6.5 shows the $h'$ error at each step of the multiscale optimization of the cost function as minimized by COBYLA. Because the daytime scenario is smooth and very slowly varying, the lowest order spatial scale already captures much of the true ionosphere. At smaller spatial scales there is some local error, but the overall region within ±0.5 km of the truth increases, as shown by the dotted contour lines. One downside to the smaller spatial scales is that with real data it could be difficult to determine if local changes in $h'$ are real or artifacts of the estimate. It may be necessary to continue running the final estimate until some convergence criteria on change in the states is reached rather than always stopping after 600 iterations. The particular configuration in iteration 4 of Fig. 6.5 may also be a local minimum of the cost function and we are effectively overfitting the data at this scale. Nonetheless, note that in iteration 3 the control points are identical to those used by the LETKF and this method has resulted in a better estimate.

Figure 6.6 shows the actual estimated $h'$ and $\beta$ ionosphere parameters using the LETKF. Below them are the ensemble standard deviations at each control point interpolated across the map. This indication of the uncertainty in the estimate has a reasonable match to areas of locally high error—where there is high error, the LETKF has also indicated higher uncertainty.

Figure 6.7 is the final $h'$ and $\beta$ estimate from the multiscale estimation. Overall $h'$ closer to the truth than the LETKF, but the regions of locally high $h'$ are evident near the center of the map. $\beta$ is slightly worse. The mean absolute deviation from the true ionosphere is 0.41 km in

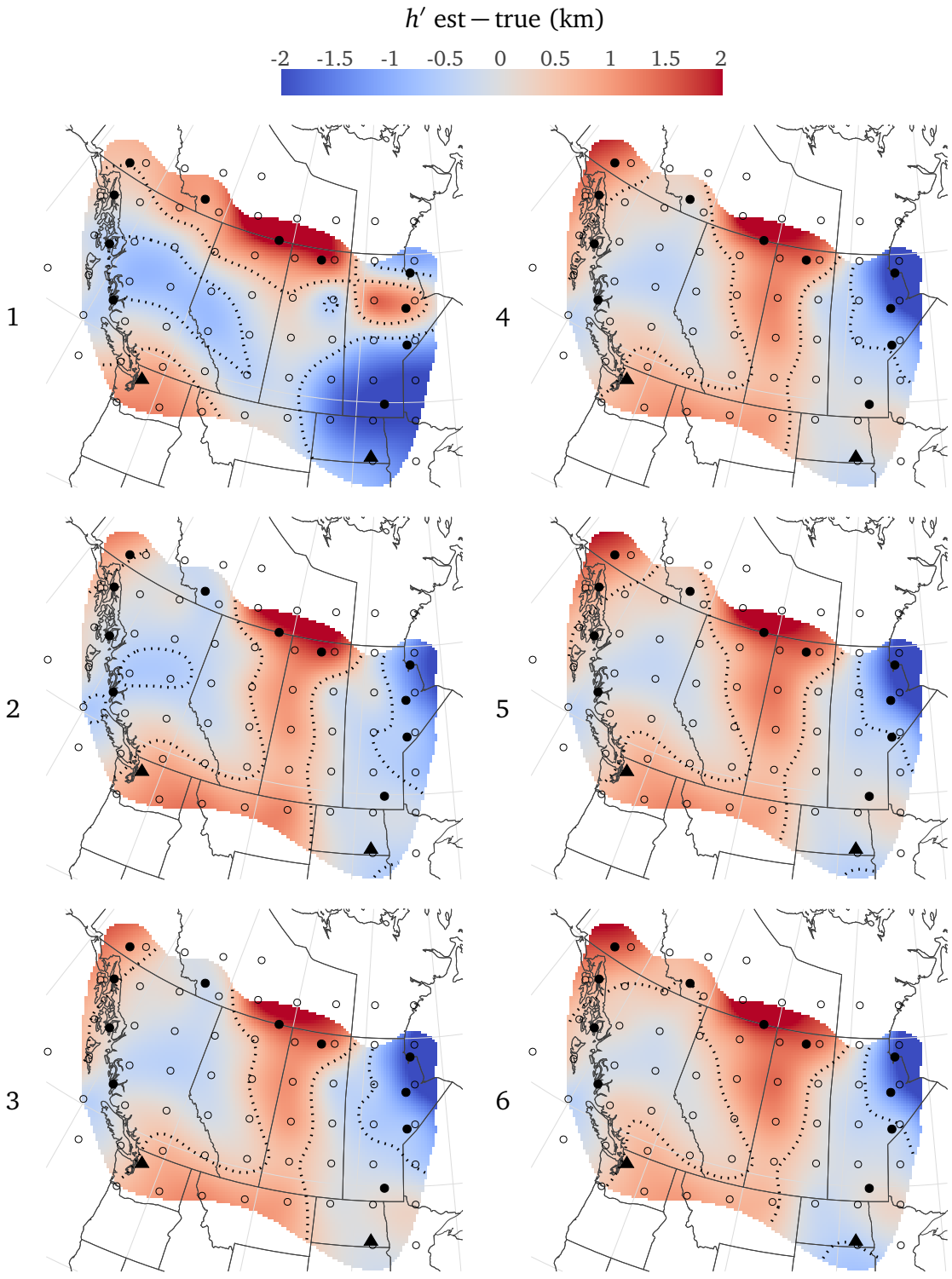


Figure 6.4: $h'$ error at each of six iterations of the LETKF for the daytime scenario. The fine dotted line is a ±0.5 km error contour.

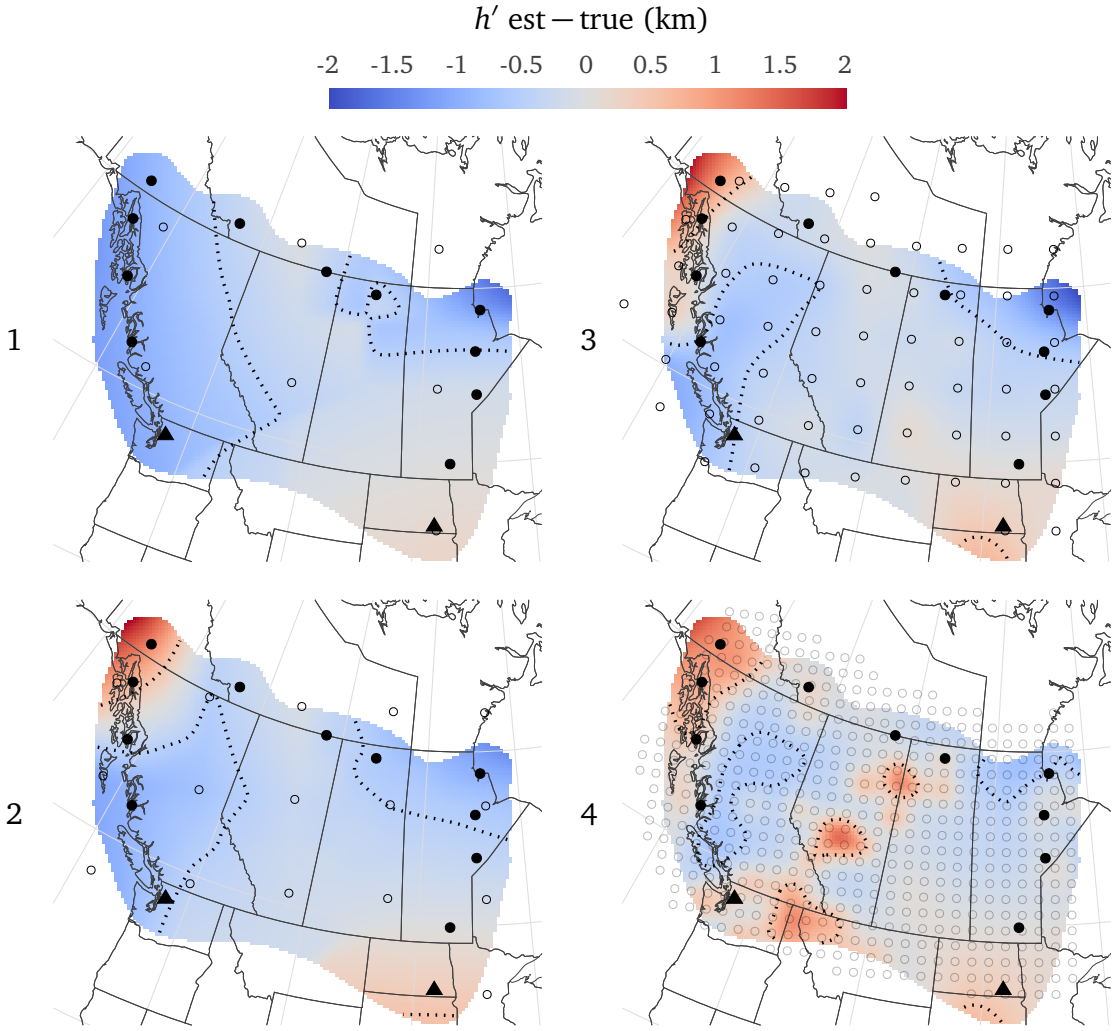


Figure 6.5: Difference between the estimated and true ionosphere at each step of the multiscale optimization of the cost function for the daytime scenario. In iteration 4 the control points (∘) are lightly colored so it is easier to see the error map.

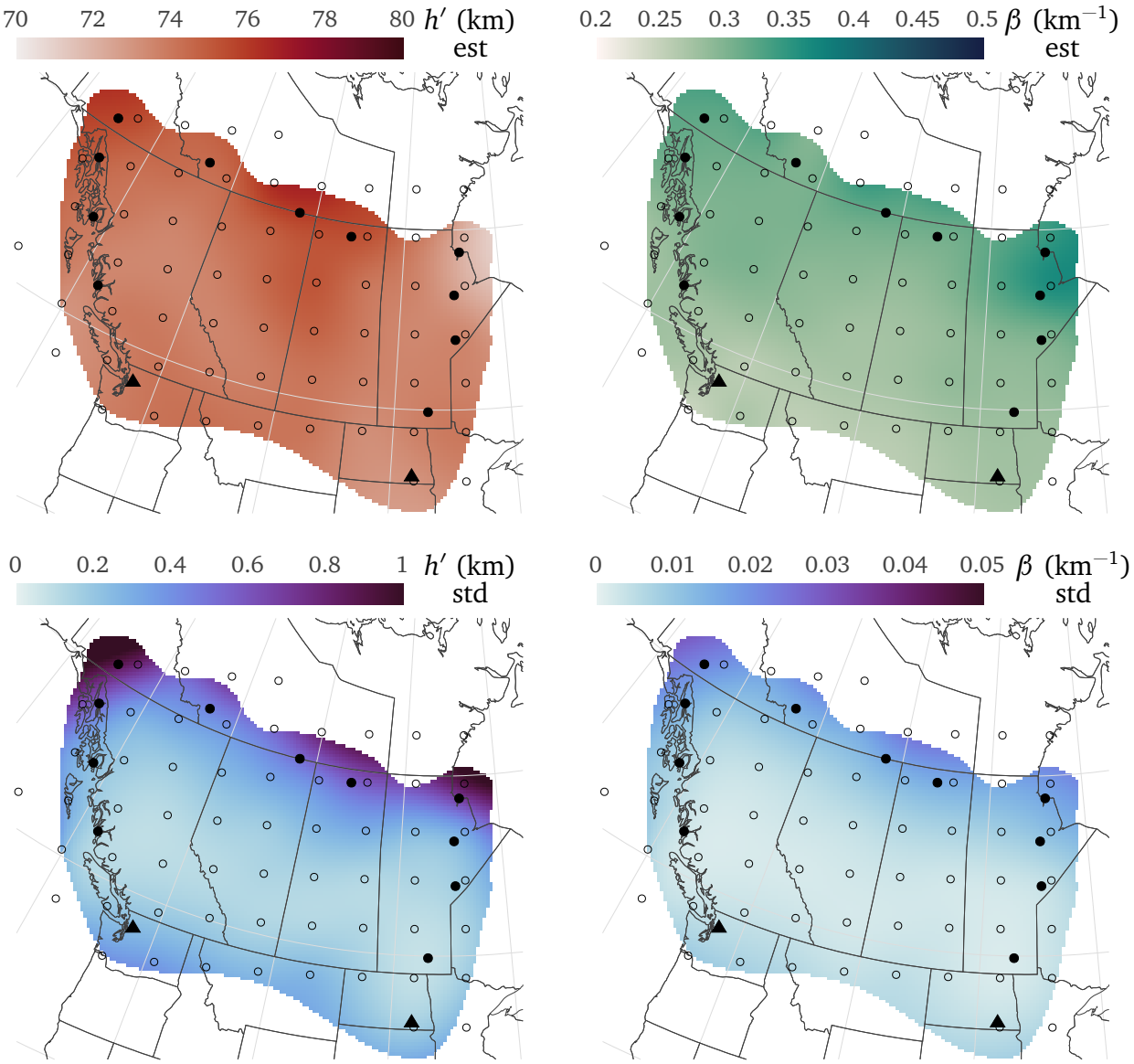


Figure 6.6: Top: Final LETKF estimate of the daytime ionosphere. Bottom: Standard deviation of the LETKF ensemble indicates the uncertainty in the estimate.

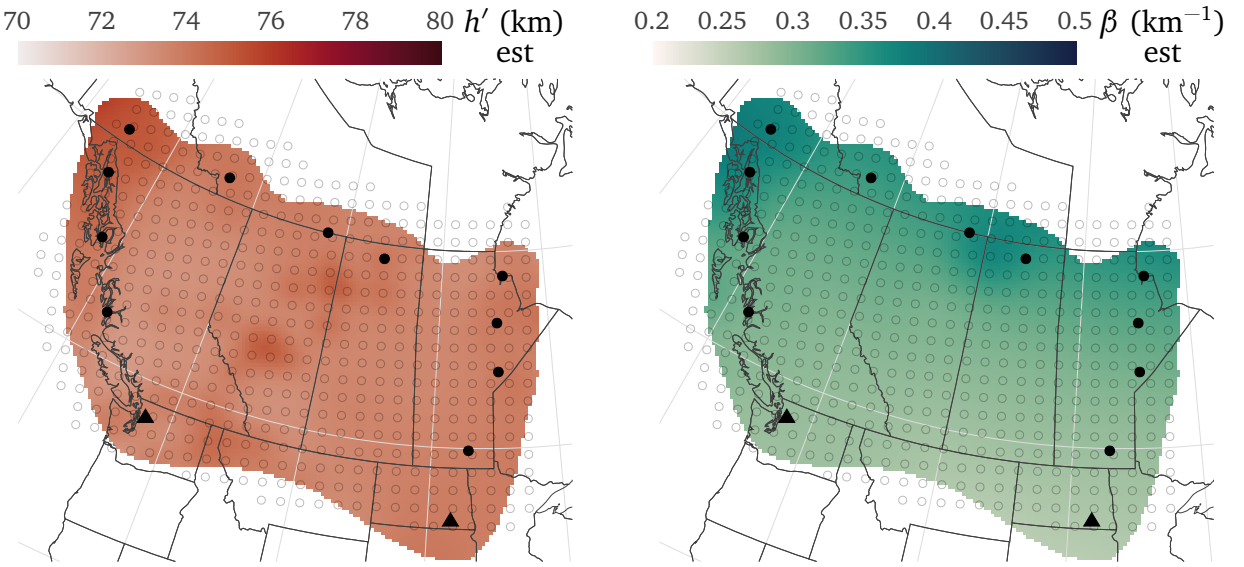


Figure 6.7: Estimated daytime $h'$ (left) and $\beta$ (right) from the multiscale optimization method.

$h'$ and 0.007 km$^{-1}$ in $\beta$ for the LETKF estimate and 0.30 km in $h'$ and 0.016 km$^{-1}$ in $\beta$ for the COBYLA estimate.

Besides the standard deviation map in Fig. 6.6, the LETKF ensemble distributions can be plotted to indicate how the filter converges at specific points. We choose the grid points labelled 1, 2, and 3, shown in Fig. 6.8, to analyze the ensemble. Figure 6.9 plots the $h'$ and $\beta$ error of each member of the ensemble at those three points as well as the Gaussian fit to the ensemble. The estimate confidence increases significantly from the prior to the first iteration and by the second iteration the estimate at each of these points is near the correct value.

Even when the truth ionosphere is unknown, the final estimate can be checked for consistency with the observations by comparing the real observations to the modeled observations through the estimated ionosphere. If these residuals are approximately equal to the measurement noise, than the estimate cannot be rejected by the observations. Figure 6.10 shows the prior and estimated amplitude and phase residuals for both techniques. The horizontal gray colored line indicates the 1- and 2-$\sigma$ noise bounds about zero. The final estimates of both the LETKF and COBYLA-minimized cost function are consistent with the observations. The

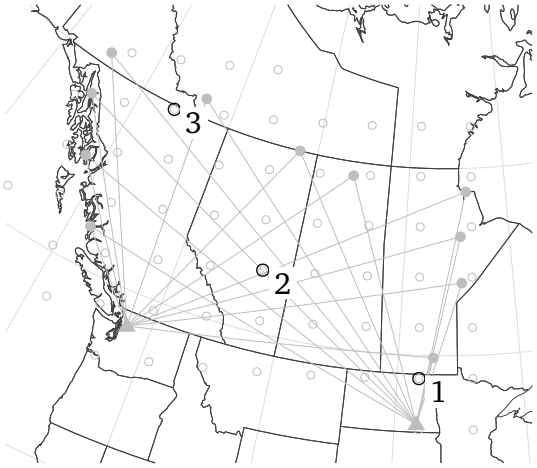


Figure 6.8: Select control points 1, 2, and 3 used for analysis of the LETKF ensemble members.

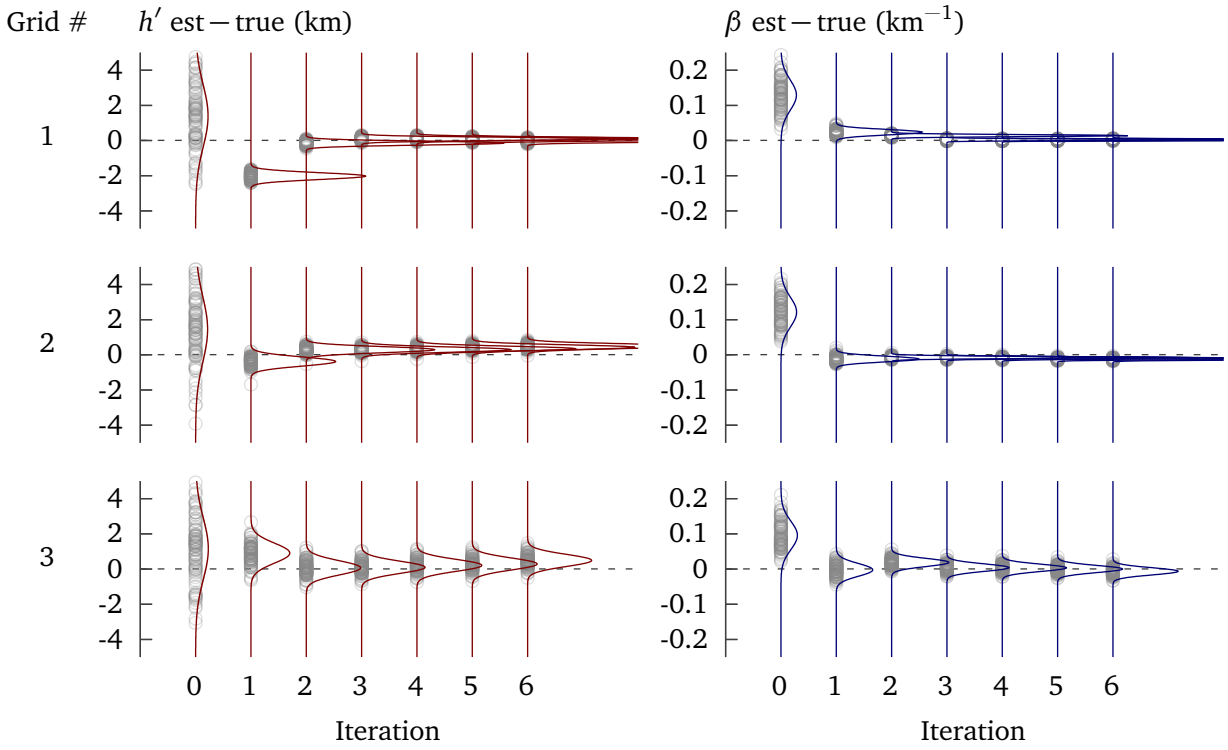


Figure 6.9: Daytime ensemble error in $h'$ (left) and $\beta$ (right) at the prior and after each iteration of the LETKF for select grid control points 1, 2, and 3 from Fig. 6.8. Point 3 has higher variance because of the lower path density in its region, i.e. point 3 is localized by fewer paths than points 1 and 2.

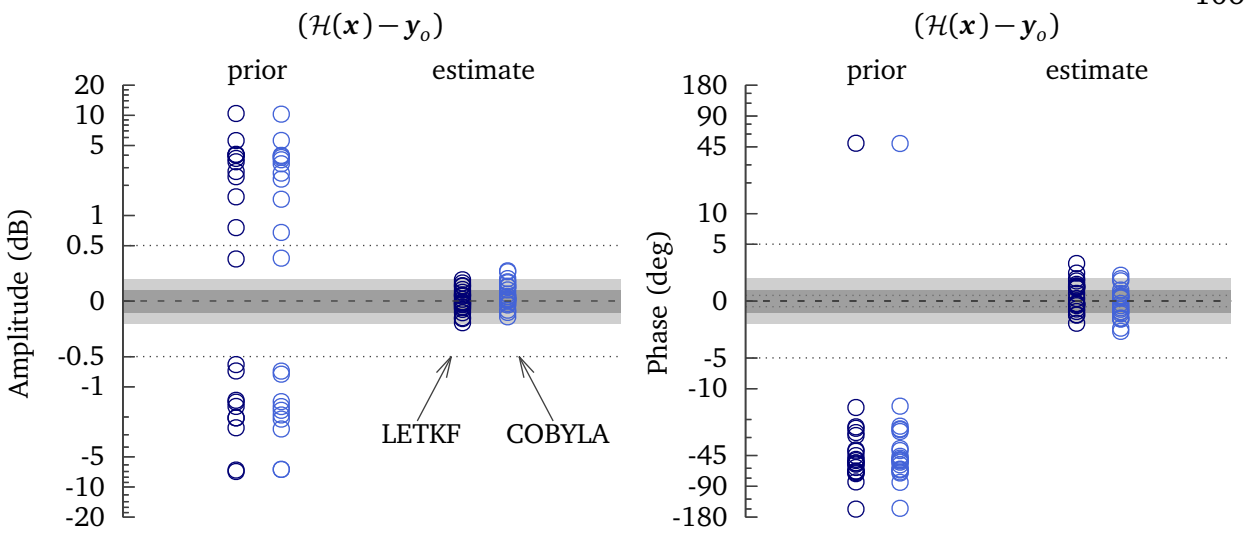


Figure 6.10: Measurement residuals for the prior and estimated daytime ionospheres using both estimation methods. The y-axis of the plot is linear between the horizontal dotted lines and logarithmic outside of those lines. The tiered solid gray bars about zero indicate the first $2\sigma$ of measurement noise.

small differences in the prior residuals between the LETKF and COBYLA are due to the different interpolation methods used for each method.

## 6.2 Terminator scenario

### 6.2.1 Truth model

The terminator scenario uses a coarse model for the terminator ionosphere that transitions from a constant daytime ionosphere to a constant nighttime ionosphere through a logistic curve in solar zenith angle $\chi$:

$$h'(\chi) = \frac{h'_{\text{night}} - h'_{\text{day}}}{1 + \exp\big(-0.7(\chi - 95°)\big)} + h'_{\text{day}} \tag{6.1}$$

where $h'_{\text{day}} = 74\,\text{km}$ and $h'_{\text{night}} = 86\,\text{km}$. This curve is shown in Fig. 6.11. The $\beta$ ionosphere follows an identical curve using $\beta_{\text{day}} = 0.3\,\text{km}$ and $\beta_{\text{night}} = 0.5\,\text{km}$ in place of $h'_{\text{day}}$ and $h'_{\text{night}}$.

This scenario occurs at 2020-03-01 0200 UTC. The truth map is shown in Fig. 6.13. The

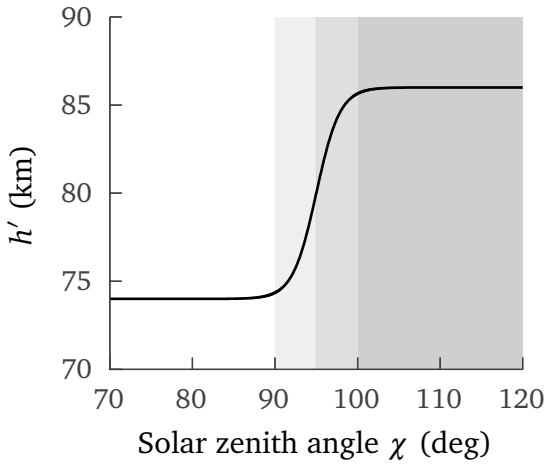


Figure 6.11: $h'$ model for the terminator scenario. $\beta$ follows the same logistic curve adjusted for typical day and night values of $\beta$.

scenario is challenging because the terminator transition occurs over only two columns of the 300 km control point grid for the LETKF method.

### 6.2.2 Prior model

The prior ionosphere for both methods is the Ferguson ionosphere (Section 2.1.2.3) at 2020-03-01 0200 UTC. As evident from Fig. 6.13, the Ferguson ionosphere does not capture the rapid day/night transition of the truth ionosphere. Additionally note that the prior error map in Fig. 6.13 is scaled differently than most of the other error maps. The $h'$ prior is biased from the truth by nearly $-6$ km on the eastern side of the map and by 2 km on the western side. To compensate for this, the prior ensemble variance for the LETKF is increased to $\sigma_{h'} = 3$ km and $\sigma_\beta = 0.05\,\text{km}^{-1}$.

### 6.2.3 Estimates

Figure 6.14 shows the $h'$ and $\beta$ estimates and standard deviations from the LETKF method. Clearly this is a better estimate of the truth than the prior, but there is some local structure on the map. The standard deviation map resembles the standard deviation for the daytime

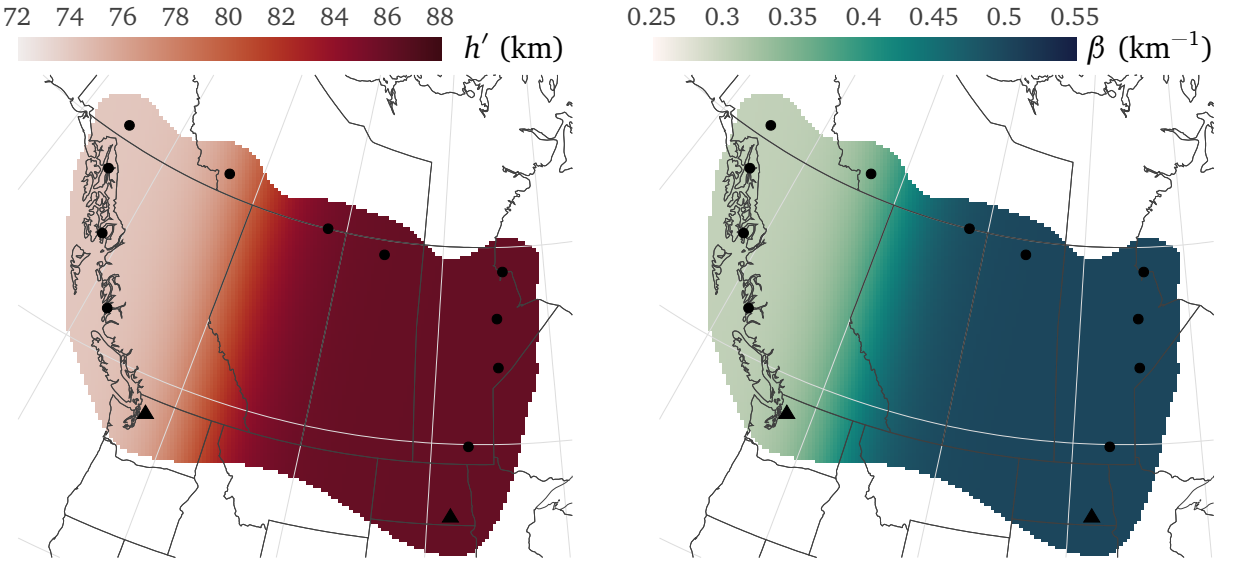


Figure 6.12: True $h'$ and $\beta$ ionosphere parameters for the terminator scenario.

scenario because the paths and the measurement noise for each path are the same.

Figure 6.15 shows the multiscale COBYLA estimate and estimate error. The COBYLA estimate is closer to the truth and has fewer local perturbations compared to the LETKF estimate. The mean absolute deviation from the true ionosphere is 0.85 km in $h'$ and 0.024 km$^{-1}$ in $\beta$ for the LETKF estimate and 0.41 km in $h'$ and 0.017 km$^{-1}$ in $\beta$ for the COBYLA estimate. Both methods are compared at the 300 km control point spacing (step 3 for the multiscale optimization) in Fig. 6.16. COBYLA is closer to the truth at this stage of the optimization as well.

The LETKF $h'$ and $\beta$ ensemble members at control points 1, 2, and 3 from Fig. 6.8 are shown in Fig. 6.17. Although the prior ensemble is wider than the prior used for the daytime scenario and is farther from the truth, the ensemble has similar characteristics over the six iterations. The confidence greatly increases and approaches the truth at the first iteration and changes relatively little over the next six iterations.

The measurement residuals for the terminator scenario are shown in Fig. 6.18. The estimate residuals are slightly worse for this scenario than for daytime, especially for the multiscale

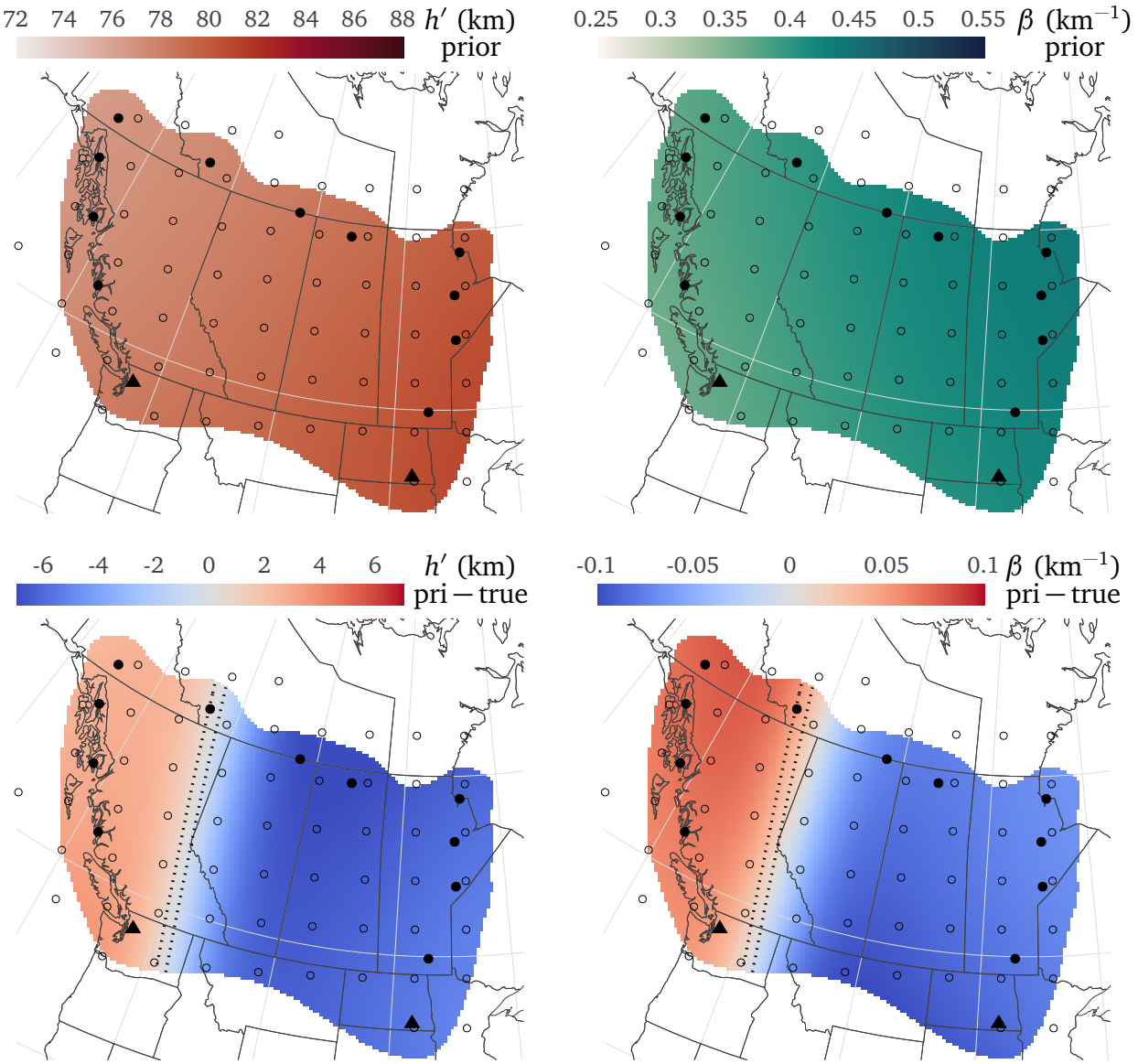


Figure 6.13: Top: Ferguson prior ionosphere for the terminator scenario. Bottom: Difference between the prior and true ionosphere. Note that the $h'$ difference plot uses a much wider scale than the corresponding plot for the daytime scenario.

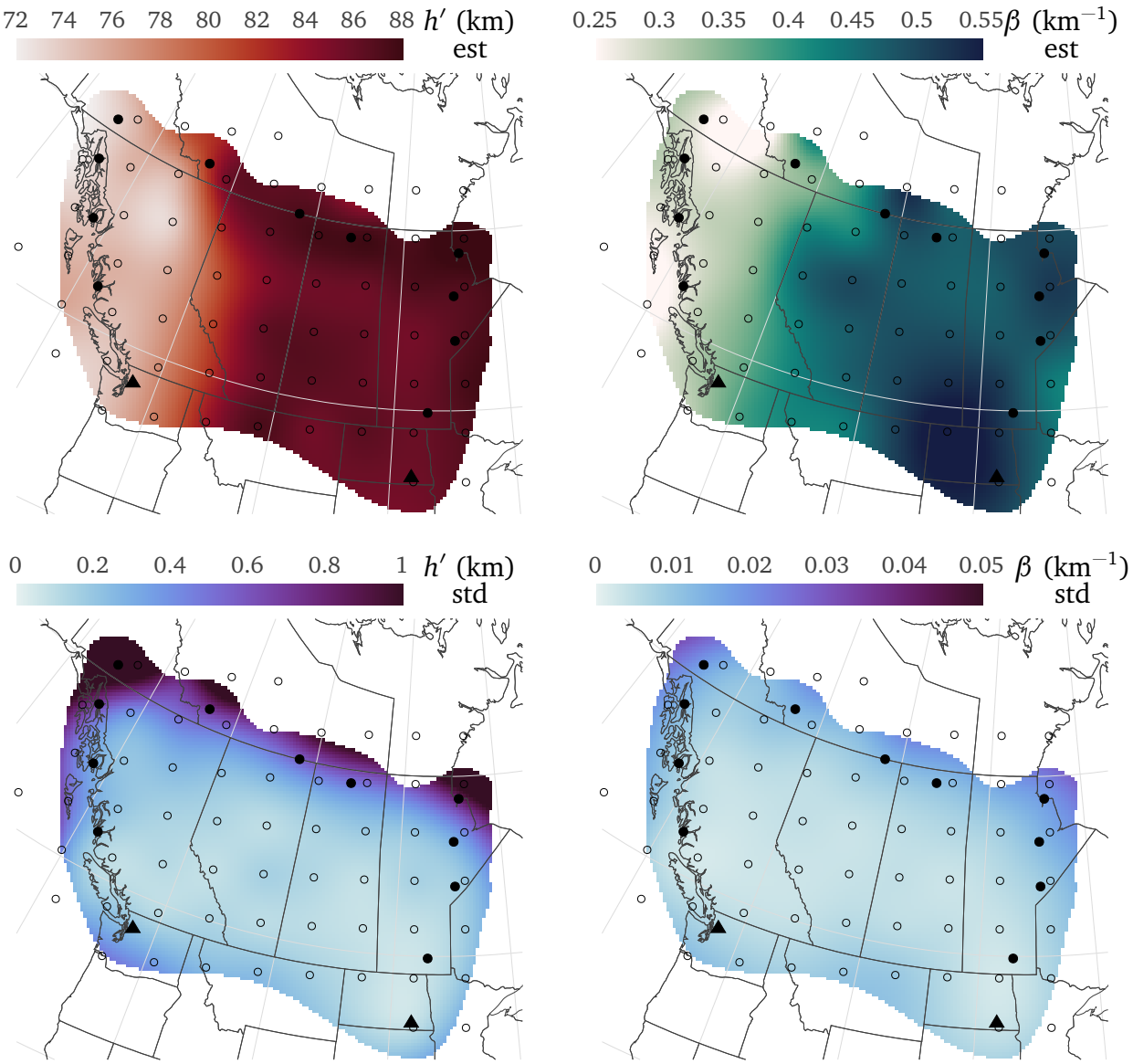


Figure 6.14: Estimated terminator ionosphere (top) and standard deviation (bottom) using the LETKF method.

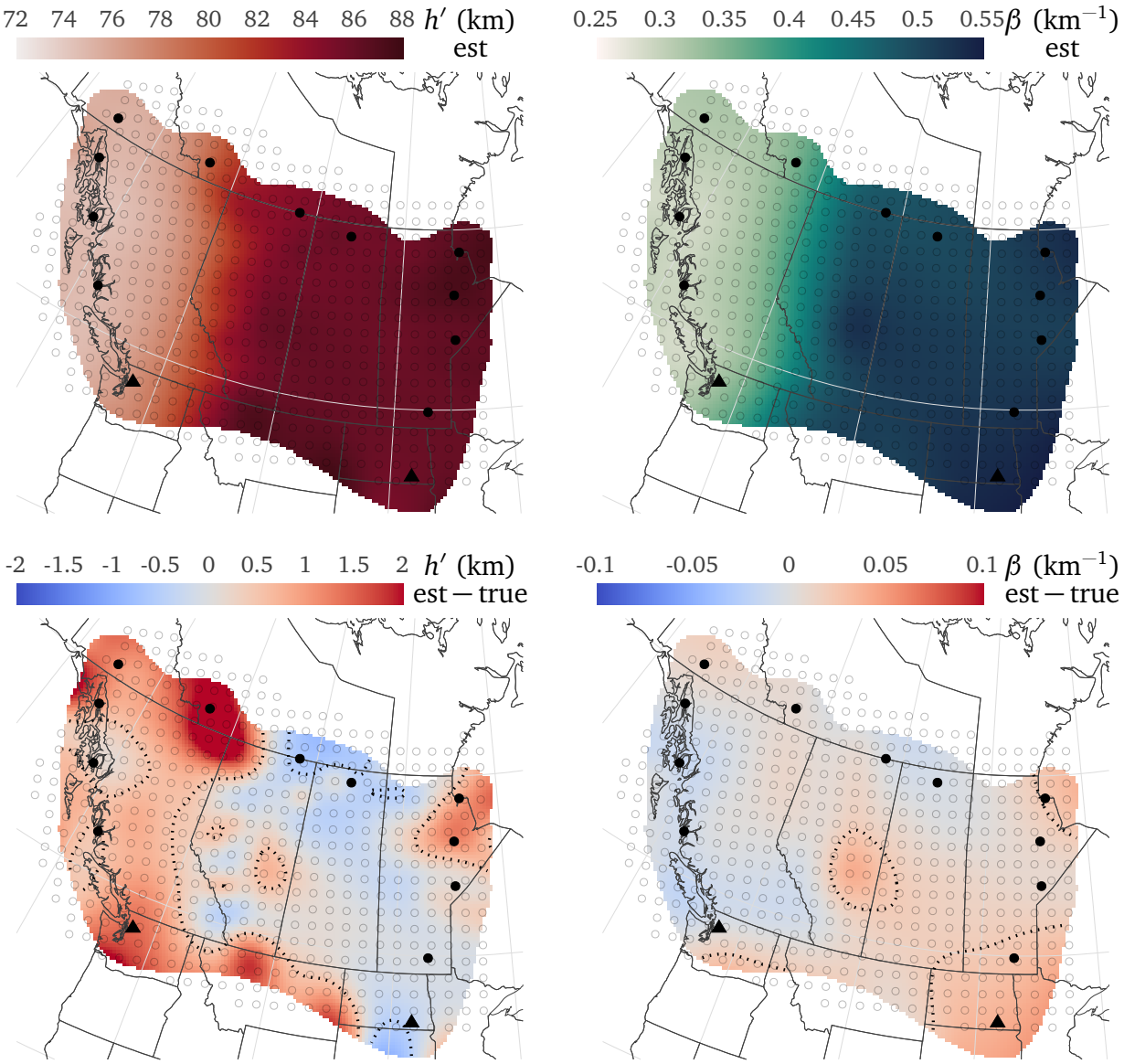


Figure 6.15: Final estimated terminator ionosphere (top) and error from the truth (bottom) using the multiscale COBYLA optimization method.

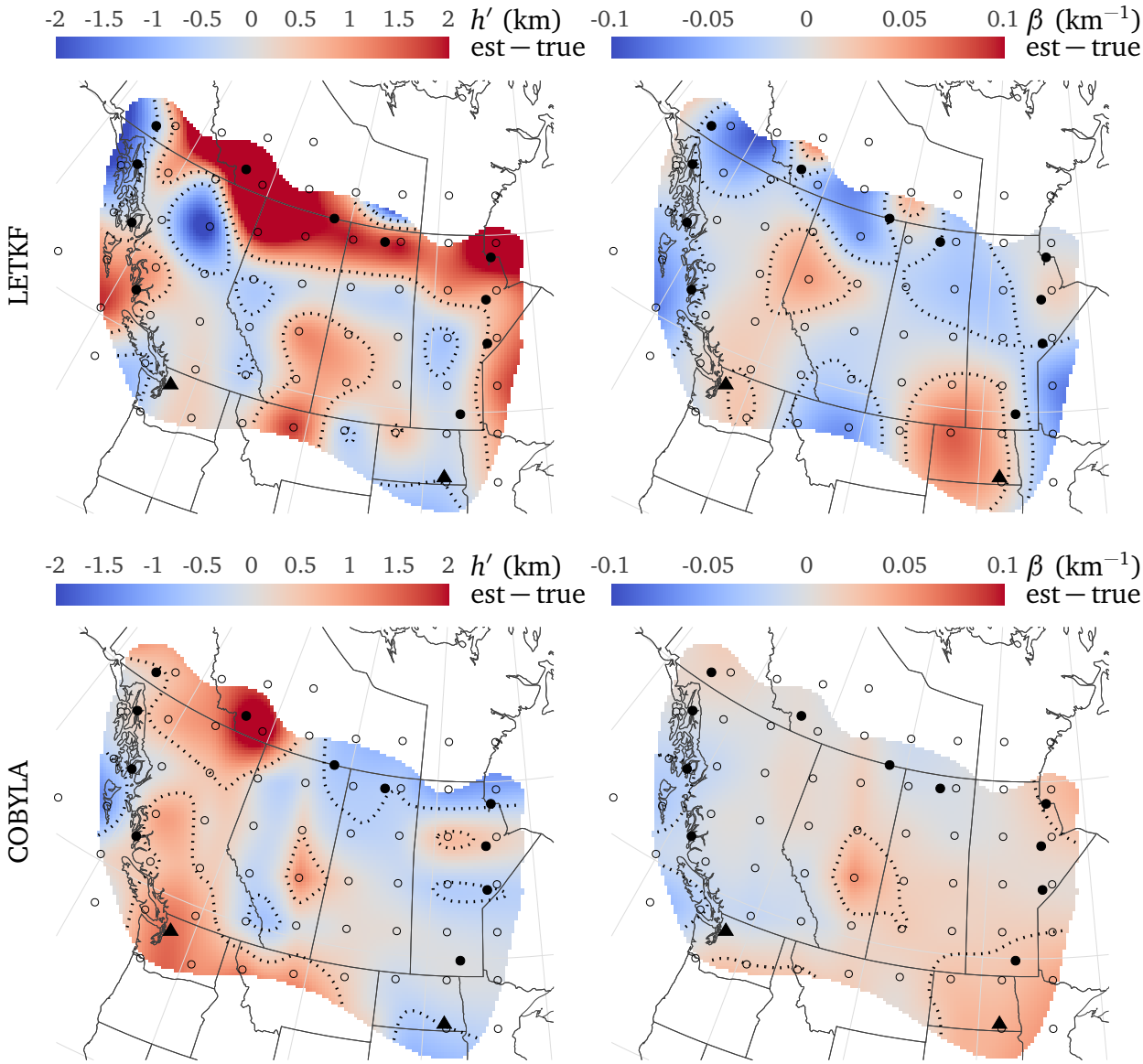


Figure 6.16: Terminator error in the $h'$ (left) and $\beta$ (right) ionosphere parameters for LETKF (top) and the third scale of the COBYLA multiscale optimization (bottom). Both of these estimates occur using the same 300 km control point spacing.

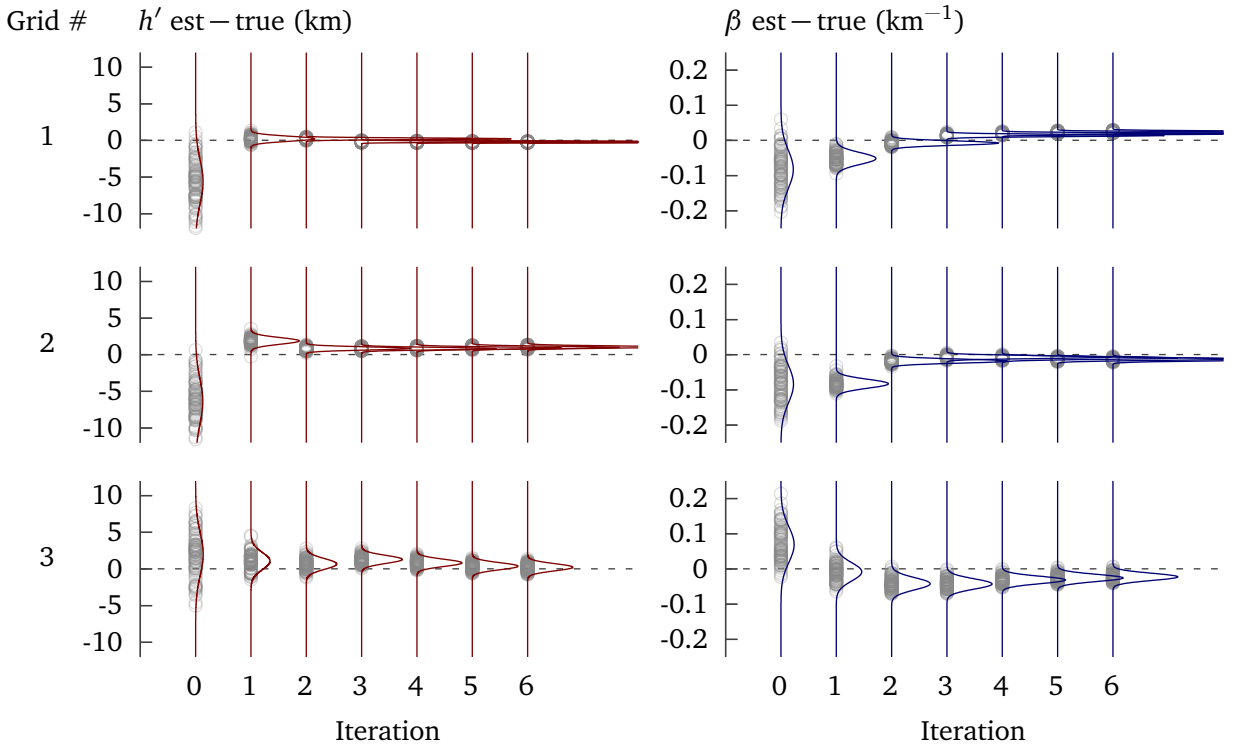


Figure 6.17: $h'$ (left) and $\beta$ (right) ensemble distributions for grid control points 1, 2, and 3 from Fig. 6.8 in the terminator scenario. The $h'$ scale in this plot is wider than the scale used for the daytime scenario in order to accommodate the wide prior estimate of the ionosphere.

COBYLA method. Although LETKF had better measurement residuals, COBYLA actually had the better final estimated ionosphere. This underscores the difficulty of ill-conditioned inverse problems. With real data and an unknown truth, it would be difficult to know which of these methods produced the better estimate.

## 6.3 Sensitivity to model errors

There are several sources contributing to the error in the electron density estimate. There is error in calibrating the received amplitude and phase to each transmitter and in calibrating the forward propagation model amplitude and phase to the transmitters. There are errors due to the discretization and representation of the truth in the estimation scheme, e.g. interpolation between the control points. There are also errors that result from the use of a two-dimensional

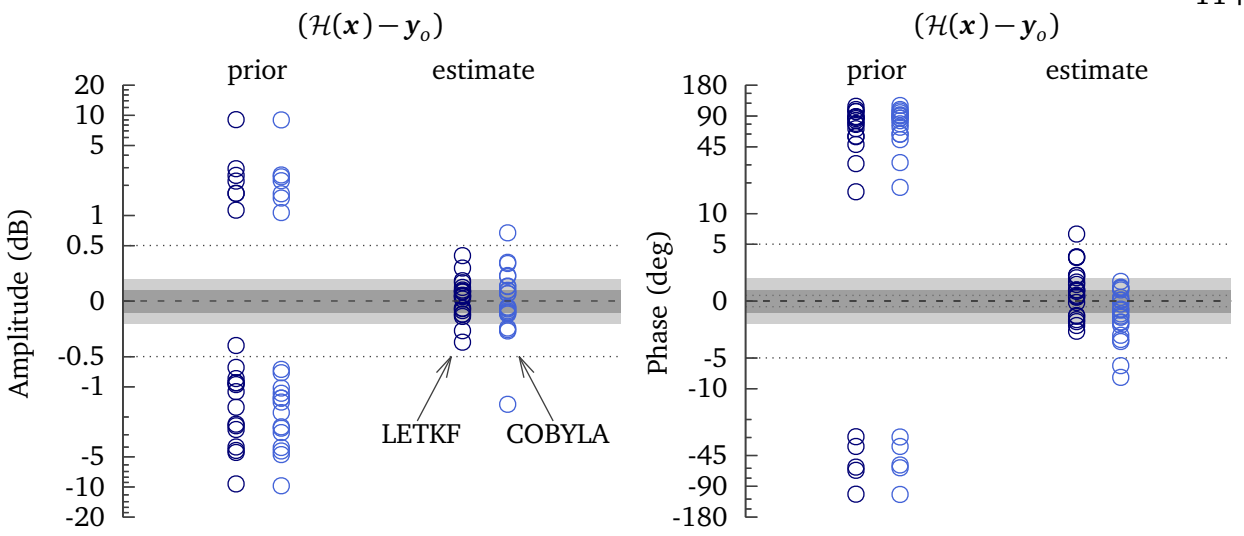


Figure 6.18: Amplitude (left) and phase (right) measurement residuals for the terminator ionosphere scenario using both methods. The y-axis of the plot is linear between the horizontal dotted lines and logarithmic outside of those lines. The tiered solid gray bars about zero indicate the first $2\sigma$ of measurement noise. Although the LETKF method has lower residuals, the multiscale COBYLA method actually produced the better estimate of the ionosphere.

forward model and that result from uncertainty in the forward propagation model parameters that are not estimated, e.g. ground conductivity. It is important that we understand the response of the estimation methods to these errors when evaluating real estimates. In this section, we examine how each of the two estimation methods respond to errors in the magnetic field, ground conductivity map, and conductivity profile for the daytime scenario shown above. Although this is not an exhaustive investigation, it gives some indication of how robust each estimation method is to these errors.

### 6.3.1 Magnetic field

In the day and terminator scenarios above, the IGRF-13 (Thébault et al., 2015) magnetic field model was used to model the simulated truth observations and by the estimation forward model, i.e. the magnetic field was perfectly known. The real magnetic field in the ionosphere is more complicated than the baseline reference model provided by IGRF. One intermediate-

fidelity geomagnetic field model is CHAOS-7 (Finlay et al., 2020), which consists of static and time-dependent internal fields, and a model of the magnetospheric field and its induced counterpart. CHAOS-7 is derived from ground and satellite-borne geomagnetic measurements. The difference between the CHAOS-7 and IGRF-13 fields at 60 km altitude at the date and time of the daytime scenario is shown in Fig. 6.19.

The difference in field strength over the northern regions of the map reaches 300 nT; however, the difference in the observed VLF transmitter signals is on the order of the typical measurement noise, as shown in Fig. 6.20. Therefore, the estimation results (Fig. 6.21) are similar to the daytime results already shown. Based on this, the IGRF-13 model is sufficient to describe real world propagation. As we will later see, this magnetic field experiment has the lowest effect on observations compared to the collision frequency and ground error experiments that follow this section. The influence of the magnetic field on the observations would be greater for a nighttime scenario (see Section 4.3) and should be investigated if performing nighttime ionosphere estimates.

### 6.3.2 Collision frequency

In the original daytime scenario the collision frequency profile assumed by Wait and Spies (1964) was used both in the estimate forward model and to generate the simulated truth observations. The real-world collision frequency profile is a function of the background number density and temperature profiles (Beharrell & Honary, 2008; Ieda, 2020). In this section we generate simulated truth observations using three different collision frequency maps that scale the exponential collision frequency profile used by Wait and Spies (1964). All three center a Gaussian collision frequency disturbance region at 54.238° N, 104.392° W and use the profile

$$\nu(z) = 1.816 \times 10^{11}(1 + W\rho)\exp(-0.15z) \tag{6.2}$$

where height $z$ is in kilometers, $W$ is a maximum weighting term and $\rho$ is a weight that varies between 0 and 1 to capture the Gaussian shape of the disturbance region. The first experiment

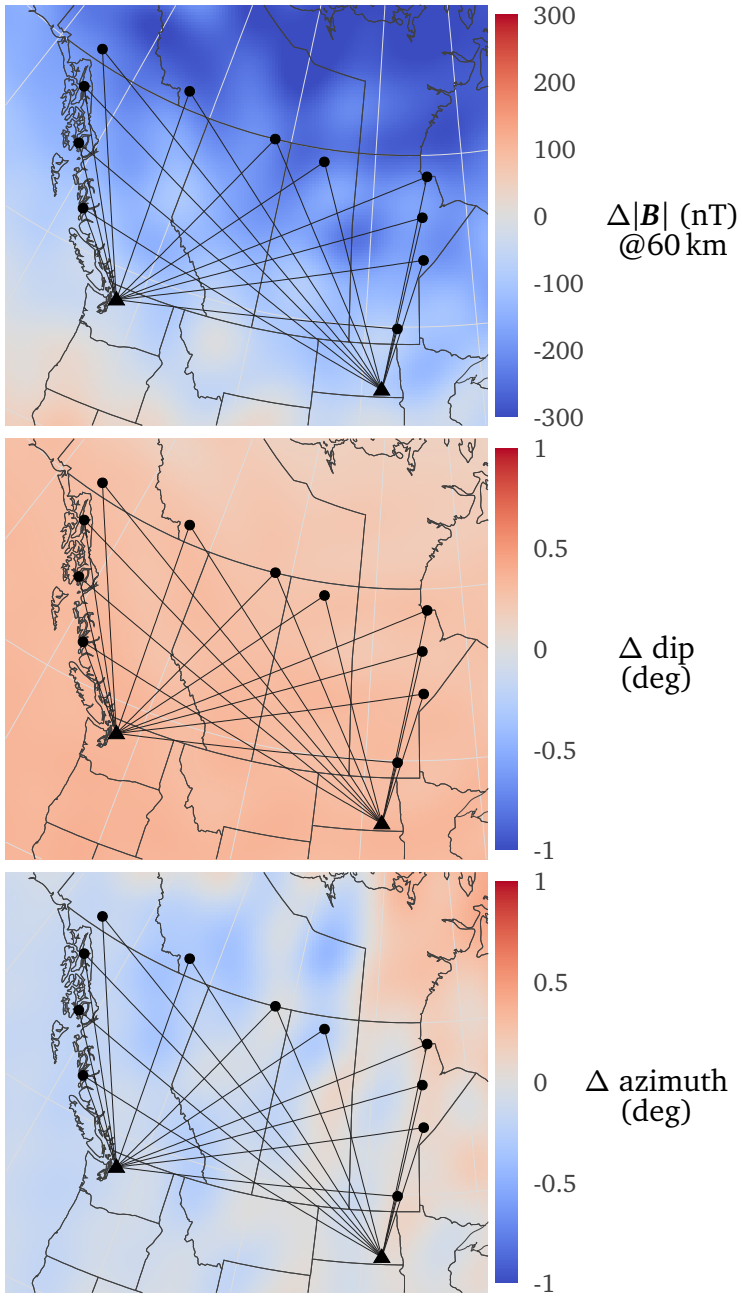


Figure 6.19: The difference in magnitude (top), dip angle (middle), and azimuth (bottom) between the CHAOS-7 and IGRF-13 magnetic field models at 60 km altitude.

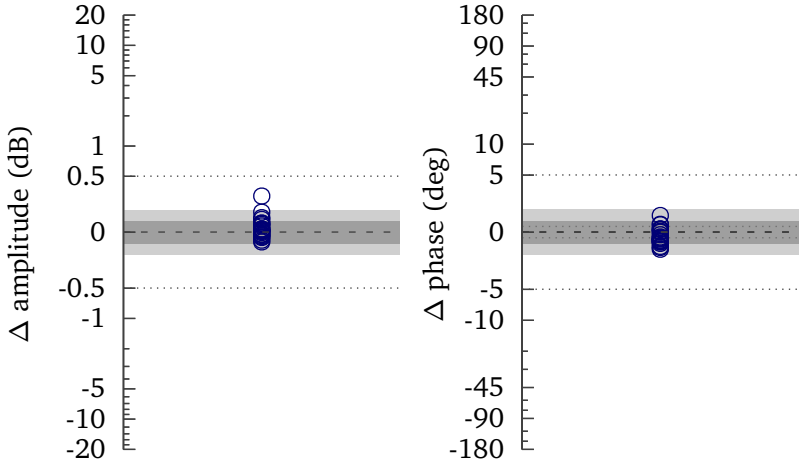


Figure 6.20: Difference in amplitude and phase observed on each propagation path between the IGRF-13 and CHAOS-7 magnetic field models for the daytime scenario.

(a) uses a disturbance width (spatial standard deviation) of 800 km and $W = 0.2$. In other words, the peak disturbance collision frequency is 120 % of the undisturbed collision frequency. The second experiment (b) increases $W$ to 0.5. The third experiment (c) uses a much larger disturbance width of 1800 km, effectively covering the entire map, and $W = 0.1$. A map of the difference in collision frequency at 75 km for these three experiments compared to the background profile is shown in Fig. 6.22.

The difference in observations between these three experiments and the undisturbed daytime scenario is shown in Fig. 6.23. There are some significant changes to the observations, well outside of the typical measurement noise. Unsurprisingly, experiment b) has a larger effect than a). Experiment c) has a slightly lower change than experiment a) and is biased similarly to the other experiments despite the fact that the entire map is affected by the disturbance.

Figure 6.24 shows the $h'$ estimate error for each experiment. One feature to point out is that the "error" is higher in the region where the disturbance is centered. This is most obvious in the LETKF estimate of experiment b). It is logical that a large collision frequency disturbance would be captured because the ionosphere conductivity is affected by both the electron density and collision frequency. A change in the collision frequency profile can be represented by

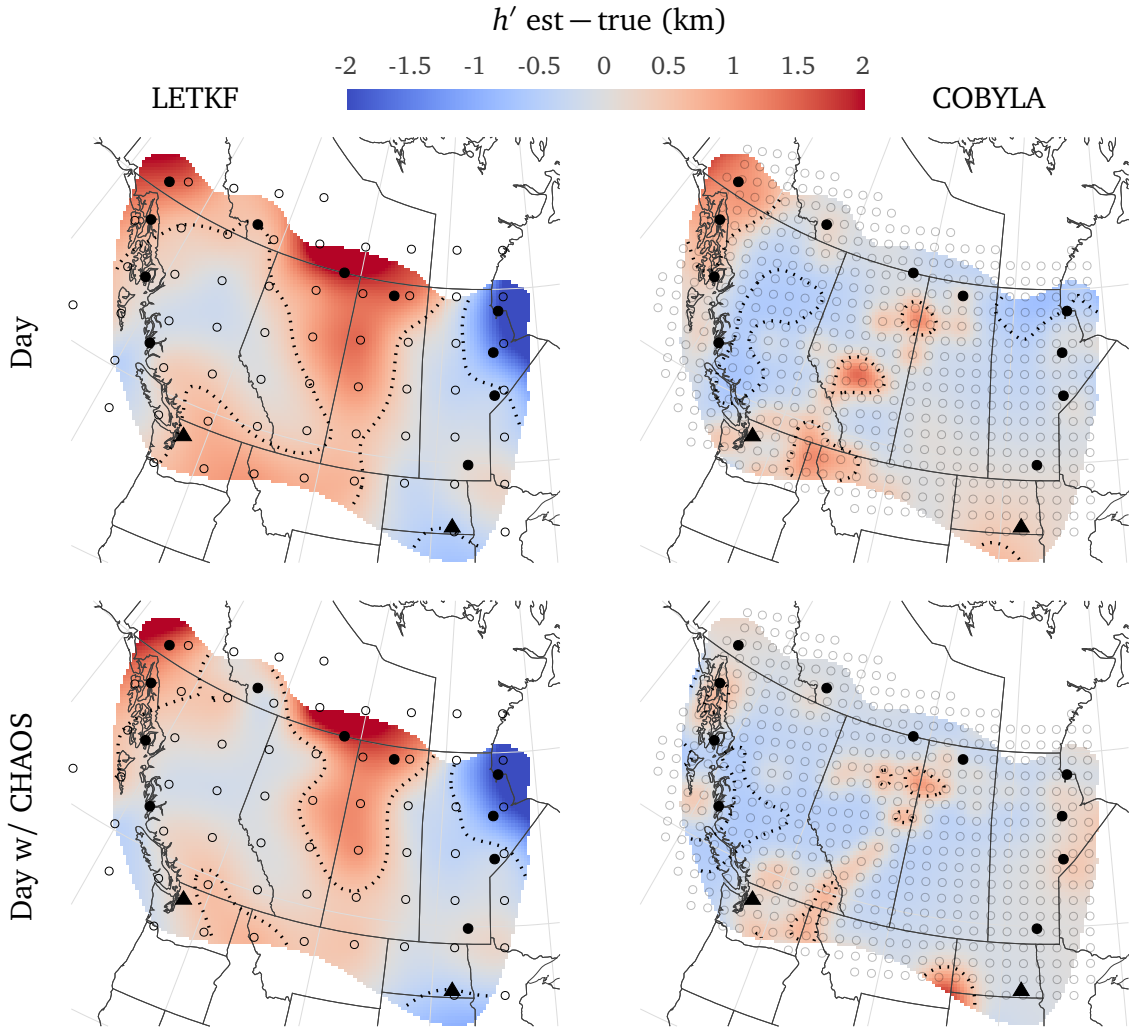


Figure 6.21: Top: The final $h'$ estimate error of the original daytime scenario using LETKF (left) and COBYLA (right), reprinted here to make visual comparison of the results easier. Bottom: $h'$ estimate error when the simulated truth observations use the CHAOS-7 magnetic field model and the estimation uses the IGRF-13 field model.

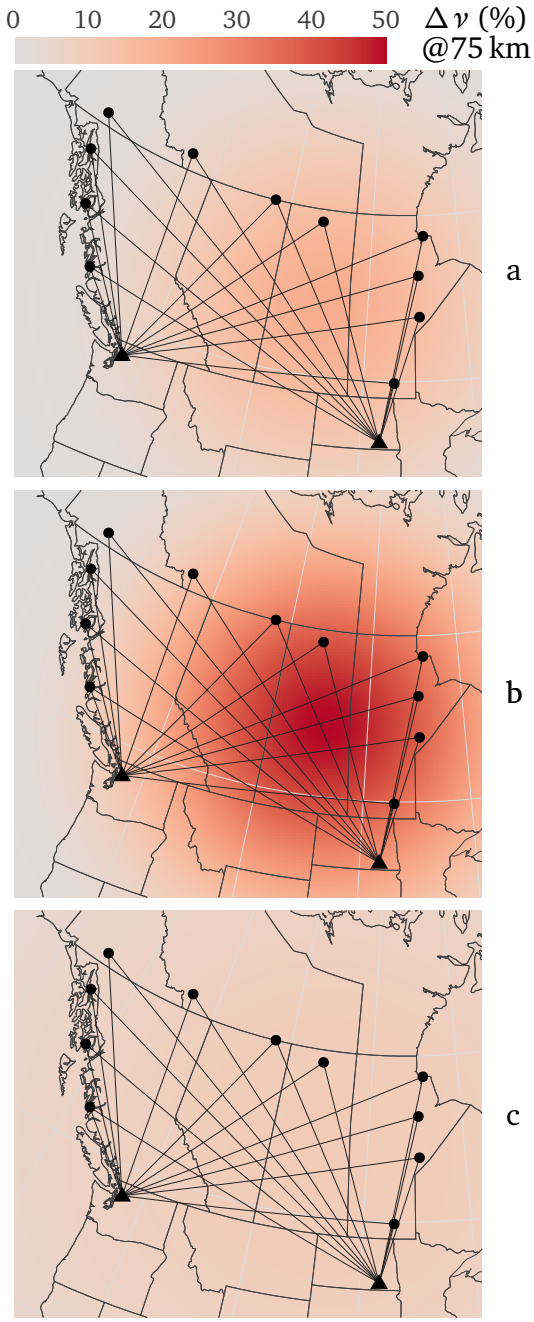


Figure 6.22: Percentage difference in collision frequency at 75 km altitude between disturbance experiments a) (top), b) (middle), and c) (bottom) and the exponential collision frequency profile assumed by Wait and Spies (1964).

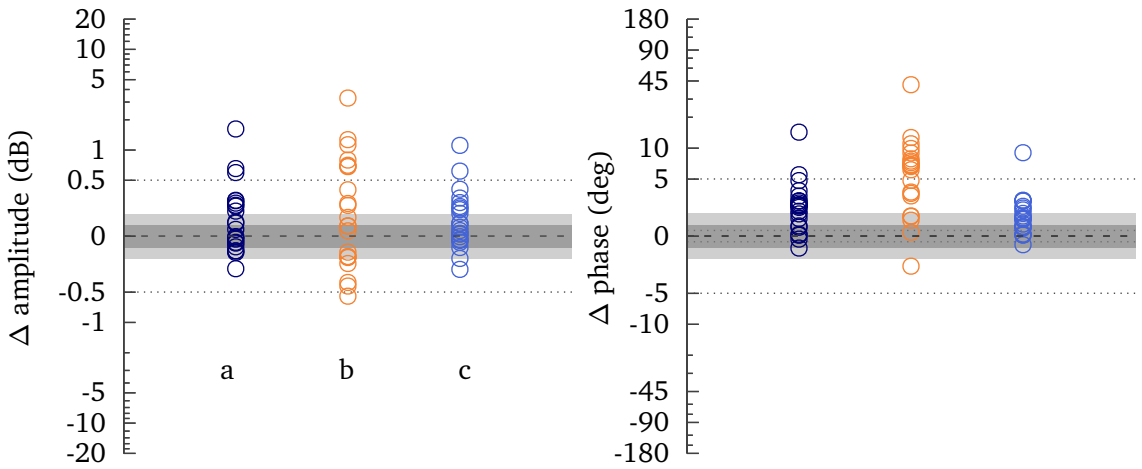


Figure 6.23: Difference in amplitude and phase observed on each propagation path between disturbed collision frequency profile experiments a), b), and c) (see Fig. 6.22) and the background exponential profile from Wait and Spies (1964).

a corresponding change in the Wait parameters for the electron density (Section 2.1.2.3). More accurate collision frequency profiles may be obtained from MSIS neutral density and temperature profiles using the work of Ieda (2020), but care must be taken when comparing the estimated $h'$ and $\beta$ values to the work of others who assume an exponential collision frequency profile.

### 6.3.3 Ground conductivity

Real-world ground conductivity has changed since the ground conductivity map distributed with LWPC was produced; sea water and polar ice boundaries have moved and industrial areas have grown. The conductivity of each of these grounds have a particularly strong influence on the propagation of VLF waves in the EIWG (see Section 4.3). Real-world ground conductivity is also more complicated than the representation in the ground conductivity map distributed with LWPC. In this section we add random errors and noise to the ground conductivity map used to build the simulated observation experiments. Figure 6.25 shows the original relative permittivity and conductivity of the ground as well as the modified versions used to generate the

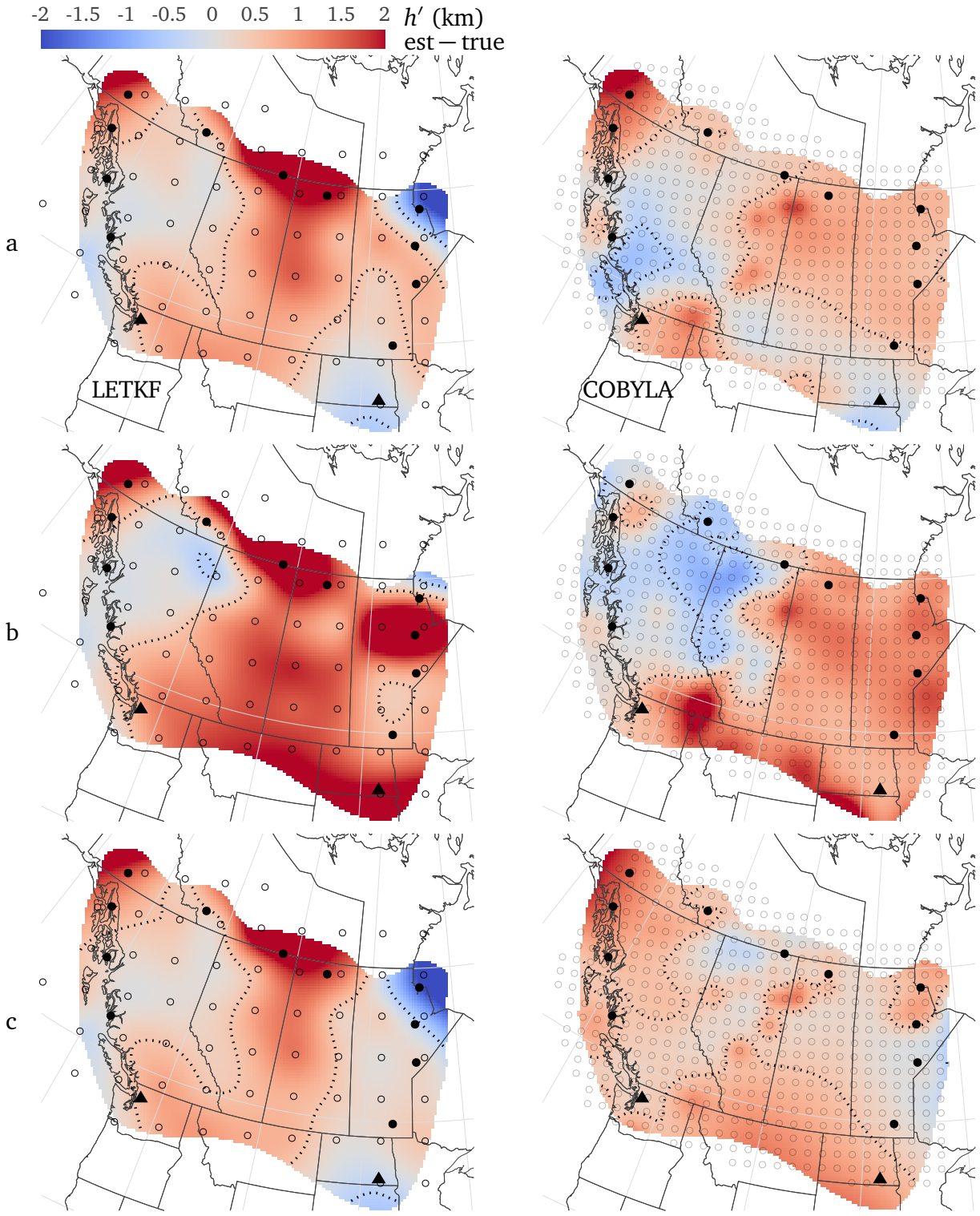


Figure 6.24: $h'$ estimate error maps for disturbed collision frequency experiments a), b), and c) (see Fig. 6.22) for LETKF (left) and COBYLA (right).

simulated truth observations. The ground conductivity map uses 10 “classes” of ground, each having different combinations of permittivity and conductivity. Ground modification experiment a) weights the original class at each point with a weight of 200 and weights all other classes with a weight of 1. The ground in a) is randomly chosen using this weighting scheme and then a 1 % Gaussian noise is applied to the conductivity. Experiment b) is similar but sets the original class weight to a value of 100 (so that it is roughly half as likely to use the original class at each geographic point). Experiment c) uses an original class weight of 100, the two nearest classes (more and less conductive) use a weight of 5 and all other classes use a weight of 1. Therefore, there is a higher probability of the ground map being off at a given point by one ground class. Finally, a 5 % Gaussian noise is added to the ground conductivity in experiment c).

The amplitude and phase difference from the original daytime scenario observed on each propagation path to each experiment is shown in Fig. 6.26. Of the three model errors examined here (magnetic field, collision frequency profile, and ground), the ground errors have the largest effect on the observations. As seen in Fig. 6.27, this translates to high $h'$ estimate errors. Unlike the previous investigations, the ground errors are randomly distributed across the map. On one hand, real-world errors in the ground conductivity map are probably mostly along transitions between different ground classes. On the other hand, industrial areas and cities are located sporadically around the map. It is difficult to evaluate how realistically the randomly distributed errors in experiments a), b), and c) capture the true errors in the map without building or obtaining a new ground conductivity map from modern Earth conductivity measurements (Kelbert et al., 2006). Given that the ground has the greatest effect on the ionosphere estimates of the model errors examined here, further work should be performed to establish a modern ground conductivity map for VLF use. Although the geophysics community has recently made great efforts to map Earth conductivity (Kelbert, 2019), their methods are very different from the methods historically used to model effective VLF ground conductivity.

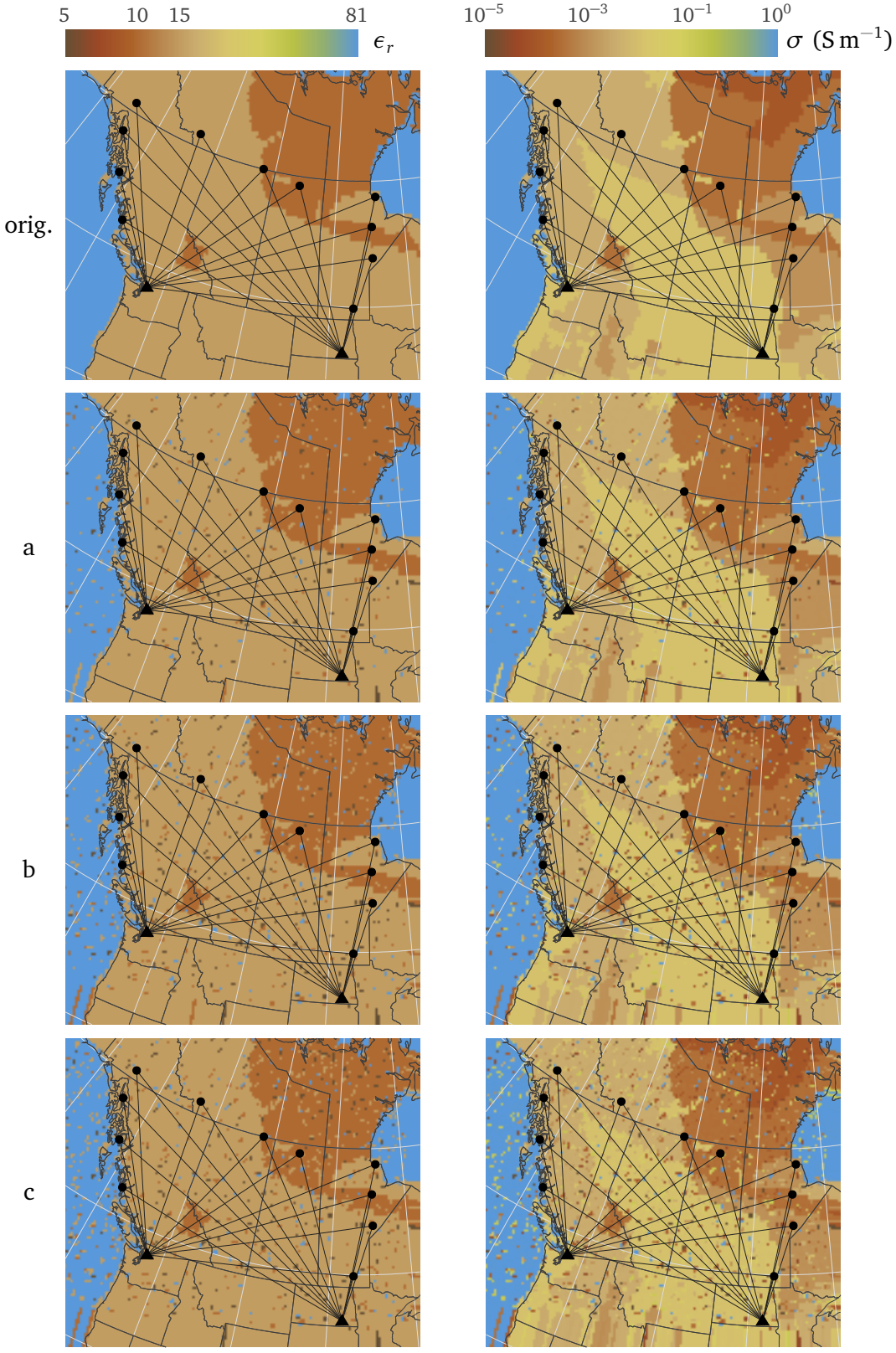


Figure 6.25: Relative permittivity (left) and conductivity (right) of the original (top) and ground error experiments a), b), and c). Only ground along each propagation path affects the simulated observations; modifications to the sea and along the outside regions of the map do not.

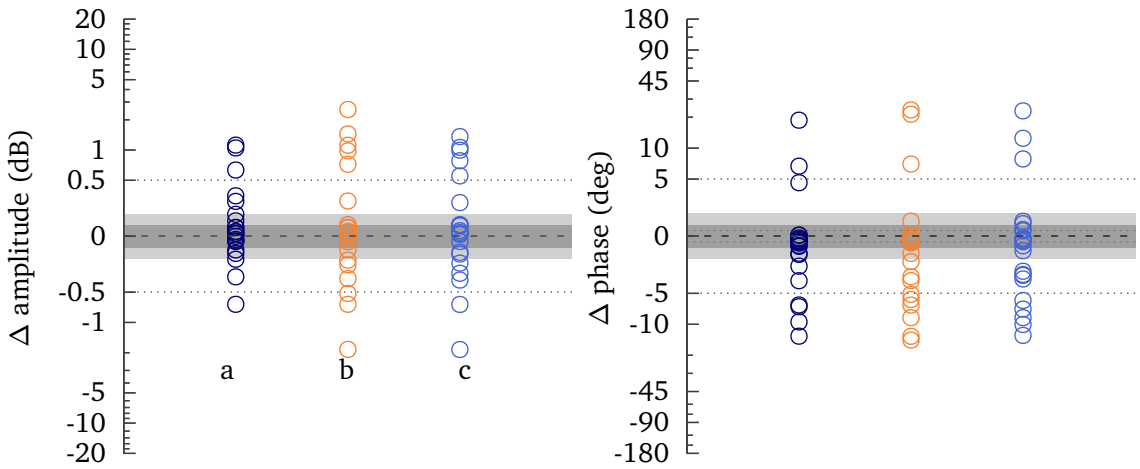


Figure 6.26: Amplitude and phase difference on each propagation path between the original ground map and modified ground experiments a), b), and c).

Otherwise, given that ground conductivity changes slowly compared to ionosphere conductivity, it may be possible to establish baseline VLF observations against which new observations are compared.

### 6.3.4 Comments

In each of the above experiments, COBYLA continued to outperform the LETKF in terms of the estimate error. The primary explanation for this consistent difference is that the cost function being minimized by COBYLA includes an explicit model cost in addition to the data cost while the LETKF measurement update is made in response only to a data cost (residual). Although the estimate covariance is inflated between each set of observations assimilated into the LETKF estimate, the filter is still prone to overfitting the data over repeated iterations. This is one explanation for why the LETKF has slightly better residuals than COBYLA in Fig. 6.18, even though COBYLA produced the lower error estimate. However, the downside to COBYLA is that the regularization parameter must be chosen by the investigator. Although we used a regularization parameter of $\lambda_{h'} = 1$ and $\lambda_{\beta} = 10$ for every experiment here, in practice the COBYLA estimate should be produced using several different regularization parameters.

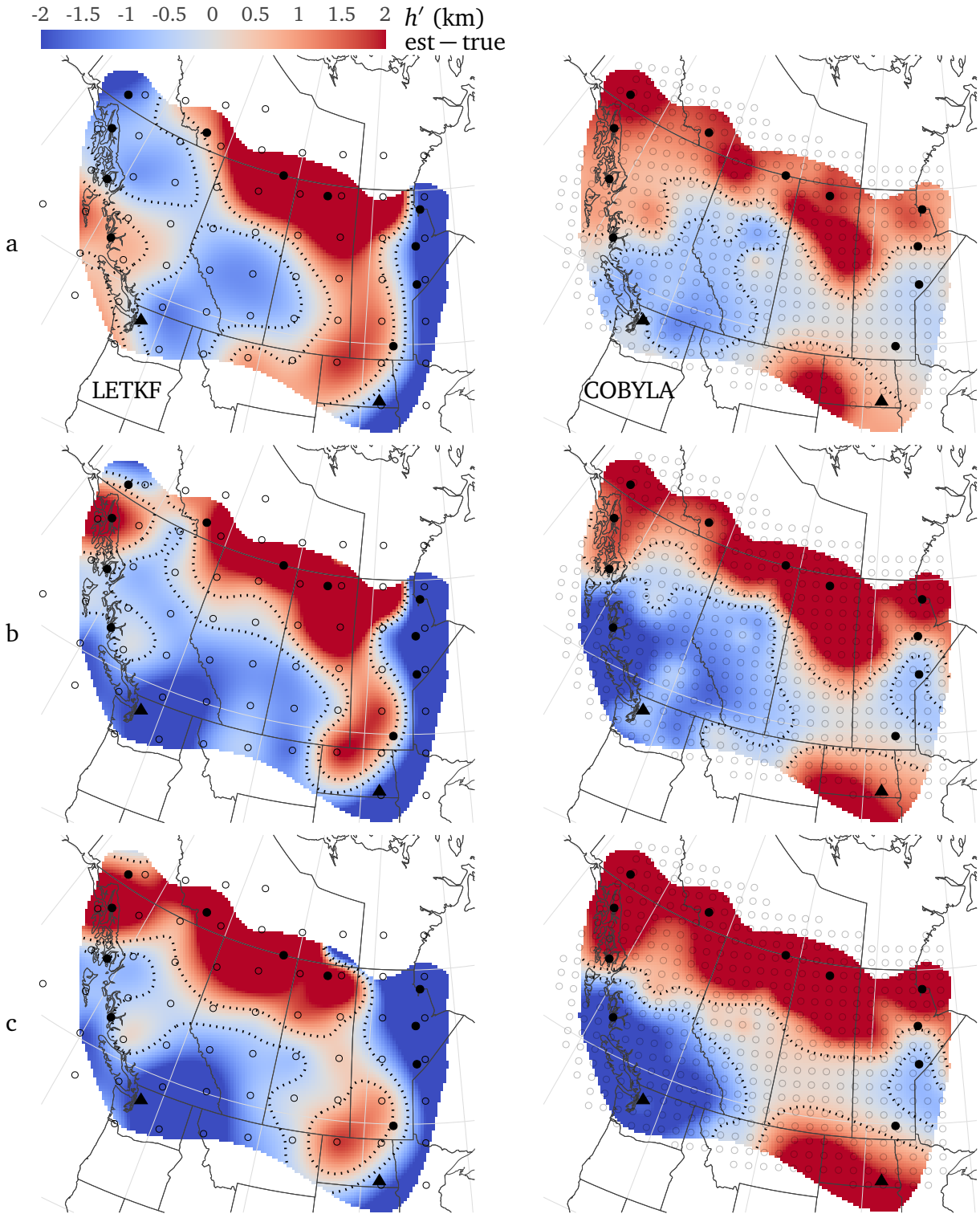


Figure 6.27: $h'$ estimate error for modified ground experiments a), b), and c), shown in Fig. 6.25, for LETKF (left) and COBYLA (right).

Not only does this greatly increase the time required to produce the final estimate, but the selection of the final estimate depends upon the opinion of the investigator. If the regularization parameter is too low, COBYLA can overfit just like the LETKF.

Given the number of uncertainties involved with forward modeling longwave propagation in the EIWG, COBYLA provides the more robust estimate of the two methods. The model cost penalizes changes in the estimate and produces the simpler model. Observation errors that increase the data cost due to forward model errors are counteracted by the model cost. The LETKF estimate includes the effect of model errors and biases that are inevitable in this problem.

## 6.4 Energetic particle precipitation

### 6.4.1 Background

One application of a two-dimensional D-region estimation technique is imaging the spatial and temporal extent of energetic particle precipitation (EPP) from the radiation belts. Charged particles trapped by Earth's magnetic field collectively form regions known as the Van Allen radiation belts (Van Allen, 1959). The inner belt, extending from 0.2 to 2 $R_e$, consists of energetic ions and electrons, and the outer belt, extending from 3 to 10 $R_e$, consists primarily of energetic electrons (Blum & Breneman, 2020). Every particle travels along a magnetic field line yet has velocity components both parallel $v_{\parallel}$ and perpendicular $v_{\perp}$ to the magnetic field. This causes the particle to have a helical trajectory as it moves along the field line. The magnetic field gets stronger near the poles and causes the particle to be repelled back towards space. This results in a particle "bouncing" between the north and south poles while traveling along a field line.

The mirror point at which a particle is repelled is determined by the particle's pitch angle $\alpha$

$$\alpha = \arctan\left(\frac{v_{\perp}}{v_{\parallel}}\right). \tag{6.3}$$

Particles with equatorial pitch angle near 90° are mirrored before particles with small pitch angles. Particles with very small pitch angles travel so far before mirroring that they may collide

with neutral species in Earth's atmosphere and are effectively lost from the radiation belt. The range of pitch angles for which collisions are expected to occur is known as the loss cone. Electrons with energies of tens of keV or higher that are in the loss cone and collide with a neutral are capable of producing ionization in the D-region ionosphere. This is called energetic particle precipitation (EPP) and is detectable with subionospheric VLF (Clilverd et al., 2009; Clilverd et al., 2017b).

### 6.4.2 Simulation

Using the tabulated EPP ionization profiles from Xu et al. (2020), we are able to construct charge density profiles for the D-region using a similar process to that shown in Marshall and Cully (2020). The process, depicted in Fig. 6.28, allows us to construct an electron density profile with EPP as a function of latitude, longitude, and date/time. Neutral density and temperature profiles are obtained from NRLMSISE-00 (Picone et al., 2002; Chagas et al., 2019) and my FIRITools.jl package is used to interpolate FIRI-2018 profiles (Friedrich et al., 2018) to the correct time, location and nearest F10.7 index over a three day window.

Whittaker et al. (2013) analyzed the energy spectra of electron precipitation measured by the DEMETER satellite and found that most distributions in the 100 keV–2.2 MeV range fit exponential or power-law energy distributions

$$f(E) = C_e \exp(-E/\beta_e) \quad \text{or} \quad f(E) = C_p E^{-\beta_p} \tag{6.4}$$

where $\beta_e$ and $\beta_p$ characterize the shape of the distribution. We assume precipitating electrons can be described by an exponential energy distribution of 90 keV to 2.2 MeV in 10 keV steps with $\beta_e = 200$ keV. They have a uniform pitch angle distribution from 0° to 90°. The ionization profile obtained from the EPP lookup table is multiplied by the precipitating flux, $10^5$ $e^-/cm^2/$ sec, say, and input to the GPI model (Lehtinen & Inan, 2007) as an external ionization source. GPI outputs the perturbed charge density profiles which can then be input to LongwaveModePropagator.jl (Gasdia & Marshall, 2021) to produce simulated observations at each receiver.

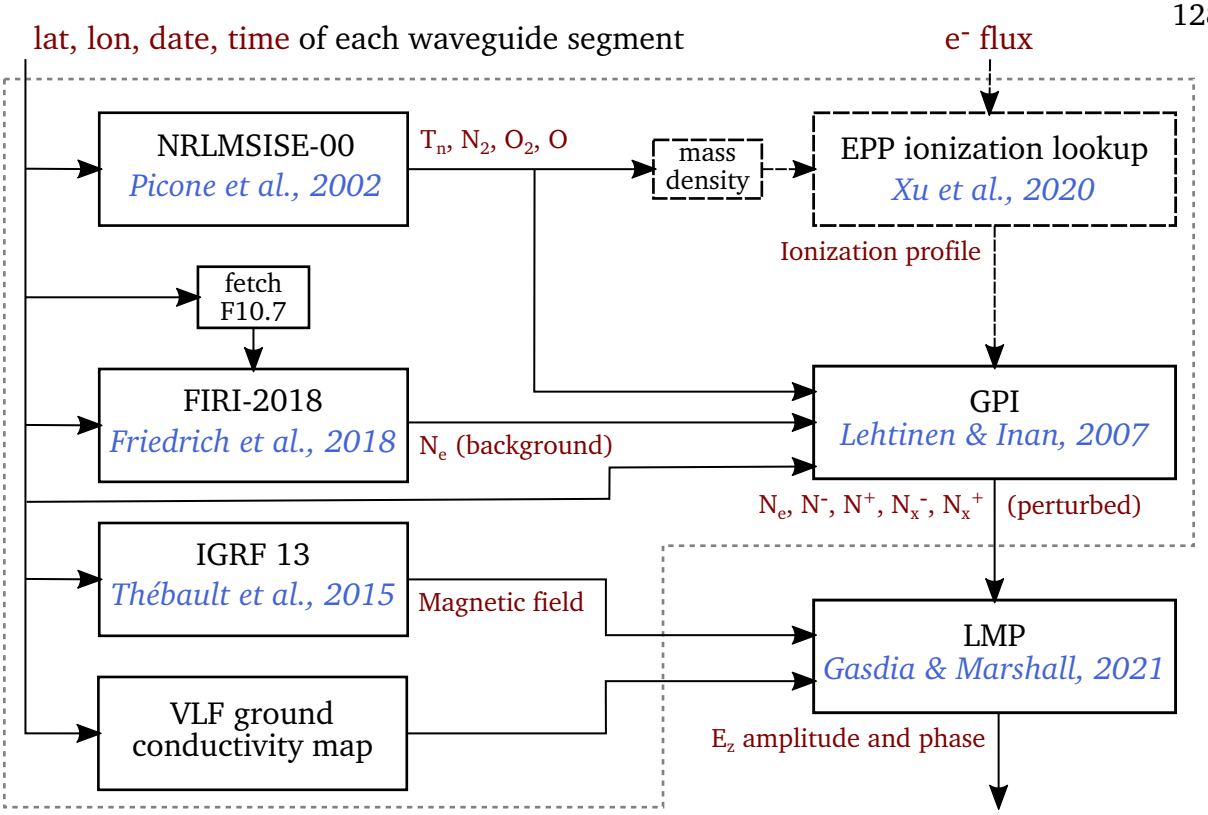


Figure 6.28: Steps used to generate the simulated receiver observations. The black dashed boxes are executed when there is precipitating electron flux. The GPI model uses a custom implementation and the FIRI and EPP ionization models are interpolated as necessary. MSIS, IGRF, and the F10.7 index is obtained from the SatelliteToolbox.jl package (Chagas et al., 2019).

An example of the background and perturbed charged profiles modeled by GPI is shown in Fig. 6.29 for an EPP flux of $10^5\,\mathrm{e}^-/\mathrm{cm}^2/\,\mathrm{sec}$. The perturbed electron density profile has a sharp shelf and is effectively 20 km lower in altitude than the background. Unlike the day and terminator scenarios, the simulated EPP observations include ions. The presence of ions has a very small effect on the observations, but increases the realism of the simulated observations.

A large EPP patch may have the appearance shown in Fig. 6.30. This patch of precipitating flux is centered on 55° N, 120° W, has a width of $\sigma_w = 3.5°$, height of $\sigma_h = 0.4°$, and a peak flux of $10^5\,\mathrm{e}^-/\mathrm{cm}^2/\mathrm{s}$. It is rotated by 1.5° from its line of latitude and has a Gaussian taper on all sides. As usual, each propagation path samples the precipitating flux from the map in Fig. 6.30 every 100 km for simulating observations using the process described above.

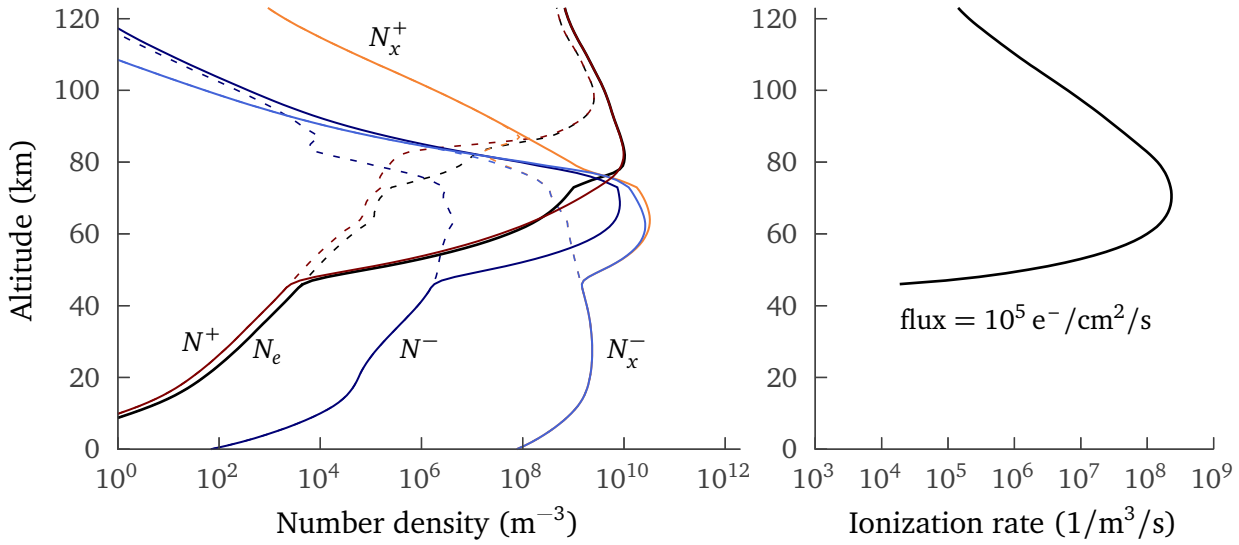


Figure 6.29: Background GPI charge profiles (dashed lines) and with $10^5$ $e^-/cm^2/$ sec EPP (solid lines) for the ionization rate profile on the right at 2020-03-01 0800 UTC, 55° N, 120° W.

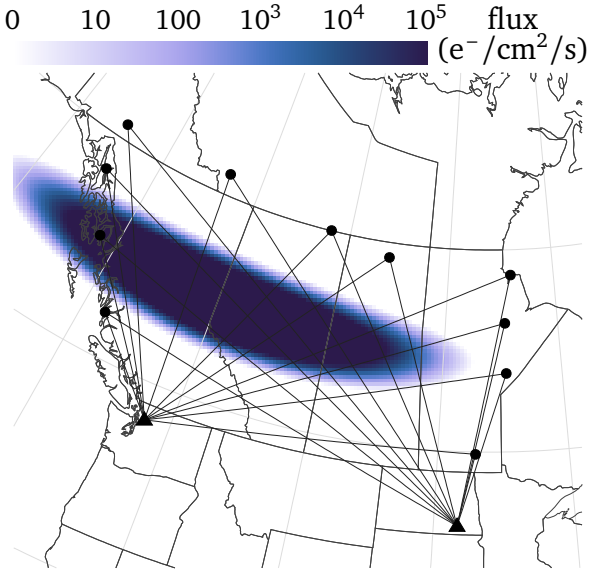


Figure 6.30: A simulated EPP patch with peak precipitating flux of $10^5$ $e^-/cm^2/s$ and Gaussian taper on all sides.

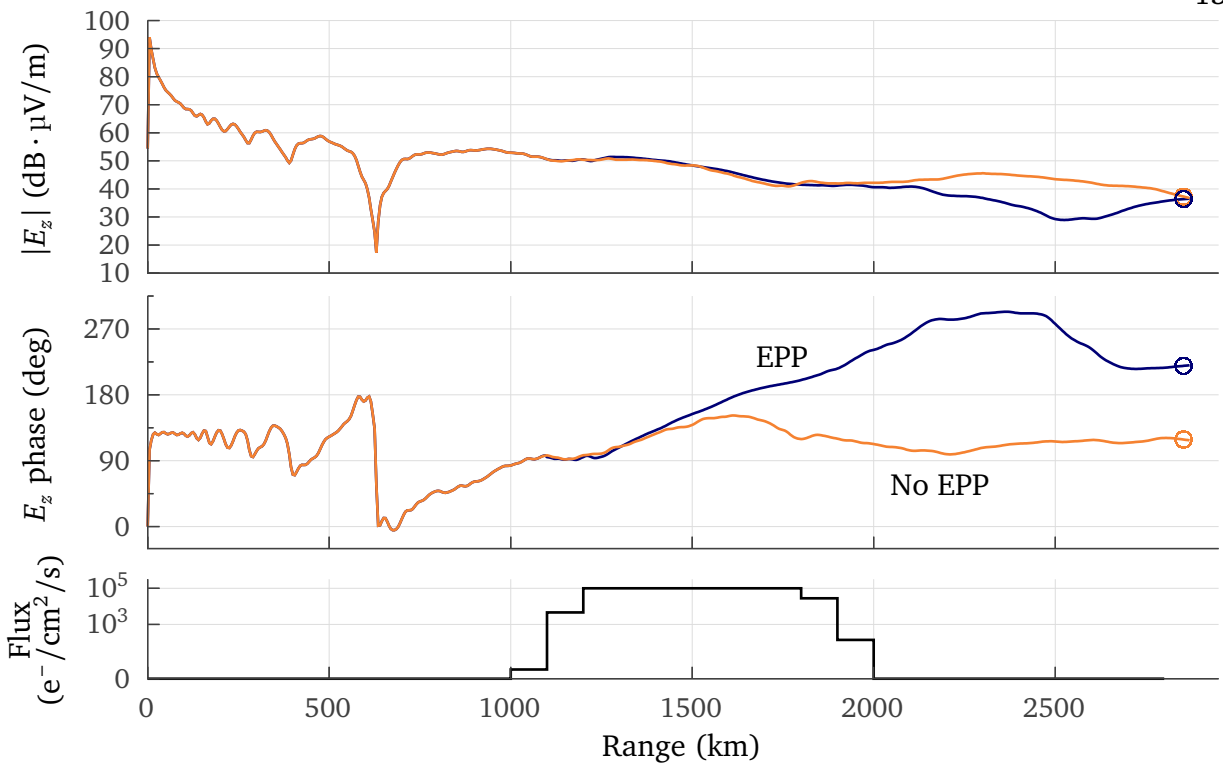


Figure 6.31: Amplitude and phase curves for the path from the NML transmitter in North Dakota to a receiver in Whitehorse, Yukon through the precipitation patch in Fig. 6.30 and through the realistic background ionosphere only. The region of peak flux occurs halfway along the propagation path, but interestingly the $E_z$ amplitude measured at the receiver (circles) is nearly identical for the EPP and no-EPP scenarios, while the phase difference is about 90°.

### 6.4.3 Analysis

A sample of the amplitude and phase curve from the NML transmitter in North Dakota to the receiver in Whitehorse, Yukon (the path also used in Figs. 5.2 and 5.3) is shown in Fig. 6.31 for both a realistic background ionosphere and the EPP patch shown in Fig. 6.30. The amplitude and phase curves both begin to diverge in the precipitating region, but the amplitude measured at the receiver is nearly identical between the EPP and no-EPP scenarios, while the phase has a 90° difference. If the receiver had been just a few hundred kilometers closer to the receiver, the amplitude could have had a difference of $\sim 10$ dB. This demonstrates why observing both amplitude and phase is useful for the estimation problem.

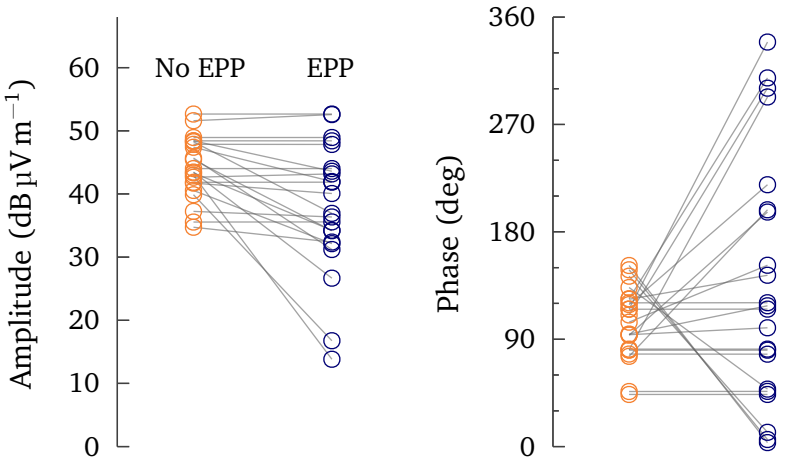


Figure 6.32: Observed amplitude and phase for each propagation path with a realistic background ionosphere and the EPP patch from Fig. 6.30. Lines connect the same path for both scenarios, showing that some do not pass through the precipitation region (horizontal lines) and others do (slanted lines).

The difference in amplitude and phase for all of the paths between EPP and no-EPP scenarios is shown in Fig. 6.32. There is considerably more spread in the observations through the perturbed ionosphere, and it is also clear that some paths do not pass through the precipitating patch while others do. The difference in observations across the network of propagation paths between the background and perturbed ionospheres is significant enough that it is unlikely for an estimation technique to confuse the two ionospheres—they have two very different minima on the cost function surface.

Figure 6.33 shows the mean absolute difference in amplitude and phase along a 3000 km propagation path between different Wait ionospheres, a realistic nighttime background ionosphere, and a $10^5$ $e^-/cm^2/s$ precipitating flux. Unlike the plots above, Fig. 6.33 uses a single homogeneous ionosphere along the path. Although that's unrealistic, it indicates the subionospheric VLF best-fit Wait profile to represent the background and perturbed ionospheres in a local segment of the waveguide.

Across the different combinations of $h'$ and $\beta$, there is a single global minimum for each of the scenarios. In fact, only the nighttime background ionosphere has a plausible second

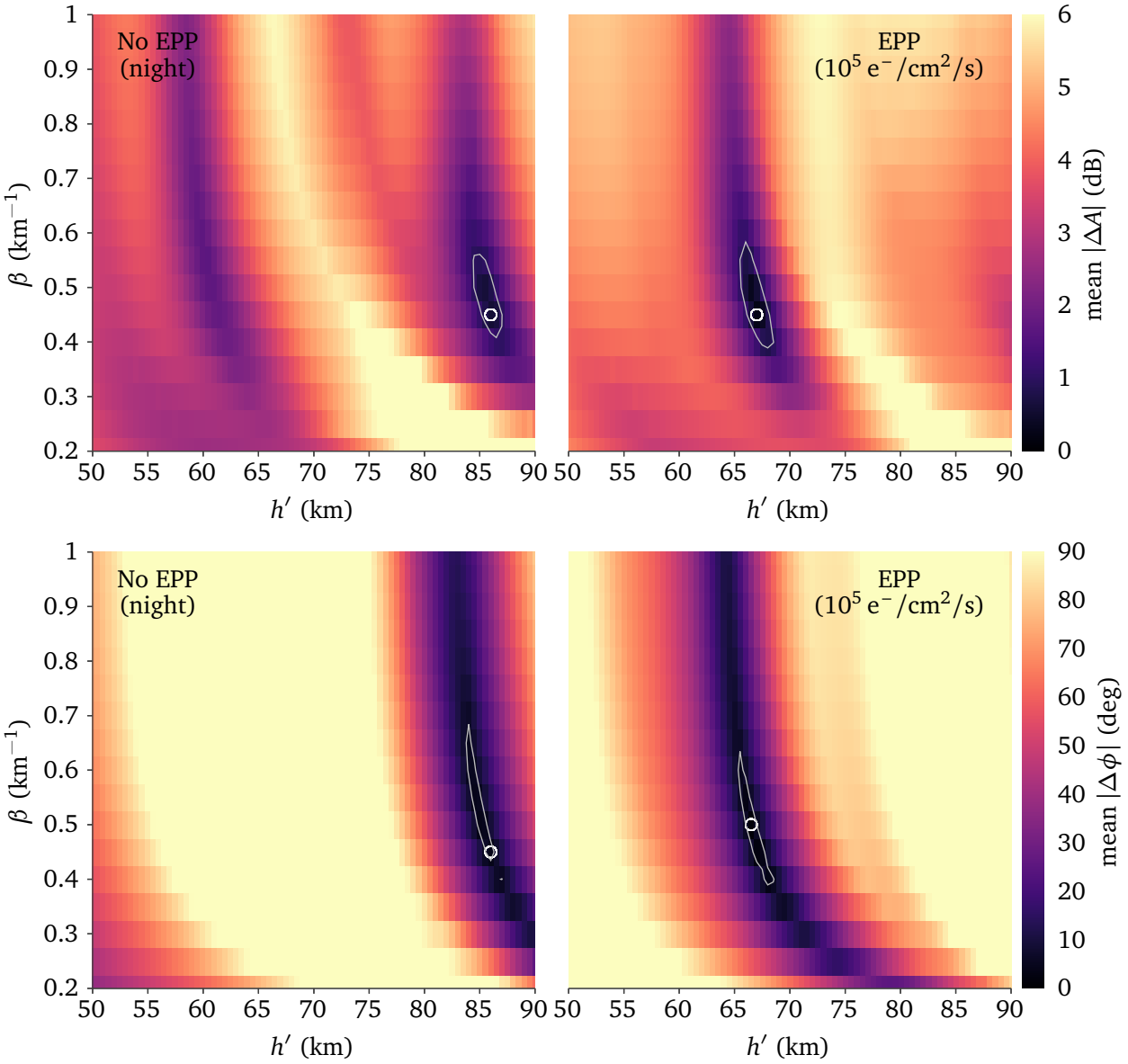


Figure 6.33: Mean absolute deviation for amplitude (top) and phase (bottom) between different Wait profiles and a realistic background ionosphere (left) and an EPP-perturbed ionosphere (right) measured over a 3000 km path with "receivers" every 5 km. The ionosphere profiles are homogeneous along the path, which has an ocean-like ground conductivity and 50 µT vertical magnetic field. Thin gray contour lines mark 1 dB error amplitude and 6° for phase.

local minimum at low $h'$. The amplitude and phase fits are approximately the same, which is why it is probably only necessary to measure one of them across the receiver array. The $\beta$ fit for both scenarios does not change from about $0.45\,\mathrm{km}^{-1}$. In fact, there is poor sensitivity to high $\beta$ values that begins as low as $0.5\,\mathrm{km}^{-1}$, much lower than typically assumed. However, $h'$ decreases by nearly 20 km. This is so large that it becomes difficult to have intuition for how the observation at the receiver should change compared to the background ionosphere. In Fig. 4.6 we showed how the amplitude and phase responds to changes in $h'$ over $\Delta 10\,\mathrm{km}$, and although that trend continues to lower $h'$, no local estimation method would be able to move from an unperturbed prior estimate to this perturbed ionosphere. As $h'$ is decreased from the background best-fit $h'$, the measurement residuals increase before decreasing again. On a single homogeneous path, a 90° change in phase can be obtained for a decrease in $h'$ of only 4 km, yet for those receivers with a $\sim 90°$ phase change in Fig. 6.32, the cause is a localized EPP patch with a much larger change in effective $h'$.

It is interesting to compare these subionospheric VLF best-fit Wait profiles to the perturbed electron density profiles for different precipitating fluxes. Figure 6.34 compares the electron density profiles under both realistic daytime and nighttime background ionospheres and with precipitating flux from $10^3$ to $10^6\,\mathrm{e}^-/\mathrm{cm}^2/\mathrm{s}$. The best-fit Wait profiles are determined using the brute force mean absolute deviation grid of Fig. 6.33. Although there is a slight increase in $\beta$ as the precipitating flux increases, it may be difficult in practice to measure this given the poor sensitivity of these observations to high $\beta$. In general the $\beta$ value is consistently $\sim 0.5\,\mathrm{km}^{-1}$ across the different scenarios. This also fits a typical nighttime ionosphere, but is high for daytime. However, there is a significant decrease in $h'$ as the precipitating flux increases. The daytime scenario with a flux of $10^6\,\mathrm{e}^-/\mathrm{cm}^2/\mathrm{s}$ has an $h'$ fit of 56 km! A heatmap of the fit for this scenario is shown in Fig. 6.35. It is also important to recognize that these fits only capture a portion of the lower shelf of the perturbed profiles and do not capture the true electron density profile.

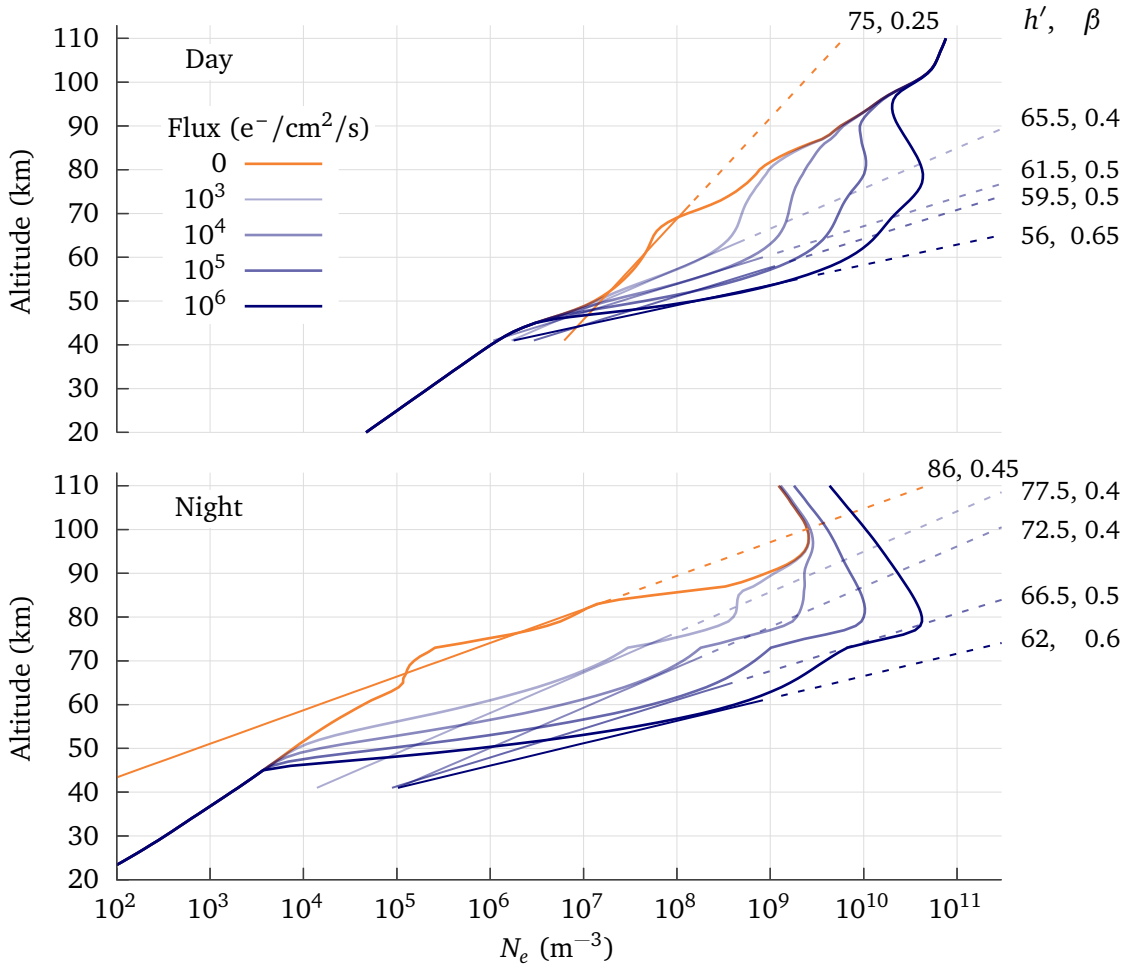


Figure 6.34: Best fit Wait profiles (straight lines) from phase only to realistic background and EPP-perturbed ionospheres for day (top) and night (bottom). The Wait profiles are dashed above the approximate reflection height at which $\omega_r = \omega$.

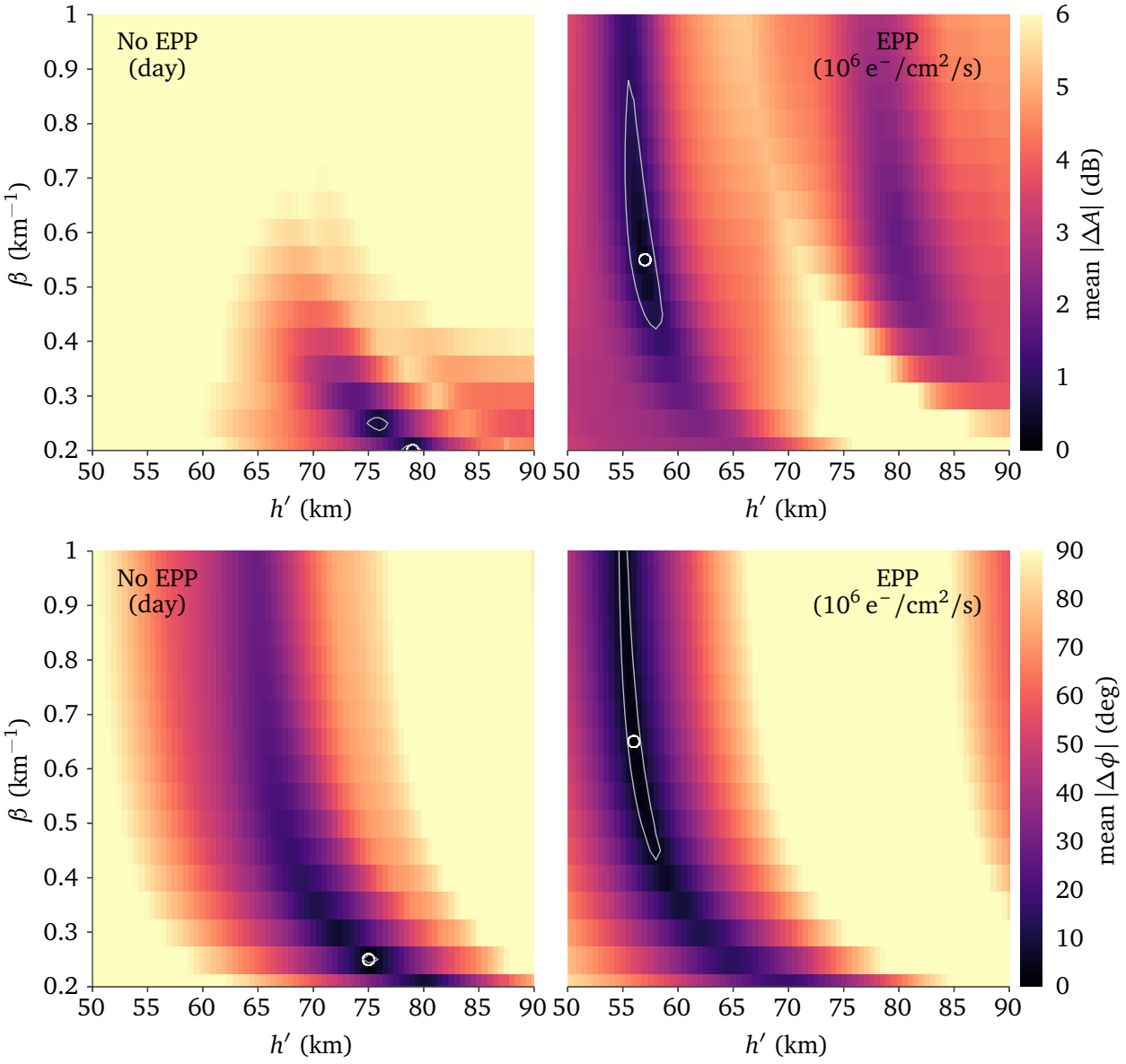


Figure 6.35: Wait profile fits for a daytime background and EPP flux of $10^6\,\mathrm{e}^-/\mathrm{cm}^2/\mathrm{s}$. The best-fit $\beta$ is similar to Fig. 6.33 for the EPP scenario. Although $\beta = 0.2\,\mathrm{km}^{-1}$ was fit for the background ionosphere using amplitude only, a second region of less than 1 dB error is identified by the contour line at $0.25\,\mathrm{km}^{-1}$. The true best-fit is probably between these values but the simulations were run only every $0.05\,\mathrm{km}^{-1}$ in $\beta$.

### 6.4.4 Comments

Both of the methods detailed in Chapter 5 fail to identify the precipitation patch of Fig. 6.30 (see Fig. 6.36). The LETKF diverges, even when using a prior with a mean shifted down to $h' = 70\,\text{km}$ and $\sigma_{h'}$ increased to 2.2 km for latitudes between 50° N and 60° N. The true perturbed ionosphere for moderate to high precipitating flux is far outside of the prior ensemble for the expected (unperturbed) ionosphere—up to about $10\sigma$ from the mean $h'$. In an ensemble Kalman filter, the nonlinear relationship between the states and observations is not captured when the truth is outside of the ensemble. At first it may seem that increasing $\sigma_{h'}$ would solve the problem, but in practice the filter continues to diverge. Over a wide range of states the likelihood function is highly non-Gaussian and the computed Kalman gain does not generate a valid update.

We have previously demonstrated that the LETKF can converge to scenarios having local geographic regions with irregular exponential electron density profiles similar to an EPP patch (Gasdia & Marshall, 2018), but the filter prior was prompted to expect the perturbed region. This previous work only modeled changes in $h'$ of $-10\,\text{km}$, representing a low precipitating flux, and did not have to contend with additional error from the use of a realistic (as opposed to exponential) truth profile. One way the LETKF might be used to estimate EPP is to use a prior that includes the EPP. This information must come from external measurements and to some extent this removes the need for imaging EPP with VLF in the first place.

The multiscale optimization approach fails because the truth is simply too far from the prior ionosphere on the cost function surface and COBYLA is a *local* optimization method. The resulting ionosphere estimate looks like a typical nighttime ionosphere (Fig. 6.36) with no indication that precipitation is present. Moving the states to the correct perturbed values incurs a large model cost. COBYLA gets stuck in a local minimum near the initial estimate and never tries to change the states as much as necessary because both the model and data costs increase as the cost function surface is traversed between the initial estimate and the correct states. On

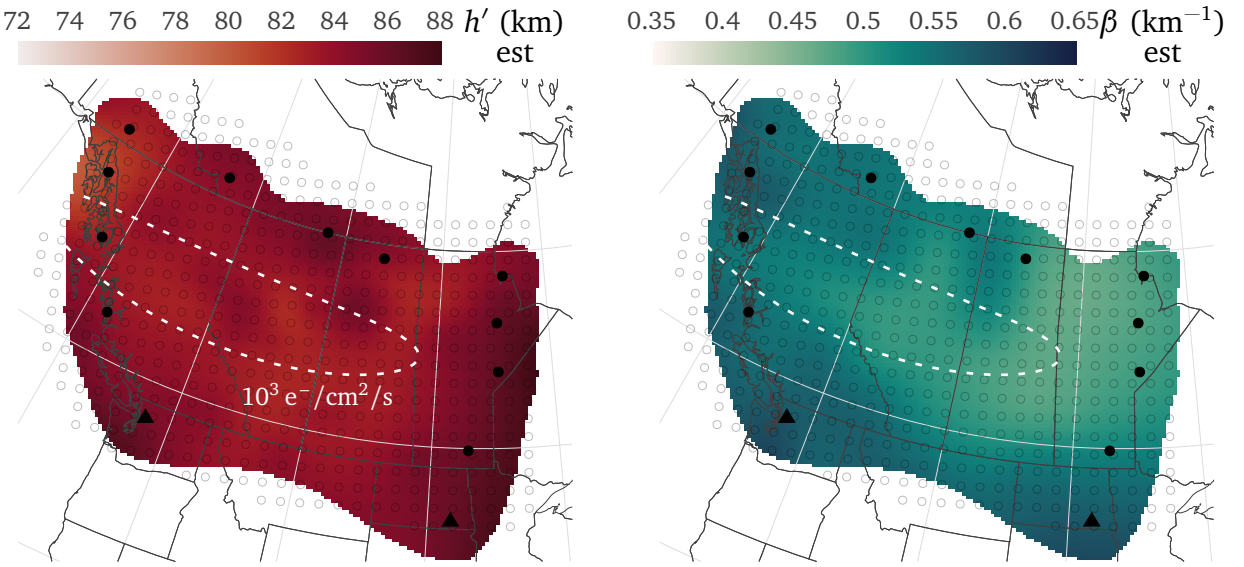


Figure 6.36: Final COBYLA multiscale estimate of the EPP ionosphere from Fig. 6.30. The method fails to capture the precipitation and moves very little from the Ferguson prior.

the upside, the failure of the method is obvious by inspecting the measurement residuals, which are very high (Fig. 6.37). This indicates that COBYLA was not able to move the estimate into a low minima of the cost function. A second factor contributing to the method's failure is that the precipitation region is not captured by the control points at the largest spatial scale. However, the method fails even when additional iterations are performed at shorter spatial scales.

Several additional approaches specific to the EPP problem have been attempted but remain active research. No matter the method, it seems most important that it explicitly test strongly perturbed states. The problem is more difficult than imaging an unperturbed ionosphere because the problem space is much larger. The challenge is to identify a method that searches the space efficiently. There are at least three ways to do this. As already mentioned, it may be possible to use external information to specify a prior that already captures a precipitation patch or other disturbance. This is the least interesting approach because it doesn't leverage the continuous, wide coverage monitoring capability of a VLF network. A second approach is to use a global cost function minimization algorithm. Unlike local algorithms, such as COBYLA, global algorithms

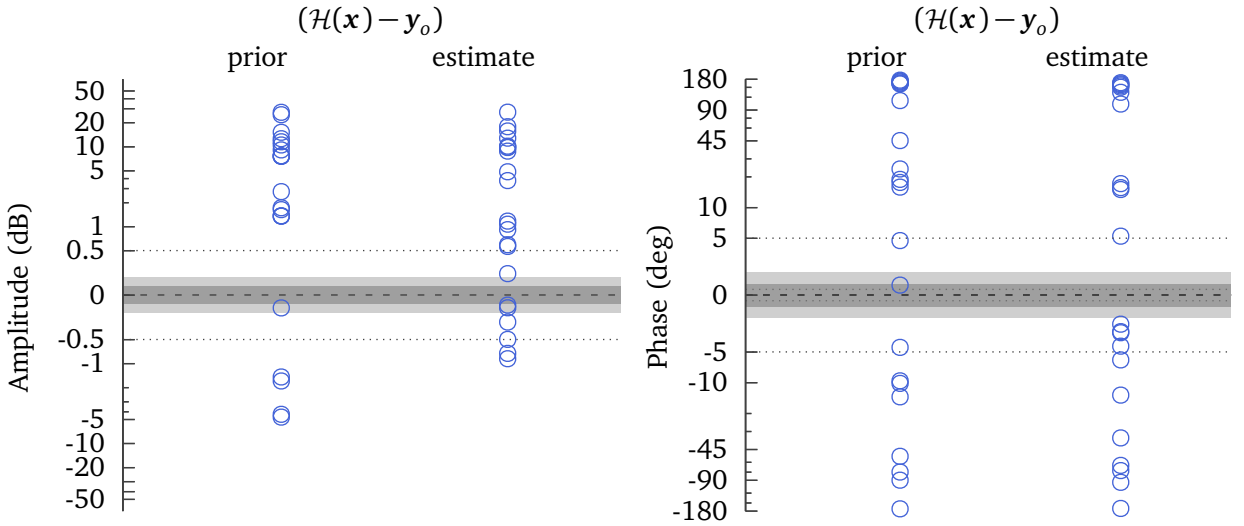


Figure 6.37: Measurement residuals of the COBYLA multiscale estimate of the EPP scenario. The method did little to improve the residuals so if this were real data and not a simulation, it would be obvious that the estimate is not correct.

compute the cost function value across the domain of states. Unfortunately, these methods require a great number of forward model evaluations as they traverse the search space. Finally, a third approach is to avoid blindly estimating a strongly perturbed ionosphere. Instead, make incremental estimates as the perturbation grows. If we begin with a good estimate of the unperturbed ionosphere, it is feasible that the full LETKF algorithm, incorporating both the forecast and measurement update steps, could be used to make estimates of the perturbation growing over time. To do this, one would need to carefully design a forecast model that captures the localized changes in the ionosphere that are associated with precipitation.

# Chapter 7

## Conclusions and future work

This dissertation has presented methods for imaging the D-region ionosphere using a network of VLF receivers, explored the response of VLF observations to different conditions in the Earth-ionosphere waveguide, and developed tools for quickly simulating observations with realistic ionospheric profiles. This chapter will summarize the conclusions of the previous chapters and provide suggestions for future work.

Chapter 2 provided background on the D-region ionosphere and VLF propagation, and introduced a tool for interpolating FIRI profiles. It also introduced profile results from our implementation of the GPI model. Chapter 3, coupled with Appendix A, is a comprehensive introduction to mode theory, with a focus on implementation, suitable for first year graduate students and others new to VLF propagation.

The Longwave Mode Propagator model is discussed and validated in Chapter 4. The model is more robust than LWPC, easier to understand because it solves the physical equations, and contains fewer lines of code. The project is open source on GitHub, thoroughly documented, has tagged versions and a test suite, yet there are a list of improvements still to be made. Here are a few:

- The code can continue to be simplified, particularly replacing some of the harder-to-understand functions containing algebraic solutions with simpler methods that may be made fast by low level (single instruction, multiple data: SIMD) vectorization.
- It may be possible to replace modified Hankel functions of order one third with the

integrated wavefields for the height gains used to compute the fields at the receiver. The integrated height gain functions are more accurate than the modified Hankel functions because they do not assume vacuum between ground and the eigenangle reference height.

- With the previous point implemented, the program could be made capable of computing the field in the ionosphere up to satellite heights. Similar extensions to LWPC predecessor codes have been made by others (Pappert, 1973; Pappert & Hitney, 1981).

- Numerical issues can arise due to the large imaginary part of the eigenangle at extremely low transmitter frequencies. ELF support can be improved by explicitly including flat Earth implementations of the algorithms and switching to them when necessary.

- Because the magnetic field is usually measured directly, the program should compute the transmitted magnetic field components at the receiver.

However, the highest priority improvement that should be made to the Longwave Mode Propagator is to reduce its runtime, which is dominated by computation of the ionosphere reflection in the mode finder. Morfitt and Shellman (1976) and Shellman (1986) designed the mode finder in LWPC to use interpolated values of the ionosphere reflection coefficient. Although this reduces the compute time (gaining an order of magnitude in speed), it reduces the robustness of the mode finder (modes may be missed). There are numerous ways this could be improved: the reflection coefficient could be interpolated across the complex plane using a function approximator, e.g. expansion of the function for the reflection coefficient using a Chebyshev series; a neural network could be trained for a large number of EIWG scenarios and used to interpolate eigenangles or ionosphere reflection coefficients; or, a new complex root finder could be developed combining aspects of the LWPC mode finder with the Cauchy argument principle for robustness to complex poles without searching a dense grid. Ideally, the new method would be automatically differentiable. This would allow faster and more accurate

D-region estimation because a change in the observation could be explicitly related to a change of the ionosphere.

Chapter 4 also examined the amplitude and phase response to a number of changes in the Earth-ionosphere waveguide, including: $h'$ and $\beta$, transmitter frequency, collision frequency profile, magnetic field direction, and ground conductivity. These are classical results, but such comparison plots are not often published. An electrons-only ionosphere is compared to an ionosphere with multiple species and it is found that ions might produce a small bias to observations relative to the electrons-only assumption. Finally, the variability of observations at a fixed receiver due to changing the location of a perturbation in the ionosphere is demonstrated.

Chapter 5 and Chapter 6 discussed the D-region imaging problem and demonstrated the use of an ensemble Kalman filter and cost function minimization method to estimate typical daytime and terminator ionospheres using a VLF receiver array. Although the Kalman filter approach is more efficient in terms of forward model runs, the prior estimate of the ionosphere must be relatively close to the true ionosphere. The second approach minimized a cost function for the problem over large to small spatial scales and produced slightly more accurate ionosphere estimates than the Kalman filter using four times the number of forward model runs. Unfortunately, neither approach successfully estimated an ionosphere perturbed by energetic particle precipitation (EPP). Nonetheless, a process for simulating realistic VLF observations through an EPP perturbed ionosphere was developed and it was found that the Wait profile fit can have a decrease in $h'$ of more than 20 km compared to the background ionosphere. This is such a large change from the unperturbed ionosphere that successfully estimating the perturbation requires the estimation method to explicitly test for it.

Although not previously discussed, a number of methods were developed and applied to the precipitation scenario. So far, none produce results that we think would be both convincing and useful to characterize the precipitation. All of the methods attempt to roughly identify the precipitation region by testing the measurement residuals or full cost function of strongly perturbed ionospheres. A local method could then be used to produce a final estimate. The

number of combinations of ionospheres that could be tested assuming coarsely defined $h'$ from 54 to 90 km in 2 km steps, $\beta$ from 0.2 to 0.8 $km^{-1}$ in 0.05 $km^{-1}$ steps, and the 60 control points in the 300 km grid is infeasibly large. A successful estimation method must reduce this search space by leveraging knowledge of the ionosphere and the geometry of the propagation paths.

One of the more promising methods shaped a covariance matrix used to generate a random perturbation map. Control points localized by paths with higher residuals were assigned higher variance than paths with lower residuals. Because all paths add some positive uncertainty to the estimate, the variance is weighted by the density of paths near each control point. A Gaussian perturbation surface, similar to those used to generate the LETKF prior ionospheres, is generated and added to the prior ionosphere estimate. If the data cost $\Phi_d$ decreases, then this new ionosphere is accepted and the process is iterated until convergence. A second promising method recognizes two or three unperturbed local ionospheres and two or three perturbed local ionospheres that can exist at each control point. These are sequentially tested at each control point and accepted if the total cost function $J$ is less than the current value. Multiple passes over the grid are required to stabilize the estimate. One can think of new methods and variations all day.

It is important to point out that applying any of these estimation methods to real data will present additional challenges. There are several systematic errors that have not been simulated, including ground conductivity, terrain, and local Earth curvature. Nighttime estimation will likely be more difficult than daytime given the high temporal variability of the observations. If the structures creating this variability are small, it will be difficult to estimate them. An interesting issue common to inversion under all conditions is calibration of the transmitters. Both the measurements and the forward model must be calibrated to the transmitter amplitude and phase. This could be explicit, or maybe the estimation method could simultaneously estimate correction terms for each transmitter. How robust are the methods to a transmitter bias in phase and/or amplitude? That has yet to be explored.

This dissertation provides guidance and inspiration to those interested in VLF propagation

and the D-region ionosphere. Numerous simulated observation experiments, the results of only a small number shown here, demonstrate that two-dimensional estimates of the D-region using networks of VLF receivers are in the near future.

## Bibliography


Anagaw, A. Y., & Sacchi, M. D. (2012). Edge-preserving seismic imaging using the total variation method. *Journal of Geophysics and Engineering*, *9*(2), 138–146. https://doi.org/10/fzpq9f

Appleton, S. E., & Piggott, W. R. (1954). Ionospheric absorption measurements during a sunspot cycle. *Journal of Atmospheric and Terrestrial Physics*, *5*(1), 141–172. https://doi.org/10/crk862

Bainbridge, G., & Inan, U. S. (2003). Ionospheric D region electron density profiles derived from the measured interference pattern of VLF waveguide modes. *Radio Science*, *38*(4), 1077. https://doi.org/10/fkmfp9

Banerjea, B. K., & Saha, M. N. (1947). On the propagation of electromagnetic waves through the atmosphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *190*(1020), 67–81. https://doi.org/10/ftq5wr

Bartholomew-Biggs, M., Brown, S., Christianson, B., & Dixon, L. (2000). Automatic differentiation of algorithms. *Journal of Computational and Applied Mathematics*, *124*(1), 171–190. https://doi.org/10/b94pb4

Beharrell, M., & Honary, F. (2008). A new method for deducing the effective collision frequency profile in the D-region. *Journal of Geophysical Research: Space Physics*, *113*(A5). https://doi.org/10/dgbchp

Belrose, J. S., & Burke, M. J. (1964). Study of the lower ionosphere using partial reflection: 1. Experimental technique and method of analysis. *Journal of Geophysical Research (1896-1977)*, *69*(13), 2799–2818. https://doi.org/10/b926hk

Best, J. E., Ratcliffe, J. A., Wilkes, M. V., & Appleton, E. V. (1936). Experimental investigations of very long waves reflected from the ionosphere. *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences*, *156*(889), 614–633. https://doi.org/10/bxvppw

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10/f9wkpj

Bilitza, D., Altadill, D., Truhlik, V., Shubin, V., Galkin, I., Reinisch, B., & Huang, X. (2017). International reference ionosphere 2016: From ionospheric climate to real-time weather predictions. *Space Weather*, *15*(2), 418–429. https://doi.org/10/ggntx9

Bilitza, D., McKinnell, L.-A., Reinisch, B., & Fuller-Rowell, T. (2011). The international reference ionosphere today and in the future. *Journal of Geodesy*, *85*(12), 909–920. https://doi.org/10/fwzx5r

Blum, L. W., & Breneman, A. W. (2020). Observations of radiation belt losses due to cyclotron wave-particle interactions. *The dynamic loss of earth's radiation belts* (pp. 49–98). Elsevier. https://doi.org/10/gkbxdv

Booker, H. G., & Appleton, E. V. (1938). Propagation of wave-packets incident obliquely upon a stratified doubly refracting ionosphere. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *237*(781), 411–451. https://doi.org/10/c9rwkd

Booker, H. G., & Walkinshaw, W. (1946). The mode theory of tropospheric refraction and its relation to wave-guides and diffraction. *Meteorological factors in radio-wave propagation*, 80–127.

Budden, K. G. (1952). CXVIII. The propagation of a radio atmospheric—II. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *43*(346), 1179–1200. https://doi.org/10/ggntzb

Budden, K. G. (1955). The numerical solution of differential equations governing reflexion of long radio waves from the ionosphere. *Proc. R. Soc. Lond. A*, *227*(1171), 516–537. https://doi.org/10/b5jw4g

Budden, K. G. (1961a). *Radio waves in the ionosphere: The mathematical theory of the reflection of radio waves from stratified ionised layers*. Cambridge University Press.

Budden, K. G. (1961b). *The waveguide mode theory of wave propagation*. Logos Press.

Budden, K. G. (1962). The influence of the earth's magnetic field on radio propagation by wave-guide modes. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *265*(1323), 538–553. https://doi.org/10/bfmvdc

Budden, K. G. (1988). *The propagation of radio waves: The theory of radio waves of low power in the ionosphere and magnetosphere* (First paperback edition). Cambridge University Press.

Budden, K. G., & Martin, H. G. (1962). The ionosphere as a whispering gallery. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *265*(1323), 554–569. https://doi.org/10/b4m62t

Budhiraja, A., Chen, L., & Lee, C. (2007). A survey of numerical methods for nonlinear filtering problems. *Physica D: Nonlinear Phenomena*, *230*(1), 27–36. https://doi.org/10/dwjj8p

Burger, W., & Burge, M. J. (2016). *Digital image processing: An algorithmic introduction using Java*. Springer London. https://doi.org/10/ghngtm

Burke, J. (2019). DARPA positioning, navigation, and timing (PNT) technology and their impacts on GPS users. https://www.gps.gov/governance/advisory/meetings/2019-06/burke.pdf

Burns, S., Gasdia, F., Simpson, J. J., & Marshall, R. (2021). 3-D FDTD modeling of long-distance VLF propagation in the Earth-ionosphere waveguide. *IEEE Transactions on Antennas and Propagation*. https://doi.org/10/gk8jc9

Chagas, R. A. J., de Sousa, F. L., Louro, A. C., & dos Santos, W. G. (2019). Modeling and design of a multidisciplinary simulator of the concept of operations for space mission pre-phase A studies. *Concurrent Engineering*, *27*(1), 28–39. https://doi.org/10/gmd6qh

Chand, A. E., & Kumar, S. (2017). VLF modal interference distance and nighttime D region VLF reflection height for west-east and east-west propagation paths to Fiji. *Radio Science*, *52*(8), 1004–1015. https://doi.org/10/gbxt5b

Chapman, S. (1931). Bakerian lecture.—some phenomena of the upper atmosphere. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *132*(820), 353–374. https://doi.org/10/b5wg59

Cheng, Z., & Cummer, S. A. (2005). Broadband VLF measurements of lightning-induced ionospheric perturbations. *Geophysical Research Letters*, *32*(8), 1–4. https://doi.org/10/cp7k7w

Chevalier, M. W., & Inan, U. S. (2006). A technique for efficiently modeling long-path propagation for use in both FDFD and FDTD. *IEEE Antennas and Wireless Propagation Letters*, *5*, 525–528. https://doi.org/10/cxnzz4

Clemmow, P. C., & Heading, J. (1954). Coupled forms of the differential equations governing radio propagation in the ionosphere. *Mathematical Proceedings of the Cambridge Philosophical Society*, *50*(2), 319–333. https://doi.org/10/cprsd6

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*(368), 829–836. https://doi.org/10/gckfrx

Clilverd, M. A., Duthie, R., Rodger, C. J., Hardman, R. L., & Yearby, K. H. (2017a). Long-term climate change in the D-region. *Scientific Reports*, *7*(1), 16683. https://doi.org/10/gkbxdx

Clilverd, M. A., Rodger, C. J., McCarthy, M., Millan, R., Blum, L. W., Cobbett, N., Brundell, J. B., Danskin, D., & Halford, A. J. (2017b). Investigating energetic electron precipitation through combining ground-based and balloon observations. *Journal of Geophysical Research: Space Physics*, *122*(1), 534–546. https://doi.org/10/f9wb77

Clilverd, M. A., Rodger, C. J., Thomson, N. R., Brundell, J. B., Ulich, T., Lichtenberger, J., Cobbett, N., Collier, A. B., Menk, F. W., Seppälä, A., Verronen, P. T., & Turunen, E. (2009). Remote sensing space weather events: Antarctic-Arctic Radiation-belt (Dynamic) Deposition-VLF Atmospheric Research Konsortium network. *Space Weather*, *7*(4). https://doi.org/10/b7bh4n

Clilverd, M. A., Rodger, C. J., Thomson, N. R., Lichtenberger, J., Steinbach, P., Cannon, P., & Angling, M. J. (2001). Total solar eclipse effects on VLF signals: Observations and modeling. *Radio Science*, *36*(4), 773–788. https://doi.org/10/ft3x2r

Clilverd, M. A., Thomson, N. R., & Rodger, C. J. (1999). Sunrise effects on VLF signals propagating over a long north-south path. *Radio Science*, *34*(4), 939–948. https://doi.org/10/d3x74b

Cohen, M. B., Gross, N. C., Higginson-Rollins, M. A., Marshall, R. A., Gołkowski, M., Liles, W., Rodriguez, D., & Rockway, J. (2018). The lower ionospheric VLF/LF response to the 2017 Great American Solar Eclipse observed across the continent. *Geophysical Research Letters*, *45*(8), 3348–3355. https://doi.org/10/gdjzfk

Cohen, M. B., Inan, U. S., Golkowski, M., & Lehtinen, N. G. (2010a). On the generation of ELF/VLF waves for long-distance propagation via steerable HF heating of the lower ionosphere. *Journal of Geophysical Research. Space Physics*, *115*(A7). https://doi.org/10/bz764c

Cohen, M. B., Inan, U. S., & Paschal, E. W. (2010b). Sensitive broadband ELF/VLF radio reception with the AWESOME instrument. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(1), 3–17. https://doi.org/10/d9sbv8

Cohen, M. B., Lehtinen, N. G., & Inan, U. S. (2012). Models of ionospheric VLF absorption of powerful ground based transmitters. *Geophysical Research Letters*, *39*(24). https://doi.org/10/gkbxdz

Constable, S. C., Parker, R. L., & Constable, C. G. (1987). Occam's inversion; a practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics*, *52*(3), 289–300. https://doi.org/10/cmvht9

Courant, R., Lewy, H., & Friedrichs, K. (1928). Über die partiellen Differenzengleichungen der mathematischen Physik. *Mathematische Annalen*, *100*, 32–74.

Crain, C. M., & Tamarkin, P. (1961). A note on the cause of sudden ionization anomalies in regions remote from high-altitude nuclear bursts. *Journal of Geophysical Research*, *66*(1), 35–39. https://doi.org/10/ckkmdq

Crombie, D. D. (1958). Differences between the east-west and west-east propagation of VLF signals over long distances. *Journal of Atmospheric and Terrestrial Physics*, *12*(2), 110–117. https://doi.org/10/bmn9gd

Crombie, D. D. (1964). Periodic fading of VLF signals received over long paths during sunrise and sunset. *Journal of Research of the National Bureau of Standards, Section D: Radio Science*, *68D*(1), 27. https://doi.org/10/gkbxd3

Crombie, D. D. (1961). Reflection from a sharply bounded ionosphere for vlf propagation perpendicular to the magnetic meridian. *Journal of Research of the National Bureau of Standards, Section D: Radio Propagation*, *65D*(5), 455. https://doi.org/10/ggntzc

Cummer, S. A., Inan, U. S., & Bell, T. F. (1998). Ionospheric D region remote sensing using VLF radio atmospherics. *Radio Science*, *33*(6), 1781–1792. https://doi.org/10/fbgqs7

Cummer, S. A. (1997). *Lightning and ionospheric remote sensing using VLF/ELF radio atmospherics* (Ph.D. Dissertation). Stanford University. Stanford, CA.

Dierckx, P. (1993). *Curve and surface fitting with splines*. Oxford University Press. http://www.netlib.org/dierckx/

Evenden, G. I. (1990). *Cartographic projection procedures for the UNIX environment; a user's manual* (USGS Numbered Series No. 90-284). U.S. Geological Survey, https://doi.org/10/gk8jpw

Evensen, G. (2003). The Ensemble Kalman Filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, *53*(4), 343–367. https://doi.org/10/ccnmtb

Fan, Y., & van den Dool, H. (2004). Climate Prediction Center global monthly soil moisture data set at 0.5° resolution for 1948 to present. *Journal of Geophysical Research: Atmospheres*, *109*(D10). https://doi.org/10/dfzd4h

Farquharson, C. G., & Oldenburg, D. W. (2004). A comparison of automatic techniques for estimating the regularization parameter in non-linear inverse problems. *Geophysical Journal International*, *156*(3), 411–425. https://doi.org/10/c537nb

Feldt, R. (2018). BlackBoxOptim.jl. https://github.com/robertfeldt/BlackBoxOptim.jl

Ferguson, J. A. (1980). *Ionospheric profiles for predicting nighttime VLF/LF propagation* (tech. rep. NOSC/TR-530). Naval Ocean Systems Center. San Diego, CA.

Ferguson, J. A. (1998). *Computer programs for assessment of long-wavelength radio communications, version 2.0: User's guide and source files* (Technical Document No. 3030). Space and Naval Warfare Systems Center. San Diego, CA.

Ferguson, J. A., & Morfitt, D. G. (1981). *Wkb mode summing program for dipole antennas of arbitrary orientation and elevation for vlf/lf propagation* (NOSC/TR-697). Naval Ocean Systems Center. San Diego, CA.

Ferguson, J. A., & Snyder, F. P. (1980). *Approximate VLF/LF waveguide mode conversion model. computer applications: FASTMC and BUMP* (NOSC-TD-400). Naval Ocean Systems Center. San Diego, CA.

Ferguson, J. A., & Snyder, F. P. (1987). *The segmented waveguide program for long wavelength propagation calculations.* (NOSC/TD-1071). Naval Ocean Systems Center. San Diego, CA.

Ferguson, J. A. (1995). Ionospheric model validation at VLF and LF. *Radio Science*, *30*(3), 775–782. https://doi.org/10/bhrm67

Fertig, E. J., Hunt, B. R., Ott, E., & Szunyogh, I. (2007). Assimilating non-local observations with a local ensemble Kalman filter. *Tellus A: Dynamic Meteorology and Oceanography*, *59*(5), 719–730. https://doi.org/10/cbxbd6

Finlay, C. C., Kloss, C., Olsen, N., Hammer, M. D., Tøffner-Clausen, L., Grayver, A., & Kuvshinov, A. (2020). The CHAOS-7 geomagnetic field model and observed changes in the South Atlantic Anomaly. *Earth, Planets and Space*, *72*(1), 1–31. https://doi.org/10/gj2v5h

Friedrich, M., Pock, C., & Torkar, K. (2018). FIRI-2018, an updated empirical model of the lower ionosphere. *Journal of Geophysical Research: Space Physics*, *123*(8), 6737–6751. https://doi.org/10/gfd8tk

Friedrich, M., & Rapp, M. (2009). News from the lower ionosphere: A review of recent developments. *Surveys in Geophysics*, *30*(6), 525–559. https://doi.org/10/dkgsd6

Friedrich, M., & Torkar, K. M. (2001). FIRI: A semiempirical model of the lower ionosphere. *Journal of Geophysical Research: Space Physics*, *106*(A10), 21409–21418. https://doi.org/10/cnwg9z

Friedrich, M., Torkar, K. M., Lehmacher, G. A., Croskey, C. L., Mitchell, J. D., Kudeki, E., & Milla, M. (2006). Rocket and incoherent scatter radar common-volume electron measurements of the equatorial lower ionosphere. *Geophysical Research Letters*, *33*(8). https://doi.org/10/c942z7

Gablonsky, J. M., & Kelley, C. T. (2001). A locally-biased form of the DIRECT algorithm. *Journal of Global Optimization*, *21*(1), 27–37. https://doi.org/10/ffnqns

Galejs, J. (1972). *Terrestrial propagation of long electromagnetic waves*. Pergamon Press.

Galuzzi, B. G., Perego, R., Candelieri, A., & Archetti, F. (2018). Bayesian optimization for full waveform inversion. In P. Daniele & L. Scrimali (Eds.), *New Trends in Emerging Complex Real Life Problems: ODS, Taormina, Italy, September 10–13, 2018* (pp. 257–264). Springer International Publishing. https://doi.org/10/gk8jp3

Gasdia, F. (2020). Fgasdia/rootsandpoles.jl: V1.4.0. https://doi.org/10/gkq5z3

Gasdia, F., & Marshall, R. A. (2018). Imaging D-region enhancement from radiation belt precipitation events. (SM43D-3614).

Gasdia, F., & Marshall, R. A. (2019). Assimilating VLF transmitter observations with an LETKF for spatial estimates of the D-region ionosphere. *IEEE Transactions on Geoscience and Remote Sensing*, 1–18. https://doi.org/10/gkbxd4

Gasdia, F., & Marshall, R. A. (2021). A new longwave mode propagator for the Earth-ionosphere waveguide. *IEEE Transactions on Antennas and Propagation*, 1–14. https://doi.org/10/gkq5z4

Gaspari, G., & Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, *125*(554), 723–757. https://doi.org/10/fb7p3z

Gebhard, L. A. (1979). *Evolution of naval radio-electronics and contributions of the Naval Research Laboratory (revision)* (NRL 7600). Naval Research Laboratory. Washington, D.C.

Glukhov, V. S., Pasko, V. P., & Inan, U. S. (1992). Relaxation of transient lower ionospheric disturbances caused by lightning-whistler-induced electron precipitation bursts. *Journal of Geophysical Research: Space Physics*, *97*(A11), 16971–16979. https://doi.org/10/d7fcth

Graf, K. L., Spasojevic, M., Marshall, R. A., Lehtinen, N. G., Foust, F. R., & Inan, U. S. (2013). Extended lateral heating of the nighttime ionosphere by ground–based VLF transmitters. *Journal of Geophysical Research: Space Physics*, *118*(12), 7783–7797.

Gross, N. C., & Cohen, M. B. (2020). VLF remote sensing of the D region ionosphere using neural networks. *Journal of Geophysical Research: Space Physics*, *125*(1). https://doi.org/10/gkbxd5

Gross, N. C., Cohen, M. B., Said, R. K., & Gołkowski, M. (2018). Polarization of narrowband VLF transmitter signals as an ionospheric diagnostic. *Journal of Geophysical Research: Space Physics*, *123*(1), 901–917. https://doi.org/10/gc5gk8

Guha, A., De, B. K., Roy, R., & Choudhury, A. (2010). Response of the equatorial lower ionosphere to the total solar eclipse of 22 July 2009 during sunrise transition period studied using VLF signal. *Journal of Geophysical Research: Space Physics*, *115*(A11), A11302. https://doi.org/10/bc2sfj

Guitton, A., & Symes, W. W. (2003). Robust inversion of seismic data using the Huber norm. *Geophysics*, *68*(4), 1310–1319. https://doi.org/10/dpq3xk

Han, F., Cummer, S. A., Li, J., & Lu, G. (2011). Daytime ionospheric D region sharpness derived from VLF radio atmospherics. *Journal of Geophysical Research: Space Physics*, *116*(A5), A05314. https://doi.org/10/cnfgjv

Hansen, N., & Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, *9*(2), 159–195. https://doi.org/10/bv87qx

Hansen, P. (2016). Developments in high power longwave radio: A brief history. *2016 IEEE International Symposium on Antennas and Propagation (APSURSI)*, 1013–1014. https://doi.org/10/gmd6qj

Hansen, P. C. (1992). Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, *34*(4), 561–580. https://doi.org/10/c2dntc

Hansen, P. C., & O'Leary, D. P. (1993). The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing*, *14*(6), 1487–1503. https://doi.org/10/cpz8wt

Hartley, R., & Zisserman, A. (2004). Appendix 6: Iterative estimation methods. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Hoffimann, J. (2018). Geostats.jl – high-performance geostatistics in julia. *Journal of Open Source Software*, *3*(24), 692. https://doi.org/10/gk8jtk

Hollingworth, J., Marchant, E. W., Rayner, E. H., Eckersley, T. L., Taylor, J. E., Smith-Rose, R. L., Moullin, E. B., & Eckersley, P. P. (1926). Discussion on "The propagation of radio waves" before the Wireless Section, 3 February, 1926. *Journal of the Institution of Electrical Engineers*, *64*(353), 589–595. https://doi.org/10/gkbxd6

Huba, J. D., Joyce, G., & Fedder, J. A. (2000). Sami2 is Another Model of the Ionosphere (SAMI2): A new low-latitude ionosphere model. *Journal of Geophysical Research: Space Physics*, *105*(A10), 23035–23053. https://doi.org/10/bkxrsv

Hunt, B. R., Kostelich, E. J., & Szunyogh, I. (2007). Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, *230*(1), 112–126. https://doi.org/10/fnmn8x

Ieda, A. (2020). Ion-neutral collision frequencies for calculating ionospheric conductivity. *Journal of Geophysical Research: Space Physics*, *125*(2), e2019JA027128. https://doi.org/10/gm7h7z

Imhof, W. L., Reagan, J. B., Voss, H. D., Gaines, E. E., Datlowe, D. W., Mobilia, J., Helliwell, R. A., Inan, U. S., Katsufrakis, J., & Joiner, R. G. (1983). Direct observation of radiation belt electrons precipitated by the controlled injection of VLF signals from a ground-based transmitter. *Geophysical Research Letters*, *10*(4), 361–364. https://doi.org/10/ftzzb6

Inan, U. S., Cummer, S. A., & Marshall, R. A. (2010). A survey of ELF and VLF research on lightning-ionosphere interactions and causative discharges. *Journal of Geophysical Research: Space Physics*, *115*(A6), 1–21. https://doi.org/10/ctfk87

Inan, U. S. (1990). VLF heating of the lower ionosphere. *Geophysical Research Letters*, *17*(6), 729–732. https://doi.org/10/dvj3s2

Inan, U. S., Bell, T. F., Pasko, V. P., Sentman, D. D., Wescott, E. M., & Lyons, W. A. (1995). VLF signatures of ionospheric disturbances associated with sprites. *Geophysical Research Letters*, *22*(24), 3461–3464.

Ingber, L. (1989). Very fast simulated re-annealing. *Mathematical and Computer Modelling*, *12*(8), 967–973. https://doi.org/10/dxmc8z

Jacobsen, T. A., & Friedrich, M. (1979). Electron density measurements in the lower D-region. *Journal of Atmospheric and Terrestrial Physics*, *41*(12), 1195–1200. https://doi.org/10/bs84v2

Jacobson, A. R., Shao, X.-M., & Holzworth, R. (2009). Full-wave reflection of lightning long-wave radio pulses from the ionospheric D region: Numerical model. *Journal of Geophysical Research: Space Physics*, *114*(A3). https://doi.org/10/cgtjb5

Johler, J. R., & Walters, L. C. (1960). On the theory of reflection of low- and very-low-radiofrequency waves from the ionosphere. *Journal of Research of the National Bureau of Standards, Section D: Radio Propagation*, *64D*(3), 269. https://doi.org/10/ggntzd

Johnson, S. G. (2021). The NLopt nonlinear-optimization package. http://github.com/stevengj/nlopt

Kaelo, P., & Ali, M. M. (2006). Some variants of the controlled random search algorithm for global optimization. *Journal of Optimization Theory and Applications*, *130*(2), 253–264. https://doi.org/10/cj45hz

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35–45. https://doi.org/10/dmftj3

Kelbert, A., Egbert, G. D., Kelbert, A., Peery, T., Clote, V., Fry, B., Erofeeva, S., staff of the National Geoelectromagnetic Facility, & their contractors. (2006). USArray TA magnetotelluric transfer functions. https://doi.org/10/gm7h73

Kelbert, A. (2019). The role of global/regional earth conductivity models in natural geomagnetic hazard mitigation. *Surveys in Geophysics*, *41*(1), 115–166. https://doi.org/10/gm22mg

Kelley, M. C. (2009). 1 introductory and background material. *The earth's ionosphere: Plasma physics and electrodynamics* (Second, pp. 1–26). Academic Press. https://doi.org/10/chw9tg

Koons, H. C., Edgar, B. C., & Vampola, A. L. (1980). *Precipitation of inner-zone electrons by whistler-mode waves from the VLF transmitters UMS and NWC* (TR-0081(6960-05)-6). The Aerospace Corporation. El Segundo, CA.

Kowalczyk, P. (2018). Global complex roots and poles finding algorithm based on phase analysis for propagation and radiation problems. *IEEE Transactions on Antennas and Propagation*, *66*(12), 7198–7205. https://doi.org/10/gfrdfc

Krige, D. G. (1951). *A statistical approach to some mine valuation and allied problems on the Witwatersrand* (Master's Thesis). University of the Witwatersrand. Johannesburg. Retrieved 2021, from http://wiredspace.wits.ac.za/handle/10539/17975

Laštovička, J., Akmaev, R. A., Beig, G., Bremer, J., & Emmert, J. T. (2006). Global change in the upper atmosphere. *Science*, *314*(5803), 1253–1254. https://doi.org/10/d432ps

Lay, E. H., & Shao, X.-M. (2011). Multi-station probing of thunderstorm-generated D-layer fluctuations by using time-domain lightning waveforms. *Geophysical Research Letters*, *38*(23). https://doi.org/10/b5dxbc

Lay, E. H., Shao, X.-M., & Jacobson, A. R. (2014). D region electron profiles observed with substantial spatial and temporal change near thunderstorms. *Journal of Geophysical Research: Space Physics*, *119*(6), 4916–4928. https://doi.org/10/gkbxd8

Lehtinen, N. G., & Inan, U. S. (2008). Radiation of ELF/VLF waves by harmonically varying currents into a stratified ionosphere with application to radiation by a modulated electrojet. *Journal of Geophysical Research: Space Physics*, *113*(A6). https://doi.org/10/bfmks4

Lehtinen, N. G., & Inan, U. S. (2007). Possible persistent ionization caused by giant blue jets. *Geophysical Research Letters*, *34*(8). https://doi.org/10/d26v3g

Lehtinen, N. G., & Inan, U. S. (2009). Full-wave modeling of transionospheric propagation of VLF waves. *Geophysical Research Letters*, *36*(3), L03104. https://doi.org/10/bmgjvm

Marsh, D. R., Mills, M. J., Kinnison, D. E., Lamarque, J.-F., Calvo, N., & Polvani, L. M. (2013). Climate change from 1850 to 2005 simulated in CESM1(WACCM). *Journal of Climate*, *26*(19), 7372–7391. https://doi.org/10/f3cvzk

Marshall, R. A. (2012). An improved model of the lightning electromagnetic field interaction with the D-region ionosphere. *Journal of Geophysical Research: Space Physics*, *117*(A3). https://doi.org/10/fxzrkc

Marshall, R. A., Inan, U. S., & Chevalier, T. W. (2008). Early VLF perturbations caused by lightning EMP-driven dissociative attachment. *Geophysical Research Letters*, *35*(21). https://doi.org/10/c4xhdx

Marshall, R. A., Inan, U. S., & Lyons, W. A. (2006). On the association of early/fast very low frequency perturbations with sprites and rare examples of VLF backscatter. *Journal of Geophysical Research: Atmospheres*, *111*(D19). https://doi.org/10/bgddx7

Marshall, R. A., & Snively, J. B. (2014). Very low frequency subionospheric remote sensing of thunderstorm-driven acoustic waves in the lower ionosphere. *Journal of Geophysical Research: Atmospheres*, *119*(9), 5037–5045. https://doi.org/10/f58qm9

Marshall, R. A., Wallace, T., & Turbe, M. (2017). Finite-difference modeling of very-low-frequency propagation in the earth-ionosphere waveguide. *IEEE Transactions on Antennas and Propagation*, *65*(12), 7185–7197. https://doi.org/10/gcqbm4

Marshall, R. A., & Cully, C. M. (2020). Atmospheric effects and signatures of high-energy electron precipitation. *The dynamic loss of earth's radiation belts* (pp. 199–255). Elsevier. https://doi.org/10/gkbxd9

Martin, H. G. (1965). The polarisation of low frequency radio waves in the terrestrial waveguide. *Journal of Atmospheric and Terrestrial Physics*, *27*(9), 995–1007. https://doi.org/10/cx9m76

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, *58*(8), 1246–1266. https://doi.org/10/fdsdjx

McCormick, J. C., & Cohen, M. B. (2018). D region ionospheric imaging using VLF/LF broadband sferics, forward modeling, and tomography, 4.

McRae, W. M., & Thomson, N. R. (2000). VLF phase and amplitude: Daytime ionospheric parameters. *Journal of Atmospheric and Solar-Terrestrial Physics*, *62*(7), 609–618. https://doi.org/10/dwdtv3

Mead, J. (2020). $\chi^2$ test for total variation regularization parameter selection. *Inverse Problems & Imaging*, *14*(3), 401–421. https://doi.org/10/gmbnnq

Mitra, A. P. (1978). The D-region of the ionosphere. *Endeavour*, *2*(1), 12–21. https://doi.org/10/b6xzmb

Miyoshi, Y., Oyama, S., Saito, S., Kurita, S., Fujiwara, H., Kataoka, R., Ebihara, Y., Kletzing, C., Reeves, G., Santolik, O., Clilverd, M., Rodger, C. J., Turunen, E., & Tsuchiya, F. (2015). Energetic electron precipitation associated with pulsating aurora: EISCAT and Van Allen Probe observations. *Journal of Geophysical Research: Space Physics*, *120*(4), 2754–2766. https://doi.org/10/gkbxfb

Modrak, R., & Tromp, J. (2016). Seismic waveform inversion best practices: Regional, global and exploration test cases. *Geophysical Journal International*, *206*(3), 1864–1889. https://doi.org/10/f859sr

Mogensen, G. T., Espinosa, H. G., & Thiel, D. V. (2014). Surface impedance mapping using sferics. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(4), 2074–2080. https://doi.org/10/gk8jp5

Morfitt, D. G. (1976). *Determination of effective ionospheric electron density profiles for VLF/LF propagation* (MEECN Technical Report DCA-C650-TP-76-4). Defense Communications Agency. Arlington, VA.

Morfitt, D. G. (1977). *Effective electron density distributions describing VLF/LF propagation data* (NOSC/TR-141). Naval Ocean Systems Center. San Diego, CA.

Morfitt, D. G., & Shellman, C. H. (1976). *'MODESRCH', an improved computer program for obtaining ELF/VLF/LF mode constants in an earth-ionosphere waveguide* (NELC/IR-77T). Naval Electronics Laboratory Center. San Diego, CA.

Morgan, R. R. (1968). *World-wide VLF effective-conductivity map* (Westinghouse Electric Corporation Report WEST-80133F-1). Westinghouse Electric Corporation.

Morris, P. B., Gupta, R. R., Warren, R. S., & Creamer, P. M. (1994). *Omega navigation system course book*. The Analytic Sciences Corp.

Mosegaard, K., & Tarantola, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research: Solid Earth*, *100*(B7), 12431–12447. https://doi.org/10/fwrdrn

Nagano, I., Yagitani, S., Miyamura, K., & Makino, S. (2003). Full-wave analysis of elves created by lightning-generated electromagnetic pulses. *Journal of Atmospheric and Solar-Terrestrial Physics*, *65*(5), 615–625. https://doi.org/10/fjtjdn

NaitAmor, S., Ghalila, H., & Cohen, M. B. (2017). TLEs and early VLF events: Simulating the important impact of transmitter-disturbance-receiver geometry. *Journal of Geophysical Research: Space Physics*, *122*(1), 792–801. https://doi.org/10/f9shtm

Naumann, U. (2019). Adjoint code design patterns. *ACM Transactions on Mathematical Software*, *45*(3), 1–32. https://doi.org/10/ggch8m

Naval Sea Systems Command (NAVSEA). (2021). Navigational positioning source using very low frequency signals. Retrieved 2021, from https://www.navysbir.com/n21_1/N211-052.htm

Navigation Innovation and Support Programme (NAVISP). (2020). 046 - combining ELF signals with GNSS for improved PNT. *European SPace Agency (ESA)*. https://navisp.esa.int/opportunity/details/68/show

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313. https://doi.org/10/bkqf

Nicolet, M., & Aikin, A. C. (1960). The formation of the D region of the ionosphere. *Journal of Geophysical Research*, *65*(5), 1469–1483. https://doi.org/10/bnm6bd

Pal, S., & Hobara, Y. (2016). Mid-latitude atmosphere and ionosphere connection as revealed by very low frequency signals. *Journal of Atmospheric and Solar-Terrestrial Physics*, *138-139*, 227–232. https://doi.org/10/f8bc5v

Pappert, R. A. (1981). *LF daytime earth ionosphere waveguide calculations* (NOSC/TR-647). Naval Ocean Systems Center. San Diego, CA.

Pappert, R. A., & Hitney, L. R. (1981). *A program to compute ELF/VLF Earth-ionosphere modal height gains via WKB methods up to satellite altitudes* (NOSC/TR-724). Naval Ocean Systems Center. San Diego, CA.

Pappert, R. A., Hitney, L. R., & Ferguson, J. A. (1983). *ELF/VLF (extremely low frequency/very low frequency) long path pulse program for antennas of arbitrary elevation and orientation.* (NOSC/TR-891). Naval Ocean Systems Center. San Diego, CA.

Pappert, R. A., & Morfitt, D. G. (1975). Theoretical and experimental sunrise mode conversion results at VLF. *Radio Science*, *10*(5), 537–546. https://doi.org/10/czvn82

Pappert, R. A., & Shockey, L. R. (1976). *Simplified VLF/LF mode conversion program with allowance for elevated, arbitrarily oriented electric dipole antennas* (Interim Report No. 771). Naval Electronics Laboratory Center. San Diego, CA.

Pappert, R. A., & Smith, R. R. (1972). Orthogonality of VLF height gains in the Earth ionosphere waveguide. *Radio Science*, *7*(2), 275–278. https://doi.org/10/b8v3k4

Pappert, R. A. (1968). A numerical study of VLF mode structure and polarization below an anisotropic ionosphere. *Radio Science*, *3*(3), 219–233. https://doi.org/10/ggntzg

Pappert, R. A. (1970). Effects of elevation and ground conductivity on horizontal dipole excitation of the earth-ionosphere waveguide. *Radio Science*, *5*(3), 579–590. https://doi.org/10/cm86r9

Pappert, R. A. (1973). Excitation of the earth-ionosphere waveguide by point dipoles at satellite heights. *Radio Science*, *8*(6), 535–545. https://doi.org/10/ftrcm3

Pappert, R. A., & Bickel, J. E. (1970). Vertical and horizontal VLF fields excited by dipoles of arbitrary orientation and elevation. *Radio Science*, *5*(12), 1445–1452. https://doi.org/10/cmn8v3

Pappert, R. A., & Ferguson, J. A. (1986). VLF/LF mode conversion model calculations for air to air transmissions in the earth-ionosphere waveguide. *Radio Science*, *21*(4), 551–558. https://doi.org/10/cdktn8

Pappert, R. A., Gossard, E. E., & Rothmuller, I. J. (1967). A numerical investigation of classical approximations used in VLF propagation. *Radio Science*, *2*(4), 387–400. https://doi.org/10/ggntzf

Pappert, R. A., & Shockey, L. R. (1971). *WKB mode summing program for VLF/ELF antennas of arbitrary length, shape and elevation* (NELC-IR-713, M402). Naval Electronics Lab Center. San Diego, CA.

Pappert, R. A., Shockey, L. R., & Moler, W. F. (1970). *A Fortran program for waveguide propagation which allows for both vertical and horizontal dipole excitation* (NELC-IR-702). Naval Electronics Lab Center. San Diego, CA.

Pappert, R. A., & Snyder, F. P. (1972). Some results of a mode-conversion program for VLF. *Radio Science*, *7*(10), 913–923. https://doi.org/10/dfdzh8

Pedersen, L. B., Persson, L., Bastani, M., & Byström, S. (2009). Airborne VLF measurements and mapping of ground conductivity in Sweden. *Journal of Applied Geophysics*, *67*(3), 250–258. https://doi.org/10/bjp7dq

Phanikumar, D. V., Maurya, A. K., Kumar, K. N., Venkatesham, K., Singh, R., Sharma, S., & Naja, M. (2018). Anomalous variations of VLF sub-ionospheric signal and mesospheric ozone prior to 2015 Gorkha Nepal earthquake. *Scientific Reports*, *8*(1), 1–9. https://doi.org/10/gdr9kd

Phelps, A. V., & Pack, J. L. (1959). Electron collision frequencies in nitrogen and in the lower ionosphere. *Physical Review Letters*, *3*(7), 340–342. https://doi.org/10/bhnq7c

Picone, J. M., Hedin, A. E., Drob, D. P., & Aikin, A. C. (2002). NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues. *Journal of Geophysical Research: Space Physics*, *107*(A12), 1–16. https://doi.org/10/ckftkq

Pitteway, M. L. V. (1965). The numerical calculation of wave-fields, reflexion coefficients and polarizations for long radio waves in the lower ionosphere. I. *Phil. Trans. R. Soc. Lond. A*, *257*(1079), 219–241. https://doi.org/10/fgdhvg

Plessix, R.-E. (2006). A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, *167*(2), 495–503. https://doi.org/10/bbmzk7

Poulsen, W. (1991). *Modeling of very low frequency wave propagation and scattering within the earth-ionosphere waveguide in the presence of lower ionospheric disturbances* (Dissertation). Stanford University.

Powell, M. J. D. (1994). A direct search optimization method that models the objective and constraint functions by linear interpolation. In S. Gomez & J.-P. Hennart (Eds.), *Advances in Optimization and Numerical Analysis* (pp. 51–67). Springer Netherlands. https://doi.org/10/gft3s8

Powell, M. J. D. (2007). A view of algorithms for optimization without derivatives. *Mathematics Today-Bulleting of the Institute of Mathematics and its Applications*, *43*(5).

Powell, M. J. D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives* (Technical Report NA2009/06). Department of Applied Mathematics and Theoretical Physics. Cambridge, England.

Qin, J., Pasko, V. P., McHarg, M. G., & Stenbaek-Nielsen, H. C. (2014). Plasma irregularities in the D-region ionosphere in association with sprite streamer initiation. *Nature Communications*, *5*, 3740. https://doi.org/10/f56s9f

Rackauckas, C., & Nie, Q. (2017). DifferentialEquations.jl – a performant and feature-rich ecosystem for solving differential equations in julia. *Journal of Open Research Software*, *5*. https://doi.org/10/ggfnfj

Ratcliffe, J. A. (1959). *The magneto-ionic theory & its applications to the ionosphere*. Cambridge University Press.

Reagan, J. B., Meyerott, R. E., Gunton, R. C., Imhof, W. L., Gaines, E. E., & Larsen, T. R. (1981). Modeling of the ambient and disturbed ionospheric media pertinent to elf/vlf propagation. *Medium, long and very long wave propagation (at frequencies less than 3000 kHz)*, *305*.

Richter, J. H. (1966). Application of conformal mapping to earth-flattening procedures in radio propagation problems. *Radio Science*, *1*(12), 1435–1438. https://doi.org/10/ggntzh

Ridley, A. J., Deng, Y., & Tóth, G. (2006). The global ionosphere–thermosphere model. *Journal of Atmospheric and Solar-Terrestrial Physics*, *68*(8), 839–864. https://doi.org/10/ftvkcz

Rogers, N., & Honary, F. (2014). D-region HF absorption models incorporating real-time riometer measurements. *2014 XXXIth URSI General Assembly and Scientific Symposium (URSI GASS)*, 1–2. https://doi.org/10/gk8jp6

Rowan, T. H. (1990). *Functional stability analysis of numerical algorithms* (Doctoral dissertation). University of Texas at Austin.

Rozhnoi, A., Solovieva, M., Levin, B., Hayakawa, M., & Fedun, V. (2014). Meteorological effects in the lower ionosphere as based on VLF/LF signal observations. *Natural Hazards and Earth System Sciences*, *14*(10), 2671–2679. https://doi.org/10/f6p5mx

Samanes, J. E., Raulin, J.-P., Macotela, E. L., & Day, W. R. G. (2015). Estimating the VLF modal interference distance using the South America VLF Network (SAVNET). *Radio Science*, *50*(2), 122–129. https://doi.org/10/f66thj

SAS Institute Inc. (2021). *SAS/IML(R) 9.3 User's Guide*. https://support.sas.com/documentation/cdl/en/imlug/64248/HTML/default/imlug_langref_sect208.htm

Schaller, R. R. (1997). Moore's law: Past, present and future. *IEEE Spectrum*, *34*(6), 52–59. https://doi.org/10/c6mrsz

Sheddy, C. H. (1968). A general analytic solution for reflection from a sharply bounded anisotropic ionosphere. *Radio Science*, *3*(8), 792–795. https://doi.org/10/ggntzk

Sheddy, C. H., Pappert, R., Gough, Y., & Moler, W. (1968). *A fortran program for mode constants in an earth-ionosphere waveguide* (Interim Report No. 683). Naval Electronics Laboratory Center. San Diego, CA.

Shellman, C. H. (1986). *A new version of MODESRCH using interpolated values of the magnetoionic reflection coefficients* (NOSC/TR-1143). Naval Ocean Systems Center. San Diego, CA.

Silber, I., & Price, C. (2017). On the use of VLF narrowband measurements to study the lower ionosphere and the mesosphere–lower thermosphere. *Surveys in Geophysics*, *38*(2), 407–441. https://doi.org/10/f9w56p

Simões, F., Pfaff, R., Berthelier, J.-J., & Klenzing, J. (2012). A review of low frequency electromagnetic wave phenomena related to tropospheric-ionospheric coupling mechanisms. *Space Science Reviews*, *168*(1-4), 551–593. https://doi.org/10/c72ddw

Simpson, J. J., & Taflove, A. (2007). A review of progress in FDTD Maxwell's equations modeling of impulsive subionospheric propagation below 300 kHz. *IEEE Transactions on Antennas and Propagation*, *55*(6), 1582–1590. https://doi.org/10/b4dpx4

Siskind, D. E., Zawdie, K. A., Sassi, F., Drob, D. P., & Friedrich, M. (2018). An intercomparison of VLF and sounding rocket techniques for measuring the daytime D region ionosphere: Theoretical implications. *Journal of Geophysical Research: Space Physics*, *123*(10), 8688–8697. https://doi.org/10/gfq5pm

Siskind, D. E., Mlynczak, M. G., Marshall, T., Friedrich, M., & Gumbel, J. (2015). Implications of odd oxygen observations by the TIMED/SABER instrument for lower D region ionospheric modeling. *Journal of Atmospheric and Solar-Terrestrial Physics*, *124*, 63–70. https://doi.org/10/f3s3z8

Siskind, D. E., Stevens, M. H., Englert, C. R., & Mlynczak, M. G. (2013). Comparison of a photochemical model with observations of mesospheric hydroxyl and ozone. *Journal of Geophysical Research: Atmospheres*, *118*(1), 195–207. https://doi.org/10/gkbxfd

Siskind, D. E., Zawdie, K. A., Sassi, F., Drob, D., & Friedrich, M. (2017). Global modeling of the low- and middle-latitude ionospheric D and lower E regions and implications for HF radio wave absorption. *Space Weather*, *15*(1), 115–130. https://doi.org/10/ggntzm

Skowron, J., & Gould, A. (2012). General complex polynomial root solver and its further optimization for binary microlenses. https://arxiv.org/abs/1203.1034

Smirnova, N. V., Ogloblina, O. F., & Vlaskov, V. A. (1988). Modelling of the lower ionosphere. *Pure and Applied Geophysics*, *127*(2), 353–379. https://doi.org/10/fhf7c5

Smith, L. G. (1965). *Langmuir probes for measurements in the ionosphere* (GCA Technical Report 65-25-N). GCA Corporation. Bedford, Massachusetts.

Snyder, J. P. (1987). *Map projections: A working manual* (USGS Numbered Series No. 1395). U.S. Government Printing Office. Washington, D.C. https://doi.org/10/gk8jp7

Space Weather Prediction Center. (2020). *D region absorption predictions (D-RAP)*. Retrieved 2020, from https://www.swpc.noaa.gov/products/d-region-absorption-predictions-d-rap

Spall, J. C. (1998). An overview of the simultaneous perturbation method for efficient optimization. *Johns Hopkins APL Technical Digest*, *19*(4), 11.

Spies, K. P., & Wait, J. R. (1961). *Mode calculations for VLF propagation in the earth-ionosphere waveguide* (Technical Note No. 114). U.S. National Bureau of Standards. Boulder, CO.

Storn, R., & Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*(4), 341–359. https://doi.org/10/d8n65h

Šulić, D. M., Srećković, V. A., & Mihajlov, A. A. (2016). A study of VLF signals variations associated with the changes of ionization level in the D-region in consequence of solar conditions. *Advances in Space Research*, *57*(4), 1029–1043. https://doi.org/10/f8bkbw

Tanaka, Y. T., Raulin, J.-P., Bertoni, F. C. P., Fagundes, P. R., Chau, J., Schuch, N. J., Hayakawa, M., Hobara, Y., Terasawa, T., & Takahashi, T. (2010). First very low frequency detection of short repeated bursts from magnetar SGR J1550–5418. *The Astrophysical Journal*, *721*(1), L24–L27. https://doi.org/10/d6c669

Tarantola, A. (2006). Popper, Bayes and the inverse problem. *Nature Physics*, *2*(8), 492–494. https://doi.org/10/dzpkzm

The Staff of the Computation Library. (1945). Introduction. *Tables of the modified Hankel function of order one-third and of their derivatives* (p. 33). Harvard University Press.

Thébault, E., Finlay, C. C., Beggan, C. D., Alken, P., Aubert, J., Barrois, O., Bertrand, F., Bondar, T., Boness, A., Brocco, L., Canet, E., Chambodut, A., Chulliat, A., Coïsson, P., Civet, F., Du, A., Fournier, A., Fratter, I., Gillet, N., … Zvereva, T. (2015). International Geomagnetic Reference Field: The 12th generation. *Earth, Planets and Space*, *67*(1), 1–19. https://doi.org/10/f7fxhg

Thomson, N. R. (1993). Experimental daytime VLF ionospheric parameters. *Journal of Atmospheric and Terrestrial Physics*, *55*(2), 173–184. https://doi.org/10/b5zpqw

Thomson, N. R. (2010). Daytime tropical D region parameters from short path VLF phase and amplitude. *Journal of Geophysical Research: Space Physics*, *115*(A9). https://doi.org/10/fdbn7w

Thomson, N. R., Clilverd, M. A., & McRae, W. M. (2007). Nighttime ionospheric D region parameters from VLF phase and amplitude. *Journal of Geophysical Research: Space Physics*, *112*(A7), 1–14. https://doi.org/10/bq4q8n

Thomson, N. R., Clilverd, M. A., & Rodger, C. J. (2017). Midlatitude ionospheric D region: Height, sharpness, and solar zenith angle. *Journal of Geophysical Research: Space Physics*, *122*(8), 8933–8946. https://doi.org/10/gb2h79

Torkar, K. M., & Friedrich, M. (1983). Tests of an ion-chemical model of the D- and lower E-region. *Journal of Atmospheric and Terrestrial Physics*, *45*(6), 369–385. https://doi.org/10/ff9tfc

Tromp, J. (2020). Seismic wavefield imaging of Earth's interior across scales. *Nature Reviews Earth & Environment*, *1*(1), 40–53. https://doi.org/10/ggr3qq

Turunen, E., Verronen, P. T., Seppälä, A., Rodger, C. J., Clilverd, M. A., Tamminen, J., Enell, C.-F., & Ulich, T. (2009). Impact of different energies of precipitating particles on NOx generation in the middle and upper atmosphere during geomagnetic storms. *Journal of Atmospheric and Solar-Terrestrial Physics*, *71*(10), 1176–1189. https://doi.org/10/ckdrvh

Van Allen, J. A. (1959). The geomagnetically trapped corpuscular radiation. *Journal of Geophysical Research (1896-1977)*, *64*(11), 1683–1689. https://doi.org/10/bb2z89

Verner, J. H. (2010). Numerically optimal runge–kutta pairs with interpolants. *Numerical Algorithms*, *53*(2-3), 383–396. https://doi.org/10/b669sf

Verronen, P. T., Andersson, M. E., Marsh, D. R., Kovács, T., & Plane, J. M. C. (2016). WACCM-D—Whole Atmosphere Community Climate Model with D-region ion chemistry. *Journal of Advances in Modeling Earth Systems*, *8*(2), 954–975. https://doi.org/10/f8v5wx

Verronen, P. T., Seppälä, A., Clilverd, M. A., Rodger, C. J., Kyrölä, E., Enell, C.-F., Ulich, T., & Turunen, E. (2005). Diurnal variation of ozone depletion during the October–November 2003 solar proton events. *Journal of Geophysical Research: Space Physics*, *110*(A9). https://doi.org/10/fgw638

Vogel, C. R., & Oman, M. E. (1998). Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE Transactions on Image Processing*, *7*(6), 813–824. https://doi.org/10/fkfq3p

Volland, H. (1961). Comparison between mode theory and ray theory of VLF propagation. *Journal of Research of the National Bureau of Standards, Section D: Radio Propagation*, *65D*(4), 357. https://doi.org/10/gkbxff

Volland, H. (1995). *Handbook of atmospheric electrodynamics* (First). CRC Press. https://doi.org/10/gmbnnr

Wait, J. R. (1957). The mode theory of VLF ionospheric propagation for finite ground conductivity. *Proceedings of the IRE*, *45*(6), 760–767. https://doi.org/10/dcr43p

Wait, J. R. (1958). A study of VLF field strength data: Both old and new. *Geofisica pura e applicata*, *41*(1), 73–85. https://doi.org/10/chrc2p

Wait, J. R. (1963). Review of mode theory of radio propagation in terrestrial waveguides. *Reviews of Geophysics*, *1*(4), 481–505. https://doi.org/10/fqj3kb

Wait, J. R. (1970a). *Electromagnetic waves in stratified media*. Elsevier. https://doi.org/10/ggntzp

Wait, J. R. (1970b). Factorization method applied to electromagnetic wave propagation in a curved waveguide with nonuniform walls. *Radio Science*, *5*(7), 1059–1068. https://doi.org/10/chfjqz

Wait, J. R., & Spies, K. P. (1964). *Characteristics of the earth-ionosphere waveguide for VLF radio waves* (Technical Note No. 300). U.S. National Bureau of Standards. Boulder, CO.

Walker, D. (1965). Phase steps and amplitude fading of VLF signals at dawn and dusk. *Journal of Research of the National Bureau of Standards, Section D: Radio Science*, *69D*(11), 1435. https://doi.org/10.6028/jres.069D.155

Walsh, E. J. (1967). Full-wave solutions in terms of coupled vacuum modes. *Radio Science*, *2*(8), 913–925. https://doi.org/10/gkbxfg

Watson, G. N., & Nicholson, J. W. (1918). The diffraction of electric waves by the earth. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *95*(666), 83–99. https://doi.org/10/bhjsgk

Watson, G. N., & Nicholson, J. W. (1919). The transmission of electric waves round the earth. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *95*(673), 546–563. https://doi.org/10/fv8mhp

Watt, A. D. (1967). *VLF radio engineering* (First, Vol. 14). Pergamon Press.

Whittaker, I. C., Gamble, R. J., Rodger, C. J., Clilverd, M. A., & Sauvaud, J.-A. (2013). Determining the spectra of radiation belt electron losses: Fitting DEMETER electron flux observations for typical and storm times. *Journal of Geophysical Research: Space Physics*, *118*(12), 7611–7623. https://doi.org/10/gkbxfj

Xu, W., Marshall, R. A., Kero, A., Turunen, E., Drob, D., Sojka, J., & Rice, D. (2019). VLF measurements and modeling of the D-region response to the 2017 total solar eclipse. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(10), 7613–7622. https://doi.org/10/gkq5z5

Xu, W., Marshall, R. A., Tyssøy, H. N., & Fang, X. (2020). A generalized method for calculating atmospheric ionization by energetic electron precipitation. *Journal of Geophysical Research: Space Physics*, *125*(11), e2020JA028482. https://doi.org/10/gmd6qk

Yabroff, I. W. (1957). Reflection at a sharply-bounded ionosphere. *Proceedings of the IRE*, *45*(6), 750–753. https://doi.org/10/brsrft

Zmuda, A. J., Shaw, B. W., & Haave, C. R. (1963). Very low frequency disturbances and the high-altitude nuclear explosion of July 9, 1962. *Journal of Geophysical Research*, *68*(3), 745–758. https://doi.org/10/fmgmcb

Zmuda, A. J., Shaw, B. W., & Haave, C. R. (1964). VLF disturbances caused by trapped beta-rays from the decay of neutrons produced in high-altitude nuclear explosions. *Journal of Research of the National Bureau of Standards, Section D: Radio Science*, *68D*(1), 117–123.

# Appendix A

# Mode theory

This appendix provides additional background on mode theory for the Earth-ionosphere waveguide as implemented in the Longwave Mode Propagator (LMP). It supports Chapter 3, which walks through the components of mode theory at a higher level. This appendix begins by deriving the wave equation in Appendix A.1. With the solution to the wave equation, we formalize reflection in Appendix A.2. Appendix A.3 then introduces general waveguide concepts. Appendix A.4 is specific to the ionosphere and includes derivations of the ionosphere reflection coefficient and the Booker quartic. Appendix A.5 is devoted entirely to a derivation of the ionosphere susceptibility tensor, $\boldsymbol{M}$, and how we modify it to account for Earth curvature. The next section, Appendix A.6, derives the Fresnel reflection equations used to describe the ground reflection coefficients. Appendix A.7 provides a high level description of the mode theory provided by Budden (1962), as well as practical formulations for height gain functions and excitation factors introduced over a series of papers by Pappert (Pappert, 1968; 1970; Pappert & Bickel, 1970; Pappert & Shockey, 1971; Pappert & Shockey, 1976). Finally, Appendix A.8 presents a high level description of mode conversion between horizontally homogeneous segments of waveguide.

## A.1 Wave equation

In a homogeneous, isotropic medium

$$\boldsymbol{D} = \epsilon_0 \mathrm{n}^2 \boldsymbol{E}. \tag{A.1}$$

Assuming a plane wave is normal to the $z$-axis,

$$\frac{\partial}{\partial x} = 0, \qquad \frac{\partial}{\partial y} = 0 \tag{A.2}$$

and Maxwell's equations (Eq. (3.8)) become

$$\frac{\partial E_y}{\partial z} = ik\mathscr{H}_x, \qquad \frac{\partial \mathscr{H}_y}{\partial z} = -ik\epsilon_0^{-1} D_x \tag{A.3}$$

$$\frac{\partial E_x}{\partial z} = -ik\mathscr{H}_y, \qquad \frac{\partial \mathscr{H}_x}{\partial z} = ik\epsilon_0^{-1} D_y \tag{A.4}$$

$$\mathscr{H}_z = 0, \qquad D_z = 0 \tag{A.5}$$

Substituting Eq. (A.1) into Eqs. (A.3) and (A.4) results in two sets of equations involving only $E_y$ and $\mathscr{H}_x$ and two sets involving only $E_x$ and $\mathscr{H}_y$. These sets are independent.

Working with the set

$$\frac{\partial E_x}{\partial z} = -ik\mathscr{H}_y, \qquad \frac{\partial \mathscr{H}_y}{\partial z} = -ik\mathfrak{n}^2 E_x \tag{A.6}$$

we can eliminate $\mathscr{H}_y$

$$\begin{aligned} \frac{\partial^2 E_x}{\partial z^2} &= -ik\frac{\partial \mathscr{H}_y}{\partial z} \\ &= -ik(-ik\mathfrak{n}^2 E_x) \\ \frac{\partial^2 E_x}{\partial z^2} + k^2\mathfrak{n}^2 E_x &= 0 \end{aligned} \tag{A.7}$$

Equation (A.7) is known as the *wave equation*. Two independent solutions of Eq. (A.7) are

$$E_x = E_{x,1}\exp(-ik\mathfrak{n}z) \tag{A.8}$$

$$E_x = E_{x,2}\exp(+ik\mathfrak{n}z) \tag{A.9}$$

which are two waves travelling in the positive and negative $z$ directions, respectively.

## A.2 Defining reflection coefficients

Reflection coefficients express the ratio of the reflected to incident amplitude of an electromagnetic wave upon reflection from an impedance discontinuity. A model of reflection is

necessary to describe propagation of waves in a waveguide as they reflect off the waveguide boundaries. In Appendices A.2 and A.3 the model is geometric and the following sections detail electromagnetic wave propagation and reflection in the ionosphere.

We assume a homogeneous plane wave in the $x$–$z$ plane, which we refer to as the plane of incidence. As stated in Chapter 3, all fields are assumed to have an implicit harmonic time variation.

### A.2.1 Vertical incidence

Assume a wave (field component) $F$ is launched vertically from the ground at $z = 0$ into free space above. The field can be expressed by $F_1 \exp(-ikz)$ as it travels vertically upward. At some height, the wave is reflected from a sharp boundary and begins traveling downward as $F_2 \exp(ikz)$. As we saw in Appendix A.1, the exponential is a solution to the wave equation. This scenario is graphically depicted in Fig. A.1. $F_1$ and $F_2$ are generally complex-valued amplitudes of the waves.

The reflection coefficient of the boundary is equal to the ratio of the downgoing wave to the upgoing wave. Measured at the ground, this ratio is

$$R_0 = \frac{F_2 \exp(ikz)}{F_1 \exp(-ikz)} = \frac{F_2 \exp(0)}{F_1 \exp(0)} = \frac{F_2}{F_1}. \tag{A.10}$$

If instead the fields are measured at some height $z_1$, the reflection coefficient $R_1$ is simply

$$R_1 = \frac{F_2 \exp(ikz_1)}{F_1 \exp(-ikz_1)} = \frac{F_2}{F_1} \exp(2ikz_1) = R_0 \exp(2ikz_1). \tag{A.11}$$

The rule for referring the reflection height at one altitude to another is then (Budden, 1961b, p. 107)

$$R_2 = R_0 \exp(2ikz_2) = R_1 \exp\big(2ik(z_2 - z_1)\big). \tag{A.12}$$

Interestingly, even if the height at which the reflection coefficient is referred is within a medium of conductivity which varies with height, it is calculated in the same way—*as though this level were in free space* (Budden, 1988, p. 296). This allows us to reference reflection coefficients

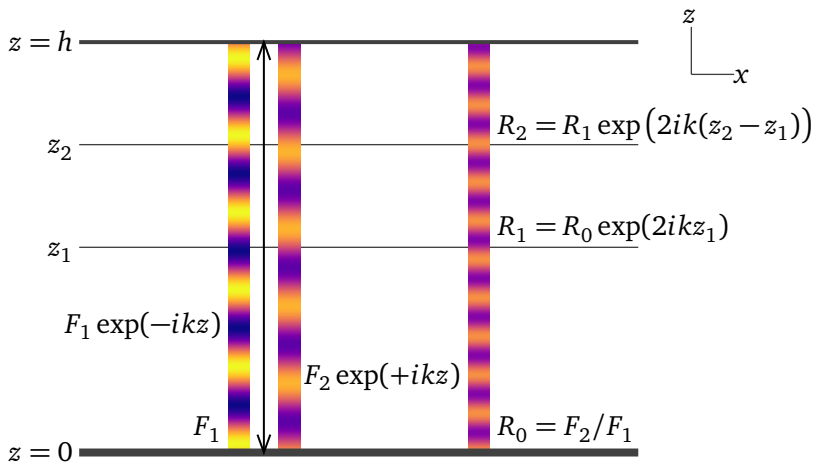


Figure A.1: A wave is launched vertically with complex amplitude $F_1$ into free space. The wave is reflected from an imperfect boundary at height $h$. The reflection coefficient $R$ is the ratio of downgoing to upgoing fields and can be referenced to any height in the guide.

to different heights for an ionosphere in which the electron density $N$ varies gradually with height, even though the fields are not truly described by $\exp(-ikz)$ and $\exp(ikz)$.

### A.2.2 Oblique incidence

Now assume the wavefront is obliquely incident on the ionosphere boundary at an angle $\theta$ measured clockwise from the $z$ direction, as depicted in Fig. A.2. The upgoing wave is now expressed as $F_1 \exp(-ik(Sx + Cz))$ and the downgoing reflected wave is $F_2 \exp(-ik(Sx - Cz))$ where $C = \cos\theta$ and $S = \sin\theta$.

Both the incident and reflected wave fields depend on $x$ through $\exp(-ikSx)$ (also see Appendix A.2.4). The reflection coefficient is defined such that both upgoing and downgoing wave fields must be measured at the same point $(x, z_1)$ in space. The reflection coefficient at height $z_1$ is

$$R_1 = \frac{F_2}{F_1} \exp(2ikCz_1) \tag{A.13}$$

so that $R_1$ is effectively independent of $x$.

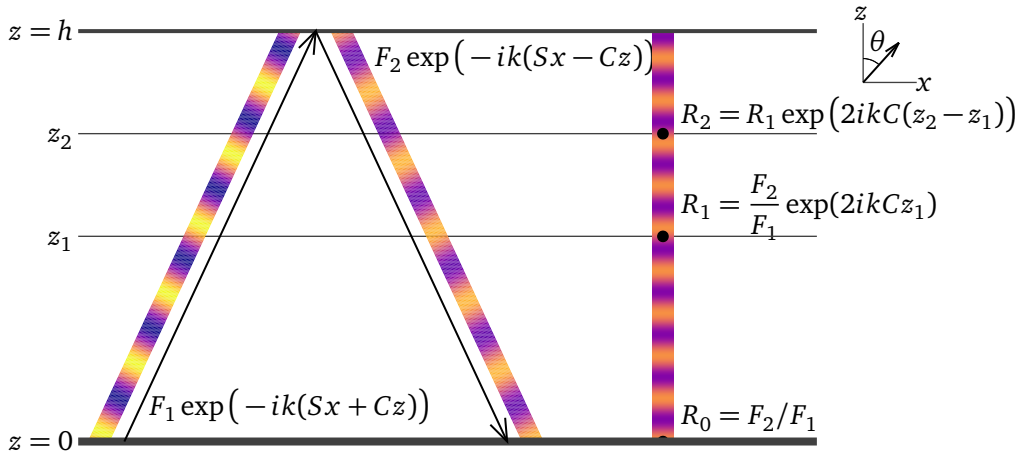


Figure A.2: A wave is launched vertically from $z = 0$ at an oblique angle to a sharp boundary at height $z = h$. A reflected wave travels downwards. Both waves have the same dependence on $x$ such that the ratio of the wavefields, i.e. the reflection coefficient $R$, effectively only depends on $z$.

As before, we can refer the reflection coefficient from one height to another through

$$R_2 = R_1 \exp\big(2ikC(z_2 - z_1)\big). \tag{A.14}$$

### A.2.3 Anisotropic media and elliptical polarization

When at least one boundary is anisotropic, energy is exchanged between field components in and perpendicular to the plane of incidence. Thus, the incident and reflected waves are generally elliptically polarized, but can be resolved into two linearly polarized components that are oriented with their electric vectors parallel and perpendicular to the plane of incidence (the $x$–$z$ plane). It then becomes convenient to introduce four reflection coefficients that describe the ratio of each reflected component to each incident component.

We use the notation ${}_aR_b$ where subscript $a$ is the incident component and subscript $b$ is the reflected component. Subscript symbol "$\parallel$" indicates the electric field is parallel to the plane of incidence and "$\perp$" indicates the electric field is perpendicular to the plane of incidence.[1]

[1] Budden (1988) suggests that the $1, 2$ notation should be preferred in modern works, but I find the explicit $\parallel, \perp$ notation more meaningful.

The total reflection matrix is

$$\boldsymbol{R} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} = \begin{pmatrix} {}_{\parallel}R_{\parallel} & {}_{\perp}R_{\parallel} \\ {}_{\parallel}R_{\perp} & {}_{\perp}R_{\perp} \end{pmatrix}. \tag{A.15}$$

In free space below the ionosphere,[2]

$$(E_{\parallel}, E_{\perp}) = (\mathscr{H}_y, E_y) \tag{A.16}$$

as depicted in Fig. A.3. Therefore, we will often use $\mathscr{H}_y$ instead of the electric field in the plane of incidence. Letting $E_y^I = \mathscr{H}_x^I = 0$,

$${}_{\parallel}R_{\parallel} = \mathscr{H}_y^R / \mathscr{H}_y^I, \quad {}_{\parallel}R_{\perp} = E_y^R / \mathscr{H}_y^I \tag{A.17}$$

and letting $E_x^I = \mathscr{H}_y^I = 0$,

$${}_{\perp}R_{\parallel} = \mathscr{H}_y^R / E_y^I, \quad {}_{\perp}R_{\perp} = E_y^R / E_y^I \tag{A.18}$$

where superscript $R$ and $I$ represent the reflected and incident fields, respectively.

Following Budden (1988, p. 299) we express the fields as

$$\boldsymbol{e}^I = \begin{pmatrix} \mathscr{H}_y^I \\ E_y^I \end{pmatrix}, \quad \boldsymbol{e}^R = \begin{pmatrix} \mathscr{H}_y^R \\ E_y^R \end{pmatrix} \tag{A.19}$$

so that the reflected components are related to the incident components through

$$\boldsymbol{e}^R = \boldsymbol{R}\boldsymbol{e}^I. \tag{A.20}$$

After two successive reflections from boundaries with reflection coefficients $\boldsymbol{R}_a$ and $\boldsymbol{R}_b$, the fields are (Budden, 1962, pp. 541–542)

$$\boldsymbol{R}_b\boldsymbol{R}_a\boldsymbol{e}^I. \tag{A.21}$$

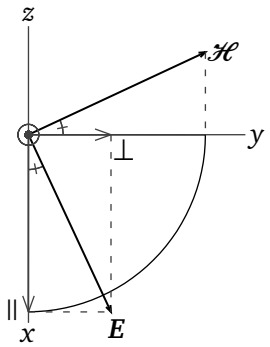


Figure A.3: Geometric relationship between the $\boldsymbol{E}$ and $\mathscr{H}$ fields with respect to the $x$, $y$ and $\parallel$, $\perp$ axes.

[2] These relations only hold in a vacuum. The relation between fields in an anisotropic plasma is complicated (Walsh, 1967).

### A.2.4 Reflection from stratified media

The real ionosphere is not a sharp boundary with free space below it, but rather a region with continuously varying refractive index as a function of height $z$. Reflection of radio waves in this medium occurs over a range of heights as the wave is gradually bent.

Assume the index of refraction of the medium can be expressed as $\mathfrak{n}(z)$. We can break the medium up into infinitely thin homogeneous slices of refractive index $\mathfrak{n}_1$, $\mathfrak{n}_2$, …, $\mathfrak{n}_n$ along the $z$ axis. As before, we have a plane wave with its wave normal in the $x$–$z$ plane at an angle to the $z$-axis which we now call $\psi_n$ and which may be complex valued.[3] The field component $F$ in this wave at slice $n$ is given by

$$F = F_n \exp\bigl(-ik\mathfrak{n}_n(x \sin\psi_n + z\cos\psi_n)\bigr) \tag{A.22}$$

Assuming the slices are infinitesimally thin and $\mathfrak{n}_n \approx \mathfrak{n}_{n+1}$, there is only a very small reflection from the next strata. The wave in that stratum is

$$F = F_{n+1} \exp\bigl(-ik\mathfrak{n}_{n+1}(x \sin\psi_{n+1} + z\cos\psi_{n+1})\bigr). \tag{A.23}$$

Snell's law applies at the boundary. Hence

$$\mathfrak{n}_n \sin\psi_n = \mathfrak{n}_{n+1} \sin\psi_{n+1}. \tag{A.24}$$

If $\mathfrak{n}$ and $\psi$ are only functions of $z$, then Eq. (A.24) shows that the product $\mathfrak{n}\sin\psi$ is constant (Budden, 1961b, p. 103).

If the wave originates in free space (i.e. $\mathfrak{n} = 1$) with $\psi = \theta$ and we call $\psi$ in the ionosphere $\theta_h$, then

$$\sin\theta = \mathfrak{n}\sin\theta_h \tag{A.25}$$

which is an extension of Snell's law to a continuously varying medium. $\theta$ is the incident angle and $\mathfrak{n}$ and $\theta_h$ are associated with a wave transmitted through the boundary.

[3] Rather than trying to interpret a complex angle as a geometric angle, it is simplest to acknowledge that the sine function is meaningfully defined across the complex plane. Physically, complex angles often manifest as an evanescent wave in propagation problems.

$\mathfrak{n}$ and $\theta$ may be unknown, but their product is not. We now introduce the variable $q$ where

$$\mathfrak{n}_n \cos\theta_{h,n} = q_n \tag{A.26}$$

so that the wave field can be expressed as

$$F = F_n \exp\bigl(-ik(x\sin\theta + q_n z)\bigr). \tag{A.27}$$

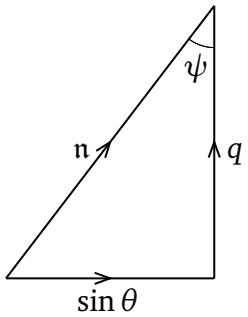


Figure A.4: $q$ may be regarded as the vertical component of $\mathfrak{n}$.

This means that for all values of $z$, the wave varies with $x$ in the same way through

$$\exp(-ikx\sin\theta). \tag{A.28}$$

The variable $q$ plays a major part in the theory of reflection from stratified media. In particular, we note that

$$\mathfrak{n}^2 = q^2 + \sin^2\theta \tag{A.29}$$

and that in general all angles and $q$ may be complex valued (Budden, 1961a, p. 121).

## A.3 Waveguides

Waveguide physics can be partially understood by examining what happens when a plane electromagnetic wave reflects off a single perfectly conducting boundary, as shown in Fig. A.5. The incident wave interferes with the reflected wave, forming a pattern of maxima and minima. Along the perfectly conducting boundary the transverse electric fields are zero. They are also zero at a distance $a$ from the boundary. If a second perfectly conducting plane were placed at $a$, the fields in the space between $a$ and the original boundary would be unaffected. Such a scenario is called a waveguide.

There is a net propagation to the right in Fig. A.5. Each of the "packets" in the imaginary guide has an effective wavelength $\lambda_g$ of

$$\lambda_g = \frac{\lambda_0}{\sin\theta}. \tag{A.30}$$

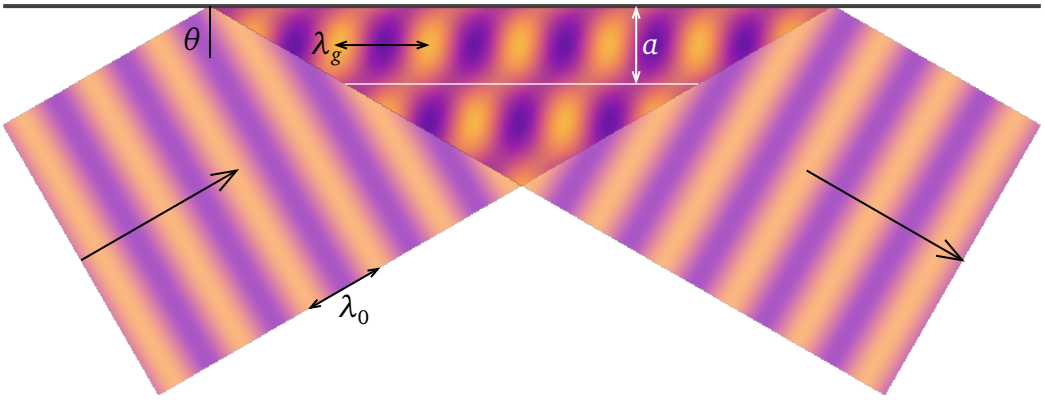


Figure A.5: A plane wave incident on a perfectly conducting boundary will interfere with its reflection. The boundary condition that the field be zero along the perfectly conducting boundary is also met at a distance $a$ from the boundary, such that if a second perfectly conducting boundary were placed there the two boundaries would form a waveguide.

If the angle of incidence $\theta$ changes, so does $a$ in order to maintain the boundary conditions

$$a = \frac{\lambda_0}{2\cos\theta}. \tag{A.31}$$

Clearly the free space wavelength also affects the values of $a$ and $\theta$ that meet the boundary conditions. Each waveguide has a cutoff frequency $f_c$

$$\cos\theta = \frac{\lambda_0}{2a} = \frac{c}{2af} = \frac{f_c}{f}, \qquad f_c = \frac{c}{2a} \tag{A.32}$$

below which propagation does not occur. Waves can exist for frequencies below the cutoff frequency at complex angles of incidence, but they are heavily attenuated and do not contribute to the field at appreciable distances from the source. They are known as evanescent waves.

### A.3.1 TM modes with perfectly reflecting boundaries

TM modes are waveguide modes in which the magnetic intensity vector is everywhere transverse to the direction of propagation, i.e. $\mathscr{H}_z = 0$. Because of anisotropy in the real ionosphere, pure TM modes cannot exist in the Earth-ionosphere waveguide. Nonetheless, the

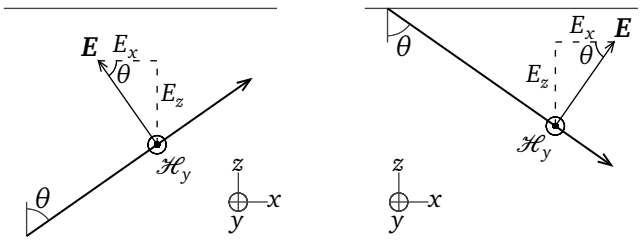


Figure A.6: Field geometry of the upgoing (left) and downgoing (right) plane waves described by Eq. (A.33) and Eq. (A.34), respectively.

real EIWG modes often resemble TM modes and are called quasi-TM. Here we assume both waveguide boundaries are perfect reflectors.

A plane electromagnetic wave travelling obliquely upwards in the waveguide with its electric vector in the $x$–$z$ plane has field components

$$\begin{aligned} E_x^I &= -E_0 \cos\theta \exp\big(-ik(Sx - Cz)\big) \\ E_z^I &= E_0 \sin\theta \exp\big(-ik(Sx - Cz)\big) \\ \mathscr{H}_y^I &= -E_0 \exp\big(-ik(Sx - Cz)\big) \end{aligned} \tag{A.33}$$

With reflection off a perfectly conducting boundary at $z = h$, $E_x$ must be 0, giving the downgoing fields as

$$\begin{aligned} E_x^R &= E_0 \cos\theta \exp\big(-ik(Sx + Cz)\big) \\ E_z^R &= E_0 \sin\theta \exp\big(-ik(Sx + Cz)\big) \\ \mathscr{H}_y^R &= -E_0 \exp\big(-ik(Sx + Cz)\big) \end{aligned} \tag{A.34}$$

The total fields are the sum of Eq. (A.33) and Eq. (A.34).

$$\begin{aligned} E_x &= -2iE_0 \cos\theta \sin(kz\cos\theta) \exp(-ikSx) \\ E_z &= 2E_0 \sin\theta \cos(kz\cos\theta) \exp(-ikSx) \\ \mathscr{H}_y &= -2E_0 \cos(kz\cos\theta) \exp(-ikSx) \end{aligned} \tag{A.35}$$

Suppose there also exists a perfectly conducting boundary at $z = 0$, such that

$$kh\cos\theta = n\pi \tag{A.36}$$

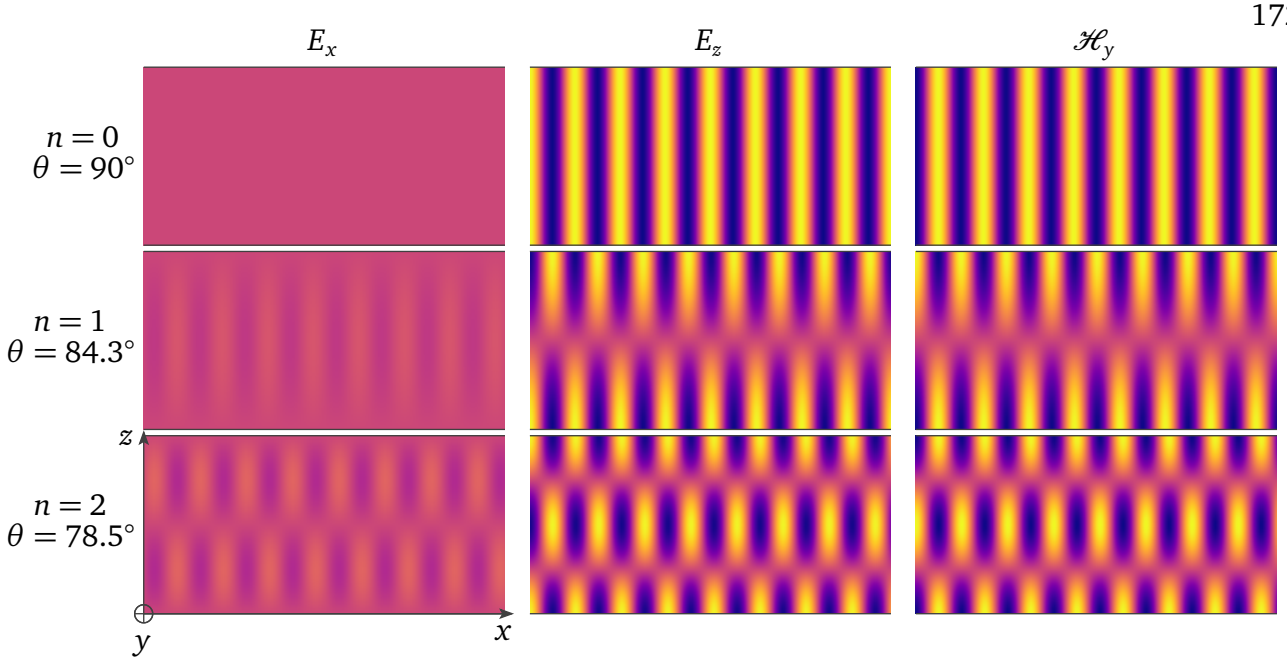


Figure A.7: Zeroth, first, and second modes for $E_x$, $E_z$, and $\mathscr{H}_y$ fields within a TM waveguide with perfectly reflecting boundaries. The field frequency is 20 kHz and waveguide height 75 km.

is satisfied. At this plane the condition $E_x = 0$ is also met, so that Eq. (A.35) are a solution of Maxwell's equations for the fields in free space between two parallel perfect conductors at $z = 0$ and $z = h$ (Budden, 1961b, p. 29). An example of the field patterns in the waveguide for the first three modes are shown in Fig. A.7.

### A.3.2 The complex $\theta$ plane

For waveguides with perfectly reflecting walls, the value of $\theta$ for a given mode are either on the real $\theta$-axis (propagated modes) or on the line $\mathrm{Re}(\theta) = 0$ (evanescent modes). When the walls are not perfect reflectors, this is no longer true. For an imperfect reflector $|R(\theta)|$ is less than 1 for real angles of incidence. Thus, the mode equation, Eq. (3.20), cannot be satisfied. To compensate for this, $C = \cos\theta$ must have an imaginary component, which means the mode is partially attenuated. When the waveguide boundaries are poor reflectors, solutions of the mode equation on the complex $\theta$ plane move away from the lines $\mathrm{Im}(\theta) = 0$ and $\mathrm{Re}(\theta) = 0$, as shown in Fig. A.8. The modes associated with $\theta$ having large imaginary part are highly

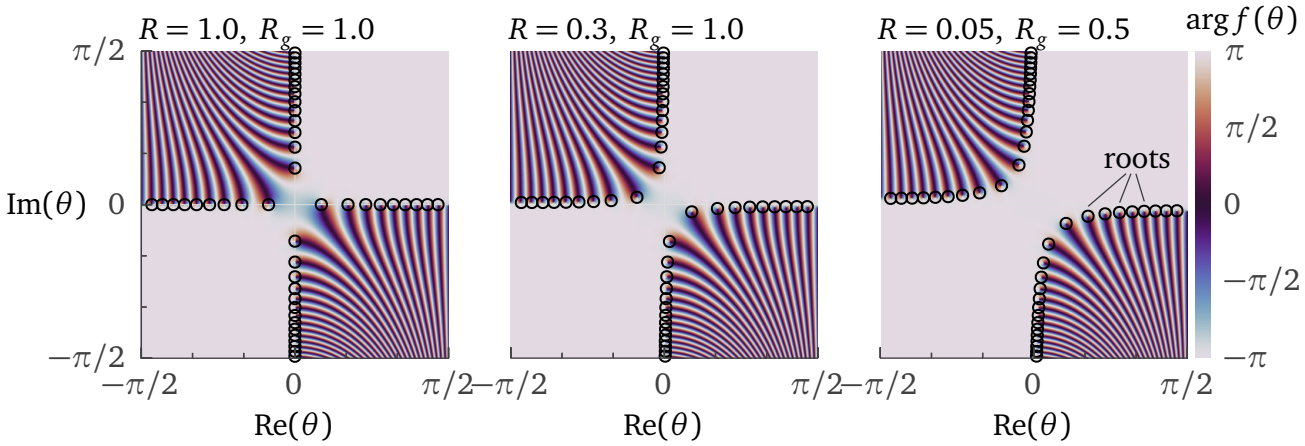


Figure A.8: When waveguide boundaries are imperfectly reflecting, solutions ("roots" or "zeros") to the fundamental equation of mode theory (Eq. (3.20)) require complex angles of incidence. Modes associated with angles having large imaginary parts are highly attenuated and make little contribution to the total field. Points representing $-\mathrm{Re}(\theta)$ provide no additional information and can be thought of as the corresponding "downgoing" wave to the $+\mathrm{Re}(\theta)$ upgoing waves (Budden, 1961b, p. 28). They may also be associated with modes of negative order which cannot be excited by a real source in the waveguide (Budden, 1961b, pp. 121–123).

attenuated; only the least attenuated modes make a significant contribution to the total field at a great distance from the radio source.

## A.4 Ionosphere reflection coefficients

### A.4.1 Differential of $R$ with respect to height

To compute the ionosphere reflection coefficient matrix we first derive the differential equations for $\boldsymbol{R}$ with respect to height. These equations come out of the definition of the reflection coefficients from Appendix A.2.3 and the matrix differential equation for the wave fields, Eq. (3.15).

The four elements of $\boldsymbol{e}$ fully describe the total electromagnetic field of the wave. Budden assumes that there is a very thin slice of vacuum (thin enough so it doesn't make a difference to the fields) parallel to the $z$ plane in the ionosphere which allows us to describe the wave fields as if they are in free space. The horizontal field components $E_x$, $-E_y$, $\mathscr{H}_x$, and $\mathscr{H}_y$ of $\boldsymbol{e}$

can be resolved into four component waves in any chosen way, as was done in Appendix A.2.3. We choose to resolve them into two upgoing and two downgoing plane waves, each with one electric field component and one magnetic field component in the plane of incidence (the $x$–$z$ plane).

Using the notation from Budden (1988, p. 303), for fields of the upgoing (incident) waves let

$$\mathscr{H}_y^I = f_1, \quad E_y^I = f_2, \quad \text{so} \quad E_x^I = Cf_1, \quad \mathscr{H}_x^I = -Cf_2 \tag{A.37}$$

and for the downgoing (reflected) waves

$$\mathscr{H}_y^R = f_3, \quad E_y^R = f_4, \quad \text{so} \quad E_x^R = -Cf_3 \quad \mathscr{H}_x^R = Cf_4 \tag{A.38}$$

where $f_1$, $f_2$, $f_3$, and $f_4$ are four complex amplitudes and $C = \cos\theta$. This form is equivalent to Budden (1955, p. 529).[4] Figure A.9 graphically depicts the relationship between $\mathscr{H}_y$ and $E_x$.

The total wave fields in free space is the sum of these four component waves

$$\boldsymbol{e} = \begin{pmatrix} E_x \\ -E_y \\ \mathscr{H}_x \\ \mathscr{H}_y \end{pmatrix} = f_1 \begin{pmatrix} C \\ 0 \\ 0 \\ 1 \end{pmatrix} + f_2 \begin{pmatrix} 0 \\ -1 \\ -C \\ 0 \end{pmatrix} + f_3 \begin{pmatrix} -C \\ 0 \\ 0 \\ 1 \end{pmatrix} + f_4 \begin{pmatrix} 0 \\ -1 \\ C \\ 0 \end{pmatrix} \tag{A.39}$$

In matrix form,

$$\boldsymbol{e} = \boldsymbol{S}\boldsymbol{f} \tag{A.40}$$

where

$$\boldsymbol{S} = \begin{pmatrix} C & 0 & -C & 0 \\ 0 & -1 & 0 & -1 \\ 0 & -C & 0 & C \\ 1 & 0 & 1 & 0 \end{pmatrix} \tag{A.41}$$

and $\boldsymbol{f} = (f_1, f_2, f_3, f_4)^\top$.

[4] The notation varies considerably between texts over time. Translating from the notation used here to Budden (1955): $\boldsymbol{S} \to \boldsymbol{L}$, $\boldsymbol{f} \to \boldsymbol{q}$, and $\boldsymbol{W} \to \boldsymbol{S}$.

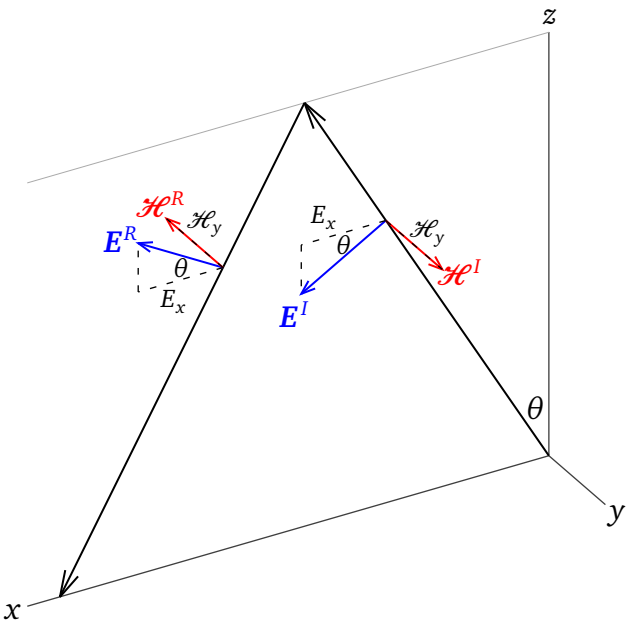


Figure A.9: Electric and magnetic intensity vectors of incident and reflected waves where $\mathscr{H}$ is aligned perpendicular to the plane of propagation. The $x$ and $y$ components are marked to demonstrate how the total $E_x = \cos\theta(\mathscr{H}_y^I - \mathscr{H}_y^R)$ in the free space of the waveguide given that $\boldsymbol{E}$ and $\mathscr{H}$ are equal magnitude. The identity holds when the intensity vectors are rotated about the wave propagation vector, but a diagram of the situation becomes difficult to interpret. The relationship between the total $\mathscr{H}_x$ field and $E_y$ can be similarly found.

Substituting Eq. (A.40) into Eq. (3.15) gives the differential equation satisfied by $\boldsymbol{f}$

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}z} = -i\boldsymbol{S}^{-1}\boldsymbol{T}\boldsymbol{S}\boldsymbol{f} = -\frac{1}{2}i\boldsymbol{W}\boldsymbol{f} \tag{A.42}$$

where

$$\boldsymbol{W} = \left(\begin{array}{cc|cc} T_{11}+T_{44}+T_{14}/C+CT_{41} & -T_{12}/C-T_{42} & -T_{11}+T_{44}+T_{14}/C-CT_{41} & -T_{12}/C-T_{42} \\ -T_{31}-T_{34}/C & C+T_{32}/C & T_{31}-T_{34}/C & -C+T_{32}/C \\ \hline -T_{11}+T_{44}-T_{14}/C+CT_{41} & T_{12}/C-T_{42} & T_{11}+T_{44}-T_{14}/C-CT_{41} & T_{12}/C-T_{42} \\ T_{31}+T_{34}/C & C-T_{32}/C & -T_{31}+T_{34}/C & -C-T_{32}/C \end{array}\right) \tag{A.43}$$

and $\boldsymbol{T}$ is a matrix that operates on $\boldsymbol{e}$ to describe the differential equations for the wavefields $\mathrm{d}\boldsymbol{e}/\mathrm{d}z$.

In Appendix A.2.3 we defined $\boldsymbol{R}$ in terms of electric and magnetic field components. We also noted that $\boldsymbol{e}^R = \boldsymbol{R}\boldsymbol{e}^I$. We can apply this exactly again as

$$\begin{pmatrix} f_3 \\ f_4 \end{pmatrix} = \boldsymbol{R} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \quad \text{or} \quad \boldsymbol{d} = \boldsymbol{R}\boldsymbol{u} \tag{A.44}$$

Then

$$\begin{pmatrix} f_1 \\ f_2 \end{pmatrix}' = \boldsymbol{u}' \quad \text{and} \quad \begin{pmatrix} f_3 \\ f_4 \end{pmatrix}' = \boldsymbol{d}' \tag{A.45}$$

where $'$ means $\frac{1}{k}\frac{\mathrm{d}}{\mathrm{d}z}$.

We combine this with the differential equation satisfied by $\boldsymbol{f}$, Eq. (A.42), where we partition $\boldsymbol{W}$, Eq. (A.43), into four $2 \times 2$ components

$$\boldsymbol{W} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \tag{A.46}$$

(A.47)

This implies

$$\boldsymbol{d}' = -\frac{1}{2} i (W_{21}\boldsymbol{u} + W_{22}\boldsymbol{d}) \tag{A.48}$$

$$\boldsymbol{u}' = -\frac{1}{2} i (W_{11}\boldsymbol{u} + W_{12}\boldsymbol{d}) \tag{A.49}$$

From the product rule, the derivative of $\boldsymbol{d}$ is

$$\boldsymbol{d}' = \boldsymbol{R}\boldsymbol{u}' + \boldsymbol{R}'\boldsymbol{u}. \tag{A.50}$$

Therefore

$$\begin{aligned} -\frac{1}{2} i (W_{21}\boldsymbol{u} + W_{22}\boldsymbol{d}) &= \boldsymbol{R}\boldsymbol{u}' + \boldsymbol{R}'\boldsymbol{u} \\ -\frac{1}{2} i (W_{21}\boldsymbol{u} + W_{22}\boldsymbol{R}\boldsymbol{u}) &= \boldsymbol{R}\left(-\frac{1}{2} i (W_{11}\boldsymbol{u} + W_{12}\boldsymbol{R}\boldsymbol{u})\right) + \boldsymbol{R}'\boldsymbol{u} \end{aligned}$$

Because every term has a $\boldsymbol{u}$ on the right side, we can omit it (Budden, 1988, p. 570).

$$-\frac{1}{2} i W_{21} - \frac{1}{2} i W_{22}\boldsymbol{R} = -\frac{1}{2} i \boldsymbol{R} W_{11} - \frac{1}{2} i \boldsymbol{R} W_{12}\boldsymbol{R} + \boldsymbol{R}'$$

Rearranging we get

$$\frac{2i}{k}\frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}z} = W_{21} + W_{22}\boldsymbol{R} - \boldsymbol{R}W_{11} - \boldsymbol{R}W_{12}\boldsymbol{R} \tag{A.51}$$

which is our differential equation for $\boldsymbol{R}$.

Remember that from the beginning, our definition of $\boldsymbol{R}$ and Eq. (A.44) has been in free space. Thus, the reflection coefficient obtained here is for the vacuum modes but not for the upgoing and downgoing characteristic waves of magnetoionic theory (Budden, 1988, p. 570). If the integration is stopped at some level in the ionosphere, this is equivalent to terminating the ionosphere by a sharp boundary at that level and the value of $\boldsymbol{R}$ that has been reached is the reflection coefficient for that boundary.

No component of $\boldsymbol{R}$ could ever have its complex norm reach or exceed unit amplitude, so there can be no failure of the integration.

### A.4.2 Booker quartic

The 4$^{\text{th}}$-order equation for wave fields in a fully homogeneous ionosphere is known as the *Booker quartic* (Yabroff, 1957). If the differential equation for the wavefields Eq. (3.15) is studied in a fully homogeneous ionosphere, the elements of $\boldsymbol{T}$ are independent of $z$ (Budden, 1988, p. 183). All field components of the wave should depend on $z$ only through a factor $\exp(-ikqz)$, as in Eq. (A.27). Combined with the differential equations for wavefields, Eq. (3.15), we get the eigenvalue problem

$$\boldsymbol{T}\boldsymbol{e} = q\boldsymbol{e} \tag{A.52}$$

or equivalently

$$(\boldsymbol{T} - q\boldsymbol{I})\boldsymbol{e} = 0 \tag{A.53}$$

for each eigenvalue $q$. The condition for a non-trivial solution is that

$$\det(\boldsymbol{T} - q\boldsymbol{I}) = 0. \tag{A.54}$$

Writing out the determinant gives

$$\begin{vmatrix} T_{11}-q & T_{12} & 0 & T_{14} \\ 0 & -q & 1 & 0 \\ T_{31} & T_{32} & -q & T_{34} \\ T_{41} & T_{42} & 0 & T_{44}-q \end{vmatrix} = 0 \tag{A.55}$$

$$\downarrow$$

$$\begin{aligned} q^4 - q^3(T_{11}+T_{44}) + q^2(T_{11}T_{44} - T_{32} - T_{14}T_{41}) + \\ q(T_{11}T_{32} + T_{32}T_{44} - T_{12}T_{31} - T_{34}T_{42}) + \\ T_{14}T_{32}T_{41} - T_{12}T_{34}T_{41} - T_{14}T_{31}T_{42} + \\ T_{11}T_{34}T_{42} + T_{12}T_{31}T_{44} - T_{11}T_{32}T_{44} = 0 \end{aligned} \tag{A.56}$$

which is a fourth degree equation for $q$ and is one form of the Booker quartic.

The wavefields $\boldsymbol{e}$ are the eigenvectors of the equation $(\boldsymbol{T} - q\boldsymbol{I})\boldsymbol{e} = 0$. These can be determined by substituting an eigenvalue $q$ into the equation and solving directly for $\boldsymbol{e}$. This can be done numerically by singular value decomposition. However, we can also generate a surprisingly simple analytical solution. We have the four equations

$$(T_{11}-q)e_1 + T_{12}e_2 + T_{14}e_4 = 0 \tag{A.57}$$

$$-qe_2 + e_3 = 0 \tag{A.58}$$

$$T_{31}e_1 + T_{32}e_2 - qe_3 + T_{34}e_4 = 0 \tag{A.59}$$

$$T_{41}e_1 + T_{42}e_2 + (T_{44}-q)e_4 = 0 \tag{A.60}$$

If we choose $e_2 = 1$, then it immediately follows from Eq. (A.58) that $e_3 = q$. We then have three equations and only two unknowns. Any two of the three equations can be used to solve

the system in the usual way, leading to

$$e_1 = \left(T_{12}(T_{44} - q) - T_{14}T_{42}\right)/d \tag{A.61}$$

$$e_2 = 1 \tag{A.62}$$

$$e_3 = q \tag{A.63}$$

$$e_4 = \left(-T_{12}T_{41} + T_{42}(T_{11} - q)\right)/d \tag{A.64}$$

$$\text{where} \qquad d = T_{14}T_{41} - (T_{11} - q)(T_{44} - q). \tag{A.65}$$

### A.4.3 Reflection from a sharply bounded ionosphere

First, it is interesting to note that if Eq. (3.29) were integrated for a great distance through a fully—horizontally and vertically—homogeneous ionosphere and the integration stopped when $\mathrm{d}\boldsymbol{R}/\mathrm{d}z = 0$, then the resulting $\boldsymbol{R}$ is the reflection coefficient for a homogeneous medium with a sharp lower free space boundary. This is precisely the solution we're looking for, and Budden (1988, p. 571) points out it is equivalent to finding the solution of

$$\boldsymbol{W}_{21} + \boldsymbol{W}_{22}\boldsymbol{R} - \boldsymbol{R}\boldsymbol{W}_{11} - \boldsymbol{R}\boldsymbol{W}_{12}\boldsymbol{R} = \boldsymbol{0} \tag{A.66}$$

which can be solved by iterative methods.

The approach used to compute the starting value of $\boldsymbol{R}$ in LongwaveModePropagator.jl is based on an approach in Budden (1988, ch. 18.7) that is very similar to the derivation of the differential equation $\mathrm{d}\boldsymbol{R}/\mathrm{d}z$. First, solve the Booker quartic, Eq. (A.56), using a polynomial root finder. Then sort the four $q$'s to select the two associated with upgoing waves (Pitteway, 1965). From Eq. (A.27), it is evident that the two complex $q$'s lying closest to the positive real and negative imaginary axis correspond to the two upgoing waves. The other two $q$'s correspond to the downgoing waves. Solve for the field component vectors associated with each, $\boldsymbol{e}_a$ and $\boldsymbol{e}_b$, from Eqs. (A.61) to (A.64). We extract the vacuum plane wave components $\boldsymbol{f}_a$ and $\boldsymbol{f}_b$ using

Eq. (A.40) and Eq. (A.41). With this transformation we have

$$\boldsymbol{f}_a = \begin{pmatrix} \boldsymbol{u}^a \\ \boldsymbol{d}^a \end{pmatrix}, \quad \boldsymbol{f}_b = \begin{pmatrix} \boldsymbol{u}^b \\ \boldsymbol{d}^b \end{pmatrix}. \tag{A.67}$$

From the definition of the reflection coefficient matrix, $\boldsymbol{d} = \boldsymbol{R}\boldsymbol{u}$ for both $a$ and $b$. Letting

$$\boldsymbol{U} = (\boldsymbol{u}^a, \boldsymbol{u}^b), \quad \boldsymbol{D} = (\boldsymbol{d}^a, \boldsymbol{d}^b) \tag{A.68}$$

leads to

$$\boldsymbol{D} = \boldsymbol{R}\boldsymbol{U} \tag{A.69}$$

which can be solved as $\boldsymbol{R} = \boldsymbol{D}\boldsymbol{U}^{-1}$. The solution does not depend on scaling of $\boldsymbol{e}^a$ or $\boldsymbol{e}^b$ because it will cancel out in $\boldsymbol{u}$, $\boldsymbol{d}$.

There are other approaches from the perspective of $\boldsymbol{M}$ rather than $\boldsymbol{T}$. Budden (1988, ch. 6) discusses stratified media and the Booker quartic in detail using a general $x$-$y$-$z$ coordinate system. Sheddy (1968) outlines an analytic solution for the reflection coefficient matrix from a sharply bounded ionosphere using a simplified coordinate system in which wave propagation is restricted to the $x$–$z$ plane. Using the simplified coordinates,

$$\frac{\partial}{\partial x} = -ikS, \quad \frac{\partial}{\partial y} = 0, \quad \frac{\partial}{\partial z} = -ikq \tag{A.70}$$

As usual, the last two of Maxwell's equations (eq. 3.7) describe the wave in the ionosphere. Beginning with the form Eq. (3.8) and making the substitutions for the partials above gives

$$qE_y = -\mathscr{H}_x, \qquad q\mathscr{H}_y = \epsilon_0^{-1} D_x \tag{A.71}$$

$$-qE_x + SE_z = -\mathscr{H}_y, \qquad -q\mathscr{H}_x + S\mathscr{H}_z = \epsilon_0^{-1} D_y \tag{A.72}$$

$$-SE_y = -\mathscr{H}_z, \qquad -S\mathscr{H}_y = \epsilon_0^{-1} D_z \tag{A.73}$$

Again we want $\boldsymbol{E}$ rather than $\boldsymbol{D}$. From the definition of electric displacement and polarization,

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{\epsilon}_r \boldsymbol{E} = \epsilon_0 \boldsymbol{E} + \epsilon_0 \boldsymbol{M} \cdot \boldsymbol{E} = \epsilon_0 (\boldsymbol{I} + \boldsymbol{M}) \boldsymbol{E} \tag{A.74}$$

Therefore, using the notation from Budden (1988, p. 145), we can write Eqs. (A.71) to (A.73) in matrix form as

$$\boldsymbol{\Gamma E} = \mathscr{H}, \qquad \boldsymbol{\Gamma}\mathscr{H} = -(\boldsymbol{I} + \boldsymbol{M})\boldsymbol{E} \tag{A.75}$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0 & -q & 0 \\ q & 0 & -S \\ 0 & S & 0 \end{pmatrix} \tag{A.76}$$

It is then trivial to remove $\mathscr{H}$

$$\begin{aligned} \boldsymbol{\Gamma}\mathscr{H} &= -(\boldsymbol{I} + \boldsymbol{M})\boldsymbol{E} \\ \boldsymbol{\Gamma\Gamma E} &= -(\boldsymbol{I} + \boldsymbol{M})\boldsymbol{E} \\ (\boldsymbol{\Gamma}^2 + \boldsymbol{I} + \boldsymbol{M})\boldsymbol{E} &= 0 \end{aligned} \tag{A.77}$$

Johler and Walters (1960) and Crombie (1961) derive the equivalent of Eq. (A.77) in their own ways. Sheddy (1968) provides no derivation, but his $\boldsymbol{L}$ is equivalent to our $\boldsymbol{\Gamma}^2$.

For a self-consistent solution $\boldsymbol{E}$, the determinant of the term in parentheses of Eq. (A.77) must be 0:

$$\det \begin{pmatrix} 1 - q^2 + M_{11} & M_{12} & Sq + M_{13} \\ M_{21} & 1 - q^2 - S^2 + M_{22} & M_{23} \\ Sq + M_{31} & M_{32} & C^2 + M_{33} \end{pmatrix} = 0 \tag{A.78}$$

When this is multiplied out, we end up with a $4^{\text{th}}$-order equation in $q$, the Booker quartic (Booker & Appleton, 1938, eq. 7),

$$B_4 q^4 + B_3 q^3 + B_2 q^2 + B_1 q + B_0 = 0 \tag{A.79}$$

where

$$B_4 = 1 + M_{33} \tag{A.80}$$

$$B_3 = S(M_{13} + M_{31}) \tag{A.81}$$

$$B_2 = -(C^2 + M_{33})(1 + M_{11}) + M_{13}M_{31} - (1 + M_{33})(C^2 + M_{22}) + M_{23}M_{32} \tag{A.82}$$

$$B_1 = S\left(M_{12}M_{23} + M_{21}M_{32} - (C^2 + M_{22})(M_{13} + M_{31})\right) \tag{A.83}$$

$$B_0 = (1 + M_{11})(C^2 + M_{22})(C^2 + M_{33}) + M_{12}M_{23}M_{31} + M_{13}M_{21}M_{32} - M_{13}(C^2 + M_{22})M_{31} - (1 + M_{11})M_{23}M_{32} - M_{12}M_{21}(C^2 + M_{33}) \tag{A.84}$$

is in the form given by Sheddy (1968, eq. 3a).

From the first two of Maxwell's equations, at the boundary of vacuum and the ionosphere the total tangential fields (i.e. in $x$ and $y$) must be continuous. For upgoing waves "1" and "2" and making use of the relationships established for incident and reflected waves in Eqs. (A.37) and (A.38), we can define the boundary equations[5]

$$E_{x,1} + E_{x,2} = C\mathscr{H}_y^I - C\mathscr{H}_y^R \tag{A.85}$$

$$E_{y,1} + E_{y,2} = E_y^I + E_y^R \tag{A.86}$$

$$\mathscr{H}_{x,1} + \mathscr{H}_{x,2} = -CE_y^I + CE_y^R \tag{A.87}$$

$$\mathscr{H}_{y,1} + \mathscr{H}_{y,2} = \mathscr{H}_y^I + \mathscr{H}_y^R \tag{A.88}$$

This system of equations above is extremely cumbersome to solve. The left hand side of Eqs. (A.85) to (A.88) can be found from Eq. (A.77) using both $q$'s. Then by definition (eq. A.17 and A.18) the components of $\boldsymbol{R}$ are ratios of terms on the right hand side of Eqs. (A.85) to (A.88). We have four equations and we need four field components for the reflection

[5] Although using $\mathscr{H}$ and slightly different notation, this is equivalent to Sheddy (1968, p. 794).

coefficients. Fortunately, Crombie (1961, pp. 457–460) has already solved them for us. Sheddy (1968) uses Crombie's solution for propagation along a line of magnetic latitude almost exactly, with a couple of sign corrections for the coordinate system. We include Sheddy's (1968) solution here for completeness:

a) Let $\boldsymbol{G} = \boldsymbol{\Gamma}^2 + \boldsymbol{I} + \boldsymbol{M}$

b) Let

$$\Delta_j = G_{11,j}G_{33} - G_{13,j}G_{31,j} \tag{A.89}$$

$$P_j = (-G_{12}G_{33} + G_{13,j}G_{32})/\Delta_j \tag{A.90}$$

$$T_j = q_j P_j - S(-G_{11,j}G_{32} + G_{12}G_{31,j})/\Delta_j \tag{A.91}$$

where $j = 1, 2$ for the two upgoing waves

c) Let $\Delta = (T_1 C + P_1)(C + q_2) - (T_2 C + P_2)(C + q_1)$

d) Then,

$$_{\parallel}R_{\parallel} = \big((T_1 C - P_1)(C + q_2) - (T_2 C - P_2)(C + q_1)\big)/\Delta \tag{A.92}$$

$$_{\perp}R_{\perp} = \big((T_1 C + P_1)(C - q_2) - (T_2 C + P_2)(C - q_1)\big)/\Delta \tag{A.93}$$

$$_{\perp}R_{\parallel} = -2C(T_1 P_2 - T_2 P_1)/\Delta \tag{A.94}$$

$$_{\parallel}R_{\perp} = -2C(q_1 - q_2)/\Delta \tag{A.95}$$

An alternate approach may be to solve for $\boldsymbol{E}$ from $\boldsymbol{\Gamma}^2 + \boldsymbol{I} + \boldsymbol{M}$ using both $q$'s, then Eq. (A.75) can be used to determine $\mathscr{H}$. Equations (A.85) to (A.88) may be solved as the matrix system[6]

$$\begin{pmatrix} C & 0 & -C & 0 \\ 0 & 1 & 0 & 1 \\ 0 & -C & 0 & C \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mathscr{H}_y^I \\ E_y^I \\ \mathscr{H}_y^R \\ E_y^R \end{pmatrix} = \begin{pmatrix} E_{x,1} + E_{x,2} \\ E_{y,1} + E_{y,2} \\ \mathscr{H}_{x,1} + \mathscr{H}_{x,2} \\ \mathscr{H}_{y,1} + \mathscr{H}_{y,2} \end{pmatrix} \tag{A.96}$$

[6] This $4 \times 4$ matrix is similar to $S$ in Eq. (A.41), except here it is formulated for $+E_y$.

using e.g. LU decomposition, although such a method was not implemented and the previous analytical solutions are more efficient.

## A.5 Ionosphere susceptibility tensor

The derivation of the ionosphere susceptibility tensor $\boldsymbol{M}$ is largely based on Budden (1988, p. 39), although previous authors, e.g. Banerjea and Saha (1947), have outlined similar steps. At each step we write the equation of motion of an electron in the ionosphere subject to an external force and then compute the overall electric polarization.

### A.5.1 Free, undamped electrons

In a large volume of plasma with a uniform electric intensity $\boldsymbol{E}$, every electron experiences the same electric force $\boldsymbol{E}e$. If $\boldsymbol{r}$ is the average vector displacement of an electron from the position it would occupy without the electric field, then its motion is harmonically varying as

$$\boldsymbol{r} = \boldsymbol{r}_0 \exp(i\omega t) \tag{A.97}$$

The constitutive relation is derived through the equation of motion of one electron:

$$\boldsymbol{E}e = m\ \partial^2\boldsymbol{r} \big/ \partial t^2 \tag{A.98}$$

where $m$ is the mass of an electron. First multiplying both sides by $Ne$ and then using Eq. (A.97) and Eq. (3.3):

$$\begin{aligned} \boldsymbol{E}Ne^2 &= Nem\ \partial^2\boldsymbol{r} \big/ \partial t^2 \\ &= -\omega^2 Nem\boldsymbol{r} \\ &= -\omega^2 m\boldsymbol{P} \end{aligned} \tag{A.99}$$

so that the constitutive relation is

$$\boldsymbol{P} = -\epsilon_0 X\boldsymbol{E} \tag{A.100}$$

where

$$X = \frac{Ne^2}{\epsilon_0 m\omega^2} = \frac{\omega_n^2}{\omega^2} \tag{A.101}$$

is the squared ratio of the angular plasma frequency to the angular wave frequency.

### A.5.2 Electron collisions

When electrons are moved by the field of a radio wave, they experience an average force caused by collisions with other particles. We assume that each electron makes $\nu$ instantaneous collisions per second with other particles. Now adding the effect of collisions to Eq. (A.98)

$$\boldsymbol{E}e = m\ \partial^2\boldsymbol{r}\big/\partial t^2 + m\nu\ \partial\boldsymbol{r}/\partial t\,. \tag{A.102}$$

Again multiplying both sides by $Ne$ (effectively this means that all the electrons experience the same collision frequency)

$$\boldsymbol{E}e = -\omega^2 m\boldsymbol{P} + i\omega m\nu Ne\boldsymbol{r} \tag{A.103}$$

$$= -\omega^2 m(1 - i\nu/\omega)\boldsymbol{P}. \tag{A.104}$$

Therefore, the constitutive relation is

$$\boldsymbol{P} = -\epsilon_0 \frac{X}{1 - iZ}\boldsymbol{E} \tag{A.105}$$

where $Z = \nu/\omega$. We will also define $U = 1 - iZ$ to be used later.

### A.5.3 Earth's magnetic field

Assuming $\boldsymbol{B}$ is the magnetic induction of Earth's magnetic field, then a charge $e$ moving with velocity $\partial\boldsymbol{r}/\partial t$ experiences a force $e\ \partial\boldsymbol{r}/\partial t \times \boldsymbol{B}$. The equation of motion of a single electron is now

$$\boldsymbol{E}e + e\ \partial\boldsymbol{r}/\partial t \times \boldsymbol{B} = m\ \partial^2\boldsymbol{r}\big/\partial t^2 + m\nu\ \partial\boldsymbol{r}/\partial t\,. \tag{A.106}$$

Multiplying both sides by $Ne/m\omega^2$ and computing $\partial/\partial t$ as multiplication by $i\omega$ results in

$$\frac{Ne^2}{m\omega^2}\boldsymbol{E} + \frac{ie}{m\omega}\boldsymbol{P} \times \boldsymbol{B} = -\boldsymbol{P}(1 - iZ). \tag{A.107}$$

We introduce the variable

$$\boldsymbol{Y} = \frac{e\boldsymbol{B}}{m\omega} = \frac{\omega_H}{\omega} \tag{A.108}$$

where $\omega_H$ is the angular gyrofrequency. Rearranging Eq. (A.107) we have

$$-\epsilon_0 X \boldsymbol{E} = U\boldsymbol{P} + i\boldsymbol{P} \times \boldsymbol{Y}. \tag{A.109}$$

Equation (A.109) is the constitutive relation used in longwave propagation mode theory. Written out as a matrix equation

$$-\epsilon_0 X \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} = \begin{pmatrix} U & iYl_z & -iYl_y \\ -Yl_z & U & iYl_x \\ iYl_y & -iYl_x & U \end{pmatrix} \begin{pmatrix} P_x \\ P_y \\ P_z \end{pmatrix} \tag{A.110}$$

where $l_x, l_y, l_z$ are direction cosines.

We can invert the matrix equation to solve for $\boldsymbol{P}$. This gives

$$\frac{1}{\epsilon_0} \begin{pmatrix} P_x \\ P_y \\ P_z \end{pmatrix} = \boldsymbol{M} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} \tag{A.111}$$

where the $3 \times 3$ matrix here is known as the susceptibility tensor $\boldsymbol{M}$ (Budden, 1988, p. 49)

$$\begin{aligned} \boldsymbol{M} &= \begin{pmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{pmatrix} \\ &= -\frac{X}{U(U^2 - Y^2)} \begin{pmatrix} U^2 - l_x^2 Y^2 & -il_z YU - l_x l_y Y^2 & il_y YU - l_x l_z Y^2 \\ il_z YU - l_x l_y Y^2 & U^2 - l_y^2 Y^2 & -il_x YU - l_y l_z Y^2 \\ -il_y YU - l_x l_z Y^2 & il_x YU - l_y l_z Y^2 & U^2 - l_z^2 Y^2 \end{pmatrix}. \end{aligned} \tag{A.112}$$

### A.5.4 Earth curvature

So far we have developed the theory assuming a waveguide with flat reflecting boundaries. Of course Earth's surface is curved, and this affects the fields in the waveguide. Rather than

rewriting the problem in spherical coordinates, we model the influence of Earth curvature by means of a fictitious media in the free space of the waveguide—a technique described by Booker and Walkinshaw (1946) for modeling tropospheric refraction and ducting and depicted in Fig. A.10. Richter (1966) applies conformal mapping to the Earth-ionosphere waveguide problem and Spies and Wait (1961) develops analytical solutions for spherical geometry as a check on the validity of the fictitious refractive index method.

In free space between ground and the ionosphere a ray travels in a straight line, yet if an observer saw the Earth as flat, the ray would appear to bend upward. If the ray is inclined at an angle $\psi$ to the flat Earth, then the ray has an upward curvature of $\cos(\psi)/R_e$. This path could be described by a medium with an index of refraction that varies with height. Budden (1961b, p. 140) supposes that we continue to assume the Earth is flat, while imagining a ray launched into a medium with real index of refraction $\mathfrak{n}(z)$. Here we'll assume Earth is a sphere.

From Appendix A.2.4 Snell's law requires that $\mathfrak{n}\cos\psi$ equals some constant $u$. Therefore,

$$\frac{\mathrm{d}\mathfrak{n}}{\mathrm{d}\psi} = u\tan\psi\sec\psi = \mathfrak{n}\tan\psi \tag{A.113}$$

If $s$ is distance along the ray, then curvature of the ray through the medium is

$$\frac{\mathrm{d}\psi}{\mathrm{d}s} = \frac{\mathrm{d}\mathfrak{n}/\mathrm{d}s}{\mathrm{d}\mathfrak{n}/\mathrm{d}\psi} = \frac{\mathrm{d}\mathfrak{n}}{\mathrm{d}z}\frac{\sin\psi}{\mathfrak{n}\tan\psi} = \frac{1}{\mathfrak{n}}\frac{\mathrm{d}\mathfrak{n}}{\mathrm{d}z}\cos\psi \tag{A.114}$$

where $\mathrm{d}\mathfrak{n}/\mathrm{d}s = \frac{\mathrm{d}\mathfrak{n}}{\mathrm{d}z}\sin\psi$. Equating the two curvatures gives

$$\frac{1}{\mathfrak{n}}\frac{\mathrm{d}\mathfrak{n}}{\mathrm{d}z}\cos\psi = \frac{\cos\psi}{R_e} \tag{A.115}$$

so

$$\mathfrak{n} = \exp\left(\frac{z-H}{R_e}\right) \tag{A.116}$$

where $H$ is a constant. The exponential expands into a power series

$$\mathfrak{n} = 1 + \frac{z-H}{R_e} + \frac{1}{2}\left(\frac{z-H}{R_e}\right)^2 + \ldots \tag{A.117}$$

Because $|z-H| \ll R_e$, we drop the higher powers of $(z-H)/R_e$. Out of convenience we work with $\mathfrak{n}^2$:

$$\mathfrak{n}^2 = 1 - \alpha(H-z) \tag{A.118}$$

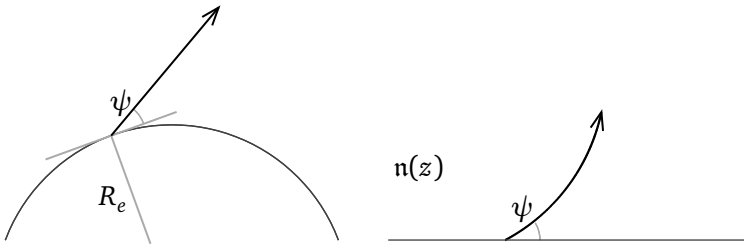


Figure A.10: Equivalence of a straight ray over curved earth with a curved ray over flat earth.

where $\alpha = 2/R_e$.

The effect of Earth curvature can be included in, e.g. the calculation of the ionosphere reflection coefficients, by modifying the susceptibility matrix (Pappert et al., 1967, p. 398). The modification is to add the curvature term to the diagonal of $\boldsymbol{M}$:

$$\boldsymbol{M} \leftarrow \boldsymbol{M} + \begin{pmatrix} -\alpha(H-z) & 0 & 0 \\ 0 & -\alpha(H-z) & 0 \\ 0 & 0 & -\alpha(H-z) \end{pmatrix}. \tag{A.119}$$

This modification makes sense given that $\mathfrak{n}^2 \approx \epsilon_r$ and $\epsilon_r = \boldsymbol{I} + \boldsymbol{M}$.

Choice of the reference height $H$ is somewhat arbitrary. It is the height at which $\mathfrak{n} = 1$ and represents true free space. $H$ is often chosen just below the ionosphere boundary (Budden, 1961b, p. 140; Pappert et al., 1967, p. 391). By definition, $H$ is the height at which the eigenangles are referenced,[7] but it is necessary to reference the eigenangles to the ground for use in excitation factors. For isotropic ionospheres, the first-order Earth curvature theory used here predicts that the ground eigenangles are invariant with $H$ (Richter, 1966). Pappert et al. (1967, p. 392) concludes that variance in the ground eigenangles with $H$ is mostly due to higher order curvature effects and confirms in Pappert (1968, p. 223) that the results are adequate even for strongly anisotropic ionospheres.

To change the reference height of $\theta$ from $H$ to ground or vice-versa, look back at Eq. (A.25). Now the situation is slightly different. Because of our use of the fictitious free space media, we

[7] The modal equation is solved with reflection coefficients defined in free space.

have $\mathfrak{n}_0 = (1-\alpha H)^{1/2}$ and $\mathfrak{n}_H = 1$, by definition.

$$\mathfrak{n}_H \sin\theta_H = \mathfrak{n}_0 \sin\theta_0 \tag{A.120}$$

$$\sin\theta_H = (1-\alpha H)^{1/2} \sin\theta_0 \tag{A.121}$$

## A.6 Ground reflection coefficients

The ground reflection is described by Fresnel reflection equations. The derivation of these reflection coefficients makes use of the law of reflection, that $\theta^R = \theta^I$, and Snell's law of refraction. We have spoken at length about reflection, but have yet to explicitly mention transmission. In general, some incident energy will be transmitted through the boundary between two different media. We will denote the angle from normal of the transmitted wave as $\theta^T$.

The reflection matrix for the ground, Eq. (3.32), is diagonal because we treat it as isotropic. Recalling our definition of the reflection coefficients,

$$_{\parallel}\overline{R}_{\parallel} = \mathscr{H}_y^R / \mathscr{H}_y^I, \qquad _{\perp}\overline{R}_{\perp} = E_y^R / E_y^I \tag{A.122}$$

so it is necessary to develop the Fresnel reflection equations for both transverse electric and transverse magnetic cases. Maxwell's equations specify the boundary conditions:

(1) the perpendicular component of $\boldsymbol{B}$ is continuous across the boundary

(2) the parallel component of $\boldsymbol{E}$ is continuous across the boundary

### A.6.1 Transverse electric Fresnel equations

Boundary conditions for the transverse electric case give:

$$E^I + E^R = E^T \tag{A.123}$$

$$B^I C^I - B^R C^R = B^T C^T \tag{A.124}$$

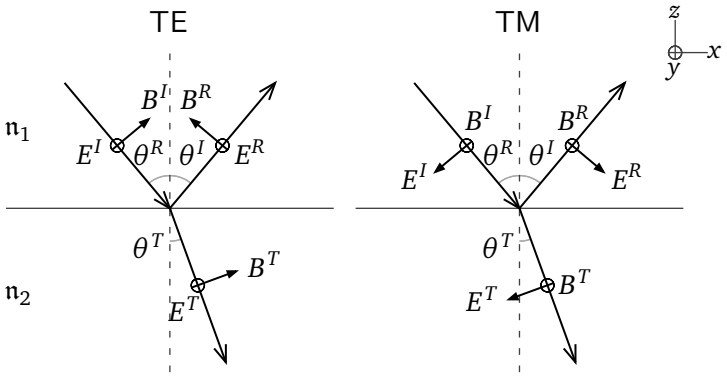


Figure A.11: Reflection and transmission geometry for transverse electric (TE) and transverse magnetic (TM) waves.

where $C^I$ and $C^T$ are $\cos\theta^I$ and $\cos\theta^T$, respectively.

We can eliminate the magnetic field through

$$B = \frac{k}{\omega}E. \tag{A.125}$$

Because $\omega$ does not change between the media it cancels out. Additionally, because of the law of reflection we can replace $\theta^R$ with $\theta^I$

$$k_1(E^I - E^R)C^I = k_2 E^T C^T. \tag{A.126}$$

Practically, we prefer to work with the index of refraction of a material, so we make the substitution

$$\mathfrak{n}_1(E^I - E^R)C^I = \mathfrak{n}_2 E^T C^T \tag{A.127}$$

where $\mathfrak{n} = ck/\omega$.

Next we apply the boundary condition for $\boldsymbol{E}$ to remove $E^T$.

$$E^I(\mathfrak{n}_1 C^I - \mathfrak{n}_2 C^T) = E^R(\mathfrak{n}_1 C^I + \mathfrak{n}_2 C^T) \tag{A.128}$$

It immediately follows that the ratio of the field amplitudes is

$$\frac{E^R}{E^I} = \frac{\mathfrak{n}_1 C^I - \mathfrak{n}_2 C^T}{\mathfrak{n}_1 C^I + \mathfrak{n}_2 C^T}. \tag{A.129}$$

We remove $C^T$ through Snell's law, $\mathfrak{n}_1 S^I = \mathfrak{n}_2 S^T$, so

$$C^T = \left(1-(S^T)^2\right)^{1/2} = \left(1-\frac{\mathfrak{n}_1^2}{\mathfrak{n}_2^2}(S^I)^2\right)^{1/2}. \tag{A.130}$$

Just above the ground of our waveguide $\mathfrak{n}_1 = 1$; $\mathfrak{n}_2 = \mathfrak{n}_g$ where $\mathfrak{n}_g$ is the refractive index of ground. Therefore, the ratio simplifies to

$$\frac{E^R}{E^I} = {}_\perp\overline{R}_\perp = \frac{C-\left(\mathfrak{n}_g^2-S^2\right)^{1/2}}{C+\left(\mathfrak{n}_g^2-S^2\right)^{1/2}} \tag{A.131}$$

where we used $\mathfrak{n}(1-S^{I^2}/\mathfrak{n}^2)^{1/2} \equiv (\mathfrak{n}^2-S^{I^2})^{1/2}$ for positive $\mathfrak{n}$.

### A.6.2 Transverse magnetic Fresnel equations

The derivation of the transverse magnetic equation is very similar to the transverse electric. This time the boundary conditions result in

$$B^I + B^R = B^T \tag{A.132}$$

$$E^I C^I - E^R C^R = E^T C^T \tag{A.133}$$

We eliminate the electric field through

$$E = \frac{\omega}{k} B \tag{A.134}$$

resulting in the second boundary condition becoming

$$\frac{1}{k_1}(B^I - B^R)C^I = \frac{1}{k_2} B^T C^T. \tag{A.135}$$

The index of refraction is substituted in through $1/k = c/(\mathfrak{n}\omega)$

$$\frac{1}{\mathfrak{n}_1}(B^I - B^R)C^I = \frac{1}{\mathfrak{n}_2} B^T C^T. \tag{A.136}$$

Application of the boundary condition for $B$ allows us to remove $B^T$

$$\frac{1}{\mathfrak{n}_1}(B^I - B^R)C^I = \frac{1}{\mathfrak{n}_2}(B^I + B^R)C^T \tag{A.137}$$

$$B^I\left(\frac{C^I}{\mathfrak{n}_1} - \frac{C^T}{\mathfrak{n}_2}\right) = B^R\left(\frac{C^I}{\mathfrak{n}_1} + \frac{C^T}{\mathfrak{n}_2}\right) \tag{A.138}$$

So that the field ratio is

$$\frac{B^R}{B^I} = \frac{C^I/\mathfrak{n}_1 - C^T/\mathfrak{n}_2}{C^I/\mathfrak{n}_1 + C^T/\mathfrak{n}_2} \tag{A.139}$$

Substituting $\mathfrak{n}_1 = 1$ and $\mathfrak{n}_2 = \mathfrak{n}_g$ and multiplying the top and bottom of the fraction by $\mathfrak{n}_g$ results in

$$\frac{B^R}{B^I} = \frac{C^I\mathfrak{n}_g - C^T}{C^I\mathfrak{n}_g + C^T} \tag{A.140}$$

Applying Snell's law and the same trigonometry identity as before gives

$$\frac{B^R}{B^I} = \frac{C^I\mathfrak{n}_g - \left(1 - (S^I)^2/\mathfrak{n}_g^2\right)}{C^I\mathfrak{n}_g + \left(1 - (S^I)^2/\mathfrak{n}_g^2\right)} \tag{A.141}$$

Multiplying both the numerator and denominator by $\mathfrak{n}_g$ gives

$$\frac{B^R}{B^I} = \frac{\mathscr{H}^R}{\mathscr{H}^I} = {}_{\|}\overline{R}_{\|} = \frac{C\mathfrak{n}_g^2 - \left(\mathfrak{n}_g^2 - S^2\right)}{C\mathfrak{n}_g^2 + \left(\mathfrak{n}_g^2 - S^2\right)} \tag{A.142}$$

### A.6.3 Refractive index of the ground

We often know the permittivity and conductivity of the ground rather than its refractive index directly. We can compute the refractive index of the ground, $\mathfrak{n}_g$, given its conductivity and relative permittivity here.

In general the refractive index $\mathfrak{n}$ is

$$\mathfrak{n} = (\epsilon_r\mu_r)^{1/2}. \tag{A.143}$$

Assuming $\mu_r \approx 1$, this simplifies to

$$\mathfrak{n}^2 \approx \epsilon_r. \tag{A.144}$$

For lossy media the total complex relative permittivity is

$$\epsilon_r = \epsilon_r' - i\frac{\sigma}{\omega\epsilon_0} \tag{A.145}$$

where $\epsilon_r'$ is the real permittivity. Then the squared refractive index of ground is

$$\mathfrak{n}_g^2 = \epsilon_r' - i\frac{\sigma}{\omega\epsilon_0} \tag{A.146}$$

used directly in Eq. (A.131) and Eq. (A.142). Typically $\epsilon_r'$ is referred to simply as $\epsilon_r$.

## A.7 Excitation of waveguide modes

Budden (1962) treats the excitation of waveguide modes in terms of line sources. We'll use notation similar to Ferguson and Snyder (1980, pp. 24–30), who presents a summary of Budden's work. To avoid confusion with previous notation, we use $\xi$ to represent the angle between the wave normal and the *horizontal*. Budden (1962) uses this angle rather than the angle $\theta$ from the vertical in order to simplify the contour integrals. Budden begins by assuming the source is a line of quadrupoles at height $d$ parallel to and aligned with the $y$-axis. If the line source has unit strength, the $y$-component of the magnetic field in free space is

$$\mathscr{H}_y = \frac{k^3}{4\pi\epsilon_0} \int_{0-i\infty}^{\pi+i\infty} \exp\bigl(ik(x\cos\xi + |z-d|\sin\xi)\bigr)\cos^2\xi \,\mathrm{d}\xi. \tag{A.147}$$

Imagine the level $d$ is in the free space of a waveguide with boundaries at $z = l$ and $z = u$ such that $l < d < u$. Each boundary has an associated reflection coefficient matrix $\overline{\boldsymbol{R}}_l$ and $\boldsymbol{R}_u$ referenced to their respective heights. The field integrand in Eq. (A.147) then represents a plane wave which will undergo reflections at $z = u$ and $z = l$ and these reflected waves must be summed to obtain the total field. There are four combinations resulting from reflections of the possible pairs of up- and downgoing waves: 1) initially upgoing to downgoing, 2) initially upgoing to upgoing, 3) initially downgoing to upgoing, and 4) initially downgoing to downgoing.[8]

The total field is written out as Budden (1962, eq. 32) and Ferguson and Snyder (1980,

[8] This is often explained using a diffraction grating model; reflected fields can be produced by image sources at heights $x - d$, $2uN \pm d$, and $2lN \pm d$ where $N$ is any positive integer.

eq. 19):

$$\begin{aligned}\mathscr{H}_y = &\frac{k^3}{4\pi\epsilon_0}\sum_{N=0}^{\infty}\int_C \exp(-ik\cos\xi)(\overline{\boldsymbol{R}}\boldsymbol{R})^N\boldsymbol{P}\cos^2\xi\,\mathrm{d}\xi \\ &+\frac{k^3}{4\pi\epsilon_0}\sum_{N=0}^{\infty}\int_C \exp(-ik\cos\xi)(\overline{\boldsymbol{R}}\boldsymbol{R})^N\overline{\boldsymbol{R}}\boldsymbol{P}\cos^2\xi\,\mathrm{d}\xi \\ &+\frac{k^3}{4\pi\epsilon_0}\sum_{N=0}^{\infty}\int_C \exp(-ik\cos\xi)(\boldsymbol{R}\overline{\boldsymbol{R}})^N\boldsymbol{P}\cos^2\xi\,\mathrm{d}\xi \\ &+\frac{k^3}{4\pi\epsilon_0}\sum_{N=0}^{\infty}\int_C \exp(-ik\cos\xi)(\boldsymbol{R}\overline{\boldsymbol{R}})^N\boldsymbol{R}\boldsymbol{P}\cos^2\xi\,\mathrm{d}\xi\end{aligned} \tag{A.148}$$

where $\boldsymbol{P}$ is a vector describing the electric and magnetic components of the source. If the second element of $\boldsymbol{P}$ is zero, it corresponds to a vertical electric dipole. The sum can be computed after careful selection of the contour of integration (Budden, 1961b; 1962). It is then convenient to let the width of the free space region tend to zero. So long as $l < d < u$ is maintained, both of the reflection coefficients are evaluated at $d$ and the subscripts $l$ and $u$ are no longer needed. The result of summing the fields is

$$\mathscr{H}_y = \frac{k^3}{4\pi\epsilon_0}\int_{0-i\infty}^{\pi+i\infty}\exp(-ikx\cos\xi)(\boldsymbol{I}+\boldsymbol{R})(\boldsymbol{I}-\overline{\boldsymbol{R}}\boldsymbol{R})^{-1}(\boldsymbol{I}+\overline{\boldsymbol{R}})\boldsymbol{P}\cos^2\xi\,\mathrm{d}\xi. \tag{A.149}$$

Budden (1962, p. 549) employs the theory of residues to deform the contour of the integral. The new contour crosses several singularities, the most important of which are poles of the factor $(\boldsymbol{I}-\overline{\boldsymbol{R}}\boldsymbol{R})^{-1}$. The contribution of a single pole at $\xi_n$ is

$$\mathscr{H}_{y,n} = \frac{ik^3}{2\epsilon_0}\exp(-ikx\cos\xi_n)\frac{(\boldsymbol{I}+\boldsymbol{R})_n\boldsymbol{X}_n(\boldsymbol{I}+\overline{\boldsymbol{R}})_n\boldsymbol{P}\cos^2\xi_n}{\left(\frac{\partial F}{\partial\xi}\right)_{\xi_n}} \tag{A.150}$$

where $F = \det(\boldsymbol{I}-\overline{\boldsymbol{R}}\boldsymbol{R})$ and $\boldsymbol{X}_n = \lim_{\xi\to\xi_n} F(\boldsymbol{I}-\overline{\boldsymbol{R}}\boldsymbol{R})^{-1}$ (Ferguson & Snyder, 1980, eq. 24). The reader should recognize $F$ as equivalent to the modal equation Eq. (3.24). Thus, the solution of the pole condition yields the waveguide modes and the residues at the poles give the excitation factors. See Budden (1952) or Budden (1961b, ch. 14) for a more thorough description of the residue series and integration path.

Performing the matrix multiplications and employing the geometrical spreading factor for

a spherical Earth yields (Ferguson & Snyder, 1980, eq. 26)

$$\mathscr{H}_{y,n} = \frac{-k^{5/2}\exp(i\pi/4)\cos^{3/2}\xi_n}{2\epsilon_0\left(2\pi R_e \sin(\rho/R_e)\right)^{1/2}} \frac{(\boldsymbol{I} + {}_{\|}\overline{R}_{\|})^2(\boldsymbol{I} - {}_{\perp}\overline{R}_{\perp\perp}R_{\perp})}{{}_{\|}\overline{R}_{\|}\left(\frac{\partial F}{\partial \xi}\right)_{\xi_n}} \exp(-ik\rho\cos\xi_n). \tag{A.151}$$

For compatibility with the rest of the formulation, it is necessary to substitute $\xi$ for $\theta$. At this stage, that only requires interchanging $\sin\xi$ for $\cos\theta = C$ and $\cos\xi$ with $\sin\theta = S$.

Equation (A.151) assumes both the source and receiver are at the height at which the reflection coefficients are determined. If the dependence of $\mathscr{H}_{y,n}$ on height $z$ is given by a function $G_n(z)$, then multiplication of Eq. (A.151) by $G_n(z_R)/G_n(d)$ where $z_R$ is the receiver height and $d$ is the reference height for the reflection elements.

### A.7.1 Height gain functions

Fundamentally, as we've previously seen in Appendix A.3, the total field in the vacuum waveguide is the sum of the upgoing and downgoing waves associated with each component. The boundary condition is that the tangential fields are continuous at transitions between free space and the waveguide boundaries. We know the fields will also satisfy the wave equation, e.g.

$$\frac{\mathrm{d}^2 E_y}{\mathrm{d}z^2} + k^2 q^2 E_y = 0. \tag{A.152}$$

From the definition of $q$, Eq. (A.29), and our use of a modified free-space index of refraction to simulate curved Earth, Eq. (A.118),

$$q^2 = \mathfrak{n}^2 - S^2 \tag{A.153}$$

$$= 1 - \alpha(H - z) - S^2 \tag{A.154}$$

$$= C^2 - \alpha(H - z) \tag{A.155}$$

By letting (Morfitt and Shellman, 1976, app. B; Budden and Martin, 1962)

$$\zeta = \left(\frac{k}{\alpha}\right)^{2/3}\left(C^2 - \alpha(H - z)\right) \tag{A.156}$$

we have the simplified form

$$\frac{\mathrm{d}^2 E_y}{\mathrm{d}\zeta^2} + \zeta E_y = 0 \tag{A.157}$$

This equation looks like the Airy function, but is actually called Stokes' equation. The solution to Stokes' equation uses the modified Hankel functions of order one-third,[9] which were described and tabulated by The Staff of the Computation Library (1945). The general solution to Eq. (A.157) is

$$E_y(\zeta) = a_1 h_1(\zeta) + a_2 h_2(\zeta) \tag{A.158}$$

where $a_1$ and $a_2$ are constants and $h_1$ and $h_2$ are the modified Hankel functions of order one-third. Calculation of the height gain functions and corresponding excitation factors requires finding $a_i$'s that are consistent with the waveguide boundary conditions. Using somewhat older notation, this process is stepped through in Budden and Martin (1962). Over time, self-consistent forms of the excitation factors and height gain functions were worked out for the different field components and collected by Pappert.

From Pappert and Shockey (1971, p. 6), the height gain functions are

$$f_1(z) = \exp\left(\frac{z}{R_e}\right)\left(F_1 h_1(q) + F_2 h_2(q)\right) \tag{A.159}$$

$$f_2(z) = \frac{1}{ik}\frac{df_1}{dz} = \frac{1}{ikR_e}\exp\left(\frac{z}{R_e}\right)\left(F_1 h_1(q) + F_2 h_2(q) + R_e\left(F_1 h_1'(q) + F_2 h_2'(q)\right)\right) \tag{A.160}$$

$$f_3(z) = F_3 h_1(q) + F_4 h_2(q) \tag{A.161}$$

where $f_1$ is the height gain for $E_z$, $f_2$ for $E_x$ and $f_3$ for $E_y$. They are also often referred to as $f_{\|}$,

[9] Despite their name, these aren't "Hankel functions" as we currently know them.

$g$, and $f_\perp$, respectively. The inner terms are (Pappert & Shockey, 1971, eq. 6–9)

$$F_1 = -H_2(q_0) + i\frac{n_0^2}{\mathfrak{n}_g^2}\left(\frac{kR_e}{2}\right)^{1/3}(\mathfrak{n}_g^2 - S^2)^{1/2}h_2(q_0) \tag{A.162}$$

$$F_2 = H_1(q_0) - i\frac{n_0^2}{\mathfrak{n}_g^2}\left(\frac{kR_e}{2}\right)^{1/3}(\mathfrak{n}_g^2 - S^2)^{1/2}h_1(q_0) \tag{A.163}$$

$$F_3 = -h_2'(q_0) + i\left(\frac{kR_e}{2}\right)^{1/3}(\mathfrak{n}_g^2 - S^2)^{1/2}h_2(q_0) \tag{A.164}$$

$$F_4 = h_1'(q_0) - i\left(\frac{kR_e}{2}\right)^{1/3}(\mathfrak{n}_g^2 - S^2)^{1/2}h_1(q_0) \tag{A.165}$$

using

$$q_z = \left(\frac{2}{kR_e}\right)^{-2/3}\left(C^2 - \frac{2}{R_e}(H - z)\right) \tag{A.166}$$

$$H_j(q) = h_j'(q) + \frac{1}{2}\left(\frac{2}{kR_e}\right)^{2/3}h_j(q), \quad j = 1, 2 \tag{A.167}$$

$$\mathfrak{n}_z^2 = 1 - \frac{2}{R_e}(H - z) \tag{A.168}$$

$$\mathfrak{n}_g^2 = \frac{\epsilon}{\epsilon_0} - i\frac{\sigma}{\omega\epsilon_0} \tag{A.169}$$

(Pappert & Shockey, 1971, eq. 10–13). The subscripts on $q$ and $\mathfrak{n}$ represent the value of $z$ at which they are evaluated. Note that $C$ is the cosine of the angle of incidence at height $H$, the height at which the modified index of refraction is equal to unity. $h_1$, $h_2$ and their derivatives, signified by the prime $'$, are again the modified Hankel functions of order one-third (The Staff of the Computation Library, 1945).

Using the excitation factors defined in the next section, the height gain functions for the receiver must be multiplied by the following factors to convert from $\mathscr{H}_y$ to the appropriate electric field component in the mode sum:

| Receiver field component | Receiver term in mode sum |
|---|---|
| $E_z$ | $-S_0 f_z(z_r)$ |
| $E_y$ | $(E_y/\mathscr{H}_y)f_y(z_r)$ |
| $E_x$ | $-f_x(z_r)$ |

$E_y/\mathscr{H}_y$ is a polarization ratio that can be computed either as

$$E_y/\mathscr{H}_y = \frac{(1+{}_{\perp}\overline{R}_{\perp})(1-{}_{\parallel}R_{\parallel\parallel}\overline{R}_{\parallel})}{(1+{}_{\parallel}\overline{R}_{\parallel})\,{}_{\parallel}R_{\perp\perp}\overline{R}_{\perp}} \tag{A.170}$$

preferred for a principally transverse magnetic mode or

$$E_y/\mathscr{H}_y = \frac{(1+{}_{\perp}\overline{R}_{\perp})\,{}_{\perp}R_{\parallel\parallel}\overline{R}_{\parallel}}{(1+{}_{\parallel}\overline{R}_{\parallel})(1-{}_{\perp}R_{\perp\perp}\overline{R}_{\perp})} \tag{A.171}$$

for a principally transverse electric mode.

### A.7.2 Alterations for ELF

The absolute value of the imaginary part of the eigenangle can become very large when in the ELF band. Pappert et al. (1970, p. 6) suggests that to avoid overflow, the flat earth analogues of the height gains Eqs. (A.159) to (A.161) should be used

$$f_1(z) = \frac{\exp(ikCz) + {}_{\parallel}\overline{R}_{\parallel}\exp(-ikCz+2ikCd)}{\exp(ikCd)(1+{}_{\parallel}\overline{R}_{\parallel})} \tag{A.172}$$

$$f_2(z) = \frac{C\left(\exp(ikCz) - {}_{\parallel}\overline{R}_{\parallel}\exp(-ikCz+2ikCd)\right)}{\exp(ikCd)(1+{}_{\parallel}\overline{R}_{\parallel})} \tag{A.173}$$

$$f_3(z) = \frac{\exp(ikCz) + {}_{\perp}\overline{R}_{\perp}\exp(-ikCz+2ikCd)}{\exp(ikCd)(1+{}_{\parallel}\overline{R}_{\parallel})} \tag{A.174}$$

### A.7.3 Excitation factors

Pappert et al. (1970, pp. 3–4) summarizes the excitation factors associated with both vertical and horizontal dipole sources and provides citations to their development. In Pappert et al. (1983, p. 10) they are presented in a convenient form, which we modify so they are defined for $\mathscr{H}_y$. This modification makes it easier to explicitly split the mode sum into a transmitter term and a receiver term.

$$\lambda = \begin{array}{ll} \text{Dipole orientation} & \\ \text{vertical} & -S_0T_1 \\ \text{end-on} & T_1 \\ \text{broadside} & T_3T_4 \end{array} \tag{A.175}$$

where the $T$'s are

$$T_1 = \frac{S^{1/2}(1+{}_{\|}\overline{R}_{\|})^2(1-{}_{\perp}R_{\perp}{}_{\perp}\overline{R}_{\perp})}{\frac{\partial F}{\partial\theta}\Big|_{\theta_n}{}_{\|}\overline{R}_{\|}D_{11}} \tag{A.176}$$

$$T_2 = \frac{S^{1/2}(1+{}_{\perp}\overline{R}_{\perp})^2(1-{}_{\|}R_{\|}{}_{\|}\overline{R}_{\|})}{\frac{\partial F}{\partial\theta}\Big|_{\theta_n}{}_{\perp}\overline{R}_{\perp}D_{22}} \tag{A.177}$$

$$T_3 = \frac{S^{1/2}(1+{}_{\|}\overline{R}_{\|})(1+{}_{\perp}\overline{R}_{\perp}){}_{\|}R_{\perp}}{\frac{\partial F}{\partial\theta}\Big|_{\theta_n}D_{12}} \tag{A.178}$$

$$T_4 = \frac{{}_{\perp}R_{\|}}{{}_{\|}R_{\perp}} \tag{A.179}$$

and

$$D_{11} = f_1^2(0) = \big(F_1h_1(0)+F_2h_2(0)\big)^2 \tag{A.180}$$

$$D_{12} = f_1f_3(0) = \big(F_1h_1(0)+F_2h_2(0)\big)\big(F_3h_1(0)+F_4h_2(0)\big) \tag{A.181}$$

$$D_{22} = f_3^2(0). \tag{A.182}$$

The $S$'s here are referenced to the curvature reference height $H$ where the reflection coefficients are defined and the modal equation solved. $S_0$ is sine of the eigenangle referenced to the ground. Also, we note $T_2$ and $D_{22}$ are not needed in our formulation.

## A.8 Mode conversion

Pappert and Snyder (1972) discuss the full mode conversion algorithm for carrying wavefields from one waveguide segment into the next. Assuming a unit amplitude wave in mode $k$ in the transmitter region (segment $M$), the $y$ and $z$ components of the electric and magnetic fields of the wave in segment $p$ is given by

$$\mathbf{e}_k^p = \frac{\exp(-ikS_{k,M}x_{M-1})}{(\sin(x/R_e))^{1/2}}\sum_j a_{jk}^p\exp(-ikS_j^p(x-x_p))\boldsymbol{f}_j^p(z) \tag{A.183}$$

where the height gain function $\boldsymbol{f}_j^p$ is

$$\boldsymbol{f}_j^p(z) = \begin{pmatrix} E_{yj}^p \\ E_{zj}^p \\ \mathscr{H}_{yj}^p \\ \mathscr{H}_{zj}^p \end{pmatrix}. \tag{A.184}$$

We use $a_{jk}^p$ to represent the coefficients associated with conversion from the $k^{\text{th}}$ to the $j^{\text{th}}$ mode.

Continuity allows us to determine the conversion coefficients $a_{jk}^p$ for the $p^{\text{th}}$ slab in terms of the $(p+1)^{\text{th}}$ slab.

$$a_{jk}^p = \begin{cases} 1 \quad \text{(conversion not applicable)} & p = M \\ I_{j,k}^{p,p+1} & p = M-1 \\ \displaystyle\sum_m a_{mk}^{p+1} \exp(-ikS_m^{p+1}(x_p - x_{p+1})) I_{j,m}^{p,p+1} & 1 \le p \le M-1 \end{cases} \tag{A.185}$$

where $I$ is the integral

$$I_{j,m}^{p,p+1} = \int_{-\infty}^{\infty} \left(\boldsymbol{g}_j^p\right)^{*\top} \boldsymbol{f}_m^{p+1} \, \mathrm{d}z. \tag{A.186}$$

$g$ is the height gain function for the *adjoint* waveguide. Equation (A.185) also makes use of the orthogonality condition between modes

$$\int_{-\infty}^{\infty} \left(\boldsymbol{g}_j^p\right)^{*\top} \boldsymbol{f}_m^p \, \mathrm{d}z = \begin{cases} 0 & j \neq m \\ 1 & j = m \end{cases} \tag{A.187}$$

where the superscript $^{*\top}$ denotes the conjugate transpose. To avoid confusion in this section, the dagger † will be used to identify parameters related to the adjoint waveguide, but *does not* act as an operator. Instead, conjugate transpose operations will be marked explicitly with $^{*\top}$.

The introduction of an adjoint waveguide may seem sudden, but is an important step in the solution of the conversion coefficients. Equation (3.38) is multiplied by the conjugate transpose of the adjoint height gain function for the $m^{\text{th}}$ forward-travelling mode in region 2, followed by integration over all space. This results in

$$\int_{-\infty}^{\infty} \boldsymbol{g}_m^{*\top} \overrightarrow{\boldsymbol{f}}_{m,1} \, \mathrm{d}z = T_{mj} \int_{-\infty}^{\infty} \boldsymbol{g}_m^{*\top} \overrightarrow{\boldsymbol{f}}_{m,2} \, \mathrm{d}z \tag{A.188}$$

which can be written as

$$I^{2,1}_{m,j} = T_{mj} I^{2,2}_{m,m}. \tag{A.189}$$

This yields the conversion coefficient

$$T_{mj} = I^{2,1}_{m,j} / I^{2,2}_{m,m} \tag{A.190}$$

which, due to the orthogonality condition, is equivalent to

$$T_{mj} = I^{2,1}_{m,j}. \tag{A.191}$$

### A.8.1 Adjoint waveguides and orthogonality

As we've already seen, modes can be obtained from fields for which

$$\frac{\partial}{\partial x} = -ikS, \quad \frac{\partial}{\partial y} = 0, \quad \frac{\partial}{\partial z} = i\omega \tag{A.192}$$

where $S$ only takes on discrete values corresponding to each mode $n$. The field components for each $n^{\text{th}}$ mode are

$$\boldsymbol{f}_n = (E_{y,n}, E_{z,n}, \mathscr{H}_{y,n}, \mathscr{H}_{z,n})^{\top} \tag{A.193}$$

Equation (A.192) can be applied to Maxwell's equations. Eliminating $E_x$ and $\mathscr{H}_x$, Maxwell's equations can be written with the $4 \times 4$ matrix operator $\boldsymbol{L}$

$$\boldsymbol{L} \begin{pmatrix} E_y \\ E_z \\ \mathscr{H}_y \\ \mathscr{H}_z \end{pmatrix} = -\frac{1}{ik} \frac{\partial}{\partial x} \begin{pmatrix} E_y \\ E_z \\ \mathscr{H}_y \\ \mathscr{H}_z \end{pmatrix} \tag{A.194}$$

which is equivalent to the eigenvalue equation

$$\boldsymbol{L}\boldsymbol{f}_n = S_n \boldsymbol{f}_n \tag{A.195}$$

There exists an adjoint matrix operator to $\boldsymbol{L}$ denoted $\boldsymbol{L}^{\dagger}$. The corresponding eigenvalue equation is

$$\boldsymbol{L}^{\dagger}\boldsymbol{g}_m = \lambda_m \boldsymbol{g}_m \tag{A.196}$$

where $\lambda_m$ are the eigenvalues and $\boldsymbol{g}_m$ is the eigenfunction of $\boldsymbol{L}^\dagger$, effectively the height gain function of the adjoint waveguide, similar to $\boldsymbol{f}$.

We now introduce a property of the adjoint. First, the general definition of the inner product is

$$(a, b) = \int_{-\infty}^{\infty} (a^*)^\top b \, \mathrm{d}z \tag{A.197}$$

The adjoint satisfies

$$(\boldsymbol{L}^\dagger a, b) = (a, \boldsymbol{L} b) \tag{A.198}$$

Thus,

$$\begin{aligned} &(\boldsymbol{L}^\dagger \boldsymbol{g}, \boldsymbol{f}) && = (\boldsymbol{g}, \boldsymbol{L}\boldsymbol{f}) \\ &= \int \left(\boldsymbol{L}^\dagger \boldsymbol{g}\right)^{*\top} \boldsymbol{f} \, \mathrm{d}z && = \int \boldsymbol{g}^{*\top} S \boldsymbol{f} \, \mathrm{d}z \\ &= \int (\lambda \boldsymbol{g})^{*\top} \boldsymbol{f} \, \mathrm{d}z && = S \int \boldsymbol{g}^{*\top} \boldsymbol{f} \, \mathrm{d}z \\ &= \lambda^* (\boldsymbol{g}, \boldsymbol{f}) && = S (\boldsymbol{g}, \boldsymbol{f}) \end{aligned} \tag{A.199}$$

This yields

$$(\lambda_m^* - S_n)(\boldsymbol{g}_m(z), \boldsymbol{f}_n(z)) = 0 \tag{A.200}$$

which gives the *biorthogonality condition*

$$(\boldsymbol{g}_m(z), \boldsymbol{f}_n(z)) = 0, \qquad p \neq q \tag{A.201}$$

and the relation

$$\lambda_m^* = S_n. \tag{A.202}$$

Pappert and Smith (1972) discuss one way to calculate the adjoint height gain function $\boldsymbol{g}$. Their method results in

$$\boldsymbol{g}^*(z) = \begin{pmatrix} \mathscr{H}_z(-l, z) \\ -\mathscr{H}_y(-l, z) \\ -E_z(-l, z) \\ E_y(-l, z) \end{pmatrix} \tag{A.203}$$

where $l$ refers to the $x$ component of the direction cosine of the geomagnetic field. Therefore, $\boldsymbol{g}(z)$ can be obtained in the same way as $\boldsymbol{f}(z)$ except assuming an "adjoint waveguide" where the $x$ component of the geomagnetic field $l$ is replaced by $-l$ and the components of the vector $\boldsymbol{g}^*$ are rearranged as above.[10]

### A.8.2 Integration of wavefields

$f$ and $g$ are determined by full wave numerical integration, as described by Pitteway (1965). Pitteway sought to determine the wavefields and reflection properties inside a horizontally stratified ionosphere. The technique used to calculate the reflection coefficient in Budden (1955) does not conveniently describe the wavefields in the ionosphere; further integration would be necessary to obtain the actual wavefields. Rather than referencing the solutions to linearly polarized waves parallel $\parallel$ and perpendicular $\perp$ to the plane of incidence, Pitteway (1965) chooses the two solutions which correspond to waves incident on the ionosphere from below with polarizations chosen to give the greatest and smallest intensity at great heights. These solutions are for the penetrating and non-penetrating modes (Fig. A.12), and come from the integration of $\mathrm{d}\boldsymbol{e}/\mathrm{d}z$, Eq. (3.15).

[10] Because Pappert and Smith (1972) solve for $\boldsymbol{g}^*$ directly, rather than $\boldsymbol{g}$, only the transpose of $\boldsymbol{g}^*$ needs to be taken when calculating $I$.

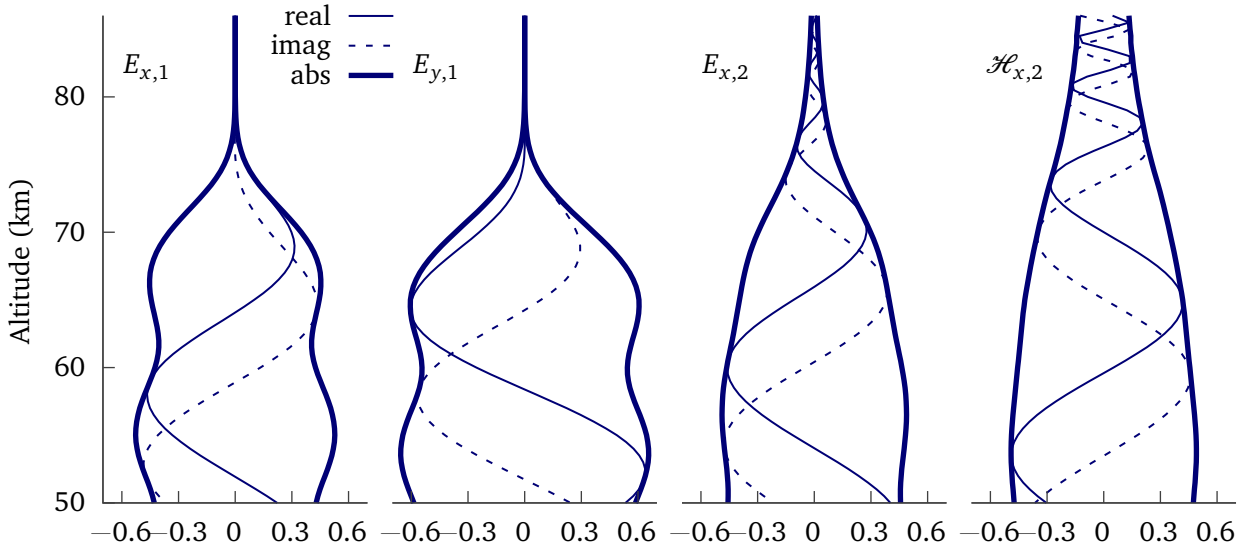


Figure A.12: Wavefields in an exponential ionosphere integrated by Pitteway integration. The two fields on the left are referred to by Pitteway as “non-penetrating” and the fields on the right are “penetrating”.

# Appendix B

# LongwaveModePropagator.jl

## B.1 Reflection coefficient solvers

Over 90 % of the runtime of LongwaveModePropagator.jl is spent searching for waveguide modes. The mode search runtime is dominated by integrating $\mathrm{d}\boldsymbol{R}/\mathrm{d}z$ through the ionosphere thousands of times for different $\theta$. Therefore, choice of the differential equations solver has a meaningful impact on the total runtime of the program. There are effectively two parameters we can choose: the solver algorithm and the tolerance. Efficient solutions are often the result of an appropriate choice of solver for a particular problem. DifferentialEquations.jl[1] has a large number of solvers specialized for solution of different types of differential equations, as well as solvers with special characteristics such as energy conservation or better handling of stiff equations. Fortunately, our equation is non-stiff and only of size 8 (4 complex equations). Additionally, no entry of $\boldsymbol{R}$ should exceed a complex norm of 1.

A set of non-stiff, low-to-moderate tolerance solvers were evaluated for default use in LongwaveModePropagator.jl. Thirty random scenarios were generated with

- $\theta \in (\pi/6, \pi/2) - (\pi/18, 0)i$ rad
- $f \in (10, 60)$ kHz
- $|\boldsymbol{B}| \in (30, 60)$ μT, dip angle $\in (0, \pi/2)$ rad, and azimuth $\in (0, 2\pi)$ rad
- Wait ionosphere with $h' \in (69, 89)$ km and $\beta \in (0.2, 1.0)$ km$^{-1}$

[1] https://diffeq.sciml.ai/stable/

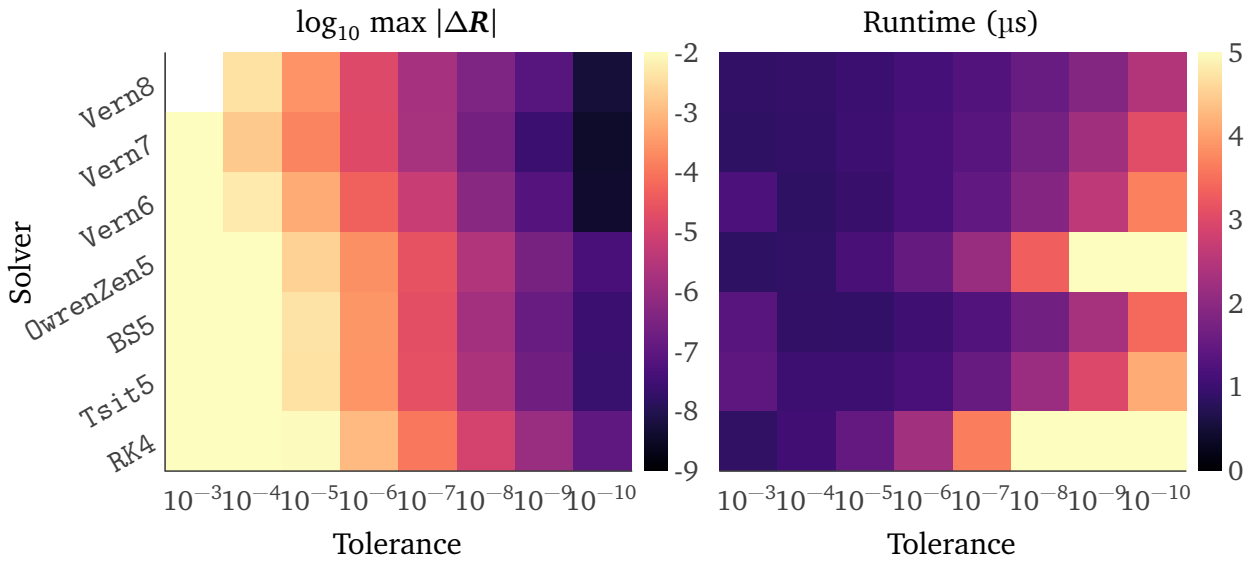


Figure B.1: Accuracy and runtime integrating $\mathrm{d}\boldsymbol{R}/\mathrm{d}z$ for a set of differential equations solvers. Left: base 10 logarithm of the maximum absolute difference between $\boldsymbol{R}$ of the solver and $\boldsymbol{R}$ determined using the `RK4` solver at a tolerance of $10^{-14}$. Right: average runtime for a single integration of $\boldsymbol{R}$. Lower values are better in each plot. A description of the solvers is available at https://diffeq.sciml.ai/stable/solvers/ode_solve.

Each solver is run for a range of tolerances on each of these ionospheres. The accuracy of the computed $\boldsymbol{R}$ is evaluated by comparing the maximum absolute difference of each of the elements of $\boldsymbol{R}$ from the test solver against the `RK4` solver using a tolerance of $10^{-14}$. To compute the runtime of each solver and tolerance, the integration is timed for 25 repetitions and averaged. The results are shown in Fig. B.1.

The tolerance used by the global complex roots and poles finding (GRPF) algorithm dominates the influence of the tolerance used to compute $\boldsymbol{R}$ when it comes to identifying the eigenangles. Here the tolerance of GRPF was $10^{-5}$ rad (0.0006°). The `v0.1` series of LongwaveModePropagator.jl explicitly filtered out candidate eigenangles based on solution of the mode equation, but Eq. (3.24) is very sensitive to $\theta$. On average, moving from a solution $\theta$ to $\theta + (10^{-3} + 10^{-3}i)$ changes the value of the mode function by several orders of magnitude. This made defining an appropriate threshold difficult and the mode finder rejected solutions that it did not need to. In fact, not a single mode of all the candidate modes in the 30 random

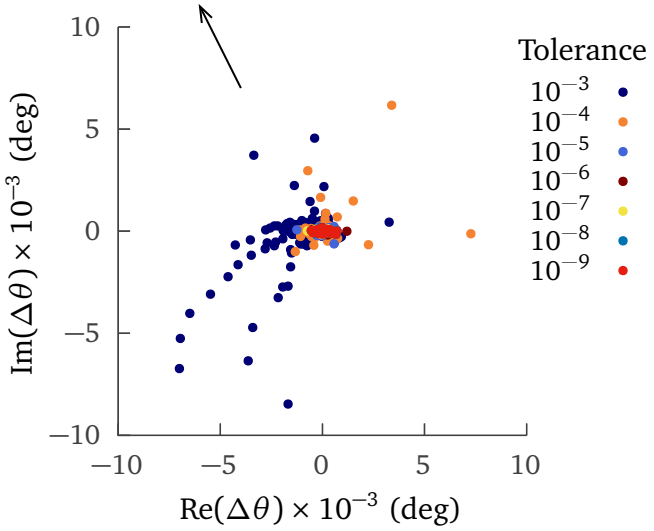


Figure B.2: Difference in modes identified over 30 random EIWG propagation scenarios for a series of solver tolerances using `Vern7` to integrate $\mathrm{d}\boldsymbol{R}/\mathrm{d}z$ compared to a tolerance of $10^{-10}$. The arrow points towards a single outlier for $10^{-3}$ at $-14.3 + 30.9i$.

scenarios used to evaluate the differential equations solvers needed to be rejected, although several would have been rejected using the criteria in `v0.1`. In `v0.2` and onward, used to produce this dissertation, this unnecessary filtering of eigenangles was removed.

Figure B.1 clearly shows `Vern7` has better accuracy relative to the other solvers for all tested tolerances. Figure B.2 shows the difference in modes identified for all 30 scenarios using `Vern7` at a series of tolerances between $10^{-3}$ and $10^{-9}$ compared to the results using $10^{-10}$. The tolerance of $10^{-3}$ and $10^{-4}$ had greater spread,[2] but by $10^{-5}$, there is little average difference from $10^{-10}$. This suggests that the default solver tolerance can be as low as about $10^{-5}$.

## B.2 Wavefield solvers

An analysis similar to the analysis of solvers for the ionosphere reflection coefficient in Appendix B.1 was conducted for integration of the wavefields $\mathrm{d}\boldsymbol{e}/\mathrm{d}z$. These height gain functions are simpler than the $\boldsymbol{R}(z)$ curves (see Fig. B.3,) but a similar set of solvers were compared. Interestingly, setting the tolerance anywhere between $10^{-3}$ and $10^{-10}$ has no influence on the

[2] The solver detected instabilities during some of the scenarios.

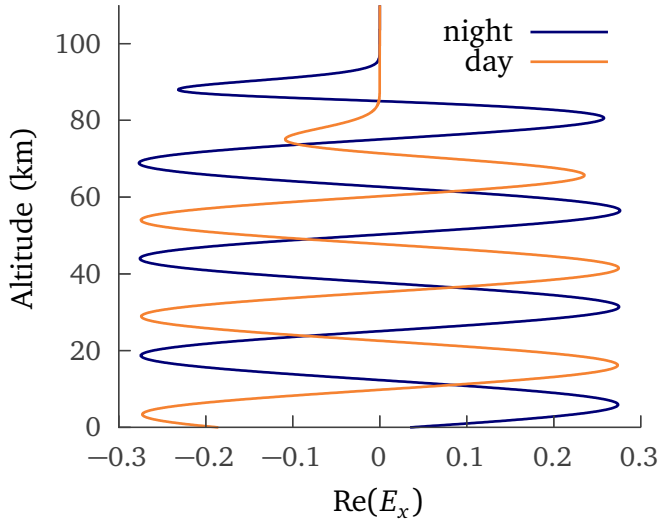


Figure B.3: The real component of the $E_x$ wavefield for the non-penetrating upgoing wave solution for a Wait daytime and nighttime ionosphere. The day ionosphere has $h' = 75\,\text{km}$ and $\beta = 0.35\,\text{km}^{-1}$ and the night ionosphere has $h' = 86\,\text{km}$ and $\beta = 0.6\,\text{km}^{-1}$. The wave frequency is 24 kHz and is at a real angle of incidence of 60° from vertical. The background magnetic field is 50 μT with a dip angle of 68° and azimuth of 111° east of the propagation direction.

accuracy of the solution or a significant affect on the runtime for a given solver algorithm. All of the solvers have similar accuracy, but vary in average runtimes, as shown in Table B.1. `Tsit5` is used by default because it has the lowest runtime.

Table B.1: Average runtime of differential equation solvers integrating $\mathrm{d}\boldsymbol{e}/\mathrm{d}z$ through a Wait daytime and nighttime ionosphere with an integration tolerance of $10^{-6}$. Smaller is better.

| Solver | Average runtime (ms) |
|---|---|
| `Tsit5` | 0.89 |
| `BS5` | 1.0 |
| `Vern6` | 1.1 |
| `OwrenZen5` | 1.1 |
| `Vern7` | 1.3 |
| `Vern8` | 1.5 |
| `Vern9` | 2.2 |