Technische Universität München

# A regional physics-motivated electron density model of the ionosphere

Dissertation
von

Wenjing Liang

Deutsches Geodätisches Forschungsinstitut
und Lehrstuhl für Geodätische Geodynamik

Ingenieurfakultät Bau Geo Umwelt

Ingenieurfakultät Bau Geo Umwelt

Deutsches Geodätisches Forschungsinstitut
und Lehrstuhl für Geodätische Geodynamik

# A regional physics-motivated electron density model of the ionosphere

Wenjing Liang

Vollständiger Abdruck der von der Ingenieurfakultät Bau Geo Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

**Doktor - Ingenieurs**

genehmigten Dissertation.

| | |
|---|---|
| Vorsitzende(r): | Univ.-Prof. Dr.-Ing. habil. Thomas Wunderlich |
| Prüfer der Dissertation: | 1. apl. Prof. Dr.-Ing. habil. Michael Schmidt |
| | 2. Univ.-Prof. Dr. phil. nat. Urs Hugentobler |
| | 3. Prof. Dr. Claudio Brunini, Universidad Nacional de La Plata, Argentinien |

Die Dissertation wurde am 08.11.2016 bei der Technischen Universität München eingereicht und durch die Ingenieurfakultät Bau Geo Umwelt am 16.01.2017 angenommen.

# Acknowledgements

First of all, I would like to express my deep gratitude to my supervisor Michael Schmidt for his supervision and invaluable suggestions on my dissertation. Without the discussions with him, this work would not have been finished. I am also very grateful to Urs Hugentobler for his suggestions for this work. I am also thankful to Claudio Brunini for his encourage and review of this work.

Additionally, I would like to thank all my colleagues at Deutsches Geodätisches Forschungsinstitut der Technischen Universität München (DGFI-TUM). Special thanks are devoted to Florian Seitz, Denise Dettmering, Marco Limberger, Eren Erdogan, Eva Börgens, Andreas Goss and Johannes Bouman.

Moreover, I would like to express my sincere appreciate to the following data providers: Laura Sánchez from DGFI-TUM; Lung-Chi Tsai from the Center for Space and Remote Sensing Research (CSRSR) of National Central University (NCU) in Taiwan; Bodo W. Reinisch from Lowell Digisonde International in U.S. and the ionosonde operators in Jicamarca and Port Stanley for making their data available via the Lowell GIRO Data Center; Center for Orbit Determination in Europe (CODE)/International GNSS Service (IGS); International Reference Ionosphere (IRI) and International Reference Ionosphere Plasmasphere (IRI-Plas) model.

I would also like to acknowledge the Deutsches Forschungsgemeinschaft, who has funded the project "MuSIK", in which part of the work has been carried on.

Finally, I would like to thank my family in particular my parents for their continuous supports and encourages during writing this thesis.

# Zusammenfassung

Die Ionosphäre ist der obere Teil der Erdatmosphäre, in dem die Anzahl der freien Elektronen ausreicht, um die Ausbreitung elektromagnetischer Signale zu beeinflussen. Die Kenntnis der Elektronendichte ist nicht nur für die Ionosphärenforschung und die Weltraumwetter-Überwachung wichtig, sondern spielt auch in vielen Anwendungsgebieten, wie z.B. der Positionierung und der Navigation, eine entscheidende Rolle. Die aktuell verfügbaren Elektronendichtemodelle erfüllen jedoch zumeist nicht die notwendigen Genauigkeitsanforderungen. Sie sind entweder rein mathematisch (z.B. voxel-basiert) oder nur klimatologisch, d.h. sie liefern nur monatliche Durchschnittswerte (z.B. die Internationale Referenz-Ionosphäre IRI oder NeQuick).

Durch die rasante Entwicklung der geodätischen Raumverfahren in den letzten Jahrzehnten sind Anzahl und Qualität der Messungen, die Ionosphärenmodelle verbessern können, drastisch angestiegen. Insbesondere liefern die bodenbasierten Zwei-Frequenz-Messungen der globalen Navigations-Satellitensysteme (GNSS) hochgenaue Information über die Elektronendichteverteilung innerhalb der Ionosphäre. Diese Information bezeichnet man als den absoluten Elektronengehalt (engl.: Slant Total Electron Content, STEC), der sich aus der Integration der Elektronendichte entlang des Signalweges zwischen Sender und Empfänger berechnet.

Die raumbasierten GNSS Radio-Okkultationsmessungen (RO) ermöglichen die Berechnung von Elektronendichteprofilen (EDPs) mit hoher vertikaler Auflösung und globaler Überdeckung. Um das Wissen über die Ionosphäre zu erweitern, ist das Zusammenspiel der einzelnen Beobachtungstechniken wichtig, da diese sich durch unterschiedliche Datenverteilungen, zeitliche und räumliche Auflösungen sowie spezifische Sensitivitäten auszeichnen.

Die Ionosphäre setzt sich in Abhängigkeit von ihrer chemischen Zusammensetzung aus verschiedenen Schichten zusammen; unter denen enthält die sogenannte F2-Schicht die maximale Elektronendichte. In dieser Arbeit wird ein regionales physikalisch-basiertes vierdimensionales (4-D) Elektronendichtemodell aus der Kombination der verschiedenen geodätischen Raumverfahren entwickelt. Dazu werden die Elektronendichteverteilungen der einzelnen Schichten, d.h. der D-, der E- sowie der F1- und der F2-Schicht durch Chapman-Funktionen beschrieben. Da Chapman-Funktionen zum Teil physikalisch interpretierbare Parameter enthalten, werden sie auch als physikalisch-motiviert bezeichnet. In größeren Höhen, zur Plasmasphäre hin, werden die Chapman-Funktionen mit einem langsam abklingenden exponentiellen Term kombiniert.

Methodisch werden anschließend mittels Sensitivitätsanalysen die sogenannten Schlüsselparameter dieser Mehrschichtendarstellung (engl.: multi-layer approach) festgelegt. Mathematisch werden die raumzeitlichen Variationen dieser Schlüsselparameter durch Reihenentwicklungen in Tensorprodukten aus 1-D lokalisierenden B-Spline Funktionen beschrieben. Die Koeffizientensätze dieser Reihenentwicklungen stellen die unbekannten Parameter dar, die mittels eines geeigneten Schätzverfahrens zu

bestimmen sind. Die Inhomogenität der geographischen Beobachtungsorte mit zum Teil großen Datenlücken stellen eine besondere Herausforderung dar und müssen gesondert, z.B. durch Einführung von Vorinformationen, behandelt werden. Darüber hinaus sei noch erwähnt, dass B-Spline-Funktionen zur Generierung einer Multiskalenrepräsentation (engl: multi-scale representation, MSR) mittels B-Spline Wavelet-Funktionen verwendet werden können. Eine MSR wird üblicherweise zur Datenkompression eingesetzt. Die Anwendung eines solchen Verfahrens ist beispielsweise bei der Nutzung sehr großer Datensätze, wie es bei der Ionosphärenmodellierung nötig ist, sinnvoll.

Der in dieser Arbeit entwickelte Auswerteansatz beinhaltet die verschiedenen stochastischen Informationen für die einzelnen Beobachtungstechniken. Um eine physikalisch sinnvolle Lösung zu erhalten, wird ein beschränktes Optimierungsverfahren in das Modellkonzept integriert.

Das entwickelte Konzept zur regionalen 4-D Elektronendichtemodellierung wird zunächst durch Modellierung der F2-Schicht mithilfe einer Chapman-Funktion untersucht. Die Schlüsselparameter der F2-Schicht, nämlich die maximale Elektronendichte, die zugeordnete Höhe und die Skalenhöhe, werden aus der Kombination von GPS-STEC-Beobachtungen und GPS-EDPs geschätzt. Anschließend wird die vertikale Beschreibung der Elektronendichte innerhalb der Ionosphäre auf den Mehrschichtenansatz erweitert. Die Trennbarkeit der ausgewählten Schlüsselparameter wird mittels simulierter EDPs untersucht. Zusammen mit einer Kreuzvalidierung haben Vergleiche mit externen Datenquellen, einschließlich Ionosonden-Messungen sowie Modellen des vertikalen absoluten Elektronengehalts (engl.: vertical total electron content, VTEC) gezeigt, dass das entwickelte Modellierungskonzept unser Wissen über die Verhältnisse innerhalb der Ionosphäre erweitert.

# Abstract

The ionosphere is the upper part of the Earth's atmosphere where the number of free electrons is sufficient to affect the propagation of the electromagnetic signals. The knowledge of the electron density is not only essential for ionosphere research and space weather studies, but it also plays a crucial role in a wide range of applications, e.g., positioning and navigation. However, the currently existing electron density models usually do not meet the necessary accuracy requirements. They are either purely mathematical (e.g., voxel-based) or only climatological, i.e., they only provide monthly averages (e.g., International Reference Ionosphere (IRI) or NeQuick).

With the rapid development of space-geodetic observation techniques, the quantity and the quality of measurements that can improve ionosphere models have increased drastically in the last decades. In particular, ground-based dual-frequency measurements of the Global Navigation Satellite System (GNSS) have provided highly accurate information about Slant Total Electron Content (STEC), i.e., the integration of the electron density along the signal path between transmitter and receiver.

Space-based GNSS Radio Occultation (RO) measurements allow the computation of Electron Density Profiles (EDPs) with high vertical resolution and global coverage. To improve our knowledge on the ionosphere, the combination of different observation techniques is an important task, as they are characterized by different data distributions, temporal and spatial sampling resolutions as well as distinct sensitivities to ionosphere parameters.

The ionosphere electron density profile exhibits various layers according to their chemical composition, where the so-called F2 layer contains the maximum electron density. The thesis addresses the development of a regional physics-motivated four-dimensional (4-D) electron density model from the combination of various space-geodetic observation techniques. For this purpose, the electron density distributions of the individual layers, i.e., the D, E, F1 and F2 layers are described by Chapman functions. Since Chapman functions contain partly physically interpretable parameters, they are also called physically-motivated. At higher altitudes, toward the plasmasphere, the Chapman functions are combined with a slowly decaying exponential term.

The so-called key parameters of this multi-layer approach are defined by means of sensitivity analysis. The spatiotemporal variations of these key parameters are described mathematically by series expansions in terms of tensor products of 1-D localizing B-spline functions. The sets of coefficients of these series expansions, which represent the unknown parameters, are determined by means of a suitable estimation method. The inhomogeneity of the geographical observation sites with partly large data gaps poses a particular challenge and must be considered appropriately, e.g., by introducing prior information. In addition, it should be noted that B-spline functions can be used to generate a Multi-Scale Representation (MSR) using B-spline

wavelet functions. A MSR is commonly used for data compression. The application of such a procedure is meaningful for handling very large data sets, as is necessary in ionospheric modeling.

The adjustment system developed in this work considers various stochastic information for the individual observation techniques. In order to obtain a physically meaningful solution, a constrained optimization method is integrated into the modeling concept.

The developed regional 4-D electron density modeling concept is firstly investigated by modeling the F2 layer using a Chapman function. The key parameters of the F2 layer, namely the maximum electron density, the corresponding height and the scale height, are estimated from the combination of GPS-STEC observations and GPS-EDPs. Subsequently, the vertical representation of the electron density within the ionosphere is extended by the multi-layer approach. The separability of the selected key parameters is studied using simulated EDPs. Together with a cross-validation, comparisons with external data sources including ionosonde measurements and a Vertical Total Electron Content (VTEC) model have shown the potential of the developed 4-D electron density modeling concept to improve our knowledge of the ionosphere.

# Abbreviations

| | |
|---|---|
| **A-S** | Anti-Spoofing |
| **BDS** | BeiDou System |
| **BDT** | BeiDou Time |
| **C/A-code** | Coarse/Acquisition Code |
| **CCIR** | International Radio Consultative Committee |
| **CDAAC** | COSMIC Data Analysis and Archive Center |
| **CDF** | Cumulative Distribution Function |
| **CGCS2000** | China Geodetic Coordinate System 2000 |
| **CHAMP** | Challenging Minisatellite Payload |
| **CIM** | COMPASS Ionospheric Model |
| **CODE** | Center for Orbit Determination in Europe |
| **CSRSR** | Center for Space and Remote Sensing Research |
| **CWB** | Central Weather Bureau |
| **DCB** | Differential Code Bias |
| **DCBs** | Differential Code Biases |
| **DE** | Dynamic Explorer |
| **DGFI-TUM** | Deutsches Geodätisches Forschungsinstitut der Technischen Universität München |
| **DOD** | Department of Defense |
| **DORIS** | Doppler Orbitography and Radiopositioning Integrated by Satellite |
| **DPB** | Differential Phase Bias |
| **DPBs** | Differential Phase Biases |
| **DPS** | Digisonde Portable Sounder |
| **EC** | European Commission |
| **ECEF** | Earth-Centered Earth-Fixed |

| | |
|---|---|
| **EDP** | Electron Density Profile |
| **EDPs** | Electron Density Profiles |
| **EIA** | Equatorial Ionization Anomaly |
| **EOF** | Empirical Orthogonal Function |
| **ESA** | European Space Agency |
| **ESOC** | European Space Operations Center |
| **ESRO** | European Space Research Organisation |
| **EUV** | Extreme Ultraviolet |
| **F3/C** | FORMOSAT-3/COSMIC |
| **F7/C2** | FORMOSAT-7/COSMIC-2 |
| **FOC** | Full Operational Capability |
| **FWT** | Fast Wavelet Transform |
| **GAIM** | Global Assimilative Ionospheric Model |
| **GAIM-FP** | Global Assimilative Ionospheric Model–Full Physics |
| **GAIM-GM** | Global Assimilative Ionospheric Model–Gauss-Markov |
| **GCPM** | Global Core Plasma Model |
| **GCV** | Generalized Cross Validation |
| **GEO** | Geostationary Earth Orbit |
| **GGOS** | Global Geodetic Observing System |
| **GIM** | Global Ionosphere Map |
| **GIRO** | Global Ionospheric Radio Observatory |
| **GLONASS** | Global'naya Navigatsionnaya Sputnikovaya Sistema |
| **GNSS** | Global Navigation Satellite System |
| **GPS** | Global Positioning System |
| **GPS/MET** | Global Positioning System/Meteorology |
| **GPST** | GPS Time |
| **GRACE** | Gravity Recovery and Climate Experiment |
| **GRACE-FO** | GRACE Follow-On |
| **GST** | Galileo System Time |
| **GTRF** | Galileo Terrestrial Reference Frame |
| **HF** | High Frequency |
| **HPD** | Highest Posterior Density |

| | |
|---|---|
| **IAACs** | Ionosphere Associate Analysis Centers |
| **IFM** | Ionosphere Forecast Model |
| **IGS** | International GNSS Service |
| **IGSO** | Inclined Geosynchronous Satellite Orbit |
| **IOV** | In-Orbit Validation |
| **IPM** | Ionosphere-Plasmasphere Model |
| **IPP** | Ionospheric Pierce Point |
| **IPPs** | Ionospheric Pierce Points |
| **IRI** | International Reference Ionosphere |
| **IRI-Plas** | International Reference Ionosphere Plasmasphere |
| **IRO** | Ionospheric Radio Occultation |
| **ISEE** | International Sun Earth Explorer |
| **ISR** | Incoherent Scatter Radar |
| **ITRF** | International Terrestrial Reference Frame |
| **ITU** | International Telecommunications Union |
| **JPL** | Jet Propulsion Laboratory |
| **KKT** | Karush–Kuhn–Tucker |
| **LEO** | Low Earth Orbiting |
| **LP** | Linear Programming |
| **LPIM** | La Plata Ionospheric Model |
| **MEO** | Medium Earth Orbit |
| **MRR** | Multi-Resolution Representation |
| **MSR** | Multi-Scale Representation |
| **NASA** | National Aeronautics and Space Administration |
| **NAVSTAR** | Navigational Satellite Timing and Ranging |
| **NCU** | National Central University |
| **NLP** | Nonlinear Programming |
| **P-code** | Precision code |
| **PCA** | Principle Component Analysis |
| **PDF** | Probability Density Function |
| **PIM** | Parameterized Ionospheric Model |
| **PNT** | Positioning, Navigation and Timing |

| | |
|---|---|
| **PPP** | Precise Point Positioning |
| **PPS** | Precise Positioning Service |
| **PRN** | Pseudorandom Noise |
| **QP** | Quadratic Programming |
| **RMS** | Root Mean Square |
| **RO** | Radio Occultation |
| **SIRGAS** | Sistema de Referencia Geocéntrico para las Américas |
| **SLM** | Single Layer Model |
| **SMI** | Russian Standard Model of the Ionosphere and Plasmasphere |
| **SPS** | Standard Positioning Service |
| **SQP** | Sequential Quadratic Programming |
| **STEC** | Slant Total Electron Content |
| **SVD** | Singular Value Decomposition |
| **TACC** | Taiwan Analysis Center for COSMIC |
| **TAI** | International Atomic Time |
| **TEC** | Total Electron Content |
| **TECU** | Total Electron Content Unit |
| **TIDs** | Traveling Ionospheric Disturbances |
| **UCAR** | University Corporation for Atmospheric Research |
| **URSI** | International Union of Radio Science |
| **USNO** | United States Naval Observatory |
| **USU** | Utah State University |
| **UT** | Universal Time |
| **UTC** | Coordinated Universal Time |
| **UV** | Ultraviolet |
| **VCE** | Variance Component Estimation |
| **VLBI** | Very Long Baseline Interferometry |
| **VTEC** | Vertical Total Electron Content |
| **WGS-84** | World Geodetic System 1984 |

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Motivation

The ionosphere is the upper region of the atmosphere extending from about 50 km above the Earth's surface and gradually merges into the plasmasphere. Free electrons and ions are present there mainly due to Extreme Ultraviolet (EUV) and X-ray solar radiation. The ionosphere is influenced by space weather[1] and highly variable. All satellite signals travel through the ionosphere and interact with the ionosphere plasma. Signals of a frequency lower than approximately 30 MHz are reflected by the ionosphere, which makes radio communication over long distances possible. Signals of a frequency above 30 MHz pass through the ionosphere and are used in communication with satellites and other spacecraft.

The rapid development of space-geodetic observation techniques has brought out a wide range of applications such as positioning and navigation. It is well known that the interaction of the signals of navigation systems such as the Global Navigation Satellite System (GNSS) with the ionosphere plasma bends the signals and causes a propagation delay, which is one of the main error sources. Dual-frequency GNSS users may eliminate the first-order ionosphere effect by using linear combinations of the measurements at two frequencies using the dispersive properties of the ionosphere. However, single frequency users cannot apply this method and may rely on ionosphere models (see, e.g., Minkwitz *et al.*, 2014) to correct the propagation delay. The propagation delay is depending mainly on the frequency of the signal and the integration of the electron density along the signal path, i.e., the Slant Total Electron Content (STEC). Therefore, the electron density is the most important and relevant ionosphere parameter from the perspective of geodetic applications (cf. Bust and Mitchell, 2008; Schmidt, 2011).

Ionosphere models used for correcting GNSS signal delay can be

- Vertical Total Electron Content (VTEC) models: an ionosphere mapping function is required to convert VTEC to STEC at the Ionospheric Pierce Point (IPP) according to the Single Layer Model (SLM) [Schaer, 1999]. The SLM (cf. Fig. 1.1) is based on the simple assumption that all free electrons are concentrated in an infinite thin shell at a certain height, whereas Ionospheric Pierce Point (IPP) is defined as the intersection between a

[1] Space weather is a term which describes variations in the Sun, solar wind, magnetosphere, ionosphere, and thermosphere, which can influence the performance and reliability of a variety of space-borne and ground-based technological systems and can also endanger human health and safety" [Cannon *et al.*, 2013].

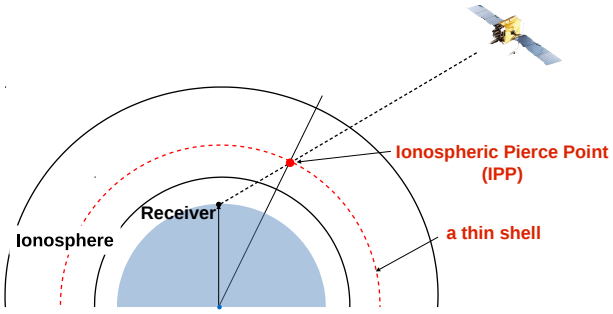



Figure 1.1: Schematic illustration of the SLM–an assumed thin shell of the ionosphere.

ray path and the ionosphere. VTEC models have been widely used and well developed (see, e.g., Hernández-Pajares *et al.*, 2009; Jakowski *et al.*, 2011), but they cannot provide height information of the ionosphere and they are mostly based on the SLM, which is not an adequate representation of the ionosphere.

- electron density models (e.g., NeQuick, Radicella, 2009): based on electron density models, STEC can be calculated by integration of the electron density along the signal path. However, the existing electron density models usually do not meet the necessary accuracy requirements. They are either purely mathematical (e.g., voxel-based) or only climatological (e.g., the empirical models IRI and NeQuick), i.e., they provide mean conditions and cannot describe instantaneous status of the ionosphere.

Consequently, development of a 4-D (space and time) electron density model which takes into account the physics is important for improving a large number of applications, as well as for ionosphere research and space weather studies.

There are several techniques to gather the information of the electron density. Before the era of GNSS, traditional techniques were much more limited [Hernández-Pajares *et al.*, 2011]. Globally distributed ground-based ionosonde/digisonde stations have been used to collect accurate ionosphere information for a long time, where the present ionosondes date back to 1932 [Committee on Solar-Terrestrial Research, 1969]. There are more than 200 ionosondes available worldwide, but only part of them are routinely operational [Liu *et al.*, 2004]. Consequently, the spatial and temporal distribution of the ionosonde measurements is rather limited. Furthermore, ground-based ionosondes/digisondes only probe the ionosphere up to the peak of the F2 layer (the ionosphere is divided into distinct layers and the F2 layer contains the maximum electron density; see Section 2.2). Incoherent Scatter Radar (ISR) can provide information about the entire ionosphere, however, only few radars are in operation worldwide due to their complex and expensive installations [Hunsucker, 1991]. Topside sounder satellites [Bilitza *et al.*, 2003] can also provide topside information from the orbit altitude down to the peak of the F2 layer, however, only a few missions, such as Alouette-1 and -2[1], ISIS-1 and -2[2] and Intercosmos-19 (IK-19)[3] have been carried out. They have collected a huge amount of topside ionograms from the sixties to the eighties, but with limited spatial coverage of various geophysical conditions (e.g., geographic location, diurnal, seasonal and solar activity) [Sibanda and Mckinnell, 2011]. The modern space-geodetic techniques have provided excellent opportunities to explore the ionosphere comprehensively and

[1]Canadian satellites, Alouette-1: 1962–1972, Alouette-2: 1965–1975.
[2]Canadian satellites, ISIS-1: 1969, ISIS-2: 1971.
[3]Russian satellite, 02.1979–04.1982.

cost-effectively (see, e.g., Dettmering *et al.*, 2011b; Jin *et al.*, 2013). On the one hand, the ionosphere may degrade the ground-based GNSS observations for applications of positioning and navigation; on the other hand, these observations provide relatively accurate high-sampling-rate STEC along the signal path from satellite to receiver. With the large amount of GNSS data available, STEC measurements have become one of the most important input for ionosphere modeling. However, the geometry of the observations does not allow to obtain adequate vertical structure of the electron density [García-Fernández *et al.*, 2003]. The Ionospheric Radio Occultation (IRO) measurements made by space-based GNSS receivers on Low Earth Orbiting (LEO) satellites, on the other hand, allow for calculating the vertical distribution of the electron density with a global coverage (e.g., Hajj and Romans, 1998; Dettmering, 2003). Each observation technique has its own unique strengths and weaknesses with respect to the sensitivity to ionosphere parameters, spatiotemporal resolutions and stochastic behavior. Improved techniques and new

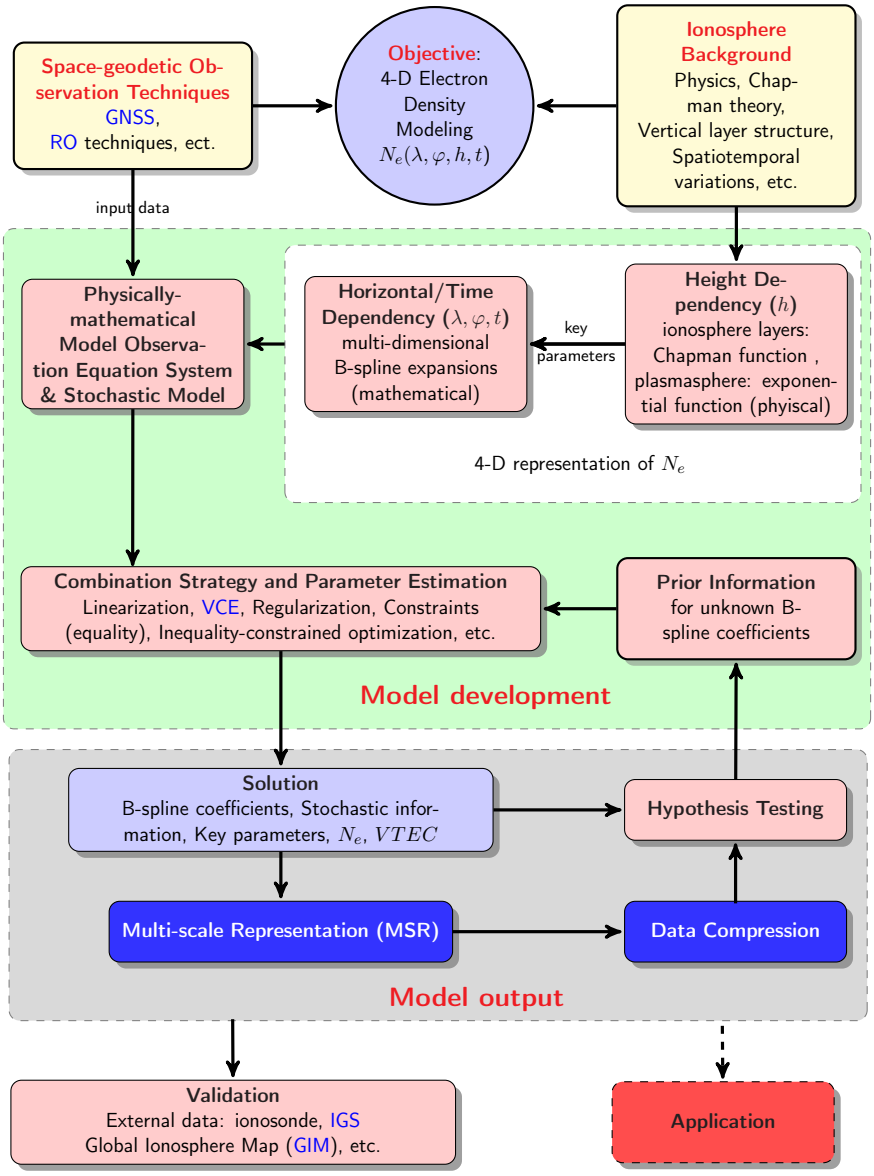



Figure 1.2: The derived 4-D electron density modeling concept.

measurements have brought out the demand of ionosphere models that are able to produce the actual condition of the ionosphere, i.e., switch from ionosphere "climate" to "weather" [Brunini *et al.*, 2004; Nava *et al.*, 2011].

As a result, a consistent combination of different observation techniques to derive a 4-D physics-motivated electron density model with high resolution and high precision has become an important task. This is in agreement with the goal of the Global Geodetic Observing System (GGOS):

> "GGOS integrates different geodetic techniques, different models, different approaches in order to ensure a long-term, precise monitoring of the geodetic observables"[1].

As shown in Fig. 1.2, this thesis will focus on this important task.

## 1.2 State of the art

Since the 1970s, many empirical models, such as the International Reference Ionosphere (IRI) [Bilitza *et al.*, 2011a] and NeQuick as mentioned previously, have been developed based on a large database of traditional ionosphere observations. They are climatological models, which can only provide monthly averages of the ionosphere behaviour. Additionally, several physical models based on ionosphere physics and chemistry have been developed. For example, the Global Assimilative Ionospheric Model (GAIM) [Schunk *et al.*, 2004] is a physics-based data assimilation model which can assimilate different types of data including ground-based GPS observations and space-based RO data. Physical models often require high computational efforts. In order to use a physical model in operational tasks, parametric models have been developed by parameterizing the physical models through simple analytical functions. One example of such models is the Parameterized Ionospheric Model (PIM) [Daniell *et al.*, 1995]; more details can be found in Section 2.5. Various modeling approaches of the electron density have been studied in the last few decades. There are voxel-based methods that divide the ionosphere into many small elements of voxels with constant electron density in each voxel (e.g., Juan *et al.*, 1997; Kuklinski, 1997; Rius *et al.*, 1997; Hernández-Pajares *et al.*, 1999; García-Fernández, 2004; van de Kamp, 2013); there are also function-based methods, where a set of 3-D basis functions has been employed to model the electron density (e.g., Brunini *et al.*, 2004; Liu, 2004; Feltens, 2007; Alizadeh, 2013; Al-Fanek, 2013). Function-based methods for 4-D electron density can be referred to, e.g., Howe *et al.* [1998], Schmidt *et al.* [2008], Zeilhofer [2008], Allain and Mitchell [2010] and Sharifi and Farzaneh [2016].

The vertical structure of the electron density has been modeled by, e.g., Empirical Orthogonal Function (EOF), localizing B-splines or profilers such as the Chapman function [Rishbeth and Garriott, 1969] and the Epstein function [Rawer, 1988], where the Chapman function is derived from a simplified aeronomic theory and frequently used. A single Chapman function has been applied to represent the electron density distribution by, e.g., Feltens [1998] and Brunini *et al.* [2004]; a multi-layer Chapman approach has been used by, e.g., Ching and Chiu [1973] and Tsai *et al.* [2011]; a Chapman function with continuously varying scale height, known as the Vary-Chap function, has been introduced by, e.g., Reinisch *et al.* [2007] and Nsumei *et al.* [2012]. The horizontal variations have been modeled by global basis functions such as spherical harmonic expansions (e.g., Brunini *et al.*, 2004; Tsai *et al.*, 2011; Alizadeh, 2013; Razin, 2016) or localizing basis functions such as B-splines (e.g., Schmidt *et al.*, 2008; Zeilhofer, 2008; Zeilhofer

[1] http://www.ggos.org/.

*et al.*, 2009). Compared to the spherical harmonic expansions, the local B-splines are compactly supported (non-zero only in a certain interval) and more appropriate for handling unevenly distributed observations and data gaps (e.g., Schmidt *et al.*, 2011, 2015). Furthermore, B-splines can construct a certain class of wavelets, which allow the application of a Multi-Scale Representation (MSR) of the signal, e.g., data compression [Schmidt, 2007, 2012]. A more comprehensive overview of different approaches will be given in Chapter 5.

Before signals of GNSS pass through the ionosphere, they first have to travel through the plasmasphere. According to Lunt *et al.* [1999], the plasmaspheric contribution is typically 10% to 30% during daytime, and can reach up to about 50% at night, particularly in winter at low solar activity. Therefore, the electron density modeling approach should also take the plasmaspheric part into account, in order to obtain a realistic description of the ionosphere. The modeling approaches considering both the ionosphere and plasmasphere can be referred to, e.g., Jakowski [2005] and Feltens [2007].

It is known that unrealistic estimates such as negative values of ionosphere parameters, e.g., VTEC or the maximum electron density, may appear in the parameter estimation. In order to obtain realistic estimates, inequality constraints must be incorporated in the estimation process. This is mathematically equivalent to solving a constrained optimization problem where a number of optimization algorithms are available. To the best of the author's knowledge, the optimization algorithms are rarely applied to ionosphere modeling. An inequality constrained least squares method has been applied by Zhang *et al.* [2013] to eliminate negative VTEC values from GPS data. Besides the estimated quantities, the quality of the estimates should also be studied. In case of the inequality constrained estimates, no analytical relationship is present between observations and the unknown parameters. Therefore, the law of error propagation cannot be applied anymore and a symmetric interval around the estimates in terms of the standard deviation is also not sufficient to quantify the uncertainty [Roese-Koerner *et al.*, 2012]. In his work (see also Roese-Koerner, 2015; Roese-Koerner *et al.*, 2015), a Monte Carlo method has been combined with a Quadratic Programming (QP) algorithm to describe the statistical information of the estimated quantities.

## 1.3 Goals and contributions

The objective of this thesis is to develop a regional physics-motivated 4-D electron density model of the ionosphere from the combination of different space-geodetic observation techniques. Compared with global models, regional ones can reach higher spatial and temporal resolutions if input data with high resolution is available in the region. To achieve this goal, different approaches are investigated, combined and adapted within the derived 4-D electron density modeling concept: as shown in the green box of Fig.1.2 with respect to the model development, the vertical layer structures of the ionosphere are described by the **physics-motivated** Chapman functions, which are combined with a slowly decaying exponential term describing the plasmaspheric electron density. Multi-dimensional **B-spline** expansions appropriate for regional modeling and handling inhomogeneous data distribution are introduced for modeling the horizontal and temporal dependencies of ionospheric key parameters. An effective **combination** strategy that takes into account the different stochastic information of the observations is used within the parameter estimation. To overcome data gaps, a **regularization** approach is introduced. An **inequality constrained optimization algorithm** is taken into account. It can be seen in the gray box of Fig.1.2 with respect to the model output that, the modeling approach allows monitoring ionosphere signals

at different resolution levels by generating a **MSR**, with which a **data compression** can be applied.

Firstly, a Chapman function representing the F2 layer combined with a slowly decaying plasmaspheric exponential term following Jakowski [2005] has been used to describe the vertical electron density distribution. The tensor product of the polynomial B-spline functions following Schmidt [2007] and Zeilhofer [2008] are taken to represent the spatiotemporal variations of the corresponding key parameters, namely, the peak density $NmF2$, the peak height $hmF2$ and the Chapman scale height $HF2$. It will be called within this thesis a **F2-layer model**. The fundamentals of the developed F2-layer model have been shown and applied to Electron Density Profiles (EDPs) from GPS RO data by Limberger [2015]. Special contributions of this thesis include:

- incorporation of *stochastic information* of the observations into covariance matrices, in order to assure realistic accuracies,
- investigation of the estimability of the three F2 Chapman key parameters from *a combination of ground-based GPS observations of STEC and EDPs from ionospheric RO data*,
- study of the *statistical information* of the estimated parameters,
- transformation of the key parameters into a *MSR*,
- implementation of *data compression*.

Following the modeling concept of Feltens [2007], the representation of the F2 layer is extended by a more sophisticated and realistic **multi-layer model**, where each of the ionosphere layers is represented by the Chapman function. The number of the key parameters is thus increased significantly, which brings new challenge, i.e., increasing difficulty in parameter estimation. The multi-layer approach has recently received a lot of attention because the representation is physically more realistic. To the best knowledge of the author, separability of the parameters of the multi-layer approach has been rarely investigated. Specific contributions of the work in this thesis are:

- investigation of the *separability of selected parameters of a multi-layer profile function* using simulated EDPs,
- application of the *constrained optimization algorithm* to obtain physically reasonable solutions,
- study of the *quality of the estimates from inequality constrained problem by a Monte Carlo method*.

These investigations have provided a basis for multi-layer electron density modeling.

## 1.4 Outline of the thesis

The organization of this thesis is following Fig. 1.2. The blue circle, i.e., the objective of the thesis, is given in the current **Chapter 1**. This chapter starts from research motivation, which is followed by state of the art and then objectives and contributions of this work.

The required background information of the ionosphere, i.e., the right yellow box, is described in **Chapter 2**. It includes the ionosphere physics with focus on the balance of the ionization

and the formulation of the Chapman layer, the vertical structure (various layers) of the ionosphere, spatiotemporal variations in the ionosphere, how the ionosphere affects the radio wave propagation, an overview of the existing ionosphere electron density models, the definition of the plasmasphere, and an introduction of few selected plasmasphere models.

The available input data for ionosphere modeling, i.e., the left yellow box, is given in **Chapter 3**. It introduces the principles of the two space-geodetic techniques used in this thesis. They are the ground-based GNSS techniques including GPS, Global'naya Navigatsionnaya Sputnikovaya Sistema (GLONASS), BeiDou System (BDS) and Galileo with special emphasis on the GPS, and space GNSS techniques, i.e., GNSS RO techniques including FORMOSAT-3/COSMIC (F3/C), Challenging Minisatellite Payload (CHAMP) and Gravity Recovery and Climate Experiment (GRACE) missions with special focus on the F3/C mission. The ionosphere observables required in this work are described in detail.

The parameter estimation theory required to solve for the model is introduced in **Chapter 4**. It firstly presents the basics of the adjustment system, which includes the concept of linearization, the definition of the Gauss-Markov model, solving unknowns within the Gauss-Markov model without/with equality constraints, rank deficiency and ill-conditioning of the problems, the corresponding solution via regularization, data combination and VCE as well as hypothesis testing. Afterwards, the basics of inequality constrained optimization is introduced and a Nonlinear Programming (NLP) algorithm is described.

The derived modeling concept, i.e., the green box, is addressed in **Chapter 5**. Firstly, an overview of previous works for modeling the height dependency of the electron density is given. Then, the well-known mathematical functions used for modeling the horizontal/temporal variations of ionosphere parameters are introduced. The localizing basis functions used in this work, i.e., B-splines, are given in details. The advantage of the B-splines that can construct B-spline wavelets for generating a MSR and the application of data compression are explained. Based on all information introduced previously, the developed 4-D model is described at last.

The numerical applications of the developed modeling approach, i.e., the gray box, is presented in **Chapter 6**. It is composed of two parts. In the first part, numerical examples of the developed F2-layer model using the combination of ground-based GPS data of STEC and EDPs from space-based GPS RO data are shown. The performance of the developed model is illustrated and discussed by validation, i.e., the bottom left pink box. In the second part, an investigation of the multi-layer model is given. Separability of selected set of key parameters are illustrated by using simulated EDPs. Application of an inequality constrained optimization algorithms is shown in order to obtain physically reasonable solutions. Finally, quality description of the estimates from the inequality constrained optimization by using a Monte Carlo method is illustrated.

The last **Chapter 7** gives some conclusions of this work and suggestions for future work, which highlights the contributions of this dissertation.

# Chapter 2

# Ionosphere Background

## 2.1 Ionosphere physics

The ionosphere is defined as the part of the Earth's upper atmosphere where ions and electrons exist, which affects the propagation of radio waves. It extends from an altitude of approximately 50 km over the Earth and has no well-defined upper boundary [Rishbeth and Garriott, 1969], but the upper boundary is generally regarded to be at an average altitude of roughly 1000 km where the plasmasphere starts (see Section 2.6) [Carpenter and Park, 1973].

### 2.1.1 Balance of ionization

The ionosphere is formed by the ionization of atmospheric gases such as $N_2$ (molecular nitrogen), $O_2$ (molecular oxygen) and O (atomic oxygen). Pairs of ions and electrons are principally produced by solar radiation in the EUV and X-ray parts of the spectrum, at least in low- and mid-latitudes. At high-latitudes and during magnetic storms, however, the production process is a collision between energetic charged particles precipitated into the atmosphere and the neutral molecules. Since the ionosphere contains a significant number of ions and free electrons, it is an electric conductor and a refracting medium for radio waves.

Once ions and electrons are existing, they tend to recombine and to react with gaseous species [Hargreaves, 1992]. Specifically, the free electrons tend to reunite with the positive ions to produce neutral atoms again (called the recombination), and attach themselves to neutral molecules to form negative ions (called the attachment). Furthermore, they can leave a given volume by movement (called the diffusion/drift) [Davies, 1965]. There is a dynamic equilibrium, which means, the net concentration of the free electrons (also called the electron density) is subject to the relative speed of the production and loss processes. Within a cell of unit volume, the rate of change $dN_e/dt$ of the electron density $N_e$ can be formulated by a continuity equation [Rishbeth and Garriott, 1969]

$$\frac{dN_e}{dt} = \overline{q} - \overline{l} - \mathrm{div}(N_e \overline{v}) \tag{2.1}$$

where

| | |
|---|---|
| $t$ | time, |
| $\overline{q}$ | production rate, |
| $\overline{l}$ | loss rate by recombination and attachment, |
| $\overline{v}$ | net drift velocity caused by the transport process, |
| $\mathrm{div}(N_e\overline{v})$ | divergence of the flux $N_e\overline{v}$, representing the loss rate of the electrons due to transport. |

Note that the continuity equation can also be written for either the positive and negative ions, or for any constituent whose concentration is subject to change.

### 2.1.2 The Chapman layer

The production rate $\overline{q}$ of pairs of ions and electrons per unit volume can be written as

$$\overline{q} = I\,\eta\,\overline{\sigma}\,\overline{n} \tag{2.2}$$

where

| | |
|---|---|
| $I$ | intensity of ionizing radiation at a certain height of the atmosphere, |
| $\eta$ | ionizing efficiency, i.e., the fraction of the absorbed radiation that goes into the ionization process, |
| $\overline{\sigma}$ | cross section for absorption of radiation in the gas, |
| $\overline{n}$ | concentration of molecules (or called number density) which are able to be ionized by that radiation. |

Starting from this equation, Sydney Chapman developed in 1931 a formula for the production function, which describes how the production rate $\overline{q}$ varies with the height $h$ and the Sun's zenith angle $\chi$. In the following part, the derivation of the so-called Chapman function, which mainly follows the descriptions of Rishbeth and Garriott [1969], will be given. The derivation is based on the assumptions:

- the radiation is monochromatic, and therefore the intensity of the radiation depends not on the wavelength, but on the height $h$, written as $I(h)$,
- the atmosphere is composed of a single absorbing gas with concentration written as $\overline{n}(h)$,
- the atmosphere is plane and horizontally stratified, i.e., no variations in the horizontal plane,
- the temperature is constant and the scale height $H$ (defined by Eq. (2.13)) is also constant.

Solar radiation attenuates as it travels downwards through the atmosphere. Figure 2.1 shows the geometry of the absorption of the solar radiation in the atmosphere, where $I_\infty$ denotes the unattenuated radiation at the top of the atmosphere before any absorption. For a path element $ds$ along the path of the radiation, the intensity of the radiation varies as

$$dI/ds = -\overline{\sigma}\,\overline{n}\,I \tag{2.3}$$

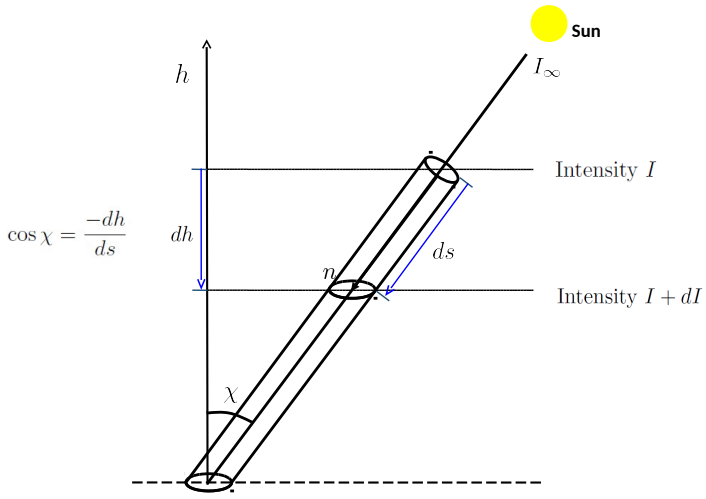



Figure 2.1: Absorption of radiation from gas, which causes the decrease in the intensity $I$ of the radiation (modified from Hargreaves, 1992). The minus sign is due to the decrease of the height along the path of the radiation.

where the minus sign indicates that the solar radiation decreases as the downward path of the radiation increases.

Rearranging Eq. (2.3) as

$$d\tau := -dI/I = \overline{\sigma}\,\overline{n}\,ds \tag{2.4}$$

yields the definition of $d\tau$, namely, an increment of the optical depth $\tau$, which specifies the attenuation of the solar radiation by the atmosphere [Ree, 1989]; see Eq. (2.8) for the mathematical definition of $\tau$. It holds that $\tau = 0$ at the top of the atmosphere where the solar radiation is unattenuated and $\tau \to \infty$ at the bottom of the atmosphere. Integrating the leftmost side and the middle term of Eq. (2.4) from the top $h_\infty$ of the atmosphere to a certain height $h$ yields[1]

$$\begin{aligned} -\int_{h_\infty}^{h} \frac{dI}{I} &= \int_{h_\infty}^{h} d\tau\,, \\ -\ln I\big|_{h_\infty}^{h} &= \tau\big|_{h_\infty}^{h}\,, \\ -(\ln I - \ln I_\infty) &= \tau\,, \\ \ln\left(\frac{I}{I_\infty}\right) &= -\tau\,. \end{aligned} \tag{2.5}$$

Thus, the variation of the intensity $I$ can be written as

$$I = I_\infty \exp(-\tau). \tag{2.6}$$

From the relation that $ds = -dh\sec\chi$[2], we can reformulate Eq. (2.4) as

$$d\tau/dh = \overline{\sigma}\,\overline{n}\,ds/dh = -\overline{\sigma}\,\overline{n}\sec\chi \tag{2.7}$$

[1] We have that $\int dx/x = \ln|x| + c$ and $\int dx = x + c$.
[2] The secant function $\sec\chi = \frac{1}{\cos\chi}$.

which is then integrated with respect to the height

$$\tau(h,\chi) = \int_{h}^{h_\infty} \overline{\sigma}\,\overline{n}(h')\sec\chi\,dh'. \tag{2.8}$$

A comparison of the Eqs. (2.2), (2.6) and (2.8) gives

$$\overline{q} = I_\infty\,\eta\,\overline{\sigma}\,\overline{n}\,\exp\left[-\tau(h,\chi)\right]\,. \tag{2.9}$$

**Hydrostatic equations**

The distribution of a neutral atmospheric gas with the height can be assumed to follow the ideal gas law

$$\overline{p} = \overline{n}\,k_B\,T \tag{2.10}$$

that supposes no intermolecular attractive forces [Dickerson *et al.*, 1979]. Herein $\overline{p}$ [Pa] denotes the pressure, the concentration $\overline{n}$ is in units of $\mathrm{m}^{-3}$, and $k_B$ is the Boltzmann's constant[1].

The hydrostatic or barometric equation of the atmosphere reads [White, 2008]

$$d\overline{p}/dh = -\rho\,g = -\overline{n}\,\overline{m}\,g \tag{2.11}$$

where $\rho$ denotes the density [$\mathrm{kg/m}^3$] and $g$ is a mean value of the gravity acceleration ($g \approx 9.81\,\mathrm{ms}^{-2}$). Let $\overline{m}$ be the molecular mass [kg], such that the relation $\rho = \overline{n}\,\overline{m}$ holds. Dividing this equation by Eq. (2.10) yields

$$\frac{1}{\overline{p}}\frac{d\overline{p}}{dh} = -\frac{\overline{m}\,g}{k_B\,T} \tag{2.12}$$

where the term of the right-hand side relates to the scale height $H$, defined as

$$H := \frac{k_B\,T}{\overline{m}\,g}\,. \tag{2.13}$$

More precisely, $H$ refers to the pressure scale height[2] (compare Section 5.2 for the definitions of the scale height of the electron density). As can be seen from Eq. (2.13), $H$ is dependent on the temperature $T$. The integration of Eq. (2.12) with respect to the height under consideration of Eq. (2.13) gives

$$-\int_{h_0}^{h}\frac{d\overline{p}}{\overline{p}} = -\ln\left(\frac{\overline{p}}{\overline{p}_0}\right) = \int_{h_0}^{h}\left(\frac{dh'}{H}\right) \tag{2.14}$$

where $\overline{p}_0$ denotes the pressure evaluated at a reference height $h_0$. In the following the subscript "0" indicates always the quantities evaluated at the height $h_0$.

Introducing a dimensionless parameter $z$ with an increment $dz$ defined as

$$dz = dh/H \tag{2.15}$$

and let $z = 0$ at some proper reference height $h_0$, the integration of Eq. (2.15) yields the expression of the so-called reduced height $z$,

$$z = \int_{h_0}^{h}\left(\frac{dh'}{H}\right)\,. \tag{2.16}$$

[1] $k_B \approx 1.38\cdot10^{-23}\,\mathrm{J/K} \approx 1.38\cdot10^{-23}\,\mathrm{m^2kgs^{-2}K^{-1}}$.

[2] A scale height represents the vertical distance within which a quantity (here refers to the pressure of the atmosphere) decreases by a factor of the base $e$ of the natural logarithm ($e \approx 2.718$).

Substituting Eq. (2.16) into Eq. (2.14) yields

$$\ln\left(\frac{\overline{p}}{\overline{p}_0}\right) = -z \quad \Longrightarrow \quad \frac{\overline{p}}{\overline{p}_0} = \exp(-z)\,. \tag{2.17}$$

Then, replacing $\overline{p}$ in this equation by the ideal gas law (2.10), we obtain the following general formulation for the variation of pressure, concentration and temperature with $z$

$$\frac{\overline{p}}{\overline{p}_0} = \frac{\overline{n}\,T}{\overline{n}_0\,T_0} = \exp(-z)\,. \tag{2.18}$$

If a single gas is considered and the variation of gravity with height is neglected, the total number of particles in a column of a unit cross section above the reference height $h_0$ is obtained

$$\begin{aligned}\int_{h_0}^{h_\infty} \overline{n}\,dh \overbrace{=}^{\text{applying Eq. (2.15)}} \int_0^\infty \overline{n}\,H\,dz \overbrace{=}^{\text{applying Eq. (2.13)}} \int_0^\infty \overline{n}\,\frac{k_B\,T}{\overline{m}\,g}\,dz \\ \overbrace{=}^{\text{applying Eq. (2.10)}} \frac{1}{\overline{m}\,g}\int_0^\infty \overline{p}\,dz \overbrace{=}^{\text{applying Eq. (2.18)}} \frac{\overline{p}_0}{\overline{m}\,g}\int_0^\infty \exp(-z)\,dz = \frac{\overline{p}_0}{\overline{m}\,g}\underbrace{[-\exp(-z)]\,|_0^\infty}_{=1} \\ = \frac{\overline{p}_0}{\overline{m}\,g} \overbrace{=}^{\text{applying Eq. (2.10)}} \frac{\overline{n}_0\,k_B\,T_0}{\overline{m}\,g} \overbrace{=}^{\text{applying Eq. (2.13)}} \overline{n}_0\,H_0\end{aligned} \tag{2.19}$$

which can be generalized that the integrated content of a column of gas with unit cross section, above any height $h$ is $\overline{n}(h)\,H(h)$.

Now we come back to Eq. (2.8), assume a plane Earth, namely that $\sec\chi$ does not change along the path, then Eq. (2.8) can be reformulated under consideration of Eq. (2.19) as

$$\tau(h,\chi) = \int_h^{h_\infty} \overline{\sigma}\,\overline{n}(h')\sec\chi\,dh' = \overline{\sigma}\sec\chi\int_h^{h_\infty}\overline{n}(h')\,dh' = \overline{\sigma}\sec\chi\,\overline{n}(h)H(h)\,. \tag{2.20}$$

If the reference height $h_0$, from where the reduced height $z$ is measured, is selected where the optical depth $\tau = 1$ when the Sun is in zenith direction (i.e., $\chi = 0°$), then Eq. (2.20) becomes

$$1 = \overline{\sigma}\,\overline{n}_0\,H_0. \tag{2.21}$$

Eliminating $T$ in Eq. (2.18) by Eq. (2.13) gives

$$\exp(-z) = \frac{\overline{p}}{\overline{p}_0} = \frac{\overline{n}\,H}{\overline{n}_0\,H_0} \tag{2.22}$$

and substituting this equation together with Eq. (2.21) into Eq. (2.20) yields

$$\tau(h,\chi) = \overline{\sigma}\sec\chi\,\overline{n}_0\,H_0\,\exp(-z) = \exp(-z)\sec\chi\,. \tag{2.23}$$

Then, the production function (2.9), can be formulated in terms of the reduced height $z$ using the Eqs. (2.20) and (2.23) as

$$\begin{aligned}\overline{q}(z,\chi) &= I_\infty\eta\,\frac{\tau(h,\chi)}{H(h)\sec\chi}\,\exp\left[-\exp(-z)\sec\chi\right] = I_\infty\eta\,\frac{\exp(-z)\sec\chi}{H(h)\sec\chi}\,\exp\left[-\exp(-z)\sec\chi\right] \\ &= \frac{I_\infty\eta}{H(h)}\,\exp(-z)\,\exp\left[-\exp(-z)\sec\chi\right] = \frac{\eta\,I_\infty}{e\,H(z)}\exp\left[1-z-\exp(-z)\sec\chi\right]\,.\end{aligned} \tag{2.24}$$

Considering the constant $H$ in the Chapman theory, we obtain the Chapman production function

$$\overline{q}(z,\chi) = q_0 \exp[1 - z - \exp(-z)\sec\chi] \quad \text{with} \quad q_0 = \frac{\eta I_\infty}{eH} \tag{2.25}$$

where $q_0$ is constant.

Figure 2.2 shows the ratio $\overline{q}(z,\chi)/\overline{q}_0$ versus the reduced height $z$ for different solar zenith angles $\chi$. As can be seen, the maximum rate of ionization decreases with increasing angles $\chi$, but the height where the maximum rate occurs increases. The ratio $\overline{q}(z,\chi)/\overline{q}_0$ reaches the maximum value when $z = 0$ and $\chi = 0°$.

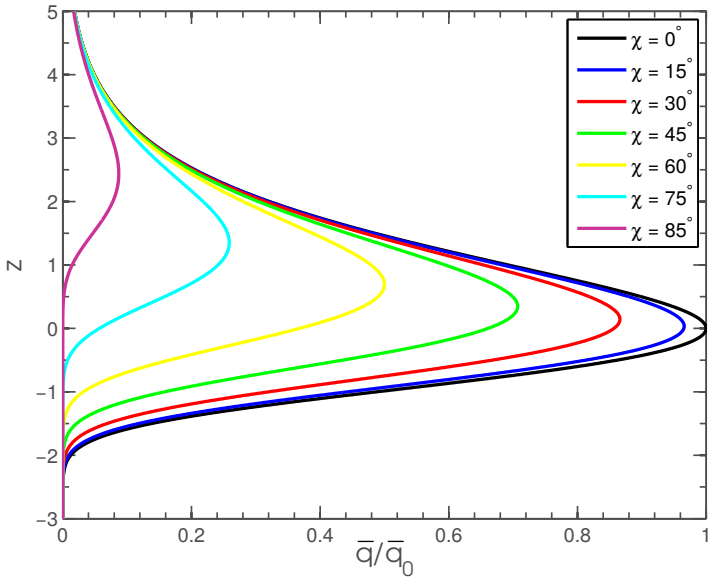



Figure 2.2: The Chapman production ratio $\overline{q}/\overline{q}_0$ with respect to the reduced height $z$ for different Sun's zenith angles $\chi$, based on Eq. (2.25).

It can be expected from Eq. (2.2) that the production rate $\overline{q}$ reaches its maximum at a certain height, since the gas concentration $\overline{n}$ decreases with height, whereas the intensity $I$ of radiation increases with height. In order to find the peak of $\overline{q}$, we can take the logarithm of $\overline{q}$ in Eq. (2.9) and solve $d\left(\ln\overline{q}\right)/dh = 0$, i.e.,

$$\frac{d\left(\ln\overline{q}\right)}{dh} = \frac{d\left\{\ln\left[I_\infty \eta\,\overline{\sigma}\,\overline{n}\,\exp\left(-\tau(h,\chi)\right)\right]\right\}}{dh} = 0\,. \tag{2.26}$$

When $I_\infty \eta\,\overline{\sigma}$ is assumed to be a constant, this equation becomes

$$\frac{d\left[\ln\left(\overline{n}\,\exp(-\tau)\right)\right]}{dh} = \frac{1}{\overline{n}\,\exp(-\tau)}\left[\exp(-\tau)\,\frac{d\overline{n}}{dh} - \overline{n}\,\exp(-\tau)\,\frac{d\tau}{dh}\right] = 0\,. \tag{2.27}$$

Therefore, the peak of $\overline{q}$ occurs where the relation

$$\frac{1}{\overline{n}}\frac{d\overline{n}}{dh} = \frac{d\tau}{dh} \tag{2.28}$$

is satisfied, which indicates that the production reaches a maximum at the level where the increasing gas concentration $\overline{n}$ and the increasing attenuation of the radiation as measured by $\tau$, with decreasing height, are compensated.

The combination of the Eqs. (2.10), (2.12) and (2.13) yields

$$-\frac{1}{H}=\frac{1}{\overline{p}}\frac{d\overline{p}}{dh}=\frac{1}{\overline{n}k_BT}\left(k_BT\frac{d\overline{n}}{dh}+\overline{n}k_B\frac{dT}{dh}\right)=\frac{1}{\overline{n}}\frac{d\overline{n}}{dh}+\frac{1}{T}\frac{dT}{dh}=\frac{1}{\overline{n}}\frac{d\overline{n}}{dh}+\frac{1}{H}\frac{dH}{dh} \tag{2.29}$$

which is then substituted into Eq. (2.28), the condition

$$1+\frac{dH}{dh}=-H\frac{d\tau}{dh} \tag{2.30}$$

can be obtained. Substituting Eq. (2.7) into the above equation under consideration of Eq. (2.20) yields

$$1+\frac{dH}{dh}=\overline{\sigma}\,\overline{n}\,H\sec\chi=\tau. \tag{2.31}$$

Then, we can find the peak value of the production, denoted as $\overline{q}_m$,

$$\overline{q}_m=\frac{I_\infty\eta\cos\chi(1+dH/dh)}{H}\exp\left[-(1+dH/dh)\right] \tag{2.32}$$

by substituting Eq. (2.31) into Eq. (2.9). Under the assumption that the scale height $H$ is constant, i.e., $H$ does not depend on $h$, we can obtain $dH/dh=0$. Then, the production peak occurs at the level where the optical depth $\tau$ is $\tau=1$ according to Eq. (2.31). Applying the definition of $\tau$ in Eq. (2.23), the location of the production peak is related to the reduced height $z_m$, written as

$$z_m=\ln\sec\chi\,. \tag{2.33}$$

The corresponding maximum production rate Eq. (2.32) can also be simplified as

$$\overline{q}_m=\frac{\eta I_\infty\cos\chi}{eH}\,. \tag{2.34}$$

The production rate $\overline{q}(z,\chi)$ can also be expressed in terms of the production peak $\overline{q}_m$ instead of $\overline{q}_0$. To do this, let $h_m$ denote the height where the production peak occurs. Since $H$ is assumed to be constant, Eq. (2.16) can be reformulated as

$$z=\frac{h-h_0}{H}\,. \tag{2.35}$$

Therefore, Eq. (2.33) can be written as

$$z_m=\frac{h_m-h_0}{H}=\ln\sec\chi \tag{2.36}$$

which yields

$$h_m=h_0+H\ln\sec\chi. \tag{2.37}$$

Let $h_m$ be chosen as the reference height, namely, $\overline{z}=0$ at the height $h_m$, we obtain

$$\overline{z}=\frac{h-h_m}{H}=\frac{h-(h_0+H\ln\sec\chi)}{H}=\frac{h-h_0}{H}-\ln\sec\chi=z-\ln\sec\chi \tag{2.38}$$

under consideration of the Eqs. (2.35) and (2.37). Substituting Eq. (2.38) into Eq. (2.25) yields

$$\begin{aligned}\overline{q}(\overline{z},\chi) &= \overline{q}_0 \exp\left\{1-(\overline{z}+\ln\sec\chi)-\exp\left[-(\overline{z}+\ln\sec\chi)\right]\sec\chi\right\}\\ &= \overline{q}_0 \exp[1-\overline{z}-\ln\sec\chi-\exp(-\overline{z})\exp(-\ln\sec\chi)\sec\chi]\\ &= \overline{q}_0 \exp(-\ln\sec\chi)\exp[1-\overline{z}-\exp(-\overline{z})\underbrace{\exp(-\ln\sec\chi)\sec\chi}_{\frac{1}{\sec\chi}\sec\chi=1}]\\ &= \overline{q}_0 \cos\chi \exp\left[1-\overline{z}+\exp(-\overline{z})\right].\end{aligned} \tag{2.39}$$

Based on the Eqs. (2.34) and (2.25), the relation

$$\overline{q}_m = \overline{q}_0 \cos\chi \tag{2.40}$$

holds. Hence Eq. (2.39) can be reformulated as

$$\overline{q}(\overline{z}) = \overline{q}_m \exp[1-\overline{z}+\exp(-\overline{z})] \tag{2.41}$$

which is an alternative form to Eq. (2.25), expressed in terms of the production peak. This expression indicates that the shape of the Chapman production function does not depend on the zenith angles $\chi$. However, the amplitude is scaled by a factor of $\cos\chi$ and the corresponding peak height is shifted (see Eq. (2.37)). Therefore, these two quantities vary with the day time.

Once the ion production is obtained, the electron density can then be derived based on Eq. (2.1). In the lower ionosphere (below 200 km) transport processes are not very important and can be neglected. Then, Eq. (2.1) becomes

$$\frac{dN_e}{dt} = \overline{q} - \overline{l}\,. \tag{2.42}$$

The photochemical equilibrium equation [Rasmussen *et al.*, 1988]

$$\overline{q} = \overline{l} \tag{2.43}$$

with $dN_e/dt = 0$ can be assumed, which means, the production rate is equal to the loss rate. This condition is generally adequate for D, E and F1 layer at daytime (see Section 2.2).

Assume that there are few negative ions compared with the electron concentration, the electrons are lost mainly by recombining with positive ions, i.e., $X^+ + e^- \rightarrow X$ with $X^+$ denoting the positive ions and $e^-$ a free electron. The loss rate $\overline{l}$ can then be written as [Davies, 1965]

$$\overline{l} = \alpha N_e^2 \tag{2.44}$$

where $\alpha$ is called the recombination coefficient. When the number of neutral molecules is significantly larger than the number of electrons, $\overline{l}$ is dominated by the process of attachment, written as, $M + e^- \rightarrow M^-$ with M denoting the neutral species. The loss rate $\overline{l}$ is then linear with $N_e$ [Davies, 1965],

$$\overline{l} = \beta N_e \tag{2.45}$$

where $\beta$ is called the attachment coefficient; see, also e.g., Ratcliffe [1960] for detailed discussions of this issue.

Now, taking the Chapman production function from the Eqs. (2.25) or (2.41), the following electron density distribution can be obtained by applying the Eqs. (2.43) and (2.44)

$$\begin{cases} N_e(z) = N_0 \exp\left\{\frac{1}{2}\left[1-z-\exp(-z)\sec\chi\right]\right\} & \text{with} \quad N_0 = (\overline{q}_0/\alpha)^{1/2} \\ N_e(z) = N_m \exp\left[\frac{1}{2}\left(1-z-\exp(-z)\right)\right] & \text{with} \quad N_m = (\overline{q}_m/\alpha)^{1/2} \end{cases} \tag{2.46}$$

or applying the Eqs. (2.43) and (2.45)

$$\begin{cases} N_e(z) = N_0 \exp\left[1 - z - \exp(-z)\sec\chi\right] & \text{with} \quad N_0 = \overline{q}_0/\beta \\ N_e(z) = N_m \exp\left[1 - z - \exp(-z)\right] & \text{with} \quad N_m = \overline{q}_m/\beta \end{cases} . \tag{2.47}$$

The distribution described by Eq. (2.46) is called the "Chapman alpha", "$\alpha$-Chapman", or simply a "Chapman" layer, and that described by Eq. (2.47) is called the "$\beta$-Chapman" layer. The Chapman theory is very useful since it reflects fundamentals of ionosphere formation and radiation absorption. It provides an invaluable reference point for interpreting observations and a starting point for ionosphere theory, although it is based on many simplifying assumptions [Hargreaves, 1992].

## 2.2 Vertical structure of the ionosphere

The ionosphere is commonly subdivided into distinct regions or layers according to different ion compositions and the wavelengths of solar radiation that are most absorbed in that regions/layers. The main layers are designated as D, E, F1 and F2 from bottom to top or regions as D, E, and F. The lower ionosphere is dominated by photochemical processes; the mid ionosphere is an area in which ionization-recombination processes together with thermal and dynamic processes play important roles; the upper ionosphere is mainly influenced by the transport processes. Actually, the real heights of the ionosphere layers vary with solar zenith angle $\chi$, time of day, seasons, solar cycles and solar activity [Zolesi and Cander, 2014]. As a result, there is not an unified designation of altitude ranges for different layers and, thus, the designations differ in the literature. The breakdown of the ionosphere into layers used in this work is shown in Table 2.1. The different layers are characterized by their critical frequencies $fo_i$ with $i \in \{D, E, F1, F2\}$, the maximum electron density value $Nm_i$ of that layer, and the peak heights $hm_i$ where the maximum electron densities are located. The critical frequency of an ionosphere layer refers to the minimum frequency of an electromagnetic wave that can penetrate that layer without being reflected back to the ground. The different regions/layers have various characteristics that affect radio communications in different ways. In the following, more details about different regions/layers will be given. Figure 2.3 shows typical day and night profiles of electron density in the mid-latitude ionosphere during the maximum and minimum sunspot (see Eq. (2.48) later), where the various layers are clearly visible.

### 2.2.1 D region

The D region ranges from an altitude of about 50 km to 90 km and it is decreased extremely after sunset. The typical electron density values are around $10^2$ to $10^3$ el/cm$^3$. The lower part of the D region is sometimes called a C layer [Hargreaves, 1992], which refers to the ionization produced by cosmic rays, and it can cause a sudden increase of electron density at sunrise and a rapid decrease at sunset [Feltens, 1998, and references therein].

Table 2.1: The main ionosphere regions.

| Layers | D | E | F1 | F2 |
|---|---|---|---|---|
| Height range [km] | 50–90 | 90–140 | 140–200 | 200–1000 |

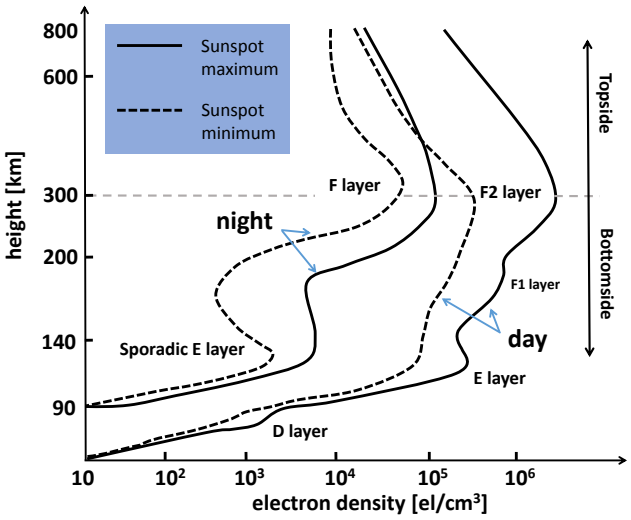



Figure 2.3: Typical day and night EDPs in the mid-latitude ionosphere [Hargreaves, 1992, and references therein]. el/cm$^3$ is an abbreviation for electrons/cm$^3$.

The number of free electrons in the D region is much smaller than the number of molecules. The collision frequency [Nicolet, 1953] between electrons and molecules is relatively high, and therefore the D region mainly has the effect of absorbing or attenuating radio communication signals. The neutral composition within the D region is virtually the same as that at the ground. However, minor constituents such as NO (nitric oxide), $O_3$ (ozone) and unstable $O_2$ (molecular oxygen, see Slanger and Copeland, 2003) play important roles. The major positive ion component is $NO^+$. The D region is the most complicated part of the ionosphere from the chemical point of view [Hargreaves, 1992]. The reason is, on the one hand, that the pressure is relatively high which causes minor and major species to be important in the photochemical reactions; on the other hand, the reason is that several sources of ionization cause ion production. These sources include: the Lyman-$\alpha$ line of the solar spectrum (ionizes the minor species NO); EUV spectrum within a certain wavelength range (ionizes unstable $O_2$); EUV radiation (also ionizes $O_2$ and $N_2$ (molecular nitrogen)); hard X-rays (ionize all constituents, affect mostly major species $O_2$ and $N_2$); galactic cosmic rays (ionize all atmospheric constituents and, thus, affect the whole atmosphere, it is a major ionization source in the lower D region); energetic particles from the Sun or of auroral origin (ionize the D region at high-latitudes).

### 2.2.2 E region

The E region ranges from an altitude of about 90 km to 140 km and includes mainly the E layer (regular) and the so-called sporadic E layer.

**E layer**

The E layer does not vanish completely at night, but is weakly ionized with electron density values about $5 \cdot 10^3$ el/cm$^3$ [Hunsucker and Hargreaves, 2003], whereas the typical electron density values at daytime are about a few $10^5$ el/cm$^3$.

The primary positive ions within the region are $NO^+$ and $O_2^+$; the secondary positive ion components are $N_2^+$ and $O^+$. The main source of ionization is the X-ray emission and the EUV

radiation. The E layer is generally well described by the $\alpha$-Chapman layer (e.g., Rishbeth and Garriott, 1969; Hunsucker and Hargreaves, 2003).

**Sporadic E layer**
The sporadic E layer, often designated as Es, is a highly variable layer in terms of spatial, diurnal and seasonal patterns, which is usually considered to be independent of the regular E layer. In low- and mid-latitudes the Es layer happens mostly during the day and prevalently during the summer, while at high-latitudes the layer is more likely to happen at night and is frequently related to aurora [Zolesi and Cander, 2014]. The most important feature is that the magnitude of the Es layer can be similar to that of the F region.

### 2.2.3 F region

The F region ranges from an altitude of about 140 km to 1000 km. A typical ionosphere EDP reaches the highest density in this region. The region is mainly responsible for the reflection of radio waves in High Frequency (HF) communication and broadcasting. Due to the complex physical mechanisms involved in its formation, the F region is split into an F1 and an F2 layer as a result of solar radiation. The F1 layer is only present during the day and disappears for high solar zenith angles.

**F1 layer**
The F1 layer lies between 140 km and 200 km and it merges with the F2 layer a few hours after sunset. The typical electron density values are about several $10^5$ to $10^6$ el/cm$^3$. The F1 layer is more striking during the summer than during the winter, at high solar activity, and during ionosphere storms [Davies, 1965]. The primary positive ions are $NO^+$ and $O_2^+$; $N^+$ and $O^+$ are the secondary positive ion components. The primary source of ionization within this layer is the EUV light, and the F1 layer is considered as the most heavily absorbed part of the solar spectrum [Hargreaves, 1992]. This layer also behaves like $\alpha$-Chapman layer [Hunsucker and Hargreaves, 2003]. However, it seldom exists as a distinct peak but a ledge; furthermore, the ledge does not always exist.

**F2 layer**
The F2 layer is present during day and night. The maximum electron concentration occurs in this layer with a value up to several $10^6$ el/cm$^3$. The peak is typically located between 300 and 500 km. This layer is highly variable with timescales ranging from the 11 years of a solar cycle or even longer, to a few seconds during the strong interactions with the plasmasphere above [Zolesi and Cander, 2014].

The major ion in the F2 layer is $O^+$ and the secondary ion components are $H^+$ (protons) and $He^+$ (helium ion). Ionization in the F region is mainly caused from the process of photo-ionization by EUV radiation. The F2 layer is different from the photochemically dominated lower layers and the diffusion dominated upper part of the topside ionosphere (topside refers to the region above $hmF2$). For layers such as D, E, and F1, there are bands of radiation (regions of the spectrum) that are strongly absorbed and thus the maximum ionization rate is produced. However, no band of radiation can produce a maximum ionization rate at heights above 180 km. The F2 layer is produced by the same band of radiation that produces the F1 layer [Barclay, 2003] and is

formed as an upward extension of the F1 layer due to the height variation of the recombination rate. Because the electron recombination rate falls off more quickly than the production rate, the electron density still increases with height above the F1 layer. To explain the peak, diffusion takes over the role with increasing heights because of the decreasing air density, where the production and loss processes become less important. The diffusion process leads to an electron density distribution that decays with height (see Eq. (2.1)), and the maximum electron concentration is thus reached when these two kinds of processes are equally important; see Hargreaves [1992] for further details. According to Dieminger *et al.* [1996], the EDP of the F2 layer is approximately a Chapman layer even if it includes diffusion processes which are not considered by the Chapman theory.

The striking characteristic of this layer are the two crests of peak electron densities $NmF2$ lying in the vicinity of $\pm 20°$ geomagnetic latitude, called the Equatorial Ionization Anomaly (EIA) (see Section 2.3.1 for more details).

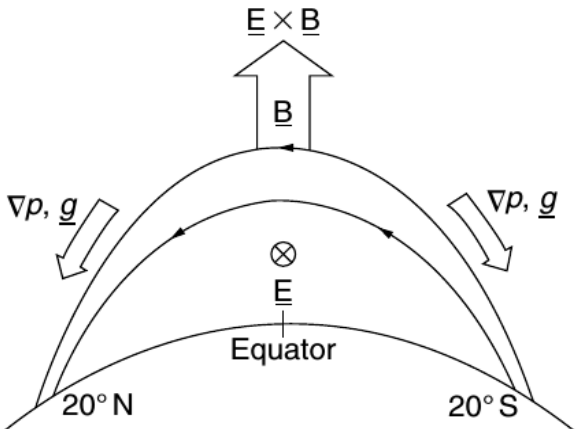
## 2.3 Spatiotemporal variations in the ionosphere

Since the ionosphere is ionized mainly by the Sun, the ionosphere in particular the electron density exhibits spatial (geographical and geomagnetic) and temporal variations (diurnal, seasonal and solar cycle). Here, only regular variations are introduced; information about ionosphere disturbances such as ionosphere storms, geomagnetic storms, Traveling Ionospheric Disturbances (TIDs) and ionosphere scintillation can be found in, e.g., Rishbeth and Garriott [1969] and Zolesi and Cander [2014]. The descriptions of the variations below are mainly following Davies [1965].

### 2.3.1 Latitudinal variations

Usually the latitudinal variation is referred to three major geographic regions with rather different features based on geomagnetic latitude, namely, the low-latitude ($0° \leq \theta < 20°$ on each side of the magnetic equator with $\theta$ denoting the geomagnetic latitude), the mid-latitude ($20° \leq \theta < 60°$ on each side of the magnetic equator) and the high-latitude ($60° \leq \theta < 90°$ on each side of the magnetic equator) region.

**Low-latitude region**
This region is characterized by the highest values of $NmF2$ and the integrated quantity Vertical Total Electron Content (VTEC); see Section 2.4.2 for details about VTEC. At low-latitudes the geomagnetic field $\mathbf{B}$ is nearly horizontal at the magnetic equator, the ionosphere there is strongly influenced by electromagnetic forces, which leads to vertical drifts of ionization. As shown in Fig. 2.4, during the daytime the eastward electric field $\mathbf{E}$ at the equator leads to an upward vertical plasma drift $\mathbf{E} \times \mathbf{B}$. The plasma is lifted up to high altitudes until the gravity ($\boldsymbol{g}$) and pressure gradient forces ($\nabla \boldsymbol{p}$) are large enough to diffuse downward along the magnetic field lines. This phenomenon is called "fountain effect". As a consequence, the electron densities are enhanced to two maxima at geomagnetic latitudes 15° to 20° on both sides of the magnetic equator, forming two crests, and the electron densities fall to a minimum (trough) near the magnetic equator. This phenomenon is called the "EIA" or the "Appleton anomaly".




(a)

Figure 2.4: A schema of the fountain effect [Kelley, 2009].

**Mid-latitude region**
Among the three regions, the mid-latitude zone is generally the least complex and variable as well as the best understood due to a relatively large number of ionosphere observations. The ionization there is caused almost completely by energetic Ultraviolet (UV) and X-ray emissions from the Sun.

**High-latitude region**
At high-latitudes the geomagnetic field is approximately vertical, which leads to that the ionosphere is much more complex than that in either the mid-latitude or low-latitude regions. Through the magnetic field lines, the high-latitude regions are connected to the outer part of the magnetosphere which is driven by the solar wind, whereas the ionosphere at mid-latitudes is connected to the inner magnetosphere that essentially rotates with the Earth [Hunsucker and Hargreaves, 2003]. The ionization at high-latitudes is dominated by energetic charged particles and solar EUV radiation. The X-rays are relatively weak due to the low elevation of the Sun.

An important feature in mid- to high-latitudes (between about 50° and 70° of geomagnetic latitude) is the "main" trough (Muldrew, 1965, or called the "mid-latitude" trough), which refers to depletion of the electron density in the F region. It happens mainly at night and more strongly in winter than in summer in the northern hemisphere. The electron density inside the trough can be drastically decreased, e.g., by as much as a factor of 2 at an altitude of 1000 km and an order of magnitude at $hmF2$ [Dieminger *et al.*, 1996]. For further details about the high-latitude ionosphere, the interested reader may refer to, e.g., Hunsucker and Hargreaves [2003] or Rodger *et al.* [1992].

### 2.3.2 Temporal variations

The ionosphere generally exhibits diurnal, seasonal, solar activity variations. Because of the Earth rotation, the relative position of the Sun and the Earth changes with time, which causes that the ionosphere exhibits a daily variation. Since ionization is mainly driven by the Sun, the ionosphere follows a seasonal variation and a long-period variation of the Sun's activity, which has a main periodicity of about 11 years. The solar activity is usually described by the Wolf (or

Zürich) sunspot[1] number $R$, which is calculated by [Hargreaves, 1992]

$$R = K(10\,G + S) \tag{2.48}$$

where $G$ denotes the number of disturbed regions (either a single spot or groups of spots), $S$ is the total number of individual spots seen, and $K$ is a constant depending on the sensitivity of the observing equipment. The solar activity can also be indicated by the solar radio flux $F10.7$, which is defined as the radio emission from the Sun at the wavelength of 10.7 cm.

There are also temporal anomalies in the ionosphere, e.g., "winter anomaly"; details can be found in e.g., Hargreaves [1992].

## 2.4 Ionospheric effects on radio wave propagation

### 2.4.1 Refractive index

The wave propagation velocity in a medium depends on the refractive index $n_r$ of that medium, which is defined as

$$n_r = \frac{c}{v} \tag{2.49}$$

where $v$ is the propagation velocity in the medium and $c$ the speed of light in vacuum.

In a non-dispersive medium, the phase velocity (the speed of propagation of a pure (unmodulated) wave) is independent of frequency, and the phase velocity and the group velocity (the speed of propagation of a modulated carrier wave, i.e., a group of waves of different frequencies are superposed centered on the carrier frequency) are identical; in a dispersive medium, the propagation velocity of an electromagnetic wave depends on its frequency, and the phase and group velocities are different. A relationship between the group refractive index $n_{gr}$ and the phase refractive index $n_{ph}$ of a dispersive medium is given by [Davies, 1990]

$$n_{gr} = n_{ph} + f\,\frac{dn_{ph}}{df} \tag{2.50}$$

where $f$ denotes the frequency of a wave.

The ionosphere is a dispersive medium. The phase refractive index $n_{ion}^{ph}$ of the ionosphere, applicable to the carrier phase observations of a GNSS signal (see Section 3.1.2), can be approximated by a series expansion of the Appleton-Hartree formula [Budden, 1985] truncated after the quadratic term, i.e.,

$$n_{ion}^{ph} = 1 - \frac{40.3N_e}{f^2}\,. \tag{2.51}$$

This equation includes the first-order ionosphere effect, i.e., the most dominant part.

Substituting Eq. (2.51) into Eq. (2.50) yields the expression of the ionosphere group refractive index $n_{ion}^{gr}$,

$$n_{ion}^{gr} = 1 + \frac{40.3N_e}{f^2} \tag{2.52}$$

which is applicable to the pseudorange observations of a GNSS signal (see Section 3.1.2).

[1] Sunspots are areas that appear visibly as dark spots, which happen because they have relatively lower temperature than the surrounding solar photosphere.

Since $N_e$ is always positive, $n_{ion}^{ph}$ is smaller than 1. Therefore, a phase travels faster than the speed of light in vacuum due to Eq. (2.49), which causes the phase advance; in contrary, $n_{ion}^{gr}$ is larger than 1 which leads to the group delay [Hofmann-Wellenhof *et al.*, 2008].

### 2.4.2 Ionospheric delay

The range between a transmitter and a receiver is defined as

$$s = c \int_D \frac{ds}{v} = \int_D n_r \, ds \tag{2.53}$$

where $D$ denotes actual signal path with the differential length element $ds$ along the signal path, $v$ is the actual signal propagation velocity at the given location. Assuming $n_r = 1$, the geometric range $s_0$ between the emitter and the receiver is obtained analogously

$$s_0 = \int_{D_0} ds_0 \tag{2.54}$$

where the integral is performed along the geometrical path $D_0$. Then, the difference between measured and geometric range, called the ionosphere range error (or called the ionosphere refraction) $\Delta\rho_{ion}$ can be defined as

$$\Delta\rho_{ion} = \int_D n_r \, ds - \int_{D_0} ds_0 \, . \tag{2.55}$$

Assuming that the geometrical path $D_0$ is the actual signal path $D$ (i.e., curvature of the signal path is neglected), the substitution of the Eqs. (2.51) and (2.52) into Eq. (2.55) yields

$$\Delta\rho_{ion}^{ph} = -\int_D \frac{40.3 N_e}{f^2} \, ds \tag{2.56}$$

and

$$\Delta\rho_{ion}^{gr} = \int_D \frac{40.3 N_e}{f^2} \, ds \tag{2.57}$$

respectively, where $\Delta\rho_{ion}^{ph}$ and $\Delta\rho_{ion}^{gr}$ are phase advance and group delay, respectively. Following Jakowski *et al.* [2004], the ray bending error caused from straight line of sight propagation assumption is in the magnitude of millimeter to centimeter range for ground-based GNSS measurements.

The integration of the electron density along the signal path is defined as Total Electron Content (TEC), which is given by

$$TEC = \int_D N_e \, ds \, . \tag{2.58}$$

TEC is measured in Total Electron Content Unit (TECU) with $1\,\text{TECU} = 10^{16}\,\text{el/m}^2$. The TEC in the slant direction of signal transmitting is called STEC whereas the one referring to the zenith direction is called VTEC. GPS generally provides information of STEC whereas satellite altimetry gives measurements of VTEC. For TEC mapping using space-geodetic techniques, VTEC is often modeled and can be obtained from STEC by applying the so-called ionosphere mapping function (see e.g., Schaer, 1999). However, VTEC modeling is not within the interests of this thesis.

The substitution of Eq. (2.58) into the Eqs. (2.56) and (2.57) gives an alternative form to the phase advance

$$\Delta\rho_{ion}^{ph} = -40.3\,\frac{TEC}{f^2} \tag{2.59}$$

and the group delay

$$\Delta\rho_{ion}^{gr} = 40.3\,\frac{TEC}{f^2} \tag{2.60}$$

respectively. The first-order ionosphere group delay (phase advance) accounts for more than 99.9% of the total ionosphere delays at the frequencies of the GNSS signals and is of the order of ∼1–50 m depending on the satellite elevation, ionosphere condition, local time, season and solar cycle for the GPS signals Kedar *et al.* [2003]. The second and third-order effects of the ionosphere on the GPS signal are typically ∼0–2 cm, and ∼0–2 mm at zenith, respectively [Bassiri and Hajj, 1993]; more details about the higher-order can be found in e.g., Fritsche *et al.* [2005], Hoque and Jakowski [2007] or Petit and Luzum [2010].

## 2.5 Electron density models – Ionosphere

Many various ionosphere models have been developed in the last few decades. These ionosphere models can be broadly categorized into three main types: empirical models, theoretical models, and parametric models (Cander *et al.*, 1999; Feltens *et al.*, 2011). Empirical models are developed by statistical analysis of long-term observations, and they are climatological models representing monthly or seasonal average conditions. They apply appropriate theory-based functions to represent spatial and temporal variations [Bilitza, 2002]. Theoretical models are also known as physical models, physics-based models, or first-principal models. They solve a set of first principles continuity, energy, and momentum equations for the electrons and ions [Cander *et al.*, 1999, and references therein]. Generally, empirical models are available in the form of computer programs and can be easily applied to a specific problem, whereas theoretical models require high computational efforts. In order to use the theoretical models in an operational setting, there are the so-called parametric models (or called the analytical models by Heise, 2002), which simplify the theoretical models by parameterizing them in terms of solar-terrestrial indices and geographical locations, aiming at representing the spatial and temporal structures of the ionosphere produced from the theoretical models by a limited number of coefficients [Cander *et al.*, 1999]. They are quite similar to the empirical models, as they use analytical functions to represent a database of measurements. One of such models is PIM [Daniell *et al.*, 1995], which is a parameterized version of four separate physical models by using EOF to represent the height profiles. The parametric models can also be attributed to the theoretical models.

With the rapid development of modern space-geodetic techniques, a large number of ionosphere models (global, regional or even local), the so-called geodetic models [Schmidt *et al.*, 2015], based on the space-geodetic techniques have been developed. Many of these models are fully analytical models, which are based on analytical functions, such as the Chapman function, exponential function and spherical harmonics (e.g., Fox, 1994; Feltens, 1998; Schaer, 1999; Alizadeh, 2013); see Chapter 5 for more details. Some of them take a physical model or an empirical model as background model and update the model parameters (e.g., Schmidt *et al.*, 2008; Zeilhofer *et al.*, 2009; Brunini *et al.*, 2011; Dettmering *et al.*, 2011b). These models are comparable to the empirical models in the type of analytical functions, however, instead of a statistical analysis

of long-term observations, they use up-to-date measurements to compute the model coefficients and describe instantaneous behaviour of the ionosphere.

A comprehensive overview of existing ionosphere models is referred to, e.g., AIAA [1999], Cander *et al.* [1999], Bilitza [2002] and Bust and Mitchell [2008]. Since the EDP is of main interest in the presented work, some existing models that specify electron density will be given below in details.

### 2.5.1 Bent model

The Bent model [Llewellyn and Bent, 1973] is an empirical global model, which predicts the EDP and the associated delay and directional changes of a wave due to refraction. The required inputs are satellite and station positions, time and specified solar data such as sunspot number and solar flux. The database which formed the basis of the model includes about 50,000 Alouette topside ionograms (1962–1966), 6,000 Ariel 3[1] in situ measurements (1967–1968), and 400,000 bottomside ionograms (1962–1969). The topside electron density is modeled by a composition of three exponential functions and a parabola; the bottomside ionosphere is represented by a so-called bi-parabola (quadratic parabola). Additionally, the topside and bottomside are fit together with a parabola function. The Bent model has been widely applied in satellite orbit determination, and has been served as a basis for the development of the topside electron density for IRI [Bilitza, 2002] introduced in Section 2.5.3.

### 2.5.2 Ching–Chiu's model

The Ching–Chiu's model [Ching and Chiu, 1973; Chiu, 1975] is a global electron density model, which was designed for global large scale variations of the electron density, i.e., the diurnal, annual and solar cycle variations. The model describes the vertical distribution of the electron density $N_e(h)$ at a certain altitude $h$ as a sum of three Chapman functions, one for each layer E, F1 and F2, which reads

$$N_e(h) = N_E(h) + N_{F1}(h) + N_{F2}(h) \tag{2.61}$$

where $N_i(h)$ of the different layers $i \in \{E, F1, F2\}$ are given in terms of Chapman functions where the peak parameters are expressed in terms of local time, annual time (days of year), altitude, geographic latitude, geomagnetic latitude and smoothed Zürich sunspot number. The model is based on monthly-averaged hourly ionosphere sounding data from 50 stations during the period 1957 to 1970.

### 2.5.3 International Reference Ionosphere (IRI)

The International Reference Ionosphere (IRI) model (Rawer *et al.*, 1978; Bilitza, 1990, 2001; Bilitza and Reinisch, 2008; Bilitza *et al.*, 2011a) is an international project sponsored by the Committee on Space Research (COSPAR) and the International Union of Radio Science (URSI). The IRI model, an internationally recognized and recommended standard for the specification of plasma parameters in the Earth's ionosphere, describes monthly averages of the electron density, electron temperature, ion temperature, ion composition, and several additional parameters

[1] U.K. satellite, 05.1967–09.1969.

in the altitude range from 50 km to 2000 km [Bilitza *et al.*, 2011b]. It has been continuously upgraded through introducing new options as new data and modeling approaches are available. Model drivers are such as solar indices, ionosphere index and magnetic indices. As an empirical model, IRI uses most of the available and reliable data sources for the ionosphere plasma. The information about the data sources is listed in Table 2.2.

Table 2.2: IRI data sources and characteristics.

| Data Source | Observed quantity | Height range | Remarks |
|---|---|---|---|
| Ionosonde | electron densities | till the F2 peak | worldwide |
| ISR | plasma densities, temperatures, velocities | the whole ionosphere | only at a few selected locations |
| Topside sounder satellite | electron densities | satellite height down to the F2 peak | global distribution |
| In situ satellite measurements | electron densities, temperatures, velocities | at the satellite orbit height | measurements are along the satellite orbit |
| Rocket | electron densities, ion composition | lower ionosphere | only reliable method for plasma parameters in the D region |

The structure of the vertical EDP of the IRI model is shown in Fig. 2.5. As can be seen, IRI divides the ionosphere into six sub-regions: from top to bottom they are the topside, the F2-bottomside, the F1 layer, the intermediate region, the E valley, and the E-bottomside/D layer. The boundaries of these sub-regions are marked by several profile anchor points including the F2, F1 and E layer peaks. IRI uses global maps of the characteristic peak densities and heights from the International Radio Consultative Committee (CCIR) [CCIR, 1996] and the International Union of Radio Science (URSI) [Rush *et al.*, 1989] as anchor points, and describes the vertical profile between these points by appropriate analytical functions. The used analytical functions include e.g., the Epstein function (see Section 5.1.1) and exponential function; see Bilitza [1990] for detailed descriptions.

### 2.5.4 NeQuick

The NeQuick (Radicella and Leitinger, 2001; Nava *et al.*, 2008) model is a quick-run electron density model of the ionosphere, developed at the Aeronomy and Radiopropagation Laboratory of the Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, Italy and at the Institute for Geophysics, Astrophysics and Meteorology of the University of Graz, Austria. It is particularly designed for transionospheric propagation applications. It allows to calculate the electron density at any given location in the ionosphere and TEC along any ground-to-satellite ray path by numerical integration. The topside model of NeQuick has been introduced as an option for the topside part of IRI [Bilitza and Reinisch, 2008]. Furthermore, the model has been adopted in the framework of the European Galileo project. NeQuick 2 is the latest version of the model, which describes the electron density of the bottomside ionosphere by a sum of five semi-Epstein layers [Rawer, 1983] with modeled thickness parameters, while the topside ionosphere

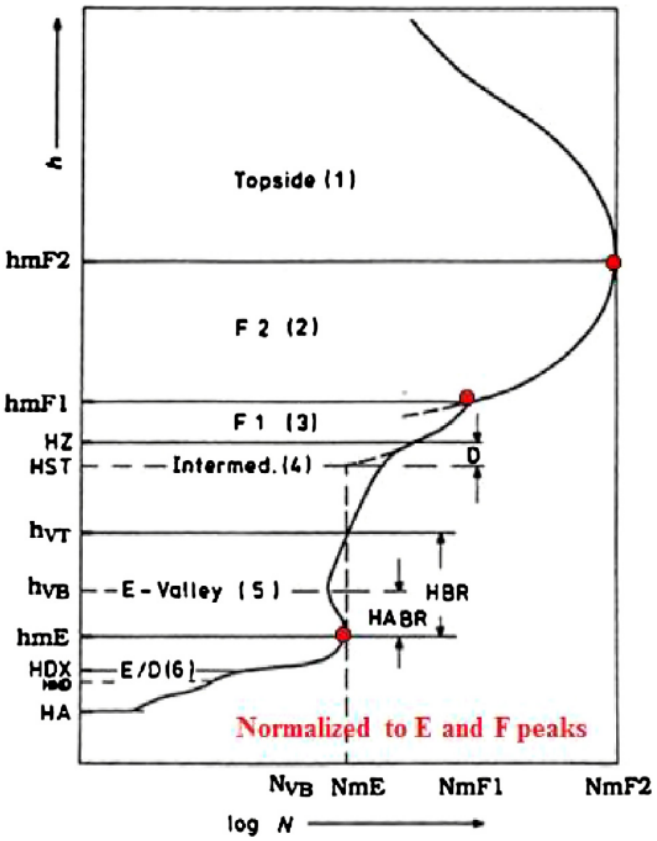



Figure 2.5: Electron density profile of IRI [Bilitza *et al.*, 2014].

is represented by a semi-Epstein layer with a height dependent thickness parameter empirically determined. The model is based on three anchor points, namely, the E, the F1 and the F2 layer peaks. Details about the complete analytical formulation of the NeQuick 2 can be found in Nava *et al.* [2008].

### 2.5.5 Global Assimilative Ionospheric Model (GAIM)

The Global Assimilative Ionospheric Model (GAIM) [Scherliess *et al.*, 2004; Schunk *et al.*, 2004] is a physics-based data assimilation model of the ionosphere and neutral atmosphere, which has been developed since 1999 by the Utah State University (USU) under the program Multidisciplinary University Research Initiatives (MURI) sponsored by the U.S. Department of Defense (DOD). It includes the Global Assimilative Ionospheric Model–Gauss-Markov (GAIM-GM) [Scherliess *et al.*, 2006] and the Global Assimilative Ionospheric Model–Full Physics (GAIM-FP) [Scherliess *et al.*, 2009]. The GAIM-GM uses a physics-based model of the ionosphere (Ionosphere Forecast Model (IFM), Schunk *et al.*, 1997) and a Kalman filter as a basis for assimilating a diverse set of real-time (or near real-time) measurements, and covers the height range of from 90 km to 1400 km. The GAIM-FP uses a physics-based Ionosphere-Plasmasphere Model (IPM) [Schunk *et al.*, 2004] and an Ensemble Kalman filter as a basis for assimilating the measurements, and covers the height range of 90 km to 30,000 km. Both GAIM-GM and GAIM-FP can assimilate different data types, e.g., in situ density measurements from satellite, bottom-side EDPs from

ionosonde, occultation data, TEC from ground-based GPS, and line-of-sight UV emissions measured by satellites. The main output is a three-dimensional (3-D) electron density distribution at user specified times. Auxiliary parameters such as $NmE$, $hmE$, $NmF2$, $hmF2$, STEC and VTEC are also provided. GAIM-FP provides additionally quantitative estimates of the accuracy of the reconstructed plasma densities and self-consistent drivers such as neutral winds and electric fields.

### 2.5.6 3-D mathematical ionosphere model at European Space Agency (ESA)/ European Space Operations Center (ESOC)

The model developed at ESA/ESOC in Darmstadt, Germany, one of the four IGS Ionosphere Associate Analysis Centers (IAACs), is an electron density model where the height distribution is represented by a sum of profile functions, one for each layer (D1, D2, E, F1, F2, plasmasphere), which follows the concept of the Ching–Chiu's model introduced in Section 2.5.2. The lowest layer D is subdivided into D2 and D1 layers, which is referred to the D and C layer introduced in Section 2.2.1. The electron density $N_e(h)$ at a certain altitude $h$ of the model reads [Feltens, 2007]

$$\begin{aligned} N_e(h) =& N_{D1}(h) + N_{D2}(h) + N_E(h) + N_{F1}(h) + N_{F2}(h) + N_p(h \geq hmF2) \\ =& NmD1 \cdot P_{D1}(h) + NmD2 \cdot P_{D2}(h) + NmE \cdot P_E(h) + NmF1 \cdot P_{F1}(h) \\ &+ NmF2 \cdot P_{F2}(h) + N_p(h \geq hmF2) \end{aligned} \tag{2.62}$$

where

| | |
|---|---|
| $N_i(h)$ | EDP of the layers $i \in \{D1, D2, E, F1, F2\}$, |
| $Nm_i$ | peak density of the layer $i$, |
| $P_i(h)$ | profile function describing the electron density distribution of the layer $i$ of the total range of the ionosphere, |
| $N_p(h \geq hmF2)$ | exponential correction to the topside part of the profile function of the F2 layer for the plasmasphere. |

Three types of profile functions are presented by Feltens [2007] for $P_i(h)$, namely, a MacLaurin series expansion[1] of the $\alpha$-Chapman layer, a superimposition of the Chapman profile function with its mirrored counterpart and a function based on the hyperbolic secant (see also Section 5.1.1). An empirical exponential correction function with a large scale height is selected for $N_p(h \geq hmF2)$.

## 2.6 Plasmasphere

Above some height level of the topside ionosphere, probably 1000 km, the domination of ions by $O^+$ is transferred to $H^+$; the corresponding region is called plasmasphere (or protonsphere, due to the domination by $H^+$). The height at which the plasmasphere starts is called the $O^+$–$H^+$ ion transition level, namely, $O^+$ and $H^+$ have a certain equilibrium relation. The upper boundary where the plasmasphere ends is called plasmapause. The plasmasphere is the inner part of the magnetosphere that co-rotates with the Earth. Typical electron density values there are about

[1] The MacLaurin series is the Taylor series centered at zero.

$10^4$ el/cm$^3$ and drop by about 1–2 orders of magnitude at the plasmapause, which can range from an L shell[1] of approximately 6 under quiet condition to a value of 2 or 3 during a large disturbance.

Actually there is no plasma production in the plasmasphere and the ionized particles have to diffuse up from the ionosphere, i.e., related to the final term of the continuity equation (2.1). The plasmasphere acts in the manner of a reservoir; during the day, it takes plasma from the ionosphere and stores it in a loss-free environment, while during the night it returns the plasma back to the ionosphere which maintains the nighttime F region [Dieminger *et al.*, 1996]. According to Hargreaves [1992], when the plasma is in equilibrium (i.e., the distribution does not change with time), the plasma density $N$ is assumed to distribute exponentially as

$$N = N_0 \exp(-h/H_P) \tag{2.63}$$

with the density $N_0$ at the ground and the plasma scale height $H_P$ defined as

$$H_P = \frac{k_B\, T_P}{m_i\, g} \tag{2.64}$$

where $k$ is the Boltzmann constant (see Eq. (2.10)) and $g = 9.81\,\mathrm{m/s^2}$ is the gravity acceleration; $m_i$ is the ion mass; the plasma temperature $T_P$ is defined as $T_P = T_e + T_i$ where $T_e$ and $T_i$ denote the electron and ion temperatures, respectively.

## 2.7 Electron density models – Plasmasphere

There exist several plasmasphere models e.g., the Global Core Plasma Model (GCPM) [Gallagher and Craven, 2000], the IZMIRAN plasmasphere model [Gulyaeva *et al.*, 2002a,b], the Global Plasmasphere Ionosphere Density (GPID) model [Webb and Essex, 2001, 2003], and the IMAGE/RPI plasmasphere model [Huang *et al.*, 2004]. They have been developed theoretically, semi-empirically or fully empirically [Goto *et al.*, 2012]. An overview of these different plasmaspheric models can be found in e.g., Gulyaeva and Bilitza [2012] and from the website `http://ccmc.gsfc.nasa.gov/modelweb/models_home.html`. In the following some details about two selective models will be given.

### 2.7.1 Global Core Plasma Model (GCPM)

The Global Core Plasma Model (GCPM) provides an empirical description of electron densities as a function of geomagnetic and solar conditions throughout the inner magnetosphere, and it merges with IRI at low altitudes. The model is based on data from the Dynamic Explorer (DE)[2] and the International Sun Earth Explorer (ISEE)[3]. It integrates region-specific models of plasma density and is composed of separate models of the ionosphere, plasmasphere, plasmapause, magnetospheric trough, and polar cap.

[1] The $L$ shell [Mcilwain, 1961] is a parameter describing the equatorial radius of a magnetic shell, e.g., $L = 6$ means that the set of the Earth's magnetic field lines which cross the Earth's magnetic equator six Earth radii from the center of the Earth.

[2] National Aeronautics and Space Administration (NASA) mission, 08.1981–02.1991.

[3] NASA and European Space Research Organisation (ESRO)/ESA program.

The plasmaspheric electron density $N_p$ at the magnetic equator is represented by

$$\begin{aligned}\log_{10}(N_p) = &(-0.79L + 5.3) \\ &+ \left[0.15\left(\cos\frac{2\pi(d+9)}{365} - 0.075\cos\frac{4\pi(d+9)}{365}\right)\right. \\ &\left. + 0.00127\overline{R} - 0.0635\right]\exp\left(\frac{L-2}{1.5}\right)\end{aligned} \tag{2.65}$$

where $L$ refers to the L shell, $d$ denotes the day of the year, and the 12-month average sunspot number is used for $\overline{R}$ in the latest model version 2.4. The plasmaspheric densities extend inward only to $\sim 2R_E$ with $R_E$ denoting the Earth's mean radius. The IRI-2007 is used to represent the electron densities at low altitudes, till an altitude of the maximum negative density gradient above the F2 peak (about 400–600 km). Above this altitude, an extrapolation is made to connect the plasmaspheric density profile by the interpolation function

$$N_e = \exp\left(\frac{h}{d_0} + d_1\right) \tag{2.66}$$

where $d_0$ and $d_1$ are constants which are fit to the slope and density at the upper boundary that IRI is used. The formulations for the plasmapause, trough and polar cap will not be included here; interested readers may refer to Carpenter and Anderson [1992] or Gallagher and Craven [2000]. The GCPM also provides the relative composition of plasmaspheric $H^+$, $He^+$ and $O^+$. The Fortran model code can be downloaded from the website `http://plasmasphere.nasa.gov/models/`.

### 2.7.2 The IZMIRAN plasmasphere model and IRI-Plas

The IZMIRAN plasmasphere model, i.e., the Russian Standard Model of the Ionosphere and Plasmasphere (SMI), is developed by IZMIRAN (the Institute of Terrestrial Magnetism, Ionosphere and Radio Waves Propagation of the Russian Academy of Sciences). It is an empirical model based on many years of measurements (whistler and satellite observations) [Gulyaeva and Gallagher, 2007], aiming at describing typical conditions as a function of geomagnetic and solar activity. The model provides global vertical EDPs towards the plasmapause up to 36,000 km. The IRI-Plas model is the IRI extension with SMI. In order to have a smooth connection between SMI and IRI, the shape of the IRI topside EDPs is switched in the IRI-Plas based on ISIS-1, ISIS-2 and IK19 satellite data.

In the IRI-Plas model, different types of expressions are used for the lower and upper part of the plasmasphere. For altitudes $h_{top} < h \leq R_E$ with $h_{top}$ denoting the height of the topside boundary of the ionosphere, the electron density $N_p(h)$ is calculated by the formulas of the SMI model (cf. Gulyaeva *et al.*, 2002a, and references therein):

$$N_p(h) = N_{top}\,\exp[(h_{top} - h)/(H_0\,L_0^2)] \tag{2.67}$$

with

$$H_0 = 0.25(R_E - h_{top})/\ln\left(N_{top}/N_{R_E}\right), \quad \text{and} \quad L_0 = 1 + h/R_E$$

where $N_{top}$ is the electron density at $h_{top}$ and $N_{R_E}$ denotes the electron density at the height $h = R_E$. At altitudes $R_E < h \leq h_{pp}$ with $h_{pp}$ referring to the height of the plasmapause, $N_p(h)$ is calculated using

$$N_p(h) = C_1 A_1 A_2 \exp(A_3\,B_L) \tag{2.68}$$

which is fit to the plasmapause electron density $N_{pp}$. Here $C_1$ and $A_1$ are functions of local time. $A_2$ and $A_3$ are functions of solar index. $B_L$ depends on $L_0$, longitude and day. Note, that the above formulas are based on the latest version of the IRI-Plas source code (modified on 24.03.2016), which have slight modifications to the formulation shown by Gulyaeva *et al.* [2002a]. The source code can be downloaded from the site `ftp://ftp.izmiran.rssi.ru/pub/izmiran/SPIM/`.

# Chapter 3

# Space-geodetic Observation Techniques

Modern space-geodetic observation techniques, such as GNSS [Hofmann-Wellenhof *et al.*, 2008], RO [Rocken *et al.*, 1997], satellite altimetry [Fu *et al.*, 1994], Very Long Baseline Interferometry (VLBI) [Schlüter and Behrend, 2007] and Doppler Orbitography and Radiopositioning Integrated by Satellite (DORIS) [Auriol and Tourain, 2010], play indispensable roles in obtaining a better understanding of the processes in the system Earth. Signals related to all these observation techniques travel through the Earth's ionosphere, and thus the measurements can be applied to obtain information about the ionosphere. In the following, GNSS and LEO RO techniques will be introduced in detail as they are the main focuses within this thesis; other space-geodetic techniques for ionospheric studies can be found in, e.g., Hobiger *et al.* [2005], Todorova [2008] or Dettmering *et al.* [2010, 2011b].

## 3.1 GNSS

The GNSS currently consists of the American Global Positioning System (GPS), the Russian GLONASS, the Chinese BeiDou System (BDS), and the European Galileo. The most important applications of GNSS are positioning and navigation. Through receiving signals from several satellites simultaneously receiver's position can be determined. An overview of different GNSS can be found in e.g., Hofmann-Wellenhof *et al.* [2008].

### 3.1.1 GPS

The GPS, officially called the Navigational Satellite Timing and Ranging (NAVSTAR) GPS, is an all-weather, space-based satellite navigation system that was developed and realized by the U.S. Department of Defense (DOD), initially for military deployment and later for civilian use as well. The GPS provides users with Positioning, Navigation and Timing (PNT) services and consists of three segments: the space, the control and the user segments.

**Space segment**
The space segment, consisting of a constellation of satellites, broadcasts radio signals towards

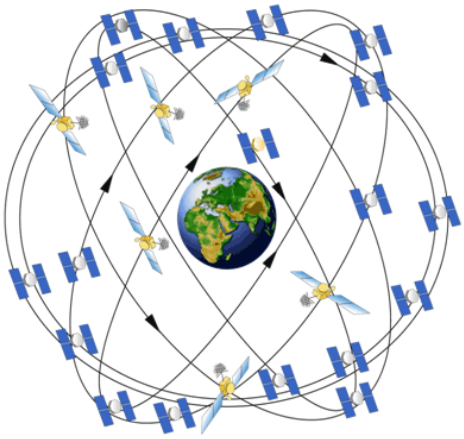


Figure 3.1: The constellation of the expandable 24 GPS satellites, source: http://www.gps.gov.

users. The GPS consists of a minimum of 24 operational satellites[1], which are distributed in six equally-spaced circular orbital planes (see Fig. 3.1) at an altitude of approximately 20,200 km above the Earth. The orbital planes have an inclination of 55° with a period of one-half of a sidereal day, i.e., about 11 h 58 min. This constellation guarantees that a global coverage with four to eight satellites above 15° elevation can be observed simultaneously at any time of day [Hofmann-Wellenhof *et al.*, 2001].

**Control segment**
The control segment, consisting of a global network of ground facilities, tracks the GPS satellites, monitors their transmissions, performs analyses, and uploads the data message to satellites. The current operational control segment contains a master control station located in Colorado Springs, Colorado, U.S.A., 15 monitor stations and a few ground control stations (ground antennas) located throughout the world.

**User segment**
The user segment, consisting of the GPS receiver equipment for military and civilian users, receives and interprets the signals broadcast by the GPS satellites.

The GPS satellites are equipped with highly accurate atomic (rubidium and cesium) clocks, which produce the fundamental L band frequency of 10.23 MHz. Each GPS satellite transmits signals continuously on frequencies which are derived from the fundamental frequency. The original GPS design includes two carriers, the main L1 carrier centered at a carrier frequency $f_1$ with wavelength $\lambda_1$:

$$f_1 = 154 \times 10.23\,\text{MHz} = 1575.42\,\text{MHz}, \quad \lambda_1 = c/f_1 \approx 19.0\,\text{cm}$$

and the L2 carrier centered at a carrier frequency $f_2$ with wavelength $\lambda_2$:

$$f_2 = 120 \times 10.23\,\text{MHz} = 1227.60\,\text{MHz}, \quad \lambda_2 = c/f_2 \approx 24.4\,\text{cm}\,.$$

[1]The Air Force has currently employed 31 operational GPS satellites (http://navcen.uscg.gov/?pageName=GPSmain).

The GPS modernization program introduced a new L5 carrier centered at a carrier frequency $f_5$ with wavelength $\lambda_5$:

$$f_5 = 115 \times 10.23\,\text{MHz} = 1176.45\,\text{MHz}, \quad \lambda_5 = c/f_5 \approx 25.5\,\text{cm}\,.$$

With such high frequencies, the effect of the ionosphere can be reduced, since the ionospheric effect becomes smaller with increasing frequencies (see the Eqs. (2.59) and (2.60)). However, the frequencies should not be too high, since a received satellite signal becomes weaker with increasing frequency for a given transmitter power. The L band frequencies are thus a good compromise between the space loss and the perturbing effect of the ionosphere [Langley, 1996].

Each carrier is modulated with Pseudorandom Noise (PRN) codes. Within this work, we have used two types of the PRN ranging codes on the L1 and L2 carriers, namely, the Coarse/Acquisition Code (C/A-code) and the Precision code (P-code):

- the C/A-code: it is available for civilian use and designated as the Standard Positioning Service (SPS). It is modulated only onto the L1 carrier;
- the P-code: it is reserved for U.S. military and other authorized users, and designated as the Precise Positioning Service (PPS). It is modulated onto both L1 and L2 carriers, which allows for removing the first-order ionospheric effect. When the "Anti-Spoofing (A-S)" is active, the P-code is encrypted into the Y-code.

In addition, both carriers are modulated by the navigation message which provides all the necessary information to allow user to calculate the position. The navigation message includes, e.g., the ephemerides of the satellites, the time parameters and clock corrections, system status, and the ionospheric parameters of the Klobuchar model [Klobuchar, 1987]. The new L5 carrier is available in new GPS satellites and is modulated with new types of PRN ranging codes, see, e.g., Kaplan and Hegarty [2006] for more details.

The reference frame of GPS is the World Geodetic System 1984 (WGS-84) developed by the U.S. DOD. It is an Earth-Centered Earth-Fixed (ECEF) coordinate system and has been refined several times to be closely coincident with the International Terrestrial Reference Frame (ITRF), which is administered by the international Association of Geodesy [Kaplan and Hegarty, 2006]. The time reference of GPS is the GPS Time (GPST), which is an atomic time scale. The initial epoch of the GPST (i.e., GPST = 0) is 0 Coordinated Universal Time (UTC) on 6 January, 1980, where UTC maintained by the United States Naval Observatory (USNO) is based on the International Atomic Time (TAI) with the relation that TAI − UTC = 19 s at the initial epoch of GPST. Since then, GPST has not been adjusted by leap seconds which are occasionally applied to UTC in order to match UTC with the Earth rotation. Therefore, it always follows the relation that GPST = TAI − 19 s. At present, GPST is ahead of UTC by 17 s due to the latest adjustment in UTC on 30 June 2015 (the next adjustment will be on 31 December, 2016).

### 3.1.2 Modeling the GPS observables

The basic GPS observables are code pseudorange and carrier phase measurements. They are obtained from comparisons of the received signal generated by the satellite clock and a replica of it generated by the receiver clock. In particular, the code pseudorange $P_{i,j}^k$ from the satellite $k$ to the receiver $j$ at a certain carrier frequency $f_i$ is derived from the traveling time $\Delta t_{i,j}^k$ of

the signal from the phase center of the satellite's antenna to the phase center of the receiver's antenna times the speed of light $c$, which reads

$$P_{i,j}^k = c\,\Delta t_{i,j}^k = c\,(t_{i,j} - t_i^k) \tag{3.1}$$

where $t_{i,j}$ represents the registered time of the signal and $t_i^k$ is the emitted time of the transmitted signal. In the following the indicators $i$, $j$ and $k$ will be kept the same for GPS observations. The traveling time $\Delta t_{i,j}^k$ is determined through a maximum correlation analysis of the PRN codes (e.g., P(Y)-code, C/A-code) generated by the GPS satellite and the replica generated by the receiver. Therefore, the calculated ranges are influenced by the satellite and receiver clock errors and "pseudorange" is named. The accuracy of pseudorange measurements is generally at the level of meter.

The carrier phase measurements are about two to three orders of magnitude more precise than the pseudorange measurements [Schaer, 1999] and can be a millimeter accuracy. The carrier phase measurements are obtained from the phase difference between the transmitted carrier wave from the satellite and the receiver's replica. They are ambiguous by an unknown integer number of wavelengths called the ambiguity, which depends on the receiver-satellite combination at the initial epoch of the signal acquisition. As long as the carrier phase tracking is continuous and not interrupted, the ambiguity remains the same. If a discontinuity that so-called cycle-slip occurs, an integer jump will be introduced and the fractional part of the phase is not influenced. Different techniques are available for cycle slip detection and repair and this issue shall not be discussed within this thesis; see, e.g., Blewitt [1990], Liu [2011] or Sharma *et al.* [2011] for details.

The code and carrier phase measurements are influenced by several different error sources. Generally, they can be categorized into satellite-related, propagation-medium-related, and receiver-related errors. The errors caused from satellites include orbital (ephemeris) errors and satellite clock errors, the errors caused from propagation medium include ionospheric effect and tropospheric effect, and the errors caused from receivers include receiver clock errors, multipath effect, receiver antenna phase center offsets and variations[1], and receiver measurement random noise. Furthermore, the carrier phase measurements are also influenced by phase wind-up effect[2]. More details about these error sources can be found in, e.g., Abdel-salam [2005] and Hofmann-Wellenhof *et al.* [2008].

### Code pseudorange observation equation

Take the most important error terms into account, the code pseudorange observation equation reads

$$P_{i,j}^k + e_{i,j}^k = \rho_j^k + c(\delta t_j - \delta t^k) + \Delta\rho_{trop,j}^k + \Delta\rho_{ion,i,j}^{gr,k} + c(b_{i,j} + b_i^k) \tag{3.2}$$

[1] Receiver antenna phase center is the point of the antenna where the GPS signal is received. Generally, it does not coincide with the antenna physical (geometrical) center and depends on the elevation, the azimuth and the frequency of the observed signal.

[2] Phase wind-up effect is caused by the property of right circularly polarized signal. Any relative rotation between satellite and receiver antennas will cause a phase variation. However, in order to obtain the maximum energy, the satellite when moving along its orbit is always rotating to keep its solar panel towards the Sun. This effect is only necessary to be corrected for high accuracy application such as Precise Point Positioning (PPP).

where

| | |
|---|---|
| $e_{i,j}^k$ | measurement random error, |
| $\rho_j^k$ | slant range, i.e., the geometric range between satellite and receiver antenna phase centers at the signal emission time and the signal reception time, respectively, |
| $\delta t_j, \delta t^k$ | receiver and satellite clock offsets with respect to the GPST, independent of frequency, |
| $\Delta\rho_{trop,j}^k$ | tropospheric delay, which is always positive and independent of frequency, |
| $\Delta\rho_{ion,i,j}^{gr,k}$ | frequency-dependent ionospheric delay, defined as Eq. (2.60), |
| $b_{i,j}, b_i^k$ | frequency-dependent receiver and satellite hardware delays on code measurements (in units of time). |

The hardware biases $b_{i,j}$ and $b_i^k$ cannot be separated from the clock offsets $\delta t_j$ and $\delta t^k$. Additionally, the hardware delays are rather stable over time, at least on time scales of several days, but they are generally varying with time (e.g., Sardón and Zarraoa, 1997; Schaer, 1999). Note, that the term $\rho_j^k$ also includes other unmodeled effects such as relativistic effects due to e.g., the satellite's velocity and the Earth's gravity field, and the receiver antenna phase center offset and variation. Furthermore, the frequency-dependent multipath effect[1] is not explicitly written in Eq. (3.2) as it is a systematic effect over a short time period of several minutes, and may be considered as measurement noise over long time spans [Schaer, 1999].

**Carrier phase observation equation**

Similarly, the carrier phase observation equation reads

$$L_{i,j}^k + \bar{e}_{i,j}^k = \rho_j^k + c(\delta t_j - \delta t^k) + \Delta\rho_{trop,j}^k + \Delta\rho_{ion,i,j}^{ph,k} + \lambda_i N_{i,j}^k - c(\bar{b}_{i,j} + \bar{b}_i^k) \tag{3.3}$$

where

| | |
|---|---|
| $L_{i,j}^k$ | carrier phase measured in units of length [m], through multiplying the phase measurements $\Phi_{i,j}^k$ (in cycles) by the wavelength $\lambda_i = c/f_i$, |
| $\bar{e}_{i,j}^k$ | measurement random error, |
| $\Delta\rho_{ion,i,j}^{ph,k}$ | frequency-dependent ionospheric delay, defined as Eq. (2.59), with the relation that $\Delta\rho_{ion,i,j}^{ph,k} = -\Delta\rho_{ion,i,j}^{gr,k}$, |
| $N_{i,j}^k$ | initial carrier phase ambiguity, which actually contains the phase wind-up effect as well, |
| $\bar{b}_{i,j}, \bar{b}_i^k$ | frequency-dependent receiver and satellite hardware delays on phase measurements (in units of time). |

The ambiguity $N_{i,j}^k$ and the phase instrumental biases $\bar{b}_{i,j}$ and $\bar{b}_i^k$ cannot be separated from each other, and they could be considered in one unknown bias [Schaer, 1999]. Note, that one ambiguity parameter $N_{i,j}^k$ has to be taken into account per satellite pass where the pass is referred

[1] Multipath refers to the phenomenon that GPS signals arrive at the receiver's antenna through two or more paths caused from reflection by reflecting surfaces surrounding the antenna, such as mountains, buildings, trees and canyon walls. The multipath effect is much more significant for code pseudorange measurements than for carrier phase measurements, and happens more frequently for signals with low elevation angles than those with high elevation angles.

to a continuous, uninterrupted arc of carrier phase tracking of a given GPS satellite on a specific receiver channel [Montenbruck, 2003]. However, if cycle-slips are detected and if they cannot be repaired, additional $N_{i,j}^k$ has to be taken into account.

### 3.1.3 Linear combination of dual-frequency GPS observables

Through appropriate linear combinations of the GPS observables, the ionospheric effect can be modeled or reduced. In the following we concentrate on the geometry-free linear combination of the pseudorange observable $P_{i,j}^k$ and the carrier phase observable $L_{i,j}^k$ measured at the same receiver to one satellite at the two carriers L1 and L2 since only these quantities have been used within this thesis. Differencing GPS data measured at one receiver to more satellites, or at more receivers to one or more satellites are also possible, see, e.g., Xu [2003] or Hofmann-Wellenhof *et al.* [2008] for details.

**Geometry-free linear combination**

For the purpose of ionospheric studies, the so-called geometry-free linear combination can be applied. The geometry-free linear combination cancels all frequency-independent terms (e.g., clock errors, tropospheric delay) in the Eqs. (3.2) and (3.3), and only the frequency-dependent terms such as the ionospheric refraction, the differential inter-frequency hardware delays of satellite and receiver, and ambiguity parameter (only for phase measurements) are left. The geometry-free linear combinations of the code and phase measurements, denoted as $P_{4,j}^k$ and $L_{4,j}^k$, read

$$P_{4,j}^k = P_{2,j}^k - P_{1,j}^k \tag{3.4}$$

$$L_{4,j}^k = L_{1,j}^k - L_{2,j}^k \,. \tag{3.5}$$

The substitution of the Eqs. (3.2) and (3.3) into the Eqs. (3.4) and (3.5) under consideration of the Eqs. (2.59) and (2.60) for slant ray paths of GPS signals yields

$$P_{4,j}^k + e_{4,j}^k = a \cdot STEC_j^k - c(\delta b_j + \delta b^k) \tag{3.6}$$

$$L_{4,j}^k + \overline{e}_{4,j}^k = a \cdot STEC_j^k - c(\delta \overline{b}_j + \delta \overline{b}^k) + N_{4,j}^k \tag{3.7}$$

where

| | |
|---|---|
| $e_{4,j}^k$, $\overline{e}_{4,j}^k$ | random error, |
| $a = 40.3 \left( \frac{f_1^2 - f_2^2}{f_1^2 f_2^2} \right)$ | frequency-dependent factor, |
| $\delta b_j = b_{1,j} - b_{2,j}$ | receiver inter-frequency or Differential Code Bias (DCB), in units of time, |
| $\delta b^k = b_1^k - b_2^k$ | satellite Differential Code Bias (DCB), in units of time, |
| $\delta \overline{b}_j = \overline{b}_{1,j} - \overline{b}_{2,j}$ | receiver inter-frequency or Differential Phase Bias (DPB) [Themens *et al.*, 2013], in units of time, |
| $\delta \overline{b}^k = \overline{b}_1^k - \overline{b}_2^k$ | satellite Differential Phase Bias (DPB), in units of time, |
| $N_{4,j}^k = \lambda_1 N_{1,j}^k - \lambda_2 N_{2,j}^k$ | ambiguity parameter, defined in units of length. |

Note that the split of signals at different frequencies is ignored here. With appropriate combination of dual-frequency measurements, the ionosphere-free linear combination can be formed and the first-order ionospheric effect can thus be eliminated. This is the main reason why at least two carrier waves are implemented in GNSS satellites [Hofmann-Wellenhof *et al.*, 2008]. Further details about the ionosphere-free linear or other combinations can be found in, e.g., Schaer [1999].

**Code pseudorange smoothing and carrier phase leveling**

As introduced previously, the phase measurements are much more precise than the code measurements, but they are ambiguous. In practice, the code and phase measurements are usually combined to take advantages of both types of measurements.

Based on a continuous time series of dual-frequency code and phase measurements, code pseudorange smoothing can be applied, which significantly reduce the noise of code measurements. Taking a certain time epoch $t$ into account, the measurements in the Eqs. (3.2) and (3.3) are reformulated as $P^k_{i,j}(t)$ and $L^k_{i,j}(t)$. The phase-smoothed code measurements $\widetilde{P}^k_{1,j}(t)$ and $\widetilde{P}^k_{2,j}(t)$ at the two frequencies (i = 1, 2) can be calculated by [Schaer, 1999]

$$\begin{aligned}\widetilde{P}^k_{1,j}(t) &= \overline{P}^k_{1,j} + \Delta L^k_{1,j}(t) + 2\frac{f_2^2}{f_1^2 - f_2^2}\left(\Delta L^k_{1,j}(t) - \Delta L^k_{2,j}(t)\right)\\ \widetilde{P}^k_{2,j}(t) &= \overline{P}^k_{2,j} + \Delta L^k_{2,j}(t) + 2\frac{f_1^2}{f_1^2 - f_2^2}\left(\Delta L^k_{1,j}(t) - \Delta L^k_{2,j}(t)\right)\end{aligned} \tag{3.8}$$

with

$$\begin{aligned}\Delta L^k_{1,j} &= L^k_{1,j}(t) - \overline{L}^k_{1,j}\\ \Delta L^k_{2,j} &= L^k_{2,j}(t) - \overline{L}^k_{2,j}\end{aligned} \tag{3.9}$$

where $\overline{P}^k_{1,j}$ and $\overline{P}^k_{2,j}$ are the mean code measurements averaged over a common time interval without a cycle-slip, at two different frequencies, respectively. Similarly, $\overline{L}^k_{1,j}$ and $\overline{L}^k_{2,j}$ are the mean phase measurements.

Following Schaer [1999], the phase measurements $L^k_{1,j}(t)$ and $L^k_{2,j}(t)$ are assumed to be error-free. The noises of the smoothed code measurements are thus reduced by a factor of about $\sqrt{n_0}$ based on Eq. (3.8), where $n_0$ is the considered total number of epochs for calculations of the mean code and phase measurements. The smoothed code measurements are highly correlated in time, which may generally result that estimated standard deviations are too optimistic if the measurements are treated as uncorrelated.

Applying the Eqs. (3.4) and (3.6) to the phase-smoothed code observables yields

$$\widetilde{P}^k_{4,j} + \widetilde{e}^k_{P,j} = \widetilde{P}^k_{2,j} - \widetilde{P}^k_{1,j} = a \cdot STEC^k_j - c(\delta b_j + \delta b^k) \tag{3.10}$$

with the random error $\widetilde{e}^k_{P,j}$, and it can be further formulated as

$$\frac{1}{a}\widetilde{P}^k_{4,j} + \frac{1}{a}\widetilde{e}^k_{P,j} = STEC^k_j - \frac{c}{a}(\delta b_j + \delta b^k)\,. \tag{3.11}$$

Another technique called carrier phase leveling can be applied to remove the phase ambiguity term to make use of high-precision phase measurements. For any phase connected arc (i.e.,

a satellite pass), the geometry-free phase observations $L^k_{4,j}$ can be leveled to the geometry-free pseudorange measurements through an offset $C^k_j$ calculated by (cf. Mannucci *et al.*, 1998; Dettmering, 2003; Themens *et al.*, 2013)

$$C^k_j = \frac{1}{n_{arc}} \sum_{m=1}^{n_{arc}} \left[ \left( P^k_{2,j}(t_m) - P^k_{1,j}(t_m) \right) - \left( L^k_{1,j}(t_m) - L^k_{2,j}(t_m) \right) \right] \tag{3.12}$$

where $n_{arc}$ is the total number of measurements over the phase connected arc with different time epochs denoted as $t_m$ $(m = 1, \ldots, n_{arc})$. $C^k_j$ can be considered as the mean difference between phase-derived and code-derived ionospheric observables. In order to simplify the notation, we leave the time $t$ out in the following. The leveled $\widetilde{L}^k_{4,j}$ at each time epoch in the arc can then be calculated by

$$\widetilde{L}^k_{4,j} = L^k_{4,j} + C^k_j \,. \tag{3.13}$$

Substituting the Eqs. (3.6) and (3.7) into Eq. (3.12) yields

$$C^k_j = \overline{-c(\delta b_j + \delta b^k)} + \overline{c(\delta \bar{b}_j + \delta \bar{b}^k)} - \overline{N^k_{4,j}} \tag{3.14}$$

where $\bar{\cdot}$ represents the operation for taking the mean value of the corresponding quantity over the phase connected arc. Over one arc, these Differential Code Biases (DCBs) and Differential Phase Biases (DPBs) can be considered to be constant since they are stable over periods much longer than an arc [Themens *et al.*, 2013, and references therein]. Furthermore, the integer ambiguity $N^k_{4,j}$ is also constant. Thus, Eq. (3.14) can be reformulated as

$$C^k_j = -c(\delta b_j + \delta b^k) + c(\delta \bar{b}_j + \delta \bar{b}^k) - N^k_{4,j} \,. \tag{3.15}$$

The substitution of the Eqs. (3.7) and (3.15) into Eq. (3.13) yields

$$\widetilde{L}^k_{4,j} + \widetilde{e}^k_{L,j} = a \cdot STEC^k_j - c(\delta b_j + \delta b^k) \tag{3.16}$$

with the random error $\widetilde{e}^k_{L,j}$. As can be seen, only ionospheric information and the satellite and receiver DCBs are left in the new observables $\widetilde{L}^k_{4,j}$. Similar to Eq. (3.11), we can reformulate Eq. (3.16) as

$$\frac{1}{a}\widetilde{L}^k_{4,j} + \frac{1}{a}\widetilde{e}^k_{L,j} = STEC^k_j - \frac{c}{a}(\delta b_j + \delta b^k) \,. \tag{3.17}$$

### 3.1.4 Other GNSS

**Global'naya Navigatsionnaya Sputnikovaya Sistema (GLONASS)**

The Russian Global'naya Navigatsionnaya Sputnikovaya Sistema (GLONASS) is operated by the Russian Space Forces. Like GPS, GLONASS also consists of the space, the control and the user segments. GLONASS was intended to be composed of at least 24 satellites, which are equally distributed in three orbital planes separated by 120° with an inclination of 64.8°. The orbits are roughly circular with the nominal radius of about 25,500 km (i.e., an orbit altitude of about 19,100 km) and an orbital period of about 11 h 15 min 44 s.

A major difference between GLONASS and GPS is the signal. The GPS satellites transmit signals on the same carrier frequencies, whereas each GLONASS satellite transmits signals on

its own carrier frequencies. The GLONASS carrier frequencies $f_{1,i}, f_{2,i}$ related to the channel number $i$ assigned to a specific satellite are [Leick *et al.*, 2015]

$$f_{1,i} = 1602 + 0.5625\, i \approx 1598.0625\text{–}1605.375 \text{ MHz},$$
$$f_{2,i} = 1246 + 0.4375\, i \approx 1242.9375\text{–}1248.625 \text{ MHz}$$

with $i = -7, -6, \ldots, +5, +6$ and $f_{1,i}/f_{2,i} = 9/7 \approx 1.2857$. Besides the above L1 and L2 sub-bands, the modernization of the GLONASS program introduced a new third L3 band for modernized satellites GLONASS-K. Modernized GLONASS-K satellites, GLONASS-KM to be launched by 2025 may transmit signals on the L5 band as the same as GPS[1]. Similar to GPS, each satellite modulates its L1 carrier with the C/A-code and the P-code, and modulates its L2 carrier only with the P-code. Since the GLONASS-M satellites, the C/A-code is also modulated on the L2 carrier. Unlike GPS, all GLONASS satellites transmit the same PRN ranging codes at different frequencies.

The differences to GPS are also related to the reference frame and the time reference. The current GLONASS coordinates are based on the "Parametry Zemli 1990" (or "The Earth Parameter 1990" in English), PZ-90.11 reference frame[2]. The GLONASS time is based on UTC as maintained at the Russian Institute of Metrology for Time and Space (UTC (SU)). GLONASS time has an offset with respect to UTC (SU) plus three hours within 1 millisecond. Unlike GPST, leap seconds are implemented in the GLONASS time and thus its time scale is not continuous. More details about GLONASS can be found in, e.g., Hofmann-Wellenhof *et al.* [2008], or the official website `https://glonass-iac.ru`.

**BeiDou System (BDS)**

The Chinese BeiDou navigation satellite system, called BeiDou System (BDS) for short, is the third navigation system after GPS and GLONASS [Montenbruck and Steigenberger, 2013]. The BDS is developed by three-step strategies. The first step, realized by BeiDou-1 (also known as BeiDou Navigation Satellite Demonstration System), is an experimental satellite navigation system, which consists of three Geostationary Earth Orbit (GEO) satellites. Unlike GPS and GLONASS, BDS-1 is an active positioning system[3]. BeiDou-1 was in full operation in 2003 and offered navigation services, providing users in China with positioning, timing, wide-area differential and short message communication services. The second step is to construct the BDS-2 (also known as COMPASS), which consists of 5 GEO, 4 Medium Earth Orbit (MEO) and 5 Inclined Geosynchronous Satellite Orbit (IGSO) satellites and has been completed by the end of 2012. The BDS-2 added the passive-positioning scheme[4] and provides positioning services for users in the Asia-Pacific region. The third step, i.e., the full constellation of the global BDS, is currently under construction and will inherit both active and passive services [CSNO, 2016]. The full deployment of BDS will be consisted of 5 GEO satellites with an orbital altitude of 35,786 km, 27 MEO satellites with an average orbital altitude of 21,528 km and an inclination of

[1] `https://en.wikipedia.org/wiki/GLONASS`.

[2] PZ-90.11 is an ECEF frame and the specifications of the parameters can be found in "PARAMETRY ZEMLI 1990" Reference Document, Moscow 2014, downloadable from `https://www.glonass-iac.ru/en/guide/navfaq.php`.

[3] It is a two-way ranging system, i.e., the users need to transmit signals to satellites at first and the signals are then retransmitted to the ground control center. The ground control center computes the users' positions and send the information back to the users via satellites. Therefore, the signals travel the distance twice.

[4] Both GPS and GLONASS are passive one-way downlink ranging systems, i.e., the signal emitted by satellite travels once from space to the Earth.

55°, and 3 IGSO satellites with an orbital altitude of 35,786 km and an inclination of 55°. The latest launch in June 2016 is the 23rd satellite. The completion of the constellation deployment is expected by the end of 2020.

All BeiDou satellites currently transmit signals on the following three nominal frequencies with corresponding wavelengths,

$$\begin{aligned} B_1 &= 1561.098\,\text{MHz}, \quad \lambda_{B_1} = 19.2\,\text{cm}, \\ B_2 &= 1207.140\,\text{MHz}, \quad \lambda_{B_2} = 24.9\,\text{cm}, \\ B_3 &= 1268.520\,\text{MHz}, \quad \lambda_{B_3} = 23.6\,\text{cm}. \end{aligned}$$

BDS broadcasts a set of eight correction parameters for single-frequency users to compute the ionospheric path delays [Montenbruck and Steigenberger, 2013]. The underlying ionospheric model is the COMPASS Ionospheric Model (CIM) [Wu *et al.*, 2013], which is similar to the GPS Klobuchar model.

The coordinate system of BDS is the China Geodetic Coordinate System 2000 (CGCS2000), see, e.g., CSNO [2013] for its definition. The time reference is the BeiDou Time (BDT), which is a continuous time scale with the initial epoch at 0 UTC on 1 January, 2006. BDT has an offset with respect to UTC within 100 ns. About further details please refer to CSNO [2013] or `http://en.beidou.gov.cn/`.

**Galileo**

The European global satellite navigation system, Galileo, was a joint initiative of the European Commission (EC) and the ESA under civilian control, with the aim of providing a highly accurate, guaranteed global positioning service and being interoperable and compatible with other GNSS.

The Galileo system also consists of three major segments: space, control and user segments. The satellite constellation will be composed of 24 operational and 3 spare satellites which are positioned in three MEO planes with an inclination of 56°. The orbits are nearly circular with an altitude of 23,222 km. They have a period of 14 hours 4 minute 45 seconds and a ground track repeat cycle of about 10 days. The full constellation of the 27 operational satellites without spare satellites will guarantee continuously global coverage, with six to eight visible satellites at any point and any time over the globe. The Galileo program is structured into three main phases: the first phase is experimental phase, where two experimental satellites were launched in December 2005 and April 2008, respectively. They are no longer operational after they finished their missions that characterized and verified the key technologies required by the Galileo system and reserved the radio frequencies at the International Telecommunications Union (ITU) as well; the second phase is In-Orbit Validation (IOV) phase, with the aim of validating the system design using a reduced constellation of four satellites (two satellites were launched on 21 October 2011 and the other two on 12 October 2012); the third phase is Full Operational Capability (FOC) phase that will deploy the remaining ground and space infrastructure, which is expected to be completed by 2020.

The Galileo system transmits signals on four different operating frequency bands with wave-

lengths:

$$\begin{aligned} \text{E1} &= 1575.42\,\text{MHz}, \quad \lambda_{\text{E1}} = 19.0\,\text{cm}, \\ \text{E6} &= 1278.75\,\text{MHz}, \quad \lambda_{\text{E6}} = 23.4\,\text{cm}, \\ \text{E5a} &= 1176.45\,\text{MHz}, \quad \lambda_{\text{E5a}} = 25.5\,\text{cm}, \\ \text{E5b} &= 1207.14\,\text{MHz}, \quad \lambda_{\text{E5b}} = 24.8\,\text{cm} \end{aligned}$$

where E5a and E5b are part of the E5 bandwidth. All satellites exploit the same carrier frequencies for signal transmission. The NeQuick model (see Section 2.5.4) is used by the Galileo single-frequency users to compute the ionospheric path delays.

The coordinate system of Galileo is the Galileo Terrestrial Reference Frame (GTRF), see, e.g., Gendt *et al.* [2011] for details. The Galileo time reference is the Galileo System Time (GST) maintained by the Galileo Central Segment and synchronized with TAI within a nominal offset of 50 ns. Like GPST, the GST is a continuous time scale and the initial epoch is at 0 UTC on 22 August, 1999. About further details please refer to `https://www.gsc-europa.eu/`.

## 3.2 Radio Occultation (RO) techniques

With a GNSS receiver onboard a LEO satellite, the signal transmitted by a GNSS satellite can be received by the LEO satellite passing through the Earth's atmosphere in a limb sounding geometry. This refers to the Radio Occultation (RO) technique, which can provide global characteristics of the Earth's neutral atmosphere and the ionosphere, with a unique combination of global coverage, high precision, high vertical resolution, long-term stability, and all-weather viewing [Kursinski *et al.*, 1997]. The RO technique dates back to the early 1960s, when a team of scientists from the Stanford University and the Jet Propulsion Laboratory (JPL) probed the atmosphere of Mars. The technique was applied to the Earth in April 1995 with the Global Positioning System/Meteorology (GPS/MET) experiment onboard the Microlab-1 satellite [Ware *et al.*, 1996], which used RO observations of GPS satellites to obtain vertical profiles of electron density in the ionosphere; refractivity, density, pressure, and temperature in the stratosphere and upper troposphere; and refractivity, density, pressure and water vapor pressure in the lower troposphere [Rocken *et al.*, 1997]. The success of the GPS/MET experiment have brought about subsequently several other successful missions with GPS occultation receiver, such as Challenging Minisatellite Payload (CHAMP), Gravity Recovery and Climate Experiment (GRACE) and FORMOSAT-3/COSMIC (F3/C). An overview of LEO missions with GNSS RO instruments can be found in Mannucci *et al.* [2014].

### 3.2.1 GPS RO missions

**Challenging Minisatellite Payload (CHAMP)**
The German Challenging Minisatellite Payload (CHAMP) mission [Reigber *et al.*, 2000] was launched on 15 July, 2000, with the aim of determining the Earth's gravity and magnetic fields and achieving global limb soundings of the Earth's neutral atmosphere and the ionosphere [Wickert *et al.*, 2001]. The orbit is almost circular and near polar (inclination = 87.2°) with an initial altitude of 454 km, which was lowered later due to atmospheric drag. The first RO measurements were available from its "BlackJack" GPS receiver on 11 February, 2001. From then on, CHAMP provided about 250 daily RO measurements continuously till October 2008. The satellite has

burnt up on 19 September, 2010. Further details about the CHAMP mission can be found at http://op.gfz-potsdam.de/champ/.

**Gravity Recovery and Climate Experiment (GRACE)**
The U.S./German Gravity Recovery and Climate Experiment (GRACE) mission [Tapley *et al.*, 2004] was launched on 17 March, 2002, with the aim of mapping the global gravity field. The GRACE mission comprises two satellites, in near-circular polar orbits (inclination = 89.5°) at an altitude of around 500 km. The two satellites are separated from each other by about 220 km along-track and connected by highly accurate inter-satellite K-band microwave ranging system. Both satellites are equipped with CHAMP-like "BlackJack" GPS flight receivers. The first occultation measurements were obtained during a 25 h period on 28/29 July, 2004 aboard the GRACE-B satellite, a longer period (41 days) of RO data were recorded between 12 January and 20 February, 2006 aboard the GRACE-A satellites, and continuous RO data were activated aboard the GRACE-A satellite since 22 May, 2006. Ionospheric profiles have been continuously provided by GRACE-A since 28 February, 2007 [Wickert *et al.*, 2009]. The GRACE Follow-On (GRACE-FO) mission is planned to be launched in August 2017. Further details about the GRACE mission can be found at http://www.csr.utexas.edu/grace/.

**FORMOSAT-3/COSMIC (F3/C)**
The Taiwan/U.S. FORMOSAT-3/COSMIC (F3/C) mission [Rocken *et al.*, 2000] was launched on April 15, 2006, with the aim of remotely sensing the Earth's atmosphere and the ionosphere. It has a constellation of six microsatellites which are distributed in separate orbital planes with an inclination of 72° and an altitude of about 800 km. F3/C can provide 2000–2500 daily profiles of key ionospheric and atmospheric properties. The primary data processing center for F3/C is the University Corporation for Atmospheric Research (UCAR) COSMIC Data Analysis and Archive Center (CDAAC) (http://cdaac-www.cosmic.ucar.edu/cdaac/). Another data processing center is the Taiwan Analysis Center for COSMIC (TACC) at Central Weather Bureau (CWB) (http://tacc.cwb.gov.tw/).

A F3/C follow-on mission named FORMOSAT-7/COSMIC-2 (F7/C2) is currently planned. It will have a constellation of twelve microsatellites, which have the capability of tracking GPS, GLONASS and Galileo signals. With the full constellation, F7/C2 can provide more than 12,000 daily ionospheric profiles. Further details about the F3/C and F7/C2 missions can be found at http://www.cosmic.ucar.edu/.

### 3.2.2 GPS RO principle

A typical geometry of the RO technique involving a GPS satellite and a LEO satellite is illustrated in Fig. 3.2. An occultation takes place when a GPS satellite sets or rises behind the Earth's ionosphere or lower neutral atmosphere as seen by a GPS receiver onboard a LEO satellite. During an occultation event, GPS and LEO satellites move relatively, and thus the ray path descends or ascends through the atmosphere. Each occultation measurement consists of a set of limb-viewing links with perigee points (or tangent points) ranging from the LEO satellite orbit height to the surface of the Earth, and an occultation event takes nearly 4–10 min [Hajj *et al.*, 2000].

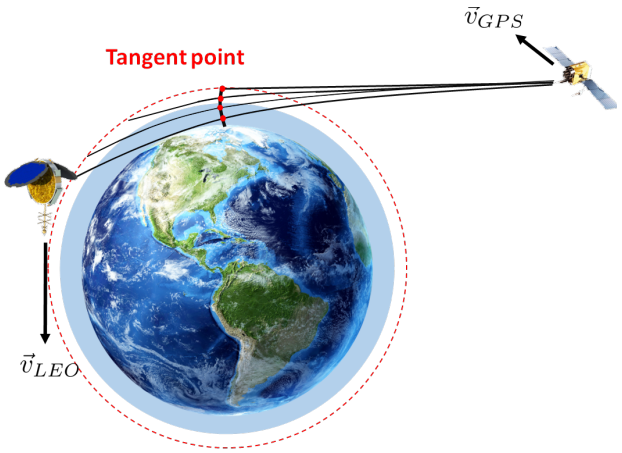



Figure 3.2: A geometry of the RO technique, adapted from http://goes-r.gov.

The signal passing through the atmosphere will not be straight but bent. The degree of bending will change due to the vertical variation of the atmospheric refractive index [Kursinski *et al.*, 1997]. For the neutral atmosphere, the refractive index is depending on pressure, humidity and temperature, whereas in case of the ionosphere it is depending on the electron density. The amount of bending (i.e., the bending angle) is a key quantity since it can be inverted to the refractive index using the Abel transform inversion [Fjeldbo and Eshleman, 1969], which in turn yields the electron density. Furthermore, similar to GPS observations measured by ground-based receivers, the signal measured by space-based GPS receivers will experience the ionospheric code delays and phase advances, and therefore the information of the ionosphere in particular $STEC$ can be extracted by dual-frequency receivers.

### 3.2.3 EDP retrieval

The most commonly used retrieval technique is the Abel transform through calculation of the two different quantities: bending angles or STEC. Other methods such as tomographic approach have also been investigated (e.g., Leitinger *et al.*, 1997; Hernández-Pajares *et al.*, 1998; Jakowski *et al.*, 2002). The tomographic approach uses a set of orthogonal functions to describe the electron density and usually requires estimation of a large number of parameters, which might lead to a low vertical resolution. The Abel inversion has a simpler performance and a better vertical resolution [Hernández-Pajares *et al.*, 2000]. However, the traditional Abel inversion technique assumes that the ionospheric electron density distribution in some region around the tangent points of rays follows spherical symmetry, which means that no horizontal variations is considered and thus the refractive index depends only on the height. However, it is well known that the ionosphere exhibits both the vertical and horizontal variations, even in the typical short time span of few minutes during an occultation event. The spherical symmetry assumption is considered to be the most significant error source for the retrieval technique (e.g., Schreiner *et al.*, 1999; García-Fernández, 2004; Wu *et al.*, 2009). Therefore, improved Abel inversions that take horizontal electron density gradient into account have been proposed (e.g., Hernández-Pajares *et al.*, 2000; García-Fernández, 2004; Tsai and Tsai, 2004; Tsai *et al.*, 2009, 2011). In the following, the traditional Abel inversion using STEC along GPS–LEO ray paths that the approaches of many research faculties as well as F3/C data processing centers are based on will be briefly described. The Abel inversion using bending angles can be found in e.g., Hajj and

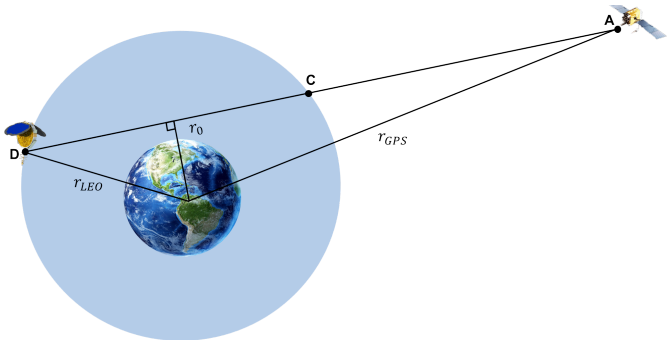



Figure 3.3: Geometry of ray path for RO technique by assuming a straight line propagation.

Romans [1998], Schreiner *et al.* [1999], García-Fernández [2004] or Limberger [2015].

The Abel inversion technique using STEC is based on that the bending angle is very small and negligible in the ionosphere, and thus straight line propagation is assumed (see Fig. 3.3). According to Schreiner *et al.* [1999], bending of signals at GPS frequencies in the ionosphere is small enough that ray separation from straight line propagation for LEO observations is only about several kilometers or less, even under the worst ionospheric conditions. The quantity is much smaller than typical vertical scales of the electron density in the F2 layer. However, this may not be valid for the layers with smaller-scale structures like the E layer, but the influence is usually less important than the error caused by the spherical symmetry assumption [García-Fernández, 2004].

The carrier phase observation equation (3.3) is also applicable to space-based GPS phase measurements. The STEC along the GPS ray path from GPS satellite $k$ to LEO satellite $j$ can be derived from the geometry-free linear combination $L^k_{4,j}$ according to Eq. (3.7), reads (cf. Hajj *et al.*, 2000)

$$STEC^k_j - a\,B^k_j = -a\,L^k_{4,j} + e'^k_j \tag{3.18}$$

where $B^k_j$ is the bias containing $N^k_{4,j}$ and the receiver and satellite DPBs; $e'^k_j$ is the random noise. In order to simplify the notation, we leave the receiver and satellite indices $j$ and $k$ out in the following. Based on the spherical symmetry assumption of the electron density distribution, STEC is related to the electron density through the following integral

$$STEC(r_0) = \left[\int_{r_0}^{r_{GPS}} + \int_{r_0}^{r_{LEO}}\right] \frac{rN_e(r)}{\sqrt{r^2 - r_0^2}}\,dr \tag{3.19}$$

where $r_0$ is the radial distance of the perigee point from the Earth's center, $r_{GPS}$ and $r_{LEO}$ are the radial distances of the GPS and the LEO, and $r$ denotes the radius, see Fig. 3.3. The derived $STEC$ (i.e., the left-hand side of Eq. (3.18)) differs from the actual $STEC$ by an offset $a\,B^k_j$, and it is calibrated so that it approximately represents the portion of $STEC$ below the orbit height of LEO satellite [Lei *et al.*, 2007], the calibrated $\widetilde{STEC}(r_0)$ reads

$$\widetilde{STEC}(r_0) = STEC_{CD}(r_0) = STEC_{AD}(r_0) - STEC_{AC}(r_0) \tag{3.20}$$

where $A$ denotes the position of the GPS, $D$ is the position of the LEO, and $C$ is an auxiliary position with the same radial distance as $D$, see Fig. 3.3; detailed calibration approach is referred to Schreiner *et al.* [1999]. Equation (3.19) can be transformed for calibrated $\widetilde{STEC}(r_0)$ to

$$\widetilde{STEC}(r_0) = 2\int_{r_0}^{r_{LEO}} \frac{rN_e(r)}{\sqrt{r^2 - r_0^2}}\,dr \tag{3.21}$$

which can then be inverted using the Abel integral transform to obtain the expression for the electron density

$$N_e(r) = -\frac{1}{\pi}\int_r^{r_{LEO}} \frac{d\widetilde{STEC}(r_0)/dr_0}{\sqrt{r_0^2 - r^2}}\, dr_0\,. \tag{3.22}$$

**Recursive Inversion of STEC data**

To solve Eq. (3.21), an onion peeling algorithm is commonly used, i.e., an onion shell structure depending on the height of the successive tangent points is introduced. For each shell we assume a constant electron density. A schematic geometry of the onion shell structure is illustrated in Fig. 3.4. Each $STEC$ observation defines a shell in the vertical profile, and the vertical resolution of EDP is determined by the sampling rate of the phase measurements [García-Fernández, 2004]. The shells starting from the outermost toward the innermost are denoted as $s = 1, \ldots, M$. In this manner, the right-hand side of Eq. (3.21) can be discretized as (cf. García-Fernández, 2004; Wu *et al.*, 2009; Limberger *et al.*, 2015)

$$\widetilde{STEC}(r_0) = 2\sum_{s=1}^{s=i} N_{e,s}\, l_{is} \tag{3.23}$$

where $N_{e,s}$ is the electron density value of the shell $s$ and $l_{is}$ denotes the length of the ray path at epoch $i$ in the $s$-th shell. The relation $s \leq i$ has to be fulfilled. Therefore, the electron density values can be calculated recursively starting from the uppermost to the bottommost observations. This method is similar to the one used at CDAAC (e.g., Lei *et al.*, 2007; Syndergaard, 2007).

The inversion errors of the above algorithm stem mainly from the spherical symmetry assumption. Other error sources include, such as the assumption of straight line propagation of the GPS-LEO ray paths in the ionosphere, the calibration technique for approximating $\widetilde{STEC}(r_0)$ and measurement errors. Further details can be found in, e.g., Schreiner *et al.* [1999], Wu *et al.* [2009] or Yue *et al.* [2010].

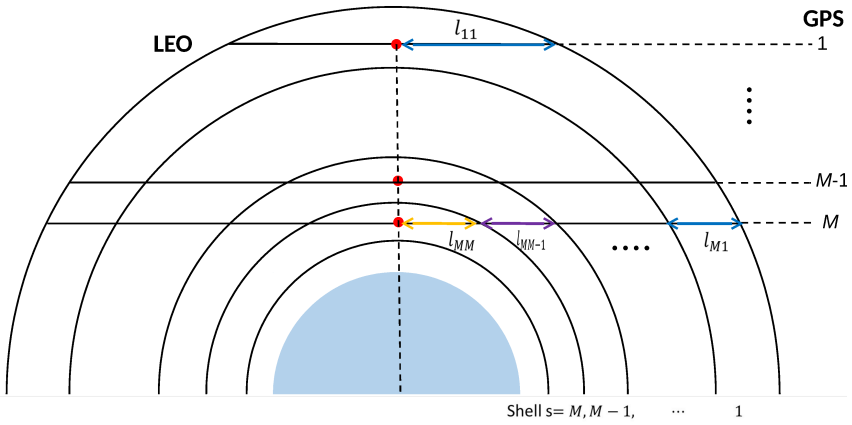



Figure 3.4: Illustration of the onion shell structure that discretizes STEC.

# Chapter 4

# Parameter Estimation

This chapter includes two parts. The first part introduces the fundamental of the adjustment theory whereas the second one explains inequality constrained optimization. In order to solve for unknown parameters, the functional relations between the unknown parameters and the observations have to be set up, and the statistical properties of the observations need to be specified. We start from a nonlinear problem, as it will be tackled within this thesis.

## 4.1 Nonlinear problem

Let $y_i^*$ $(i = 1, \ldots, n)$ be the $n$ observations with errors $e_i$ and $\beta_j$ $(j = 1, \ldots, u)$ the $u$ unknown parameters, the relations between $y_i^*$ and $\beta_j$ can generally be written as

$$\begin{cases} f_1(\beta_1, \ldots, \beta_u) & = y_1^* + e_1 \\ f_2(\beta_1, \ldots, \beta_u) & = y_2^* + e_2 \\ & \vdots \\ f_n(\beta_1, \ldots, \beta_u) & = y_n^* + e_n \end{cases} \tag{4.1}$$

where $f_i(\beta_1, \ldots, \beta_u)$ are real-valued differentiable functions of $\beta_j$. Therefore, Eq. (4.1) may be linearized by means of the Taylor series expansion, cut off at the linear term computed with the initial parameter vector $\boldsymbol{\beta}_0 = [\beta_{10}, \ldots, \beta_{u0}]^T$, which collects the initial values $\beta_{j0}$ $(j = 1, \ldots, u)$ as [Koch, 1999]

$$\begin{aligned} f_i(\beta_1, \ldots, \beta_u) &= f_i(\beta_{10} + \Delta\beta_1, \ldots, \beta_{u0} + \Delta\beta_u) \\ &\approx \underbrace{f_i(\beta_{10}, \ldots, \beta_{u0})}_{f_{i0}} + \left.\frac{\partial f_i}{\partial \beta_1}\right|_{\boldsymbol{\beta}_0} \Delta\beta_1 + \ldots + \left.\frac{\partial f_i}{\partial \beta_u}\right|_{\boldsymbol{\beta}_0} \Delta\beta_u \\ &\approx f_{i0} + \left(\nabla_{\boldsymbol{\beta}} f_i(\boldsymbol{\beta}_0)\right)^T \Delta\boldsymbol{\beta} \end{aligned} \tag{4.2}$$

with

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \Delta\boldsymbol{\beta}, \quad \boldsymbol{\beta} = [\beta_1, \ldots, \beta_u]^T, \quad \Delta\boldsymbol{\beta} = [\Delta\beta_1, \ldots, \Delta\beta_u]^T$$

and

$$\nabla_{\boldsymbol{\beta}} f_i(\boldsymbol{\beta}_0) = \left[ \left.\frac{\partial f_i}{\partial \beta_1}\right|_{\boldsymbol{\beta}_0}, \ldots, \left.\frac{\partial f_i}{\partial \beta_u}\right|_{\boldsymbol{\beta}_0} \right]^T$$

where the superscript $T$ denotes the transpose and the symbol $\nabla_{\boldsymbol{\beta}}$ represents the gradient with respect to $\boldsymbol{\beta}$. Equation (4.1) can then be reformulated under consideration of Eq. (4.2) in matrix notation as

$$\boldsymbol{X}\Delta\boldsymbol{\beta} = \boldsymbol{y}^* - \boldsymbol{y}_0 + \boldsymbol{e} = \Delta\boldsymbol{y} + \boldsymbol{e} \tag{4.3}$$

with

$$\boldsymbol{X} = \begin{pmatrix} \left.\frac{\partial f_1}{\partial \beta_1}\right|_{\boldsymbol{\beta}_0} & \cdots & \left.\frac{\partial f_1}{\partial \beta_u}\right|_{\boldsymbol{\beta}_0} \\ \left.\frac{\partial f_2}{\partial \beta_1}\right|_{\boldsymbol{\beta}_0} & \cdots & \left.\frac{\partial f_2}{\partial \beta_u}\right|_{\boldsymbol{\beta}_0} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial f_n}{\partial \beta_1}\right|_{\boldsymbol{\beta}_0} & \cdots & \left.\frac{\partial f_n}{\partial \beta_u}\right|_{\boldsymbol{\beta}_0} \end{pmatrix}, \quad \boldsymbol{y}^* = [y_1^*, \ldots, y_n^*]^T,$$

$$\boldsymbol{y}_0 = [f_{10}, \ldots, f_{n0}]^T, \quad \boldsymbol{e} = [e_1, \ldots, e_n]^T, \quad \text{and} \quad \Delta\boldsymbol{y} = [y_1^* - f_{10}, \ldots, y_n^* - f_{n0}]^T.$$

The unknown vector $\Delta\boldsymbol{\beta}$ of corrections to the initial parameter vector $\boldsymbol{\beta_0}$ has to be estimated iteratively by introducing the sum of $\boldsymbol{\beta}_0$ and the estimate $\Delta\widehat{\boldsymbol{\beta}}$ as the initial parameter vector in the next iteration step, i.e.,

$$\boldsymbol{\beta}_{it+1,0} = \boldsymbol{\beta}_{it,0} + \Delta\widehat{\boldsymbol{\beta}}_{it} \tag{4.4}$$

where $it$ denotes the iteration step. The estimated observations $\boldsymbol{y}_0$ and the partial derivatives in $\boldsymbol{X}$ also have to be recomputed according to $\boldsymbol{\beta}_{it+1,0}$.

## 4.2 Gauss-Markov model

At each iteration step, the unknown vector $\Delta\boldsymbol{\beta}$ can be estimated within a Gauss-Markov model, which is defined as (see e.g., Koch, 1999)

$$\underbrace{\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y} + \boldsymbol{e}}_{\text{functional part}} \quad \text{with} \quad \underbrace{D(\boldsymbol{e}) = D(\boldsymbol{y}) = \sigma^2 \boldsymbol{P}^{-1}}_{\text{stochastic part}} \tag{4.5}$$

where

| | |
|---|---|
| $\boldsymbol{X}$ | $n \times u$ coefficient matrix, |
| $\boldsymbol{\beta}$ | $u \times 1$ vector of unknown, |
| $\boldsymbol{y}$ | $n \times 1$ vector of observations, $\boldsymbol{y} = [y_1, \ldots, y_n]^T$, |
| $\sigma^2$ | unknown variance factor or variance of unit weight, |
| $\boldsymbol{P}$ | given weight matrix which must be positive definite. |

$E(\cdot)$ and $D(\cdot)$ denote the expectation and the covariance operator, respectively. The left set of equations $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y} + \boldsymbol{e}$ is called observation equations. By substituting $\boldsymbol{\beta} = \Delta\boldsymbol{\beta}$ and $\boldsymbol{y} = \Delta\boldsymbol{y}$, the linearized model (4.3) is reformulated as the functional part of Eq. (4.5). Usually, the number of observations $n$ is larger than the number of unknowns $u$, i.e., $n > u$. Thus, an overdetermined system is given to reduce the influence of randomness of the observations on the estimates of $\boldsymbol{\beta}$ (Koch, 1999, 2007).

**Stochastic model**
The covariance matrix $D(\boldsymbol{y})$ (also called the variance-covariance matrix or the dispersion matrix) is defined as

$$D(\boldsymbol{y}) = E\left[(\boldsymbol{y} - E(\boldsymbol{y}))(\boldsymbol{y} - E(\boldsymbol{y}))^T\right] = \begin{bmatrix} \sigma_{y_1}^2 & \sigma_{y_1 y_2} & \cdots & \sigma_{y_1 y_n} \\ \sigma_{y_2 y_1} & \sigma_{y_2}^2 & \cdots & \sigma_{y_2 y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{y_n y_1} & \sigma_{y_n y_2} & \cdots & \sigma_{y_n}^2 \end{bmatrix} \tag{4.6}$$

which is a symmetric matrix, i.e., $\sigma_{y_i y_j} = \sigma_{y_j y_i}$ $(i, j \in \{1, \ldots, n\})$. The diagonal elements denote the variances of the individual observations, whereas the off-diagonal elements indicate covariances between the observations. The covariances reflect the correlations; the correlation coefficient $\rho_{y_i y_j}$ of the observations $y_i$ and $y_j$ is defined as

$$\rho_{y_i y_j} = \sigma_{y_i y_j} / (\sigma_{y_i} \cdot \sigma_{y_j}) \,. \tag{4.7}$$

When $\sigma_{y_i y_j} = \rho_{y_i y_j} = 0$, the observations $y_i$ and $y_j$ are said to be uncorrelated, i.e., independent. The weight matrix $\boldsymbol{P}$ is then a diagonal matrix. If all observations are uncorrelated and of equal weight, $\boldsymbol{P}$ can be chosen as the identity matrix $\mathbf{I}$.

### 4.2.1 Method of least squares for Gauss-Markov model

Since nonlinear relations may generally be transferred into a linear model via a linearization, the adjustment theory will be introduced based on the Gauss-Markov model (4.5) in the following. The method of least squares means to find $\boldsymbol{\beta}$ such that the quadratic function $S(\boldsymbol{\beta})$ is minimized, i.e.,

$$\begin{aligned} \min_{\boldsymbol{\beta} \in \mathbb{R}^u} \quad S(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{P} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) / \sigma^2 \\ &= (\boldsymbol{y}^T \boldsymbol{P} \boldsymbol{y} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{P} \boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{P} \boldsymbol{X} \boldsymbol{\beta}) / \sigma^2 \end{aligned} \tag{4.8}$$

which is obtained by setting the gradient $\nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$ equal to zero, i.e.,

$$\nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = \left[\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_u}\right]^T = (-2\boldsymbol{X}^T \boldsymbol{P} \boldsymbol{y} + 2\boldsymbol{X}^T \boldsymbol{P} \boldsymbol{X} \boldsymbol{\beta}) / \sigma^2 = \mathbf{0} \,. \tag{4.9}$$

This leads to the normal equations

$$\underbrace{\boldsymbol{X}^T \boldsymbol{P} \boldsymbol{X}}_{\boldsymbol{N}} \widehat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}^T \boldsymbol{P} \boldsymbol{y}}_{\boldsymbol{b}} \tag{4.10}$$

where $\widehat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$, the matrix $\boldsymbol{N}$ is called the matrix of normal equations, and $\boldsymbol{b}$ is the right-hand side of the normal equations.

When the coefficient matrix $\boldsymbol{X}$ is of full column rank (see Section 4.3 later), i.e., $\operatorname{rank}\boldsymbol{X} = u$, $\boldsymbol{N}$ is invertible and a unique estimate $\widehat{\boldsymbol{\beta}}$ is obtained through

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{P} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{P} \boldsymbol{y} = \boldsymbol{N}^{-1} \boldsymbol{b}. \tag{4.11}$$

Substituting the estimate $\widehat{\boldsymbol{\beta}}$ into the observation equations in (4.5), the vector $\widehat{\boldsymbol{e}}$ of residuals is obtained as

$$\widehat{\boldsymbol{e}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y}\,. \tag{4.12}$$

The variance factor can then be estimated by the weighted sum of squares of the residuals $\Omega$

$$\Omega = \widehat{\boldsymbol{e}}^T\,\boldsymbol{P}\,\widehat{\boldsymbol{e}}\,. \tag{4.13}$$

Since the expected value $E(\Omega)$ fulfills (cf. Koch, 2007)

$$E(\Omega) = \sigma^2\,(n-u) \tag{4.14}$$

the unbiased estimate $\widehat{\sigma}^2$ of $\sigma^2$ can be obtained by

$$\widehat{\sigma}^2 = \frac{\Omega}{r} \tag{4.15}$$

with the degree of freedom (i.e., the number of redundant equations in the model)

$$r = n - u\,. \tag{4.16}$$

The covariance matrix $D(\widehat{\boldsymbol{\beta}})$ of the estimate $\widehat{\boldsymbol{\beta}}$ can then be calculated in order to extract the error information of the estimate. Applying the law of error propagation to Eq. (4.11), one can obtain $D(\widehat{\boldsymbol{\beta}})$ by

$$D(\widehat{\boldsymbol{\beta}}) = \sigma^2\left(\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X}\right)^{-1} = \sigma^2\,\boldsymbol{N}^{-1}. \tag{4.17}$$

Inserting $\widehat{\sigma}^2$ from Eq. (4.15) into the equation above, we can thus obtain $\widehat{D}(\widehat{\boldsymbol{\beta}})$ by

$$\widehat{D}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2\left(\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X}\right)^{-1} = \widehat{\sigma}^2\boldsymbol{N}^{-1}. \tag{4.18}$$

According to the definition of the covariance matrix (cf. Eq. (4.6)), $\widehat{D}(\widehat{\boldsymbol{\beta}})$ can be written as

$$\widehat{D}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2\begin{bmatrix} \sigma^2_{\beta_1} & \sigma_{\beta_1\beta_2} & \dots & \sigma_{\beta_1\beta_u} \\ \sigma_{\beta_2\beta_1} & \sigma^2_{\beta_2} & \dots & \sigma_{\beta_2\beta_u} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\beta_u\beta_1} & \sigma_{\beta_u\beta_2} & \dots & \sigma^2_{\beta_u} \end{bmatrix}. \tag{4.19}$$

The estimate $\widehat{\sigma}^2_{\beta_j}$ of the variance of the $j$-th unknown $\beta_j$ is the $j$-th diagonal element of $\widehat{D}(\widehat{\boldsymbol{\beta}})$, i.e.,

$$\widehat{\sigma}^2_{\beta_j} = \widehat{\sigma}^2 \cdot \sigma^2_{\beta_j}. \tag{4.20}$$

Similar to Eq. (4.7), the correlation coefficients $\rho_{\beta_i\beta_j}$ $(i,j \in \{1,\dots,u\})$ between the $i$-th and $j$-th unknown parameters can be estimated by

$$\rho_{\beta_i\beta_j} = \sigma_{\beta_i\beta_j}/(\sigma_{\beta_i}\cdot\sigma_{\beta_j}) \tag{4.21}$$

which can then be collected in the so-called correlation matrix, written as

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{\beta_1\beta_2} & \dots & \rho_{\beta_1\beta_u} \\ \rho_{\beta_2\beta_1} & 1 & \dots & \rho_{\beta_2\beta_u} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{\beta_u\beta_1} & \rho_{\beta_u\beta_2} & \dots & 1 \end{bmatrix}. \tag{4.22}$$

Since $\sigma_{\beta_{j_1}\beta_{j_2}} = \sigma_{\beta_{j_2}\beta_{j_1}}$, the matrix $\boldsymbol{R}$ is also symmetric. Note, that for a linearized problem as given in Eq. (4.2), covariance and correlation matrices should be taken from the computations according to the Eqs. (4.18) and (4.22) when convergency is reached.

### 4.2.2 Gauss-Markov model with equality constraints

Assume the Gauss-Markov model (4.5) is subject to equality constraints

$$\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{w} \tag{4.23}$$

where $\boldsymbol{H}$ is a $p \times u$ matrix of known coefficients with rank$\boldsymbol{H} = p$ and $p < u$, and $\boldsymbol{w}$ is a given $p \times 1$ vector.

Solving for the unknown vector $\boldsymbol{\beta}$ by the method of least squares means to find a solution $\boldsymbol{\beta}$ such that $S(\boldsymbol{\beta})$ (cf. Eq. (4.8)) is minimized subject to Eq. (4.23). The Lagrangian function reads [Koch, 1999]

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{k}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{P} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) / \sigma^2 + 2\boldsymbol{k}^T (\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{w})/\sigma^2 \tag{4.24}$$

where the $p \times 1$ vector $2\boldsymbol{k}/\sigma^2$ contains the Lagrange multipliers. Equation (4.24) must be minimized by setting the gradients of $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{k})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{k}$ both equal to zero

$$\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{k}) = \left(-2\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} + 2\boldsymbol{H}^T\boldsymbol{k}\right)/\sigma^2 = \boldsymbol{0} \tag{4.25a}$$

$$\nabla_{\boldsymbol{k}} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{k}) = 2(\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{w})/\sigma^2 = \boldsymbol{0}\,. \tag{4.25b}$$

This leads to the extended normal equations

$$\underbrace{\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X} & \boldsymbol{H}^T \\ \boldsymbol{H} & \boldsymbol{0} \end{bmatrix}}_{\widetilde{\boldsymbol{N}}} \begin{bmatrix} \widetilde{\boldsymbol{\beta}} \\ \boldsymbol{k} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{P}\boldsymbol{y} \\ \boldsymbol{w} \end{bmatrix} \tag{4.26}$$

where $\widetilde{\boldsymbol{\beta}}$ is the estimate of the unknown parameter vector $\boldsymbol{\beta}$ subject to the constraints (4.23). Since $\boldsymbol{N} = \boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X}$ is invertible and we have rank$\boldsymbol{H} = p$, it follows that $\det(\widetilde{\boldsymbol{N}}) \neq 0$[1]. Therefore, the values for $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{k}$ are uniquely determined. Similar to Eq. (4.15), the unbiased estimate $\widetilde{\sigma}^2$ of the variance factor $\sigma^2$ follows as [Koch, 1999]

$$\widetilde{\sigma}^2 = \frac{\Omega_H}{r'} \quad \text{with} \quad \Omega_H = \widetilde{\boldsymbol{e}}^T \boldsymbol{P} \widetilde{\boldsymbol{e}} \tag{4.27}$$

where the vector $\widetilde{\boldsymbol{e}}$ of residuals is obtained from

$$\widetilde{\boldsymbol{e}} = \boldsymbol{X}\widetilde{\boldsymbol{\beta}} - \boldsymbol{y} \tag{4.28}$$

and the degree of freedom $r'$ reads

$$r' = n + p - u\,. \tag{4.29}$$

The covariance matrix of the estimated parameter vector $\widetilde{\boldsymbol{\beta}}$ can be extracted from the inverse of the extended normal equation matrix $\widetilde{\boldsymbol{N}}$ following Eq. (4.18).

[1] Applying the rule that the determinant $\det\left(\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}\right) = \det(\boldsymbol{A}) \det(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})$ when $\boldsymbol{A}$ is invertible (cf. Brookes, 2011).

## 4.3 Rank deficiency and ill-conditioning

So far, the coefficient matrix $\boldsymbol{X}$ has been assumed to be of full column rank. When rank$\boldsymbol{X} < u$, $\boldsymbol{X}$ is not of full column rank, i.e., rank deficient. The relation rank$\boldsymbol{N}$ = rank$(\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X})$ = rank$\boldsymbol{X} < u$ holds, and therefore $\boldsymbol{N}$ is singular and the inverse does not exist. Rank deficiency implies that there exist rows or columns which are linear combinations of some or all of the remaining rows and columns. As for an ill-conditioned system of linear equations, the solution is highly sensitive to the values of the observations.

### 4.3.1 Singular Value Decomposition (SVD)

A very powerful numerical tool to analyze rank deficient and/or ill-conditioned systems is the Singular Value Decomposition (SVD). The SVD of a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times u}$ (assume $n \geq u$) reads (e.g., Golub and Van Loan, 1996)

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \sum_{m=1}^{u} \boldsymbol{u}_m \, s_m \, \boldsymbol{v}_m^T \tag{4.30}$$

where

| | |
|---|---|
| $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n] \in \mathbb{R}^{n\times n}$ | $n \times n$ orthogonal matrix with orthonormal columns $\boldsymbol{u}_i$ $(i = 1, \ldots, n)$, with $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \mathbf{I}$, |
| $\boldsymbol{u}_i$ | called left singular vector, with $\boldsymbol{u}_{i_1}^T \boldsymbol{u}_{i_2} = \delta_{i_1 i_2} = \begin{cases} 0 & i_1 \neq i_2 \\ 1 & i_1 = i_2 \end{cases}$ $(i_1, i_2 \in \{1, \ldots, n\})$, |
| $\boldsymbol{S} \in \mathbb{R}^{n\times u}$ | $n \times u$ matrix with non-negative elements $s_m$ called singular values on its main diagonal $(m = 1, \ldots, u)$, |
| $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_u] \in \mathbb{R}^{u\times u}$ | $u \times u$ orthogonal matrix with orthonormal columns $\boldsymbol{v}_j$, with $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^T = \mathbf{I}$, |
| $\boldsymbol{v}_j$ | called right singular vector, with $\boldsymbol{v}_{j_1}^T \boldsymbol{v}_{j_2} = \delta_{j_1 j_2}$ $(j_1, j_2 \in \{1, \ldots, u\})$. |

The singular values $s_m$ along the main diagonal of $\boldsymbol{S}$ appear in a decreasing order, i.e.,

$$s_1 \geq s_2 \geq \cdots \geq s_u \geq 0. \tag{4.31}$$

The rank of the matrix $\boldsymbol{X}$ equals to the number of non-zero singular values, which means, if there exist singular values that are equal to zero, the matrix $\boldsymbol{X}$ is not of full column rank and is said to be rank deficient. Additionally, very small singular values will cause instabilities in the least squares estimation since it can amplify the noise in the data significantly; see e.g., Hansen [1990].

### 4.3.2 Condition number

A measure of the instability of the solution of the linear equation system is the condition number. Assume that the coefficient matrix $\boldsymbol{X}$ is square and invertible (i.e., nonsingular), the condition number of $\boldsymbol{X}$ is defined as (e.g., Higham, 1987)

$$\text{cond}(\boldsymbol{X}) = \|\boldsymbol{X}\| \cdot \|\boldsymbol{X}^{-1}\| \tag{4.32}$$

where $\|\cdot\|$ denotes a matrix norm[1].

Based on the rule that $\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|$, one can easily obtain that

$$\text{cond}(\boldsymbol{X}) = \|\boldsymbol{X}\| \cdot \|\boldsymbol{X}^{-1}\| \geq \|\boldsymbol{X}\boldsymbol{X}^{-1}\| = \|\mathbf{I}\| = 1. \tag{4.33}$$

Therefore, the condition number is always not smaller than one. When the condition number is not too much larger than one, the system of equations is said to be well-conditioned, otherwise the system is ill-conditioned. A singular matrix has condition number of infinity. In the following we will see how the condition number reflects the sensitivity of the solution to perturbations in the data.

Assume that the data distorted by the error vector $\boldsymbol{e}$ will cause a change of $\delta\boldsymbol{\beta}$ in the corresponding solution, i.e.,

$$\boldsymbol{X}(\boldsymbol{\beta} + \delta\boldsymbol{\beta}) = \boldsymbol{y} + \boldsymbol{e} \tag{4.34}$$

or written as the two equations

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}, \tag{4.35a}$$

$$\boldsymbol{X}\delta\boldsymbol{\beta} = \boldsymbol{e} \Rightarrow \delta\boldsymbol{\beta} = \boldsymbol{X}^{-1}\boldsymbol{e}\,. \tag{4.35b}$$

Taking the norm of the two equations above and applying the rule $\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|$ yields

$$\|\boldsymbol{y}\| = \|\boldsymbol{X}\boldsymbol{\beta}\| \leq \|\boldsymbol{X}\| \cdot \|\boldsymbol{\beta}\|, \tag{4.36a}$$

$$\|\delta\boldsymbol{\beta}\| = \|\boldsymbol{X}^{-1}\boldsymbol{e}\| \leq \|\boldsymbol{X}^{-1}\| \cdot \|\boldsymbol{e}\|\,. \tag{4.36b}$$

Multiplying the corresponding sides of the two equations and then dividing both sides by $\|\boldsymbol{\beta}\| \cdot \|\boldsymbol{y}\|$ gives

$$\frac{\|\delta\boldsymbol{\beta}\|}{\|\boldsymbol{\beta}\|} \leq \underbrace{\|\boldsymbol{X}\| \cdot \|\boldsymbol{X}^{-1}\|}_{\text{cond}(\boldsymbol{X})} \cdot \frac{\|\boldsymbol{e}\|}{\|\boldsymbol{y}\|} \tag{4.37}$$

which implies that the relative error $\|\delta\boldsymbol{\beta}\|/\|\boldsymbol{\beta}\|$ of the solution can be amplified by a factor of $\text{cond}(\boldsymbol{X})$ with respect to the relative error $\|\boldsymbol{e}\|/\|\boldsymbol{y}\|$ of the observations. Therefore, if the condition number is very large, a small perturbation in the data vector $\boldsymbol{y}$ can cause a large deviation in the estimate of $\boldsymbol{\beta}$.

When the Euclidean norm ($\ell_2$ norm) is used, the condition number can be formulated in terms of the singular values [Golub and Van Loan, 1996]

$$\text{cond}(\boldsymbol{X}) = \|\boldsymbol{X}\|_2 \cdot \|\boldsymbol{X}^{-1}\|_2 = \frac{s_1}{s_u} \tag{4.38}$$

[1]For example, the $p$-norm $\|\cdot\|_p$ with $p = 1, 2, \ldots, \infty$ can be used. The $p$-norms of a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ can be obtained from [Golub and Van Loan, 1996]

$$\|\boldsymbol{A}\|_p = \max_{\|\boldsymbol{x}\|_p=1} \|\boldsymbol{A}\boldsymbol{x}\|_p$$

with vector $\boldsymbol{x} \in \mathbb{R}^n$. The $p$-norms of the vector $\boldsymbol{x}$ is defined as

$$\|\boldsymbol{x}\|_p = (|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}}.$$

where the smallest singular value $s_u > 0$ holds due to the nonsingularity of $\boldsymbol{X}$. In other words, the condition number of nonsingular matrix $\boldsymbol{X}$ is equal to the ratio between the largest and the smallest singular values of $\boldsymbol{X}$. Since the method of least squares is applied to solve for the unknown parameter vector $\boldsymbol{\beta}$, it makes more sense to consider the condition of the matrix $\boldsymbol{N}$ in Eq. (4.10). In a special case that the weight matrix $\boldsymbol{P} = \mathbf{I}$, $\boldsymbol{N}$ can be written as

$$\boldsymbol{N} = \boldsymbol{X}^T\boldsymbol{X} = (\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T)^T(\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T) = \boldsymbol{V}\boldsymbol{S}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \boldsymbol{V}\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{V}^T = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T \tag{4.39}$$

under consideration of Eq. (4.30). Herein $\boldsymbol{\Lambda}$ is the diagonal matrix whose diagonal elements are called the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$. Following Eq. (4.38), the condition number of $\boldsymbol{N}$ can be obtained by

$$\text{cond}(\boldsymbol{N}) = \|\boldsymbol{N}\| \cdot \|\boldsymbol{N}^{-1}\| = \frac{s_1^2}{s_u^2} = \frac{\lambda_1}{\lambda_u} \tag{4.40}$$

where $\lambda_1$ and $\lambda_u$ denote the maximum and the minimum eigenvalues and both values are larger than zero[1]; see e.g., Kalman [1996], Strang and Borre [1997] or Naeimi [2013] for details.

## 4.4 Regularization

To solve rank deficient and/or ill-conditioned problems, the method of regularization is often applied. A detailed overview of different regularization methods can be found in, e.g., Bouman [1998, 2000]. The most well-known and commonly used method is probably the Tikhonov regularization (Phillips, 1962; Tikhonov, 1963; Tikhonov and Arsenin, 1977), which stabilizes the solution by adding a penalty term to the normal equation matrix. The penalty term is a positive definite matrix times the regularization parameter which is generally unknown. The Tikhonov regularization can be interpreted by Bayesian statistics as estimation with prior information [Koch, 1999, 2007; Koch and Kusche, 2002, and references therein].

Within this thesis, we introduce the prior information of $\boldsymbol{\beta}$ to stabilize the estimation process, i.e., the $u \times 1$ expectation vector $E(\boldsymbol{\beta}) = \boldsymbol{\mu}$ and the positive definite covariance matrix $D(\boldsymbol{\beta}) = \boldsymbol{\Sigma}_\beta$ are given. Considering the prior information as an additional observation technique, we can formulate the linear model

$$\boldsymbol{\beta} = \boldsymbol{\mu} + \boldsymbol{e}_\mu \quad \text{with} \quad D(\boldsymbol{\mu}) = \sigma_\mu^2 \boldsymbol{P}_\mu^{-1} \tag{4.41}$$

with the error vector $\boldsymbol{e}_\mu$ of the prior information, the unknown variance factor $\sigma_\mu^2$, and the given $u \times u$ positive definite weight matrix $\boldsymbol{P}_\mu = \boldsymbol{\Sigma}_\beta^{-1}$.

Assume that the prior information $\boldsymbol{\mu}$ and the observation vector $\boldsymbol{y}$ from Eq. (4.5) are independent, we can formulate [Koch, 2007]

$$\overline{\boldsymbol{X}}\boldsymbol{\beta} = \overline{\boldsymbol{y}} + \overline{\boldsymbol{e}} \tag{4.42}$$

$$\text{with} \quad \overline{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X} \\ \mathbf{I} \end{bmatrix}, \overline{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{\mu} \end{bmatrix}, \overline{\boldsymbol{e}} = \begin{bmatrix} \boldsymbol{e} \\ \boldsymbol{e}_\mu \end{bmatrix},$$

$$\text{and} \quad D(\overline{\boldsymbol{y}}) = \sigma^2 \begin{bmatrix} \boldsymbol{P}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \sigma_\mu^2 \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{P}_\mu^{-1} \end{bmatrix}.$$

[1] $\boldsymbol{N}$ is a positive definite matrix since $\boldsymbol{X}$ is invertible. All eigenvalues of a positive definite matrix are positive.

Substituting Eq. (4.42) in Eq. (4.10), the normal equations

$$\left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X}+\frac{1}{\sigma_\mu^2}\boldsymbol{P}_\mu\right)\widehat{\boldsymbol{\beta}}=\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{y}+\frac{1}{\sigma_\mu^2}\boldsymbol{P}_\mu\boldsymbol{\mu} \tag{4.43}$$

can be obtained. By introducing $\overline{\lambda}=\sigma^2/\sigma_\mu^2$, Eq. (4.43) can be reformulated as

$$\left(\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X}+\overline{\lambda}\boldsymbol{P}_\mu\right)\widehat{\boldsymbol{\beta}}=\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{y}+\overline{\lambda}\boldsymbol{P}_\mu\boldsymbol{\mu} \tag{4.44}$$

which leads to the Tikhonov regularization with the Tikhonov regularization parameter $\overline{\lambda}$ if $\boldsymbol{\mu}$ is assumed to be the zero vector. Consequently, the Tikhonov regularization parameter can be estimated through Variance Component Estimation (VCE) (Koch, 1999, 2007; Koch and Kusche, 2002; see Section 4.5), since $\overline{\lambda}$ is the ratio of $\sigma^2$ and $\sigma_\mu^2$. Small values of $\sigma_\mu^2$ give strong regularizations, whereas large values lead to weak regularizations. Many other strategies are also available for choosing $\overline{\lambda}$, such as the L-curve criterion [Hansen and O'leary, 1993] and the Generalized Cross Validation (GCV) [Golub *et al.*, 1979]. However, these different methods are beyond the scope of the thesis; the interested reader may refer to, e.g., Hansen [1996] or Xu *et al.* [2016].

## 4.5 Data combination and Variance Component Estimation (VCE)

Usually, different types (based on e.g., techniques or missions) of observations are combined to solve the unknown model parameters. The variance factors belonging to different types of observations are called variance components. The Gauss-Markov model (4.5) can be established individually for each observation type

$$\boldsymbol{X}_o\boldsymbol{\beta}=\boldsymbol{y}_o+\boldsymbol{e}_o \quad \text{with} \quad D(\boldsymbol{y}_o)=\sigma_o^2\,\boldsymbol{P}_o^{-1} \tag{4.45}$$

where $o\in\{1,\ldots,\overline{c}\}$ denotes the different observation techniques with total number $\overline{c}$.

If the different types of observations $\boldsymbol{y}_o$ and the prior information $\boldsymbol{\mu}$ from Eq. (4.41) are assumed to be independent, the combination of the $\overline{c}+1$ observation types yields the following extended linear model similar to Eq. (4.42) (cf. Koch and Kusche, 2002)

$$\underbrace{\begin{bmatrix}\boldsymbol{X}_1\\ \vdots\\ \boldsymbol{X}_{\overline{c}}\\ \mathbf{I}\end{bmatrix}}_{\boldsymbol{X}}\boldsymbol{\beta}=\underbrace{\begin{bmatrix}\boldsymbol{y}_1\\ \vdots\\ \boldsymbol{y}_{\overline{c}}\\ \boldsymbol{\mu}\end{bmatrix}}_{\boldsymbol{y}}+\underbrace{\begin{bmatrix}\boldsymbol{e}_1\\ \vdots\\ \boldsymbol{e}_{\overline{c}}\\ \boldsymbol{e}_\mu\end{bmatrix}}_{\boldsymbol{e}} \tag{4.46}$$

with

$$D(\boldsymbol{y})=D\left(\begin{bmatrix}\boldsymbol{y}_1\\ \vdots\\ \boldsymbol{y}_{\overline{c}}\\ \boldsymbol{\mu}\end{bmatrix}\right)=\sigma_1^2\boldsymbol{M}_1+\ldots+\sigma_{\overline{c}}^2\boldsymbol{M}_{\overline{c}}+\sigma_\mu^2\boldsymbol{M}_\mu$$

and

$$\boldsymbol{M}_1 = \begin{bmatrix} \boldsymbol{P}_1^{-1} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \cdots, \boldsymbol{M}_{\bar{c}} = \begin{bmatrix} \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{P}_{\bar{c}}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \boldsymbol{M}_{\mu} = \begin{bmatrix} \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{P}_{\mu}^{-1} \end{bmatrix}.$$

Similarly, the following normal equations

$$\left( \sum_{o=1}^{\bar{c}} \frac{1}{\sigma_o^2} \boldsymbol{X}_o^T \boldsymbol{P}_o \boldsymbol{X}_o + \frac{1}{\sigma_\mu^2} \boldsymbol{P}_\mu \right) \widehat{\boldsymbol{\beta}} = \sum_{o=1}^{\bar{c}} \frac{1}{\sigma_o^2} \boldsymbol{X}_o^T \boldsymbol{P}_o \boldsymbol{y}_o + \frac{1}{\sigma_\mu^2} \boldsymbol{P}_\mu \boldsymbol{\mu} \tag{4.47}$$

are obtained. When the different types of observations are dependent, $D(\boldsymbol{y})$ is no longer a diagonal block matrix and unknown covariance matrices between different types of observations have to be taken into account and estimated as well [Koch, 1999].

Since the weight matrices $\boldsymbol{P}_o$ are positive definite, the matrices $\boldsymbol{X}_o^T \boldsymbol{P}_o \boldsymbol{X}_o$ are at least positive semidefinite. Furthermore, the weight matrix $\boldsymbol{P}_\mu$ of the prior information is positive definite, and therefore, the matrix of normal equations

$$\boldsymbol{N}_c = \frac{1}{\sigma_1^2} \boldsymbol{X}_1^T \boldsymbol{P}_1 \boldsymbol{X}_1 + \ldots + \frac{1}{\sigma_{\bar{c}}^2} \boldsymbol{X}_{\bar{c}}^T \boldsymbol{P}_{\bar{c}} \boldsymbol{X}_{\bar{c}} + \frac{1}{\sigma_\mu^2} \boldsymbol{P}_\mu \tag{4.48}$$

is regular, i.e., invertible, if appropriate variance components $\sigma_o^2$ ($o = \{1, \ldots, \bar{c}\}$) and $\sigma_\mu^2$ are chosen. The estimate $\widehat{\boldsymbol{\beta}}$ of the unknown vector $\boldsymbol{\beta}$ in case of the model (4.46) according to Eq. (4.11) is then obtained by

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{N}_c^{-1} \left( \frac{1}{\sigma_1^2} \boldsymbol{X}_1^T \boldsymbol{P}_1 \boldsymbol{y}_1 + \ldots + \frac{1}{\sigma_{\bar{c}}^2} \boldsymbol{X}_{\bar{c}}^T \boldsymbol{P}_{\bar{c}} \boldsymbol{y}_{\bar{c}} + \frac{1}{\sigma_\mu^2} \boldsymbol{P}_\mu \boldsymbol{\mu} \right). \tag{4.49}$$

The unknown variance components can be chosen manually, i.e., empirically, or estimated within the Variance Component Estimation (VCE). The VCE is an effective approach to estimate the unknown variance components when different data types have to be combined in a parameter estimation. There are a number of methods to estimate variance components [Xu *et al.*, 2006, and references therein], here the method introduced by Koch and Kusche [2002] will be given briefly.

The estimate $\widehat{\sigma}_o^2$ of $\sigma_o^2$ and the estimate $\widehat{\sigma}_\mu^2$ of $\sigma_\mu^2$ can be obtained through

$$\widehat{\sigma}_o^2 = \frac{\widehat{\boldsymbol{e}}_o^T \boldsymbol{P}_o \widehat{\boldsymbol{e}}_o}{r_o}, \quad \text{with} \quad o \in \{1, \ldots, \bar{c}\}, \quad \widehat{\boldsymbol{e}}_o = \boldsymbol{X}_o \widehat{\boldsymbol{\beta}} - \boldsymbol{y}_o \tag{4.50}$$

and

$$\widehat{\sigma}_\mu^2 = \frac{\widehat{\boldsymbol{e}}_\mu^T \boldsymbol{P}_\mu \widehat{\boldsymbol{e}}_\mu}{r_\mu} \quad \text{with} \quad \widehat{\boldsymbol{e}}_\mu = \widehat{\boldsymbol{\beta}} - \boldsymbol{\mu} \tag{4.51}$$

where $\widehat{\boldsymbol{e}}_o$ and $\widehat{\boldsymbol{e}}_\mu$ denote the vectors of residuals. The denominators $r_o$ and $r_\mu$ are the partial redundancies, namely, the contributions of $\boldsymbol{y}_o$ and $\boldsymbol{\mu}$ to the total redundancy $r_c$ of the model (4.46). The total redundancy $r_c$ is defined as

$$r_c = \sum_{o=1}^{\bar{c}} n_o + u - u = \sum_{o=1}^{\bar{c}} n_o \tag{4.52}$$

where $n_o$ denotes the number of observations collected in the vector $\boldsymbol{y}_o$. Let $\sum_{o=1}^{\bar{c}} n_o = n$, the relation

$$\sum_{o=1}^{\bar{c}} r_o + r_\mu = r_c = n \tag{4.53}$$

holds based on Eq. (4.52). The partial redundancies $r_o$ and $r_\mu$ can be obtained by

$$r_o = n_o - \mathrm{Tr}\left(\frac{1}{\sigma_o^2}\boldsymbol{X}_o^T\boldsymbol{P}_o\boldsymbol{X}_o\boldsymbol{N}_c^{-1}\right), \tag{4.54a}$$

$$r_\mu = u - \mathrm{Tr}\left(\frac{1}{\sigma_\mu^2}\boldsymbol{P}_\mu\boldsymbol{N}_c^{-1}\right) \tag{4.54b}$$

where the symbol Tr denotes the trace. As can be seen from the Eqs. (4.49), (4.50) and (4.51), the estimated variance components are depending on the observation residuals estimated from the vector $\widehat{\boldsymbol{\beta}}$, which in turn depends on the estimated variance components. Thus, the estimation of variance components has to be performed iteratively. For very large systems of normal equations, computations of the inverse $\boldsymbol{N}_c^{-1}$ in Eq. (4.54) are rather expensive. A stochastic trace estimator can then be used instead (e.g., Hutchinson, 1990; Koch and Kusche, 2002). It is worth noting that, as VCE interprets the prior information as an additional noisy observation technique, if the residual vector $\widehat{\boldsymbol{e}}_\mu$ as defined in Eq. (4.51) has no random character, then it cannot be used to compute a reliable estimation of the variance component $\sigma_\mu^2$, and thus for the regularization parameter.

## 4.6 Hypothesis testing

A hypothesis is a claim or a statement about a characteristic of one or more populations. Hypothesis testing is used to accept or reject a hypothesis under a given error probability. It can be done to test the significance of the model parameters. For simplifications, we will introduce the hypothesis testing based on the Gauss-Markov model (4.5) of full rank.

When a single parameter of the model is tested, we can set up the null hypothesis $H_0$ and the alternative hypothesis $H_1$

$$H_0 : \beta_j = \beta_{oj} \quad \text{for a fixed} \quad j \in \{1, \ldots, u\} \qquad \text{versus} \qquad H_1 : \beta_j \neq \beta_{oj} \tag{4.55}$$

where $\beta_{oj}$ is the given value. The test statistic $T$ is defined as [Koch, 1999]

$$T = \frac{(\widehat{\beta}_j - \beta_{oj})^2}{\widehat{\sigma}_{\beta_j}^2} \tag{4.56}$$

if $\widehat{\sigma}_{\beta_j}^2 \neq 0$ where $\widehat{\sigma}_{\beta_j}^2$ is calculable from Eq. (4.20). If the null hypothesis is true, the distribution of $T$ follows the $F$-distribution[1] [Phillips, 1982] by

$$T \sim F(1, n' - q) \tag{4.57}$$

[1]The $F$-distribution with $\nu_1$ degree of freedom in the numerator and $\nu_2$ degree of freedom in the denominator refers to the variable derived from the ratio of two independent chi-squared distributed random variables $u_1 \sim \chi^2(\nu_1)$ and $u_2 \sim \chi^2(\nu_2)$ with degrees of freedom $\nu_1$ and $\nu_2$, respectively. It is written as $\frac{u_1/\nu_1}{u_2/\nu_2} \sim F(\nu_1, \nu_2)$.

where $n'$ denotes the total number of observations in the model and $q = \text{rank}\boldsymbol{X} = u$. Then, the calculated value of $T$ is compared to a critical value from the $F$-distribution for a specific level $\gamma$ of significance (e.g., $\gamma = 0.1, 0.05, 0.01$), and the decision rule is

$$\begin{cases} \text{if } T \geq F(1-\gamma; 1, n'-q) & \text{reject } H_0; \\ \text{if } T < F(1-\gamma; 1, n'-q) & \text{fail to reject } H_0. \end{cases} \tag{4.58}$$

Note, that a $t$-test can also be used for the null hypothesis (4.55) of a single parameter. Instead of Eq. (4.57), we have $\sqrt{T} \sim t(n'-q)$ with $t(n'-q)$ denoting the Student's t-distribution [Fisher, 1925] with $n'-q$ degree of freedom; the outcome will be the same as for the $F$-test.

We can also set up the joint null hypothesis and the alternative hypothesis

$$\begin{aligned} &H_0 : \beta_j = \beta_{oj} \quad \text{for all} \quad j = j', \ldots, k \quad \text{versus} \\ &H_1 : \beta_j \neq \beta_{oj} \quad \text{for at least one} \quad j \in \{j', \ldots, k\}. \end{aligned} \tag{4.59}$$

The test statistic $T$ can be calculated through [Koch, 1999]

$$T = \frac{1}{(k-j'+1)\,\widehat{\sigma}^2} \left(\widehat{\boldsymbol{\beta}}_{j',\ldots,k} - \boldsymbol{\beta}_{o,\,j',\ldots,k}\right)^T \left(\boldsymbol{\Sigma}_\beta\right)^{-1}_{j',\ldots,k} \left(\widehat{\boldsymbol{\beta}}_{j',\ldots,k} - \boldsymbol{\beta}_{o,\,j',\ldots,k}\right) \tag{4.60}$$

with

$$\left(\boldsymbol{\Sigma}_\beta\right)_{j',\ldots,k} = \left(\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X}\right)^{-1}_{j',\ldots,k} = \begin{bmatrix} \sigma^2_{\beta_{j'}} & \cdots & \sigma_{\beta_{j'}\beta_k} \\ \vdots & \ddots & \vdots \\ \sigma_{\beta_k\beta_{j'}} & \cdots & \sigma^2_{\beta_k} \end{bmatrix} \tag{4.61}$$

which is the part of the covariance matrix in Eq. (4.19) related to the parameters $\beta_j$ ($j = j', \ldots, k$). Note that the estimate $\widehat{\sigma}^2$ of the variance factor of the observation residuals in Eq. (4.19) is moved to the denominator of Eq. (4.60). If the null hypothesis is true, the distribution of $T$ follows by

$$T \sim F(k-j'+1, n'-q)\,. \tag{4.62}$$

The calculated value of $T$ is then compared to a critical value from the $F$-distribution for a specific level $\gamma$ of significance

$$\begin{cases} \text{if } T \geq F(1-\gamma; k-j'+1, n'-q) & \text{reject } H_0; \\ \text{if } T < F(1-\gamma; k-j'+1, n'-q) & \text{fail to reject } H_0. \end{cases} \tag{4.63}$$

If the null hypothesis $H_0$ is not rejected, none of the parameters $\beta_j$ ($j \in \{j', \ldots, k\}$) is significantly different from $\beta_{oj}$; if $H_0$ is rejected, the parameters $\beta_j$ ($j = j', \ldots, k$) are jointly statistically significant and at least one of the parameters is not equal to $\beta_{oj}$. However, it does not indicate which parameter is not equal to $\beta_{oj}$.

For the model (4.46), the testing approach is also valid, only the variance factors and the covariance matrix used to compute $T$ have to be replaced by the corresponding quantities.

## 4.7 Constrained optimization

As introduced previously in Section 4.2.2, equality constraints can easily be incorporated into the adjustment system by the method of Lagrange multipliers. However, when the constraints are

in the form of inequalities, the problems become more complicated and will lead to optimization problems, also known as mathematical programming. Mathematically speaking, optimization is to minimize or maximize a function subject to constraints on its variables, which can be formulated generally as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^u} \quad S(\boldsymbol{\beta}) \tag{4.64a}$$

$$\text{s.t.} \quad \boldsymbol{h}(\boldsymbol{\beta}) = \mathbf{0} \tag{4.64b}$$

$$\boldsymbol{g}(\boldsymbol{\beta}) \geq \mathbf{0} \tag{4.64c}$$

where $\boldsymbol{\beta}$ is the vector of unknowns (optimization variables), $S(\boldsymbol{\beta})$ is the real-valued objective function of $\boldsymbol{\beta}$ to be minimized, the $p \times 1$ vector-valued functions $\boldsymbol{h}(\boldsymbol{\beta}) = [h_1(\boldsymbol{\beta}), \ldots, h_p(\boldsymbol{\beta})]^T$ are the equality constraints, and the $\overline{p} \times 1$ vector-valued functions $\boldsymbol{g}(\boldsymbol{\beta}) = [g_1(\boldsymbol{\beta}), \ldots, g_{\overline{p}}(\boldsymbol{\beta})]^T$ are the inequality constraints. The symbol large than or equal to "$\geq$" in Eq. (4.64c) represents a component-wise operator.

Based on characteristics of the objective function, the constraints and the unknowns, optimization problems can be classified into different categories; see, e.g., Nocedal and Wright [1999] and Kumar [2015] for various taxonomies. It is important to identify the type of the problem firstly, since different optimization algorithms can be suited for a particular type of problems. Within this thesis, we focus on (inequality) constrained and continuous optimization, which means the optimization variables are all real values whereas the objective and constraint functions are continuous. Moreover, both the objective and constraint functions are assumed to be twice continuously differentiable. According to the nature of the involved expressions in the objective function and the constraints, the optimization problems can be classified into Linear Programming (LP), QP, and NLP problems [Dutta, 2016]. Various characteristics of these problems are illustrated in Table 4.1. Note that a QP problem is a special case of the NLP problem.

Table 4.1: Classification of optimization problems based on the nature of the equations involved.

| Classification | Objective function (standard form) | Constraints (equality/inequality) |
|---|---|---|
| LP | linear ($\boldsymbol{c}^T\boldsymbol{\beta}, \quad \boldsymbol{c} \in \mathbb{R}^u$) | linear |
| QP | quadratic ($\boldsymbol{c}^T\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{Q}\boldsymbol{\beta}, \quad \boldsymbol{Q} \in \mathbb{R}^{u\times u}$) | linear |
| NLP | at least one of the functions (objective and constraint functions) are nonlinear | |

The method (4.8) of least squares adjustment in a Gauss-Markov model can actually be formulated as a QP problem, since the objective function $S(\boldsymbol{\beta})$ has the quadratic form. As we work on the nonlinear model (4.1), we can directly set up the objective function in the least squares sense similar to Eq. (4.8) as

$$S(\boldsymbol{\beta}) = (\boldsymbol{y}^* - \boldsymbol{f}(\boldsymbol{\beta}))^T\boldsymbol{P}(\boldsymbol{y}^* - \boldsymbol{f}(\boldsymbol{\beta}))/\sigma^2 \tag{4.65}$$

which should be minimized. The problem (4.64) under consideration of the objective function (4.65) belongs to a NLP problem.

Speaking of constrained optimization, several basic concepts are known. A point that satisfies all $p + \overline{p}$ constraints in the Eqs. (4.64b) and (4.64c) is said to be feasible. The set of all feasible

points forms a feasible set $\mathcal{F}$, i.e.

$$\mathcal{F} = \{\boldsymbol{\beta} | \boldsymbol{h}(\boldsymbol{\beta}) = \mathbf{0}; \quad \boldsymbol{g}(\boldsymbol{\beta}) \geq \mathbf{0}\}. \tag{4.66}$$

At a feasible point, the inequality constraint is said to be active if the equal sign in Eq. (4.64c) holds (i.e., $g_i(\boldsymbol{\beta}) = 0$, $i \in \{1, \ldots, \overline{p}\}$), whereas the inequality constraint is said to be inactive if strict inequality is fulfilled (i.e., $g_i(\boldsymbol{\beta}) > 0$, $i \in \{1, \ldots, \overline{p}\}$).

To solve the constrained NLP problems, there are different methods, such as the penalty and augmented Lagrangian methods (e.g., Rockafellar, 1973; Fiacco and McCormick, 1990) that transform the constrained problem into a sequence of unconstrained ones by minimizing a penalty function that penalizes the constraint violations (i.e., at any infeasible point), the Sequential Quadratic Programming (SQP) methods (e.g., Han, 1977; Powell, 1978) that transform the original problem into a sequence of QP subproblems, and the interior-point methods (also called barrier methods, e.g., Fiacco and McCormick, 1990; Forsgren *et al.*, 2002) where the classical ones[1] transform the constrained problem into a sequence of unconstrained ones by minimizing a barrier function that introduces a barrier on the boundary of the feasible set, which prevents iterates starting from the interior of the feasible set leaving it. All these methods apply quadratic approximations to a function combining the objective function and constraints [Goldsmith, 1999]. An overview of the optimization techniques for NLP problems can be found in, e.g., Venter [2010]. Among those, the Sequential Quadratic Programming (SQP) methods are probably the most preferable methods in NLP (see e.g., Schittkowski, 1985; Bogg and Tolle, 1996) and will be introduced in Section 4.7.2. In the following, the optimality conditions for constrained optimization problem will be given firstly, since many algorithms are based on them.

### 4.7.1 Karush–Kuhn–Tucker (KKT) optimality conditions

For a constrained optimization problem, the optimality conditions called the Karush–Kuhn–Tucker (KKT), also known as the Kuhn–Tucker conditions [Kuhn and Tucker, 1951], have to be fulfilled at the constrained optimum point. The KKT approach generalizes the Lagrangian approach that allows only equality constraints (see Eq. (4.24)). The Lagrangian function for the constrained optimization problem (4.64) is defined as

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{k}, \overline{\boldsymbol{k}}) = S(\boldsymbol{\beta}) + \boldsymbol{k}^T \boldsymbol{h}(\boldsymbol{\beta}) - \overline{\boldsymbol{k}}^T \boldsymbol{g}(\boldsymbol{\beta}) \tag{4.67}$$

where the $p \times 1$ vector $\boldsymbol{k}$ and the $\overline{p} \times 1$ vector $\overline{\boldsymbol{k}}$ consist of the Lagrange multipliers.

Suppose $\boldsymbol{\beta}^*$ is a local minimum, there exist vectors $\boldsymbol{k}^* = [k_1^*, \ldots, k_p^*]^T$ and $\overline{\boldsymbol{k}}^* = [\overline{k}_1^*, \ldots, \overline{k}_{\overline{p}}^*]^T$ such that the following conditions are satisfied (cf. Nocedal and Wright, 1999; Luenberger and Ye, 2008)

$$\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}^*, \boldsymbol{k}^*, \overline{\boldsymbol{k}}^*) = \nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}^*) + \boldsymbol{H}^T \boldsymbol{k}^* - \boldsymbol{G}^T \overline{\boldsymbol{k}}^* = \mathbf{0} \tag{4.68a}$$

$$\boldsymbol{h}(\boldsymbol{\beta}^*) = \mathbf{0} \tag{4.68b}$$

$$\boldsymbol{g}(\boldsymbol{\beta}^*) \geq \mathbf{0} \tag{4.68c}$$

$$g_i(\boldsymbol{\beta}^*)\, \overline{k}_i^* = 0, \quad i = 1, \ldots, \overline{p} \tag{4.68d}$$

$$\overline{k}_i^* \geq 0, \quad i = 1, \ldots, \overline{p} \tag{4.68e}$$

[1] Most modern interior-point methods are infeasible, i.e., they can start from any initial point. They also introduce slack variables to transform inequality constraints into equalities [Nocedal and Wright, 1999].

where $\boldsymbol{H}$ and $\boldsymbol{G}$ are the Jacobian matrices of the vector-valued constraint functions $\boldsymbol{h}(\boldsymbol{\beta})$ and $\boldsymbol{g}(\boldsymbol{\beta})$, i.e.,

$$\begin{aligned} \boldsymbol{H} &= [\nabla_{\boldsymbol{\beta}}\, h_1(\boldsymbol{\beta}^*), \ldots, \nabla_{\boldsymbol{\beta}}\, h_p(\boldsymbol{\beta}^*)]^T \\ \boldsymbol{G} &= [\nabla_{\boldsymbol{\beta}}\, g_1(\boldsymbol{\beta}^*), \ldots, \nabla_{\boldsymbol{\beta}}\, g_{\overline{p}}(\boldsymbol{\beta}^*)]^T . \end{aligned} \tag{4.69}$$

Equation (4.68) is known as the KKT conditions, which are the first-order (due to gradient) necessary conditions for a point to be a constrained local optimum. A point that satisfies these conditions is known as a KKT point. Equation (4.68a) indicates that the gradient of the Lagrangian must vanish at the optimum point. The Eqs. (4.68b) and (4.68c) suggest that all constraints are fulfilled, i.e., the optimum point is feasible with respect to all constraints. Equation (4.68d) indicates that if the $i$-th inequality constraint is inactive, i.e., $g_i(\boldsymbol{\beta}^*) > 0$, then the corresponding Lagrange multiplier holds $\overline{k}_j^* = 0$. Therefore, the inactive constraints can be taken out from Eq. (4.68a). Equation (4.68e) states that the Lagrange multipliers associated with inequality constraints must be nonnegative. Note, that there is no restriction on the sign of the Lagrange multipliers associated with equality constraints. If the active set, i.e., the set of constraints that holds as equality, is known, then the problem (4.64) can be transformed into an equality constrained problem which can easily be solved.

### 4.7.2 Sequential Quadratic Programming (SQP) method

The SQP method is an iterative method, where the update of the estimates $\boldsymbol{\beta}_{it}$ from the current iteration to the next iteration for the problem (4.64) is obtained by solving a QP subproblem (cf. Han, 1977; Powell, 1978; Bogg and Tolle, 1996; Nocedal and Wright, 1999)

$$\min_{\boldsymbol{p}\in\mathbb{R}^u} \quad \overline{S}(\boldsymbol{p}) = \left(\nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}_{it})\right)^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})\, \boldsymbol{p} \tag{4.70a}$$

$$\text{s.t.} \quad \boldsymbol{H}\,\boldsymbol{p} + \boldsymbol{h}(\boldsymbol{\beta}_{it}) = \boldsymbol{0} \tag{4.70b}$$

$$\boldsymbol{G}\,\boldsymbol{p} + \boldsymbol{g}(\boldsymbol{\beta}_{it}) \geq \boldsymbol{0} \tag{4.70c}$$

where $\overline{S}(\boldsymbol{p})$ denotes the objective function with the vector $\boldsymbol{p} = \boldsymbol{\beta} - \boldsymbol{\beta}_{it}$ of optimization variables, $\boldsymbol{\beta}_{it}$, $\boldsymbol{k}_{it}$ and $\overline{\boldsymbol{k}}_{it}$ are the estimates of parameter vector and vectors of multipliers in the current iteration, and $\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})$ is the Hessian of the Lagrangian function (4.67) with respect to $\boldsymbol{\beta}$, which is defined as

$$\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it}) = \begin{bmatrix} \frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})}{\partial \beta_1 \partial \beta_1} & \cdots & \frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})}{\partial \beta_1 \partial \beta_u} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 L(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})}{\partial \beta_u \partial \beta_1} & \cdots & \frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})}{\partial \beta_u \partial \beta_u} \end{bmatrix} . \tag{4.71}$$

The Eqs. (4.70b) and (4.70c) are obtained by linearizing the constraints (4.64b) and (4.64c), whereas the form of Eq. (4.70c) is related to the quadratic Taylor series approximation for the Lagrangian function (4.67)[1].

As can be seen from Eq. (4.70), the computation of the Hessian $\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})$ is required. Analytical Hessian might be unavailable or it is computationally expensive in each iteration,

[1] At the current iterate $(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})$, the quadratic (second-order) Taylor series expansion for the Lagrangian reads $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it}) \approx \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it}) + \left(\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})\right)^T (\boldsymbol{\beta} - \boldsymbol{\beta}_{it}) + \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_{it})^T \nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it}) (\boldsymbol{\beta} - \boldsymbol{\beta}_{it})$.

and therefore, it is usually replaced by an approximation $\boldsymbol{B}_{it}$ using, e.g., a finite difference approximation or a quasi-Newton method; further details can be found in, e.g., Powell [1978] or Nocedal and Wright [1999].

The QP subproblem (4.70) can then be simplified as

$$\min_{\boldsymbol{p}\in\mathbb{R}^u} \quad \overline{S}(\boldsymbol{p}) = \boldsymbol{c}^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{Q}\,\boldsymbol{p} \tag{4.72a}$$

$$\text{s.t.} \quad \boldsymbol{H}\,\boldsymbol{p} + \overline{\boldsymbol{h}} = \boldsymbol{0} \tag{4.72b}$$

$$\boldsymbol{G}\,\boldsymbol{p} + \overline{\boldsymbol{g}} \geq \boldsymbol{0} \tag{4.72c}$$

where the iteration index $it$ is dropped for a better readability and

$$\boldsymbol{c} = \nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}_{it}), \quad \boldsymbol{Q} = \nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it}) = \boldsymbol{B}_{it},$$
$$\overline{\boldsymbol{h}} = \boldsymbol{h}(\boldsymbol{\beta}_{it}), \quad \overline{\boldsymbol{g}} = \boldsymbol{g}(\boldsymbol{\beta}_{it}).$$

There are a variety of algorithms to solve Eq. (4.72) and the algorithms will be introduced in Section 4.7.2.1. The solution $\boldsymbol{p}_{it}$ of the QP subproblem (4.72) is then used to form a new iterate

$$\boldsymbol{\beta}_{it+1} = \boldsymbol{\beta}_{it} + \alpha_{it}\,\boldsymbol{p}_{it} \tag{4.73}$$

applying a line search strategy, in order to force convergence from poor starting points [Powell, 1978]. The positive scalar $\alpha_{it}$, called the step length, gives the size of the step taken from the current iterate to the next one. For a constrained problem, the step length has to be determined such that not only the objective function (4.64a) has a sufficient decrease but also the constraints (4.64b) and (4.64c) are satisfied. This is achieved by a line search to reduce a merit function; see Section 4.7.2.2 for details.

#### 4.7.2.1 Search direction – Active-set methods

Most of the algorithms that solve the inequality constrained optimization problem (4.72) can be subdivided into two categories: active-set methods and interior-point methods (e.g., Wong, 2011; Roese-Koerner, 2015). The active-set methods try to predict the true active set and solve a sequence of equality constrained subproblems iteratively, which follow the boundary of the feasible set, whereas the interior-point methods compute iterates which follow the interior of the feasible set. An advantage of the active-set methods is that they allow a "warm start", namely, any given active set is allowed to be specified to start the algorithm. This is very useful within the SQP methods for NLP problem, since each QP subproblem in the sequence is related to the previous QP subproblem (e.g., Maes, 2011; Wong, 2011).

Thus, we will focus on the active-set methods in particular the primal active-set methods here, which means that all iterates remain feasible with respect to the primal problem (4.72). Since

Taking a quadratic model of the Lagrangian as the objective function gives

$$\min_{\boldsymbol{p}\in\mathbb{R}^u} \quad \left(\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})\right)^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T\,\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})\,\boldsymbol{p}\,.$$

Equation (4.70a) is equivalent to the equation above for problems with only equality constraints, since the gradient of the Lagrangian $\left(\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}_{it}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})\right)^T \boldsymbol{p}$ contains an additional term $\boldsymbol{k}\,\boldsymbol{H}\,\boldsymbol{p}$ whereas $\boldsymbol{H}\,\boldsymbol{p}$ is constant according to Eq. (4.70b). If inequality constraints exist, the two forms are equivalent if the Lagrange multipliers corresponding to inactive linearized constraints are zero; details can be found in, e.g., Bogg and Tolle [1996].

the true active set (if any) is generally unknown, a "working set" (i.e., all the equality constraints Eq. (4.72b) and some of the inequality constraints in Eq. (4.72c) which are regarded as equalities), is used to predict the correct active set and is updated during the iterations. Therefore, the equality constraints (4.72b) should always be included in the working set, whereas the inequality constraints can be altered during the iterations. We use $\mathcal{W}^k$ to refer to those active constraints from (4.72c) in the working set in the $k$-th iteration. We use the superscript $k$ to denote the iteration step in the active-set methods (minor iterations), in order to avoid confusion with the subscript $it$ for the iterative computations (major iterations) in the SQP methods. We define the step $\boldsymbol{q}$ to the iterate $\boldsymbol{p}^k$ as

$$\boldsymbol{q} = \boldsymbol{p} - \boldsymbol{p}^k \,. \tag{4.74}$$

Substituting Eq. (4.74) into Eq. (4.72a), we obtain the objective function $\widetilde{S}(\boldsymbol{q})$ as

$$\begin{aligned} \widetilde{S}(\boldsymbol{q}) &= \boldsymbol{c}^T (\boldsymbol{p}^k + \boldsymbol{q}) + \frac{1}{2} (\boldsymbol{p}^k + \boldsymbol{q})^T \boldsymbol{Q} (\boldsymbol{p}^k + \boldsymbol{q}) \\ &= \overline{\boldsymbol{c}}^T \boldsymbol{q} + \frac{1}{2} \boldsymbol{q}^T \boldsymbol{Q} \boldsymbol{q} + \boldsymbol{l}^k \end{aligned} \tag{4.75}$$

with

$$\overline{\boldsymbol{c}} = \boldsymbol{c} + \boldsymbol{Q} \boldsymbol{p}^k, \quad \boldsymbol{l}^k = \boldsymbol{c}^T \boldsymbol{p}^k + \frac{1}{2} \boldsymbol{p}^k \boldsymbol{Q} \boldsymbol{p}^k$$

where $\boldsymbol{l}^k$ is independent of $\boldsymbol{q}$ and can be taken out from the objective function $\widetilde{S}(\boldsymbol{q})$. Therefore, solving the QP problem (4.72) is transformed into solving the equality constrained subproblem (cf. Nocedal and Wright, 1999; Roese-Koerner, 2015)

$$\min_{\boldsymbol{q} \in \mathbb{R}^u} \quad \overline{\boldsymbol{c}}^T \boldsymbol{q} + \frac{1}{2} \boldsymbol{q}^T \boldsymbol{Q} \boldsymbol{q} \tag{4.76a}$$

$$\text{s.t.} \quad \boldsymbol{H} (\boldsymbol{p}^k + \boldsymbol{q}) + \overline{\boldsymbol{h}} = \boldsymbol{0} \implies \overline{\boldsymbol{H}} \boldsymbol{q} = \boldsymbol{0} \tag{4.76b}$$

$$\widetilde{\boldsymbol{G}}^k (\boldsymbol{p}^k + \boldsymbol{q}) + \widetilde{\boldsymbol{g}} = \boldsymbol{0} \implies \widetilde{\boldsymbol{G}}^k \boldsymbol{q} = \boldsymbol{0} \tag{4.76c}$$

where the matrix $\widetilde{\boldsymbol{G}}^k$ is the part of $\boldsymbol{G}$ related to the active constraints in the working set $\mathcal{W}^k$. If we write $\boldsymbol{G}$ as $\boldsymbol{G} = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{\overline{p}}]^T$ with column vector $\boldsymbol{g}_i$ $(i = 1, \ldots, \overline{p})$, then we can formulate $\widetilde{\boldsymbol{G}}^k = [\ldots, \boldsymbol{g}_j, \ldots]^T$ with $j \in \mathcal{W}^k$.

The equality constrained problem (4.76) can be solved by using, e.g., the KKT equations of the Lagrangian of the problem (similar to the procedure introduced in Section 4.2.2).

Let $\boldsymbol{q}^k$ be the solution of Eq. (4.76), $\boldsymbol{p}^k$ will be updated through

$$\boldsymbol{p}^{k+1} = \boldsymbol{p}^k + \alpha^k \boldsymbol{q}^k \tag{4.77}$$

similar to Eq. (4.73), where $\alpha^k$ denotes the step length, which should be determined such that it does not violate any constraint that is not in the working set [Gill *et al.*, 1981], namely, the feasibility of the next iteration is maintained. The full step length, i.e., $\alpha^k = 1$ is taken if the solution $\boldsymbol{p}^k + \boldsymbol{q}^k$ is feasible with respect to all the constraints. Different cases are distinguished to determine $\alpha^k$.

**1)** Case 1: $\boldsymbol{q}^k \neq \boldsymbol{0}$
It can easily be obtained that

$$\begin{aligned} \boldsymbol{H} \boldsymbol{p}^{k+1} + \overline{\boldsymbol{h}} &= \boldsymbol{H} (\boldsymbol{p}^k + \alpha^k \boldsymbol{q}^k) + \overline{\boldsymbol{h}} = \boldsymbol{H} \boldsymbol{p}^k + \overline{\boldsymbol{h}} = \boldsymbol{0} \\ \widetilde{\boldsymbol{G}}^k \boldsymbol{p}^{k+1} + \widetilde{\boldsymbol{g}} &= \widetilde{\boldsymbol{G}}^k (\boldsymbol{p}^k + \alpha^k \boldsymbol{q}^k) + \widetilde{\boldsymbol{g}} = \widetilde{\boldsymbol{G}}^k \boldsymbol{p}^k + \widetilde{\boldsymbol{g}} = \boldsymbol{0} \end{aligned} \tag{4.78}$$

since the constraints in the working set are satisfied at $\boldsymbol{p}^k$. Therefore, the constraints in the working set will also be satisfied at $\boldsymbol{p}^{k+1}$ for any value of $\alpha^k$. In other words, the search moves along the boundary of feasible set formed from the working set.

Let $i$ denote the index of a constraint that is not in $\mathcal{W}^k$ (i.e., $i \notin \mathcal{W}^k$) and $\overline{g}_i$ be the $i$-th element of $\overline{\boldsymbol{g}}$ in Eq. (4.72c). When the relation $\boldsymbol{g}_i^T \boldsymbol{q}^k \geq 0$ holds, the inactive constraint, i.e., $\boldsymbol{g}_i^T \boldsymbol{p}^k + \overline{g}_i > 0$, will not be violated through a positive step in the direction $\boldsymbol{q}^k$ because

$$\begin{aligned} \boldsymbol{g}_i^T \boldsymbol{p}^{k+1} + \overline{g}_i &= \boldsymbol{g}_i^T \left( \boldsymbol{p}^k + \alpha^k \boldsymbol{q}^k \right) + \overline{g}_i \\ &= \underbrace{\boldsymbol{g}_i^T \boldsymbol{p}^k + \overline{g}_i}_{>0} + \underbrace{\alpha^k}_{>0} \underbrace{\boldsymbol{g}_i^T \boldsymbol{q}^k}_{\geq 0} > 0 \,. \end{aligned} \tag{4.79}$$

Therefore, the inactive constraint $i$ ($i \notin \mathcal{W}^k$) will remain inactive in the next iteration step $k+1$. If $\boldsymbol{g}_i^T \boldsymbol{q}^k \geq 0$ holds for every $i : i \notin \mathcal{W}^k$, the inequality constraints that are not in the working set will not impose any restriction on the step length. Then, the unit step length $\alpha^k = 1$ along $\boldsymbol{q}^k$ is the solution of Eq. (4.76), which achieves the best decrease in the objective function.

However, if there exist some $i$ ($i \notin \mathcal{W}^k$) satisfy $\boldsymbol{g}_i^T \boldsymbol{q}^k < 0$, the corresponding constraints can become active or violated at the next iterate. The constraints will not be violated only if $\boldsymbol{g}_i^T \left( \boldsymbol{p}^k + \alpha^k \boldsymbol{q}^k \right) + \overline{g}_i \geq 0$, i.e.,

$$\alpha^k \leq \frac{-\overline{g}_i - \boldsymbol{g}_i^T \boldsymbol{p}^k}{\boldsymbol{g}_i^T \boldsymbol{q}^k} \,. \tag{4.80}$$

The optimal step length $\alpha_{opt}^k$, which is the distance to the nearest constraint, is given by

$$\alpha_{opt}^k = \min \quad \left( \frac{-g_i - \boldsymbol{g}_i^T \boldsymbol{p}^k}{\boldsymbol{g}_i^T \boldsymbol{q}^k} \right), \quad \forall i : i \notin \mathcal{W}^k \quad \text{and} \quad \boldsymbol{g}_i^T \boldsymbol{q}^k < 0 \,. \tag{4.81}$$

When $\alpha_{opt}^k$ is taken as the step length $\alpha^k$ (i.e., $\alpha^k = \alpha_{opt}^k$), the nearest inactive constraint that prevents a full step holds as equality. This constraint is then added in the working set $\mathcal{W}^{k+1}$ for the next iteration and a new equality constrained optimization problem has to be solved.

When $\alpha^k = 1$, $\boldsymbol{p}^{k+1} = \boldsymbol{p}^k + \boldsymbol{q}^k$ is the solution that minimizes the objective function in Eq. (4.76). Therefore, it is not possible to further decrease the objective function without removing a constraint from the working set. Then, the KKT condition introduced previously in Eq. (4.68e) has to be checked, i.e., the Lagrange multipliers corresponding to the inequality constraints in the working set must be nonnegative at the optimum point. Actually a negative Lagrange multiplier associated to an inequality constraint implies that the objective function can be minimized by removing that constraint (see e.g., Roese-Koerner, 2015 for more details). Therefore, if some of the multipliers related to the inequality constraints in the working set are negative, one of these constraints are removed from the working set and a new problem (4.76) in a new iteration is solved. Note, that equality constraints will never be removed. On the other hand, if all Lagrange multipliers associated to inequality constraints are nonnegative, it is not possible to further minimize the objective function.

**2) Case 2**: $\boldsymbol{q}^k = \boldsymbol{0}$

If the Lagrangian multipliers $\overline{k}_j \geq 0$ for all $j \in \mathcal{W}^k$, $\boldsymbol{p}^{k+1} = \boldsymbol{p}^k$ is the KKT point and the iteration terminates; otherwise, if there exists $\overline{k}_j < 0$ for $j \in \mathcal{W}^k$, one of the constraints corresponding to the negative multipliers is dropped from the current working set and a new problem (4.76) is solved.

#### 4.7.2.2 Step length – Line search and merit function

As introduced previously, the solution of the above QP subproblem (4.72) is the search direction $\boldsymbol{p}_{it}$ in Eq. (4.73). The step length parameter $\alpha_{it}$ has to be determined, such that not only the objective function $S(\boldsymbol{\beta})$ has a sufficient decrease at subsequent iterations but also all the constraints are satisfied. As these two goals often conflict, a merit or penalty function that weights their relative importance can be used as a criterion to decide whether one point is better than another [Moré and Wright, 1993]. The merit function of the form [Powell, 1978, and references therein]

$$\Psi(\boldsymbol{\beta}, \boldsymbol{\nu}) = S(\boldsymbol{\beta}) + \sum_{i=1}^{i=p} \nu_i \cdot |h_i(\boldsymbol{\beta})| + \sum_{i=p+1}^{i=p+\overline{p}} \nu_i \cdot \left| \min\left(0, g_i(\boldsymbol{\beta})\right) \right| \tag{4.82}$$

can be used demanding that

$$\Psi(\boldsymbol{\beta}_{it} + \alpha_{it}\, \boldsymbol{p}_{it}, \boldsymbol{\nu}) < \Psi(\boldsymbol{\beta}_{it}, \boldsymbol{\nu}) \tag{4.83}$$

where $\nu_i > 0$ $(i = 1, \ldots, \overline{p})$ are penalty parameters. Equation (4.82) is called the $\ell_1$ penalty function; other merit functions can be found in, e.g., Moré and Wright [1993] or Nocedal and Wright [1999]. The choice of the penalty parameters can be based on the Lagrange multipliers at the solution of the QP subproblem (4.72) that defines $\boldsymbol{p}_{it}$; further details can be found in, e.g., Powell [1978], Coleman *et al.* [1999] and Matlab [2008].

# Chapter 5

# Electron Density Modeling

There are various approaches available for modeling the electron density of the ionosphere. Generally, these approaches can be categorized into the voxel-based approach and the function-based approach for ionosphere tomography[1] [Liu *et al.*, 2006]. The voxel-based one assumes that the ionosphere can be subdivided into cells with constant electron density (see e.g., Juan *et al.*, 1997; Kuklinski, 1997; Hernández-Pajares *et al.*, 1999). The function-based one employs a set of basis functions modeling the spatiotemporal variations of the ionosphere, respectively. For examples, Liu and Gao [2004] modeled the electron density correction term that is defined to be relative to an a priori electron density reference, horizontally by a series of harmonics functions and vertically by EOF; Schmidt *et al.* [2008] represented the electron density correction term by a 2-D B-spline (longitude and latitude)/1-D height-dependent EOF tensor product with initially unknown time-dependent coefficients; an approach of 3-D B-spline expansions with respect to longitude, latitude and height with initially unknown time-dependent scaling coefficients was also applied (e.g., Schmidt, 2007; Zeilhofer *et al.*, 2009); Schmidt *et al.* [2008] proposed a 4-D B-spline modeling approach, where the time-dependency is also represented by a B-spline function. In addition, the height-dependency of the electron density is also modeled by so-called ionosphere "profilers" (in the sense that they use "anchor points", i.e., layer peaks and provide the vertical EDP between the anchor points, see Leitinger *et al.* [2001]), such as the Chapman (both $\alpha$ and $\beta$ type), exponential, Epstein (also called sech-squared) and parabolic functions (see details later in Section 5.1.1). For example, Sharifi and Farzaneh [2016] modeled the electron density vertically by the Chapman function and horizontally by a Slepian function, whereas the B-spline was used to account for temporal variations.

Generally, compared with the function-based approach, the voxel-based one is more flexible, as the spatial and temporal resolutions are not fixed within the region under investigation [Feltens *et al.*, 2011]. However, the function-based approach allows the computation of the electron density everywhere within the region under investigation, whereas the voxel-based approach requires additionally interpolation [Liu *et al.*, 2006].

Within this thesis, the function-based approach with vertical representation in terms of profilers will be exploited, since these profilers provide directly the key ionosphere parameters with physical meanings, such as peak densities and peak heights. The vertical profile functions have been used to model a certain ionosphere layer, part of the ionosphere, or even the whole ionosphere.

[1] Ionospheric tomography refers to the reconstruction of ionosphere parameters (e.g., the electron density) from measured signals of the line integrated parameters in various directions [Austen *et al.*, 1988].

To be specific, the modeling approach using such profilers can be generally divided into three groups: (1) a one-layer model (e.g., Feltens, 1998), which have limited usage especially during the day when several layers exist; (2) a two-layer model, i.e., a combination of the bottomside and topside layers (e.g., Fox, 1994; Ezquer *et al.*, 1996; Alizadeh, 2013); (3) a composite model considering the various ionosphere layers, i.e., a combination of several profilers (e.g., Rawer *et al.*, 1978; Radicella and Leitinger, 2001; Feltens, 2007; Tsai *et al.*, 2009; Brunini *et al.*, 2013b).

## 5.1 General modeling of the vertical electron density distribution

Following the modeling concept introduced in Sections 2.5.2 and 2.5.6, the vertical distribution of the electron density $N_e(h)$ can be modeled as a superposition of different layers, written as

$$N_e(h) = N_D(h) + N_E(h) + N_{F1}(h) + N_{F2}(h)\,. \tag{5.1}$$

$N_i(h)$ with $i \in \{D, E, F1, F2\}$ denotes the electron density of the various layers, each of which can be modeled by means of a certain profile function.

### 5.1.1 Profile functions for the ionosphere layers

Mathematical representations of the most often used profile functions are listed below. All these functions are characterized by the peak density $Nm_i$, the peak height $hm_i$, and a scale height $H_i$ describing the shape (thickness) of the ionospheric EDP (see Section 5.2 for more details). All functions express the altitude in terms of the dimensionless parameters, the so-called dimensionless reduced height $z_i = (h - hm_i)/H_i$ (cf. Eq. (2.35)).

1. Epstein (sech-squared) function (e.g., Booker, 1977; Rawer, 1983; Giovanni and Radicella, 1990)

$$N_i(h) = Nm_i\,\mathrm{sech}^2(z_i/2)^{1} = Nm_i\,\frac{4\exp(z_i)}{\left[1+\exp(z_i)\right]^2}\,. \tag{5.2}$$

2. Exponential function

$$N_i(h) = Nm_i\,\exp\left(-z_i\right)\,. \tag{5.3}$$

   Note that the exponential function cannot represent the full range of a certain layer $i$ due to its monotonicity (below $hm_i$, the electron densities will exceed $Nm_i$), and therefore, it can be partly used e.g., for the topside ionosphere, i.e., when $h > hmF2$.

3. Chapman function

$$N_i(h) = Nm_i\,\exp\left\{\bar{c}\left[1 - z_i - \exp\left(-z_i\right)\right]\right\}. \tag{5.4}$$

   As introduced previously in Section 2.1.2, a Chapman-layer shape is predicted by a simplified aeronomic theory, with the assumption that photoionization in a one-species neutral gas, neglecting transport processes [Rishbeth and Garriott, 1969]. This function is depending on a factor $\bar{c}$, called $\alpha$-Chapman for $\bar{c} = 0.5$ and $\beta$-Chapman for $\bar{c} = 1$. The factor $\bar{c}$ depends on the assumptions regarding the principles of chemical recombination [Hargreaves, 1992].

[1] The hyperbolic secant function is defined as $\mathrm{sech}(x) = \frac{1}{\cosh(x)} = \frac{2}{e^x + e^{-x}}$

4. Parabolic function

$$N_i(h) = Nm_i \left[1 - (z_i/2)^2\right]. \tag{5.5}$$

According to Stankov *et al.* [2003], this function is appropriate for modeling the profile near $hmF2$.

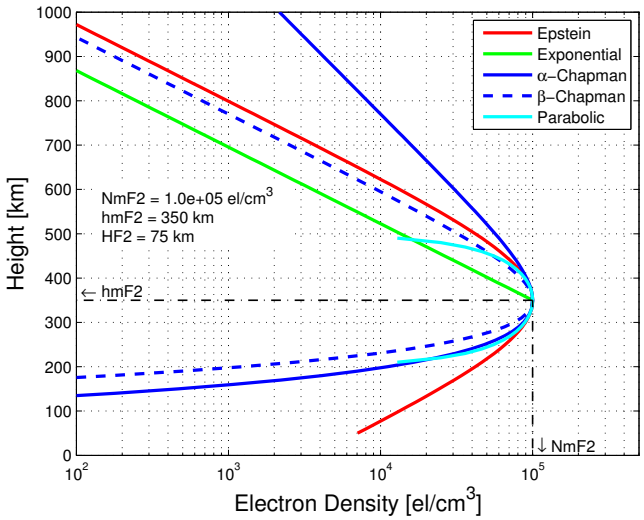



Figure 5.1: Comparisons of the shapes of different profilers.

A comparison of the shapes of these functions is shown in Fig. 5.1 exemplarily modeled for the F2 layer. To produce these curves, the same values of $NmF2$, $hmF2$ and $HF2$ are used. Note that, the exponential function below $hmF2$ is not shown, as the values exceed $NmF2$. The parabolic function is also shown partly, since it has nonpositive values when $h \leq hmF2 - 2 \cdot HF2$ or $h \geq hmF2 + 2 \cdot HF2$. As can be seen, the shapes of the resulting vertical distributions of the electron density are quite different. Specifically, the $\alpha$-Chapman function is the steepest in the topside, while the Epstein function is the steepest in the bottomside.

It is not an easy task to find out which profile function is the most suitable mathematical representation. Some work has compared the performance of the different functions, in particular for the topside ionosphere. For example, Fonda *et al.* [2005] compared the Chapman and Epstein functions using topside EDPs obtained from measurements of IK-19 and ISIS-2 satellites, and showed the best performance of the $\alpha$-Chapman function. A two-layer topside formulation (i.e., a Chapman function with $\alpha < 0.5$ below $hmF2 + 400$ and a different function for the higher part up to 1000 km above $hmF2$) was proposed in order to better reproduce the experimental shape of the topside. More recently, Verhulst and Stankov [2014] evaluated profilers including the exponential, Chapman and Epstein functions by fitting them to EDPs obtained from ionograms recorded by the topside sounders onboard the Alouette and ISIS satellites, and found that the exponential profiler has the best fit in most cases, followed by the Chapman profiler.

These functions with a constant scale height for the whole topside F2 layer are deduced for a single-constituent ($O^+$) [Nsumei *et al.*, 2010], which is a simplified representation. However, $H^+$ is also the major ion species in the topside ionosphere by assuming that $He^+$ has little influence on the EDPs under most conditions [Sibanda and Mckinnell, 2011, and references therein]. Consequently, a combination of two profile functions with two scale heights individually for $O^+$ and $H^+$ were applied by e.g., Jakowski [2005] (see Section 5.1.2 for details) and Sibanda and Mckinnell [2011]. Some authors applied representations with a height-variable scale height,

such as the well-known Vary-Chap (e.g., Huang and Reinisch, 2001; Reinisch and Huang, 2001) and Semi-Epstein (e.g., Nava *et al.*, 2008) functions. The Vary-Chap function, i.e., a general $\alpha$-Chapman function with a height-dependent scale height $H(h)$, is defined as

$$N_e(h) = NmF2 \cdot \left(\frac{H_m}{H(h)}\right)^{\frac{1}{2}} \exp\left\{\frac{1}{2}\left[1 - y - \exp(-y)\right]\right\} \tag{5.6}$$

$$\text{with} \quad y = \int_{hmF2}^{h} \frac{dh}{H(h)} ,$$

where $H_m$ denotes the scale height at $hmF2$. When $H(h)$ is assumed to be constant along the height and with $y = z_i$, Eq. (5.6) becomes the $\alpha$-Chapman function (5.4). The solution of $H(h)$ can be found in e.g., Huang and Reinisch [2001].

The Semi-Epstein function, i.e., an Epstein layer function (5.2) with a height-dependent scale height $H(h)$, is used for the topside representation in the NeQuick model. The function $H(h)$ is then defined as

$$H(h) = H_0\left[1 + \frac{r \cdot g(h - h_m)}{r \cdot H_0 + g(h - h_m)}\right]$$

with the constant parameters $r = 100$ and $g = 0.125$. The parameter $H_0$ is a quantity related to the scale height of the bottomside F2-layer; for further details see Nava *et al.* [2008].

### 5.1.2 Plasmasphere extension

As introduced in Section 2.6, the plasmasphere starts from the height level where the domination of ion composition changes from $O^+$ to $H^+$. There is actually no clear upper boundary between ionosphere and plasmasphere. As mentioned above, the electron density $N_e^{top}(h)$ in the topside ionosphere can be regarded as a sum of the ion densities of $O^+$ and $H^+$ by assuming that $He^+$ is negligible (e.g., Jakowski, 2005; Stankov *et al.*, 2011; González-Casado *et al.*, 2013). The corresponding ion composition is shown schematically in Fig. 5.2. Therefore, we can write

$$N_e^{top}(h) = N_{O^+}(h) + N_{H^+}(h) \qquad (h > hmF2) . \tag{5.7}$$

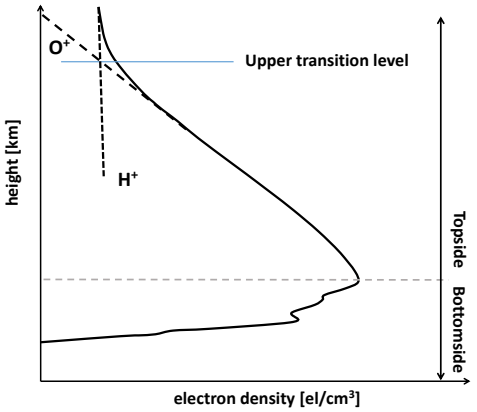



Figure 5.2: A schema of $N_e$ profile and ion composition in the upper ionosphere and plasmasphere, based on Stankov *et al.* [2003].

As a consequence, Eq. (5.1) can be extended by introducing a plasmaspheric term $N_{H^+}(h > hmF2)$, i.e.

$$N_e(h) = N_D(h) + N_E(h) + N_{F_1}(h) + N_{F2}(h) + N_{H^+}(h > hmF2) \tag{5.8}$$

For the plasmaspheric term $N_{H^+}(h)$, the exponential function (2.63) is often used (see e.g., Jakowski, 2005; Feltens, 2007). The so-called Topside Ionosphere and Plasmasphere model, developed by Jakowski [2005] (see also Stankov and Jakowski, 2006), applied a combination of the $\alpha$-Chapman function and the exponential function, reads

$$N_e(h) = NmF2 \exp\left[0.5(1 - z - \exp(-z))\right] + N_{P0} \exp(-h/H_p) \qquad (h > hmF2) \tag{5.9}$$

with $z = (h - hmF2)/H_{TS}$. $H_{TS}$ is the scale height in the topside ionosphere, $N_{P0}$ is the plasmaspheric basis density, and $H_p$ is the scale height of the plasmasphere.

## 5.2 Ionospheric key parameters and modeling approaches

Ionospheric key parameters, e.g., the peak parameters and scale heights are essential for using the above profile functions. Since the F2 layer peak characteristics play a crucial role in ionosphere dynamics as well as in radio communication and positioning applications, more details about $NmF2$, $hmF2$ and $HF2$ will be given in the following.

**NmF2**

$NmF2$ can be obtained through its linear relation with the square of the critical frequency $foF2$ of the F2 layer (see Section 2.2) derived from ionograms recorded by ionosondes[1]

$$\underbrace{NmF2}_{[\mathrm{m}^{-3}]} = \underbrace{1.24 \cdot 10^{10}}_{[s^2/m^3]} \cdot \underbrace{(foF2)^2}_{[MHz^2]} . \tag{5.10}$$

The relation also holds for other ionosphere layers.

**hmF2**

$hmF2$ can be obtained through its relation with the propagation factor $M(3000)F2$[2] derived from ionograms. Schimazak [1995] proposed an empirical relationship between $hmF2$ and $M(3000)F2$, given as

$$\underbrace{hmF2}_{[km]} = \frac{\overbrace{1490}^{[km]}}{\underbrace{M(3000)F2}_{[-]}} - \underbrace{176}_{[km]} . \tag{5.11}$$

[1] When the wave cannot propagate any further and is reflected, the refractive index of the ionosphere becomes zero. It holds that $f^2 = \frac{N_e\, e^2}{4\,\pi^2 \varepsilon_0\, m_e}$ where $f$ is the wave frequency; the natural constants $e$, $\varepsilon_0$ and $m_e$ are electron charge, permittivity in vacuum and electron mass, respectively. Equation (5.10) is thus derived; see, e.g., Basu *et al.* [1985] or Hargreaves [1992] for details.

[2] $M(3000)F2 = \frac{MUF(3000)}{foF2}$ where $MUF(3000)$ is the highest frequency at which a radio wave can be received over a distance of 3000 km after reflection in the ionosphere [Bradley and Dudeney, 1973].

However, many authors have found unsatisfactory results and further modified it. For example, Bilitza *et al.* [1979] introduced a correction factor $DM$ into Eq. (5.11) based on incoherent scatter radar measurements, and obtained

$$\underbrace{hmF2}_{[km]} = \frac{\overbrace{1490}^{[km]}}{\underbrace{M(3000)F2}_{[\text{-}]} + \underbrace{DM}_{[\text{-}]}} - \underbrace{176}_{[km]} \tag{5.12}$$

where the correction factor $DM$ is dependent on the ratio $\frac{foF2}{foE}$, the solar activity in terms of the sunspot number and the geomagnetic latitude; $foE$ is the critical frequency of the E layer.

**HF2**

The scale height is a key ionosphere characteristic that reflects the shape (thickness) of the ionosphere EDP and connects the ionosphere dynamics, plasma thermal structure and compositions [Liu *et al.*, 2007a]. Many works investigate the diurnal and seasonal variations as well as solar activity dependences of the scale height (e.g., Lei *et al.*, 2005; Belehaki *et al.*, 2006; Liu *et al.*, 2006; Stankov and Jakowski, 2006; Zhang *et al.*, 2006; Liu *et al.*, 2007a,b; Ram *et al.*, 2009). It is worth noting that there exist various definitions of the scale height in the literature (see Belehaki *et al.*, 2006; Liu *et al.*, 2007a; Ram *et al.*, 2009). For example, the plasma scale height in Eq. (2.64) is based on theoretical considerations. The effective scale height or sometimes referred to the Chapman scale height, is defined as the scale height by fitting the Chapman function to the EDPs.

Various representations have been used to model the horizontal and temporal variations of these key parameters, such as spherical harmonics (Tsai *et al.*, 2009; Alizadeh, 2013; Brunini *et al.*, 2013b), EOF (Zhang *et al.*, 2011, 2009; Yu *et al.*, 2015), B-splines (Limberger *et al.*, 2013; Liang *et al.*, 2015b), and polynomials (Hoque and Jakowski, 2011; Jakowski *et al.*, 2011). In the following, we will focus on three frequently used methods, i.e., spherical harmonics, B-spline representation and EOF. Table 5.1 gives an overview of the essential characteristics of the three methods.

### 5.2.1 Spherical harmonics

An ionosphere key parameter $\kappa \in \{NmE, hmE, HE, NmF2, hmF2, HF2, \ldots\}$ can be represented spatially as a series expansion [Schmidt *et al.*, 2011]

$$\kappa(\lambda, \varphi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_{n,m}\, Y_{n,m}(\lambda, \varphi) \tag{5.13}$$

in terms of spherical harmonics $Y_{n,m}(\lambda, \varphi)$ which are calculable from

$$Y_{n,m}(\lambda, \varphi) = \begin{cases} P_{n,m}(\sin(\varphi)) \cos(m \cdot \lambda) & \text{for } m = 0, \ldots, n \\ P_{n,|m|}(\sin(\varphi)) \sin(|m| \cdot \lambda) & \text{for } m = -n, \ldots, -1 \end{cases} \tag{5.14}$$

Table 5.1: The characteristics of the spherical harmonics, the B-spline representation and EOF, cf. Jekeli [2005].

| | Spherical harmonics | B-splines | EOF |
|---|---|---|---|
| Defined domain | surface of a sphere | Euclidean space | multi-dimensional vector space |
| Formula | mathematically given | mathematically given | empirically defined by data |
| Orthogonality | yes | no | yes |
| Nature of support | global support[a] | local support[b] | depending on the area of investigation |
| Update capability (e.g., model update with new data) | all coefficients need to be recomputed | only coefficients in the relevant region need to be recomputed | EOF modes have to be recomputed |
| Generating wavelet functions | no | yes | no |

[a] Each spherical harmonic function is significantly different from zero almost everywhere on the entire sphere [Jekeli, 2005].

[b] The basis function is zero everywhere except in a local region.

with

| | |
|---|---|
| $\lambda$ | longitude, |
| $\varphi$ | latitude, |
| $P_{n,m}(\sin\varphi)$ | associated Legendre functions of degree $n$ and order $m$, |
| $d_{n,m}$ | unknown spherical harmonic coefficients. |

The expansion (5.13) has to be truncated at a finite degree $n = N$, and the total number of unknown coefficients yields $(N+1)^2$, see e.g., Schmidt *et al.* [2008]. A maximum harmonic degree $N = 15$ within a Sun-fixed coordinate system is set up in e.g., IGS daily GIM [Schaer, 1999]. Generally, spherical harmonics are used to model a function on a sphere, and the temporal variations can be considered by introducing time-dependent spherical harmonic coefficients $d_{n,m}(t)$.

### 5.2.2 Empirical Orthogonal Function (EOF)

Empirical Orthogonal Function (EOF) analysis, also known as Principle Component Analysis (PCA), has been widely used in multivariate data analysis and is suitable for identifying dominant spatial structures and their temporal evolution (e.g., Divinskikh, 1988; Schmidt *et al.*, 2008, and references therein). Suppose that the spatial coordinates $(\lambda, \varphi)$ are discretized with $\lambda_i$, $(i = 1, \ldots, p_1)$ and $\varphi_j$ $(j = 1, \ldots, p_2)$, and the time coordinate is discretized to $t_n$ $(n = 1, \ldots, m)$, $\boldsymbol{K} = \big(\kappa(\lambda_i, \varphi_j, t_n)\big)$ is a $m \times p$ matrix with $p = p_1 \cdot p_2$, i.e., the two spatial dimensions $\lambda$ and $\varphi$ are concatenated together. Take the time average of each spatial grid point, i.e., take the mean value of each column of $\boldsymbol{K}$, we can obtain the residual map $\boldsymbol{K'} = \big(\kappa(\lambda_i, \varphi_j, t_n) - \overline{\kappa}(\lambda_i, \varphi_j)\big)$ with

$\overline{\kappa}(\lambda_i, \varphi_j) = \frac{1}{m} \sum_{n=1}^{m} \kappa(\lambda_i, \varphi_j, t_n)$. The data set $\boldsymbol{K'}$ can be generally decomposed into a set of basis functions with associated coefficients, reads (cf. Zhang *et al.*, 2009)

$$\boldsymbol{K'} = \sum_{k=1}^{N} \boldsymbol{K'}_k(\lambda, \varphi, t) = \sum_{k=1}^{N} \boldsymbol{a}_k \cdot \boldsymbol{e}_k^T \tag{5.15}$$

where $\boldsymbol{e}_k = [e_k^1, \ldots, e_k^p]^T$ is the $k$-th EOF pattern representing the spatial structure of the ionosphere parameter $\kappa$, and $\boldsymbol{a}_k = [a_k^1, \ldots, a_k^m]^T$ is the associated vector of coefficients representing the temporal behavior. $N$ is the truncation order of the EOF modes. The basis "functions"[1] $\boldsymbol{e}_k$ $(k = 1, \ldots, N)$ are orthogonal. The key parameter $\kappa(\lambda_i, \varphi_j, t_n)$ at a certain position and time can then be written as

$$\kappa(\lambda_i, \varphi_j, t_n) = \overline{\kappa}(\lambda_i, \varphi_j) + \sum_{k=1}^{N} a_k(t_n) \cdot e_k(\lambda_i, \varphi_j) \tag{5.16}$$

where $a_k(t_n)$ and $e_k^T(\lambda_i, \varphi_j)$ are the elements of $\boldsymbol{a}_k$ and $\boldsymbol{e}_k$ related to the time $t_n$ and the spatial position $(\lambda_i, \varphi_j)$, respectively. The orthogonal functions $\boldsymbol{e}_k$ are arranged in a descending order in terms of the variances, which means, the first mode captures the most energy of the original data set, and the second one captures the second most energy and so on [Zhang *et al.*, 2011, and references therein]. Usually the first few modes of EOFs capture most of the energy of the data, and therefore the number of coefficients is drastically reduced. To use the EOF method, a background model of the key parameter $\kappa$ is required. A disadvantage of the EOF method is that the shape of the EOFs is strongly influenced by the background model and may not be changed within the adjustment [Schmidt *et al.*, 2008].

### 5.2.3 B-splines

The main advantage of B-spline functions is that they are compactly supported, namely, they are non-zero only in a finite interval and zero elsewhere. The key parameter $\kappa$ can be represented by a series expansion [Schmidt, 2007]

$$\kappa(\lambda, \varphi, t) = \sum_{k_1=0}^{K_1-1} \sum_{k_2=0}^{K_2-1} \sum_{k_3=0}^{K_3-1} d_{k_1,k_2,k_3}^{J_1,J_2,J_3} \, \phi_{k_1,k_2,k_3}^{J_1,J_2,J_3}(\lambda, \varphi, t) \tag{5.17}$$

in terms of 3-D basis function

$$\phi_{k_1,k_2,k_3}^{J_1,J_2,J_3}(\lambda, \varphi, t) = \phi_{J_1;k_1}(\lambda) \, \phi_{J_2;k_2}(\varphi) \, \phi_{J_3;k_3}(t) \tag{5.18}$$

as the tensor product of three 1-D basis functions $\phi_{J;k}(x)$ depending on $x \in \{\lambda, \varphi, t\}$ with unknown series coefficients $d_{k_1,k_2,k_3}^{J_1,J_2,J_3}$. $\phi_{J;k}(x)$ is of resolution level $J \in \mathbb{N}_0$ $(J \in \{J_1, J_2, J_3\})$ and shift $k \in \mathbb{N}_0$ $(k \in \{k_1, k_2, k_3\})$.

The normalized quadratic B-splines $N_{J;k}^m(x)$ with $m = 2$ can be chosen as $\phi_{J;k}(x)$, i.e.,

$$\phi_{J;k}(x) = N_{J;k}^2(x) \, . \tag{5.19}$$

[1] Note that $\boldsymbol{e}_k$ are called functions here, in fact they are vectors.

Let $K_J$ be a positive integer value, then $\phi_{J;k}(x)$ is defined recursively by a sequence of non-decreasing values $t_0^J, t_1^J, \ldots, t_{K_J+2}^J$, called knots, as

$$N_{j;k}^m(x) = \frac{x - t_k^J}{t_{k+m}^J - t_k^J} N_{J;k}^{m-1}(x) + \frac{t_{k+m+1}^J - x}{t_{k+m+1}^J - t_{k+1}^J} N_{J;k+1}^{m-1}(x) \tag{5.20}$$

with initial values

$$N_{J;k}^0(x) = \begin{cases} 1 & \text{if } t_k^J \leqslant x < t_{k+1}^J \text{ and } t_k^J < t_{k+1}^J \\ 0 & \text{otherwise} \end{cases}$$

where $m = 1, 2$ and $k = 0, \ldots, K_J - 1$; see e.g., Stollnitz *et al.* [1995b]. Note that the fractions are set to be 0 when their denominators are 0.

Within this thesis we will deal with a regional problem, the endpoint-interpolating normalized quadratic B-splines, defined on the unit interval $\mathbb{I} = [0, 1]$, are thus introduced (see e.g., Stollnitz *et al.*, 1995b; Lyche and Schumaker, 2001; Schmidt, 2007). For that purpose, we restrict ourselves at first to the unit interval and set the first three knots to the value zero and the last three knots to the value one, and define equally spaced B-splines, i.e.,

$$0 = t_0^J = t_1^J = t_2^J < t_3^J < \ldots < t_{K_J-1}^J < t_{K_J}^J = t_{K_J+1}^J = t_{K_J+2}^J = 1 \tag{5.21}$$

with

$$t_{k+1}^J - t_k^J = 2^{-J} \quad \text{for} \quad k = 2, \ldots, K_J - 1 \tag{5.22}$$

and

$$K_J = 2^J + 2\,. \tag{5.23}$$

Assume a region defined as $\Omega^3 = [\lambda_{\min}, \lambda_{\max}] \times [\varphi_{\min}, \varphi_{\max}] \times [t_{\min}, t_{\max}]$ is handled, transformation equations have to be considered to normalize $\Omega^3$ to the unit cube $\mathbb{I}^3 = [0, 1] \times [0, 1] \times [0, 1]$ through

$$x' = (x - x_{\min})/(x_{\max} - x_{\min}) \tag{5.24}$$

with $x \in [\lambda, \varphi, t]$. Then Eq. (5.19) under consideration of the Eqs. (5.20) and (5.21) has to be inserted into Eq. (5.17) for all the variables $\lambda$, $\varphi$ and $t$.

Figure 5.3 shows the family of 1-D endpoint-interpolating normalized quadratic B-spline functions $N_{j;k}^2(x)$ exemplarily for the level $J = 3$. As can be seen, the first two functions on the left and the last two on the right are modified due to the endpoint-interpolating procedure, whereas the other 6 inner basis functions are not affected. According to Eq. (5.20), it can be derived that $N_{j;k}^2(x)$ is non-zero on the interval $[t_k^J, t_{k+3}^J)$. Therefore, each B-spline has an influence zone, i.e. a non-zero zone $\mathbb{I}_k^J = [t_k^J, t_{k+3}^J) \subset \mathbb{I}$. The length of the subintervals $\mathbb{I}_k^J$ is about $3 \cdot d_J$ where

$$d_J = \frac{1}{K_J - 1} = \frac{1}{2^J + 1} \tag{5.25}$$

represents the approximate distance between two adjacent B-spline functions. Take B-splines of $J = 3$ shown in the figure as example, the length of the non-zero zone is about $3 \cdot d_3 = 3 \times 1/9 = 0.33$. Furthermore, on any interval $[t_k^J, t_{k+1}^J)$, at most three B-spline functions are non-zero and the sum of them is equal to one (e.g., de Boor, 1993). The higher the level $J$ is chosen the larger is the number $K_J$ and the more narrow or sharp is each basis function, and thus finer structure can be modeled. However, the selection of the level $J$ should depend on the sampling

interval of the input data. Assume that the average sampling interval with respect to $\lambda$ within an observation interval of length $\Lambda = \lambda_{\max} - \lambda_{\min}$ is denoted as $\Delta_\lambda$, the condition [Schmidt *et al.*, 2011]

$$\Delta_\lambda < d_{J_1} \tag{5.26}$$

has to be fulfilled. Therefore, the maximum level $J_1$ can be determined under consideration of Eq. (5.25) by

$$J_1 \leq \log_2 \left( \frac{\Lambda}{\Delta_\lambda} - 1 \right) \tag{5.27}$$

cf. Schmidt *et al.* [2011]. The level values $J_2$ and $J_3$ for latitude and time can be chosen in the same manner.

It is worth to mention that a global application can be obtained by replacing the normalized quadratic B-splines with the trigonometric B-splines for the longitudinal dependence. The trigonometric B-splines are characterized by its "wrapping around" effect, see e.g., Schumaker and Traas [1991] and Schmidt *et al.* [2011] for details.

#### 5.2.3.1 Multi-Scale Representation (MSR)

An important feature of the B-spline basis functions is that they can be used to construct B-spline wavelets (Stollnitz *et al.*, 1995a,b). Wavelets are mathematical functions for representing and analyzing multi-scale structures, i.e., to perform a MSR, also known as multi-resolution analysis or Multi-Resolution Representation (MRR). Since both the ionosphere structures and the observation distributions are heterogeneous over the globe, the MSR is rather useful in the field of ionosphere modeling. Such applications towards ionosphere modeling can be found in, e.g., Schmidt [2007] and Zeilhofer [2008] for $VTEC$ signals, and Liang *et al.* [2015a] for the $NmF2$ signal. The construction of wavelets and the basic theory of a MSR will be given below following Stollnitz *et al.* [1995b]; Schmidt [2007, 2012]; Zeilhofer [2008]; Koch [2011]; Koch and Schmidt [2011]; Schmidt *et al.* [2015].

**1-D scaling and wavelet functions**

Within MSR, the basis functions (5.19) are called scaling functions of level $J$. The 1-D scaling

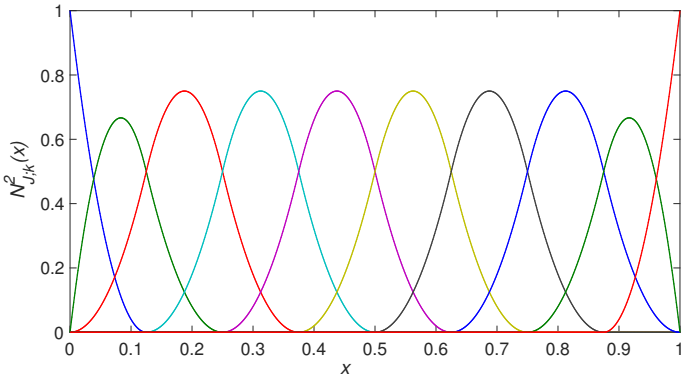



Figure 5.3: 1-D endpoint-interpolating normalized quadratic B-spline functions $N^2_{J;k}(x)$ of resolution level $J = 3$ for $k = 0, \ldots, 9$ identified by different colors.

functions $\phi_{J;k}(x)$ set up a basis for the scaling space $V_J$. The MSR generates a nested set of spaces

$$V_0 \subset V_1 \subset V_2 \cdots \subset V_{J-1} \subset V_J \subset V_{J+1} \ldots . \tag{5.28}$$

Therefore, the scaling functions $\phi_{J-1;l}(x) \in V_{J-1}$ with $l = 0, \ldots, K_{J-1} - 1$ can be represented as a linear combination of the scaling functions $\phi_{J;k}(x) \in V_J$ with $k = 0, \ldots, K_J - 1$, i.e., a linear relation so-called two-scale relation[1]

$$\boldsymbol{\phi}_{J-1}^T(x) = \boldsymbol{\phi}_J^T(x)\, \boldsymbol{P}_J \tag{5.29}$$

holds, where the $K_J \times 1$ scaling vector $\boldsymbol{\phi}_J^T(x)$ collects all the scaling functions $\phi_{J;k}(x)$ of a given resolution level $J$, i.e.,

$$\boldsymbol{\phi}_J(x) = [\phi_{J;0}(x), \ldots, \phi_{J;K_J-1}(x)]^T. \tag{5.30}$$

Similarly, it holds that

$$\boldsymbol{\phi}_{J-1}(x) = [\phi_{J-1;0}(x), \ldots, \phi_{J-1;K_{J-1}-1}(x)]^T. \tag{5.31}$$

The matrix $\boldsymbol{P}_J$ is a $K_J \times K_{J-1}$ matrix of constants that can be determined by solving Eq. (5.29) for certain values of $x$.

A detail space $W_{J-1}$ is defined as the orthogonal complement of the scaling space $V_{J-1}$ in the scaling space $V_J$, which means, $W_{J-1}$ contains all the functions in $V_J$ that are orthogonal to all the functions in $V_{J-1}$, written as

$$V_J = V_{J-1} \oplus W_{J-1} \tag{5.32}$$

where the symbol $\oplus$ denotes the direct sum, cf. Koch [1999]. The corresponding basis functions that constitute a complete basis for $W_{J-1}$ are called wavelet functions which are also compactly supported, denoted as $\psi_{J;l}(x)$ with $l = 0, \ldots, L_J - 1$ where the relation

$$L_J = K_{J+1} - K_J \tag{5.33}$$

holds. Based on the relations $W_{J-1} \subset V_j$, $\psi_{J-1;l}(x) \in W_{J-1}$ and $\phi_{J;k}(x) \in V_J$, a second two-scale relation

$$\boldsymbol{\psi}_{J-1}^T(x) = \boldsymbol{\phi}_J^T(x)\, \boldsymbol{Q}_J \tag{5.34}$$

can be established where the $L_{J-1} \times 1$ wavelet vector

$$\boldsymbol{\psi}_{J-1}(x) = [\psi_{J-1;0}(x), \ldots, \psi_{J-1;L_{J-1}-1}(x)]^T \tag{5.35}$$

is introduced. $\boldsymbol{Q}_J$ is a $K_J \times L_{J-1}$ matrix of initially unknown elements that can be determined from the matrix $\boldsymbol{P}_J$. The derivation of $\boldsymbol{Q}_j$ will not be presented within this thesis, details can be found in e.g., Zeilhofer [2008] or Schmidt [2012].

Here the $\boldsymbol{P}_J$ and $\boldsymbol{Q}_J$ matrices, calculated by Stollnitz *et al.* [1995b] for endpoint-interpolating

[1] They connect the two resolution levels (also refer to scales) $J$ and $J - 1$.

quadratic B-spline wavelets till level $J = 3$, are listed below:

$$\boldsymbol{P}_1 = \frac{1}{2}\begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}, \qquad \boldsymbol{P}_2 = \frac{1}{4}\begin{bmatrix} 4 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 4 \end{bmatrix}, \qquad \boldsymbol{P}_3 = \frac{1}{4}\begin{bmatrix} 4 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 & 0 \\ 0 & 1 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 & 0 \\ 0 & 0 & 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix},$$

$$\boldsymbol{Q}_1 = \sqrt{\frac{5}{4}}\begin{bmatrix} -2 \\ 3 \\ -3 \\ 2 \end{bmatrix}, \qquad \boldsymbol{Q}_2 = \sqrt{\frac{3}{4936}}\begin{bmatrix} -144 & 0 \\ 177 & 21 \\ -109 & -53 \\ 53 & 109 \\ -21 & -177 \\ 0 & 144 \end{bmatrix},$$

$$\boldsymbol{Q}_3 = \sqrt{\frac{1}{713568}}\begin{bmatrix} -4283.828550 & 0 & 0 & 0 \\ 5208.746077 & 780 & & 0 \\ -3099.909150 & -1949 & -11 & 0 \\ 1300.002166 & 3481 & 319 & 0 \\ -253.384964 & -3362 & -1618 & -8.737413 \\ 8.737413 & 1618 & 3362 & 253.384964 \\ 0 & -319 & -3481 & -1300.002166 \\ 0 & 11 & 1949 & 3099.909150 \\ 0 & & -780 & -5208.746077 \\ 0 & 0 & 0 & 4283.828550 \end{bmatrix}. \tag{5.36}$$

Figure 5.4 shows the family of 1-D endpoint-interpolating quadratic B-spline wavelets $\psi_{J;l}(x)$ exemplarily for the level $J = 3$.

Now the decomposition equation that is required for the MSR can be derived. According to the relations $\phi_J(x) \in V_J$, $\psi_{J-1}(x) \in W_{J-1}$ and $\phi_{J-1}(x) \in V_{J-1}$, together with Eq. (5.32), the decomposition equation

$$\boldsymbol{\phi}_J^T(x) = \boldsymbol{\phi}_{J-1}^T(x)\overline{\boldsymbol{P}}_J + \boldsymbol{\psi}_{J-1}^T(x)\overline{\boldsymbol{Q}}_J \tag{5.37}$$

can be established. The $K_{J-1} \times K_J$ matrix $\overline{\boldsymbol{P}}_J$ and $L_{J-1} \times K_J$ matrix $\overline{\boldsymbol{Q}}_J$ are initially unknown coefficient matrices. Now inserting the Eqs. (5.29) and (5.34) into Eq. (5.37) gives

$$\boldsymbol{\phi}_J^T(x) = \boldsymbol{\phi}_J^T(x)\,\boldsymbol{P}_J\,\overline{\boldsymbol{P}}_J + \boldsymbol{\phi}_J^T(x)\,\boldsymbol{Q}_J\,\overline{\boldsymbol{Q}}_J. \tag{5.38}$$

Since the scaling functions are defined uniquely, the identity

$$\mathbf{I} = \boldsymbol{P}_J\,\overline{\boldsymbol{P}}_J + \boldsymbol{Q}_J\,\overline{\boldsymbol{Q}}_J = \begin{bmatrix} \boldsymbol{P}_J & \boldsymbol{Q}_J \end{bmatrix}\begin{bmatrix} \overline{\boldsymbol{P}}_J \\ \overline{\boldsymbol{Q}}_J \end{bmatrix} \tag{5.39}$$

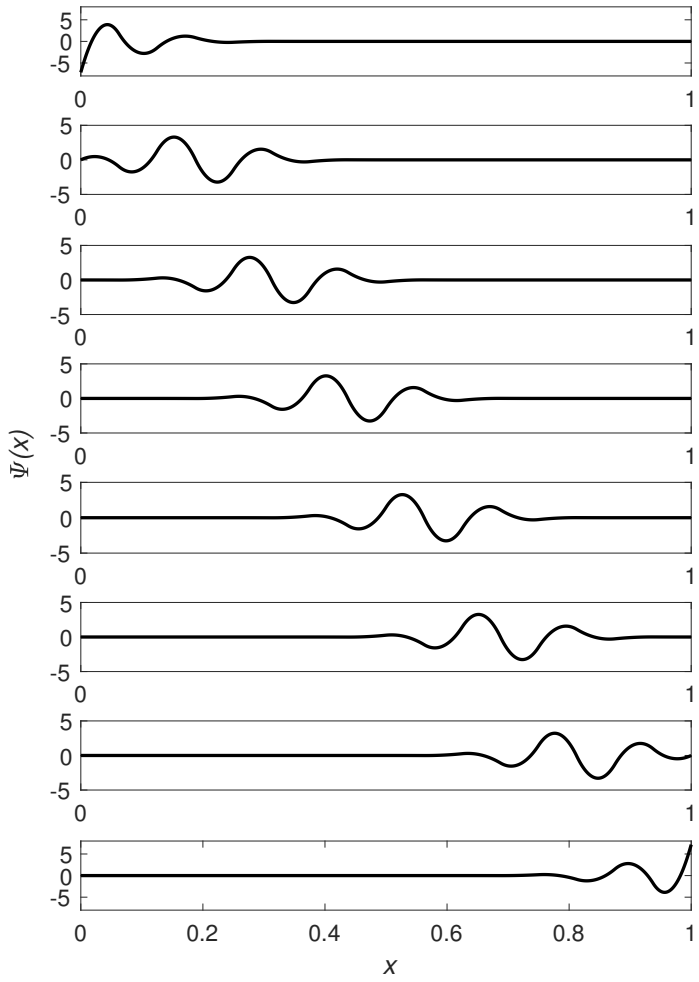



Figure 5.4: 1-D endpoint-interpolating quadratic B-spline wavelets $\psi_{J;l}(x)$ of resolution level $J = 3$ for $l = 0, \ldots, 7$ arranged from top to bottom. The first and last two wavelets are affected by the endpoint-interpolation procedure.

follows. Due to the relation $L_{J-1} = K_J - K_{J-1}$ according to Eq. (5.33), the matrices $\begin{bmatrix} \boldsymbol{P}_J & \boldsymbol{Q}_J \end{bmatrix}$ and $\begin{bmatrix} \overline{\boldsymbol{P}}_J \\ \overline{\boldsymbol{Q}}_J \end{bmatrix}$ are both square matrices. $\overline{\boldsymbol{P}}_J$ and $\overline{\boldsymbol{Q}}_J$ can be derived according to

$$\begin{bmatrix} \overline{\boldsymbol{P}}_J \\ \overline{\boldsymbol{Q}}_J \end{bmatrix} = \begin{bmatrix} \boldsymbol{P}_J & \boldsymbol{Q}_J \end{bmatrix}^{-1}. \tag{5.40}$$

### MSR of 1-D signal

The basic principle of the MSR is to split an input signal into a smoothed version and a certain number of detail signals by successive low-pass filtering [Schmidt, 2007]. Thus, the MSR of a 1-D input signal denoted as $f_J(x)$ with $f_J \in V_j$ can be written as

$$f_J(x) = f_{J'}(x) + \sum_{j=J'}^{J-1} g_j(x) \quad \text{with} \quad J' \in \{0, \cdots, J-1\} \tag{5.41}$$

where $f_{J'}(x)$ with $f_{J'} \in V_{J'}$ means a smoothed, i.e., a low-pass filtered version of $f_J(x)$ and $g_j(x)$ with $g_j \in W_j$ denotes a detail signal of level $j$, which is a band-pass filtered version of $f_J(x)$.

The signal $f_{J'}(x)$ can be represented as a series expansion in terms of scaling functions $\phi_{J';k}(x)$ $(k = 0, \ldots, K_{J'} - 1)$ with unknown scaling coefficients $d_{J';k}$ as

$$f_{J'}(x) = \sum_{k=0}^{K_{J'}-1} d_{J';k}\, \phi_{J';k}(x) = \boldsymbol{\phi}_{J'}^T(x)\, \boldsymbol{d}_{J'} \tag{5.42}$$

where the $K_{J'} \times 1$ scaling coefficient vector

$$\boldsymbol{d}_{J'} = [d_{J';0}, \ldots, d_{J';K_{J'}-1}]^T \tag{5.43}$$

collects all the scaling coefficients of level $J'$.

Similarly, the detail signal $g_j(x)$ can be represented as a series expansion in terms of wavelet functions $\psi_{j;k}(x)$ $(k = 0, \ldots, L_j - 1)$ with unknown wavelet coefficients $c_{j;k}$ as

$$g_j(x) = \sum_{k=0}^{L_j-1} c_{j;k}\, \psi_{j;k}(x) = \boldsymbol{\psi}_j^T(x)\, \boldsymbol{c}_j \tag{5.44}$$

where the $L_j \times 1$ wavelet coefficient vector of level $j$ is defined as

$$\boldsymbol{c}_j = [c_{j;0}, \ldots, c_{j;L_j-1}]^T. \tag{5.45}$$

According to Eq. (5.41), one can write

$$f_J(x) = f_{J-1}(x) + g_{J-1}(x). \tag{5.46}$$

Now inserting the Eqs. (5.42) and (5.44) into Eq. (5.46) yields

$$\boldsymbol{\phi}_J^T(x)\, \boldsymbol{d}_J = \boldsymbol{\phi}_{J-1}^T(x)\, \boldsymbol{d}_{J-1} + \boldsymbol{\psi}_{J-1}^T(x)\, \boldsymbol{c}_{J-1}. \tag{5.47}$$

A comparison of Eq. (5.47) with Eq. (5.37) yields the relations

$$\boldsymbol{d}_{J-1} = \overline{\boldsymbol{P}}_J\, \boldsymbol{d}_J \tag{5.48a}$$

$$\boldsymbol{c}_{J-1} = \overline{\boldsymbol{Q}}_J\, \boldsymbol{d}_J \tag{5.48b}$$

which reflect the dependency of the scaling and wavelet coefficient vectors between adjacent levels. This decomposition process, i.e., a downsampling procedure, can be numerically realized by the highly effective pyramid algorithm (Fast Wavelet Transform (FWT)), which is illustrated in Fig. 5.5. Starting from the scaling coefficient vector $\boldsymbol{d}_J$ of the maximum level $J$, the scaling and wavelet coefficient vectors of the lower levels can be successively obtained and, thus, the MSR can be achieved.

Conversely, it is also possible to reconstruct the scaling coefficient vector of a higher level through scaling and wavelet coefficient vectors of the lower level. To do this, Eq. (5.48) can be written as

$$\begin{bmatrix} \boldsymbol{d}_{J-1} \\ \boldsymbol{c}_{J-1} \end{bmatrix} = \begin{bmatrix} \overline{\boldsymbol{P}}_J \\ \overline{\boldsymbol{Q}}_J \end{bmatrix} \boldsymbol{d}_J, \tag{5.49}$$

and $\boldsymbol{d}_j$ can then be solved under consideration of Eq. (5.40) through

$$\boldsymbol{d}_J = \begin{bmatrix} \boldsymbol{P}_J & \boldsymbol{Q}_J \end{bmatrix} \begin{bmatrix} \boldsymbol{d}_{J-1} \\ \boldsymbol{c}_{J-1} \end{bmatrix} = \boldsymbol{P}_J\, \boldsymbol{d}_{J-1} + \boldsymbol{Q}_J\, \boldsymbol{c}_{J-1}. \tag{5.50}$$

This reconstruction process, i.e., an upsampling procedure, is illustrated in Fig. 5.6.

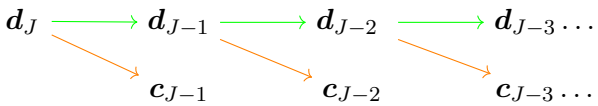



Figure 5.5: 1-D pyramid algorithm (from left to right): scaling coefficient vectors (top) and wavelet coefficient vectors (bottom). The green arrows indicate low-pass filtering, and the orange arrows denote band-pass filtering.

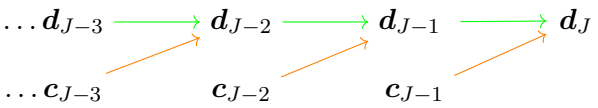



Figure 5.6: 1-D reconstruction algorithm (from left to right): scaling coefficient vectors (top) and wavelet coefficient vectors (bottom).

**3-D scaling and wavelet functions**

As shown in Eq. (5.17), ionosphere key parameters are modeled with respect to $\lambda$, $\phi$ and $t$, and thus we have to deal with the MSR of a 3-D signal. As mentioned before, a 3-D B-spline scaling function $\phi^{J_1,J_2,J_3}_{k_1,k_2,k_3}(\lambda,\varphi,t)$ is defined as the tensor product of three 1-D scaling function $\phi_{J;k}(x)$ with $x \in \{\lambda,\varphi,t\}$ and $J \in \{J_1,J_2,J_3\}$. The 3-D scaling functions $\phi_{\boldsymbol{J};\boldsymbol{k}}(\lambda,\varphi,t) = \phi^{J_1,J_2,J_3}_{k_1,k_2,k_3}(\lambda,\varphi,t)$ with $\boldsymbol{J} = [J_1,J_2,J_3]^T$ and $\boldsymbol{k} = [k_1,k_2,k_3]^T$ for $k_1 = 0,\ldots,K_{J_1}-1$, $k_2 = 0,\ldots,K_{J_2}-1$ and $k_3 = 0,\ldots,K_{J_3}-1$ are collected in the $K_{\boldsymbol{J}} \times 1$ vector $\boldsymbol{\phi}_{\boldsymbol{J}}(\lambda,\varphi,t)$ with $K_{\boldsymbol{J}} = K_{J_1}\cdot K_{J_2}\cdot K_{J_3}$, and can be defined as

$$\begin{aligned}\boldsymbol{\phi}_{\boldsymbol{J}}(\lambda,\varphi,t) &= [\phi_{\boldsymbol{J};0,0,0}(\lambda,\varphi,t),\ \ldots\ ,\phi_{\boldsymbol{J};K_{J_1}-1,K_{J_2}-1,K_{J_3}-1}(\lambda,\varphi,t)]^T \\ &= \boldsymbol{\phi}_{J_3}(t)\otimes\boldsymbol{\phi}_{J_2}(\varphi)\otimes\boldsymbol{\phi}_{J_1}(\lambda)\,,\end{aligned} \tag{5.51}$$

wherein the symbol $\otimes$ denotes the Kronecker product of the three vectors $\phi_J(x)$ introduced in Eq. (5.30) (cf. Koch, 1999). Corresponding to the 3-D scaling functions $\phi_{\boldsymbol{J};\boldsymbol{k}}(\lambda,\varphi,t)$, a 3-D scaling space $\boldsymbol{V}_{\boldsymbol{J}}$ is defined by the tensor product of three 1-D scaling spaces $V_{J_1}$, $V_{J_2}$ and $V_{J_3}$, i.e.,

$$\boldsymbol{V}_{\boldsymbol{J}} = V_{J_1}\otimes V_{J_2}\otimes V_{J_3}. \tag{5.52}$$

According to Eq. (5.32), Eq. (5.52) can be reformulated as

$$\boldsymbol{V}_{\boldsymbol{J}} = (V_{J_1-1}\oplus W_{J_1-1})\otimes(V_{J_2-1}\oplus W_{J_2-1})\otimes(V_{J_3-1}\oplus W_{J_3-1}) \tag{5.53}$$

which leads to altogether $2^3$ tensor product spaces, denoted as

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{J-1}} &= V_{J_1-1}\otimes V_{J_2-1}\otimes V_{J_3-1},\\ \boldsymbol{W}^1_{\boldsymbol{J-1}} &= V_{J_1-1}\otimes V_{J_2-1}\otimes W_{J_3-1},\\ \boldsymbol{W}^2_{\boldsymbol{J-1}} &= V_{J_1-1}\otimes W_{J_2-1}\otimes V_{J_3-1},\\ \boldsymbol{W}^3_{\boldsymbol{J-1}} &= W_{J_1-1}\otimes V_{J_2-1}\otimes V_{J_3-1},\\ \boldsymbol{W}^4_{\boldsymbol{J-1}} &= V_{J_1-1}\otimes W_{J_2-1}\otimes W_{J_3-1},\\ \boldsymbol{W}^5_{\boldsymbol{J-1}} &= W_{J_1-1}\otimes V_{J_2-1}\otimes W_{J_3-1},\\ \boldsymbol{W}^6_{\boldsymbol{J-1}} &= W_{J_1-1}\otimes W_{J_2-1}\otimes V_{J_3-1},\\ \boldsymbol{W}^7_{\boldsymbol{J-1}} &= W_{J_1-1}\otimes W_{J_2-1}\otimes W_{J_3-1}.\end{aligned} \tag{5.54}$$

Consequently, Eq. (5.53) can also be written as

$$\boldsymbol{V_J} = \boldsymbol{V_{J-1}} \oplus \boldsymbol{W^1_{J-1}} \oplus \ldots \oplus \boldsymbol{W^7_{J-1}} \,. \tag{5.55}$$

The corresponding basis functions for the above spaces can be formulated as

$$\begin{aligned}
&\text{scaling function} \quad && \phi_{\boldsymbol{J};\boldsymbol{k}}(\lambda,\varphi,t) = \phi_{J_1;k_1}(\lambda)\cdot\phi_{J_2;k_2}(\phi)\cdot\phi_{J_3;k_3}(t)\,, \\
&\text{wavelet function} \quad && \begin{cases}
\psi^1_{\boldsymbol{J};\boldsymbol{k}}(\lambda,\varphi,t) = \phi_{J_1;k_1}(\lambda)\cdot\phi_{J_2;k_2}(\phi)\cdot\psi_{J_3;k_3}(t) \\
\psi^2_{\boldsymbol{J};\boldsymbol{k}}(\lambda,\varphi,t) = \phi_{J_1;k_1}(\lambda)\cdot\psi_{J_2;k_2}(\phi)\cdot\phi_{J_3;k_3}(t) \\
\psi^3_{\boldsymbol{J};\boldsymbol{k}}(\lambda,\varphi,t) = \psi_{J_1;k_1}(\lambda)\cdot\phi_{J_2;k_2}(\phi)\cdot\phi_{J_3;k_3}(t) \\
\psi^4_{\boldsymbol{J};\boldsymbol{k}}(\lambda,\varphi,t) = \phi_{J_1;k_1}(\lambda)\cdot\psi_{J_2;k_2}(\phi)\cdot\psi_{J_3;k_3}(t) \\
\psi^5_{\boldsymbol{J};\boldsymbol{k}}(\lambda,\varphi,t) = \psi_{J_1;k_1}(\lambda)\cdot\phi_{J_2;k_2}(\phi)\cdot\psi_{J_3;k_3}(t) \\
\psi^6_{\boldsymbol{J};\boldsymbol{k}}(\lambda,\varphi,t) = \psi_{J_1;k_1}(\lambda)\cdot\psi_{J_2;k_2}(\phi)\cdot\phi_{J_3;k_3}(t) \\
\psi^7_{\boldsymbol{J};\boldsymbol{k}}(\lambda,\varphi,t) = \psi_{J_1;k_1}(\lambda)\cdot\psi_{J_2;k_2}(\phi)\cdot\psi_{J_3;k_3}(t)
\end{cases} .
\end{aligned} \tag{5.56}$$

**MSR of 3-D signal**

According to Eq. (5.17), a 3-D signal $f_{\boldsymbol{J}}(\lambda,\varphi,t)$ can be represented as

$$f_{\boldsymbol{J}}(\lambda,\varphi,t) = \sum_{k_1=0}^{K_{J_1}-1}\sum_{k_2=0}^{K_{J_2}-1}\sum_{k_3=0}^{K_{J_3}-1} d^{J_1,J_2,J_3}_{k_1,k_2,k_3}\,\phi^{J_1,J_2,J_3}_{k_1,k_2,k_3}(\lambda,\varphi,t) = \boldsymbol{\phi_J}(\lambda,\varphi,t)\,\boldsymbol{d_J} \tag{5.57}$$

wherein the $K_{\boldsymbol{J}} \times 1$ scaling coefficient vector is defined as

$$\boldsymbol{d_J} = \left[d_{\boldsymbol{J};0,0,0},\ \ldots,\ d_{\boldsymbol{J};K_{J_1}-1,K_{J_2}-1,K_{J_3}-1}\right]^T . \tag{5.58}$$

Similar to Eq. (5.41), the 3-D MSR of a signal $f_{\boldsymbol{J}}(\lambda,\varphi,t)$ can be defined under consideration of Eq. (5.55) as

$$f_{\boldsymbol{J}}(\lambda,\varphi,t) = f_{\boldsymbol{J-m}}(\lambda,\varphi,t) + \sum_{i=1}^{m}\sum_{n=1}^{2^3-1} g^n_{\boldsymbol{J-i}}(\lambda,\varphi,t) \tag{5.59}$$

with

$$\begin{gathered}
\boldsymbol{J-m} = [J_1-m, J_2-m, J_3-m]^T = [J'_1, J'_2, J'_3]^T,\ m \in \{1,\ldots,\min(J_1,J_2,J_3)\}, \\
\text{and} \quad \boldsymbol{J-i} = [J_1-i, J_2-i, J_3-i]^T = [\overline{J}_1, \overline{J}_2, \overline{J}_3]^T
\end{gathered} \tag{5.60}$$

where $\min\{\cdot\}$ is the operation for taking the minimum value.

The smoothed version $f_{\boldsymbol{J'}}(\lambda,\varphi,t)$ of level $\boldsymbol{J'}$ $(= \boldsymbol{J-m})$ is defined as

$$f_{\boldsymbol{J'}}(\lambda,\varphi,t) = \sum_{k_1=0}^{K_{J'_1}-1}\sum_{k_2=0}^{K_{J'_2}-1}\sum_{k_3=0}^{K_{J'_3}-1} d^{J'_1,J'_2,J'_3}_{k_1,k_2,k_3}\,\phi^{J'_1,J'_2,J'_3}_{k_1,k_2,k_3}(\lambda,\varphi,t) = \boldsymbol{\phi}^T_{\boldsymbol{J'}}(\lambda,\varphi,t)\,\boldsymbol{d_{J'}} \tag{5.61}$$

and the detail signal $g^n_{\overline{\boldsymbol{J}}}(\lambda,\varphi,t)$ of the level $\overline{\boldsymbol{J}}$ $(= \boldsymbol{J-i})$is defined as

$$g^n_{\overline{\boldsymbol{J}}}(\lambda,\varphi,t) = \sum_{k_1=0}^{L_{\overline{J}_1}-1}\sum_{k_2=0}^{L_{\overline{J}_2}-1}\sum_{k_3=0}^{l_{\overline{J}_3}-1} c^n_{\overline{\boldsymbol{J}};k_1,k_2,k_3}\,\psi^n_{\overline{\boldsymbol{J}};k_1,k_2,k_3}(\lambda,\varphi,t) = \left(\boldsymbol{\psi}^n_{\overline{\boldsymbol{J}}}(\lambda,\varphi,t)\right)^T \boldsymbol{c}^n_{\overline{\boldsymbol{J}}} \tag{5.62}$$

where

$$\boldsymbol{\psi}^n_{\overline{\boldsymbol{J}}}(\lambda,\varphi,t) = [\psi^n_{\overline{\boldsymbol{J}};0,0,0},\ldots,\psi^n_{\overline{\boldsymbol{J}};L_{\overline{J}_1}-1,L_{\overline{J}_2}-1,L_{\overline{J}_3}-1}]^T \tag{5.63}$$

and $\boldsymbol{c}^n_{\overline{\boldsymbol{J}}}$ with $n = 1,\ldots,2^3-1$ is the wavelet coefficient vector, reads

$$\boldsymbol{c}^n_{\overline{\boldsymbol{J}}} = [c^n_{\overline{\boldsymbol{J}};0,0,0},\ \cdots,\ c^n_{\overline{\boldsymbol{J}};L_{\overline{J}_1}-1,L_{\overline{J}_2}-1,L_{\overline{J}_3}-1}]^T. \tag{5.64}$$

The 1-D decomposition equation Eq. (5.48) of coefficient vectors can be extended to the 3-D case as follows

$$\begin{aligned}
\boldsymbol{d}_{\boldsymbol{J-1}} &= (\overline{\boldsymbol{P}}_{J_3}\otimes\overline{\boldsymbol{P}}_{J_2}\otimes\overline{\boldsymbol{P}}_{J_1})\,\boldsymbol{d}_{\boldsymbol{J}},\\
\boldsymbol{c}^1_{\boldsymbol{J-1}} &= (\overline{\boldsymbol{Q}}_{J_3}\otimes\overline{\boldsymbol{P}}_{J_2}\otimes\overline{\boldsymbol{P}}_{J_1})\,\boldsymbol{d}_{\boldsymbol{J}},\\
\boldsymbol{c}^2_{\boldsymbol{J-1}} &= (\overline{\boldsymbol{P}}_{J_3}\otimes\overline{\boldsymbol{Q}}_{J_2}\otimes\overline{\boldsymbol{P}}_{J_1})\,\boldsymbol{d}_{\boldsymbol{J}},\\
\boldsymbol{c}^3_{\boldsymbol{J-1}} &= (\overline{\boldsymbol{P}}_{J_3}\otimes\overline{\boldsymbol{P}}_{J_2}\otimes\overline{\boldsymbol{Q}}_{J_1})\,\boldsymbol{d}_{\boldsymbol{J}},\\
\boldsymbol{c}^4_{\boldsymbol{J-1}} &= (\overline{\boldsymbol{Q}}_{J_3}\otimes\overline{\boldsymbol{Q}}_{J_2}\otimes\overline{\boldsymbol{P}}_{J_1})\,\boldsymbol{d}_{\boldsymbol{J}},\\
\boldsymbol{c}^5_{\boldsymbol{J-1}} &= (\overline{\boldsymbol{Q}}_{J_3}\otimes\overline{\boldsymbol{P}}_{J_2}\otimes\overline{\boldsymbol{Q}}_{J_1})\,\boldsymbol{d}_{\boldsymbol{J}},\\
\boldsymbol{c}^6_{\boldsymbol{J-1}} &= (\overline{\boldsymbol{P}}_{J_3}\otimes\overline{\boldsymbol{Q}}_{J_2}\otimes\overline{\boldsymbol{Q}}_{J_1})\,\boldsymbol{d}_{\boldsymbol{J}},\\
\boldsymbol{c}^7_{\boldsymbol{J-1}} &= (\overline{\boldsymbol{Q}}_{J_3}\otimes\overline{\boldsymbol{Q}}_{J_2}\otimes\overline{\boldsymbol{Q}}_{J_1})\,\boldsymbol{d}_{\boldsymbol{J}}.
\end{aligned} \tag{5.65}$$

Similar to Fig. 5.5, this decomposition process for 3-D signal is illustrated in Fig. 5.7. Once

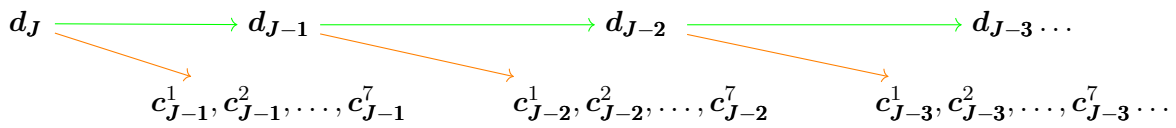



Figure 5.7: 3-D pyramid algorithm.

the level-dependent scaling and wavelet coefficient vectors are derived, the smoothed version of the 3-D signal can be computed according to Eq. (5.61), and the level-$\overline{\boldsymbol{J}}$ detail signal can be computed by

$$\boldsymbol{g}_{\overline{\boldsymbol{J}}}(\lambda,\varphi,t) = \sum_{n=1}^{7}\boldsymbol{g}^n_{\overline{\boldsymbol{J}}}(\lambda,\varphi,t) \tag{5.66}$$

under consideration of Eq. (5.62).

Similar to the Eqs. (5.49) and (5.50), the reconstruction, i.e., the computation of the level-$\overline{\boldsymbol{J}}$ scaling coefficient vector $\boldsymbol{d}_{\overline{\boldsymbol{J}}}$ can be derived from the level-$(\overline{\boldsymbol{J}}-\boldsymbol{1})$ vector $\boldsymbol{d}_{\overline{\boldsymbol{J}}-\boldsymbol{1}}$ and $\boldsymbol{c}^n_{\overline{\boldsymbol{J}}-\boldsymbol{1}}$ through the inversion of the linear equation system (5.65).

**MSR application – data compression**

Wavelet functions are also localizing (cf. Fig. 5.4), and thus many wavelet coefficients in the vector $\boldsymbol{c}^n_{\boldsymbol{J-i}}$ are numerically very small and negligible [Schmidt, 2007]. The wavelet decomposition can thus be applied for data compression, mainly through choosing empirical thresholds or statistical hypothesis testing on significance.

- Empirical thresholds:
  Since detail signals of higher-level contain more small-scale structures, more wavelet coefficients can usually be neglected. Therefore, level-dependent thresholds can be applied. Specifically, the wavelet coefficient $c^n_{\boldsymbol{J};\boldsymbol{k}}$ of a certain resolution level $\boldsymbol{J}$ may be neglected (i.e., $c^n_{\boldsymbol{J};\boldsymbol{k}}$ is set to 0) if $|c^n_{\boldsymbol{J};\boldsymbol{k}}| < \varrho_{\boldsymbol{J}}$ where $\varrho_{\boldsymbol{J}}$ denotes the level-dependent threshold.
- Statistical hypothesis testing on significance:
  Applying the law of error propagation to the decomposition equation (5.65), the estimated covariance matrices of the estimated scaling and wavelet coefficient vectors can be derived, from the estimated covariance matrix $\widehat{D}(\widehat{\boldsymbol{d}}_{\boldsymbol{J}})$, which is calculable from Eq. (4.18). Afterwards, hypothesis testing can be performed to check the wavelet coefficients on significance following the Eqs. (4.56) and (4.57); see, e.g., Koch [1999] for details. Coefficients that fail the significance test are then neglected.

Other data compression algorithms can be found in e.g., Ogden [1997]. Through the new wavelet coefficient vectors $\overline{\boldsymbol{c}}^n_{\boldsymbol{j};\boldsymbol{k}}$ whose smallest/most nonsignificant wavelet coefficients are neglected, a compressed signal $\overline{f}_{\boldsymbol{J}}$ can then be reconstructed. In an ideal case of data compression, the reconstructed signal $\overline{f}_{\boldsymbol{J}}$ does not lose significant information in $f_{\boldsymbol{J}}$; very often only a small number of coefficients need to be stored.

## 5.3 Set-up of an electron density model

### 5.3.1 Linearized observation equation system

When the electron density is modeled by Eq. (5.8) with discrete layers being represented by profilers introduced in Section 5.1.1, the electron density can thus be represented as a function of the key parameters such as $NmF2$, $hmF2$, $HF2$, $NmE$ etc. Exploiting the localizing feature of the B-spline basis functions, each of the key parameters can be spatially and temporally modeled as a series expansion in terms of the 3-D B-spline functions according to Eq. (5.17). Therefore, STEC observations and electron density observations can be finally represented by a set of unknown B-spline coefficients. However, mathematical expressions such as the Chapman or Epstein functions are nonlinear with respect to the parameters peak heights and scale heights, a nonlinear scheme is thus required to solve the unknown scaling coefficients. Since an explicit formula (see the Eqs. (6.1) and (6.15) later) is available, the partial derivatives of the electron density with respect to the key parameters can be computed analytically and easily obtained. Therefore, a linearization can be applied and an iterative algorithm is performed. For this purpose, the unknown coefficient vector $\boldsymbol{d}_{\kappa_q}$ of a certain key parameter $\kappa_q$ ($q = 1, \ldots, p$ with $p$ denoting the maximum number of key parameters to be modeled) is decomposed into an initial coefficient vector $\boldsymbol{d}_{0,\kappa_q}$ and a correction term $\Delta\boldsymbol{d}_{\kappa_q}$.

Let

$$\begin{aligned} \boldsymbol{d} &= [\boldsymbol{d}^T_{\kappa_1}, \cdots, \boldsymbol{d}^T_{\kappa_p}]^T, \\ \boldsymbol{d}_0 &= [\boldsymbol{d}^T_{0,\kappa_1}, \cdots, \boldsymbol{d}^T_{0,\kappa_p}]^T, \\ \Delta\boldsymbol{d} &= [\Delta\boldsymbol{d}^T_{\kappa_1}, \cdots, \Delta\boldsymbol{d}^T_{\kappa_p}]^T, \end{aligned} \tag{5.67}$$

we have

$$\boldsymbol{d} = \boldsymbol{d}_0 + \Delta\boldsymbol{d} \tag{5.68}$$

where the vector $\boldsymbol{d}_0$ of initial coefficients can be derived from an external model such as the IRI. According to Eq. (5.17), we can write the following scalar product

$$\kappa_q(\lambda,\varphi,t) = \boldsymbol{\phi}_{\kappa_q}^T(\lambda,\varphi,t)\,\boldsymbol{d}_{\kappa_q} \tag{5.69}$$

with

$$\boldsymbol{\phi}_{\kappa_q}(\lambda,\varphi,t) = [\boldsymbol{\phi}_{\kappa_q\,0,0,0}^{J_1,J_2,J_3}(\lambda,\varphi,t),\,\cdots,\,\boldsymbol{\phi}_{\kappa_q\,K_{J_1}-1,K_{J_2}-1,K_{J_3}-1}^{J_1,J_2,J_3}(\lambda,\varphi,t)]^T. \tag{5.70}$$

for the key parameter $\kappa_q$. Note that the level values $J_1$, $J_2$ and $J_3$ can be different for each key parameter $\kappa_q$.

### $N_e$ model of GNSS IRO observations

Assume that $m$ profiles $p_i$ $(i = 1,\ldots,m)$ with discrete $N_e$ observations at $n$ positions $(\lambda_{ij}, \varphi_{ij}, h_{ij}, t_{ij})$ $(j = 1,\ldots,n)$ lie in the region under investigation. Applying the chain rule $\frac{\partial u(x,y)}{\partial r} = \frac{\partial u(x,y)}{\partial x}\cdot\frac{\partial x}{\partial r} + \frac{\partial u(x,y)}{\partial y}\cdot\frac{\partial y}{\partial r}$ under consideration of the Eqs. (4.2), (5.68) and (5.69), a system of linearized observation equations

$$\begin{bmatrix} N_e(\lambda_{11},\varphi_{11},h_{11},t_{11}) \\ \vdots \\ N_e(\lambda_{mn},\varphi_{mn},h_{mn},t_{mn}) \end{bmatrix} + \begin{bmatrix} e(\lambda_{11},\varphi_{11},h_{11},t_{11}) \\ \vdots \\ e(\lambda_{mn},\varphi_{mn},h_{mn},t_{mn}) \end{bmatrix} = \begin{bmatrix} \overline{N}_e(\lambda_{11},\varphi_{11},h_{11},t_{11})\big|_{\boldsymbol{d}_0} \\ \vdots \\ \overline{N}_e(\lambda_{mn},\varphi_{mn},h_{mn},t_{mn})\big|_{\boldsymbol{d}_0} \end{bmatrix} +$$

$$\begin{bmatrix} \frac{\partial\overline{N}_e(\lambda_{11},\varphi_{11},h_{11},t_{11})}{\partial\kappa_1}\big|_{\boldsymbol{d}_0}\cdot\frac{\partial\kappa_1(\lambda_{11},\varphi_{11},t_{11})}{\partial\boldsymbol{d}_{\kappa_1}} & \cdots & \frac{\partial\overline{N}_e(\lambda_{11},\varphi_{11},h_{11},t_{11})}{\partial\kappa_p}\big|_{\boldsymbol{d}_0}\cdot\frac{\partial\kappa_p(\lambda_{11},\varphi_{11},t_{11})}{\partial\boldsymbol{d}_{\kappa_p}} \\ \vdots & \ddots & \vdots \\ \frac{\partial\overline{N}_e(\lambda_{mn},\varphi_{mn},h_{mn},t_{mn})}{\partial\kappa_1}\big|_{\boldsymbol{d}_0}\cdot\frac{\partial\kappa_1(\lambda_{mn},\varphi_{mn},t_{mn})}{\partial\boldsymbol{d}_{\kappa_1}} & \cdots & \frac{\partial\overline{N}_e(\lambda_{mn},\varphi_{mn},h_{mn},t_{mn})}{\partial\kappa_p}\big|_{\boldsymbol{d}_0}\cdot\frac{\partial\kappa_p(\lambda_{mn},\varphi_{mn},t_{mn})}{\partial\boldsymbol{d}_{\kappa_p}} \end{bmatrix}\Delta\boldsymbol{d} \tag{5.71}$$

can be derived. In order to avoid confusions, $N_e$ on the left-hand side is assumed to be an observation with the observation error $e(\cdot)$, whereas $\overline{N}_e$ represents the quantity calculated by a model, e.g., Eq. (5.8). The first partial derivative term $\partial\overline{N}_e/\partial\kappa_q$ depends on the used profilers, and the second term is in fact $\partial\kappa_q(\lambda_{ij},\varphi_{ij},t_{ij})/\partial\boldsymbol{d}_{\kappa_q} = \boldsymbol{\phi}^T(\lambda_{ij},\varphi_{ij},t_{ij})$ according to Eq. (5.69). As introduced previously, the variables $\lambda$, $\varphi$ and $t$ have to be transformed to $\lambda',\varphi',t' \in \mathbb{I}$. Here the two notations will not be distinguished to avoid confusions, namely, the observations at $(\lambda_{ij},\varphi_{ij},t_{ij})$ within the investigated region $\Omega^3$ have already been normalized into the unit cube $\mathbb{I}^3$ in a preprocessing step before calculating the B-splines. Assume that there are altogether $N_1$ discrete $N_e$ discrete observations, denoted as $N_{e,1},\ldots,N_{e,N_1}$ with the corresponding observation errors $e_1,\ldots,e_{N_1}$, Eq. (5.71) can be formulated in the form of the Gauss-Markov model (4.5) as

$$\boldsymbol{X}_1\,\Delta\boldsymbol{\beta}_1 = \boldsymbol{y}_1 + \boldsymbol{e}_1 \quad \text{with} \quad D(\boldsymbol{y}_1) = \sigma_1^2\,\boldsymbol{P}_1^{-1} \tag{5.72}$$

where

$$\boldsymbol{X}_1 = \begin{pmatrix} \frac{\partial\overline{N}_{e,1}}{\partial\boldsymbol{d}_{\kappa_1}}\big|_{\boldsymbol{d}_0} & \cdots & \frac{\partial\overline{N}_{e,1}}{\partial\boldsymbol{d}_{\kappa_p}}\big|_{\boldsymbol{d}_0} \\ \frac{\partial\overline{N}_{e,2}}{\partial\boldsymbol{d}_{\kappa_1}}\big|_{\boldsymbol{d}_0} & \cdots & \frac{\partial\overline{N}_{e,2}}{\partial\boldsymbol{d}_{\kappa_p}}\big|_{\boldsymbol{d}_0} \\ \vdots & \ddots & \vdots \\ \frac{\partial\overline{N}_{e,N_1}}{\partial\boldsymbol{d}_{\kappa_1}}\big|_{\boldsymbol{d}_0} & \cdots & \frac{\partial\overline{N}_{e,N_1}}{\partial\boldsymbol{d}_{\kappa_p}}\big|_{\boldsymbol{d}_0} \end{pmatrix},\quad \Delta\boldsymbol{\beta}_1 = \Delta\boldsymbol{d}$$

$$\boldsymbol{y}_1 = [N_{e,1} - \overline{N}_{e,1}\big|_{\boldsymbol{d}_0},\,\ldots,\,N_{e,N_1} - \overline{N}_{e,N_1}\big|_{\boldsymbol{d}_0}]^T,\quad \boldsymbol{e}_1 = [e_1,\,\ldots,\,e_{N_1}]^T.$$

The stochastic model of the $N_e$ observations is defined by the covariance matrix $D(\boldsymbol{y}_1)$ which includes the given positive definite $N_1 \times N_1$ weight matrix $\boldsymbol{P}_1$ and the unknown variance factor $\sigma_1^2$.

**Stochastic model of $Ne$ observations**

A realistic solution can only be obtained, if the stochastic model in the Gauss-Markov model (4.5) is set up realistically. In most cases, the weight matrix $\boldsymbol{P}_1$ is simply set to the identity matrix, e.g., Alizadeh [2013], Limberger *et al.* [2013] or Liang *et al.* [2015a].

Recall the onion peeling algorithm Eq. (3.23), collecting all data during an occultation event, a system of linear equations

$$\underbrace{\begin{bmatrix} \widetilde{STEC}(r_{01}) \\ \widetilde{STEC}(r_{02}) \\ \vdots \\ \widetilde{STEC}(r_{0M}) \end{bmatrix}}_{\widetilde{\boldsymbol{STEC}}} = \underbrace{\begin{bmatrix} 2\,l_{11} & 0 & \cdots & 0 \\ 2\,l_{21} & 2\,l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 2\,l_{M1} & 2\,l_{M2} & \cdots & 2\,l_{MM} \end{bmatrix}}_{\boldsymbol{B}} \underbrace{\begin{bmatrix} N_{e,1} \\ N_{e,2} \\ \vdots \\ N_{e,M} \end{bmatrix}}_{\boldsymbol{N}_e} \tag{5.73}$$

can be established. $\boldsymbol{B}$ is a lower triangular matrix. The unknown vector $\boldsymbol{N}_e$ can thus be determined by

$$\widehat{\boldsymbol{N}}_e = \boldsymbol{B}^{-1}\,\widetilde{\boldsymbol{STEC}}\,. \tag{5.74}$$

Applying to the law of error propagation to Eq. (5.74), the covariance matrix $D(\boldsymbol{N}_e)$ can be derived through

$$D(\widehat{\boldsymbol{N}}_e) = \boldsymbol{B}^{-1}\,D(\widetilde{\boldsymbol{STEC}})\,(\boldsymbol{B}^{-1})^T\,. \tag{5.75}$$

If $D(\widetilde{\boldsymbol{STEC}})$ is known, the standard deviations of $N_{e,i}$ $(i = 1, \ldots, M)$ can be obtained and the vertical error correlation can also be calculated according to Eq. (4.7). It can be expected that errors are increasing from the outermost toward the innermost shells (cf. Fig. 3.4). The weight matrix $\boldsymbol{P}_1$ is then given as

$$\boldsymbol{P}_1 = \left(D(\widehat{\boldsymbol{N}}_e)\right)^{-1}\,. \tag{5.76}$$

**STEC model of GNSS observations**

Within this thesis, carrier phase leveling is applied to GPS observations. According to Eq. (3.17), we can establish the observation equation as

$$STEC(R,S,t) + e(R,S,t) = \int_R^S \overline{N}_e(\lambda,\varphi,h,t)\,ds + \beta_R + \beta^S \tag{5.77}$$

where $STEC(R,S,t)$ is assumed to be the observation from a satellite $S$ to a receiver $R$ with observation error $e(R,S,t)$. $\beta_R$ and $\beta^S$ are biases given as

$$\beta_R = -\frac{c}{a}\,\delta b_R \quad \text{and} \quad \beta^S = -\frac{c}{a}\,\delta b^S \tag{5.78}$$

where $\delta b_R$ and $\delta b^S$ are the DCBs of receiver $R$ and satellite $S$, respectively. A linearization of Eq. (5.77) gives

$$\begin{aligned} & STEC(R,S,t) + e(R,S,t) \\ =& \int_R^S \left( \overline{N}_e(\lambda,\varphi,h,t)\big|_{\boldsymbol{d}_0} + \sum_{q=1}^{p} \left.\frac{\partial \overline{N}_e(\lambda,\varphi,h,t)}{\partial \kappa_q}\right|_{\boldsymbol{d}_{0,\kappa_q}} \cdot \boldsymbol{\phi}^T(\lambda,\varphi,t) \cdot \Delta \boldsymbol{d}_{\kappa_q} \right) ds + \beta_R + \beta^S \\ =& \overline{STEC}_0(R,S,t) + \sum_{q=1}^{p} \left[ \int_R^S \left( \left.\frac{\partial \overline{N}_e(\lambda,\varphi,h,t)}{\partial \kappa_q}\right|_{\boldsymbol{d}_{0,\kappa_q}} \cdot \boldsymbol{\phi}^T(\lambda,\varphi,t) \right) ds \cdot \Delta \boldsymbol{d}_{\kappa_q} \right] + \Delta\beta_R + \Delta\beta^S \end{aligned} \tag{5.79}$$

where $STEC_0 = \overline{STEC}\big|_{\boldsymbol{d}_0}$ is the initial $STEC$ value calculated from $\boldsymbol{d}_0$ and the initial values of receiver bias $\beta_R^0$ and satellite bias $\beta_0^S$, and $\Delta\beta_R$ and $\Delta\beta^S$ are the corresponding corrections. We assume now that GNSS ray paths with altogether $m$ receivers and $n$ satellites lie in the region under investigation, written as $STEC(R_i, S_j, t_{i,j})$. Herein the receiver positions are defined as $R_i$ $(i = 1,\ldots,m)$ and the satellite positions as $S_j$ $(j = 1,\ldots,n)$. It should be noted that it is not necessary that each receiver $i$ can observe all the $n$ satellite position. Therefore, the total number of ray paths, denoted as $N_2$, should not exceed $m \cdot n$. For simplicity, let $STEC_1,\ldots,STEC_{N_2}$ denote the $N_2$ distinct observations, with the corresponding observation errors $e_1,\ldots,e_{N_2}$. Similar to Eq. (5.72), the Gauss-Markov model can be established as

$$\boldsymbol{X}_2\,\Delta\boldsymbol{\beta}_2 = \boldsymbol{y}_2 + \boldsymbol{e}_2 \quad \text{with} \quad D(\boldsymbol{y}_2) = \sigma_2^2\,\boldsymbol{P}_2^{-1} \tag{5.80}$$

where

$$\boldsymbol{X}_2 = \Bigg( \underbrace{\begin{matrix} \frac{\partial STEC_1}{\partial \boldsymbol{d}_{\kappa_1}}\big|_{\boldsymbol{d}_{0,\kappa_1}} & \cdots & \frac{\partial STEC_1}{\partial \boldsymbol{d}_{\kappa_p}}\big|_{\boldsymbol{d}_{0,\kappa_p}} \\ \vdots & \ddots & \vdots \\ \frac{\partial STEC_{N_2}}{\partial \boldsymbol{d}_{\kappa_1}}\big|_{\boldsymbol{d}_{0,\kappa_1}} & \cdots & \frac{\partial STEC_{N_2}}{\partial \boldsymbol{d}_{\kappa_p}}\big|_{\boldsymbol{d}_{0,\kappa_p}} \end{matrix}}_{\boldsymbol{X}_{21}} \quad \underbrace{\begin{matrix} \frac{\partial STEC_1}{\partial \beta_{R_1}} & \cdots & \frac{\partial STEC_1}{\partial \beta_{R_r}} & \frac{\partial STEC_1}{\partial \beta^{S_1}} & \cdots & \frac{\partial STEC_1}{\partial \beta^{S_s}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial STEC_{N_2}}{\partial \beta_{R_1}} & \cdots & \frac{\partial STEC_{N_2}}{\partial \beta_{R_r}} & \frac{\partial STEC_{N_2}}{\partial \beta^{S_1}} & \cdots & \frac{\partial STEC_{N_2}}{\partial \beta^{S_s}} \end{matrix}}_{\boldsymbol{X}_{22}} \Bigg), \\$$

$$\Delta\boldsymbol{\beta}_2 = [\Delta\boldsymbol{d}^T, \Delta\boldsymbol{\beta}_{DCB}^T]^T \quad \text{with} \quad \Delta\boldsymbol{\beta}_{DCB} = [\Delta\beta_{R_1},\ldots,\Delta\beta_{R_r},\Delta\beta^{S_1},\ldots,\Delta\beta^{S_s}]^T,$$
$$\boldsymbol{y}_2 = [STEC_1 - \overline{STEC}_1\big|_{\boldsymbol{d}_0},\ldots,STEC_{N_2} - \overline{STEC}_{N_2}\big|_{\boldsymbol{d}_0}]^T, \quad \boldsymbol{e}_2 = [e_1,\ldots,e_{N_2}]^T.$$

The matrix $\boldsymbol{X}_{21}$ is the Jacobian matrix of the STEC observation vector with respect to the scaling coefficient vector, whereas the matrix $\boldsymbol{X}_{22}$ is the Jacobian matrix with respect to the DCBs. Both the receiver and satellite DCBs can be estimated as additional unknowns and daily constants can be assumed. $\Delta\boldsymbol{\beta}_{DCB}$ is the correction vector to be estimated, which is relative to the initial vector $\boldsymbol{\beta}_{DCB}^0 = [\beta_{R_1}^0,\ldots,\beta_{R_r}^0,\beta_0^{S_1},\ldots,\beta_0^{S_s}]^T$ collecting initial values of the biases. The stochastic model of $STEC$ observations is defined by the covariance matrix $D(\boldsymbol{y}_2)$ which includes the given positive definite $N_2 \times N_2$ weight matrix $\boldsymbol{P}_2$ and the unknown variance factor $\sigma_2^2$.

In order to avoid rank deficiency and separate receiver and satellite DCBs from each other, an additional constraint, e.g., the zero-mean condition, must be applied to the DCBs of all GPS satellites, following the IGS strategy [Schaer, 1999]. With the total number of GPS satellites $s'$, the constraint reads

$$\delta b^{S_1} + \ldots + \delta b^{S_{s'}} = 0\,. \tag{5.81}$$

According to Eq. (5.78), we have

$$\beta^{S_1} + \ldots + \beta^{S_{s'}} = 0\,. \tag{5.82}$$

Note, that it might happen that not all GPS satellites are observed over a region within a short time period, i.e., $s < s'$. Then the values of the DCBs of the satellite that are not in view are set as known and can be taken from IGS. Since DCBs are estimated together with the unknown coefficients of the key parameters, DCBs also have to be estimated iteratively. Introducing the decomposition of the DCBs of the satellites due to the linearization into Eq. (5.82) and assuming $s = s'$ give

$$\Delta\beta^1 + \ldots + \Delta\beta^{S_s} = -(\beta_0^1 + \ldots + \beta_0^{S_s}) = -\sum_{k=1}^{s} \beta_0^{S_k}\,. \tag{5.83}$$

Introducing the equality constraint (5.83) leads to the following normal equation

$$\begin{bmatrix} \boldsymbol{X}_2^T \boldsymbol{P}_2 \boldsymbol{X}_2 & \boldsymbol{h} \\ \boldsymbol{h}^T & 0 \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\beta}_2 \\ k \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_2^T \boldsymbol{P}_2\, \boldsymbol{y}_2 \\ -\sum\limits_{k=1}^{s} \beta_0^{S_k} \end{bmatrix} \tag{5.84}$$

based on Eq. (4.26) with

$$\boldsymbol{h} = [\underbrace{0, \ldots, 0}_{K}, \underbrace{0, \ldots, 0}_{r}, \underbrace{1, \ldots, 1}_{s}]^T \tag{5.85}$$

where $K$ denotes the total number of unknown B-spline coefficients and the total number of unknowns in the adjustment system is $u = K + r + s$.

**Stochastic model of STEC observations**

GPS observations have a high sampling rate, e.g., the IGS daily Receiver Independent Exchange Format (RINEX) files provide measurements of 30 s sampling rate, therefore the data should be highly correlated. In case that the correlations are not known, the amount of the data can be reduced to a lower sampling rate and the weight matrix could be set to $\boldsymbol{P}_2 = \mathbf{I}$, cf. Dettmering *et al.* [2011a] or Liang *et al.* [2015a,b].

### 5.3.2 Gauss integration along the ray-paths

As can be seen from Eq. (5.79), integrals have to be performed along GNSS ray paths from satellite to receiver, i.e., through the ionosphere and plasmasphere from around 50 km to 20,200 km height above the Earth's surface. This can be done with the help of numerical integration techniques (called quadrature), that is, the integrand is evaluated at a finite set of points (called integration knots) and a weighted sum of these values is used to approximate the integral. The Gauss quadrature is extremely accurate in most cases [Kendall, 1989]. It allows for choosing the weighting coefficients as well as the locations of the abscissa values at which the function is evaluated.

**Gauss–Legendre quadrature**

The original Gauss–Legendre formula is defined on the interval $[-1, 1]$ by [Kendall, 1989]

$$\int_{-1}^{1} l(x)\, dx \simeq \sum_{i=1}^{n} \omega_i\, l(x_i) \tag{5.86}$$

where $l(x)$ is the function to be integrated, and the integration nodes $x_i$ are equal to the zeros of the degree $n$ of the Legendre polynomial $P_n(x)$ on $[-1, 1]$. The coefficients $\omega_i$ are called the integration weights or quadrature weights, given by

$$\omega_i = \frac{2}{(1 - x_i^2)[P_n'(x_i)]^2} \,. \tag{5.87}$$

where $P_n'$ denotes the derivative of the Legendre polynomial.

When the integrals are defined on any finite interval $[a, b]$, the following linear transformation

$$\int_a^b l(x)\, dx = \frac{b-a}{2} \int_{-1}^{1} l\left(\frac{b-a}{2}x + \frac{a+b}{2}\right) dx \tag{5.88}$$

has to be performed. Thus, applying the Gauss quadrature to an integral on an interval $[a, b]$ yields

$$\int_a^b l(x)\, dx \simeq \frac{b-a}{2} \sum_{i=1}^{n} \omega_i \cdot l\left(\frac{b-a}{2}x_i + \frac{a+b}{2}\right) . \tag{5.89}$$

For the Gauss–Legendre nodes and weights of different degrees $n$ we refer e.g., to Kendall [1989] and Stroud and Secrest [1966]. According to Eq. (5.79), the electron density and the B-splines have to be computed for each evaluation point $x_i$ along the ray path. These two quantities have been both expressed in terms of spherical coordinates, but it is more convenient to use the Cartesian coordinate system to describe the positions of the evaluation points; please refer to Eq. (5.90) for the transformation from spherical coordinates to the ECEF coordinates.

**Transformation between Cartesian and spherical coordinates**
According to Eq. (5.79), the electron density and the B-splines have to be computed for each evaluation point $x_i$ along the ray path. These two quantities are both expressed in terms of spherical coordinates $(\lambda, \phi, h)$. However, it is more convenient to use the Cartesian coordinate system such as the Earth-Centered Earth-Fixed (ECEF) coordinate system to describe the positions of the evaluation points. The ECEF coordinate system is a right-handed Cartesian system $(x, y, z)$. Its origin is at the center of mass of the Earth, while the z-axis is along the spin axis of the Earth and pointing to the north pole; the x-axis is pointing to the mean Greenwich meridian; and the y-axis is orthogonal to the z- and x-axes following the right-hand rule. The transformation between spherical coordinates $(\lambda, \phi, r)$ and the Cartesian coordinates $(x, y, z)$ reads

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r\,\cos\varphi\,\cos\lambda \\ r\,\cos\varphi\,\sin\lambda \\ r\,\sin\varphi \end{pmatrix} \quad \text{and} \quad \begin{cases} \lambda = \arctan(y/x) \\ \varphi = \arctan(z/\sqrt{x^2+y^2}) \\ r = \sqrt{x^2+y^2+z^2} \end{cases} . \tag{5.90}$$

where $r$ is the radial distance of the point $(x, y, z)$, i.e., $r = R_E + h$ with $R_E \approx 6371\,\text{km}$ the Earth's mean radius and $h$ the height over the sphere.

### 5.3.3 Data combination

As introduced in Section 5.2.3, each B-spline function has a compact influence zone. The corresponding coefficient is computable when there is at least one profile lying in or one ray path penetrating the influence zone. However, even if ground-based GPS observations and IRO data

are combined, there are still data gaps in particular over the oceans, which will cause that the corresponding coefficients cannot be supported or only weakly be supported by the observations and the normal equation system is singular or at least ill-conditioned. Consequently, not all of the coefficients can be estimated. Solving such a problem, the corresponding columns of $\boldsymbol{X}_i$ $(i = 1, 2)$ in the Eqs. (5.72) and (5.80), which correspond to unsupported coefficients, may be eliminated to form reduced design matrices. Here we introduce prior information as introduced in Section 4.4 to stabilize the estimation process. In particular, the expectation vector $E(\boldsymbol{d}) = \boldsymbol{\mu}_d$ for the vector of the B-spline coefficients with the positive definite covariance matrix $D(\boldsymbol{d}) = \boldsymbol{\Sigma}_d$ of the unknown coefficient vector $\boldsymbol{d}$ are introduced as an additional observation technique by means of

$$\boldsymbol{d} = \boldsymbol{\mu}_d + \boldsymbol{e}_\mu \quad \text{with} \quad D(\boldsymbol{\mu}_d) = \sigma_\mu^2 \boldsymbol{P}_\mu^{-1} \tag{5.91}$$

following Eq. (4.41). Herein $\boldsymbol{e}_\mu$ denotes the error vector of the prior information, $\sigma_\mu^2$ is the unknown variance factor, and $\boldsymbol{P}_\mu = \boldsymbol{\Sigma}_d^{-1}$ is the given positive definite $K \times K$ weight matrix. The vector $\boldsymbol{\mu}_d$ can be approximated from the key parameters provided by/derived from a selected background model according to Eq. (5.69). It should be pointed that the choice of a background model is quite important, because the prior information is used to overcome data gaps and can only be improved in areas and at periods where observations are available. Moreover, the prior information may be different from the background model, since the accuracy of the representation of the background model depends on the B-spline levels. Note that the prior information is always referred to the set of the coefficients $\boldsymbol{d}$, and thus, not changing during the iterative procedure. Since the decomposition Eq. (5.68) is performed due to linearization, Eq. (5.91) can be reformulated as

$$\Delta\boldsymbol{d} = (\boldsymbol{\mu}_d - \boldsymbol{d}_0) + \boldsymbol{e}_\mu \quad \text{with} \quad D(\boldsymbol{\mu}_d - \boldsymbol{d}_0) = \sigma_\mu^2 \boldsymbol{P}_\mu^{-1}\,. \tag{5.92}$$

Let

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \Delta\boldsymbol{\beta}, \quad \boldsymbol{\beta}_0 = [\boldsymbol{d}_0^T, (\boldsymbol{\beta}_{DCB}^0)^T]^T, \quad \Delta\boldsymbol{\beta} = [\Delta\boldsymbol{d}^T, \Delta\boldsymbol{\beta}_{DCB}^T]^T, \tag{5.93}$$

and assume that the different observation techniques are uncorrelated, based on Eq. (4.46), the combination of the linear models (5.72), (5.80) and (5.92) yields the extended Gauss-Markov model

$$\begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{0}_{N_1,r+s} \\ \boldsymbol{X}_{21} & \boldsymbol{X}_{22} \\ \boldsymbol{I} & \boldsymbol{0}_{K,r+s} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{d} \\ \\ \Delta\boldsymbol{\beta}_{DCB} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{\mu}_d - \boldsymbol{d}_0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \boldsymbol{e}_\mu \end{bmatrix} \tag{5.94}$$

with the extended stochastic model of the observations $\bar{\boldsymbol{y}} = [\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, (\boldsymbol{\mu}_d - \boldsymbol{d}_0)^T]^T$

$$D(\bar{\boldsymbol{y}}) = \sigma_1^2 \boldsymbol{M}_1 + \sigma_2^2 \boldsymbol{M}_2 + \sigma_\mu^2 \boldsymbol{M}_\mu \tag{5.95}$$

where

$$\boldsymbol{M}_1 = \begin{bmatrix} \boldsymbol{P}_1^{-1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{M}_2 = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_2^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{M}_\mu = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_\mu^{-1} \end{bmatrix}.$$

Since no prior information is introduced for the DCBs, the DCB constraint introduced in the Eqs. (5.83) to (5.85) has to be considered to overcome rank deficiencies of the normal equation

system for Eq. (5.94), the unknown correction vector $\Delta\boldsymbol{\beta}_{it}$ at an iteration step $it$ can be solved by the extended normal equation system

$$\underbrace{\begin{bmatrix} \left(\sum_{i=1}^{2} \frac{1}{\sigma_i^2}\overline{\boldsymbol{X}}_i^T \boldsymbol{P}_i \overline{\boldsymbol{X}}_i + \frac{1}{\sigma_\mu^2}\overline{\boldsymbol{P}}_\mu\right) & \boldsymbol{h} \\ \boldsymbol{h}^T & 0 \end{bmatrix}}_{\overline{\boldsymbol{N}}} \begin{bmatrix} \Delta\widehat{\boldsymbol{\beta}}_{it} \\ k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{2} \frac{1}{\sigma_i^2}\overline{\boldsymbol{X}}_i^T \boldsymbol{P}_i \boldsymbol{y}_i + \frac{1}{\sigma_\mu^2}\overline{\boldsymbol{P}}_\mu(\overline{\boldsymbol{\mu}}_d - \boldsymbol{\beta}_{it,0}) \\ -\sum_{k=1}^{s} \beta_{it,0}^{S_k} \end{bmatrix} \tag{5.96}$$

where

$$\overline{\boldsymbol{X}}_1 = [\boldsymbol{X}_1 \quad \boldsymbol{0}_{N_1,r+s}], \quad \overline{\boldsymbol{X}}_2 = [\boldsymbol{X}_{21} \quad \boldsymbol{X}_{22}],$$
$$\overline{\boldsymbol{\mu}}_d = [\boldsymbol{\mu}_d^T \quad \boldsymbol{0}_{r+s,1}^T]^T, \quad \overline{\boldsymbol{P}}_\mu = \begin{bmatrix} \boldsymbol{P}_\mu & \boldsymbol{0}_{K,r+s} \\ \boldsymbol{0}_{r+s,K} & \boldsymbol{0}_{r+s,r+s} \end{bmatrix}.$$

As mentioned previously, the variance factors $\sigma_1^2$, $\sigma_2^2$ and $\sigma_\mu^2$ are chosen manually by experience or automatically by using VCE, see Section 4.5.

The initial vector $\boldsymbol{\beta}_{it,0}$ has to be updated during the iteration step through

$$\boldsymbol{\beta}_{it,0} = \boldsymbol{\beta}_0 + \sum_{it'=0}^{it-1} \Delta\widehat{\boldsymbol{\beta}}_{it'} \tag{5.97}$$

and thus the right-hand side of the linear model (5.92), i.e., $\boldsymbol{\mu}_d - \boldsymbol{d}_0$, has also to be updated. Then, the final estimation is

$$\widehat{\boldsymbol{\beta}}_{it_{max}} = \boldsymbol{\beta}_{it_{max},0} + \Delta\widehat{\boldsymbol{\beta}}_{it_{max}} = \boldsymbol{\beta}_0 + \sum_{it'=0}^{it_{max}} \Delta\widehat{\boldsymbol{\beta}}_{it'} \tag{5.98}$$

where $it_{max}$ signifies the total number of required iterations for convergence point.

The estimated covariance matrix of $\widehat{D}(\widehat{\boldsymbol{\beta}}_{it_{max}})$ can be extracted by the inverse of the matrix $\overline{\boldsymbol{N}}$ of normal equations in Eq. (5.96) at the last iteration step since the variance components of the observation techniques have been already included in $\overline{\boldsymbol{N}}$. The key parameters $\kappa_q(\lambda, \varphi, t)$ can then be computed everywhere within the 3-D volume $\Omega^3$ by introducing the corresponding B-spline coefficient vector $\boldsymbol{d}_{\kappa_q}$ from Eq. (5.98) into Eq. (5.69), and the variance and covariance information for the vector $\widehat{\boldsymbol{\kappa}}_q$, which collects the estimated $\widehat{\kappa}_q(\lambda_i, \varphi_i, t_i)$ at a set of points within $\Omega^3$, can be derived by applying the law of error propagation to Eq. (5.69) as

$$\widehat{D}\,(\widehat{\boldsymbol{\kappa}}_q) = \boldsymbol{Y}\,\widehat{D}(\widehat{\boldsymbol{d}}_{\kappa_q})\,\boldsymbol{Y}^T \tag{5.99}$$

where $\boldsymbol{Y}$ collects $\boldsymbol{\phi}_{\kappa_q}^T(\lambda, \varphi, t)$ at all the points $(\lambda_i, \varphi_i, t_i)$ and $\widehat{D}(\widehat{\boldsymbol{d}}_{\kappa_q})$ is extracted from $\widehat{D}(\widehat{\boldsymbol{\beta}}_{it_{max}})$. Starting with the correction vector $\Delta\boldsymbol{d} = \sum_{it'=0}^{it_{max}} \Delta\boldsymbol{d}_{it'}$ of the selected level $\boldsymbol{J}$, the pyramid algorithm 5.65 can be applied to obtain the MSR.

It is possible to introduce other space-geodetic observation techniques (e.g., satellite altimetry, VLBI) by stacking their observation equations and stochastic models into Eq. (5.95). However, it should be noted that there might exist offsets between the observation techniques, e.g., between GPS and satellite altimetry, see Brunini *et al.* [2005], Todorova [2008] or Dettmering *et al.* [2011a]. The inter-techniques biases have to be considered as additional unknowns by extending the columns of the design matrices in the combination procedure.

### 5.3.4 Procedure of model calculation

The main steps of the computation procedure for the discussed electron density model are shown in Fig. 5.8. The iterative algorithm is terminated when convergence is reached. The convergence criteria can be chosen so that the corrections of the coefficients become numerically small. In order to obtain stable and realistic solutions, some issues have to be considered. For example, the unknown parameters such as peak densities, peak heights and scale heights have very different magnitudes. The large differences of magnitude may cause that the magnitudes of the singular values of the matrix of normal equations are very different, and thus, produce a large condition number according to Eq. (4.40). Therefore, it is necessary to scale the parameters to assure a well-conditioned numerical estimation, see details below. Furthermore, the ionosphere key parameters have physical meanings and should be strictly positive, and some key parameters should fall within a certain interval, and thus inequality constraints might be necessary to be imposed in case that the estimated parameters are not realistic. An overview about the difficulties of such inverse problems can be found in, e.g., Garcia and Crespon [2008].

**Parameter scaling**

For the linear equation system in Eq. (4.5), parameter scaling can be done by right multiplication of the design matrix $\boldsymbol{X}$ by a weighting matrix $\boldsymbol{W}$ through [Siciliano and Khatib, 2008]

$$(\boldsymbol{X}\boldsymbol{W})\left(\boldsymbol{W}^{-1}\boldsymbol{\beta}\right) = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}} \tag{5.100}$$

where $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{W}$ and $\widetilde{\boldsymbol{\beta}} = \boldsymbol{W}^{-1}\boldsymbol{\beta}$ denote the scaled design matrix and parameter vector, respectively. There are many approaches to choose the scaling matrix $\boldsymbol{W}$ where the most common one is column scaling. If the matrix $\boldsymbol{X}$ is of full column rank, $\boldsymbol{W}$ can be defined as a diagonal matrix

$$\boldsymbol{W} = \mathrm{diag}\{w_1, \ldots, w_k\} \quad \text{with} \quad w_j = \|\boldsymbol{x}_j\|^{-1} \tag{5.101}$$

where $\boldsymbol{x}_j$ denotes the j-th column of $\boldsymbol{X}$ and $\|\cdot\|$ indicates the Euclidean norm. After the estimation of the new variable $\widetilde{\boldsymbol{\beta}}$ is obtained, the parameter vector $\widehat{\boldsymbol{\beta}}$ can be obtained through $\widehat{\boldsymbol{\beta}} = \boldsymbol{W}\widetilde{\boldsymbol{\beta}}$.

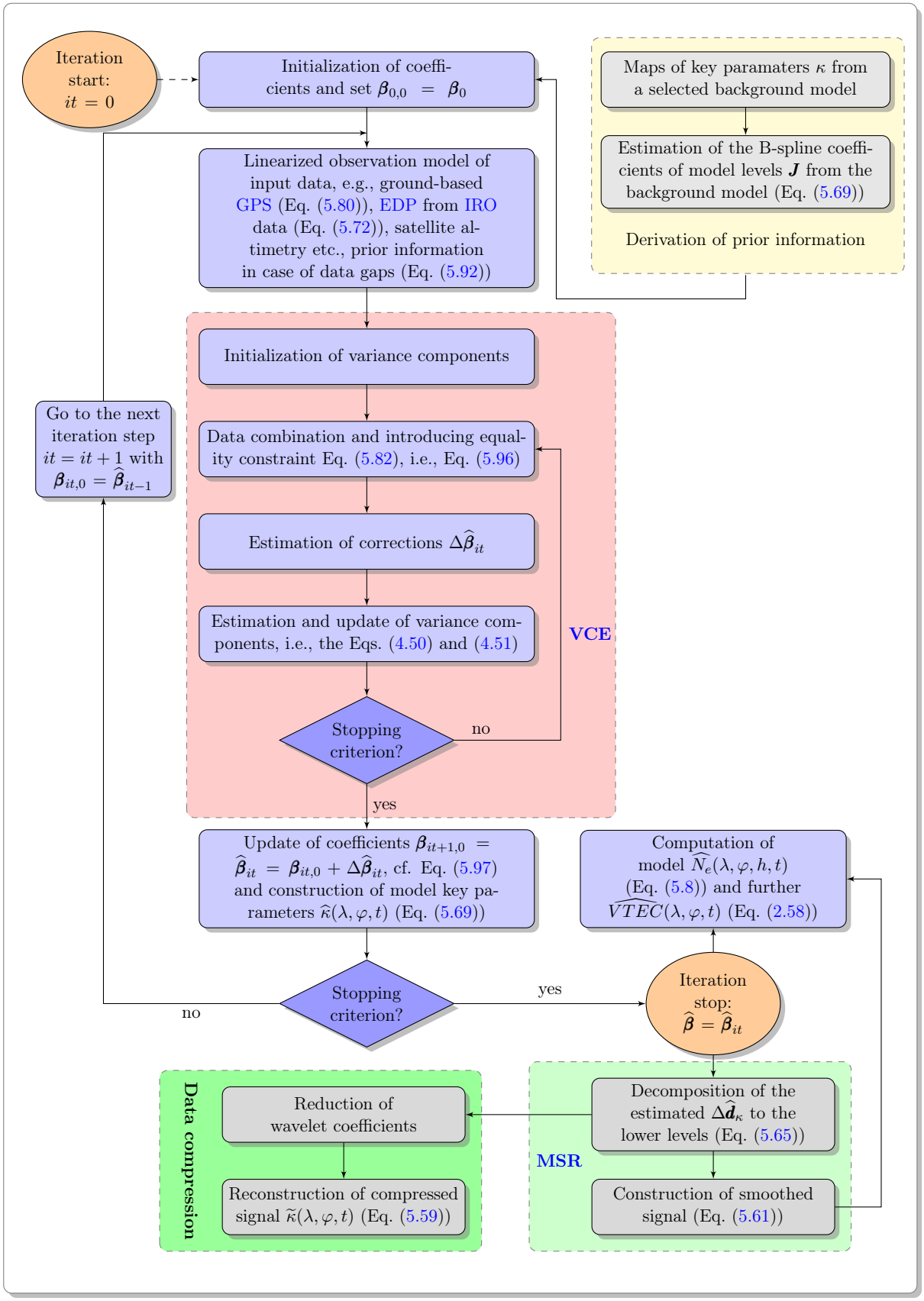



Figure 5.8: Flowchart of the main steps of calculating 4-D $N_e$ model.

# Chapter 6

# Numerical Analysis

This chapter is composed of two sections: the F2-layer modeling and the multi-layer modeling. The first section is a modified and extended version of the article *Regional modeling of ionospheric peak parameters using GNSS data – an update for IRI* [Liang *et al.*, 2015b]. The MSR and data compression follows the method described in the article *Combination of ground- and space-based GPS data for the determination of a multi-scale regional 4-D ionosphere model* [Liang *et al.*, 2015a].

## 6.1 F2-layer modeling

### 6.1.1 Introduction of a regional F2-Chapman/plasmasphere model

Since the F2 layer is the most important part of the ionosphere, the development of the profile function used in this thesis is firstly started with modeling the F2 layer and the plasmasphere, i.e., the first three terms of Eq. (5.8) are vanishing. Furthermore, the plasmaspheric term of Eq. (5.8) is adapted to allow the validity of the whole function over the full height range. Among the different profilers introduced in Section 5.1.1, the $\alpha$-Chapman function, i.e., Eq. (5.4) with $\overline{c} = 0.5$, is taken here to model the F2 layer, because the Chapman function is derived by the simplified aeronomic theory and most modelers have found that $\alpha$-Chapman function provides a closer match with observations [Bilitza, 2002]. The plasmspheric representation used here is based on the exponential function (2.63) and (5.9). The profile function (see Fig. 6.1) reads

$$N_e(h) \;=\; NmF2\,\exp\left[0.5\,(1-z-\exp(-z))\right] \;+\; N_{P0}\,\exp\left(-\frac{|h-hmF2|}{H_p}\right) \tag{6.1}$$

where $z = (h - hmF2)/HF2$. $N_{P0}$ is called the plasmaspheric basis density, and $H_p$ is the plasmaspheric scale height. Compared with the plasmaspheric term in Eq. (2.63) or Eq. (5.9), the exponential part in Eq. (6.1) is taking $hmF2$ as a base point, and a small $H_p$ value below $hmF2$ is introduced to ensure a continuous curve for the plasmaspheric part with a very small contribution below $hmF2$. It is worth noting that, even though only one constant scale height $HF2$ is considered in the Chapman layer for the bottomside and topside, a variation of the shape of the resultant profile will be indirectly considered by the superposed exponential decaying function. In the following, the three ionospheric parameters, written in a vector $\boldsymbol{\kappa}_{F2} := [NmF2, hmF2, HF2]^T$, will be handled, and the left two plasmaspheric parameters are

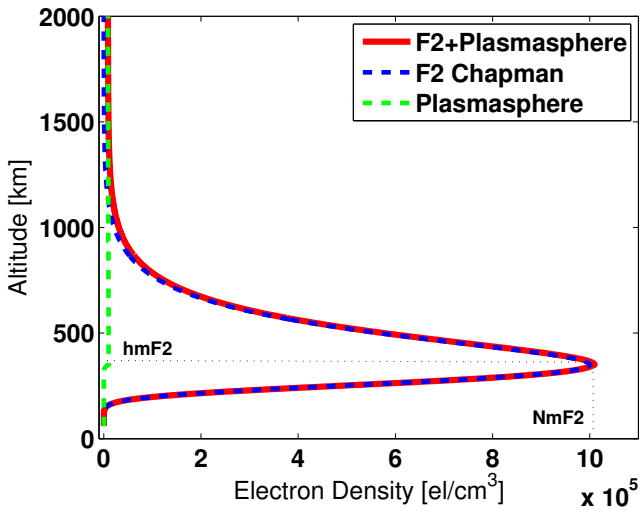



Figure 6.1: Vertical distribution of the electron density (Eq. (6.1)) with exemplary values $NmF2 = 1 \cdot 10^6\,\mathrm{el/cm^3}$, $hmF2 = 350\,\mathrm{km}$, $HF2 = 75\,\mathrm{km}$, and $N_{P0} = 1 \cdot 10^4\,\mathrm{el/cm^3}$.

assumed to be known. Following Jakowski [2005], $N_{P0}$ is assumed to be proportional to $NmF2$. $H_p$ is set to a physically plausible value of $10^4\,\mathrm{km}$ for $h \geq hmF2$ (10 km for $h < hmF2$).

To model the spatiotemporal variations, each key parameter $\kappa \in \boldsymbol{\kappa}_{F2}$ is represented regionally by a series expansion in terms of tensor products of B-splines according to Eq. (5.17). Following the Eqs. (5.71) and (5.79), the partial derivatives of Eq. (6.1) with respect to each key parameter $\kappa \in \boldsymbol{\kappa}_{F2}$ have to be calculated. They are obtained from

$$\frac{\partial N_e}{NmF2} = \exp\left[0.5\left(1 - z - \exp(-z)\right)\right] \tag{6.2}$$

$$\frac{\partial N_e}{hmF2} = \begin{cases} \frac{NmF2}{2HF2}\,\exp\left[0.5\left(1 - z - \exp(-z)\right)\right]\left(1 - \exp(-z)\right) + \frac{N_{P0}}{H_p}\,\exp\left(-\frac{|h - hm_i|}{H_p}\right) & \text{if } h \geq hmF2 \\ \frac{NmF2}{2HF2}\,\exp\left[0.5\left(1 - z - \exp(-z)\right)\right]\left(1 - \exp(-z)\right) - \frac{N_{P0}}{H_p}\,\exp\left(-\frac{|h - hm_i|}{H_p}\right) & \text{else} \end{cases} \tag{6.3}$$

$$\frac{\partial N_e}{HF2} = NmF2\,\exp\left[0.5\left(1 - z - \exp(-z)\right)\right]\,\frac{h - hmF2}{2\,(HF2)^2}\left(1 - \exp(-z)\right). \tag{6.4}$$

The general process of the model calculation follows Fig. 5.8; see also Limberger *et al.* [2013] and Liang *et al.* [2015a,b] for more details. In the following, numerical applications of the developed model will be given.

### 6.1.2 Study area and input data

The equatorial region is of primary interest to ionosphere scientists, since the ionosphere characteristics have large spatial gradients [Coco, 1991]. The study area is thus chosen in a South

American region, within [60°S, 30°N] and [250°E, 340°E] (a region of 90° × 90°), which includes the EIA. The developed model is tested for three selected days – a day under high solar activity (July 1, 2012, solar $F10.7$ flux[1] = 137.9, $Kp$[2] = 3.3), a day under moderate solar activity (July 16, 2011, solar $F10.7$ flux = 96.9, $Kp$ = 1.0), and a day under low solar activity (July 16, 2008, solar $F10.7$ flux = 66.7, $Kp$ = 1.3), respectively. Two types of GNSS data, namely, ground-based two-frequency GPS observations of STEC and EDPs retrieved from ionospheric GPS RO measurements acquired by the F3/C mission, are used.

**Ground-based GPS data**

The GPS observables are provided by the regional densification network Sistema de Referencia Geocéntrico para las Américas (SIRGAS)[3]. It comprises about 350 continuously operating GNSS stations distributed along Latin America and the Caribbean [Sánchez *et al.*, 2013]. As mentioned previously, leveled phase data after data screening are used. Therefore DCBs have to be considered in the GPS ionospheric observation (see Eq. (5.77)) and are estimated as a by-product with daily constants in this study. Additionally, a zero-mean condition (cf. Eq. (5.82)) is applied to the observed satellite DCBs during one day as explained previously. GPS data are available in a 30-second sampling rate, which are highly correlated. Therefore, the amount of GPS data is reduced using a lower time sampling rate. The weighting matrix, i.e., $\boldsymbol{P}_2$ in Eq. (5.80) is thus set to the identity matrix.

**EDPs from ionospheric GPS RO data**

The EDPs retrieved from RO data of the F3/C mission are used and kindly provided by the CSRSR of NCU in Taiwan. Their retrieval technique is an improved Abel inversion which considers horizontal gradients of the electron density distribution; for details see e.g., Tsai *et al.* [2009]. A data screening is performed in a preprocessing step, in order that only profiles with reasonable shapes around the F2 peak are used as input. In particular, profiles with $hmF2$ values locating in a physically reasonable height range of [140 km, 500 km] and with data availability around $hmF2$ within ±50 km are used. In addition, profiles dominated by high noises are considered as outliers and removed. Since the electron densities are retrieved from the GPS observations which have a high sampling rate during an occultation event, the electron densities are densely distributed along the profile of about 2 km, and height-dependent correlations are assumed to exist due to the retrieval algorithm. Let $N_{e,i}$ and $N_{e,j}$ denote the electron densities at the height $h_i$ and $h_j$, respectively, the correlation between the two individual electron densities $N_{e,i}$ and $N_{e,j}$ are approximated by the simple and commonly used positive definite Hirvonen's covariance function (e.g., Moritz, 1976, and references therein; Koch, 1999)

$$\widehat{\sigma}_{i,j} = \widehat{\sigma}(\Delta h) = \frac{\widehat{\sigma}(0)}{1 + (\Delta h/a)^2} \quad \text{with} \quad \Delta h = |h_j - h_i| \tag{6.5}$$

where $\widehat{\sigma}(0) = \widehat{\sigma}_{ii} = \widehat{\sigma}_{jj}$, namely, the standard deviations of the electron densities along the profile are assumed to be equal. The constant $a$ is the correlation length, which is set to the value 10 such that the correlation of the electrons with a distance of 200 km is close to zero (see Fig. 6.2).

[1] The solar $F10.7$ flux and the $Kp$ index are obtainable from the Goddard Space Flight Center/Space Physics Data Facility (GSFC/SPDF) OMNIWeb interface (`http://omniweb.gsfc.nasa.gov`).

[2] $Kp$ index is a global geomagnetic storm index, which is calculated by an average of K indices from a network of geomagnetic observatories. $Kp$ index ranges from 0 to 9 where $0-1$ means quiet geomagnetic conditions and $Kp \geq 4$ refers to disturbed geomagnetic conditions [Bartels *et al.*, 1939].

[3] `www.sirgas.org`.

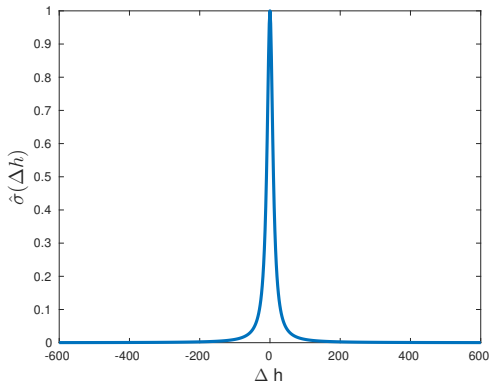



Figure 6.2: Covariance function with the correlation length $a = 10$ and $\widehat{\sigma}(0) = 1$.

Figure 6.3 shows the distribution of input data exemplarily for July 1, 2012, where the gray triangles depict the positions of the GPS stations; the red dots indicate the locations of EDPs (after data screening) labeled with the F3/C satellite identifier C# and the observed time in UT (hour:minute:second).

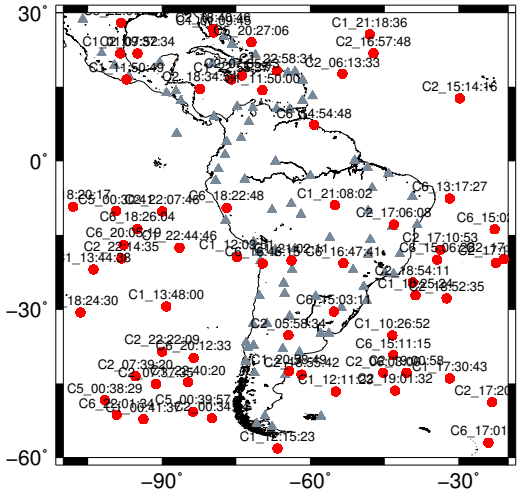



Figure 6.3: Data distribution within the study area on July 1, 2012.

**IRI – Background model**

As explained previously, the prior information is required to overcome data gaps. In this work, IRI-2012 is taken as the background model to derive the prior information $\boldsymbol{\mu}_d$ of the coefficients. For this purpose, the B-spline levels in Eq. (5.17) have to be specified. According to Eq. (5.27), $J_1 = J_2 = J_3 = 3$ is chosen here, for all three key parameters in $\boldsymbol{\kappa}_{F2}$. In this manner, the model resolution is adapted to the spatiotemporal distribution of the combined two types of GNSS data during one day. Based on Eq. (5.23), $\boldsymbol{\mu}_d$ is composed of $3 \cdot K_{J_1} \cdot K_{J_2} \cdot K_{J_3} = 3 \cdot (2^3 + 2)^3 = 3000$ B-spline coefficients with 1000 for each parameter $\kappa_q$ ($q = 1 :- NmF2, 2 :- hmF2, 3 :- HF2$). To derive $\boldsymbol{\mu}_d$, time-varying maps of $\boldsymbol{\kappa}_{F2}$ are required in the study area. $NmF2$ and $hmF2$ can be directly predicted by IRI, whereas $HF2$ can be indirectly obtained from IRI. According to Jayachandran *et al.* [2004, and references therein], the relation between the slab thickness $\tau_s$ and

the scale height $HF2$ of an $\alpha$-Chapman layer of the ionosphere

$$\tau_s = 4.13 \cdot HF2 \tag{6.6}$$

holds for an $\alpha$-Chapman layer, where $\tau_s$[1] is defined as the ratio [Davies, 1990]

$$\tau_s = \frac{VTEC}{NmF2} \tag{6.7}$$

which represents the equivalent slab thickness of the ionosphere having a constant uniform electron density equal to $NmF2$. $\tau_s$ is a significant parameter that contains information about the neutral temperature and can be related directly to the scale height for an assumed EDP [Jayachandran *et al.*, 2004; Jin *et al.*, 2007]. Therefore, the relation

$$HF2 = \frac{VTEC}{4.13 \cdot NmF2} \tag{6.8}$$

can easily be obtained, namely, $HF2$ can be indirectly derived as IRI predicts VTEC values. $\boldsymbol{\kappa}_{F2}$ is stored on a 3-D grid with a resolution of $2.5^\circ \times 2.5^\circ \times 6\,\text{min}$ with respect to $\lambda$, $\varphi$ and $t$. Therefore, totally $(90/2.5+1)\times(90/2.5+1)\times(24\cdot 60/6+1) = 329{,}929$ grid points are used as the size of the study area is $90^\circ(\lambda) \times 90^\circ(\varphi) \times 24\,\text{h}(t)$. Let $\boldsymbol{\kappa}_q^{IRI}$ be the vector of a certain parameter $\kappa_q$ which collects the corresponding values from background model of all grid points, $\boldsymbol{e}_{\kappa_q^{IRI}}$ be the error vector including errors resulting from, e.g., modeling approximations, and $\boldsymbol{\mu}_{d_{\kappa q}}$ be the vector of the prior information of the 1000 coefficients of $\kappa_q$, we can establish the equation

$$\boldsymbol{\Phi}\,\boldsymbol{\mu}_{d_{\kappa q}} = \boldsymbol{\kappa}_q^{IRI} + \boldsymbol{e}_{\kappa_q^{IRI}} \tag{6.9}$$

based on Eq. (5.17) where

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_{0,0,0}^{J_1,J_2,J_3}(\lambda_1,\varphi_1,t_1) & \cdots & \phi_{K_{J_1}-1,K_{J_2}-1,K_{J_3}-1}^{J_1,J_2,J_3}(\lambda_1,\varphi_1,t_1) \\ \vdots & \ddots & \vdots \\ \phi_{0,0,0}^{J_1,J_2,J_3}(\lambda_n,\varphi_n,t_n) & \cdots & \phi_{K_{J_1}-1,K_{J_2}-1,K_{J_3}-1}^{J_1,J_2,J_3}(\lambda_n,\varphi_n,t_n) \end{bmatrix}. \tag{6.10}$$

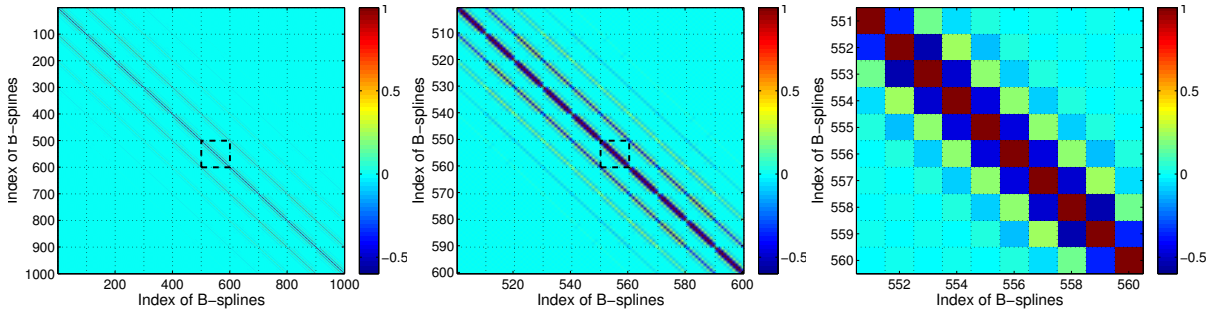
Note that the selected spatiotemporal region $[250^\circ, 340^\circ] \times [-60^\circ, 30^\circ] \times [0\,\text{h}, 24\,\text{h}]$ has to be transformed to the unit cube, in order to calculate $\boldsymbol{\Phi}$. Then, $\boldsymbol{\mu}_{d_{\kappa q}}$ can be estimated by the method of least squares through

$$\widehat{\boldsymbol{\mu}}_{d_{\kappa q}} = (\boldsymbol{\Phi}^T\,\boldsymbol{\Phi})^{-1}\,\boldsymbol{\Phi}^T\,\boldsymbol{\kappa}_q^{IRI}\,. \tag{6.11}$$

Note, the prior information generally differs from the background model, since the accuracy of the representation of the background model depends on the selected B-spline levels. If Eq. (6.11) is applied to all three parameters of $\boldsymbol{\kappa}_{F2}$, the vector $\widehat{\boldsymbol{\mu}}_d = [\widehat{\boldsymbol{\mu}}_{d_{\kappa 1}}^T, \widehat{\boldsymbol{\mu}}_{d_{\kappa 2}}^T, \widehat{\boldsymbol{\mu}}_{d_{\kappa 3}}^T]^T$ is obtained.

It can be expected from the characteristics of the B-splines that neighboring overlapped B-splines coefficients are correlated, see Fig. 5.3. The corresponding correlation matrix of the B-splines can be calculated according to the Eqs. (4.18), (4.21) and (4.22). Since the correlation matrix of the B-splines for a certain $\kappa_q$ depends only on the covariance matrix $(\boldsymbol{\Phi}^T\,\boldsymbol{\Phi})^{-1}$, which is determined by positions and times of the grid points and the B-spline levels, the correlations of the B-splines for a certain $\kappa_q$ do not depend on $\kappa_q$ itself. Figure 6.4(a) shows the correlation matrix of the 1000 B-spline coefficients, which are arranged as follows: firstly, the shift $k_1$ of the B-splines in the longitude dimension, i.e., $\phi_{J_1;k_1}(\lambda)$ (see Eq. (5.18)), is varied, then followed by the shift $k_2$

[1] A typical value of the slab thickness is 300 km [Davies, 1990].




(a) (b) (c)

Figure 6.4: Correlation matrix of the 1000 B-spline coefficients $d_{k_1,k_2,k_3}^{J_1,J_2,J_3}$ of the levels $J_1 = J_2 = J_3 = 3$ (a) with the black box ($100 \times 100$ coefficients) zoomed in (b), and the black box ($10 \times 10$ coefficients) of (b) zoomed in (c). The scales of the three color bars are the same.

of $\phi_{J_2;k_2}(\varphi)$, finally $k_3$ of $\phi_{J_3;k_3}(t)$ is changed, i.e., the shifts of the coefficients are ordered by

$$
\begin{aligned}
1-100 \text{ coefficients} \begin{cases} k_1=0,k_2=0,k_3=0 \rightarrow & \cdots k_1=9,k_2=0,k_3=0 \rightarrow \\ k_1=0,k_2=1,k_3=0 \rightarrow & \cdots k_1=9,k_2=1,k_3=0 \rightarrow \\ & \vdots \\ k_1=0,k_2=9,k_3=0 \rightarrow & \cdots k_1=9,k_2=9,k_3=0 \rightarrow \end{cases} \\
101-1000 \text{ coefficients} \begin{cases} k_1=0,k_2=0,k_3=1 \rightarrow & \cdots k_1=9,k_2=9,k_3=1 \rightarrow \\ k_1=0,k_2=0,k_3=9 \rightarrow & \cdots k_1=9,k_2=9,k_3=9\,. \end{cases}
\end{aligned}
$$

Therefore, the ten blocks along the diagonal direction of Fig. 6.4(a) from left to right represent the changes of the B-splines in the time dimension, corresponding to $k_3 = 0, \ldots, 9$. To have a closer look at a certain block, e.g., the $100 \times 100$ coefficients in the black box of Fig. 6.4(a), it is zoomed in Fig. 6.4(b). Again, the ten blocks along the diagonal direction from left to right indicate the variations of the shift $k_2$. The $10 \times 10$ coefficients in the black box of Fig. 6.4(b) is zoomed in Fig. 6.4(c), where the ten elements along the diagonal direction refer to the changes of $k_1$. As can be seen, for a certain B-spline, the correlation structures are visible between it and its three adjacent B-splines. In particular, a negative correlation of about $-0.5$ appears between the nearest neighboring B-splines, which means, an increase of the value of a specific B-spline will lead to an decrease of the values of the nearest neighbors. This can be expected from the property of the B-splines introduced in Section 5.2.3, that at most three B-splines are non-zero on any interval between two neighboring knots and their sum is equal to one. A positive correlation of approximately 0.2 exists between the second nearest neighbors, which is followed by the correlation of about $-0.1$ between the third nearest neighbors. The derived stochastic information is then introduced into the weight matrix $\boldsymbol{P}_{\mu_d}$ for model calculation, however, correlations between the coefficients of the different parameters are not incorporated here, i.e.,

$$
\boldsymbol{P}_{\mu_d} = \begin{bmatrix} \boldsymbol{P}_{\mu,\,d_{NmF2}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_{\mu,\,d_{hmF2}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{\mu,\,d_{HF2}} \end{bmatrix} \tag{6.12}
$$

where the assumptions $\boldsymbol{P}_{\mu,\,d_{NmF2}} = \boldsymbol{P}_{\mu,\,d_{hmF2}} = \boldsymbol{P}_{\mu,\,d_{HF2}}$ holds for the weight matrices of the

three parameters.

### 6.1.3 Results of the F2 layer parameters and discussions

Since the establishment of the adjustment model considering $STEC$ observations requires a huge number of integration, the calculation is very time-consuming. According to Stankov *et al.* [2003], the electron densities above approximately 2000 km have little contribution to the integrated electron content. In order to integrate efficiently, the numerical integration is performed from 80 km to 2000 km, as electron content below 80 km is also negligible. Three "integration layers" with layer-dependent step sizes are introduced to consider the different electron density distribution, see Fig. 6.5. In particular, the F2 layer has the maximum electron density values

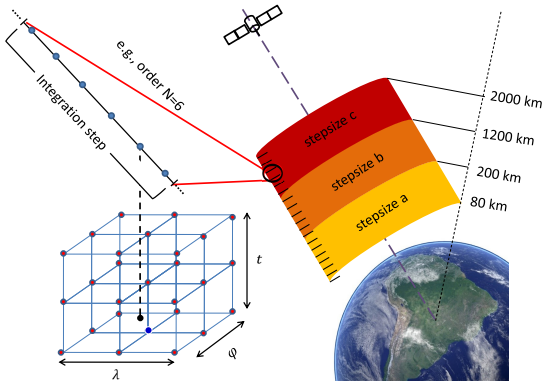



Figure 6.5: Illustration of the developed integration procedure using the Gauss-Legendre quadrature method [Liang *et al.*, 2015b]. Each key parameter is stored as a 3-D grid with a resolution of 2.5° × 2.5° × 6 min. Interpolation should be performed based on the 3-D grid, in order to calculate electron density values at the evaluation points (in blue) along the ray path.

and the largest variations, and thus the smallest step size is selected for the "integration layer" covering the F2 layer. The selection of the steps sizes is in consideration of computation time and accuracy. The specific height boundaries of the introduced "integration layers" and the step sizes used for integration are shown in Table 6.1. Within each step size (cf. the interval in the

Table 6.1: Introduced "integration layers" and layer-dependent step sizes for the numerical integration of the electron densities along a certain ray path.

| "Integration layer" | Height range [km] | Step size [km] |
|---|---|---|
| "Integration layer" 1 | $[80, 200]$ | 50 |
| "Integration layer" 2 | $(200, 1200]$ | 40 |
| "Integration layer" 3 | $(1200, 2000]$ | 100 |

black circle of Fig. 6.5), the Gauss–Legendre quadrature (Eq. (5.89)) is performed. The number of nodes or the degree $n$ is set to 6 for the same consideration of the step sizes. To carry out model computations, each key parameter $\kappa_q \in \boldsymbol{\kappa}_{F2} = [NmF2, hmF2, HF2]^T$ is stored in a 3-D grid of resolution 2.5° × 2.5° × 6 min with respect to $\lambda$, $\varphi$ and $t$, respectively (see Fig. 6.5). In

this way, the electron density values at the integration nodes are calculable using an interpolation algorithm.

Since linearization is performed, initial values of the unknowns are required for the iterative algorithm. The initial values of the coefficients can be derived from an external model, e.g., IRI. Note, that the external model used to estimate the initial condition needs not be the same as the background model for deriving the prior information.

To estimate $\boldsymbol{\kappa}_{F2}$, a sequential estimation similar to Alizadeh [2013] is performed in this thesis, in order to increase the stability of the linearized model, namely

- in the first step, only $NmF2$ is estimated by assuming $hmF2$ and $HF2$ as known,
- in the second step, the estimated $NmF2$ values are used to estimate $hmF2$ and $HF2$ jointly.

The parameter estimation follows Eq. (5.96), where the unknown variance components $\sigma_1^2$ of F3/C observations and $\sigma_2^2$ of GPS observations are determined by VCE introduced in Section 4.5. We introduce different variance components $\sigma^2_{\mu,\,NmF2}$, $\sigma^2_{\mu,\,hmF2}$ and $\sigma^2_{\mu,\,HF2}$ for the prior information of the three key parameters. As introduced previously at the end of Section 4.5, a reliable estimation of the variance component of the prior information cannot be obtained if the residual vector in Eq. (4.51) has no random character. This is the case when IRI is used as a given background model. Thus, the three different variance components for the prior information are fixed to empirical values. After the unknowns are estimated, $\widehat{\boldsymbol{\kappa}}_{F2}$ can be constructed everywhere over the study area during the whole day. Figure 6.6 shows exemplarily the estimated key parameters (bottom) for $\widehat{NmF2}$ (left), $\widehat{hmF2}$ (mid) and $\widehat{HF2}$ (right), which are constructed on the defined 3-D grid (cf. Fig. 6.5) at 16 UT on July 1, 2012. The used input data as shown in Fig. 6.3 is also illustrated in the figures, where white dots instead red ones are used to denote locations of EDPs. The corresponding key parameters derived from IRI-2012 are shown in the top panels. Comparing the top and the bottom maps, it can clearly be seen that the estimated signals describe finer structures than IRI, and the differences to IRI appear mainly over the continent and partly over the ocean where observations are available. Note that, as introduced previously, each B-spline has an influence zone, and therefore a single observation can influence only its surrounding area in both, the spatial and temporal domains. The size of the influence area is based on the resolution of the model according to the selected B-spline levels. For the selected levels $J_1 = J_2 = J_3 = 3$, the resolution of the model is about $10^\circ \times 10^\circ \times 2.7\,\mathrm{h}$. Therefore, the two profiles at about 17 h 1 min and 17 h 20 min at the bottom right corner can affect the maps at 16 UT. It can also be seen that, for regions where no observations are available, especially over the ocean areas, the developed model depends on the quality of IRI. Taking a closer look to the $NmF2$ maps, we can see that the estimated $\widehat{NmF2}$ values seem to smooth out the EIA structure, namely, the estimated $\widehat{NmF2}$ yields lower values than IRI for the double EIA peaks and higher values for the trough between the two peaks. This phenomenon is consistent with the studies by, e.g., Yue *et al.* [2012]. According to Yue *et al.* [2012, and references therein], the Abel retrieval from F3/C RO measurements can tend to smooth out the EIA structure as a result of the assumption of spherical symmetry. Azpilicueta *et al.* [2015] estimated the F2 peak parameters using EDPs derived from F3/C, and they also found that the developed model gives systematically higher values than IRI for the valleys between the two EIA peaks. The estimated standard deviations of the parameters are computed according to Eq. (5.99) and shown in Fig. 6.7. It can be seen that the precision could only be improved over regions where data are available. In particular, the precisions of about $1.3 \cdot 10^4\,\mathrm{el/cm^3}$ for $\widehat{NmF2}$, 2.9 km for

$\widehat{hmF2}$ and 0.9 km for $\widehat{HF2}$ are reached.

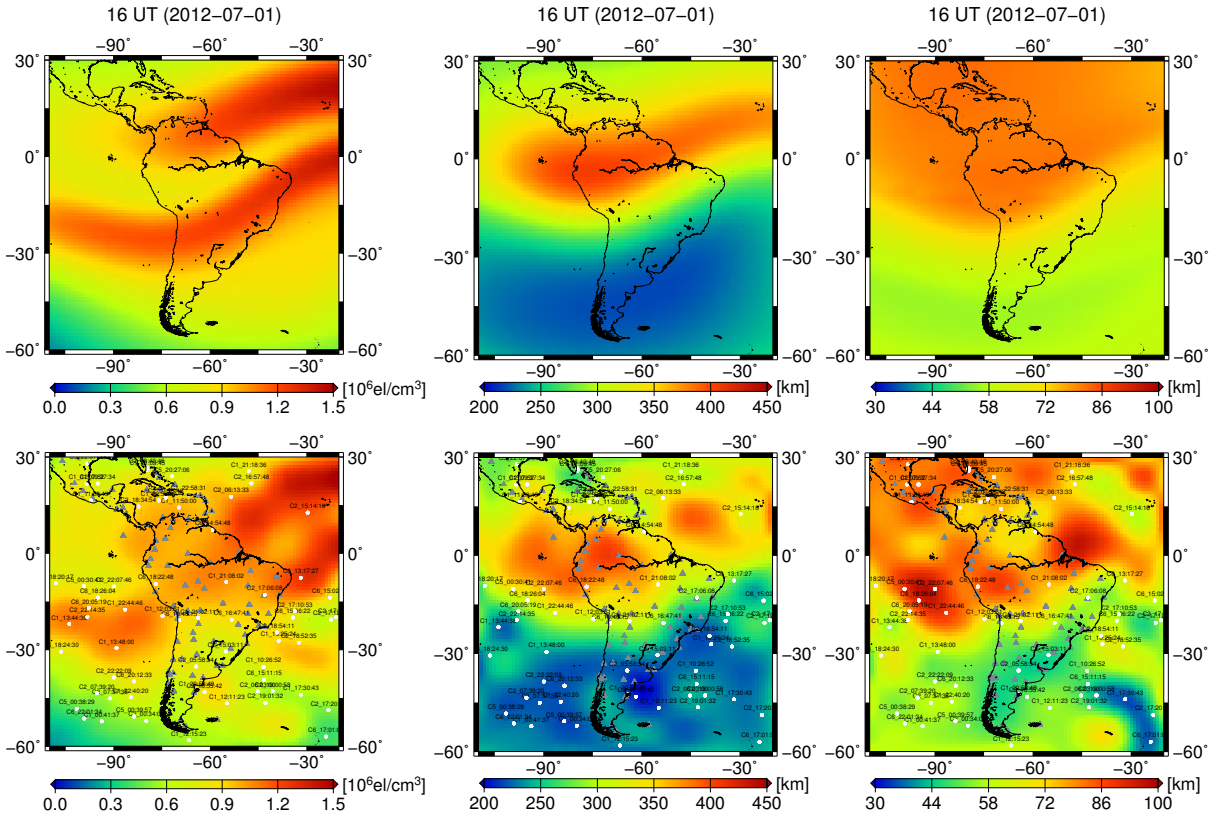



Figure 6.6: IRI-2012 $NmF2$ (left), $hmF2$ (mid) and $HF2$ (right) at 16 UT on July 1, 2012 ($1^{st}$ row); estimated final parameters $\widehat{NmF2}$ (left), $\widehat{hmF2}$ (mid) and $\widehat{HF2}$ (right) ($2^{nd}$ row).

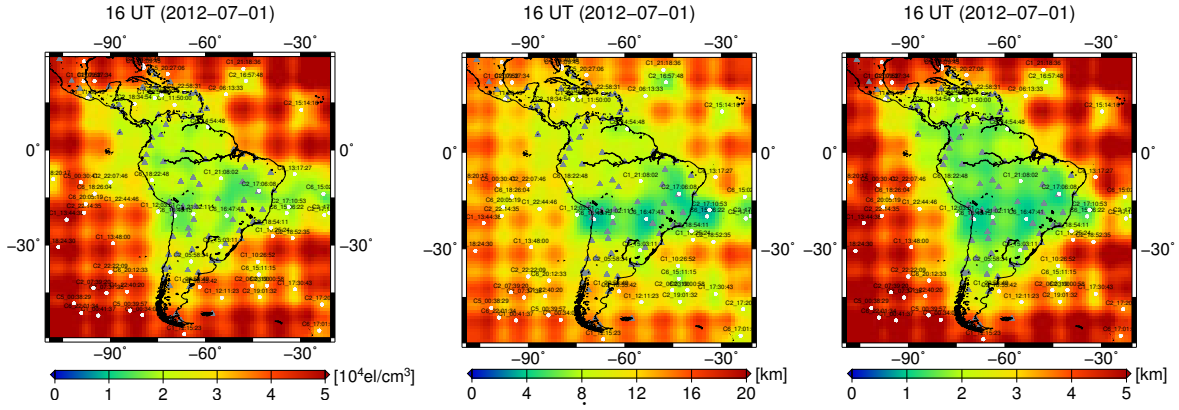



Figure 6.7: Estimated standard deviations of $\widehat{NmF2}$ (left), $\widehat{hmF2}$ (mid) and $\widehat{HF2}$ (right).

In the following, comparisons between the developed model and other data sources will be carried on. For the purpose of quantifying the comparisons, some useful metrics are defined now. Let $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ be two vectors of the same size, where $\boldsymbol{\beta}$ is the model vector and $\boldsymbol{\beta}^*$ is treated as the reference vector. Furthermore, $\bar{\cdot}$ and $\tilde{\cdot}$ are used to denote the operations for taking the mean value and the standard deviation of a vector. We define the following statistical quantities:

- mean difference: $\overline{\boldsymbol{\beta}-\boldsymbol{\beta}^*}$,
- standard deviation of difference: $\widetilde{\boldsymbol{\beta}-\boldsymbol{\beta}^*}$,
- mean of relative difference: $\overline{(\frac{\boldsymbol{\beta}-\boldsymbol{\beta}^*}{\boldsymbol{\beta}^*})}$,
- standard deviation of relative difference: $\widetilde{(\frac{\boldsymbol{\beta}-\boldsymbol{\beta}^*}{\boldsymbol{\beta}^*})}$,
- correlation coefficient [Press *et al.*, 2007]: $\frac{\sum_i(\beta_i-\overline{\boldsymbol{\beta}})(\beta_i^*-\overline{\boldsymbol{\beta}^*})}{\sqrt{\sum_i(\beta_i-\overline{\boldsymbol{\beta}})^2}\sqrt{\sum_i(\beta_i^*-\overline{\boldsymbol{\beta}^*})^2}}$.

Note that these definitions are always referred to a type of variable, which can be one of the three key parameters or VTEC. Furthermore, when we talk about differences, there is always a reference model as $\boldsymbol{\beta}^*$ depending on the place and context in this thesis.

The estimated key parameters are compared with the reference values predicted by IRI at 16 UT, i.e., for a certain key parameter, $\boldsymbol{\beta}$ is the vector collecting estimated model values of all grid points over the study area at 16 UT, while $\boldsymbol{\beta}^*$ is the vector collecting the corresponding values from IRI. The statistical analyses show that there is no significant difference between the developed model and IRI. For example, the differences of $NmF2$ range from $-4.3 \cdot 10^5$ el/cm$^3$ to $2.4 \cdot 10^5$ el/cm$^3$, with a mean value of $-3.1 \cdot 10^4$ el/cm$^3$ ($-2.6\%$, mean of relative difference) and a standard deviation of $1.2 \cdot 10^5$ el/cm$^3$ (12.6%, standard deviation of relative difference). The differences of $hmF2$ range from $-58.7$ km to 32.6 km, with a mean value of $-2.5$ km ($-0.5\%$, mean of relative difference) and a standard deviation of 14.8 km (4.9%, standard deviation of relative difference). Then the statistical quantities of deviations between the developed model and IRI are also calculated considering all the constructed grid points during the three selected days, indicating no significant difference as well. The conclusions are compatible with the studies by e.g., Azpilicueta *et al.* [2015]: they concluded that there is no significant systematic bias on both $NmF2$ and $hmF2$ between IRI and their La Plata Ionospheric Model (LPIM) [Brunini *et al.*, 2013a] obtained from F3/C EDPs. Furthermore, the ranges of the deviations of the two parameters between the developed model and IRI are comparable with the ones shown by Azpilicueta *et al.* [2015]. Bilitza *et al.* [2012] have also reported the discrepancy of IRI $hmF2$ with two Brazilian ionosonde stations in the low latitude region, and have revealed the overestimation of the IRI $hmF2$ at Sao Luis (near the magnetic equator, geographical coordinate: $2.3°$S, $316°$E)) by about 40 km (cf. Fig. 6.8(B)(a)). It should be kept in mind that IRI is a climatological model which provides monthly averages based on a large volume of past ground and space data. Since IRI-2012 has currently not assimilated any space-geodetic observation, the modeling approach can serve to update the parameters of IRI.

### 6.1.4 MSR and data compression application

As introduced previously, B-splines can construct wavelet functions which generate a MSR. Based on the vectors $\Delta\widehat{\boldsymbol{d}}_{NmF2}$, $\Delta\widehat{\boldsymbol{d}}_{hmF2}$ and $\Delta\widehat{\boldsymbol{d}}_{HF2}$ (cf. Eq. (5.68)) of the estimated coefficient corrections, a MSR is performed. Figure 6.8 presents a graphical demonstration of the MSR exemplarily applied to the estimated $\Delta\widehat{NmF2}$ (Fig. 6.8(A)(a)), $\Delta\widehat{hmF2}$ (Fig. 6.8(B)(a)), and $\Delta\widehat{HF2}$ (Fig. 6.8(C)(a)) values of the highest levels $J_1 = J_2 = J_3 = 3$ at 16 UT on July 1, 2012. The ordering of the panels of each subfigure is the same as the arrangement of the coefficient vectors of the pyramid algorithm explained in Fig. 5.5 (1-D case there, but 3-D case

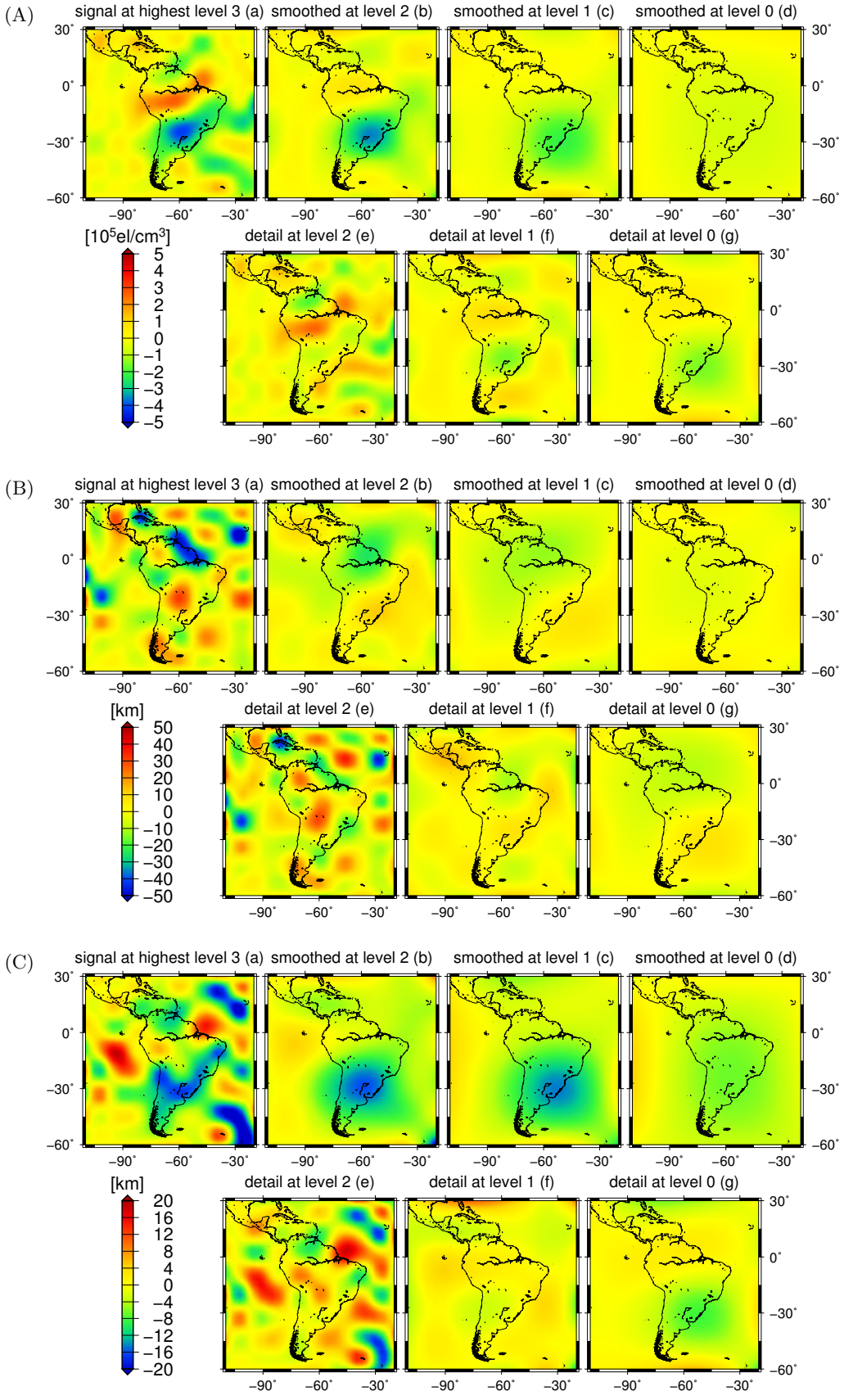



Figure 6.8: MSR of the estimated values $\Delta\widehat{NmF2}$ (A), $\Delta\widehat{hmF2}$ (B) and $\Delta\widehat{HF2}$ (C) at 16 UT on July 1, 2012: low-pass filtered smoothed signals (top: from left to right), estimated band-pass filtered detail signals (bottom: from left to right). Note that the level refers to all longitude, latitude and time dimensions.

here). In particular, the panels (b)-(d) depict the low-pass filtered (smoothed) signals, which are computed based on the Eqs. (5.42) and (5.48a). The band-pass filtered detail signals are shown in the panels (e)-(g), and they are computed according to the Eqs. (5.44) and (5.48b). The detail signals contain the information of the difference between the two adjacent smoothed signals, i.e., $(e) = (a) - (b)$, $(f) = (b) - (c)$, and $(g) = (c) - (d)$. Consequently, it can easily be deduced that $(d) + (e) + (f) + (g) = (a)$, namely, the sum of the smoothed signal at the lowest level (panel (d)) and the three detail signals (panels (e)-(g)) yields the signal at the highest level (panel (a)), which reflects the principle of the MSR. It can be seen from the figure that the structures of both smoothed and detail signals become gradually coarser with decreasing levels.

The number $K_{\boldsymbol{J}}$ of scaling coefficients at different levels $\boldsymbol{J}$, and the total number $L_{\boldsymbol{J}}$ of wavelet coefficients as well as the number $L^n_{\boldsymbol{J}}$ $(n = 1, \ldots, 7)$ of wavelet coefficients of different subbands (see Eq. (5.56)), are listed in Table 6.2. It is clear from the table that the number of both scaling and wavelet coefficients decrease with decreasing levels, namely, less coefficients are required to describe signals with coarser structures. Furthermore, the number of the scaling coefficients of level $\boldsymbol{J}$ equals the sum of the number of the scaling coefficients and the total number of the wavelet coefficients of the adjacent lower level $\boldsymbol{J-1}$ (see Eq. (5.33)), which means, the number of coefficients during the decomposition process of MSR is also kept.

As an important application of the MSR, data compression can be performed. Since all small-scale structures are contained in the detail signals of the higher levels, level-dependent thresholds are applied and the wavelet coefficients whose absolute values are smaller than the specified threshold are neglected. Following Zeilhofer [2008], a relation between the level-dependent

Table 6.2: The number $K_{\boldsymbol{J}}$ of scaling coefficients of levels $\boldsymbol{J} \in \{\{3,3,3\}, \ldots \{0,0,0\}\}$, the total number $L_{\boldsymbol{J}}$ of wavelet coefficients with the number $L^n_{\boldsymbol{J}}$ $(n = 1 \ldots, 7)$ of wavelet coefficients of the 7 subbands.

| level $\boldsymbol{J}$ | $K_{\boldsymbol{J}}$ | Total $L_{\boldsymbol{J}}$ | $L^1_{\boldsymbol{J}}$ | $L^2_{\boldsymbol{J}}$ | $L^3_{\boldsymbol{J}}$ | $L^4_{\boldsymbol{J}}$ | $L^5_{\boldsymbol{J}}$ | $L^6_{\boldsymbol{J}}$ | $L^7_{\boldsymbol{J}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 3, 3, 3 | 1000 | – | – | – | – | – | – | – | – |
| 2, 2 ,2 | 216 | 784 | 144 | 144 | 144 | 96 | 96 | 96 | 64 |
| 1, 1, 1 | 64 | 152 | 32 | 32 | 32 | 16 | 16 | 16 | 8 |
| 0, 0, 0 | 27 | 37 | 9 | 9 | 9 | 3 | 3 | 3 | 1 |

thresholds $\varrho^{0,0,0}_{\kappa_q} = 0.5 \cdot \varrho^{1,1,1}_{\kappa_q} = 0.25 \cdot \varrho^{2,2,2}_{\kappa_q}$ are assumed and the threshold $\varrho^{2,2,2}_{\kappa_q}$ for the detail signal of the levels $J_1 = J_2 = J_3 = 2$ is chosen empirically. The more the wavelet coefficients are neglected, the higher the compression rate. This is clearly reflected from Table 6.3, which illustrates the variations of the number of neglected wavelet coefficients, compression rate, Root Mean Square (RMS) value of the deviations between the compressed and the original signal of the highest level, with respect to different thresholds, exemplarily applied for $\Delta\widehat{\boldsymbol{d}}_{NmF2}$. The compression rate $\xi$ is calculated by

$$\xi = \frac{\sum\limits_{\boldsymbol{J'}} m_{\boldsymbol{J'}}}{\sum\limits_{\boldsymbol{J'}} L_{\boldsymbol{J'}}} \times 100\% \quad \text{with} \quad \boldsymbol{J'} \in \{\{2,2,2\},\{1,1,1\},\{0,0,0\}\} \tag{6.13}$$

where the numerator is the total number of neglected wavelet coefficients with $m_{\boldsymbol{J'}}$ the number of neglected wavelet coefficients of level $\boldsymbol{J'}$, and the denominator denotes the total number of

Table 6.3: Data compression with different level-dependent thresholds: $m_{\boldsymbol{J'}}$ denotes the total number of neglected wavelet coefficients of the 7 subbands of level $\boldsymbol{J'}$.

| $\varrho^{2,2,2}_{NmF2}$ [el/cm$^3$] | Total # $\sum_{\boldsymbol{J'}} m_{\boldsymbol{J'}}$ | $m_{2,2,2}$ | $m_{1,1,1}$ | $m_{0,0,0}$ | $\xi$ | RMS [el/cm$^3$] |
|---|---|---|---|---|---|---|
| $5 \cdot 10^4$ | 844 | 722 | 105 | 17 | 86.7% | $5.4 \cdot 10^4$ |
| $4 \cdot 10^4$ | 809 | 701 | 93 | 15 | 83.1% | $5.0 \cdot 10^4$ |
| $3 \cdot 10^4$ | 751 | 658 | 81 | 12 | 77.2% | $4.6 \cdot 10^4$ |
| $2 \cdot 10^4$ | 666 | 589 | 69 | 8 | 68.4% | $3.8 \cdot 10^4$ |
| $1 \cdot 10^4$ | 507 | 456 | 47 | 4 | 52.1% | $2.8 \cdot 10^4$ |
| $9 \cdot 10^3$ | 482 | 435 | 43 | 4 | 49.5% | $2.7 \cdot 10^4$ |
| $8 \cdot 10^3$ | 458 | 415 | 39 | 4 | 47.1% | $2.5 \cdot 10^4$ |
| $7 \cdot 10^3$ | 433 | 393 | 36 | 4 | 44.5% | $2.4 \cdot 10^4$ |

wavelet coefficients. According to Table 6.2, we have that $\sum_{\boldsymbol{J'}} L_{\boldsymbol{J'}} = 784 + 152 + 37 = 973$. The compressed signal is computed on the defined 3-D grids according to the Eqs. (5.44), (5.57) and (5.59) from the reduced set of coefficients. As can be seen from Table 6.3, with increasing threshold, more wavelet coefficients are neglected and thus the compression rate becomes larger. However, larger RMS value is the deviation between the compressed and the original signals. Usually, the threshold has to be chosen so that a good balance between lost information and compression rate is achieved.

Taking the threshold $\varrho^{2,2,2}_{NmF2} = 8 \cdot 10^3$ el/cm$^3$ for example, the compressed signal of $\Delta\widehat{NmF2}$ at 16 UT on July 1, 2012 is shown in the left panel of Fig. 6.9(a) and its deviation to the original signal (Fig. 6.8(A)(a)) is illustrated in the right panel. As can be seen, there is no large difference, and therefore no dominant signal is lost. The RMS value of the deviation map is approximately $3.1 \cdot 10^4$ el/cm$^3$. At the selected threshold, it can be seen from Table 6.3 that almost half of the wavelet coefficients are neglected out of the total 973 wavelet coefficients. Compared with the 1000 scaling coefficients for $NmF2$ at the highest levels, totally $1000 - 458 = 542$ coefficients needs to be stored. Figure 6.9(b) shows the results for $\Delta\widehat{hmF2}$ at the threshold $\varrho^{2,2,2}_{hmF2} = 1$ km with the compression rate of 36.3%. The RMS value of the deviation map is about 3.4 km, and 3.8 km considering all the grid points during the whole day. Figure 6.9(c) depicts the results for $\Delta\widehat{HF2}$ at the threshold $\varrho^{2,2,2}_{HF2} = 0.3$ km with the compression rate of 30.5%. The RMS of the deviations there equals 1.3 km at 16 UT and 1.2 km during one day.

It should be kept in mind that the positions of the nonzero coefficients have also to be stored besides the nonzero coefficients. Since the input data sets in this scenario are small, the chosen B-spline levels are not high, i.e., the number of the wavelet coefficients is a few. Consequently, the effect of data compression cannot be achieved well. The purpose of the examples is to demonstrate the application of MSR for data compression. An efficient data compression could be achieved when handling huge data sets.

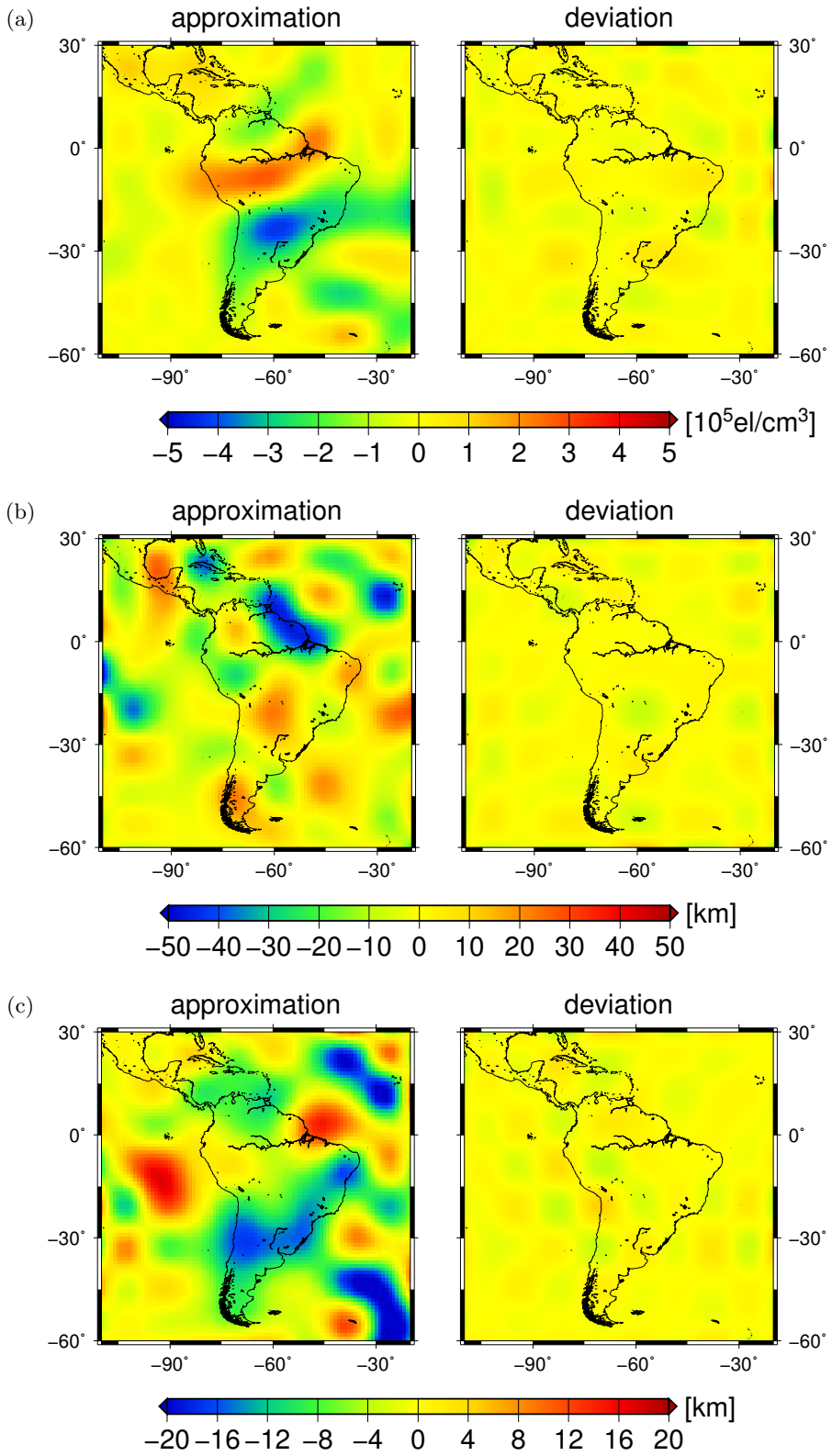



Figure 6.9: Approximation of the compressed signal of $\Delta\widehat{NmF2}$ with the threshold $\varrho^{2,2,2}_{NmF2} = 8 \cdot 10^3\,\mathrm{el/cm^3}$ (a, left), $\Delta\widehat{hmF2}$ with the threshold $\varrho^{2,2,2}_{hmF2} = 1\,\mathrm{km}$ (b, left) and $\Delta\widehat{HF2}$ with the threshold $\varrho^{2,2,2}_{HF2} = 0.3\,\mathrm{km}$ (c, left) at 16 UT on July 1, 2012 based on reduced number of wavelet coefficients; Deviations between the compressed and original signals of $\Delta\widehat{NmF2}$ (a, right), $\Delta\widehat{hmF2}$ (b, right) and $\Delta\widehat{HF2}$ (c, right).

### 6.1.5 Validation

In the following, the results of the developed model will be validated by independent datasets: ionosonde data (Section 6.1.5.1) and VTEC maps computed by a different approach (Section 6.1.5.2). Additionally, the developed model is evaluated by a cross validation with GPS observations (Section 6.1.5.3).

#### 6.1.5.1 Ionosonde measurements

The principle of ionosonde is to transmit a radio pulse vertically and to measure the travel time, that is taken to receive the echo because of the reflection of the ionosphere in the different layers. As the ionosonde transmits signals of different frequencies from low to high, the critical frequencies (see the introduction of Section 2.2) of the different layers can be deduced from the so-called ionograms (a graph of "virtual" heights[1] versus frequencies), since a signal that has a higher frequency than the critical frequency of a certain layer will not be reflected from that layer and it will pass to a denser layer. It is generally stated that ionosondes provide reliable values of $foF2$. According to Eq. (5.10), $NmF2$ can be derived from $foF2$. The ground-based ionosonde data from two stations in the study area are used: Jicamarca (geographical coordinate: $(12.0°\text{S}, 283.2°\text{E})$) and Port Stanley (geographical coordinate: $51.6°\text{S}, 302.1°\text{E}$)). The data can be downloaded from the Global Ionospheric Radio Observatory (GIRO) [Reinisch and Galkin, 2011]. All ionogram data from Jicamarca and Port Stanley were manually rescaled using the visualization and editing tool SAO-Explorer [Khmyrov *et al.*, 2008], to correct the unreliable autoscaling by the outdated Automatic Real-Time Ionogram Scaler with True height (ARTIST) 4 software [Galkin *et al.*, 2008] that is still being used in the old digisondes at these stations.

Figure 6.10 shows the $NmF2$ comparison of the developed model and IRI with the ionosonde measurements as the reference, for the three selected days under high solar activity (top), moderate solar activity (mid) and low solar activity (bottom), respectively. The left column illustrates the comparison at station Jicamarca and the right for station Port Stanley. The uncertainty of the estimates (i.e., the standard deviations), computed according to Eq. (4.18), of the developed model is plotted along with the model estimations. It can be seen that the developed model for $NmF2$ generally fits to ionosonde data better than IRI before 14 UT at Jicamarca for July 1, 2012, whereas most of the ionosonde data lie between the developed model and IRI afterwards. As for Port Stanley, the model estimation is similar to IRI, whereas the ionosonde measurements almost fall within the model uncertainty range from 0 UT till 11 UT. For July 16, 2011, disturbances such as Traveling Ionospheric Disturbances (TIDs)[2] can be observed (Reinisch, B. W., personal communication, 2014). A good match between the results of the developed model and the ionosonde measurements can still be seen from 0 UT to 2 UT at Jicamarca, and it is also apparent from the figure that the model estimations are much closer to the ionosonde data than IRI around the daily maximum of $NmF2$ at Port Stanley. Furthermore, on July 16, 2008, although there was a strong spread F[3] at night at Jicamarca when the F layer had no well-defined $NmF2$

[1] Radio waves travel more slowly in the ionosphere than in free space, however, the radio waves are assumed to travel at the speed of light to derive the height of the layers from the travel time.

[2] TIDs refer to irregularities of the F region exhibited as the wave-like structures of electron density oscillations [Zolesi and Cander, 2014].

[3] Spread F refers to plasma instability phenomenon occurring in the F region, which stems from the earliest observations using ionosondes, showing that the reflected echo did not display a well-behaved pattern but was "spread" in range or frequency [Kelley, 2009, and references therein].

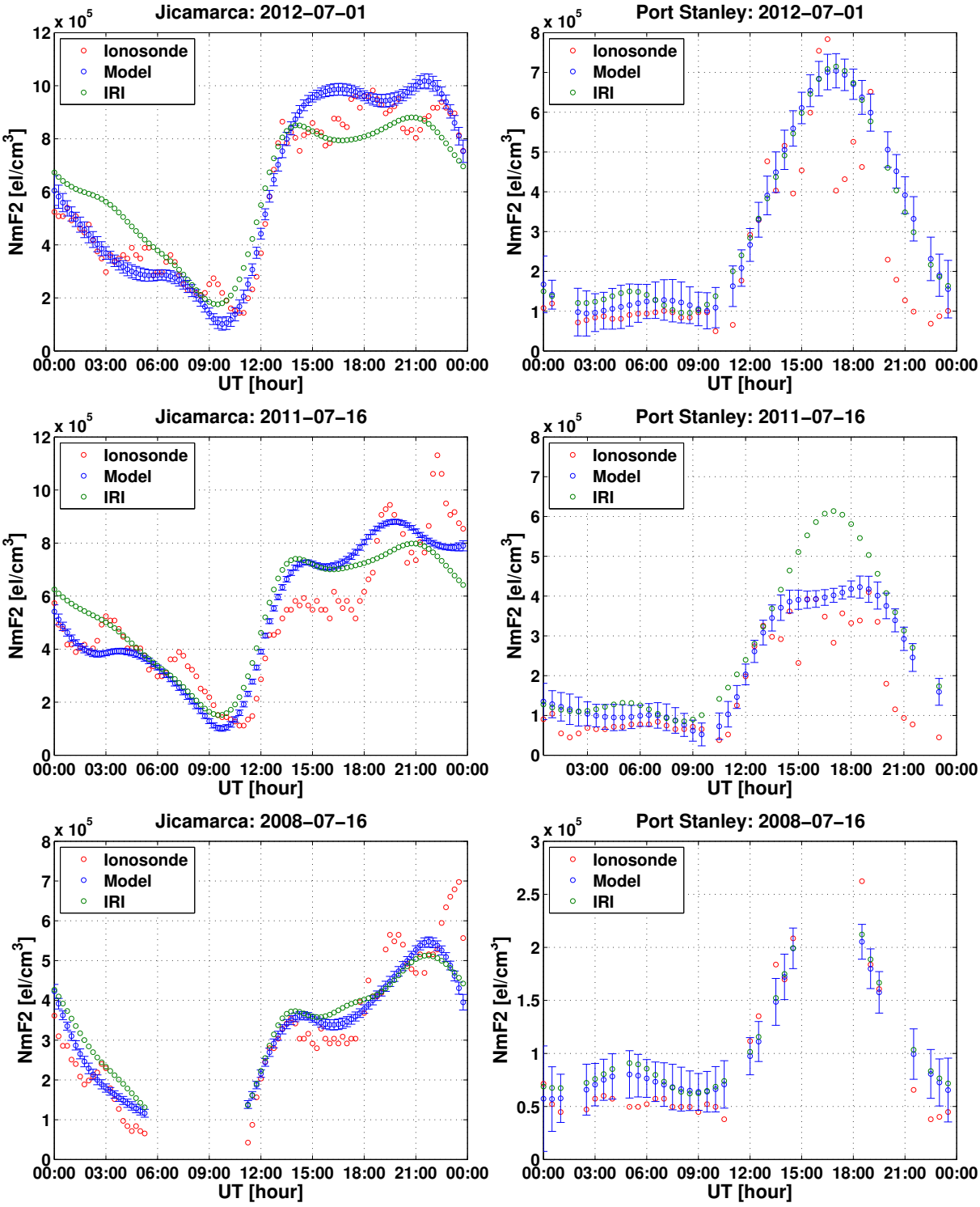



Figure 6.10: Comparisons of $NmF2$ estimated from the developed model (blue circles), IRI-2012 (green circles) and ionosonde observations (red circles) at Jicamarca (left) and at Port Stanley (right) for July 1, 2012 (top), July 16, 2011 (mid) and July 16, 2008 (bottom). The standard deviations (i.e., the formal errors) are plotted along with the model estimations [Liang *et al.*, 2015b].

(Reinisch, B. W., personal communication, 2014), it can still be seen that the developed model is generally closer to the ionosonde data than IRI from 0 UT to 5 UT. Regarding Port Stanley, a number of ionograms were unfortunately missing (it is an old Digisonde Portable Sounder (DPS) installed in 1993), the figure reveals that the developed model for $NmF2$ is slightly closer to the ionosonde measurements till 7 UT whereas the developed model and IRI are comparable for the rest of the day. In general, the modeling approach shows the potential to update IRI $NmF2$ at both stations.

With the ionosonde measurements as the reference, some statistical quantities are calculated. The standard deviations of the difference and the relative difference between the model estimations and IRI are shown in Table 6.4. It is obvious that for all three days, at Jicamarca, the standard deviations of the differences between the developed model and ionosonde data are smaller than the ones between IRI and ionosonde data. At Port Stanley, the developed model also yields smaller standard deviation of relative differences than IRI for the day under moderate solar activity (July 16, 2011) and the day under low solar activity (July 16, 2008). However, the developed model gives a slightly larger standard deviation than IRI for the day under high solar activity (July 1, 2012). In particular, at Jicamarca, the developed model achieves the data from 2012, 2011 and 2008, respectively, a smaller variation by 10.6%, 6.9% and 5.4% than IRI. These numbers may indicate that the variations gradually decrease with decreasing solar activity, which means that the improvements probably increase with increasing solar activity. At Port Stanley, the developed model reaches the ionosonde measurements with a smaller variation of about 8.6% than IRI for the day under moderate solar activity, however, the developed model and IRI are comparable for the other two days.

The relative uncertainty of the model estimations of $NmF2$ is then computed. We use the term relative uncertainty of a set of model parameters, which is referred to its standard deviation divided by the corresponding model estimate. The mean of relative uncertainties of $NmF2$ estimations of the developed model reaches 6.1% at Jicamarca, and amounts to about 27.0% at Port Stanley for July 1, 2012. The values are 2.9% at Jicamarca and 21.3% at Port Stanley for July 16, 2011, and 3.9% and 28.1% at Jicamarca and at Port Stanley, respectively, for July 16, 2008. It is apparent that the uncertainties of the developed model at Port Stanley are larger than those at Jicamarca, which might be attributed to a better distribution of observations at Jicamarca than at Port Stanley (cf. gray triangles in Fig. 6.3, less GPS stations exist in the surrounding area of the station Port Stanley).

Based on the validation by ionosonde data, the following conclusions can be drawn:

- in general, the developed model approaches the ionosonde data better than IRI,
- the improvements are different for the two stations, which is likely due to different data distributions,
- a clear correlation of the improvements with solar activity is not detectable, which is probably due to the small database.

#### 6.1.5.2 VTEC

The model is validated additionally by a comparison with the VTEC computed by the IGS IAACs [Hernández-Pajares *et al.*, 2009]. Based on $\widehat{NmF2}$, $\widehat{hmF2}$ and $\widehat{HF2}$ values of the developed model, the vertical EDPs are constructed on the defined 3-D grid points. Then VTEC values

Table 6.4: Standard deviations of the $NmF2$ differences (absolute and relative) between IRI or the developed model and the ionosonde data

| | **Jicamarca** | | |
|---|---|---|---|
| | **2012-07-01** | **2011-07-16** | **2008-07-16** |
| IRI [el/cm$^3$] | $1.1 \cdot 10^5$ (29.0%) | $1.3 \cdot 10^5$ (33.5%) | $7.8 \cdot 10^4$ (41.2%) |
| Model [el/cm$^3$] | $8.0 \cdot 10^4$ (18.4%) | $1.1 \cdot 10^5$ (26.6%) | $7.0 \cdot 10^4$ (35.8%) |
| | **Port Stanley** | | |
| | **2012-07-01** | **2011-07-16** | **2008-07-16** |
| IRI [el/cm$^3$] | $9.4 \cdot 10^4$ (61.7%) | $8.9 \cdot 10^4$ (74.8%) | $2.2 \cdot 10^4$ (35.4%) |
| Model [el/cm$^3$] | $1.0 \cdot 10^5$ (63.7%) | $6.0 \cdot 10^4$ (66.2%) | $2.2 \cdot 10^4$ (32.7%) |

are computed using the Gauss–Legendre quadrature. For more details the VTEC product from CODE, as one of the IGS IAACs, is taken as a reference for comparison. Their VTEC maps are based on spherical harmonic representations using ground-based GNSS observations [Schaer, 1999]. The comparison is shown in Fig. 6.11 exemplarily for 16 UT of July 1, 2012. The top panels (from left to right) illustrate the VTEC maps of the developed model, IRI and CODE, respectively. The VTEC difference between the estimated model and CODE is shown in the bottom left panel, whereas the one between IRI and CODE is illustrated in the bottom right panel. As can be seen from the difference maps, IRI generally has larger differences to CODE than the developed model in particular over the continent. It can also be seen that the two difference maps are similar such as the top left corner. This is because that almost no observation is available there around 16 UT. For those regions without any observation, the developed model values are dependent on the prior information derived from IRI. The RMS value of the differences at 16 UT yields 6.8 TECU between IRI and CODE, which is improved to 5.6 TECU by the developed model. Regarding all the grid points during the whole day, the RMS values of the differences to CODE is 6.4 TECU for IRI, and 5.7 TECU for the developed model. For July 16, 2011, the RMS value is 4.7 TECU for IRI, which is slightly improved to 4.2 TECU by the developed model, while for July 16, 2008, the RMS values are comparable, 2.8 TECU for IRI and 2.7 TECU for the developed model.

It would be interesting to see if there is a discrepancy between the continental and the ocean areas. For that purpose, all the constructed grid points are divided into either continental or oceanic points. Then the correlation coefficient is computed to have a quantitative comparison. At 16 UT the correlation coefficient between IRI and CODE is 0.53 over the continent, increased significantly to 0.94 by the developed model. The correlation coefficient between IRI and CODE over the ocean areas is 0.84, compared to 0.86 between the developed model and CODE. As for all the continental grid points during the whole day, the correlation coefficient between IRI and CODE is increased by the developed model from 0.82 to 0.92 for July 1, 2012, from 0.87 to 0.94 for July 16, 2011, and from 0.90 to 0.92 for July 16, 2008. However, for the ocean areas, the developed model and IRI are similar when compared with CODE. Specifically, the correlation coefficient of the developed model is 0.89 compared to 0.87 of IRI for 2012, and both are about 0.90 for 2011 and 2008. The relatively higher similarity between the developed model and CODE over the continent is due to the fact that both the developed model and CODE use ground-based GPS observations as input.

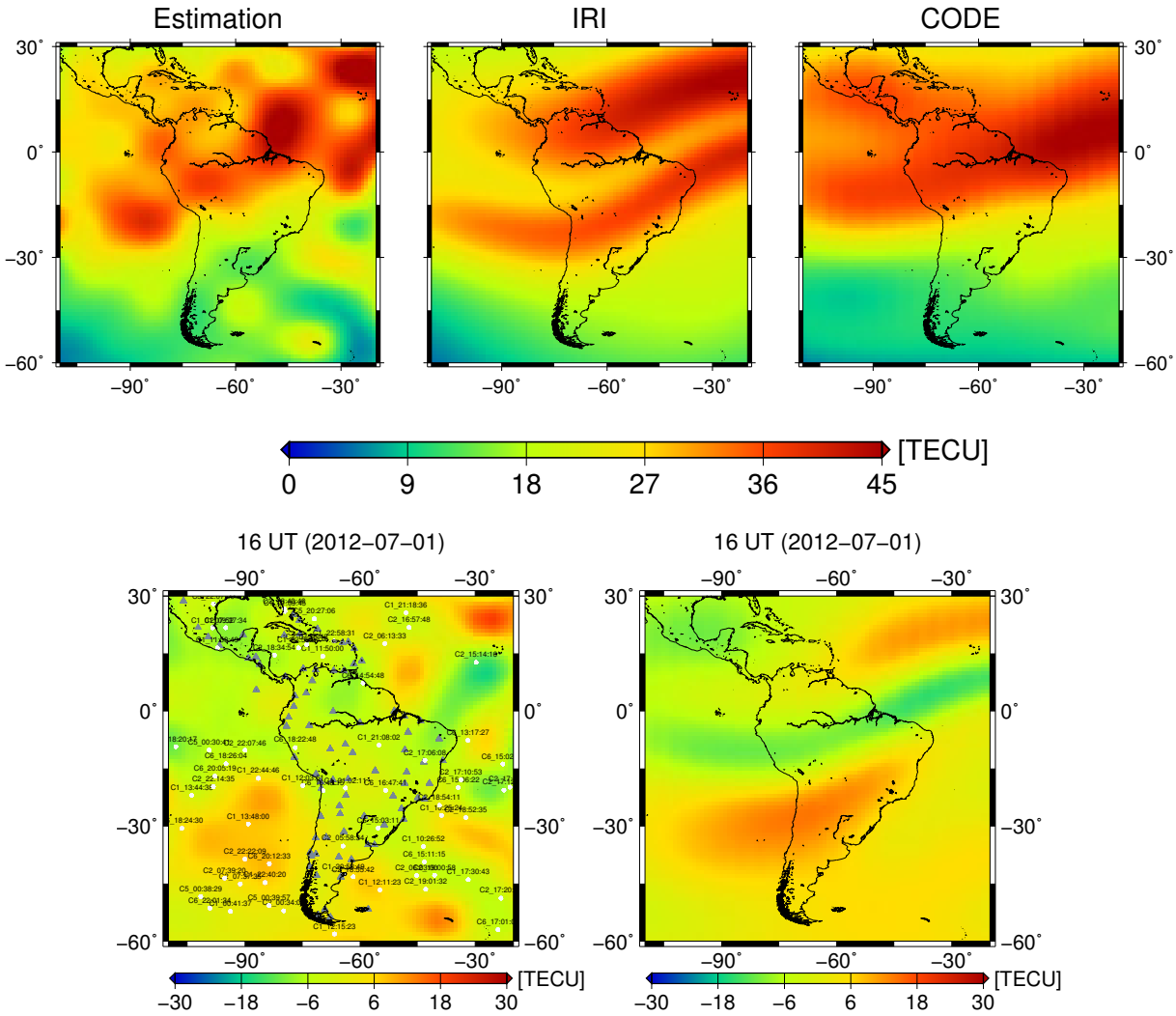



Figure 6.11: VTEC comparisons (July 1, 2012): VTEC maps at 16 UT from the developed model, IRI and CODE (top, from left to right); VTEC differences to CODE of the developed model (bottom left), IRI (bottom right).

Based on the above comparisons, the following statements might be concluded:

- VTEC difference to CODE can be improved by the developed model by up to 10% with respect to IRI,
- the improvements decrease with decreasing solar activity,
- main improvements can be reached over the continents, since most observations are ground-based GPS data which are located there.

#### 6.1.5.3 Cross-validation

A simple cross-validation by GPS observations is also applied to evaluate the performance of the developed model. For this purpose, the input GPS dataset is divided randomly into two parts with roughly equal number of ray paths: one is used for the parameter estimation (called

the training set) and the other for validation (called the testing set). Figure 6.12(a, b) display the geographical distribution of the two datasets exemplarily for July 1, 2012, whereas IPPs are shown in blue for the training set (a) and in red for the testing set (b), at a selected height of 450 km. Firstly, the developed model is estimated using the dataset (a), then STEC values of the ray paths of the dataset (b) (as the reference data for comparisons) are calculated based on the model parameters estimated from (a). Figure 6.12(c) shows the STEC comparison of the dataset (b) plotted with respect to UT. The blue dot is the calculated model $STEC$, the red dot represents the measurements with DCBs subtracted based on the developed model estimation, and their difference is shown in green. As can be seen, $STEC$ values are larger between 15 UT and 20 UT, which fits to the local maximum of the ionosphere at about 14 LT. It can also be seen that the differences of the model and the measurements do not differ with respect to the time, and have a mean value of approximately 0.3 TECU. The RMS value of the differences is about 3.9 TECU, which is comparable to the estimated mean standard deviation 3.0 TECU of the residuals of the GPS observations. For both 2011 and 2008, the RMS values of the difference of the model $STEC$ and the reference data are about 2.5 TECU with respect to the estimated mean standard deviations of 2.3 TECU.

It can be concluded from the cross-validation that:

- the STEC residuals have no dependency on UT and solar activity,
- a STEC accuracy of about 4 TECU can be reached by the model.

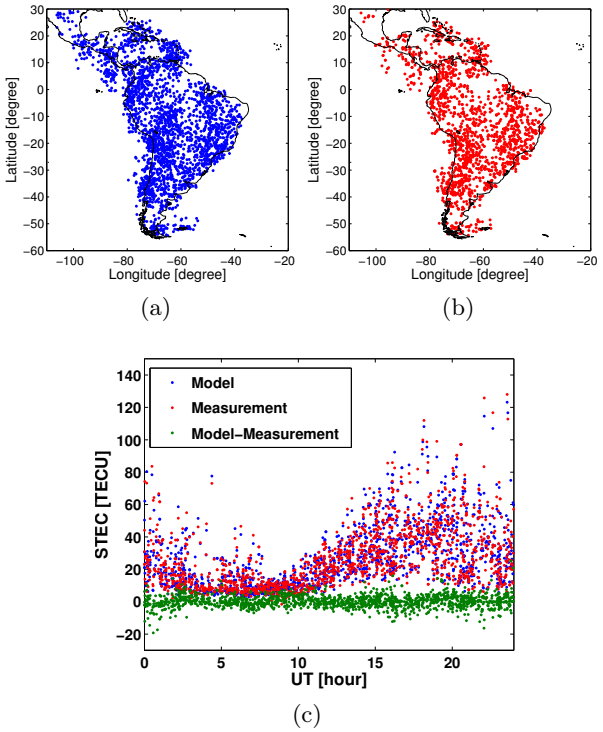



Figure 6.12: Cross-validation: datasets for model calculation (a) and for model validation (b) for July 1, 2012. Blue and red dots show the IPPs of the GPS observations at 450 km; STEC comparisons (c): computed STEC (blue dots), measurements (red dots) and differences (green dots) [Liang *et al.*, 2015b].

## 6.2 Multi-layer modeling

The model (6.1) includes only the most important F2 layer. To have a physically more realistic representation, the profile function is extended by considering a multi-layer approach according to Eq. (5.8). However, this makes parameter estimation more difficult as the number of key parameters is increased significantly. In order to consider the most important layers, the significance of the distinct ionosphere layers is investigated firstly. To this end, the contributions of the different layers in Eq. (5.8) to the VTEC till GPS orbit height are quantified, as GPS data are one of the main input data for ionosphere modeling.

### 6.2.1 Significance of distinct layers

For this purpose, EDPs spanning the height range till 20,200 km are required. Since RO data provide EDPs only till orbit height of LEO satellites, empirical models will thus be applied to obtain EDPs for the investigation. As introduced in Section 2.5, IRI predicts the electron density at the height range from 50 km to 2000 km, whereas IRI-Plas provides the electron density towards the plasmaspause up to 36,000 km. Assuming that IRI with the NeQuick topside option provides a more reliable electron density distribution than IRI-Plas below 2000 km, IRI with the NeQuick topside model and IRI-Plas are therefore combined. In particular, the electron densities till 2000 km are taken directly from IRI and those above 2000 km are obtained from IRI-Plas. In order to avoid jumps which are caused by an inconsistency of the two models, gradients of EDP from the IRI-Plas model in the area around 2000 km are compared with the gradient of EDP from IRI at 2000 km. The used electron density points from IRI-Plas are starting at the height when the absolute value of the gradient of EDP from IRI-Plas is larger than the value from IRI at 2000 km. Fig. 6.13 shows exemplarily a profile located at (20°S, 300°E) on 20 June, 2012, where the electron density values predicted from IRI are indicated as blue dashed line, the black dashed line with crosses represents the electron densities obtained from IRI-Plas, and the red dashed line with circles shows the combined ones.

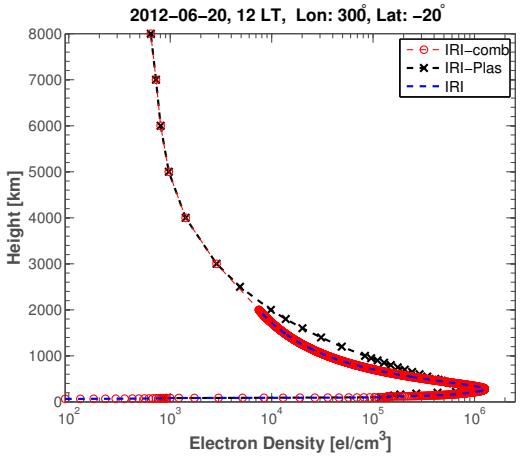



Figure 6.13: EDP from IRI (blue dashed line), IRI-Plas (black dashed line with crosses) and the final combined one (red dashed line with circles). Here only the electron densities below 8000 km are shown. Note that the x-axis is in logarithmic scale.

Since electron densities are given at discrete heights, interpolation is required at the nodes when

using the Gauss-Legendre formula (5.86). For convenience, the trapezoidal rule is applied. To have a quantitative overview at different geographical positions and ionosphere conditions, EDPs are generated on a regular grid with a resolution of $10° \times 10°$ over the globe. Four days with different seasons are selected during 2008 (low solar activity) and 2012 (high solar activity), respectively. The nighttime condition is not considered because the D and F1 layers disappear at night [Hargreaves, 1992]. The computation is performed exemplarily at 12 LT.

Figure 6.14 shows the contributions (in percentage) of the different layers, of all grid points over the globe arranged along the x-axis, exemplarily for the selected day June 20, 2012. It is clearly visible that the largest contribution comes from the F2 layer, followed by the plasmaspheric part and the F1 layer. As expected, the smallest contribution is from the D layer. Average contribu-

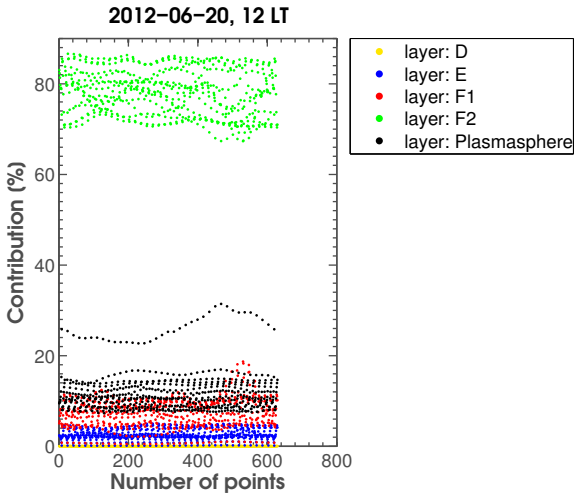



Figure 6.14: Contributions of the different layers to VTEC integrated till GPS orbit height (2012-06-20), at 12 LT.

tions of the different layers over the globe during the selected days are shown in Table 6.5(a) for 2008 and Table 6.5(b) for 2012, where the colors red, green, yellow mark the first three largest contributions, respectively. The results in various seasons are consistent. The F1 layer seems to be more significant during the summer and at low solar activity, which is in agreement with the statement in, e.g., Pezzopane and Scotto [2008]. The contributions of the D layer are within 0.03%.

### 6.2.2 Introduction of a multi-layer/plasmasphere profile

According to the above results, we neglect the D layer, i.e., the general model (5.8) becomes

$$N_e(h) = N_E(h) + N_{F1}(h) + N_{F2}(h) + N_{H^+}(h > hmF2)\,. \tag{6.14}$$

Like in Eq. (6.1), the $\alpha$-Chapman function is used to describe the F2 layer. As introduced previously in Section 2.2, the E and F1 layers behave like $\alpha$-Chapman layer. Therefore, three $\alpha$-Chapman functions are applied for modeling the E, F1 and F2 layers, respectively. The plasmaspheric term $N_{H^+}$ uses the exponential function (2.63), and it starts above $hmF2$ following

Table 6.5: Average contributions of the different ionosphere layers and plasmasphere over the globe during selected days of 2008 under low solar activity (a) and 2012 under high solar activity (b).

(a)

| Day | D-layer 50–90 km | E-layer 90–140 km | F1 layer 140–200 km | F2 layer 200–1000 km | Plasmasphere 1000–20,200 km |
|---|---|---|---|---|---|
| 2008-03-21 | 0.02% | 3.67% | 8.53% | 76.25% | 11.53% |
| 2008-06-20 | 0.03% | 4.54% | 10.93% | 71.14% | 13.35% |
| 2008-09-19 | 0.03% | 4.30% | 9.67% | 73.73% | 12.27% |
| 2008-12-19 | 0.02% | 3.15% | 8.36% | 75.20% | 13.26% |

(b)

| Day | D-layer 50–90 km | E-layer 90–140 km | F1 layer 140–200 km | F2 layer 200–1000 km | Plasmasphere 1000–20,200 km |
|---|---|---|---|---|---|
| 2012-03-21 | 0.01% | 1.92% | 5.13% | 82.39% | 10.55% |
| 2012-06-20 | 0.02% | 2.59% | 7.45% | 78.10% | 11.84% |
| 2012-09-19 | 0.02% | 2.16% | 5.76% | 80.94% | 11.12% |
| 2012-12-19 | 0.02% | 1.97% | 5.84% | 79.81% | 12.37% |

Eq. (5.9). Therefore, Eq. (6.14) is modeled by

$$N_e(h) = \sum_{i=1}^{3} Nm_i \exp\left\{0.5\left[1 - \frac{h - hm_i}{H_i} - \exp\left(-\frac{h - hm_i}{H_i}\right)\right]\right\} + N_{P0} \exp\left(-\frac{h}{H_p}\right) \quad (6.15)$$

with altogether eleven key parameters, collected in the $11 \times 1$ vector

$$\boldsymbol{\kappa} = [NmE, hmE, HE, NmF1, hmF1, HF1, NmF2, hmF2, HF2, N_{P0}, H_p]^T. \quad (6.16)$$

According to Liu *et al.* [2006, and references therein], the Chapman function, even with a constant scale height, fits the topside ionosphere profile well up to several hundred kilometers above the F2 peak. Within this thesis, the plamaspheric term $N_{H^+}$ is considered to start at $hmF2+200$ km whereas 200 km is chosen empirically to better fit the IRI topside. Figure 6.15 shows the profile function (6.15) with exemplary $\boldsymbol{\kappa}$ vector.

When applying a multi-layer modeling approach, a problem will arise whether such many key parameters can be separated from each other. In order to check if the key parameters can be estimated jointly, we use simulated EDPs to estimate the parameters in the following.

### 6.2.3 Simulation

**Investigation of error covariance matrix of EDP**

Before simulating EDPs, the error propagation through the onion peeling algorithm is investigated following the strategy introduced in Section 5.3.1. We assume that the errors of the STEC observables derived from carrier phase observations are statistically independent to each other

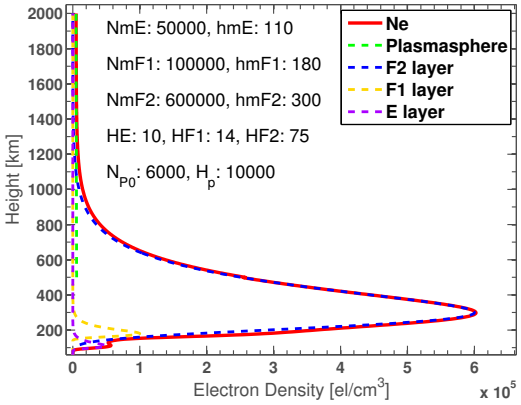



Figure 6.15: Vertical distribution of the electron density with exemplary $\boldsymbol{\kappa}$ vector (Eq. (6.15)).

and are random, Gaussian-distributed. Therefore, the error covariance matrix $D(\widetilde{\boldsymbol{STEC}})$ can be assumed to be a diagonal matrix. Additionally, the standard deviations of $\widetilde{\boldsymbol{STEC}}$ along the height are considered to be equal to 1 TECU. Therefore, $D(\widetilde{\boldsymbol{STEC}}) = \mathbf{I}$ is used in the following numerical computations.

The $N_e$ observations along the profile are considered in a height range of $[50, 800]$ km with a constant resolution $\Delta h = 2\,\text{km}$. This leads to altogether 376 $N_e$ observations. The covariance matrix $D(\boldsymbol{N_e})$ is then calculated according to Eq. (5.75). The standard deviation $\sigma_{N_e,i}$ of $N_e$ observations along the height $h$ is shown in Fig. 6.16. As can be seen, $\sigma_{N_e,i}$ is increasing from top to bottom, i.e., the errors get larger from top to bottom. This clearly reflects that the errors of the observations at higher altitudes propagate to the ones at lower altitudes. Except that from the first to the second observation, the standard deviation increases linearly from top to bottom. However, $\sigma_{N_e,i}$ do not vary strongly along the height where the ratio between the highest and the lowest points is about 1.3. Consequently, they can be assumed to be of equal quality.

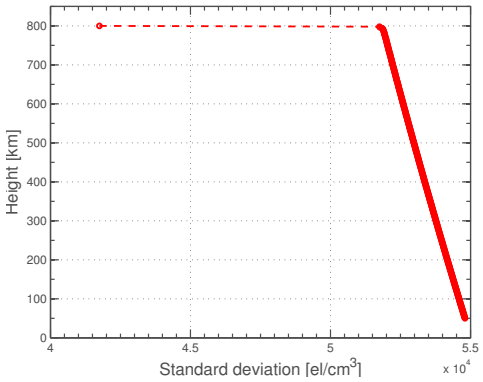



Figure 6.16: Variations of $\sigma_{N_e,i}$ as a function of height.

Based on Eq. (4.7), the correlation coefficient is calculated and organized in the correlation matrix following Eq. (4.22). The correlation matrix is illustrated in the left panel of Fig. 6.17. The x- and y-axis indicate the index of the $N_e$ observations along the height, where the top-left corner denotes the observation at 800 km and the bottom-right corner denotes the observation

at 50 km. To have a closer look, the part in the black dashed box is zoomed in the right panel. It

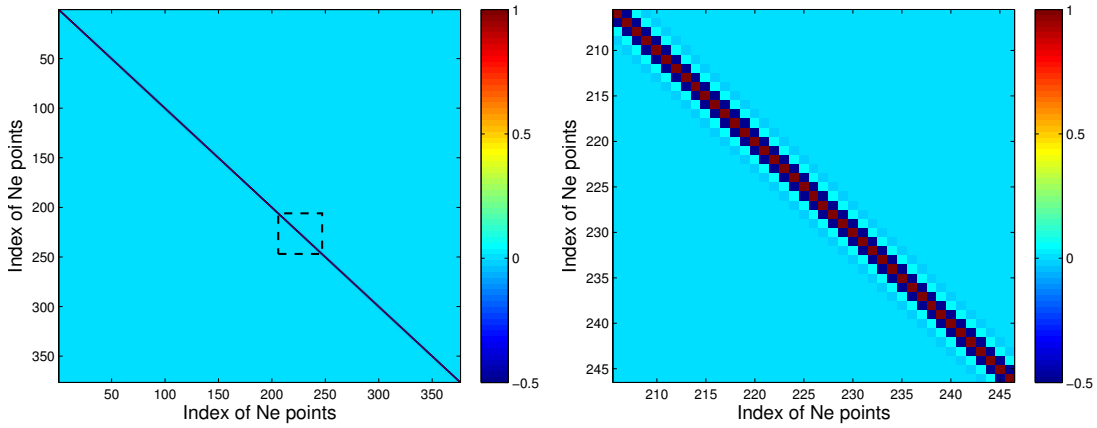



Figure 6.17: Correlation matrix of $N_e$ observations along the height (left), where the black dashed box is zoomed in the right panel.

can be visible that the strongest correlation occurs between the nearest neighboring observations, with a negative correlation approaching $-0.5$. This is understandable from the second row of the triangular matrix $\boldsymbol{B}$ in Eq. (5.73). An increased value of an electron density observation will decrease the value of the nearest observation as the sum of the linear combination of the two $N_e$ observations is fixed. The second strongest correlation occurs between the second nearest neighboring observations with a value of about 0.1. Furthermore, nearly no correlation appears. Generally, a correlation less than 0.5 can be regarded as week correlation, and therefore, the correlations between $N_e$ observations can be considered to be independent from each other.

A closed-loop simulation is then performed to verify the ability of the adjustment system to solve for the key parameters simultaneously. Specifically, noisy $N_e$ measurements are simulated and achieved by a sum of "true" electron density values calculated from a set of predefined "true" values of parameters according to Eq. (6.15) and random noises. Since Eq. (6.15) depends on the vector $\kappa$, we can write

$$N_{e,\,j} = N_e(h_j) = N_e(h_j, \boldsymbol{\kappa}_{true}) + e_j \tag{6.17}$$

where $j = 1, \ldots, n$ indicates the index of the measurements along a profile. The vector $\boldsymbol{\kappa}_{true}$ collects predefined "true" values of the parameters in $\boldsymbol{\kappa}$. Based on the numerical investigations above, the random measurement noise is assumed to be independently and normally distributed with a mean value of zero and a fixed standard deviation $\sigma_e$, i.e., the noise vector $\boldsymbol{e} = [e_1, \ldots, e_n]^T$ follows

$$\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2\,\mathbf{I}) \quad \text{with} \quad \sigma_e = \frac{\epsilon}{n} \cdot \sum_{i=1}^{n} N_e(h_j, \boldsymbol{\kappa}_{true})\,. \tag{6.18}$$

By setting $\boldsymbol{y}_{true} = [N_e(h_1, \boldsymbol{\kappa}_{true}), \ldots, N_e(h_n, \boldsymbol{\kappa}_{true})]^T$, $\boldsymbol{y} = [N_{e,1}, \ldots, N_{e,n}]^T$, we can formulate

$$\boldsymbol{y} = \boldsymbol{y}_{true} + \boldsymbol{e} \quad \text{with} \quad \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{y}_{true}, \sigma_e^2\,\mathbf{I})\,. \tag{6.19}$$

Taking the parameters of the plasmaspheric part into account, measurements will be simulated in the height range from 90 km to 1200 km with a constant resolution of 2 km. Therefore, we have a total number $n = 556$ of observations.

Figure 6.18(a-c) shows an example of simulated electron density measurements (black dots) at the different noise level $\epsilon = 2\%$ (a), $\epsilon = 5\%$ (b) and $\epsilon = 10\%$ (c) around the "true" profile (blue line), where the E, F1, and F2 layers are present. As can be seen, at the relatively low noise level $\epsilon = 2\%$, most of the measurements follow the trace of the "true" profile well. The measurements at the upper part of the profile are more noisy. This is due to relatively smaller electron density values there and thus relatively larger noise-to-signal ratio, as $\sigma_e$ of the random noises added in the simulated data is based on the mean of the total signal (see Eq. (6.18)). This phenomenon becomes more prominent in the panels (b) and (c), where the electron densities of the upper part deviate from the "true" more significantly with increasing noise levels. Furthermore, the lower part of the profile is more noisy than the middle region including the F2 layer. It can be expected that, parameter estimation will become more difficult with increasing noise levels in particular for the parameters of the E, F1 layer and the plasmaspheric part.

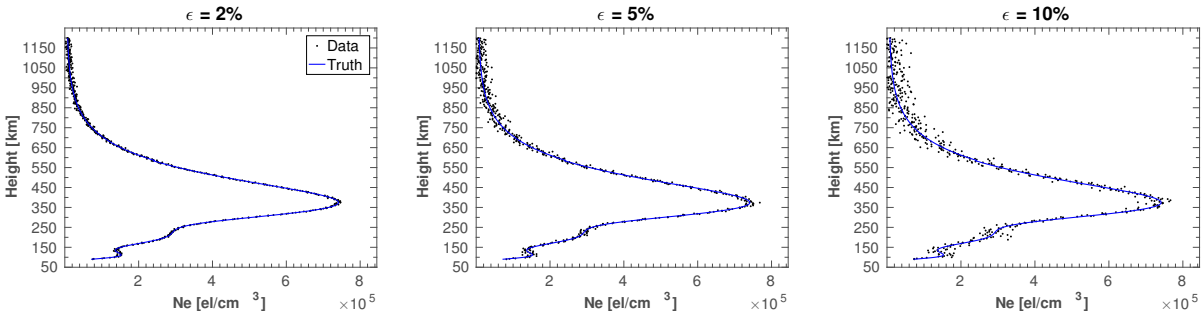



Figure 6.18: Simulated EDPs with different noise levels $\epsilon$ around the "true" profiles.

Within this thesis, we will focus on the seven key parameters, written in the unknown parameter vector

$$\boldsymbol{\beta} = [NmE, hmE, NmF1, NmF2, hmF2, HF2, N_{P0}]^T . \tag{6.20}$$

The other four parameters, i.e., $HE$, $hmF1$, $HF1$ and $H_p$ are assumed to be known and without bias, i.e., no systematic error is introduced. Similar to Eq. (5.72), the Gauss-Markov model of the linearized observation equation can be established as

$$\boldsymbol{X}\Delta\boldsymbol{\beta} = \boldsymbol{y} + \boldsymbol{e} \quad \text{with} \quad D(\boldsymbol{y}) = \sigma^2 \boldsymbol{P}^{-1} \tag{6.21}$$

where

$$\boldsymbol{X} = \begin{pmatrix} \frac{\partial \overline{N}_{e,1}}{\partial NmE}\big|_{\boldsymbol{\beta}_0} & \cdots & \frac{\partial \overline{N}_{e,1}}{\partial N_{P0}}\big|_{\boldsymbol{\beta}_0} \\ \frac{\partial \overline{N}_{e,2}}{\partial NmE}\big|_{\boldsymbol{\beta}_0} & \cdots & \frac{\partial \overline{N}_{e,2}}{\partial N_{P0}}\big|_{\boldsymbol{\beta}_0} \\ \vdots & \ddots & \vdots \\ \frac{\partial \overline{N}_{e,n}}{\partial NmE}\big|_{\boldsymbol{\beta}_0} & \cdots & \frac{\partial \overline{N}_{e,n}}{\partial N_{P0}}\big|_{\boldsymbol{\beta}_0} \end{pmatrix}, \quad \boldsymbol{\beta}_0 = [NmE^0, \ldots, N_{P0}^0],$$

$$\Delta\boldsymbol{\beta} = [\Delta NmE, \ldots, \Delta N_{P0}]^T, \quad \boldsymbol{y} = [N_{e,1} - \overline{N}_{e,1}|_{\boldsymbol{\beta}_0}, \ldots, N_{e,n} - \overline{N}_{e,n}|_{\boldsymbol{\beta}_0}]^T$$

by decomposing the unknown parameter vector $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \Delta\boldsymbol{\beta}$ into the initial vector $\boldsymbol{\beta}_0$ and the correction part $\Delta\boldsymbol{\beta}$. The partial derivatives in the design matrix $\boldsymbol{X}$ can easily be obtained from Eq. (6.15) in a similar way as the Eqs. (6.2) to (6.4). According to Eq. (6.18), the weight

matrix $\boldsymbol{P} = \mathbf{I}$ will be used. Since the F1 layer does not always exist, two different scenarios are simulated, one with F1 layer during the day and the other without F1 layer at night. Note, that the unknown parameter vector is always referred to Eq. (6.20), namely, $NmF1$ has to be estimated whether the F1 layer exist or not.

#### 6.2.3.1 Scenario 1 – with F1 layer

As stated previously, the design matrix should be scaled at first if the key parameters have very different magnitudes. Figure 6.19 shows the importance for introducing a scaling exemplarily using the simulated data in Fig. 6.18(b), where the eigenvalues of the normal equation matrix, i.e., $\boldsymbol{X}^T\boldsymbol{X}$ ($\boldsymbol{P} = \mathbf{I}$), and the scaled one, i.e., $\boldsymbol{X}_s^T\boldsymbol{X}_s$ with $\boldsymbol{X}_s = \boldsymbol{X}\boldsymbol{W}$ (see Eq. (5.100)), are shown in the left and right panels, respectively. As can be seen from the left panel, the maximum

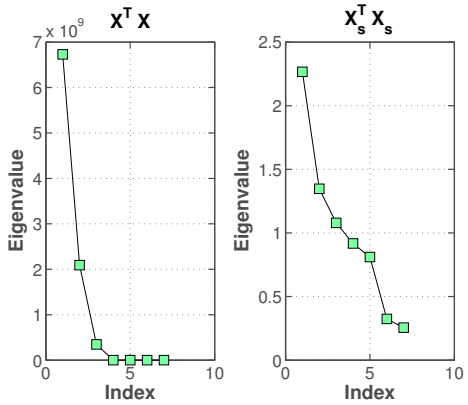
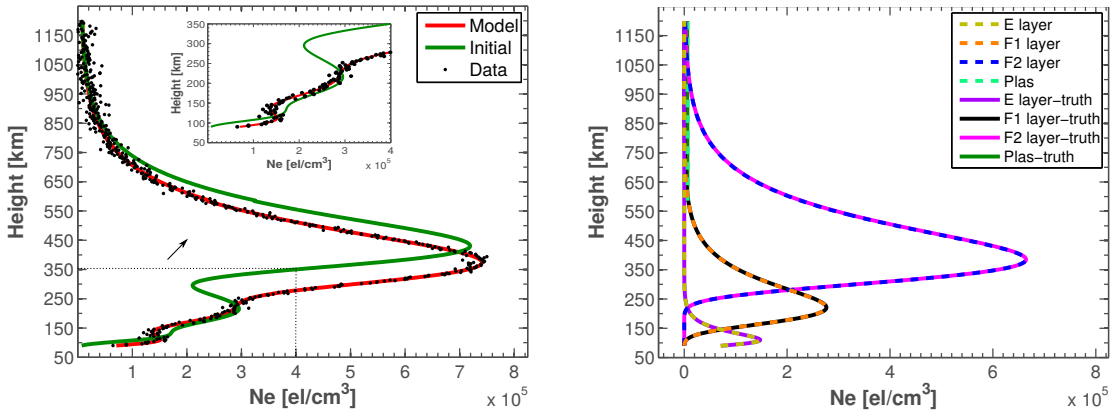



Figure 6.19: Eigenvalues of the normal equation matrix: original (left); with scaling (right).

eigenvalue is in the magnitude of $10^9$ and the minimum value is in the magnitude of 10. This will lead to a large condition number according to Eq. (4.40). However, the eigenvalues of the scaled normal equation matrix are in the same order of magnitude. Therefore, a small condition number is derived and a stable solution can be obtained. In the following calculations, scaling will always be introduced in the adjustment system.

The simulated profiles in Fig. 6.18 will be used as input data. The result of profile fitting is shown in Fig. 6.20 exemplarily at the noise level of $\epsilon = 5\%$ . The panel (a) illustrates the total profile, where the estimated one is marked by the red line, the profile calculated by $\boldsymbol{\beta}_0$ is indicated by the green line, and the input data are given by the black dots. For a better visualization, the lower part of the profile is zoomed in. It can be seen that the estimated profile generally follows along the data well. The standard deviation $\widehat{\sigma}$ of observation residuals, defined as the differences between estimated and input electron density values (i.e., $\widehat{\boldsymbol{e}} = \widehat{\boldsymbol{y}} - \boldsymbol{y}$), and $\sigma_e$ of the simulated input noise are both about $1.13 \cdot 10^4\,\mathrm{el/cm^3}$. To further look if the various layers can be partitioned correctly, the estimated and the "true" layers are plotted in the panel (b). It is clearly visible that the estimation and the "truth" are overlapping with each other. Thus, the different layers are partitioned fairly well.

The relative error $\delta_\beta$, defined as $\delta_\beta = |\widehat{\beta} - \beta_{true}|/\beta_{true}$, the estimate $\widehat{\beta}$, and the "true" value $\beta_{true}$ of all seven unknown parameters $\beta \in \boldsymbol{\beta}$ are illustrated in Table 6.6. It can be seen that small estimation errors for the parameters of the E, F1 layer and in particular the F2 layer are obtained.




(a) (b)

Figure 6.20: Profile fitting from the simulated data (with F1 layer) at $\epsilon = 5\%$ (a); separation into distinct layers (b).

However, $\widehat{N}_{P0}$ has relatively larger error. This can be expected, because relatively larger noise-to-signal ratio is in the upper region of the profile, which has an important contribution to estimate $N_{P0}$. The error information of the estimated parameters is derived from the covariance matrix calculated following Eq. (4.18). The estimated standard deviation $\widehat{\sigma}_\beta$, together with the relative standard deviation, defined as $\widehat{\sigma}_\beta/\widehat{\beta}$, is also illustrated in Table 6.6 for all seven unknown parameters. It can be seen that the parameters of the F2 layer have the highest relative precisions, whereas $N_{P0}$ has the largest uncertainty. This is again because that the noise-to-signal ratio of the input data in the F2 layer is the smallest and that in the upper part of the profile is the largest, with the mean of the absolute values of the noise-to-signal ratios of the input data in the height range of [600, 1200] km up to 43.3%. In can be seen from error intervals, defined by $\widehat{\beta} \pm \widehat{\sigma}_\beta$, that the "true" values for all parameters except $hmF2$ fall within them. The "true" value of $hmF2$ is slightly outside one-$\sigma$ interval.

Table 6.6: Comparisons of $\widehat{\beta}$, $\beta_{true}$, $\delta_\beta$ [%] together with the error information $\widehat{\sigma}_\beta$, $\widehat{\sigma}_\beta/\widehat{\beta}$ [%] from the simulated data (with F1 layer) at $\epsilon = 5\%$.

| | $NmE$ [el/cm$^3$] | $hmE$ [km] | $NmF1$ [el/cm$^3$] | $NmF2$ [el/cm$^3$] | $hmF2$ [km] | $HF2$ [km] | $N_{P0}$ [el/cm$^3$] |
|---|---|---|---|---|---|---|---|
| $\widehat{\beta}$ | $1.475 \cdot 10^5$ | 110.08 | $2.755 \cdot 10^5$ | $6.638 \cdot 10^5$ | 384.59 | 64.88 | $7.128 \cdot 10^3$ |
| $\beta_{true}$ | $1.478 \cdot 10^5$ | 110.00 | $2.765 \cdot 10^5$ | $6.644 \cdot 10^5$ | 385.00 | 65.00 | $6.644 \cdot 10^3$ |
| $\delta_\beta$ [%] | 0.24 | 0.07 | 0.33 | 0.09 | 0.11 | 0.18 | 7.28 |
| $\widehat{\sigma}_\beta$ | $2.597 \cdot 10^3$ | 0.63 | $2.106 \cdot 10^3$ | $1.597 \cdot 10^3$ | 0.29 | 0.22 | 830.09 |
| $\widehat{\sigma}_\beta/\widehat{\beta}$ [%] | 1.76 | 0.57 | 0.76 | 0.24 | 0.08 | 0.35 | 11.65 |
| Error interval | $[1.449 \cdot 10^5, 1.501 \cdot 10^5]$ | [109.45, 110.71] | $[2.734 \cdot 10^5, 2.776 \cdot 10^5]$ | $[6.622 \cdot 10^5, 6.654 \cdot 10^5]$ | [384.30, 384.88] | [64.66, 65.10] | $[6.298 \cdot 10^3, 7.958 \cdot 10^3]$ |

The correlation matrix calculated based on the Eqs. (4.21) and (4.22) is shown in Fig. 6.21. As can be seen, the parameter pair $NmF2$ and $HF2$ as well as $HF2$ and $N_{P0}$ have moderate

negative correlations, both with a value of about -0.5. The most positive correlation appears between $NmF1$ and $hmF2$, with the value of about 0.5. This could explain why the estimated $hmF2$ is not within one-$\sigma$ interval.

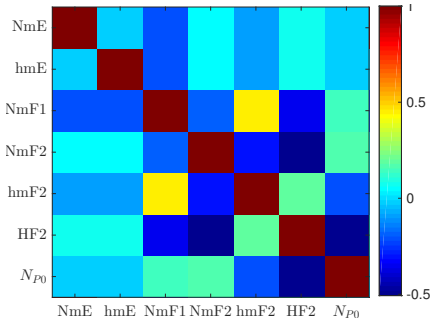



Figure 6.21: Correlation matrix of the seven key parameters from the simulated data (with F1 layer) in Fig. 6.20(a).

**Hypothesis testing**

A hypothesis testing is carried out to check if the difference of the parameter $\beta$ for $hmF2$ and the corresponding "true" value $\beta_{true}$ is statistically different form zero, i.e., $H_0 : \beta = \beta_{true}$ versus $H_1 : \beta \neq \beta_{true}$ for $hmF2$. Following Eq. (4.56), the test statistic $T = \frac{(384.59-385.00)^2}{0.29^2} \approx 2.0$. At the significance level $\gamma = 0.05$, the critical value of the $F$-distribution is equal to 3.9. According to Eq. (4.58), we fail to reject the null hypothesis because $2.0 < 3.9$. Then, the test statistics are calculated for any pair in the parameter set based on Eq. (4.60), for the null hypothesis $H_0 : \beta_i = \beta^i_{true}, \beta_j = \beta^j_{true}$ with $i, j \in \{1, \ldots, 7\}$ $(i \neq j)$ and the alternative hypothesis $H_1 : \beta \neq \beta_{true}$ with at least one parameter in $(i, j)$. The largest test statistic is $T = 1.4$ corresponding to the parameters $NmF2$ and $hmF2$. At the significance level $\gamma = 0.05$, the critical value of the $F$-distribution is equal to 3.0. Therefore, we do not reject the null hypothesis following Eq. (4.63). Afterwards, the statistic tests are calculated for any combination of three parameters, where the largest statistic test value is compared with the critical value; this procedure is performed till all seven parameters are jointly tested. The final results indicate that model parameters are consistent with the "truth".

**Convergence test**

For nonlinear problems, there might exist multiple minima. In order to check stability of the adjustment system, the parameter estimation is performed with changing initial values. Figure 6.22 shows comparisons of the parameter values from 30 runs with different initial values, where the red line indicates $\beta_{true}$, the blue dots are $\widehat{\beta}$ and the green dots represent $\beta_0$. As can be seen, the estimations follow the "true" values during 30 runs. This might reflect that a global minimum (i.e., the results are regardless of initial values) is achieved.

**Performance at different noise levels**

To have a quantitative comparison of parameter estimations at different noise levels, $\sigma_e$, $\widehat{\sigma}_e$ and the relative standard deviations $\widehat{\sigma}_\beta/\widehat{\beta}$ are shown in Table 6.7, at $\epsilon = 2\%$, $\epsilon = 5\%$ and $\epsilon = 10\%$. It is clearly visible that $\widehat{\sigma}_e$ and $\sigma_e$ fit well with each other at all different noise levels. Furthermore,

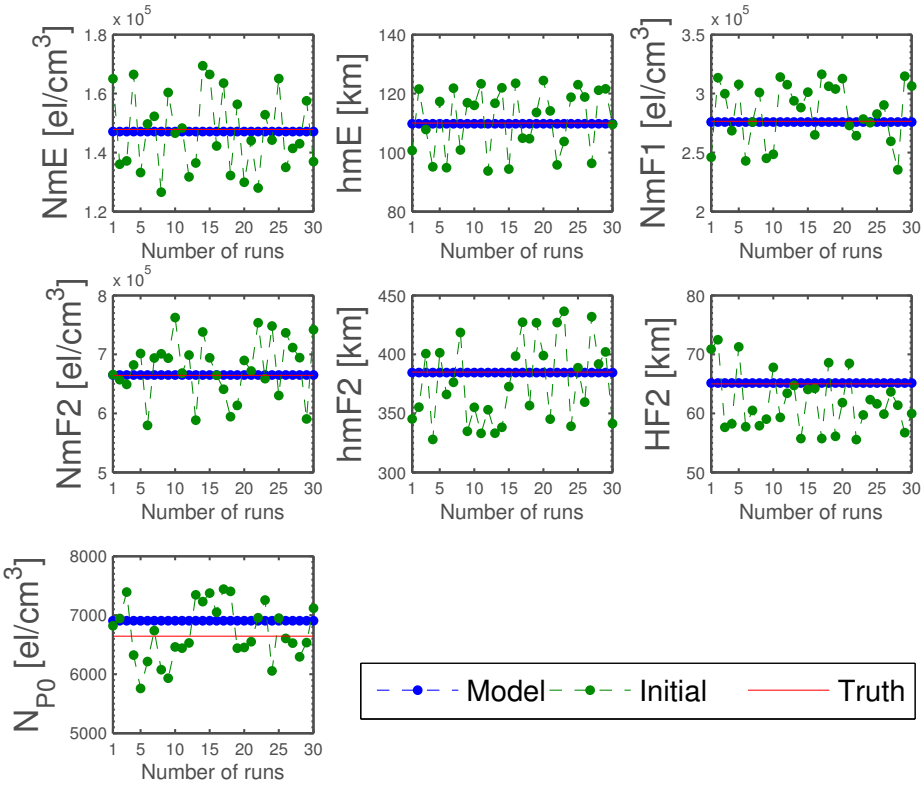



Figure 6.22: Estimates $\widehat{\beta}$ from the simulated data (with F1 layer) at $\epsilon = 5\%$ during 30 runs with different initial values.

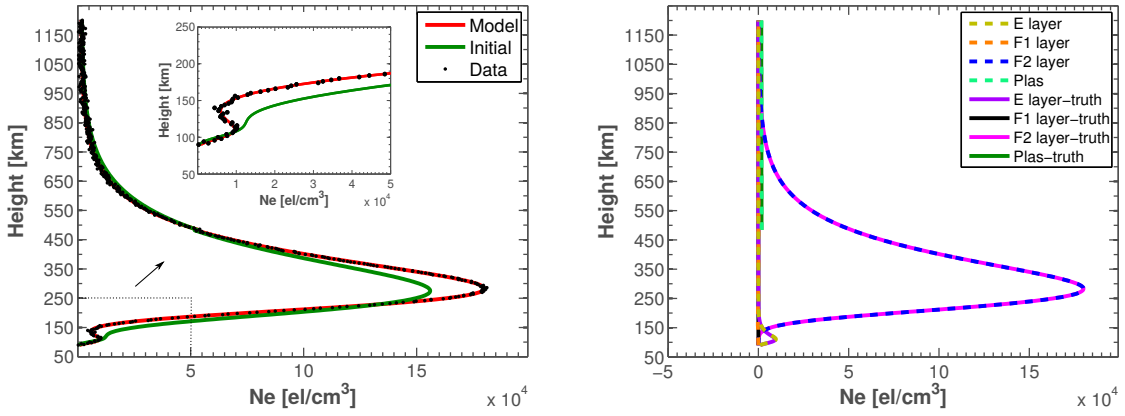
the relative standard deviations of all parameters increase with the noise levels, indicating larger uncertainties.

Table 6.7: Comparisons of the error information of the estimates from the simulated data (with F1 layer) at different noise levels.

| Noise level | $\sigma_e$ | $\widehat{\sigma}_e$ | $\widehat{\sigma}_\beta/\widehat{\beta}$ [%] | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | $NmE$ | $hmE$ | $NmF1$ | $NmF2$ | $hmF2$ | $HF2$ | $N_{P0}$ |
| $\epsilon = 2\%$ | $4.34 \cdot 10^3$ | $4.30 \cdot 10^3$ | 0.65 | 0.22 | 0.29 | 0.09 | 0.03 | 0.13 | 4.78 |
| $\epsilon = 5\%$ | $1.13 \cdot 10^4$ | $1.13 \cdot 10^4$ | 1.76 | 0.57 | 0.76 | 0.24 | 0.08 | 0.35 | 11.65 |
| $\epsilon = 10\%$ | $2.17 \cdot 10^4$ | $2.15 \cdot 10^4$ | 3.08 | 1.17 | 1.42 | 0.46 | 0.14 | 0.66 | 24.84 |

#### 6.2.3.2 Scenario 2 – without F1 layer

In this scenario, nighttime EDPs without F1 layer are simulated. Figure 6.23 shows one possible result obtained from a simulated profile at $\epsilon = 2\%$. As can be seen, both the estimated total profile and individual layers fit to the simulated data well. The standard deviations $\widehat{\sigma}_e$ and $\sigma_e$ are both about $7.79 \cdot 10^2$ el/cm$^3$. The corresponding $\widehat{\beta}$, $\beta_{true}$, $\delta_\beta$ and their error information are displayed in Table 6.8. It can be seen that the relative errors of the estimated parameters for the




(a) (b)

Figure 6.23: An example of profile fitting (unreasonable estimations with negative $NmF1$) from the simulated data (without F1 layer) at $\epsilon = 2\%$ (a); separation into distinct layers (b).

F2 layer are close to the "true" values and high precisions are obtained. Instead of $N_{P0}$, $NmE$ has the largest relative error, since the effective region that determines $NmE$ is located in the smallest electron density part of the profile for this scenario (see Fig. 6.23(a)). It is also noticed that $\widehat{NmF1}$ is negative which is physically not meaningful. Since its "true" value 0 is within the error interval, the estimated $NmF1$ is not significant and $NmF1$ can be set to zero.

Table 6.8: Comparisons of $\widehat{\beta}$, $\beta_{true}$ and $\delta_\beta$, together with the error information from the simulated data (without F1 layer) at $\epsilon = 2\%$.

| | $NmE$ [el/cm$^3$] | $hmE$ [km] | $NmF1$ [el/cm$^3$] | $NmF2$ [el/cm$^3$] | $hmF2$ [km] | $HF2$ [km] | $N_{P0}$ [el/cm$^3$] |
|---|---|---|---|---|---|---|---|
| $\widehat{\beta}$ | $9.70 \cdot 10^3$ | 110.50 | -300.66 | $1.80 \cdot 10^5$ | 282.80 | 57.96 | $1.79 \cdot 10^3$ |
| $\beta_{true}$ | $9.31 \cdot 10^3$ | 110.00 | 0 | $1.80 \cdot 10^5$ | 283.00 | 58.00 | $1.80 \cdot 10^3$ |
| $\delta_\beta$ [%] | 4.14 | 0.45 | – | 0.07 | 0.07 | 0.07 | 0.71 |
| $\widehat{\sigma}_\beta$ | $3.40 \cdot 10^2$ | 0.47 | 331.08 | 168.51 | 0.11 | 0.06 | 50.15 |
| $\widehat{\sigma}_\beta/\widehat{\beta}$ [%] | 3.51 | 0.43 | -110.12 | 0.09 | 0.04 | 0.10 | 2.81 |
| Error interval | [$9.36 \cdot 10^3$, $1.00 \cdot 10^4$] | [110.03, 110.97] | [-631.74, 30.42] | [$1.798 \cdot 10^5$, $1.802 \cdot 10^5$] | [282.69, 282.91] | [57.90, 58.02] | [$1.740 \cdot 10^3$, $1.840 \cdot 10^3$] |

The corresponding correlation matrix is shown in Fig. 6.24. As can be seen, more parameters are correlated compared to Fig. 6.21. The peak parameters of the neighboring ionosphere layers tend to have strong negative correlation. In particular, $NmE$ and $NmF1$ have a negative correlation of about $-0.8$. $NmF1$ and $NmF2$ have a correlation of about $-0.7$. The peak height $hmF2$ has also strong correlations with other parameters. A negative correlation appears between $hmF2$ and $NmF2$, with a value of about $-0.7$. A positive large correlation exists between $hmF2$ and $NmF1$, with a value of about 0.8.

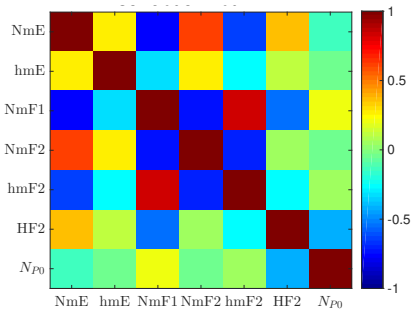
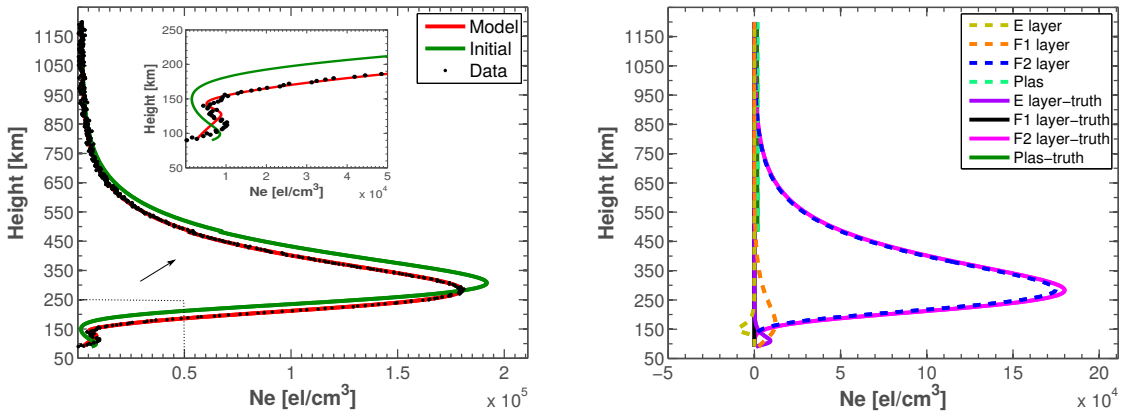



Figure 6.24: Correlation matrix of the seven key parameters obtained from the simulated data (without F1 layer) in Fig. 6.23(a).

For this scenario, however, iteratively least squares algorithm is not stable anymore, even at the relatively low noise level $\epsilon = 2\%$. It may yield unrealistic parameter estimates, in particular for $NmE$, $hmE$ and $NmF1$. Figure 6.25 shows one of the possible results. Generally, most part of the total fitted profile seem to be fine. However, if we look at the zoomed figure of the lower part shown in the panel (a), it can be noticed that the estimated profile does not actually fit to the observations. If we look further at the separate layers shown in the panel (b), it can be seen that the estimated E and F1 layers are far from the "true" layers. The estimated $NmE$ is equal to $-9.27 \cdot 10^3$, which is negative; the estimated $NmF1$ is equal to $1.23 \cdot 10^4$, which is much larger than the "true" value of zero.




(a) (b)

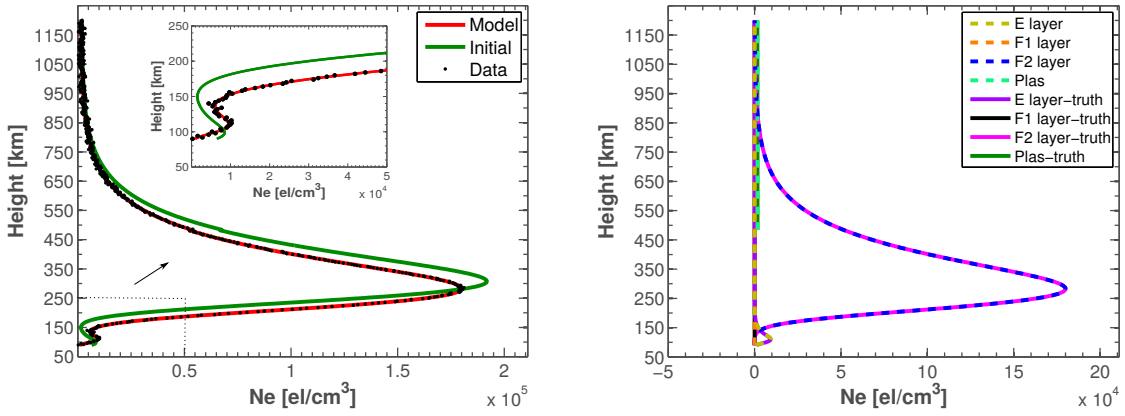
Figure 6.25: An example of profile fitting (unreasonable estimations with negative $NmE$ and large value of $NmF1$) from the simulated data (without F1 layer) at the noise level $\epsilon = 2\%$ (a); separation into distinct layers (b).

In order to obtain realistic results, inequality constraints must be introduced. Within this thesis inequality constraints are put on the three parameters $NmE$, $hmE$ and $NmF1$. According to Section 2.2, the E layer is weakly ionized at night, with electron density values about $5 \cdot 10^3$ el/cm$^3$. The height range of $hmE$ is about 90 km to 140 km (see Table 2.1). As also introduced previously, there is not an unified designation of altitude ranges for different layers because the real heights of the ionosphere layers vary with many factors such as season, solar

activity and so on. Here we set a looser range, namely,

$$NmE \geq 10^3, \quad 70 \leq hmE \leq 160 \quad \text{and} \quad NmF1 \geq 0\,. \tag{6.22}$$

The problem is then mathematically attributed to optimization problems with inequality constraints. Figure 6.26 shows the results from the same input data and the same initial vector $\boldsymbol{\beta}_0$ as




(a) (b)

Figure 6.26: An example of profile fitting (reasonable estimations) from the same input data as in Fig. 6.25 by using optimization method (a); separation into distinct layers (b).

in Fig. 6.25 by using the method introduced in Section 4.7. As can been seen, the estimated total profile fits to the observations well, and the estimated E layer is consistent with the simulated "truth". The estimated $\widehat{\beta}$ and $\delta_\beta$ is shown in Table 6.9. As can be seen, unrealistic parameter estimates are avoided.

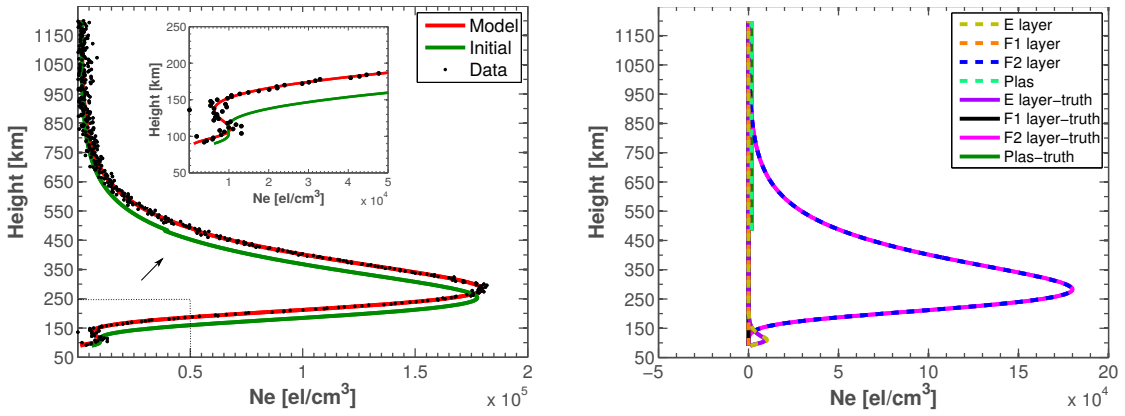
Table 6.9: Comparisons of $\widehat{\beta}$, $\beta_{true}$ and $\delta_\beta$ from the simulated data (without F1 layer) at $\epsilon = 2\%$ using optimization method.

| | $NmE$ [el/cm$^3$] | $hmE$ [km] | $NmF1$ [el/cm$^3$] | $NmF2$ [el/cm$^3$] | $hmF2$ [km] | $HF2$ [km] | $N_{P0}$ [el/cm$^3$] |
|---|---|---|---|---|---|---|---|
| $\widehat{\beta}$ | $9.45 \cdot 10^3$ | 110.35 | 0 | $1.80 \cdot 10^5$ | 282.88 | 57.93 | $1.80 \cdot 10^3$ |
| $\delta_\beta$ [%] | 1.55 | 0.32 | – | 0.00 | 0.04 | 0.11 | 0.18 |

Figure 6.27 shows one example of the final estimated profile from the simulated data at increasing noise level of $\epsilon = 5\%$ using optimization method. As can be seen, the estimated profile also fits to the observations and the various layers can be separated correctly. The corresponding parameter information are illustrated in Table 6.10, where the parameters of the F2 layer have the best estimations and $NmE$ has the largest relative error.

**Stochastic information**

For an inequality constrained optimization implementation, no analytical relationship is present between observations and the unknown parameters. Therefore, we cannot calculate the covariance matrix based on Eq. (4.18) following the law of error propagation. Furthermore, symmetric




(a) (b)

Figure 6.27: An example of profile fitting using optimization algorithm at $\epsilon = 5\%$ (a); separation into distinct layers (b).

Table 6.10: Comparisons of $\widehat{\beta}$, $\beta_{true}$ and $\delta_\beta$ from the simulated data (without F1 layer) at $\epsilon = 5\%$ using optimization method.

| | $NmE$ [el/cm$^3$] | $hmE$ [km] | $NmF1$ [el/cm$^3$] | $NmF2$ [el/cm$^3$] | $hmF2$ [km] | $HF2$ [km] | $N_{P0}$ [el/cm$^3$] |
|---|---|---|---|---|---|---|---|
| $\widehat{\beta}$ | $1.02 \cdot 10^4$ | 109.87 | 0 | $1.80 \cdot 10^5$ | 282.99 | 58.02 | $1.80 \cdot 10^3$ |
| $\beta_{true}$ | $9.31 \cdot 10^3$ | 110.00 | 0 | $1.80 \cdot 10^5$ | 283.00 | 58.00 | $1.80 \cdot 10^3$ |
| $\delta_\beta$ [%] | 9.86 | 0.12 | – | 0.13 | 0.00 | 0.04 | 0.04 |

interval about the estimates is not appropriate to describe the uncertainty anymore, because a parameter space may be truncated by inequality constraints. In order to describe the quality of the estimates of an inequality constrained problem using the optimization method, confidence regions[1] can be applied. For this purpose, the Probability Density Function (PDF) of the estimated parameters must be known [Roese-Koerner *et al.*, 2012].

Let $Z$ be a continuous random variable with values $z \in \mathbb{R}$, the Cumulative Distribution Function (CDF), denoted as $F(z)$, is defined as

$$F(z) = P(Z < z) \tag{6.23}$$

where $P(Z < z)$ denotes the probability for $Z < z$. Let $F(z)$ be continuously differentiable, the PDF, denoted as $p(z)$, is defined as

$$p(z) = dF(z)/dz \tag{6.24}$$

and the CDF can be written as

$$F(z) = \int_{-\infty}^{z} p(z')\,dz' \quad \text{with} \quad p(z) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} p(z')\,dz' = 1 \tag{6.25}$$

where $z'$ is a variable of integration.

[1] A confidence region is a subspace of the parameter space in which the parameter vector $\boldsymbol{z}$ is located with a given probability (e.g., 95%).

For an $n$-D continuous random variable $(Z_1, \ldots, Z_n)$ that take on the values $z_1, \ldots, z_n \in \mathbb{R}$, the joint CDF, denoted as $F(z_1, \ldots, z_n)$, of the multi-dimensional variable is defined as

$$F(z_1, \ldots, z_n) = P(Z_1 < z_1, \ldots, Z_n < z_n) \tag{6.26}$$

where $P(Z_1 < z_1, \ldots, Z_n < z_n)$ denotes the probability for $Z_1 < z_1, \ldots, Z_n < z_n$. The corresponding joint PDF is formulated as

$$p(z1, \ldots, z_n) = \partial^n F(z_1, \ldots, z_n)/\partial z_1 \ldots \partial z_n \,. \tag{6.27}$$

Similar to Eq. (6.25), it fulfills that

$$\begin{aligned} F(z_1, \ldots, z_n) &= \int_{-\infty}^{z_n} \ldots \int_{-\infty}^{z_1} p(z'_1, \ldots, z'_n)\, dz'_1 \ldots dz'_n \\ \text{with} \quad p(z_1, \ldots, z_n) \geq 0 \quad &\text{and} \quad \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(z'_1, \ldots, z'_n)\, dz'_1 \ldots dz'_n = 1 \end{aligned} \tag{6.28}$$

where $z'_1, \ldots, z'_n$ denote the variables of integration. For a certain $Z_i$, the marginal distribution function $q(z_i)$ can be defined as

$$q(z_i) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(z'_1, \ldots, z'_i, \ldots, z'_n)\, dz'_1 \ldots dz'_{i-1}\, dz'_{i+1} \ldots dz'_n \tag{6.29}$$

which is only dependent on $z_i$, since it is obtained by integrating the joint PDF with respect to all variables except $z_i$.

In order to obtain an approximation of the PDF of the model parameters, a Monte Carlo method can be applied. As we work on simulated data, the probability distribution of the observations is directly known, i.e., a normal distribution as in Eq. (6.19). Using Monte Carlo method, the observations are simulated $M$ times where $M$ refers to the number of Monte Carlo iterations. In particular, $M$ samples $\boldsymbol{y}^m$ $(m = 1, \ldots, M)$ are drawn randomly and independently from the normal distribution $\mathcal{N}(\boldsymbol{y}_{true}, \sigma_e^2\, \mathbf{I})$ at a certain noise level $\epsilon$ (2%, 5% or 10%, cf. Eq. (6.18)). The unknown parameters are then determined from each of the $M$ samples via the inequality constrained optimization algorithm. It will finally yield $M$ independent estimates $\widehat{\boldsymbol{\beta}}^m$. The histogram of the parameters can then be computed and normalized (e.g., through dividing the counts in each bin[1] by the total number of samples times the width of bin), and it can be regarded as a discrete approximation of the joint PDF of the parameters. The marginal PDF can be approximated by adding up the corresponding rows of the hypermatrix (7-D matrix) of the multi-dimensional histogram. If the number $M$ is selected sufficiently large, the histogram of the parameters will be adequately representative for the PDF of the parameters. However, it will be computationally more expensive. Within this thesis, $M$ = 10,000 is chosen to keep the accuracy and computation time acceptable. Furthermore, from the second Monte Carlo iteration on, the initial values can be identified with the estimations from the previous iteration, in order to speed up the convergence of the optimization algorithm, similar to the strategy used by Roese-Koerner *et al.* [2012, 2015]. Figure 6.28 depicts the estimated marginal distribution of a parameter $\beta \in \boldsymbol{\beta}$ from $M$ = 10,000 Monte Carlo iterations with observations at noise levels $\epsilon = 2\%$ (left), $\epsilon = 5\%$ (mid) and $\epsilon = 10\%$ (right). As can be seen, all parameters except $NmF1$ tend to be normal distributed. The distributions of $NmF1$ are skewed right, which results from the constraint of the zero lower bound, see Eq. (6.22). As expected, the distributions of all parameters have wider range with respect to the x-axis with increasing noise levels. Based on $\widehat{\boldsymbol{\beta}}^m$, some statistics are computed and illustrated in Table 6.11 with different noise levels. It can

[1] Bins are the intervals that are taken to count how many values fall into when constructing a histogram.

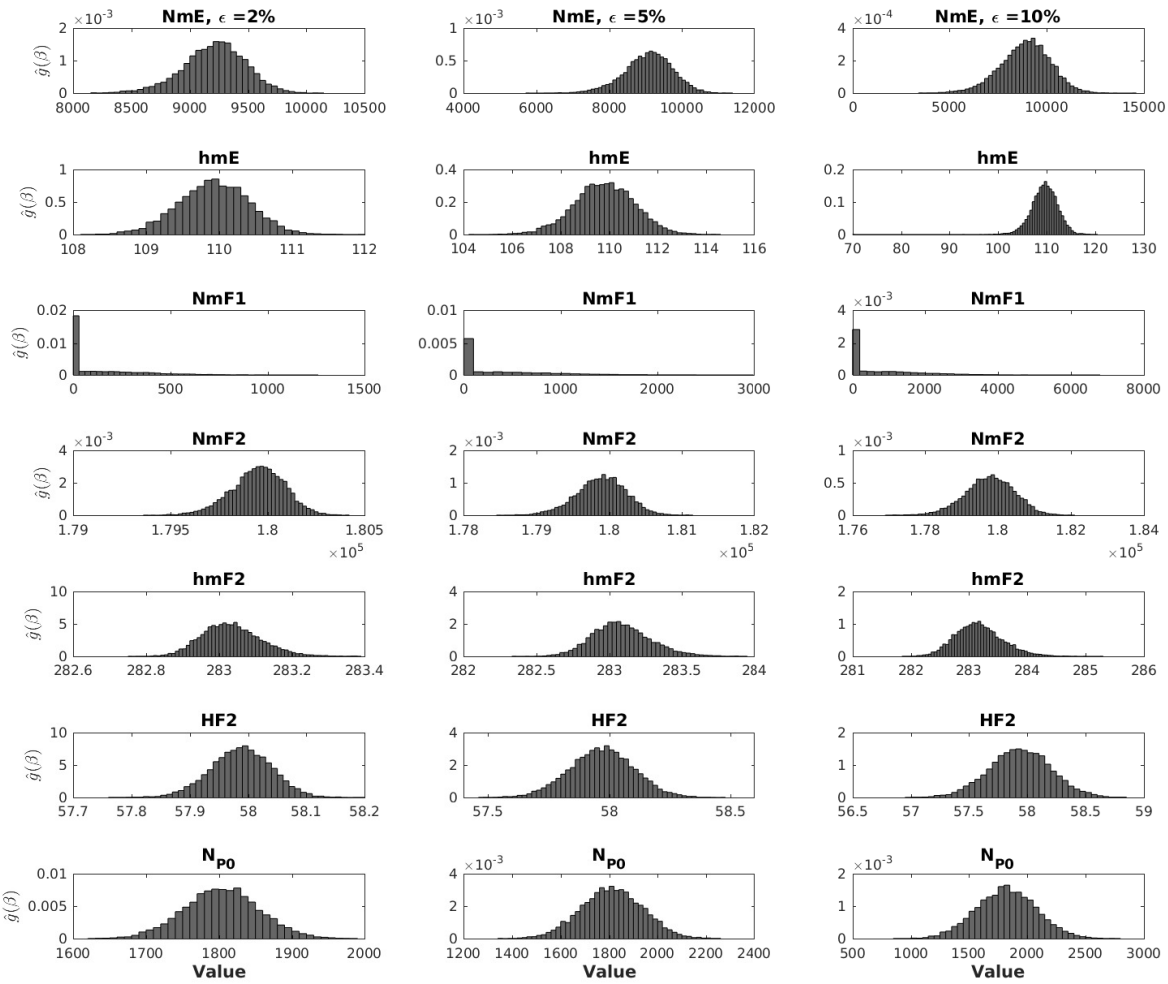



Figure 6.28: Estimated marginal PDF of $\beta$ from $M$ = 10,000 Monte Carlo realizations with different noise levels.

be seen from the minimum and maximum values of the estimates that, the estimation of $hmE$ may go to the lower bound of 70 km at the highest noise level $\epsilon = 10\%$.

Now one may think what the optimal estimate is. Based on PDF, at least four different estimates, i.e., the mean, the median, the mode and the point that minimizes the original inequality constrained problem (i.e., the established problem before starting the Monte Carlo method), can be used (cf. Zhu *et al.*, 2005). The mean and the median coincide with each other in case of a symmetric density function. This has been reflected in Table 6.11, where the mean and the median of all parameters except $NmF1$ fit to each other well, since the corresponding marginal distributions in Fig. 6.28 are almost symmetric. If the symmetric density function has a single mode, the mean, the median, and the mode are all identical. Following Roese-Koerner *et al.* [2012], mean and mode are inappropriate when inequality constraints are introduced. The solution that solves original inequality constrained problem is used, e.g., the solution shown in Fig. 6.26 for the noise level $\epsilon = 2\%$ or Fig. 6.27 for the noise level $\epsilon = 5\%$. The estimate is optimal in the sense that it minimizes the sum of squared residuals in the feasible region.

The parameters are correlated, as shown in Fig. 6.24. Based on the joint PDF of any pair of

parameters, we can obtain the corresponding correlations. Figure 6.29 shows exemplarily the estimated joint PDF for parameter pairs $NmE$ and $NmF1$ (a), $NmF2$ and $NmF1$ (b), $NmF1$ and $hmF2$ (c) and $NmF2$ and $hmF2$ (d) at $\epsilon = 2\%$ (left), $\epsilon = 5\%$ (mid) and $\epsilon = 10\%$ (right). As it should be, the correlation does not change with the noise levels. It can also be seen that negative correlations appear between $NmE$ and $NmF1$ (computed correlation coefficient: -0.6), between $NmF1$ and $NmF2$ (computed correlation coefficient: -0.5), and between $NmF2$ and $hmF2$ (computed correlation coefficient: -0.5), because the trend of the corresponding plot is down; a positive correlation between $NmF1$ and $hmF2$ (computed correlation coefficient: 0.6) is also visible, since the trend of the plot is up. These relationships are consistent with the ones in Fig. 6.24.

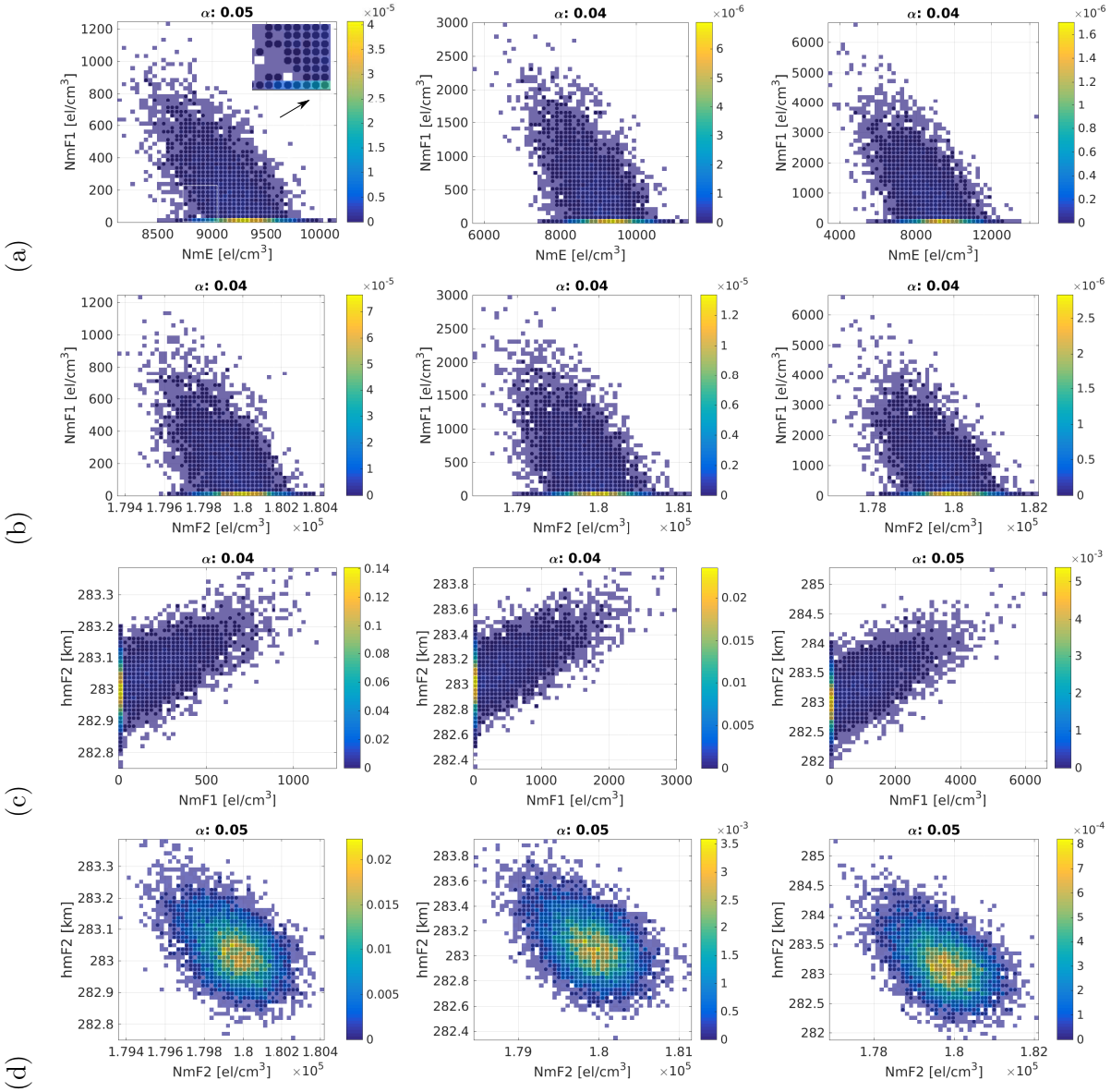



Figure 6.29: Estimated joint PDF of parameter pairs $NmE$–$NmF1$ (a), $NmF2$–$NmF1$ (b), $NmF1$–$hmF2$ (c), and $NmF2$–$hmF2$ (d) from $M = 10{,}000$ Monte Carlo realizations with different noise levels.

As stated previously, the standard deviation (i.e., the second central moment) is not appropriate to describe the uncertainty for truncated PDF, which is the case here for $NmF1$. Following Roese-Koerner *et al.* [2012, 2015], the concept that is called the Highest Posterior Density (HPD) region is appropriate for obtaining the confidence interval of the estimates from Monte Carlo method. The HPD region is defined as the region $\Theta$ including $1-\gamma$ percent of the samples, i.e.,

$$P(\boldsymbol{\beta} \in \Theta \,|\, \boldsymbol{y}) = 1 - \gamma \tag{6.30}$$

with $1-\gamma$ the level of significance. $P(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ is the conditional probability that takes the observations into account. The HPD region for a 1-D case can be obtained by sorting the estimates from the Monte Carlo method and removing the smallest and largest $\gamma/2$ percent of the samples. For a multi-dimensional problem, the HPD region can be obtained in a similar manner. Instead of sorting the estimated values, the values of the multi-dimensional histogram are sorted from the largest to the smallest ones. Then the cumulative sum is calculated until $1-\gamma$ is achieved, and the confidence region is derived from all bins of the histogram that form the probability of $1-\gamma$. The confidence regions are computed for all the panels in Fig. 6.29, where the bins in the confidence region are marked with dots and the corresponding $\gamma$ values are illustrated in the title. As can be seen, most of the bins fall within the confidence region, and the bins that are outside the confidence interval are almost on the boundary of the joint PDF.

Table 6.11: Statistics of the estimates from $M = 10{,}000$ Monte Carlo iterations with different noise levels.

| $\epsilon$ | | 2% | 5% | 10% |
|---|---|---|---|---|
| $NmE$ [el/cm$^3$] true: $9.31 \cdot 10^3$ | min | $8.15 \cdot 10^3$ | $5.78 \cdot 10^3$ | $3.53 \cdot 10^3$ |
| | max | $1.01 \cdot 10^4$ | $1.13 \cdot 10^4$ | $1.44 \cdot 10^4$ |
| | range | $1.99 \cdot 10^3$ | $5.54 \cdot 10^3$ | $1.09 \cdot 10^4$ |
| | mean | $9.21 \cdot 10^3$ | $9.07 \cdot 10^3$ | $8.89 \cdot 10^3$ |
| | median | $9.22 \cdot 10^3$ | $9.11 \cdot 10^3$ | $8.96 \cdot 10^3$ |
| $hmE$ [km] true: 110.00 | min | 108.17 | 104.29 | 70.00 |
| | max | 111.91 | 114.40 | 120.21 |
| | range | 3.74 | 10.11 | 50.21 |
| | mean | 109.93 | 109.79 | 109.56 |
| | median | 109.93 | 109.80 | 109.64 |
| $NmF1$ [el/cm$^3$] true: 0 | min | 0 | 0 | 0 |
| | max | $1.24 \cdot 10^3$ | $2.95 \cdot 10^3$ | $6.61 \cdot 10^3$ |
| | range | $1.24 \cdot 10^3$ | $2.95 \cdot 10^3$ | $6.61 \cdot 10^3$ |
| | mean | 130.06 | 309.60 | 628.58 |
| | median | 0 | 0 | 0 |
| $NmF2$ [el/cm$^3$] true: $1.80 \cdot 10^5$ | min | $1.79 \cdot 10^5$ | $1.78 \cdot 10^5$ | $1.77 \cdot 10^5$ |
| | max | $1.80 \cdot 10^5$ | $1.81 \cdot 10^5$ | $1.82 \cdot 10^5$ |
| | range | $1.05 \cdot 10^3$ | $2.69 \cdot 10^3$ | $5.12 \cdot 10^3$ |
| | mean | $1.80 \cdot 10^5$ | $1.80 \cdot 10^5$ | $1.80 \cdot 10^5$ |
| | median | $1.80 \cdot 10^5$ | $1.80 \cdot 10^5$ | $1.80 \cdot 10^5$ |
| $hmF2$ [km] true: 283.00 | min | 282.76 | 282.35 | 281.89 |
| | max | 283.38 | 283.93 | 285.28 |
| | range | 0.62 | 1.58 | 3.39 |
| | mean | 283.03 | 283.09 | 283.17 |
| | median | 283.03 | 283.07 | 283.15 |
| $HF2$ [km] true: 58.00 | min | 57.76 | 57.47 | 56.97 |
| | max | 58.19 | 58.47 | 58.84 |
| | range | 0.43 | 1.00 | 1.87 |
| | mean | 57.99 | 57.97 | 57.94 |
| | median | 57.99 | 57.97 | 57.94 |
| $N_{P0}$ [el/cm$^3$] true: $1.80 \cdot 10^3$ | min | $1.63 \cdot 10^3$ | $1.34 \cdot 10^3$ | $8.87 \cdot 10^2$ |
| | max | $1.99 \cdot 10^3$ | $2.25 \cdot 10^3$ | $2.76 \cdot 10^3$ |
| | range | $3.61 \cdot 10^2$ | $9.12 \cdot 10^2$ | $1.87 \cdot 10^3$ |
| | mean | $1.80 \cdot 10^3$ | $1.81 \cdot 10^3$ | $1.82 \cdot 10^3$ |
| | median | $1.80 \cdot 10^3$ | $1.81 \cdot 10^3$ | $1.82 \cdot 10^3$ |

# Chapter 7

# Summary, Conclusion and Outlook

A regional physics-motivated 4-D modeling concept of the electron density has been derived and investigated within this thesis. Various methods have been tested, combined and adapted in the developed 4-D ionosphere modeling approach. It considers the physics-motivated Chapman function for the electron density distribution of ionosphere layers and a slowly decaying exponential term for the plasmaspheric part. Due to the ability for regional modeling, series expansions in compactly supported multi-dimensional B-spline functions based on tensor products are applied to model the spatiotemporal variations of the ionosphere key parameters. For each space-geodetic observation technique, linearized observation equation system together with its stochastic information is set up within the Gauss-Markov model. The combination of different space-geodetic observation techniques is performed within a joint adjustment system including prior information for handling data gaps. The individual weighting of each observation technique is considered by applying the method of VCE where the unknown B-spline coefficients are solved. Benefiting from the B-spline functions, the modeling approach allows transforming the model results into a MSR and further the application of the MSR for data compression. Furthermore, an inequality constrained optimization method has been considered partly in this work in order to avoid physically unreasonable parameter estimates. The approach of quantifying the uncertainty of the inequality constrained estimates has also been studied.

Some issues of the 4-D electron density modeling concept have been investigated numerically in detail through two different scenarios.

**F2-layer model**
Firstly, the most important ionosphere layer, namely, the F2 layer, is considered and modeled by the Chapman function. The characteristics of the F2 layer, namely, the peak density $NmF2$, the peak height $hmF2$ and the Chapman scale height $HF2$ play essential roles in ionosphere dynamics. Each of these three parameters has been modeled by a series expansion as tensor products of three 1-D polynomial B-spline basis functions depending on longitude, latitude and time with initially unknown series coefficients. The developed model is applied to two types of GPS data, namely, ground-based dual-frequency GPS observations of STEC and EDPs retrieved from ionospheric GPS RO measurements acquired by the F3/C mission. A South American region has been selected as the study area where the EIA is included. Although two observation techniques have been combined, there are still data gaps, which have caused that some coefficients are not supported and thus resulted in an ill-conditioned and even singular stacked normal

equation system. To overcome data gaps, prior information derived from the background model IRI-2012 has been introduced in the adjustment system, i.e., a kind of regularization is performed to stabilize the solutions. Linearized observation equation systems of the two types of GPS data and linearized prior information as well as their stochastic information have been put into a joint adjustment process to solve for the unknown series coefficients as well as by-products such as the satellite and receiver DCBs. The individual weighting of the two GPS observation techniques has been determined by VCE, whereas the weighting of the prior information has been selected empirically. In order to increase the stability of the linearized model, a two-step approach has been performed within the parameter estimation. In particular, only $NmF2$ is estimated firstly, and $hmF2$ and $HF2$ are then estimated jointly. The estimated series coefficients have been then used to obtain the representations of the three key parameters and their precisions have been derived by applying the law of error propagation, and further electron densities as well as VTEC have been constructed everywhere within the study area under the investigated time period.

The developed model has been applied exemplarily on three selected days under low solar activity (July 16, 2008), moderate solar activity (July 16, 2011) and high solar activity (July 1, 2012), respectively. The estimated model parameters shown in Fig. 6.6 have clearly demonstrated that finer structures are obtainable in areas with input data. The results have indicated no significant difference between the model estimations and IRI. The estimated model key parameters have been validated by independent data sources, namely, $NmF2$ from ionosonde data over Jicamarca and Port Stanley, and VTEC maps from CODE Analysis Center of the IGS. A cross-validation by STEC values has also been performed. The comparison of $NmF2$ have shown that a better consistency with ionosonde data is achieved by the developed model compared with IRI. The improvements differ for the two stations and it seems that the most significant improvements can be achieved at Jicamarca under high solar activity. The constructed VTEC map can approach the CODE map better than IRI, and the improvements decrease with decreasing solar activity. As ground-based GNSS data are also the input data for CODE GIM and most GPS observations are over the continents, main improvements have been obtained over the continents. The cross validation with GPS observations has suggested that the STEC residuals have no dependency on UT and solar activity. The developed model has reached a STEC accuracy of about 3–4 TECU. In general, the numerical examples have shown the potential of the derived modeling approach to update the key parameters of IRI in different solar activities, i.e., providing instantaneous behavior of the ionosphere using up-to-date observations.

MSR has been applied to the estimated model results to calculate all scaling and wavelet coefficients of the lower levels using the pyramid algorithm. The MSR for the three key parameters shown in Fig. 6.8 have suggested that the signal structures become coarser with decreasing resolution levels. A simple level-dependent thresholding method is then applied to the wavelet coefficients and a data compression has been demonstrated.

**Multi-layer model**

Various ionospheric layers have been then considered in the modeling approach. The sensitivities of the various ionosphere layers, i.e., the D, E, F1 and F2 layers and the plasmaspheric part to the VTEC till GPS orbit height have been investigated by using simulated EDPs from the combination of IRI and IRI-Plas. The numerical results have clearly demonstrated the significance of the F2 layer, the plasmaspheric part and the F1 layer. As the multi-layer approach significantly increases the number of the unknowns and thus brings great challenge to the parameter estimation, the D layer has been neglected in the multi-layer modeling approach. The separa-

bility of the selected seven parameters ($NmE$, $hmE$, $NmF1$, $NmF2$, $hmF2$, $HF2$, $N_{P0}$) has been investigated by using a closed-loop simulation. In order to have a more realistic stochastic model of the EDPs, the error propagation of the electron density observations through the onion peeling algorithm has been studied, based on the assumption that STEC observables during an occultation event are independent and follow Gaussian distribution. The results have shown that the qualities of the electron densities do not degrade significantly along the profile and no strong height correlations between successive electron densities are introduced. Thus, the identity matrix has been used for the weighting matrix of EDPs. This study gives an understanding how the errors propagate vertically and how correlations are introduced between the electron densities from the onion peeling algorithm.

To investigate the separability of the parameters of the multi-layer approach, two different scenarios have been simulated: daytime EDPs with F1 layer and nighttime EDPs without F1 layer. Additive white Gaussian noises with different noise levels have been put on the simulated "truth". The numerical results in the scenario with F1 layer have shown that the selected seven parameters can be separated within the adjustment system. In particular, the three key parameters of the F2 layer can be best estimated. The calculated correlations between the parameters have shown that moderate correlations exist. For the simulated profiles without F1 layer, correlations between the parameters become stronger and correlations appear between more parameter pairs. The adjustment system has even failed to separate the E and F1 layers: negative value of $NmE$ as well as unreasonable $hmE$ and $NmF1$ appear. As a consequence, inequality constraints have been imposed on these three parameters. Since inequality constraints cannot easily be incorporated into the adjustment system, the problem has been solved by applying a constrained optimization algorithm by which reasonable solutions have been obtained. Using inequality constrained optimization algorithm, it is not possible to obtain the quality of the estimates by calculating the covariance matrix, because the relations between observations and the unknowns cannot be described analytically. In addition, a symmetric standard deviation around estimates is not sufficient to describe the uncertainty, since the parameter space can be truncated by inequality constraints. To obtain the quality information of the estimates, a Monte-Carlo method has been combined with the optimization algorithm, in order to obtain approximations of the PDF of the model parameters. The estimated marginal PDF in Fig. 6.28 has shown that the range of the distribution is larger with increasing noise levels for all parameters. This indicates an increase in uncertainty of the estimates. Furthermore, the PDF of $NmF1$ is accumulated on the boundary because of the placed nonnegative constraint. The joint PDF of parameter pairs has reflected the correlation between the parameters. The derived correlations are consistent with the ones derived from the adjustment system.

## 7.1 Conclusions and future works

Based on the numerical examples presented in this work, the following conclusions have been drawn and future works have been suggested.

#### F2-layer model

- The combination of ground-based GPS data of STEC and EDPs from GPS RO data has made a large contribution to 4-D electron density modeling. The derived modeling ap-

proach enables the electron density to be constructed everywhere over the study area and continuously during the time period under investigation, which is required for almost all geodetic applications. Furthermore, since a physics-motivated function is considered, the model outputs allow some physical interpretation of the ionosphere. This is essential for ionosphere research and for a better understanding of the space weather.

- Improvement of model resolution and accuracy: the resolution of the developed model is depending on the levels of the B-spline functions, whereas the levels have to be chosen according to the average sampling interval of input data. The resolution of the model is mainly limited by the inhomogeneously distributed data distribution. In general, the ground-based GNSS observations allow for a relatively high resolution level over the continent. However, the number of usable EDPs retrieved from F3/C mission is rather limited as many EDPs have been rejected as outlier profiles. Therefore, the chosen B-spline levels for model computation are not very high. The resolution and accuracy of the developed model are expected to be improved when more EDPs are available. With the full constellation, F7/C2 can provide more than 12,000 daily ionosphere profiles. Therefore, it will have great potential to improve the 4-D electron density modeling in near future.

- Regularization: the weighting of the prior information has been chosen empirically. Conventional methods for determining the regularization parameter, such as L-curve and GCV, are recommended. Specifically, during each iteration step of the loop with respect to VCE shown in Fig. 5.8, an optimal regularization parameter is chosen by the conventional methods based on current variance components of the observation techniques. Then, the correction vector of the coefficients is estimated. Finally, variance components of the observation techniques are computed and updated.

- Incorporation of more space-geodetic techniques: the developed model can be easily extended by introducing other observation types (i.e., techniques or missions) by stacking the normal equations of each data type. Combining more observation techniques are absolutely recommended to fully exploit the strengths of various space-geodetic observations and improve resolution and accuracy of the ionosphere model.

- Data compression: the application of data compression has been demonstrated through a simple level-dependent thresholding algorithm. Data compression will be particularly useful for handling the huge ionosphere data sets in terms of reducing storage space, and for real time application in terms of increasing transmission speed. A more reasonable data compression technique should be performed in future. As introduced in Section 5.2.3.1, data compression can be performed statistically by hypothesis testing to separate significant from non-significant model parts.

- Regional densification and a data-adapted model: the presented model results have a unique resolution for the area under investigation, where the resolution is limited by the distribution of the input data. A multi-resolution model, namely, a data-adapted model can be obtained from the derived modeling approach, by means of a combination of a regional model part and a series of densification areas. Specifically, parameter estimation is performed separately for the entire regional part and the densification areas. Then, the pyramid algorithm is applied to the results (scaling coefficients) for all sets of the scaling coefficients to generate the MSR. Afterwards, image fusion algorithm known from digital image processing can be applied to the scaling and wavelet coefficients to merge all relevant information. In this manner, a data-adapted model is achieved by reconstruction of the

signal.

### Multi-layer model

- The multi-layer approach is physically more realistic and should be the main focus in future.
- It is necessary to apply inequality constrained optimization algorithm to the multi-layer modeling approach.
- The Monte Carlo method is computationally time consuming, but it has provided a means to obtain the quality of the estimates.
- The performed investigations can serve as a solid foundation for multi-layer modeling approach. In the next step, the separability of the parameters should be studied by applying real observations. Then, spatiotemporal variations of the key parameters are modeled and finally a physically more realistic 4-D regional electron density model of the ionosphere/plasmasphere is achieved from the combination of various space-geodetic observation techniques.
- Besides space-geodetic observation techniques, ionosonde measurements are suggested to be included, since they will help resolving various layers in the bottomside of the ionosphere.

It can be expected that, the ultimate realization of the physically more realistic 4-D regional electron density model and with more space-geodetic observations available due to, e.g., the full constellation of F7/C2, our knowledge on the ionosphere will be improved significantly in future. This in turn will help to improve a wide range of applications in space geodesy.

# References


Abdel-salam, M.A. (2005). *Precise Point Positioning Using Un-Differenced Code and Carrier Phase Observations*. Master's thesis, University of Calgary, Department of Geomatics Engineering. 34

AIAA (1999). Guide to Reference and Standard Ionosphere Models. *American Institute of Aeronautics and Astronautics, Virginia, USA*. 24

Al-Fanek, O.J.S. (2013). *Ionospheric Imaging for Canadian Polar Regions*. Ph.D. thesis, University of Calgary, Alberta, Canada, Depts. of Geomatics Engineering. 4

Alizadeh, M.M. (2013). *Multi-dimensional modeling of the ionosphere parameters using space geodetic techniques*. Ph.D. thesis, Vienna University of Technology, Faculty of Mathematics and Geoinformation. 4, 23, 66, 70, 84, 99

Allain, D.J. and Mitchell, C.N. (2010). Comparison of 4D tomographic mapping versus thin-shell approximation for ionospheric delay corrections for single-frequency GPS receivers over North America. *GPS Solut.*, **14**, 279–291. 4

Auriol, A. and Tourain, C. (2010). Doris system: The new age. *Adv. Space Res.*, **46(12)**, 1484–1496. 31

Austen, J.R., Franke, S.J. and Liu, C.H. (1988). Ionospheric imaging using computerized tomography. *Radio Sci.*, **23(3)**, 299–307. 65

Azpilicueta, F., Altadill, D., Brunini, C., Torta, J.M. and Blanch, E. (2015). A comparison of the LPIM-COSMIC F2 peak parameters determinations against the IRI(CCIR). *Adv. Space Res.*, **55(8)**, 2012–2019. 99, 101

Barclay, L.W.E. (2003). *Propagation of Radiowaves (2nd Edition)*. The Institution of Engineering and Technology, London. 18

Bartels, J., Heck, N.H. and Johnston, H.F. (1939). The three-hour-range index measuring geomagnetic activity. *J. Geophys. Res.*, **44 (4)**, 411–454. 94

Bassiri, S. and Hajj, G.A. (1993). Higer-Order Ionospheric Effects on the GPS Observable and Means of Modeling Them. *Manuscripta Geodaetica*, **18**, 280–289. 23

Basu, S., Buchau, J., Rich, F.J., Weber, E.J., Field, E.C., Heckscher, J.L., Kossey, P.A., Lewis, E.A., Dandekar, B.S., McNamara, L.F., Cliver, E.W., Millman, G.H., Aarons, J., Klobuchar, J.A. and Mendillo, M.F. (1985). *Chapter 10 Ionospheric radio wave propagation, In: Jursa A.S. (eds.) Handbook of Geophysics and the Space Environment*.

Air Force Geophysics Laboratory, NTIS Document Accession Number ADA 167000, Springfield. 69

Belehaki, A., Marinov, P., Kutiev, I., Jakowski, N. and Stankov, S. (2006). Comparison of the topside ionosphere scale height determined by topside sounders model and bottomside digisonde profiles. *Adv. Space Res.*, **37**, 963–966. 70

Bilitza, D. (1990). International Reference Ionosphere 1990. *NSSDC 90-22, Greenbelt, Maryland*. 24, 25

Bilitza, D. (2001). International Reference Ionosphere 2000. *Radio Sci.*, **36(2)**, 261–275. 24

Bilitza, D. (2002). Ionospheric Models for Radio Propagation Studies. *in Review of Radio Science: 1999-2002, edited by Stone, W. R., John Wiley & Sons*, 625–679. 23, 24, 92

Bilitza, D. and Reinisch, B.W. (2008). International Reference Ionosphere 2007: Improvements and new parameters. *Adv. Space Res.*, **42(4)**, 599–609. 24, 25

Bilitza, D., Eyfrig, R. and Sheikh, N.M. (1979). A global model for the height of the F2-peak using M3000 values from the CCIR numerical map. *ITU Telecommunication Journal*, **46**, 549–553. 70

Bilitza, D., Reinisch, B., Benson, R., Grebowsky, J., Napitashvili, N., Huang, X., Schar, W. and Hills, K. (2003). Online data base of satellite sounder and insitu measurements covering two solar cycles. *Adv. Space Res.*, **31(3)**, 769–774. 2

Bilitza, D., Altadill, D., Zhang, Y., Mertens, C., Truhlik, V., Richards, P., Mckinnell, L.A. and Reinisch, B.W. (2011a). The international Reference Ionosphere 2012 – A MODEL. *J. Geod.*, **85**, 909–920. 4, 24

Bilitza, D., Mckinnell, L.A., Reinisch, B.W. and Fuller-Rowell, T. (2011b). The international reference ionosphere today and in the future. *J. Geod.*, **85**, 909–920. 25

Bilitza, D., Brown, S.A., Wang, M.Y., Souza, J.R. and Roddy, P.A. (2012). Measurements and {IRI} model predictions during the recent solar minimum. *Journal of Atmospheric and Solar-Terrestrial Physics*, **86**, 99 – 106. 101

Bilitza, D., Altadill, D., Zhang, Y., Mertens, C., Truhlik, V., Richards, P., Mckinnell, L.A. and Reinisch, B.W. (2014). The International Reference Ionosphere 2012 – a model of international collaboration. *J. Space Weather Space Clim.*, **4(A07)**. 15, 26

Blewitt, G. (1990). An Automatic Editing Algorithm for GPS Data. *Geophys. Res. Lett.*, **17(3)**, 199–202. 34

Bogg, P.T. and Tolle, J.W. (1996). Sequential quadratic programming. *Acta Numerica*, 1–51. 59, 60, 61

Booker, H.G. (1977). Fitting of multi-region ionospheric profiles of electron density by a single analytic function of height. *J. Atmos. Terr. Phys.*, **39**, 619–623. 66

Bouman, J. (1998). Quality of regularization methods. *DEOS Report no. 98.2, Delft University Press*. 53

Bouman, J. (2000). Quality assessment of satellite-based global gravity field models. *NGG Nederlandse Commissie voor Geodesie Netherlands Geodetic Commission, Delft, The Netherlands*. 53

Bradley, P. and Dudeney, J. (1973). A simple model of the vertical distribution of electron concentration in the ionosphere. *J. Atmos. Sol. Terr. Phys.*, **35**, 2131 – 2146. 69

Brookes, M. (2011). The matrix reference manual, [online] http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html. 50

Brunini, C., Meza, A., Azpilicueta, F., Gende, M., Van Zele, M.A., Gende, M. and Díaz, A. (2004). A New Ionosphere Monitoring Technology Based on GPS. *Astrophys. and Space Sci.*, **290(3)**, 415–429. 4

Brunini, C., Meza, A. and Bosch, W. (2005). Temporal and spatial variability of the bias between TOPEX- and GPS-derived total electron content. *J Geod*, **79**, 175–188. 89

Brunini, C., Azpilicueta, F., Gende, M., Aragón-Ángel, A., Hernández-Pajares, M., Juan, J.M. and Sanz, J. (2011). Toward a sirgas service for mapping the ionosphere's electron density distribution. *International Association of Geodesy Symposia*, **136**, 753–760. 23

Brunini, C., Azpilicueta, F. and Nava, B. (2013a). A technique for routinely updating the ITU-R database using radio occultation electron density profiles. *J Geod*, **87**, 813–823. 101

Brunini, C., Conte, J.F., Azpilicueta, F. and Bilitza, D. (2013b). A different method to update monthly median hmF2 values. *Adv. Space Res.*, **51**, 2322–2332. 66, 70

Budden, K.G. (1985). *The propagation of radio waves*. Cambridge University Press, New York. 21

Bust, G.S. and Mitchell, C.N. (2008). History, current state, and future directions of ionospheric imaging. *Rev. Geophys.*, **46**, RG1003. 1, 24

Cander, L.R., Leitinger, R. and Levy, M.F. (1999). Ionospheric models including the auroral environment. *Workshop on Space Weather. The Netherlands: European Space Agency*, 135–142. 23, 24

Cannon, P., Angling, M., Barclay, L., Curry, C., Dyer, C., Edwards, R., Greene, G., Hapgood, M., Horne, R., Jackson, D., Mitchell, C., Owen, J., Richards, A., Rogers, C., Ryden, K., Saunders, S., Sweeting, M., Tanner, R., Thomson, A. and Underwood, C. (2013). Extreme space weather: impacts on engineered systems and infrastructure. Tech. rep., London. 1

Carpenter, D.L. and Anderson, R.R. (1992). An ISEE/Whistler model of equatorial electron density in the magnetosphere. *J. Geophys. Res.*, **97(A2)**, 1097–1108. 29

Carpenter, D.L. and Park, C.G. (1973). On what ionospheric workers should know about the plasmapause-plasmasphere. *Reviews of Geophysics and Space Physics*, **11**, 133–154. 8

CCIR (1996). CCIR atlas of ionospheric characteristics. *Report 340-1, 340-6, Comité Consultatif International des Radiocommunications, Genève, Switzerland*. 25

Ching, B.K. and Chiu, Y.T. (1973). A Phenomenological Model of Global Ionospheric Electron Density in the E-, F1- and F2-regions. *J. Atmos. Sol. Terr. Phys.*, **35**, 1615–1630. 4, 24

Chiu, Y.T. (1975). An improved phenomenological model of ionospheric density. *J. Atmos. Sol. Terr. Phys.*, **37**, 1563–1570. 24

Coco, D. (1991). GPS–Satellites of Opportunity for Ionospheric Monitoring. *GPS World, October*, 47–50. 93

Coleman, T.F., Branch, M.A. and Grace, A. (1999). Optimization toolbox–for use with matlab, user's guide, 2nd edn. *The MathWorks, Inc., Natick*. 64

Committee on Solar-Terrestrial Research (1969). Physics of the earth in space: The role of ground-based research. Tech. rep., Washington, D. C. 2

CSNO (2013). BeiDou Navigation Satellite System Signal In Space Interface Control Document–Open Service Signal (Version 2.0). *China Satellite Navigation Office (CSNO), December*. 40

CSNO (2016). China's BeiDou Navigation Satellite System. *The State Council Information Office of the PeopleâĂŹs Republic of China (SCIO), June*. 39

Daniell, R.E., Brown, L.D., Anderson, D.N., Fox, M.W., Doherty, P.H., Decker, D.T., Sojka, J.J. and Schunk, R.W. (1995). Parameterized Ionospheric Model: A Global Ionospheric Parameterization Based on First Principle Models. *Radio Sci.*, **30(5)**, 1499–1510. 4, 23

Davies, K. (1965). *Ionospheric Radio Propagation*. U. S. Dept. of Commerce, National Bureau of Standards, Washington, D. C. 8, 15, 18, 19

Davies, K. (1990). *Ionospheric Radio*. Peter Peregrinus Ltd., London. 21, 96

de Boor, C. (1993). *B(asic)-Spline basics*, 27–49. Academic Press, London. 73

Dettmering, D. (2003). *Die Nutzung des GPS zur dreidimensionalen Ionosphärenmodellierung*. Ph.D. thesis, Universität Stuttgart, Department of Geodesy and Geoinformatics. 3, 38

Dettmering, D., Limberger, M. and Schmidt, M. (2010). Using DORIS measurements for modeling the vertical total electron content of the Earth's ionosphere. *J Geod*, **88(12)**, 1131–1143. 31

Dettmering, D., Heinkelmann, R. and Schmidt, M. (2011a). Systematic differences between VTEC obtained by different space-geodetic techniques during CONT08. *J Geod*, **85**, 443–451. 86, 89

Dettmering, D., Schmidt, M., Heinkelmann, R. and Seitz, M. (2011b). Combination of different space-geodetic observations for regional ionosphere modeling. *J Geod*, **85(12)**, 989–998. 3, 23, 31

Dickerson, R.E., Gray, H.B. and Haight, G.P. (1979). *Chemical principles*. The Benjamin/Cummings Publishing Company, Menlo Park, CA. 11

Dieminger, W., Hartmann, G.K. and Leitinger, R.e. (1996). *The Upper Atmosphere: Data Analysis and Interpretation*. Springer, Heidelberg. 19, 20, 28

Divinskikh, N.I. (1988). Expansion of ionospheric characteristics fields in empirical orthogonal functions. *Adv. Space Res.*, **8(4)**, 179–187. 71

Dutta, S. (2016). *Optimization in Chemical Engineering*. Cambridge University Press, India. 58

Ezquer, R.G., Mosert de González, M. and Heredia, T. (1996). Electron density profile modeling. *Ann. Geofis.*, **XXXIX (3)**, 539–542. 66

Feltens, J. (1998). Chapman profile approach for 3-d global TEC representation. *In proceedings of the 1998 IGS Analysis Center Workshop, ESOC, Darmstadt, Germany, February 9-11*, 285–297. 4, 16, 23, 66

Feltens, J. (2007). Development of a new three-dimensional mathematical ionosphere model at european space agency/european space operations centre. *Space Weather*, **5(12)**, S12002. 4, 5, 6, 27, 66, 69

Feltens, J., Angling, M., Jackson-Booth, N., Jakowski, N., Hoque, M., Hernández-Pajares, M., Aragón-Ángel, A., Orús, R. and Zandbergen, R. (2011). Comparative testing of four ionospheric models driven with GPS measurements. *Radio Sci.*, **46(6)**, RS0D12. 23, 65

Fiacco, A.V. and McCormick, G.P. (1990). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. SIAM, Philadelphia. 59

Fisher, R.A. (1925). Applications of "student's" distribution. *Metron.*, **5**, 90–104. 57

Fjeldbo, G. and Eshleman, V.R. (1969). Atmosphere of Venus as studied with the Mariner V dual radio frequency occultation experiment. *Radio Sci.*, **4**, 879–897. 43

Fonda, C., Coïsson, P., Nava, B. and Radicella, S.M. (2005). Comparison of analytical functions used to describe topside electron density profiles with satellite data. *Ann. Geophys.*, **48 (3)**, 491–495. 67

Forsgren, A., Gill, P.E. and Wright, M.H. (2002). Interior Methods for Nonlinear Optimization. *SIAM Review*, **44(4)**, 525–597. 59

Fox, M.W. (1994). A simple, convenient formalism for electron density profiles. *Radio Sci.*, **29(6)**, 1473–1491. 23, 66

Fremouw, E.J. and Secan, J.A. (1992). Application of stochastic inverse theory to ionospheric tomography. *Radio Sci.*, **27 (5)**, 721–732.

Fritsche, M., Dietrich, R., Knüfel, C., Rülke, A., Vey, S., Rothacher, M. and Steigenberger, P. (2005). Impact of higher-order ionospheric terms on GPS estimates. *Geophys. Res. Lett.*, **32**, n/a–n/a, l23311. 23

Fu, L.L., Christensen, E.J., Yamarone Jr., C.A., Lefebvre, M., Ménard, Y., Dorrer, M. and Escudier, P. (1994). Topex/poseidon mission overview. *J. Geophys. Res.*, **99(C12)**, 24369–24381. 31

Galkin, I.A., Khmyrov, G.M., Kozlov, A.V., Reinisch, B.W., Huang, X. and Paznukhov, V.V. (2008). The ARTIST 5. *Radio Sounding and Plasma Physics, AIP Conf. Proc.*, **974**, 150–159. 106

Gallagher, D.L. and Craven, P.D. (2000). Global core plasma model. *J. Geophys. Res.*, **105(A8)**, 18,819–18,833. 28, 29

Garcia, R. and Crespon, F. (2008). Radio tomography of the ionosphere: Analysis of an underdetermined, ill-posed inverse problem, and regional application. *Radio Sci.*, **43**, RS2014. 90

García-Fernández, M. (2004). *Contributions to the 3D ionospheric sounding with GPS data*. Ph.D. thesis, Universitat Politècnica de Catalunya (UPC), Spain, Depts. of Applied Mathematics IV and Applied Physics. 4, 43, 44, 45

García-Fernández, M., Hernández-Pajares, M., Juan, J.M., Sanz, J., Orús, R., Coisson, P., Nava, B. and Radicella, S.M. (2003). Combining ionosonde with ground GPS data for electron density estimation. *J. Atmos. Sol. Terr. Phys.*, **65**, 683–691. 3

Gendt, G., Altamimi, Z., Dach, R., Söhne, W. and Springer, T. (2011). GGSP: Realisation and maintenance of the Galileo Terrestrial Reference Frame. *Adv. Space Res.*, **47(2)**, 174–185. 41

Gill, P.E., Murray, W. and Wright, M.H. (1981). *Practical Optimization*. Academic Press, London. 62

Giovanni, G.D. and Radicella, S.M. (1990). An analytical model of the electron density profile in the ionosphere. *Adv. Space Res.*, **10(11)**, (11)27–(11)30. 66

Goldsmith, M.J. (1999). *Sequential Quadratic Programming Methods based on Indefinite Hessian Approximations*. Ph.D. thesis, Stanford University, Department of Operations Research. 59

Golub, G.H. and Van Loan, C.F. (1996). *Matrix Computation*. The Johns Hopkins University Press, Baltimore and London. 51, 52

Golub, G.H., Heath, M. and Wahba, G. (1979). Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, **21(2)**, 215–223. 54

González-Casado, G., Juan, J.M., Hernández-Pajares, M. and Sanz, J. (2013). Two-component model of topside ionosphere electron density profiles retrieved from global navigation satellite systems radio occultation. *J. Geophys. Res.*, **118**, 7348–7359. 68

Goto, Y., Kasahara, Y. and Ide, T. (2012). improvement of equatorial density distribution of the global core plasma model using GPS-derived TEC. *Radio Sci.*, **47**, RS0F12. 28

Gulyaeva, T. and Bilitza, D. (2012). Towards ISO Standard Earth Ionosphere and Plasmasphere Model. *edited by Larsen, R. J., NOVA, Hauppauge, New York*, 1–39. 28

Gulyaeva, T.L. and Gallagher, D.L. (2007). Comparison of two IRI electron-density plasmasphere extensions with GPS-TEC observations. *Adv. Space Res.*, **39**, 744–749. 29

Gulyaeva, T.L., Huang, X. and Reinisch, B.W. (2002a). Ionosphere-Plasmasphere Model Software for ISO. *Acta Geod. Geophys. Hungarica*, **37**, 143–152. 28, 29, 30

Gulyaeva, T.L., Huang, X. and Reinisch, B.W. (2002b). Plasmaspheric extension of topside electron density profiles. *Adv. Space Res.*, **29(6)**, 825–831. 28

Hajj, G.A. and Romans, L.J. (1998). Ionospheric electron density profiles obtained with the Global Positioning System: Results from the GPS/MET experiment. *Radio Sci.*, **33(1)**, 175–190. 3, 43

Hajj, G.A., Lee, L.C., Pi, X., Romans, L.J., Schreiner, W.S., Straus, P.R. and Wang, C. (2000). COSMIC GPS Ionospheric Sensing and Space Weather. *Terrestrial, Atmospheric and Oceanic Sciences*, **11(1)**, 235–272. 42, 44

Han, S.P. (1977). A Globally Convergent Method for Nonlinear Programming. *J Optim Theory Appl*, **22**, 297. 59, 60

Hansen, P.C. (1990). Truncated Singular Value Decomposition Solutions to Discrete Ill-Posed Problems with Ill-Determined Numerical Rank. *SIAM J. Sci. and Stat. Comput.*, **11(3)**, 503–518. 51

Hansen, P.C. (1996). *Rank-Deficient and Discrete Ill-posed Problems*. Polyteknisk Forlag, Lyngby, Denmark. 54

Hansen, P.C. and O'leary, D.P. (1993). The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems. *SIAM J. Sci. Comput.*, **14(6)**, 1487–1503. 54

Hargreaves, J.K. (1992). *The Solar-Terrestrial Environment*. Cambridge University Press, New York. 15, 8, 10, 16, 17, 18, 19, 21, 28, 66, 69, 113

Heise, S. (2002). *Rekonstruktion dreidimensionaler Elektronendichteverteilungen basierend auf CHAMP-GPS-Messungen*. Ph.D. thesis, Freie Universität Berlin (in German), Institut für Meteorologie, Fachbereich Geowissenschaften. 23

Hernández-Pajares, M., Juan, J.M., Sanz, J. and Solé, J.G. (1998). Global observation of the ionospheric electronic response to solar events using ground and LEO GPS data. *J. Geophys. Res.*, **103(A9)**, 20789–20796. 43

Hernández-Pajares, M., Juan, J.M. and Sanz, J. (1999). New approaches in global ionospheric determination using ground GPS data. *J. Atmos. Sol. Terr. Phys.*, **61(1999)**, 1237–1247. 4, 65

Hernández-Pajares, M., Juan, J.M. and Sanz, J. (2000). Improving the Abel inversion by adding ground GPS data to LEO radio occultations in ionospheric sounding. *Geophys. Res. Lett.*, **27(16)**, 2473–2476. 43

Hernández-Pajares, M., Juan, J.M., Sanz, J., Orus, R., Garcia-Rigo, A., Feltens, J., Komjathy, A., Schaer, S.C. and Krankowski, A. (2009). The IGS VTEC maps: a reliable source of ionospheric information since 1998. *J Geod*, **83(3-4)**, 263–275. 2, 108

Hernández-Pajares, M., Juan, J.M., Sanz, J., Aragón-Ángel, A., A., G.R., Salazar, D. and Escudero, M. (2011). The ionosphere: effects, GPS modeling and the benefits for space geodetic techniques. *J Geod*, **85(12)**, 887–907. 2

Higham, N.J. (1987). A Survey of Condition Number Estimation for Triangular Matrices. *SIAM Review*, **29(4)**, 575–596. 51

Hobiger, T., Boehm, J. and Schuh, H. (2005). Vlbionos – Probing the Ionosphere by Means of Very Long Baseline Interferometry. *International Association of Geodesy Symposia*, **128**, 131–136. 31

Hofmann-Wellenhof, B., Lichtenegger, H. and Collins, J. (2001). *GPS Theory and Practice*. Springer-Verlag, Wien. 32

Hofmann-Wellenhof, B., Lichtenegger, H. and Wasle, E. (2008). *GNSS – Global Navigation Satellite Systems, GPS, GLONASS, Galileo &more*. SpringerWienNewYork, Wien. 22, 31, 34, 36, 37, 39

Hoque, M.M. and Jakowski, N. (2007). Higher order ionospheric effects in precise gnss positioning. *J Geod*, **81**, 259–268. 23

Hoque, M.M. and Jakowski, N. (2011). A new global empirical NmF2 model for operational use in radio systems. *Radio Sci.*, **46**, RS6015. 70

Howe, B.M., Runciman, K. and Secan, J.A. (1998). Tomography of the ionosphere: Four-dimensional simulations. *Radio Sci.*, **33(1)**, 109–128. 4

Huang, X. and Reinisch, B.W. (2001). Vertical electron content from ionograms in real time. *Radio Sci.*, **36(2)**, 335–342. 68

Huang, X., Reinisch, B.W., Song, P., Green, J.L. and Gallagher, D.L. (2004). Developing an empirical density model of the plasmasphere using IMAGE/RPI observations. *Adv. Space Res.*, **33**, 829–832. 28

Hunsucker, R.D. (1991). *Radio Techniques for Probing the Terrestrial Ionosphere*. Springer-Verlag, Berlin Heidelberg. 2

Hunsucker, R.D. and Hargreaves, J.K. (2003). *The high-latitude ionosphere and its effects on radio propagation*. Cambridge University Press, New York. 17, 18, 20

Hutchinson, M.F. (1990). A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines. *Commun Stat- Sim*, **19**, 433–450. 56

Jakowski, N. (2005). Ionospheric GPS radio occultation measurements on board CHAMP. *GPS Solut.*, **9**, 88–95. 5, 6, 67, 68, 69, 93

Jakowski, N., Wehrenpfennig, A., Heise, S., Reigber, C., Lühr, H., Grunwaldt, L. and Meehan, T.K. (2002). GPS radio occultation measurements of the ionosphere from CHAMP: Early results. *Geophys. Res. Lett.*, **29(10)**, 95–1–95–4. 43

Jakowski, N., Leitinger, R. and Ciraolo, L. (2004). Behaviour of large scale structures of the electron content as a key parameterfor range errors in gnss applications. *Annals of Geophysics*, **47**. 22

Jakowski, N., Hoque, M.M. and Mayer, C. (2011). A new global tec model for estimating transionospheric radio wave propagation errors. *Journal of Geodesy*, **85**, 965–974. 2, 70

Jayachandran, B., Krishnankutty, T.N. and Gulyaeva, T.L. (2004). Climatology of ionospheric slab thickness. *Ann. Geophys.*, **22**, 25–33. 95, 96

Jekeli, C. (2005). Spline Representations of Functions on a Sphere for Geopotential Modeling. *Report 475, Geod and Geoinf Sci, Ohio State University, Columbus, Ohio*. 18, 71

Jin, S., Cho, J.H. and Park, J.U. (2007). Ionospheric slab thickness and its seasonal variations observed by GPS. *J. Atmos. Terr. Phys.*, **69**, 1864–1870. 96

Jin, S., van Dam, T. and Wdowinski, S. (2013). Observing and understanding the earth system variations from space geodesy. *Journal of Geodynamics*, **72**, 1 – 10, sI: Geodetic Earth System. 3

Juan, J.M., Rius, A., Hernández-Pajares, M. and Sanz, J. (1997). A two-layer model of the ionosphere using Global Positioning System data. *Geophys. Res. Lett.*, **24(4)**, 393–396. 4, 65

Kalman, D. (1996). A Singularly Valuable Decomposition: The SVD of a Matrix. *The College Mathematics Journal*, **27(1)**, 2–23. 53

Kaplan, E.D. and Hegarty, C.J. (2006). *Understanding GPS, principles and Applications, Second Edition*. Artech House, Boston, London. 33

Kedar, S., Hajj, G.A., Wilson, B.D. and Heflin, M.B. (2003). The effect of the second order GPS ionospheric correction on receiver positions. *Geophys. Res. Lett.*, **30(16)**, 1829. 23

Kelley, M. (2009). *The Earth's Ionosphere: plasma physics and electrodynamics*. Academic Press, USA. 15, 20, 106

Kendall, E.A. (1989). *An introduction to numerical analysis, Second Edition*. John Wiley & Sons, Canada. 86, 87

Khmyrov, G.M., Galkin, I.A., Kozlov, A.V., Reinisch, B.W., McElroy, J. and Dozois, C. (2008). Exploring digisonde ionogram data with SAO-X and DIDBase. *Radio Sounding and Plasma Physics, AIP Conf. Proc.*, **974**, 175–185. 106

Klobuchar, J. (1987). Ionospheric time-delay algorithm for single-frequency GPS users. *IEEE Trans. Aerosp. Electron. Syst.*, **AES-23**, 325–331. 33

Koch, K.R. (1999). *Parameter Estimation and Hypothesis Testing in Linear Models*. Springer. 46, 47, 50, 53, 54, 55, 56, 57, 75, 79, 82, 94

Koch, K.R. (2007). *Introduction to Bayesian Statistics*. Springer. 47, 49, 53, 54

Koch, K.R. (2011). Data compression by multi-scale represenation of signals. *Journal of Applied Geodesy*, **5**, 1–12. 74

Koch, K.R. and Kusche, J. (2002). Regularization of geopotential determination from satellite data by variance components. *J. Geod.*, **76**, 259–268. 53, 54, 55, 56

Koch, K.R. and Schmidt, M. (2011). N-dimensional B-spline surface estimated by lofting for locally improving iri. *J. Geodetic Science*, **1(1)**, 41–51. 74

Kuhn, H.W. and Tucker, A.W. (1951). Nonlinear programming. *Proc. Second Berkeley Symp. on Math. Statist. and Prob., University of California Press, Berkeley*, 481–492. 59

Kuklinski, W.S. (1997). Ionospheric tomography via iterative cross-entropy minimization. *Radio Sci.*, **32(3)**, 1037–1049. 4, 65

Kumar, D.N. (2015). Introduction and basic concepts: (iii) classification of optimization problems, lecture Notes, Indian Institute of Science Bangalore, Department of Civil Engineering. 58

Kursinski, E.R., Hajj, G.A., Schofield, J.T., Linfield, R.P. and Hardy, K.R. (1997). Observing Earth's atmosphere with radio occultation measurements using the Global Positioning System. *J. Geophys. Res.*, **102(D19)**, 23,429–23,465. 41, 43

Langley, R.B. (1996). Propagation of the GPS signals. *In: Kleusberg A., Teunissen P. J. G, GPS for Geodesy, Springer Berlin Heidelberg*, 103–140. 33

Lei, J., Liu, L., Wan, W. and Zhang, S.R. (2005). Variations of electron density based on long-term incoherent scatter radar and ionosonde measurements over millstone hill. *Radio Sci.*, **40**, RS2008. 70

Lei, J., Syndergaard, S., Bruns, A.G., Solomon, S.C., Wang, W., Zeng, Z., Roble, R.G., Wu, Q., Kuo, Y.H., Holt, Z.S.R., J. M., Hysell, D.L., Rodrigues, F.S. and Lin, C.H. (2007). Comparison of COSMIC ionospheric measurements with ground-based observations and model predictions: Preliminary results. *J. Geophy. Res.*, **112**, A07308. 44, 45

Leick, A., Rapoport, L. and D., T. (2015). *GPS Satellite Surveying, 4th Edition*. John Wiley & Sons, Hoboken, New Jersey. 39

Leitinger, R., Ladreiter, H.P. and Kirchengast, G. (1997). Ionosphere tomography with data from satellite reception of Global Navigation Satellite System signals and ground reception of Navy Navigation Satellite System signals. *Radio Sci.*, **32(4)**, 1657–1669. 43

Leitinger, R., Nava, B. and Hochegger, S., G.and Radicella (2001). Ionospheric profilers using data grids. *Phys Chem Earth*, **26(5)**, 293–301. 65

Liang, W., Limberger, M., Schmidt, M., Dettmering, D. and Hugentobler, U. (2015a). Combination of ground- and space-based gps data for the determination of a multiscale regional 4-d ionosphere model. *International Association of Geodesy Symposia*. 74, 84, 86, 92, 93

Liang, W., Limberger, M., Schmidt, M., Dettmering, D., Hugentobler, U., Bilitza, D., Jakowski, N., Hoque, M., Wilken, V. and Gerzen, T. (2015b). Regional modeling of ionospheric peak parameters using GNSS data – an update for IRI. *Adv. Space Res.*, **55(8)**, 1981–1993. 16, 70, 86, 92, 93, 98, 107, 111

Limberger, M. (2015). *Ionosphere modeling from GPS radio occultations and complementary data based on B-splines*. Ph.D. thesis, Technische Universität Müchen, Institut für Astronomische und Physikalische Geodäsie. 6, 44

Limberger, M., Liang, W., Schmidt, M., Dettmering, D. and Hugentobler, U. (2013). Regional representation of F2 Chapman parameters based on electron density profiles. *Ann. Geophys.*, **31(12)**, 2215–2227. 70, 84, 93

Limberger, M., Hernández-Pajares, M., Aragón-Ángel, A., Altadill, D. and Dettmering, D. (2015). Long-term comparison of the ionospheric F2 layer electron density peak derived from ionosonde data and Formosat-3/COSMIC occultations. *J. Space Weather Space Clim.*, **5**, A21. 45

Liu, J.Y., Chuo, Y.J., Shan, S.J., Tsai, Y.B., Chen, Y.I., Pulinets, S.A. and Yu, S.B. (2004). Pre-earthquake ionospheric anomalies registered by continuous gps tec measurements. *Annales Geophysicae*, **22**, 1585–1593. 2

Liu, L., Wan, W. and Ning, B. (2006). A study of the ionogram derived effective scale height around the ionospheric hmF2. *Ann. Geophys.*, **24**, 851–860. 65, 70, 114

Liu, L., Le, H., Wan, W., Sulzer, M.P., Lei, J. and Zhang, M.L. (2007a). An analysis of the scale heights in the lower topside ionosphere based on the Arecibo incoherent scatter radar measurements. *J. Geophys. Res.*, **112**, A06307. 70

Liu, L., Wan, W., Zhang, M.L., Ning, B., Zhang, S.R. and Holt, J.M. (2007b). Variations of topside ionospheric scale heights over millstone hill during the 30-day incoherent scatter radar experiment. *Ann. Geophys.*, **25**, 2019–2027. 70

Liu, Z. (2011). A new automated cycle slip detection and repair method for a single dual-frequency GPS receiver. *J Geod*, **85(3)**, 171–183. 34

Liu, Z.Z. (2004). *Ionospheric Tomographic Modeling and Applications Using Global Positioning System (GPS) Measurements*. Ph.D. thesis, University of Calgary, Department of Geomatics Engineering. 4

Liu, Z.Z. and Gao, Y. (2004). Development and Evaluation of a New 3-D Ionospheric Modeling Method. *Navigation*, **51(4)**, 311–329. 65

Llewellyn, S.K. and Bent, R.B. (1973). Documentation and Description of the Bent Ionospheric Model. *Air Force Geophysics Laboratory, Report AFCRL-TR-73-0657, Hanscom AFB, Massachusetts*. 24

Luenberger, D.G. and Ye, Y. (2008). *Linear and Nonlinear Programming, Third Edition*. Springer, New York. 59

Lunt, N., Kersley, L. and Bailey, G.J. (1999). The influence of the protonosphere on GPS observations: Model simulations. *Radio Sci.*, **34(3)**, 725–732. 5

Lyche, T. and Schumaker, L.L. (2001). A Multiresolution Tensor Spline Method for Fitting Functions on the Sphere. *SIAM J. Sci. Comput.*, **22(2)**, 724–746. 73

Maes, C.M. (2011). *A Regularized Active-set Method for Sparse Convex Quadratic Programming*. Ph.D. thesis, Stanford University, Institute for Computational and Mathematical Engineering. 61

Mannucci, A.J., Wilson, B.D., Yuan, D.N., Ho, C.H., Lindqwister, U.J. and Runge, T.F. (1998). A global mapping technique for GPS-derived ionospheric total electron content measurements. *Radio Sci.*, **33(3)**, 565–582. 38

Mannucci, A.J., Ao, C.O., Young, L.E. and Meehan, T.K. (2014). Studying the Atmosphere Using Global Navigation Satellites. *EOS, Trans. Am. Geophys. Union*, **95(43)**, 389–391. 41

Matlab (2008). Optimization toolbox$^{TM}$ 4, user's guide. *The MathWorks, Inc., Natick*. 64

Mcilwain, C.E. (1961). Coordinates for Mapping the Distribution of Magnetically Trapped Particles. *J. Geophys. Res.*, **66(11)**, 3681–3691. 28

Minkwitz, D., Gerzen, T., Wilken, V. and Jakowski, N. (2014). Application of swaci products as ionospheric correction for single-point positioning: a comparative study. *Journal of Geodesy*, **88**, 463–478. 1

Montenbruck, O. (2003). Kinematic GPS positioning of LEO satellites using ionosphere-free single frequency measurements. *Aerosp Sci Technol*, **7**, 396–405. 36

Montenbruck, O. and Steigenberger, P. (2013). The BeiDou Navigation Message. *IGNSS Symposium 2013, Outrigger Gold Coast, Qld, Australia, 16-18 July*. 39, 40

Moré, J.J. and Wright, S.J. (1993). *Optimization Software Guide*. SIAM, Philadelphia. 64

Moritz, H. (1976). Covariance functions in least-squares collocation. Tech. Rep. MSU-CSE-00-2, Department of Geodetic Science, Ohio State University, Columbus, Ohio. 94

Muldrew, D.B. (1965). F-layer ionization troughs deduced from Alouette data. *J. Geophys. Res.*, **70(11)**, 2635–2650. 20

Naeimi, M. (2013). *Inversion of satellite gravity data using spherical radial base functions*. Ph.D. thesis, University of Hanover, Institute of Geodesy. 53

Nava, B., Coïsson, P. and Radicella, S.M. (2008). A new version of the nequick ionosphere electron density model. *J. Atmos. Sol. Terr. Phys.*, **70**, 1856–1862. 25, 26, 68

Nava, B., Coïsson, P. and Radicella, S.M. (2011). Data ingestion into NeQuick 2. *Radio Sci.*, **46**, RS0D17. 4

Nicolet, M. (1953). The collision frequency of electrons in the ionosphere. *J. Atmos. Sol. Terr. Phys.*, **3**, 200–211. 17

Nocedal, J. and Wright, S.J. (1999). *Numerical Optimization*. Springer-Verlag, New York. 58, 59, 60, 61, 62, 64

Nsumei, P., Reinisch, B.W., Huang, X. and Bilitza, D. (2012). New Vary-Chap profile of the topside ionosphere electron density distribution for use with the IRI model and the GIRO real time data. *Radio Sci.*, **47**, RS0L16. 4

Nsumei, P.A., Reinisch, B.W. and Huang, D., X.and Bilitza (2010). Comparing topside and bottomside-measured characteristics of the f2 layer peak. *Adv. Space Res.*, **46**, 974–983. 67

Ogden, R.T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser, Bosten, Basel, Berlin. 82

Petit, G. and Luzum, B.e. (2010). IERS Conventions (2010). *IERS Technical Note N0. 36, Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt am Main, Germany*. 23

Pezzopane, M. and Scotto, C. (2008). A method for automatic scaling of F1 critical frequencies from ionograms. *Radio Sci.*, **43**, RS2S91. 113

Phillips, D.L. (1962). A Technique for the Numerical Solution of Certain Integral Equations of the First Kind. *J. ACM*, **9(1)**, 84–97. 53

Phillips, P.C.B. (1982). The True Characteristic Function of the F Distribution. *Biometrika*, **69(1)**, 261–264. 56

Powell, M.J.D. (1978). *A fast algorithm for nonlinearly constrained optimization calculations, In: Watson G.A. (eds.) Numerical Analysis*. Springer, Berlin Heidelberg. 59, 60, 61, 64

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York. 101

Radicella, S.M. (2009). The NeQuick model genesis, uses and evolution. *Ann. Geophys.*, **52(3/4)**, 417–422. 2

Radicella, S.M. and Leitinger, R. (2001). The evolution of the DGR approach to model electron density profiles. *Adv. Space Res.*, **27(1)**, 35–40. 25, 66

Ram, S.T., Su, S.Y., Liu, C.H., Reinisch, B.W. and Mckinnell, L.A. (2009). Topside ionospheric effective scale heights ($H_T$) derived with ROCSAT-1 and ground-based ionosonde observations at equatorial and midlatitude stations. *J. Geophys. Res.*, **114**, A10309. 70

Rasmussen, C.E., Schunk, R.W. and Wickwar, V.B. (1988). A Photochemical Equilibrium Model for Ionospheric Conductivity. *J. Geophys. Res.*, **93 (A9)**, 9831–9840. 15

Ratcliffe, J.A. (1960). *Physics of the Upper Atmosphere*. Academic Press, New York. 15

Rawer, K. (1983). Replacement of the present sub-peak plasma density profile by a unique expression. *Adv. Space Res.*, **2(10)**, 183–190. 25, 66

Rawer, K. (1988). Synthesis of ionospheric electron density profiles with Epstein functions. *Adv. Space Res.*, **8(4)**, 191–199. 4

Rawer, K., Bilitza, D. and Ramakrishnan, S. (1978). Goals and status of the International Reference Ionosphere. *Rev. Geophys.*, **16(2)**, 177–181. 24, 66

Razin, M.R.G. (2016). Development and analysis of 3d ionosphere modeling using base functions and gps data over iran. *Acta Geodaetica et Geophysica*, **51**, 95–111. 4

Ree, M.H. (1989). *Physics and chemistry of the upper atmosphere*. Cambridge University Press, Cambridge. 10

Reigber, C., Lühr, H. and Schwintzer, P. (2000). CHAMP Mission Status and Perspectives. *Suppl. To EOS, Transactions, AGU*, **81, 48, F307**. 41

Reinisch, B.W. and Galkin, I.A. (2011). Global Ionospheric Radio Observatory (GIRO). *Earth Planets Space*, **63**, 337–381. 106

Reinisch, B.W. and Huang, X. (2001). Deducing topside profiles and total electron content from bottomside ionograms. *Adv. Space Res.*, **27(1)**, 23–30. 68

Reinisch, B.W., Nsumei, P., Huang, X. and Bilitza, D. (2007). Modeling the F2 topside and plasmasphere for IRI using IMAGE/RPI and ISIS data. *Adv. Space Res.*, **39**, 731–738. 4

Rishbeth, H. and Garriott, O.K. (1969). *Introduction to Ionospheric Physics*. Academic Press, New York and London. 4, 8, 9, 18, 19, 66

Rius, A., Ruffini, G. and Cucurull, L. (1997). Improving the vertical resolution of ionospheric tomography with GPS Occultations. *Geophys. Res. Lett.*, **24(18)**, 2291–2294. 4

Rockafellar, R.T. (1973). The multiplier method of Hestenes and Powell applied to convex programming. *J. Optimiz. Theory APP.*, **12(6)**, 555–562. 59

Rocken, C., Anthes, R., Exner, M., Hunt, D., Sokolovskiy, S., Ware, R., Gorbunov, M., Schreiner, W.S., Feng, D., Herman, B., Kuo, Y.H. and Zou, X. (1997). Analysis and validation of GPS/MET data in the neutral atmosphere. *J. Geophys. Res.*, **102(D25)**, 29849–29866. 31, 41

Rocken, C., Kuo, Y.H., Schreiner, W.S., Hunt, D., Sokolovskiy, S. and McCormick, C. (2000). Cosmic system description. *Terr. Atmos. Ocean. Sci.*, **11(1)**, 21–52. 42

Rodger, A.S., Moffett, R.J. and Quegan, S. (1992). The role of ion drift in the formation of ionisation troughs in the mid- and high-latitude ionosphere–a review. *J. Atmos. Terr. Phys.*, **54(1)**, 1–30. 20

Roese-Koerner, L. (2015). *Convex Optimization for Inequality Constrained Adjustment Problems*. Ph.D. thesis, University of Bonn, Institute of Geodasy and Geoinformation. 5, 61, 62, 63

Roese-Koerner, L., Devaraju, B., Sneeuw, N. and Schuh, W.D. (2012). A stochastic framework for inequality constrained estimation. *J Geod*, **86(11)**, 1005–1018. 5, 125, 126, 127, 129

Roese-Koerner, L., Devaraju, B., Schuh, W.D. and Sneeuw, N. (2015). Describing the quality of inequality constrained estimates. *IAG Symposia*, **140**, 15–20. 5, 126, 129

Rush, C., Fox, M., Bilitza, D., Davies, K., McNamara, L., Stewart, F. and PoKempner, M. (1989). Ionospheric mapping–an update of foF2 coefficients. *Telecomm. J.*, **56**, 179–182. 25

Sánchez, L., Seemüller, W., Drewes, H., Mateo, L., González, G., Silva, A., Pampillón, J., Martínez, W., Cioce, V., Cisneros, D. and S., C. (2013). Long-term stability of the sirgas reference frame and episodic station movements caused by the seismic activity in the sirgas region. *IAG Symposia*, **138**, 153–161. 94

Sardón, E. and Zarraoa, N. (1997). Estimation of total electron content using gps data: How stable are the differential satellite and receiver instrumental biases? *Radio Sci.*, **32(5)**, 1899–1910. 35

Schaer, S. (1999). *Mapping and Predicting the Earth's Ionosphere Using the Global Positioning System*. Ph.D. thesis, Bern University, Astronomical Institute. 1, 22, 23, 34, 35, 37, 71, 85, 109

Scherliess, L., Schunk, R.W., Sojka, J.J. and Thompson, D.C. (2004). Development of a physics-based reduced state kalman filter for the ionosphere. *Radio Sci.*, **39(1)**, RS1S04. 26

Scherliess, L., Schunk, R.W., Sojka, J.J., Thompson, D.C. and Zhu, L. (2006). Utah State University Global Assimilation of Ionospheric Measurements Gauss-Markov Kalman filter model of the ionosphere: Model description and validation. *J. Geophys. Res.*, **111**, A11315. 26

Scherliess, L., Thompson, D.C., and Schunk, R.W. (2009). Ionospheric dynamics and drivers obtained from a physics-based data assimilation model. *Radio Sci.*, **44(1)**, RS0A32. 26

Schimazak, T. (1995). World-wide daily variation in the height of the maximum electron density of the ionospheric f2-layer. *J. Radio Res. Labs.*, **2(7)**, 85–97. 69

Schittkowski, K. (1985). NLQPL: A FORTRAN-Subroutine Solving Constrained Nonlinear Programming Problems. *Ann. Oper. Res.*, **5(2)**, 485–500. 59

Schlüter, W. and Behrend, D. (2007). The International VLBI Service for Geodesy and Astrometry (IVS): current capabilities and future prospects. *J Geod*, **81(6-8)**, 379–387. 31

Schmidt, M. (2007). Wavelet modeling in support of IRI. *Adv. Space Res.*, **39 (5)**, 932–940. 5, 6, 65, 72, 73, 74, 77, 81

Schmidt, M. (2011). Preface. *J Geod*, **85**, 885–886. 1

Schmidt, M. (2012). Towards a Multi-Scale Representation of Multi-dimensional Signals. *International Association of Geodesy Symposia*, **137**, 119–127. 5, 74, 75

Schmidt, M., Bilitza, D., K., S.C. and Zeilhofer, C. (2008). Regional 4-d modeling of the ionospheric electron density. *Adv. Space Res.*, **42 (2008)**, 782–790. 4, 23, 65, 71, 72

Schmidt, M., Dettmering, D., Mößmer, M., Wang, W. and Zhang, J. (2011). Comparison of spherical harmonic and B spline models for the vertical total electron content. *Radio Sci.*, **46**, RS0D11. 5, 70, 74

Schmidt, M., Dettmering, D. and Seitz, F. (2015). Using B-Spline Expansions for Ionosphere Modeling. *In: Freeden W., Nashed M.Z., Sonar T. (Eds.) Handbook of Geomathematics (Second Edition)*, 939–983. 5, 23, 74

Schreiner, W.S., Sokolovskiy, S.V. and Rocken, C. (1999). Analysis and validation of GPS/MET radio occultation data in the ionosphere. *Radio Sci.*, **34(4)**, 949–966. 43, 44, 45

Schumaker, L.L. and Traas, C. (1991). Fitting scattered data on spherelike surfaces using tensor products of trigonometric and polynomial splines. *Numer. Math.*, **60**, 133–144. 74

Schunk, R.W., Sojka, J.J. and Eccles, J.V. (1997). Expanded Capabilities for the Ionospheric Forecast Model. *Air force materiel command*. 26

Schunk, R.W., Scherliess, L., Sojka, J.J., Thompson, D.C., Anderson, D.N., Codrescu, M., Minter, C., Fuller-Rowell, T.J., Heelis, R.A., Hairston, M. and Howe, B.M. (2004). Global Assimilation of Ionospheric Measurements (GAIM). *Radio Sci.*, **39(1)**, RS1S02. 4, 26

Sharifi, M.A. and Farzaneh, S. (2016). The ionosphere electron density spatio-temporal modeling based on the slepian basis functions. *Acta Geodaetica et Geophysica*, 1–14. 4, 65

Sharma, S., Dashora, N., Galav, P. and Pandey, R. (2011). Cycle slip detection, correction and phase leveling of RINEX formatted GPS observables. *Curr. Sci.*, **100(2)**, 205–212. 34

Sibanda, P. and Mckinnell, L.A. (2011). Topside ionospheric vertical electron density profile reconstruction using GPS and ionosonde data: possibilities for South Africa. *Ann. Geophys.*, **29**, 229–236. 2, 67

Siciliano, B. and Khatib, O. (2008). *Springer Handbook of Robotics*. Springer Berlin Heidelberg, New York. 90

Slanger, T.G. and Copeland, R.A. (2003). Energetic Oxygen in the Upper Atmosphere and the Laboratory. *Chem. Rev.*, **103 (12)**, 4731–4765. 17

Stankov, S.M. and Jakowski, N. (2006). Topside ionospheric scale height analysis and modelling based on radio occultation measurements. *J. Atmos. Sol. Terr. Phys.*, **68**, 134–162. 69, 70

Stankov, S.M., Jakowski, N., Heise, S., Muhtarov, P., Kutiev, I. and Warnant, R. (2003). A new method for reconstruction of the vertical electron density distribution in the upper ionosphere and plasmasphere. *J. Geophys. Res.*, **108 (A5)**, 1164. 15, 67, 68, 98

Stankov, S.M., Stegen, K., Muhtarov, P. and Warnant, R. (2011). Local ionospheric electron density profile reconstruction in real time from simultaneous ground-based GNSS and ionosonde measurements. *Adv. Space Res.*, **47 (7)**, 1172–1180. 68

Stollnitz, E.J., DeRose, T.D. and Salesin, D.H. (1995a). Wavelets for Computer Graphics: A prime, Part 1. *IEEE Comput. Graph. Appl.*, **15(3)**, 76–84. 74

Stollnitz, E.J., DeRose, T.D. and Salesin, D.H. (1995b). Wavelets for Computer Graphics: A prime, Part 2. *IEEE Comput. Graph. Appl.*, **15(4)**, 75–85. 73, 74, 75

Strang, G. and Borre, K. (1997). *Linear Algebra, Geodesy, and GPS*. Wellesley-Cambridge Press, USA. 53

Stroud, A. and Secrest, D. (1966). *Gaussian Quadrature Formulas*. Prentice-Hall, Inc., Englewood Cliffs, N. J. 87

Syndergaard, S. (2007). Retrieval of electron density profiles at CDAAC. *CEDAR-DASI Workshop, Santa Fe, New Mexico*. 45

Tapley, B.D., Bettadpur, S., Watkins, M. and Reigber, C. (2004). The gravity recovery and climate experiment: Mission overview and early results. *Geophys. Res. Lett.*, **31**, L09607. 42

Themens, D.R., Jayachandran, P.T., Langley, R.B., MacDougall, J.W. and Nicolls, M.J. (2013). Determining receiver biases in GPS-derived total electron content in the auroral oval and polar cap region using ionosonde measurements. *GPS Solut.*, **17(3)**, 357–369. 36, 38

Tikhonov, A.N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, **5**, 1035–1038. 53

Tikhonov, A.N. and Arsenin, V.Y. (1977). *Solutions of ill-posed problems*. V. H. Winston & Sons, Washington, D. C. 53

Todorova, S. (2008). *Combination of space geodetic techniques for global mapping of the ionosphere*. Ph.D. thesis, Vienna University of Technology, Faculty of Mathematics and Geoinformation. 31, 89

Tsai, L.C. and Tsai, W.H. (2004). Improvement of GPS/MET Ionospheric Profiling and Validation Using the Chung-Li Ionosonde Measurements and the IRI model. *Terr. Atmos. Ocean Sci.*, **15(4)**, 589–607. 43

Tsai, L.C., Liu, C.H. and Hsiao, T.Y. (2009). Profiling of Ionospheric Electron Density Based on FormoSat-3/COSMIC Data: Results from the Intense Observation Period Experiment. *Terr. Atmos. Ocean Sci.*, **20(1)**, 181–191. 43, 66, 70, 94

Tsai, L.C., Cheng, K.C. and Liu, C.H. (2011). GPS radio occultation measurements on ionospheric electron density from low Earth orbit. *J Geod*, **85**, 941–948. 4, 43

van de Kamp, M.M.J.L. (2013). Medium-scale 4-D ionospheric tomography using a dense GPS network. *Ann. Geophys.*, **31**, 75–89. 4

Venter, G. (2010). *Review of Optimization Techniques, In: Blockley R., Shyy W. (eds.) Encyclopedia of Aerospace Engineering*. John Wiley & Sons, Ltd., Hoboken. 59

Verhulst, T. and Stankov, S.M. (2014). Evaluation of ionospheric profilers using topside sounding data. *Radio Sci.*, **49**, 181–195. 67

Ware, R., Rocken, C., Solheim, F., Exner, M., Schreiner, W., Anthes, R., Feng, D., Herman, B., Gorbunov, M., Sokolovskiy, S., Hardy, K., Kuo, Y., Zou, X., Trenberth, K., Meehan, T., Melbourne, W. and Businger, S. (1996). GPS Sounding of the Atmosphere from Low Earth Orbit: Preliminary Results. *Bull. Amer. Meteor. Soc.*, **77**, 19–40. 41

Webb, P.A. and Essex, E.A. (2001). A dynamic diffusive equilibrium model of the ion densities along plasmaspheric magnetic flux tubes. *J. Atmos. Sol. Terr. Phys.*, **63**, 1249–1260. 28

Webb, P.A. and Essex, E.A. (2003). Modifications to the Titheridge upper ionosphere and plasmasphere temperature model. *J. Geophys. Res.*, **108(A10)**, 1359. 28

White, F.M. (2008). *Fluid Mechanics*. McGraw-Hill, New York. 11

Wickert, J., Reigber, C., Beyerle, G., König, R., Marquardt, C., Schmidt, T., Grunwaldt, L., Galas, R., Meehan, T.K., Melbourne, W.G. and Hocke, K. (2001). Atmosphere sounding by GPS radio occultation: First results from CHAMP. *Geophys. Res. Lett.*, **28(17)**, 3263–3266. 41

Wickert, J., Michalak, G., Schmidt, T., Beyerle, G., Cheng, C.Z., Healy, S.B., Heise, S., Huang, C.Y., Jakowski, N., Köhler, W., Mayer, C., Offiler, D., Ozawa, E., Pavelyev, A.G., Rothacher, M., Tapley, B. and Arras, C. (2009). GPS Radio Occultation: Results from CHAMP, GRACE and FORMOSAT-3/COSMIC. *Terr. Atmos. Ocean. Sci.*, **20(1)**, 35–50. 42

Wong, E. (2011). *Active-Set Methods for Quadratic Programming*. Ph.D. thesis, University of California, San Diego. 61

Wu, X., Hu, X., Gong, X., Zhang, X. and Wang, X. (2009). Analysis of inversion errors of ionospheric radio occultation. *GPS Solut.*, **13**, 231–239. 43, 45

Wu, X., Hu, X., Wang, G., Zhong, H. and Tang, C. (2013). Evaluation of COMPASS ionospheric model in GNSS positioning. *Adv. Space Res.*, **51**, 959–968. 40

Xu, G. (2003). *GPS Theory, Algorithms and Applications*. Springer, Berlin Heidelberg New York. 36

Xu, J., Schreier, F., Doicu, A. and Trautmann, T. (2016). Assessment of tikhonov-type regularization methods for solving atmospheric inverse problems. *Journal of Quantitative Spectroscopy and Radiative Transfer*, **184**, 274–286. 54

Xu, P., Shen, Y. and Fukuda, Y. (2006). Variance component estimation in linear inverse ill-posed models. *J Geod*, **80**, 69–81. 55

Yu, Y., Wan, W., Xiong, B., Ren, Z., Zhao, B., Zhang, Y., Ning, B. and Liu, L. (2015). Modeling chinese ionospheric layer parameters based on eof analysis. *Space Weather*, **13**, 339–355. 70

Yue, X., Schreiner, W.S., Lei, J., Sokolovskiy, S.V., Rocken, C., Hunt, D.C. and Kuo, Y.H. (2010). Error analysis of Abel retrieved electron density profiles from radio occultation measurements. *Ann. Geophys.*, **28**, 217–222. 45

Yue, X., Schreiner, W.S., Kuo, Y.H., Hunt, D.C., Wang, W., Solomon, S.C., Burns, A.G., Bilitza, D., Liu, J.Y., Wan, W. and Wickert, J. (2012). Global 3-D ionospheric electron density reanalysis based on multisource data assimilation. *J. Geophys. Res.*, **117**, A09325. 99

Zeilhofer, C. (2008). *Multi-dimensional B-spline Modeling of Spatio-temporal Ionospheric Signals*. Master's thesis, DGK Reihe A123, Deutsche Geodätische Kommission, Müchen. 4, 6, 74, 75, 103

Zeilhofer, C., Schmidt, M. and Bilitza, C.K., D.and Schum (2009). Regional 4-D modeling of the ionospheric electron density from satellite data and IRI. *Adv. Space Res.*, **43(11)**, 1669–1675. 4, 23, 65

Zhang, E.A., Xiao, D.H., Hao, Y.Q., Ridley, A.J. and Moldwin, M. (2011). Modeling ionospheric fof2 by using empirical orthogonal function analysis. *Ann. Geophys.*, **29**, 1501–1515. 70, 72

Zhang, H., Xu, P., Han, W., Ge, M. and Shi, C. (2013). Eliminating negative VTEC in global ionosphere maps using inequality-constrained least squares. *Adv. Space Res.*, **51**, 988–1000. 5

Zhang, M.L., Reinisch, B.W., Shi, J.K., Wu, S.Z. and Wang, X. (2006). Diurnal and seasonal variation of the ionogram-derived scale height at the F2 peak. *Adv. Space Res.*, **37(5)**, 967–971. 70

Zhang, M.L., Liu, C., Wan, W., Liu, L. and Ning, B. (2009). A global model of the ionospheric F2 peak height based on EOF analysis. *Ann. Geophys.*, **27**, 3203–3212. 70, 72

Zhu, J., Santerre, R. and Chang, X.W. (2005). A bayesian method for linear, inequality-constrained adjustment and its application to gps positioning. *J Geod*, **78**, 528–534. 127

Zolesi, B. and Cander, L.R. (2014). *Ionospheric Prediction and Forecasting*. Springer, Heidelberg. 16, 18, 19, 106