Technical University of Munich
School of Engineering and Design
Deutsches Geodätisches Forschungsinstitut (DGFI-TUM)
Prof. Dr.-Ing. habil. Florian Seitz

# Estimation of Electron Density Key Parameters Using the Multi-Layer Chapman Model Considering Inequality Constraints from Ionospheric Radio Occultation Measurements

**Qiu, Fenghe**

Master Thesis

**Study Program:** Geodesy and Geoinformation

**Supervisors:** Prof. Dr.-Ing. habil. Michael Schmidt
Deutsches Geodätisches Forschungsinstitut (DGFI-TUM)
Technical University of Munich

M.Sc. Ganesh Lalgudi-Gopalakrishnan
Deutsches Zentrum für Luft- und Raumfahrt (DLR)

**19/10/2022**

## Declaration of Authorship

I confirm that this Master's thesis is my own work and I have documented all sources and material used.

This thesis was not previously presented to another examination board and has not been published.

| | |
|---|---|
| _____ | _____ |
| Place and date | Signature |

# Abstract

The four-dimensional (4D) space and time-dependent electron density needs to be accurately known for precise point positioning, satellite navigation, and telecommunication. If a high-precision and high-resolution model of the electron density of the ionosphere and plasmasphere is available globally, each measurement of space-geodetic observation techniques such as the Global Navigation Satellite System, Satellite Altimetry or the satellite tracking system (like Doppler Orbitography and Radiopositioning Integrated by Satellite) could be corrected for the plasmaspheric and ionospheric impact. Since the development of such a model still relies on data with insufficient and unevenly global coverage, it would be beneficial to introduce equality and inequality constraints. In this work, we model the electron density using a multi-layer Chapman model based on 2D series expansions of B-spline functions. Different scenarios are selected containing various key parameters with equality and inequality constraints and estimated under the consideration of inequality constrained optimization. The algorithm is applied to simulated data and semi-simulated input data. Finally, the results are evaluated via the comparisons between the input and reconstructed electron densities.

**Keywords:** ionosphere modeling; plasmasphere modeling; multi-layer Chapman model; inequality constrained optimization; B-spline expansions

# Acknowledgements

First of all, I would like to give my heartfelt thanks to my academic supervisor Prof. Schmidt, for his invaluable instruction, and inspiration. Without his previous advice and guidance, this study could not have been completed. Also, I must express my sincere thanks to my co-supervisor Ganesh, for their enlightening ideas and advice.

My acknowledgments also go to my colleagues. Their support and generosity move me a lot. Last but not least, I have to thank my family members especially my parents.

# Contents


**1 Introduction** **1**
- 1.1 Motivation . . . 1
- 1.2 State of the art . . . 2
- 1.3 Goals and contributions . . . 4
- 1.4 Outline of the thesis . . . 4

**2 Ionosphere Background** **7**
- 2.1 Ionosphere structure . . . 7
  - 2.1.1 Structure of the ionosphere . . . 7
  - 2.1.2 Spatial variations in the ionosphere . . . 9
  - 2.1.3 Temporal variations in the ionosphere . . . 10
- 2.2 Mechanism of ionization . . . 11
- 2.3 The Chapman layer . . . 12
- 2.4 International Reference Ionosphere . . . 14
- 2.5 Plasmasphere . . . 14

**3 Parameter Estimation** **17**
- 3.1 Nonlinear problem . . . 17
- 3.2 Optimization with inequality constraints . . . 18
- 3.3 Karush–Kuhn–Tucker optimality conditions . . . 19
- 3.4 Sequential Quadratic Programming method . . . 20
  - 3.4.1 Search direction – Quadratic Programming subproblem . . . 20
  - 3.4.2 Active-set methods . . . 20
  - 3.4.3 Step length – Line search and merit function . . . 26

**4 Electron Density Modeling** **29**
- 4.1 General modeling of the vertical electron density distribution . . . 29
- 4.2 Parametrization for key parameters - B-splines . . . 30
  - 4.2.1 Normalized quadratic polynomial B-splines . . . 30
  - 4.2.2 Normalized trigonometric B-splines . . . 32
  - 4.2.3 B-spline tensor products . . . 33
- 4.3 Linearized observation equation system . . . 35
- 4.4 Procedure of modeling . . . 36

**5 Numerical Analysis** **39**
- 5.1 One parameter with inequality constraints . . . 39
- 5.2 Three parameters with inequality constraints . . . 43
- 5.3 Nine parameters with inequality constraints . . . 56
- 5.4 Solution with separability approach . . . 66

**6 Conclusions and Outlook** **73**
- 6.1 Conclusions . . . 73
- 6.2 Outlook . . . 74

# Acronyms

**AM** Amplitude Modulated . . . 8
**CCIR** Committee Consultative for Ionospheric Radiowave . . . 2
**COSMIC** Constellation Observing System for Meteorology Ionosphere and Climate . . . 74
**COSPAR** Committee on Space Research . . . 3
**EDP** Electron Density Profile . . . 14
**EDU** Electron Density Unit . . . 41
**EOF** Empirical Orthogonal Function . . . 30
**EUV** Extreme Ultraviolet . . . 11
**FIR** Finite Impulse Response . . . 3
**FM** Frequency modulated . . . 8
**GCPM** Global Core Plasma Model . . . 15
**GIM** Global Ionosphere Map . . . 69
**GNSS** Global Navigation Satellite System . . . 1
**GPID** Global Plasmasphere Ionosphere Density . . . 15
**GRACE** Gravity Recovery and Climate Experiment . . . 3
**IAAC** Ionospheric Associated Analysis Centers . . . 69
**IAG** International GNSS Service . . . 69
**ICLS** inequality constrained least squares . . . 3
**ICO** Inequality Constrained Optimization . . . 36
**ICTP** International Centre for Theoretical Physics . . . 3
**IGAM** Institute for Geophysics, Astrophysics and Meteorology . . . 3
**IRI** International Reference Ionosphere . . . 3
**ISR** Incoherent Scatter Radar . . . 15
**KKT** Karush–Kuhn–Tucker . . . 17
**LSM** Least Squares Method . . . 42
**LT** Local Time . . . 9
**MLCM** Multi-Layer Chapman Model . . . 4
**NEDM** Neutrelitz electron density model . . . 15
**NLP** Non-Linear Programming . . . 18
**NPDM** Neustrelitz Peak Density Model . . . 16
**NPHM** Neustrelitz Peak Height Model . . . 16
**NPSM** Neustrelitz Plasmasphere Model . . . 16
**NTCM** Neustrelitz TEC model . . . 16
**QP** Quadratic Programming . . . 20
**RPI** radio plasma imager . . . 15
**SH** Spherical Harmonics . . . 30
**SQP** Sequential Quadratic Programming . . . 17
**TEC** Total Electron Content . . . 16
**URSI** International Union of Radio Science . . . 3
**VTEC** Vertical Total Electron Content . . . 3

# List of Figures


1.1 Ionospheric layer composition based on the vertical electron density distribution [Limberger, 2015] . . . 5

2.1 Typical day and night electron density profiles in the mid-latitude ionosphere [Hargreaves, 1992]. el/cm$^3$ is an abbreviation for electrons/cm$^3$. . . . 8
2.2 A schema of the fountain effect where **B** the geomagnetic field, **E** the electric field, $\boldsymbol{g}$ the gravity and $\boldsymbol{\nabla p}$ the pressure gradient forces [Kelley, 2009]. . . . 10
2.3 Loss of radiation intensity with respect to a path element $ds$ or height interval $dh$ [Limberger, 2015] . . . 13
2.4 Electron density profiles for different function parameters. Dependency with respect to the maximum electron density $N_m = \left[10^5 \text{ el/cm}^3, 2 \cdot 10^5 \text{ el/cm}^3, ..., 6 \cdot 10^5 \text{ el/cm}^3\right]$ (top left), peak heights $h_m = [200 \text{ km}, 250 \text{ km}, ..., 400 \text{ km}]$ (top right), and scale height $H = [60 \text{ km}, 70 \text{ km}, ..., 100 \text{ km}]$ (bottom left) [Limberger, 2015]. . . . 14

4.1 Normalized quadratic polynomial B-splines with different levels $J_\Phi = 0, 1, 2, 3$ and accordingly different number of B-splines $K_\Phi = 3, 4, 6, 10$. . . . 31
4.2 Trigonometric B-splines with different levels $J_T = 0, 1, 2, 3$ and accordingly different number of B-splines $K_T = 3, 6, 12, 24$. . . . 33
4.3 Polynomial B-splines $\Phi^3_{k_1}(x_1)$ and $\Phi^2_{k_2}(x_2)$ of degree $m = 2$ with different levels $J_1 = 3$ and $J_2 = 2$. $\phi^2_2(x_2)$ (left), $\phi^2_0(x_2)$ (right) and $\phi^3_4(x_1)$ are emphasized to show the support area. . . . 34
4.4 Combination of trigonometric B-splines with level $J_1 = 3$ and polynomial B-splines with level $J_2 = 2$. A specific spline combination identified by $k_1 = 4$ and $k_2 = 7$ has been highlighted and plotted in the center part of the left subplot. Accordingly, a 3D representation of the tensor product is given on the right hand side. . . . 35
4.5 Procedure developed within this thesis . . . 37

5.1 Flow chart of closed loop validation . . . 40
5.2 Maps of the original key parameter $N_m^{F_2}$, the estimated key parameter from the least squares method and the differences between them. . . . 41
5.3 Maps of the original key parameter $N_m^{F_2}$, the estimated key parameter in scenario 1-1 and the differences between them. . . . 42
5.4 Maps of the original key parameter $N_m^{F_2}$, the estimated key parameter in scenario 1-2 and the differences between them . . . 43
5.5 Original electron densities vs. reconstructed electron densities in scenario 2-1 . 45
5.6 Differences between original electron densities and reconstructed electron densities in scenario 2-1 . . . 45
5.7 Differences between original electron densities and reconstructed electron densities in scenario 2-2. . . . 46
5.8 Maps of the original key parameter $N_m^{F_2}$, the estimated key parameter in scenario 2-3 and the differences between them. . . . 47

5.9 Maps of the original key parameter $h_m^{F_2}$, the estimated key parameter in scenario 2-3 and the differences between them. . . . 48
5.10 Maps of the original key parameter $H^{F_2}$, the estimated key parameter in scenario 2-3 and the differences between them. . . . 48
5.11 Differences between original electron densities and reconstructed electron densities in scenario 2-3 . . . 49
5.12 Map of average absolute electron density differences along vertical profiles in scenario 2-3 . . . 50
5.13 Maps of the original key parameter $N_m^{F_2}$, the estimated key parameter in scenario 2-4 and the differences between them. . . . 50
5.14 Maps of the original key parameter $h_m^{F_2}$, the estimated key parameter in scenario 2-4 and the differences between them. . . . 51
5.15 Maps of the original key parameter $H^{F_2}$, the estimated key parameter in scenario 2-4 and the differences between them. . . . 51
5.16 Maps of upper bounds of the key parameters in scenario 2-5 . . . 52
5.17 Maps of lower bounds of the key parameters in scenario 2-5 . . . 53
5.18 Maps of estimated results of the key parameters in scenario 2-5 . . . 53
5.19 Differences between upper bounds and estimated values in scenario 2-5 . . . 54
5.20 Differences between lower bounds and estimated values in scenario 2-5 . . . 54
5.21 Differences between noise-free electron densities and reconstructed electron densities in scenario 2-5 . . . 55
5.22 Map of average absolute electron density differences along vertical profiles in scenario 2-5 . . . 55
5.23 Maps of standard deviations of estimated parameters in scenario 2-5 . . . 56
5.24 Original electron densities vs. reconstructed electron densities in scenario 3-1 . . . 57
5.25 Differences between original electron densities and reconstructed electron densities in scenario 3-1 . . . 58
5.26 Original electron densities vs. reconstructed electron densities in scenario 3-2 . . . 59
5.27 Original electron densities vs. reconstructed electron densities in scenario 3-3 . . . 61
5.28 Original electron densities vs. reconstructed electron densities in scenario 3-4 . . . 61
5.29 Maps of lower bounds of key parameters in scenario 3-5 . . . 62
5.30 Maps of upper bounds of key parameters in scenario 3-5 . . . 63
5.31 Maps of estimated results of the key parameters in scenario 3-5 . . . 63
5.32 Differences between upper bounds and estimated values in scenario 3-5 . . . 64
5.33 Differences between lower bounds and estimated values in scenario 3-5 . . . 64
5.34 Differences between original electron densities and reconstructed electron densities in scenario 3-5 . . . 65
5.35 Map of average absolute electron density difference along vertical profiles in scenario 3-5 . . . 65
5.36 Maps of estimated results of the key parameters in scenario 3-6 . . . 66
5.37 Differences between upper bounds and estimated values in scenario 3-6 . . . 67
5.38 Differences between lower bounds and estimated values in scenario 3-6 . . . 67
5.39 4: Differences between noise-free electron densities and reconstructed electron densities in scenario 3-6 . . . 68
5.40 Map of average absolute electron density differences along vertical profiles in scenario 3-6 . . . 68
5.41 Maps of estimated results of the key parameters in scenario 4-1 . . . 69
5.42 Differences between upper bounds and estimated values in scenario 4-1 . . . 70
5.43 Differences between lower bounds and estimated values in scenario 4-1 . . . 70

5.44 Differences between original electron densities and reconstructed electron densities in scenario 4-1 . . . . . 71
5.45 Map of average absolute electron density differences along vertical profiles in scenario 4-1 . . . . . 71

# List of Tables


2.1 Approximate values of height ranges for different ionospheric layers . . . . . 7
2.2 IRI data sources and characteristics [Liang, 2017] . . . . . 15

5.1 Contents of the scenarios . . . . . 40
5.2 Values of the key parameters . . . . . 44
5.3 Relevant values in scenario 2-1 . . . . . 44
5.4 Relevant values in scenario 2-2 . . . . . 46
5.5 Constraints of the key parameters in scenario 2-3 . . . . . 47
5.6 Relevant values in scenario 3-1 . . . . . 57
5.7 Relevant values in scenario 3-2 . . . . . 59
5.8 Relevant values in scenario 3-3 . . . . . 60
5.9 Relevant values in scenario 3-4 . . . . . 60

# 1 Introduction

## 1.1 Motivation

The ionosphere means the ionized part of the atmosphere from 50 km from the Earth's surface up to the plasmasphere, where plasma consists of neutral molecules, positive ions, free electrons and atoms. The plasma is formed in the high-altitude atmosphere by the photoionization of solar radiation (ultraviolet rays, X-rays and solar wind) and the collision of energetic particles produced by the Sun or other stars [Davies, 1990]. The electron density in the ionosphere varies with height, thus dividing the ionosphere from the bottom up into layers $D$, $E$, $F_1$ and $F_2$, where the electron density reaches its maximum in the $F_2$ layer [Schunk and Nagy, 2009]. The region above the peak height of the $F_2$ layer is called the top side of the ionosphere, while the lower region is called the bottom side of the ionosphere. In a narrow sense, the ionosphere refers to the atmospheric space around 50-1000 km above ground level, while the region from 1000 km to the top of the magnetosphere is known as the plasmasphere, where the magnetospheric boundary extends to a height of about 3-5 Earth radii [Lunt et al., 1999; Bishop et al., 2009; Teunissen and Montenbruck, 2017]. Although the electron density of the plasmasphere is much lower than that of the ionosphere, it still contains a large proportion of electrons due to its large altitude range: around 10% during the day and up to 60% at night [Yizengaw et al., 2008]. Although we describe and model the ionosphere in layers, in fact, the temporal and spatial changes of the ionosphere are quite complex, not only with periods such as day, season, and year but also with changes in latitude and longitude. It is also affected by various factors such as solar activity and geomagnetism, resulting in very complex changes [Cander, 2019]. Therefore, the study of the ionosphere has become one of the hotspots in related scientific fields such as atmospheric research and climate monitoring.

When studying and analyzing the ionosphere, multiple ionospheric parameters are generally used to characterize the ionospheric structure, which include the electron density, ion density, electron temperature, and ion temperature. From the point of view of geodetic applications, the electron density among these ionospheric parameters is the most important and relevant one [Bust and Mitchell, 2008]. It has a significant influence on the propagation of radio waves causing bending the signal and producing a propagation delay. For example, when an electromagnetic wave is emitted by the Global Navigation Satellite System (GNSS), it propagates through the ionosphere and produces an error of up to tens of meters, which is one of the main error sources that restrict the high-precision positioning of GNSS users (especially single-frequency GNSS users) [Macalalad et al., 2013; Choy et al., 2008; Øvstedal, 2002]. For multi-frequency GNSS users, since the ionosphere is a dispersive medium, the first-order ionosphere effect can be eliminated by using linear combinations of the measurements at two frequencies. However, single-frequency GNSS users may rely on ionospheric models to correct those propagation errors [Minkwitz et al., 2014]. Generally, ionospheric models can be divided into three categories, empirical models, theoretical models, and parametric models [Cander et al., 1999; Feltens et al., 2011]. A more specific description of them will be given in the next section.

However, due to the complexity of spatio-temporal variations in the ionosphere and the strong correlations among different parameters, the classical least squares estimation may lead to physically unrealistic results, e.g., the maximum electron density in a certain region is negative. As a consequence, when we are estimating the parameters of models, inequality constraints need to be considered in order to give feasible regions for parameters and avoid those unrealistic results [Koch, 1985; Roese-Koerner, 2015; Nocedal and Wright, 1999]. Furthermore, in the field of GNSS, models with inequality constraints help to refine and improve regional and global ionospheric delay correction models and improve GNSS real-time positioning accuracy.

## 1.2 State of the art

Electron density modeling has always been an important direction of ionosphere research, mainly by means of processing and analyzing observation data, so as to be able to grasp spatio-temporal variations in the ionosphere. Although a large number of different ground-based and space-based observing technologies can largely provide information for ionospheric research, ionospheric variations are affected by multiple factors. Therefore, it is necessary to use ionosphere models to quantitatively describe and analyze the ionosphere. These ionosphere models can be broadly categorized into three main types: theoretical models, empirical models, and parametric models. [Cander et al., 1999; Feltens et al., 2011]. For detailed descriptions of the first two models, see Liang [2017]. In this thesis, we will concentrate on the empirical models.

The empirical models are models based on statistical analysis of large amounts of observations and use certain mathematical functions to quantitatively describe the physical parameters of ionospheric manifestations in response to various external factors (solar activity, temporal periodicity, geomagnetic activity, etc.) and to describe the "meteorology" and behavior of the ionosphere. The main of these quantitative descriptions of the formulation physical parameters are Chapman functions (see Section 2.3), spherical harmonic functions, polynomials, and spline functions [Schaer, 1999; Hu et al., 2014]. Usually, the accuracy of empirical models is affected by factors such as the number and quality of observed data samples and modeling methods. In the following, detailed descriptions of some famous empirical models will be given and the definitions of the $D$, $E$, $F_1$ and $F_2$ layers are introduced in Section 2.1.

The Bent model [Llewellyn and Bent, 1973] describes the ionospheric electron density as a function of latitude, longitude, time, season, and solar radio flux. The topside is represented by a parabola and three exponential profile segments, and the bottomside by a bi-parabola. The model is based on about 50000 Alouette topside ionograms (1962-1966), 6000 Ariel 3 in-situ measurements (1967-1968), and 400000 bottomside ionograms (1962-1969). For the $F_2$-peak the Committee Consultative for Ionospheric Radiowave (CCIR) maps are used. The model has been widely used for ionospheric refraction corrections in satellite tracking. It does not include the lower layers ($D$, $E$, $F_1$) and uses a simple quadratic relationship between factor in CCIR map and the height of the $F_2$-peak·

The Ching–Chiu's model [Ching and Chiu, 1973; Chiu, 1975] is a global phenomenologica1 mode1 of the large-scale variations of ionospheric electron density with the annual, diurnal and solar activity cycles has been constructed from monthly averaged hourly ionospheric sounding data from some 50 stations spanning the years 1957-1970 and is specifically designed for global thermospheric and ionospheric dynamical calculations. Compared to the

Bent model, the Ching–Chiu's model covers $E$, $F_1$ and $F_2$ layers using the standard Chapman function. However, it is less detailed in latitude and diurnal variations than large computer models because it summarizes the observations in simple empirical formulas. But it is more comprehensive in large annual and solar cycle variations.

The International Reference Ionosphere (IRI) model [Bilitza et al., 1990, 2014] is an international project sponsored by the Committee on Space Research (COSPAR) and the International Union of Radio Science (URSI). Famous versions are IRI79, IRI2012, IRI2016 and IRI2020. It divides electron density profile into four subregions covering $D$, $E$, $F_1$ and $F_2$ layers in the altitude range from 50 km to 2000 km. As an empirical model, IRI uses most of the available and reliable data sources for the ionosphere plasma and keeps continuously upgrading through introducing new options as new data and modeling approaches are available. Because of this, in this thesis, we select the IRI model and solve the inequality constrained optimization problems for the key parameters. For further details about IRI model, see Section 2.4.

The NeQuick [Nava et al., 2008] is an ionospheric electron density model 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 (IGAM) of the University of Graz, Austria. It is particularly adopted by the Galileo system in order to compute the ionospheric delay corrections [Aragon-Angel et al., 2019]. It allows to calculate the electron density at any given location in the ionosphere and the Vertical Total Electron Content (VTEC) along any ground-to-satellite ray path by numerical integration. In recent years, several changes have been introduced in the version 1 of the NeQuick model and the NeQuick 2 is the latest version. It uses a modified profile formulation, which includes five semi-Epstein layers with modeled thickness parameters. Three profile anchor points are used; namely the $E$ layer peak, the $F_1$ peak and the $F_2$ peak. The model topside is represented by a semi-Epstein layer with a height-dependent thickness parameter empirically determined.

For the study of inequality constrained optimization, it should be noted that inequality constrained optimization problems are usually difficult to solve compared to unconstrained or equality constrained optimization problems [Nocedal and Wright, 1999]. Fortunately, there are already optimization procedures available(see e.g. Lötstedt [1984]; Gill et al. [1984]; Coleman and Li [1996]; Mead and Renaut [2010]). In our field, Koch [1981] and Fritsch [1983a] were the first to address inequality constrained least squares (ICLS) problems in geodesy. While the former formulated the design of optimal Finite Impulse Response (FIR) filters as ICLS problem, the latter examined hypothesis testing with inequality constraints. Later on, Fritsch [1982] and Koch [1985] transformed the quadratic programming problem resulting from the first and second-order design of a geodetic network into a linear complementarity problem and solved it via Lemke's algorithm. Fritsch [1983b, 1985] examined further possibilities resulting from the use of ICLS for the design of FIR filters and other geodetic applications.

A more recent approach stems from Peng et al. [2006] who established a method to express many simple inequality constraints as one intricate equality constraint in a least squares context. Tang et al. [2012] used inequalities as smoothness constraints to improve the estimated mass changes in Antarctica from Gravity Recovery and Climate Experiment (GRACE) observations, which leads again to a quadratic program. Liang [2017] used inequality constraints optimization to directly estimate key parameters for IRI model along vertical profile.

## 1.3 Goals and contributions

The objective of this thesis is to achieve an estimation of certain elements in the empirical Multi-Layer Chapman Model (MLCM) (see Section 4.1) of the electron density under the consideration of inequality constrained optimization, which results should be physically realistic. And we choose the IRI model for testing the algorithm since it has obvious advantages in data sources and coverage over other electron density models described in Section 1.2. To achieve this goal, different approaches are investigated, combined and adapted within the derived 3D electron density modeling concept. In the vertical profile, we select the standard Chapman function to describe the distribution of electron density in $D$, $E$, $F_1$ and $F_2$ layers, see Fig. 1.1. For the plasmasphere, we apply the exponential function. In total, we cover the electron density from 50 km to 1000 km. Since B-spline expansions are appropriate for handling inhomogeneous data distribution even data gaps, it is introduced to model the horizontal key parameters distributed globally.

In the presentation of MLCM, there are totally 14 key parameters and the goal of this thesis is to select the largest but appropriate subset of parameters to be estimated. Since the core of the thesis is inequality constrained optimization, we start the numerical estimation with a closed loop simulation and then apply the optimization procedure to a combination of simulated and real data based on the separability approach. To evaluate the quality of the estimation, we adopt the Monte Carlo method to compute the standard deviation of the estimated parameters and make a comparison between the original electron density and the reconstructed electron density in a closed loop simulation. Specific contributions of the work in this thesis are:

- Perform the estimation with one parameter with inequality constraints based on simulated data and compare the result with the classical least squares method.
- Perform the estimation with three parameters with inequality constraints based on simulated data and compute the standard deviation using the Monte Carlo method.
- Perform the estimation with nine parameters with inequality constraints based on simulated data and data from the separability approach.

## 1.4 Outline of the thesis

In Chapter 1, the objective of this thesis is given. We start from the motivation of the research and continued by the state of the art as well as goals and contributions.

The background of the ionosphere is introduced in Chapter 2, which includes physical information about the ionosphere and plasmasphere. The vertical structure, spatio-temporal variations in the ionosphere, and the mechanism of ionization in ionosphere and plasmasphere are explained. Furthermore, we address the background of the Chapman function and the IRI model since they are used in this thesis for 3D electron density modeling.

The procedure of inequality constrained optimization is covered in Chapter 3, where we first explain the definition of nonlinear problem. For inequality constrained problem, we first introduce the Karush–Kuhn–Tucker optimality conditions (see Section 3.3) and choose the sequential quadratic programming method which is explained in the last section.

Chapter 4 mainly gives the configurations of our contributions. We select Chapman functions to describe the vertical electron density distribution of the $D$, $E$, $F_1$, and $F_2$ layers and use the

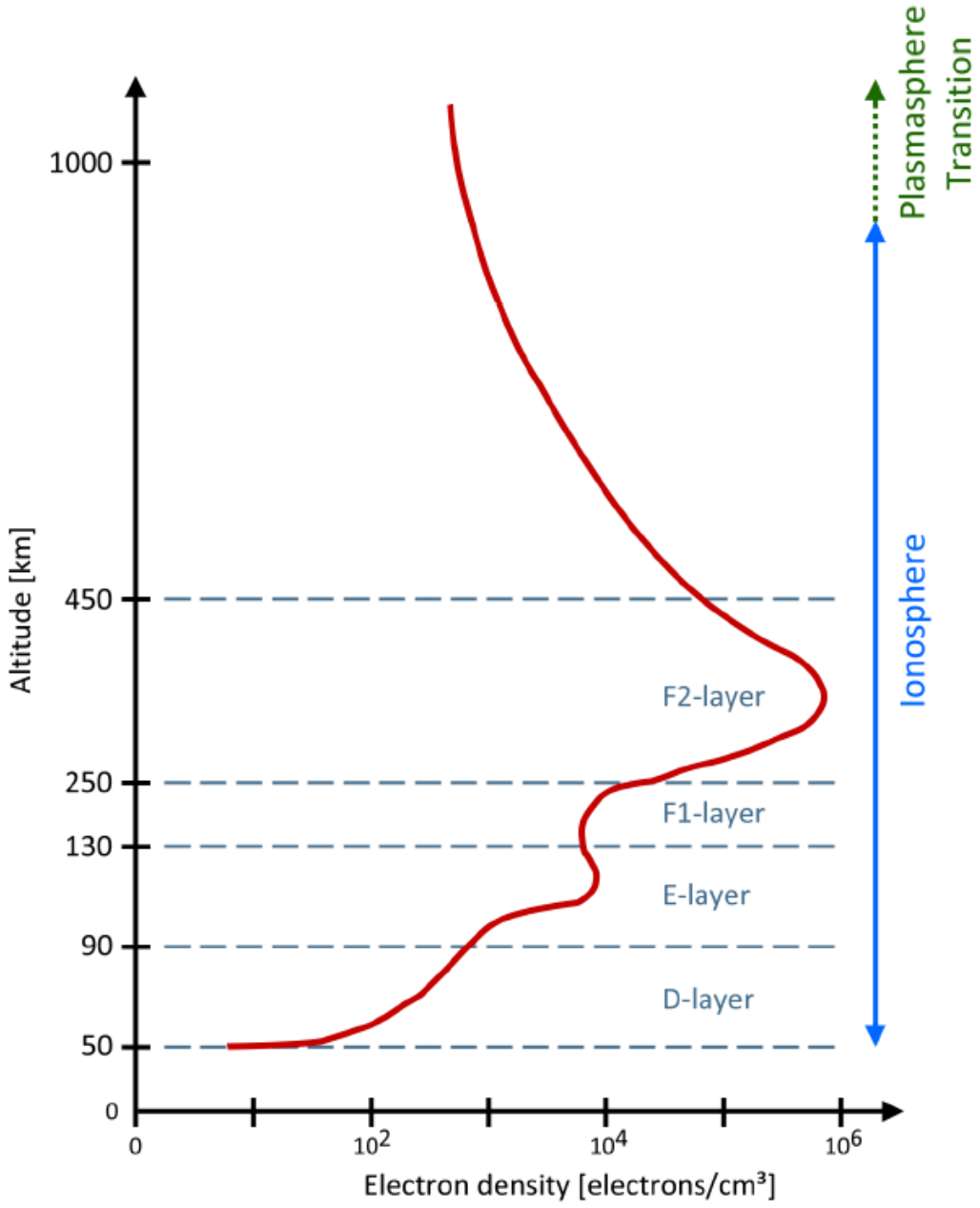


Figure 1.1: Ionospheric layer composition based on the vertical electron density distribution [Limberger, 2015]

exponential function for plasmasphere. The Section 4.2 covers the parametrization with the use of B-splines for the key parameters. The linearization of the observation equation system and the modeling procedure are described in the last two sections.

The contributions in this study are summarized in Chapter 5, which consists of four sections corresponding to different scenarios. In the first scenario, only one parameter is estimated with inequality constraints and the result is compared with the classical least squares method. Estimations with three inequality constrained parameters and nine inequality constrained parameters are presented in the following scenarios where the standard deviations of the results consisting of three inequality constrained parameters are also included. The last scenario is in response to nine inequality constrained parameters estimation with the input from the separability approach.

The major findings and conclusions of this thesis are summarized in the last part: conclusion and outlook (Chapter 6). Furthermore, some thoughts for further works are raised.

# 2 Ionosphere Background

## 2.1 Ionosphere structure

The ionosphere refers to the region of the Earth's upper atmosphere where ions and electrons exist. In a broad sense, the ionosphere also includes the part of the Earth's atmospheric space up to the top of the magnetosphere, and the upper part of the ionosphere is usually referred to as the plasmasphere [Kelley, 2009]. Because the ionosphere covers a wide height range, the top and bottom ionosphere exhibit different properties such as electron temperature and density, ion composition and density. Therefore, it is of great importance to study the structure and properties of the ionosphere and plasmasphere.

### 2.1.1 Structure of the ionosphere

The chemical composition is usually not evenly distributed in space, resulting in an altitude-dependent structure of atmospheric molecules when ionized by solar radiation. According to those different structures and wavelengths of solar radiation that are most absorbed in certain regions, the ionosphere can be subdivided into distinct regions. A widely used strategy is that the ionosphere is divided into $D$, $E$, $F_1$ and $F_2$ layer from the bottom to the top, see Fig. 1.1. Table 2.1 gives the approximate altitude range for the layers used in this thesis and Fig. 2.1 describes the typical day and night electron density in the mid-latitude ionosphere [Hargreaves, 1992]. The dashed lines depict the electron density distribution for the minimum sunspot case and the solid lines depict the maximum sunspot case. In the following, more detailed descriptions of each layer will be given.

#### 2.1.1.1 $D$ layer

This layer is the closest layer to the ground, with an altitude range of 50-90 km and an electron concentration of about $10^2 - 10^4$ el/cm$^3$. The layer is subject to multiple variables under the influence of solar activity and the local atmosphere, such as the electron density in high solar activity years is about two to three times higher than in low solar activity years; the electron density is usually larger in summer than in winter; the electron density is usually the largest

Table 2.1: Approximate values of height ranges for different ionospheric layers

| Layer | Height range [km] |
|---|---|
| $D$ | 50-90 |
| $E$ | 90-140 |
| $F_1$ | 140-200 |
| $F_2$ | 200-500 |
| Plasmasphere | 500-1000 |

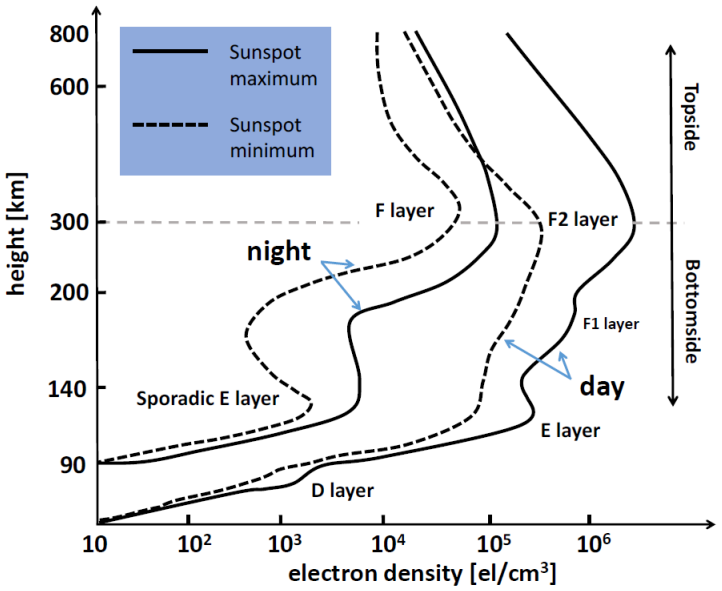


Figure 2.1: Typical day and night electron density profiles in the mid-latitude ionosphere [Hargreaves, 1992]. el/cm$^3$ is an abbreviation for electrons/cm$^3$.

in the afternoon and decreases after sunset; it largely disappears at night. The electron density is usually the largest in the afternoon and decreases after sunset; largely disappears at night. Frequency modulated (FM) based radio signals can be scattered in this area [Liang, 2017].

#### 2.1.1.2 *E* layer

The *E* layer is a layer containing molecules and ions located at about 90 km to 140 km from the ground, containing $NO^+$, $O_2^+$, and absorbing mainly soft X-rays from solar radiation. The electron density in this layer is higher than in the *D* layer, about $10^3 - 10^5$ el/cm$^3$, and there are significant diurnal, seasonal and solar activity cycle variations [Liang, 2017]. Its peak is located at about 100 km to 110 km. Due to the presence of different molecular gases at this altitude, the ion production and loss rates are independent of the altitude and are dominated by molecular-ion decomposition and compounding processes, resulting in a strong correlation between the electron concentration and the ion production and loss rates in this region. The *E* layer reflects radio-standard Amplitude Modulated (AM) signals from the ground back to the ground. During the night the *E* layer does not disappear, but the electron concentration is somewhat weakened. In addition to the regular ionospheric *E* layer, there is a sporadic *E* layer in this region called the *Es* layer [Haldoupis, 2011; García-Rigo et al., 2011], which generally occurs at 90 km to 120 km and higher. 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 the aurora. The most important feature is that the magnitude of the *Es* layer can be similar to that of the *F* region, which will be given in the following.

#### 2.1.1.3 $F$ layer

This layer is the mid region of the ionosphere, about 140-1000 km above the ground, with the largest electron concentration of about $10^5 - 10^6$ el/cm$^3$, and the electron distribution is mainly influenced by the neutral wind and the Earth's magnetic field [Liang, 2017]. The $F$ layer can reflect radio shortwave signals emitted from the ground, while visible light, radar, TV and FM wavelengths are too short to be reflected by the ionosphere and penetrate the ionosphere. At the same time, changes in the larger scales of the ionosphere are associated with changes in this layer, as it has the largest proportion of electron density in the entire ionosphere.

Due to the complex physical mechanisms involved in its formation, the $F$ layer can be subdivided into the $F_1$ layer and the $F_2$ layer. The $F_1$ layer is more striking during the summer than during the winter, at high solar activity, and during ionosphere storms. The $F_2$ layer is at a higher altitude than the $F_1$ layer. The peak is typically located between 300 km and 500 km. It exists during day and night but 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]. However, the $F_1$ layer disappears during the night.

### 2.1.2 Spatial variations in the ionosphere

Globally, differences in atmospheric density, atmospheric composition and exposure to solar radiation at different locations result in an uneven horizontal distribution of the electron density. In general, the ionosphere is active and varies strongly at low latitudes; at mid-latitudes it is relatively calm, with significant seasonal and diurnal variations; at high latitudes (especially the poles) it is influenced by polar days and nights, with less significant seasonal and diurnal variations. In addition, the ionosphere is subject to other regular and irregular variations due to solar activity, solar radiation and solar altitude angle variations, as described below.

The first thing we need to note is that, in low latitude regions, there are usually two maxima of the electron density occurring at about 15° to 20° north and south of the Earth's magnetic equator [Nishida, 1968; Maeda and Badillo, 1966]. This is the so-called "Appleton anomaly" affected by "fountain effect", see Fig. 2.2. The charged particles in the bottom region of the ionosphere are continuously moving upwards due to the combined effect of solar radiation and tides, and that the ionization of atmospheric molecules between the latitude of $\pm 20°$ are enhanced by the interaction of the horizontal geomagnetic field in the equatorial region.

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, which generally start at 9:00 to 11:00 Local Time (LT) and can continue until 22:00 LT, after which the phenomenon disappears. The peak electron density in the summer hemisphere is usually lower than that in the winter hemisphere around the winter/summer solstice.

In mid-latitude regions, the electron density in the $F_2$ layer is higher in winter than that in summer [Barab, 1962]. The reason is that, under the influence of seasonal summer airflow, the increase of atmospheric molecules is more than the increase of single atoms at mid-latitudes, resulting in higher ion capture rates in summer than in winter. Furthermore, the increase in capture rates in summer is much higher than the increase due to ionization by solar radiation, which in turn results in lower electron concentrations in summer than in winter. This anomaly

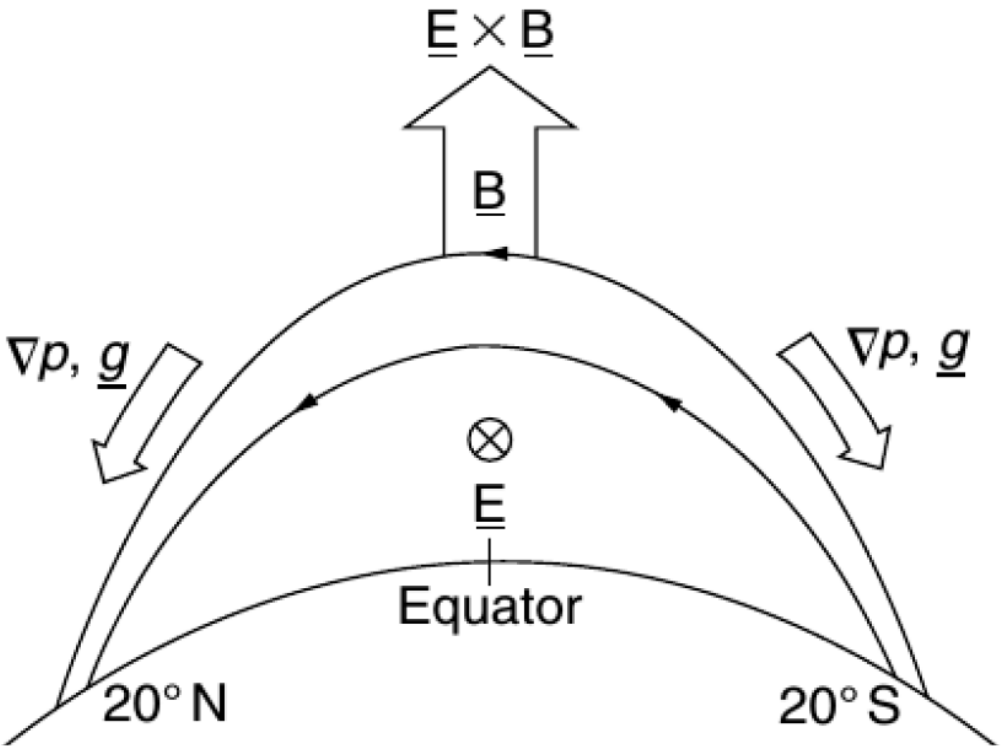


Figure 2.2: A schema of the fountain effect where **B** the geomagnetic field, **E** the electric field, $g$ the gravity and $\nabla\ p$ the pressure gradient forces [Kelley, 2009].

can occur annually in the northern hemisphere but is not present in the southern hemisphere during low solar activity years.

Bellchambers and Piggott [1958] as well as Penndorf [1965] found that the electron density does not always decrease monotonically at night, but sometimes appears to increase. This phenomenon is possible in all regions except high latitudes. At the same time, the probability of occurrence before midnight is comparable to that after midnight and it is more likely to occur in summer than that in winter. Besides, the higher the solar and geomagnetic activity, the less this phenomenon will occur.

In addition to the above solar and geomagnetic activities that can cause ionospheric anomalies, some other natural phenomena and human activities can cause ionospheric anomalies, such as typhoons [Bauer, 1957], earthquakes [Davies and Baker, 1965], lightning [Kelley et al., 1985; Holzworth et al., 1985], tsunamis [Tang et al., 2018] and nuclear experiments [Zhang and Tang, 2015; Park et al., 2011], among others.

### 2.1.3 Temporal variations in the ionosphere

The ionospheric variability is influenced by solar activity and can be similar to that of the Sun, e.g., daily cycles, quasi-27-day cycles [Rich et al., 2003], seasonal cycles [Balan et al., 1998; Da Rosa et al., 1973; Titheridge, 1973], On the other hand, ionospheric variability is influenced by solar-terrestrial variations and solar-lunar variations. In addition to these periodic variations, atmospheric molecular concentrations and solar radiation are also influenced by other

factors, resulting in the ionosphere being characterized by periodic variations in the mixing of multiple frequency signals [Amiri-Simkooei and Asgari, 2012].

## 2.2 Mechanism of ionization

The physical processes in the ionosphere consist of two parts: the photochemical process, which leads to the production and loss of ionized particles, and the transport process, which describes the movement of ionized particles for example diffusion and drift [Davies and Baker, 1965]. The balance of ionization is in fact a dynamic equilibrium, which means, the electron density is affected by the relative speed of the production and loss processes, as well as the movement of the particles. The ionized continuity equation can be adopted to describe the rate of change $dN_e/dt$ of the electron density $N_e$

$$\frac{dN_e}{dt} = Q - L - \mathrm{div}(N_e v) \tag{2.1}$$

where $t$ is the time, $Q$ and $L$ present the production and loss rate, respectively. $\mathrm{div}(N_e v)$ describes the loss rate of the electrons due to the transport process with $v$ the net drift velocity.

Here, we will discuss more details about $Q$ and $L$, since they are affected by two important ionization processes: photoionization and recombination.

The amount of ion-electron pairs lost per unit volume and per unit time is defined as the loss rate $L$, and it is mainly caused by two processes, which are called recombination and attachment. Recombination describes the fact that the free electrons tend to reunite with the positive ions to produce neutral atoms again, whereas attachment means the free electrons tend to attach themselves to neutral molecules to form negative ions. The loss rate can be represented by

$$L = \frac{dI}{ds} = -\sigma_f n I \tag{2.2}$$

where $I$ denotes the loss of radiation intensity along the Line-of-Sight. $n$ is the density of a neutral particle and $\sigma_f$ identifies the cross section of the radiation or photon absorption rate for the frequency $f$.

The production rate $Q$ is the number of ion-electron pairs produced caused by photoionization per unit volume per unit time. The mechanisms of photoionization are different depending on the latitude. In low- and mid-latitude, solar radiation in the Extreme Ultraviolet (EUV) and X-ray parts of the spectrum are the main cause for the production of ions and electron pairs. However, in high latitude regions or during magnetic storms, the production process is caused by the fact that charged particles settle into the atmosphere and collide with neutral molecules to produce particle ionization, and produce negative ions during the attachment at the bottom ionosphere. Since the ionization is balanced, The production rate $Q$ of ions or electrons can be obtained as

$$Q = -\frac{dI}{ds} = \sigma_f n I. \tag{2.3}$$

## 2.3 The Chapman layer

The Chapman layer modeling, the IRI model as well as the plasmasphere described in the current and following sections mainly follow the work in Limberger [2015] and Liang [2017]. In order to approximate the variations of free electrons along the vertical, Chapman [1931a,b] gave a formulation which is generally called the Chapman function. Its methodology can also be found in Chapman and Mian [1942a,b]. The Chapman function is built based on the following assumptions:

- the atmosphere is composed of only one chemical element, i.e., as an isothermal one-component gas,
- the atmosphere is horizontally stratified and can be described by an undisturbed layer structure without diffusion or horizontal variations,
- the radiation is monochromatic and parallel,
- the temperature $T$, gravitational acceleration $g$ and molecular mass $m$ are constant, so that the scale height $H = kT/(mg)$ ($k$ is the Boltzmann's constant) is constant, too, and
- the electron production is caused only by photoionization and electron loss only through recombination.

Considering the last assumption, recombination is the only reason for electron loss and can be described as

$$v_L = \alpha N_e^2 \tag{2.4}$$

where $\alpha$ is introduced as the recombination coefficient [Hargreaves, 1992].

Therefore, the electron density variation with respect to time can be explained as

$$\frac{dNe}{dt} = Q - \alpha N_e^2 = \sigma_f n I - \alpha N_e^2 \tag{2.5}$$

according to the Eqs. (2.3) and (2.4). This shows that now the electron density variation is in fact the difference between ionization and recombination [Davies, 1990]. Here, in order to describe $n$, we start from a known neutral density value $n_0$ at $h_0$ so that

$$n = n_0 \exp(-\frac{h - h_0}{H}). \tag{2.6}$$

It is worth mentioning that the Chapman function is dependent on the Sun's position since the sun is of significant importance in influencing the density of electrons. Therefore, the solar zenith angle $\chi$ is taken into account. According to Fig. 2.3, we have

$$ds = \frac{dh}{\cos(\chi)}. \tag{2.7}$$

We define $z = h - h_0$ and substitute all the expressions above into Eq. (2.3), which leads to

$$Q = \sigma_f \; n_0 \; I_\infty \exp(-\frac{z}{H} + \frac{\sigma_f n_0 H}{cos(\chi)} \exp(-\frac{z}{H})). \tag{2.8}$$

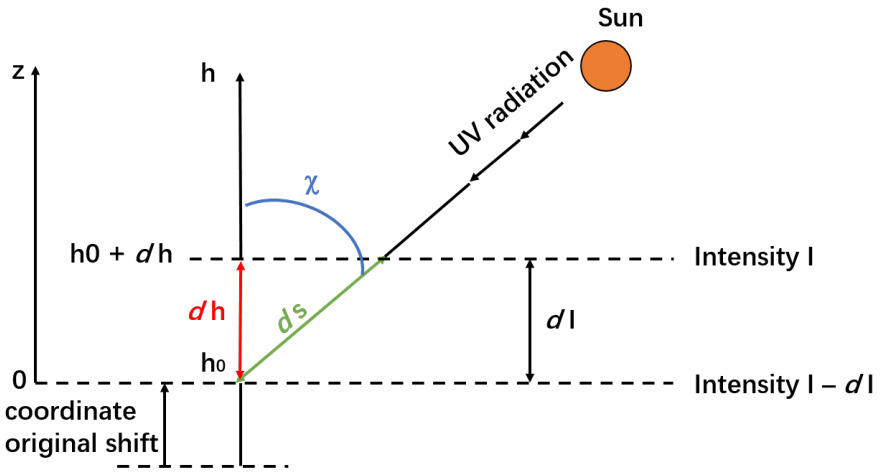


Figure 2.3: Loss of radiation intensity with respect to a path element $ds$ or height interval $dh$ [Limberger, 2015]

Here, the derivation of the equation above is calculated with the purpose of finding the maximum production rate and its corresponding peak height. We adopt $\chi = 0$ because $Q$ will reach its maximum when the sun is at its zenith and obtain

$$\frac{dQ}{dz} = \underbrace{\sigma_f \; n_0 \; I_\infty}_{\text{term1}} \underbrace{\exp(-\frac{z}{H} + \sigma_f \; n_0 \; H \exp(-\frac{z}{H}))}_{\text{term2}} \underbrace{-\frac{1}{H}(1 + \sigma_f \; n_0 \; H \exp(-\frac{z}{H}))}_{\text{term3}} \stackrel{!}{=} 0. \tag{2.9}$$

It can be proved that the terms 1 and 2 are always larger than 0. Therefore, only term 3 = 0 can be expected, which leads to

$$\sigma_f \; n_0 \; H = -1 \tag{2.10}$$

under the consideration that the maximum ion production can only occur at $h_0$ or $z = 0$. Substituting the above into Eq. (2.8), we reach the expression of the maximum ion production

$$\begin{aligned} Q_{max} &= \sigma_f \; n_0 \; I_\infty \exp(-1) \\ &= Q \exp(-1 + \frac{z}{H} + \frac{1}{cos(\chi)} \exp(-\frac{z}{H})). \end{aligned} \tag{2.11}$$

Solving the above equations, we finally get the Chapman product function

$$Q = Q_{max} \exp(1 - \frac{h - h_0}{H} - \frac{1}{\cos(\chi)} \exp(-\frac{h - h_0}{H})). \tag{2.12}$$

We assume $N_e = \sqrt{Q/\alpha}$, $\chi = 0$ and $h_m = h_0$, which leads to

$$N_e(h) = N_m \exp(\frac{1}{2}(1 - \frac{h - h_m}{H} - \exp(-\frac{h - h_m}{H}))). \tag{2.13}$$

This equation shows that there are three key parameters in each layer:

- $N_m$ is the maximum electron density of the current layer,
- $h_m$ corresponds to the peak height, and
- $H$ denotes the scale height.

Fig. 2.4 shows how the three parameters influence the shape of the Chapman function profiles.

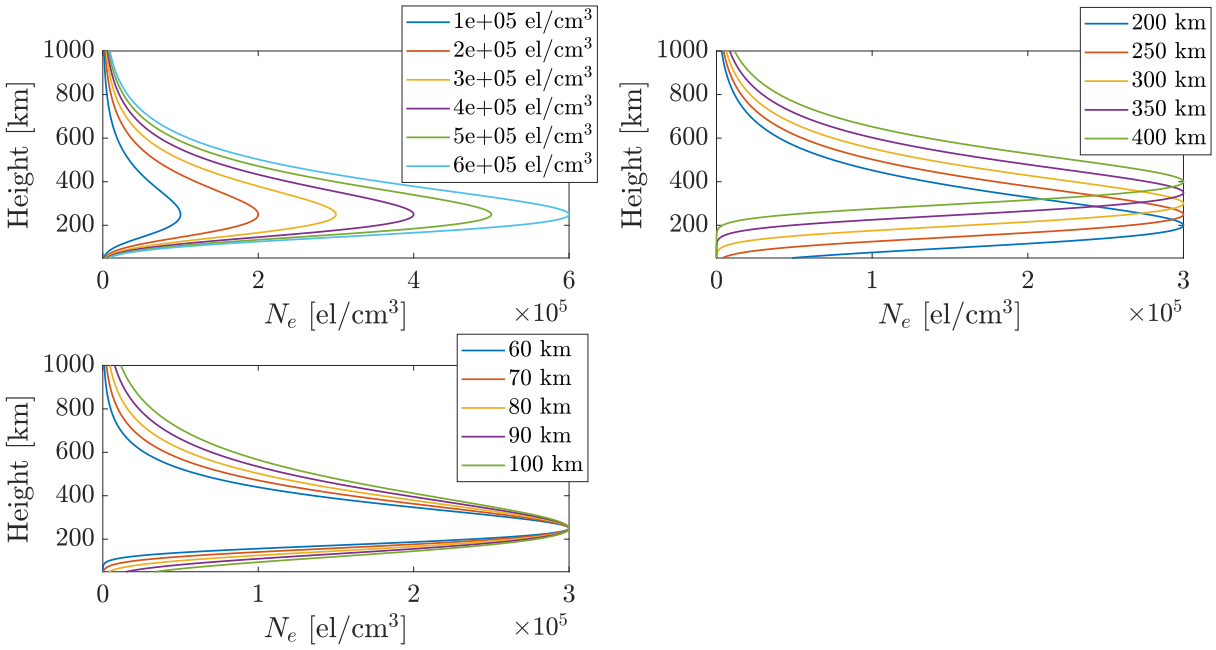


Figure 2.4: Electron density profiles for different function parameters. Dependency with respect to the maximum electron density $N_m = [10^5 \text{ el/cm}^3, 2 \cdot 10^5 \text{ el/cm}^3, ..., 6 \cdot 10^5 \text{ el/cm}^3]$ (top left), peak heights $h_m = [200 \text{ km}, 250 \text{ km}, ..., 400 \text{ km}]$ (top right), and scale height $H = [60 \text{ km}, 70 \text{ km}, ..., 100 \text{ km}]$ (bottom left) [Limberger, 2015].

## 2.4 International Reference Ionosphere

The IRI model [Bilitza et al., 1990; Bilitza and Reinisch, 2008; Bilitza et al., 2014] is an international project sponsored by COSPAR and 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 in the altitude range from 50 km to 2000 km [Bilitza et al., 2011]. 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.

The structure of the vertical Electron Density Profile (EDP) of the IRI model is shown in Fig. 1.1. As can be seen, IRI divides the ionosphere into four layers: from top to bottom they are the $D$, $E$, $F_1$ and $F_2$ layer. IRI uses global maps of the characteristic peak densities and heights from the CCIR and the URSI [Rush, 1989] as anchor points, and describes the vertical profile between these points by appropriate analytical functions.

## 2.5 Plasmasphere

The plasmasphere refers to the region above the topside ionosphere probably 1000 km where the domination of ions is $H^+$ instead of $O^+$. The outer boundary of the plasmasphere is known as the plasmapause, which is defined by an order of magnitude drop in plasma density. The

Table 2.2: IRI data sources and characteristics [Liang, 2017]

| Data Source | Observed quantity | Height range | Remarks |
|---|---|---|---|
| Ionosonde | electron densities | till the F2 peak | worldwide |
| Incoherent Scatter Radar (ISR) | plasma densities, temperatures, velocities | the whole ionosphere | only at a few selected locations |
| Topside souder 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 |

plasmasphere is the inner part of the magnetosphere that co-rotates with the Earth. Typical electron density values there are about $10^4$ el/cm$^3$ and drop by about 1-2 orders of magnitude at the plasmapause.

Actually, no plasma production occurs 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). According to Hargreaves [1992], when the distribution of the plasma does not change with time, in other words, the plasma is in equilibrium, the plasma density $N$ is represented with an exponential function

$$N = N_0 \exp(-h/H_P). \tag{2.14}$$

Here exist several plasmasphere models e.g., the Global Core Plasma Model (GCPM) [Gallagher et al., 2000], the IZMIRAN plasmasphere model [Gulyaeva, 2002], the Global Plasmasphere Ionosphere Density (GPID) model [Webb and Essex, 2001, 2003], the IMAGE/RPI plasmasphere model [Hu et al., 2014], and Neutrelitz electron density model (NEDM) [Hoque et al., 2022]. They have been developed theoretically, semi-empirically or fully empirically [Goto et al., 2012].

The GCPM-2000 of Gallagher et al. [2000] is an empirical description of thermal plasma densities in the plasmasphere, plasmapause, magnetospheric trough, and polar cap. It has been developed from retarding ion mass spectrometer data collected by the Dynamic Explorer satellite, includes several previously published regional models, and represents the low energy plasma distribution along the field lines from 0 to 24 hours magnetic local time world-wide. GCPM-2000 is smoothly coupled to IRI in the transition region of 400-600 km altitude. It was applied also for the plasmasphere extension of NeQuick model [Cueto et al., 2007].

The GPID model is a semi-empirical representation that was developed by Webb and Essex [2001, 2003]. GPID includes IRI below about 500 km to 600 km and extends with a theoretical plasmasphere electron density description along the magnetic field lines. Authors report on drawbacks of merging of the IRI with the plasmasphere part of GPID.

The IMAGE/RPI plasmasphere model [Huang et al., 2004] is based on radio plasma imager (RPI) [Reinisch et al., 2000] measurements of the electron density distribution along magnetic field lines. A plasmaspheric model is evolving for up to about four earth radii. The depletion

and refilling of the plasmasphere during and after magnetic storms is described in Reinisch et al. [2004]. A power profile model as function of magnetic activity was developed from RPI observations for the polar cap region [Nsumei et al., 2003].

The IZMIRAN plasmasphere model [Gulyaeva, 2002; Gulyaeva et al., 2002] is an empirical model based on whistler and satellite observations. IZMIRAN is the Institute of Terrestrial Magnetism, Ionosphere and Radiowaves Propagation. The IZMIRAN plasmasphere model presents global vertical analytical profiles of electron density (Ne) smoothly linked with the IRI electron density profile at altitude of one basis scale height above the F2 peak (400 km for electron temperature) and extended towards the plasmapause up to 36000 km (IRI-Plas).

The NEDM model is developed by superposing the Neustrelitz Plasmasphere Model (NPSM) to an ionosphere model composed of separate F and E-layer distributions. It uses the Neustrelitz TEC model (NTCM), the Neustrelitz Peak Density Model (NPDM), and the Neustrelitz Peak Height Model (NPHM) for the Total Electron Content (TEC), peak ionization, and peak height information. These models describe the spatial and temporal variability of the key parameters as a function of local time, geographic/geomagnetic location, solar irradiation, and activity. The model is developed to calculate the electron concentration at any given location and time in the ionosphere for trans-ionospheric applications.

# 3 Parameter Estimation

In order to estimate the unknown parameters, we need to build the functional relationship between the observations and the unknown parameters, besides, the statistical properties of the observations should also be determined. Therefore, in this chapter, we first introduce the fundamentals of the nonlinear problem and then followed by the inequality constrained optimization, including the Karush–Kuhn–Tucker (KKT) conditions and the Sequential Quadratic Programming (SQP) method. Our problem can be specified as a nonlinear one, thus, we will start to tackle the nonlinear problem. For more details in this Chapter, see Liang [2017].

## 3.1 Nonlinear problem

The relations between the observations and the unknown parameters can be generally defined as

$$\begin{cases} f_1(\beta_1, ..., \beta_u) = y_1^* + e_1 \\ f_2(\beta_1, ..., \beta_u) = y_2^* + e_2 \\ \qquad\qquad \vdots \\ f_n(\beta_1, ..., \beta_u) = y_n^* + e_n \end{cases} \tag{3.1}$$

where $y_i^*$ $(i = 1, ..., n)$ are the $n$ observations, $\beta_j$ $(j = 1, ..., u)$ denote the $u$ unknown parameters, $f_i(\beta_1, ..., \beta_u)$ correspond to the real-valued differentiable functions of $\beta_j$, and we have $e_i$ for observation errors.

Then, we define

$$\boldsymbol{X} = \begin{pmatrix} \frac{\partial f_1}{\partial \beta_1}\big|_{\boldsymbol{\beta}_0} & \cdots & \frac{\partial f_1}{\partial \beta_u}\big|_{\boldsymbol{\beta}_0} \\ \frac{\partial f_2}{\partial \beta_1}\big|_{\boldsymbol{\beta}_0} & \cdots & \frac{\partial f_2}{\partial \beta_u}\big|_{\boldsymbol{\beta}_0} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial \beta_1}\big|_{\boldsymbol{\beta}_0} & \cdots & \frac{\partial f_n}{\partial \beta_u}\big|_{\boldsymbol{\beta}_0} \end{pmatrix}$$

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

where the superscript $T$ means transpose. By means of the Taylor series expansion, Eq. (3.1) can be written in matrix notation as

$$\boldsymbol{X}\boldsymbol{\Delta\beta} = \boldsymbol{y}^* - \boldsymbol{y}_0 + \boldsymbol{e} = \boldsymbol{\Delta y} + \boldsymbol{e} \tag{3.2}$$

To solve such a problem, the iteration method can be adopted, which means during each iteration, the unknown vector $\boldsymbol{\Delta\beta}$, denoting the correction to the initial parameter vector $\boldsymbol{\beta}_0$, has to be estimated. Together with $\boldsymbol{\beta}_{it,0}$ from the previous iteration step, the sum of them is considered as the initial parameter vector in the next step, i.e.,

$$\boldsymbol{\beta}_{it+1,0} = \boldsymbol{\beta}_{it,0} + \boldsymbol{\Delta}\hat{\boldsymbol{\beta}}_{it} \tag{3.3}$$

where $it$ describes the iteration step,and $\Delta\hat{\boldsymbol{\beta}}_{it}$ represents the estimate of the unknown vector $\Delta\boldsymbol{\beta}$. During the iteration, the $\boldsymbol{y}_0$ and the partial derivatives in the Jacobian matrix $\boldsymbol{X}$ have to be updated according to $\boldsymbol{\beta}_{it+1,0}$.

## 3.2 Optimization with inequality constraints

According to Liang [2017], equality constraints can easily be incorporated into the adjustment system by the method of Lagrange multipliers. However, the problems get more complicated if inequality constraints need to be introduced, which 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

$$\min_{\Delta\boldsymbol{\beta}\in\mathbb{R}^u} \quad S(\Delta\boldsymbol{\beta}) \tag{3.4a}$$

$$s.t. \quad \boldsymbol{h}(\Delta\boldsymbol{\beta}) = \boldsymbol{0} \tag{3.4b}$$

$$\boldsymbol{g}(\Delta\boldsymbol{\beta}) \leq \boldsymbol{0} \tag{3.4c}$$

where $S(\Delta\boldsymbol{\beta})$ is the real-valued objective function of $\Delta\boldsymbol{\beta}$ to be minimized, the $p \times 1$ vector-valued functions $\boldsymbol{h}(\Delta\boldsymbol{\beta})$ are the equality constraints, and the $\overline{p} \times 1$ vector-valued functions $\boldsymbol{g}(\Delta\boldsymbol{\beta})$ are the inequality constraints. The symbol smaller than or equal to "$\leq$" in Eq. (3.4c) represents a component-wise operator.

It is worth mentioning that Eq. (3.4) contains both types of constraints. In the following, problems of this type will be referred to as "inequality constrained problems" and not as "inequality and equality constrained problems". This abbreviation of notation is legitimated by the fact that if both types of constraints appear, the inequalities are the much more challenging ones. Furthermore, it is easy to incorporate equality constraints in almost any algorithm for inequality constrained estimation.

Maybe the largest difference between unconstrained (or equality constrained) optimization and inequality constrained optimization is that for the latter, it is not known beforehand which constraints will influence the result. Equality constraints in general influence the result. However, this is not necessarily the case for inequalities. Due to this fact, there exist only iterative algorithms to solve inequality constrained problems. In general, such a problem is much harder to solve than an equality or unconstrained one.

As we work on the nonlinear model (3.1), we can directly set up the objective function in the least squares sense as

$$S(\Delta\boldsymbol{\beta}) = (\Delta\boldsymbol{y} - \boldsymbol{f}(\Delta\boldsymbol{\beta}))^T \boldsymbol{P} (\Delta\boldsymbol{y} - \boldsymbol{f}(\Delta\boldsymbol{\beta}))/\sigma^2 \tag{3.5}$$

where $\boldsymbol{P}$ is the weight matrix of the observations and $\sigma$ means the unknown variance factor of variance of unit weight. The problem (3.4) under consideration of the objective function (3.5) belongs to a Non-Linear Programming (NLP) problem.

In order to solve the constrained NLP problem, there are different methods, such as the penalty and augmented Lagrangian methods (e.g., Fiacco and McCormick [1990]; Rockafellar [1973]), the SQP methods (e.g., Han [1977]; Powell [1978]) and the interior-point methods (also called barrier methods, e.g., Fiacco and McCormick [1990]; Forsgren et al. [2002];) 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 SQP methods are probably the most preferable methods in NLP (see e.g., Boggs and Tolle [1995]; Schittkowski [1986]) and will be introduced in Section 3.4. In the following, the optimality conditions for constrained optimization problems will be given first, since many algorithms are based on them.

## 3.3 Karush–Kuhn–Tucker optimality conditions

For a constrained problem, the optimality conditions called KKT conditions, also known as the Kuhn–Tucker conditions [Kuhn and Tucker, 2014], have to be fulfilled at the constrained optimum point. The KKT approach generalizes the Lagrangian approach that allows only equality constraints. The Lagrangian function for the constrained optimization problem (3.4a) is defined as

$$L(\boldsymbol{\Delta\beta}, \boldsymbol{k}, \overline{\boldsymbol{k}}) = S(\boldsymbol{\Delta\beta}) + \boldsymbol{k}^T \boldsymbol{h}(\boldsymbol{\Delta\beta}) - \overline{\boldsymbol{k}}^T \boldsymbol{g}(\boldsymbol{\Delta\beta}) \tag{3.6}$$

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

Suppose $\boldsymbol{\Delta\beta}^*$ is a local minimum, there exist vectors $\boldsymbol{k}^* = [k_1^*, ..., k_p^*]^T$ and $\overline{\boldsymbol{k}}^* = [\overline{k}_1^*, ..., \overline{k}_p^*]^T$ such that the following conditions are satisfied (cf. Luenberger et al. [1984]; Nocedal and Wright [1999])

$$\nabla_{\boldsymbol{\Delta\beta}} L(\boldsymbol{\Delta\beta}^*, \boldsymbol{k}^*, \overline{\boldsymbol{k}}^*) = \nabla_{\boldsymbol{\Delta\beta}} S(\boldsymbol{\Delta\beta}^*) + \boldsymbol{H}^T \boldsymbol{k}^* - \boldsymbol{G}^T \overline{\boldsymbol{k}}^* = 0 \tag{3.7a}$$

$$\boldsymbol{h}(\boldsymbol{\Delta\beta}^*) = 0 \tag{3.7b}$$

$$\boldsymbol{g}(\boldsymbol{\Delta\beta}^*) \leq 0 \tag{3.7c}$$

$$g_i(\boldsymbol{\beta}^*)\overline{k}_i^* = 0, \; i = 1, ..., \overline{p} \tag{3.7d}$$

$$\overline{k}_i^* \geq 0, \; i = 1, ..., \overline{p} \tag{3.7e}$$

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.,

$$\begin{aligned} \boldsymbol{H} &= [\nabla_{\boldsymbol{\Delta\beta}} h_1(\boldsymbol{\Delta\beta}^*), ..., \nabla_{\boldsymbol{\Delta\beta}} h_p(\boldsymbol{\Delta\beta}^*)]^T \\ \boldsymbol{G} &= [\nabla_{\boldsymbol{\Delta\beta}} g_1(\boldsymbol{\Delta\beta}^*), ..., \nabla_{\boldsymbol{\Delta\beta}} g_{\overline{p}}(\boldsymbol{\Delta\beta}^*)]^T. \end{aligned} \tag{3.8}$$

Eqs. (3.7) are 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. Eq. (3.7a) indicates that the gradient of the Lagrangian must vanish at the optimum point. The Eqs. (3.7b) and (3.7c) suggest that all constraints are fulfilled, i.e., the optimum point is feasible with respect to all constraints. Eq. (3.7d) indicates that if the $i$-th inequality constraint is inactive, i.e. $g_i(\boldsymbol{\Delta\beta}^*) < 0$, then the corresponding Lagrange multiplier holds $\overline{k}_j^* = 0$. Therefore, the inactive constraints can be taken out from Eq. (3.7a). Eq. (3.7e) states that the Lagrange multipliers associated with inequality constraints must be non-negative. 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 (3.4) can be transformed into an equality constrained problem which can easily be solved.

## 3.4 Sequential Quadratic Programming method

### 3.4.1 Search direction – Quadratic Programming subproblem

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 (3.4) is obtained by solving the Quadratic Programming (QP) subproblem (cf. Boggs and Tolle [1995]; Han [1977]; Nocedal and Wright [1999]; Powell [1978])

$$\begin{aligned} \underset{\Delta\boldsymbol{\beta}\in\mathbb{R}^u}{min} \quad & \overline{S}(\boldsymbol{p}) = (\nabla_{\Delta\boldsymbol{\beta}} S(\Delta\boldsymbol{\beta}_{it,0}))^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \nabla^2_{\Delta\boldsymbol{\beta}\Delta\boldsymbol{\beta}} L(\Delta\boldsymbol{\beta}_{it,0}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})\boldsymbol{p} && \text{(3.9a)} \\ s.t. \quad & \boldsymbol{H}\boldsymbol{p} + \boldsymbol{h}(\Delta\boldsymbol{\beta}_{it,0}) = 0 && \text{(3.9b)} \\ & \boldsymbol{G}\boldsymbol{p} + \boldsymbol{g}(\Delta\boldsymbol{\beta}_{it,0}) \leq 0 && \text{(3.9c)} \end{aligned}$$

where $\overline{S}(\boldsymbol{p})$ denotes the objective function of optimization variables, $\boldsymbol{p}$ is the solution of the current QP problem. $\Delta\boldsymbol{\beta}$, $\boldsymbol{k}_{it}$ and $\overline{\boldsymbol{k}}_{it}$ are the estimates of parameter vector and vectors of multipliers in the current iteration, and $\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} L(\Delta\boldsymbol{\beta}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it})$ is the Hessian of the Lagrangian function (3.6).

The QP subproblem (3.9) can then be simplified as

$$\begin{aligned} \underset{\boldsymbol{\beta}\in\mathbb{R}^u}{min} \quad & \overline{S}(\boldsymbol{p}) = \mathbf{c}^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \boldsymbol{Q}\boldsymbol{p} && \text{(3.10a)} \\ s.t. \quad & \boldsymbol{H}\boldsymbol{p} + \overline{\boldsymbol{h}} = 0 && \text{(3.10b)} \\ & \boldsymbol{G}\boldsymbol{p} + \overline{\boldsymbol{g}} \leq 0 && \text{(3.10c)} \end{aligned}$$

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

$$\begin{aligned} & \mathbf{c} = \nabla_{\Delta\boldsymbol{\beta}} S(\Delta\boldsymbol{\beta}_{it,0}), \quad \boldsymbol{Q} = \nabla_{\Delta\boldsymbol{\beta}\Delta\boldsymbol{\beta}} L(\Delta\boldsymbol{\beta}_{it,0}, \boldsymbol{k}_{it}, \overline{\boldsymbol{k}}_{it}), \\ & \overline{\boldsymbol{h}} = \boldsymbol{h}(\Delta\boldsymbol{\beta}_{it,0}), \quad \overline{\boldsymbol{g}} = \boldsymbol{g}(\Delta\boldsymbol{\beta}_{it,0}). \end{aligned} \tag{3.11}$$

There are a variety of algorithms to solve Eq. (3.10) and the algorithms will be introduced in the next sections. The solution $\boldsymbol{p}_{it}$ of the QP subproblem (3.10) is then used to form a new iterate

$$\boldsymbol{\beta}_{it} = \boldsymbol{\beta}_{it,0} + \alpha_{it}\boldsymbol{p}_{it} \tag{3.12}$$

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 (3.4a) has a sufficient decrease but also the constraints (3.4b) and (3.4c) are satisfied. This is achieved by a line search to reduce a merit function; see Section 3.4.3 for details [Liang, 2017].

### 3.4.2 Active-set methods

There exists a wide variety of algorithms for solving inequality constrained optimization problems of type (3.10) [Liang, 2017]. Most of them can be subdivided into two main classes:

active-set and interior-point methods. While the algorithms of the first type follow the boundary of the feasible set, the latter follows a central path through the interior of the feasible region. Furthermore, in active-set algorithms, the constraints enter in an exact way, while interior-point methods use relaxed constraints, which are tightened in each iteration. Compared with active-set approaches, interior-point methods usually need fewer, but more expensive, iterations to solve an optimization problem [Gould, 2003]. An advantage of the active-set methods is that they allow a "warm start" (i.e. allow to specify an initial solution). This is very useful within the SQP methods for NLP problem, since each QP subproblem in the sequence is related to the previous QP problem(e.g. Maes [2010]; Wong [2011]).

Thus, we will focus on the active-set methods here. The main idea of it is to follow the boundary of the region in the parameter space, where all constraints are satisfied (i.e. the feasible region), in an iterative approach until the optimal solution is reached. This is done by extracting the constraints that hold as equality constraints (called active set) at the point of the current solution and therefore solve a sequence of equality constrained subproblems. If at least one constraint is active, the point of the optimal solution will always be at the boundary of the feasible region.

In the following, we will describe the binding-direction primal active-set method, which is a combined version of the algorithms introduced in Gill et al. [1981] and Best [1984]. Firstly, two concepts for different types of constraints and directions on which the method relies will be explained.

#### 3.4.2.1 Active, Inactive and Violated Constraints

An inequality constraint

$$g_{i1}p_1 + g_{i2}p_2 + ... + g_{iu}p_u \leq -\bar{g}_i \tag{3.13}$$

is called active if it holds as the equality constraint

$$g_{i1}p_1 + g_{i2}p_2 + ... + g_{iu}p_u = -\bar{g}_i \tag{3.14}$$

where $g_{ij}$ and $\overline{g}_i$ are taken from $\boldsymbol{G}$ and $\overline{\boldsymbol{g}}$ in Eq. (3.10c). This is equivalent to the statement that $\boldsymbol{p}$ is on the boundary of the feasible set. A constraint is inactive if the strict inequality

$$g_{i1}p_1 + g_{i2}p_2 + ... + g_{iu}p_u < -\bar{g}_i \tag{3.15}$$

holds. In this case, there is a "buffer" between $\boldsymbol{p}$ and the constraint. The constraint is called violated if

$$g_{i1}p_1 + g_{i2}p_2 + ... + g_{iu}p_u > -\bar{g}_i, \tag{3.16}$$

which must not happen throughout the iterations of the algorithm. For each feasible point $\boldsymbol{p}$, if we comprise all active inequality constraints (3.14) and all equality constraints in the $u \times p_w$ matrix $\boldsymbol{W}$ and the $p_w \times 1$ vector $\boldsymbol{w}$, a working set can be assembled, i.e.

$$\boldsymbol{W}^T\boldsymbol{p} = \boldsymbol{w}. \tag{3.17}$$

All $p_v$ inactive inequality constraints are combined to

$$\boldsymbol{V}^T\boldsymbol{p} = \boldsymbol{v}, \tag{3.18}$$

with the $u \times p_v$ matrix $\boldsymbol{V}$ and the $p_v \times 1$ vector $\boldsymbol{v}$. Taken together, both sets yield the original set of constraints

$$\underbrace{\begin{bmatrix} \boldsymbol{W}^T \\ \boldsymbol{V}^T \end{bmatrix}}_{\boldsymbol{G}} \boldsymbol{p} \leq \underbrace{\begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{v} \end{bmatrix}}_{-\bar{\boldsymbol{g}}} \tag{3.19}$$

Each point $\boldsymbol{p}(k)$ has its specific set of active and inactive constraints. Thus, it would be consequent to write

$$\boldsymbol{W}^{(k)^T} \boldsymbol{p}^{(k)} = \boldsymbol{w}^{(k)}, \tag{3.20}$$

$$\boldsymbol{V}^{(k)^T} \boldsymbol{p}^{(k)} = \boldsymbol{v}^{(k)}, \tag{3.21}$$

However, we decide to drop the iteration index $(k)$ of the sets in the following whenever it seems appropriate to keep the formulas tidy. Therefore, one should keep in mind that these sets change at each iteration.

#### 3.4.2.2 Binding, Non-binding and Infeasible Directions

Within the algorithms, the parameter vector $\boldsymbol{p}$ will be updated repeatedly with a search direction $\boldsymbol{q}$ and a step length $\alpha$

$$\boldsymbol{p}^{(k+1)} = \boldsymbol{p}^{(k)} + \alpha^{(k)} \boldsymbol{q}^{(k)}, \quad \alpha^{(k)} > 0 \tag{3.22}$$

Therefore, the determination if a step in a potential search direction $\boldsymbol{p}^{(k)}$ can violate a constraint is crucial. For inactive constraints, if $\alpha^{(k)}$ is chosen small enough, a step in any direction is possible. Thus, it is sufficient to examine solely the active constraints as they can be violated through an update in the incorrect direction. A direction is called binding if

$$\boldsymbol{W}^T \boldsymbol{q}^{(k)} = 0 \tag{3.23}$$

holds concerning all active constraints, which means, $\boldsymbol{q}^{(k)}$ is in the null-space of $\boldsymbol{W}$ and due to

$$\begin{aligned} \boldsymbol{W}^T \boldsymbol{p}^{(k+1)} &= \boldsymbol{W}^T (\boldsymbol{p}^{(k)} + \alpha^{(k)} \boldsymbol{q}^{(k)}) \\ &= \boldsymbol{W}^T \boldsymbol{p}^{(k)} + \alpha^{(k)} \underbrace{\boldsymbol{W}^T \boldsymbol{q}^{(k)}}_{=\boldsymbol{0}} \\ &= \boldsymbol{W}^T \boldsymbol{p}^{(k)} \\ &= \boldsymbol{w}, \end{aligned} \tag{3.24}$$

all constraints active in $\boldsymbol{p}^{(k)}$ will stay active after the next iteration. A direction is called non-binding concerning the constraint $i$ if

$$\boldsymbol{W}(:,i)^T \boldsymbol{q}^{(k)} < 0 \tag{3.25}$$

holds where $\boldsymbol{W}(:,i)$ means the $i$th column in $\boldsymbol{W}$. As a consequence, the constraint $i$ will become inactive in the point $\boldsymbol{p}^{(k+1)}$ (cf. Eq. (3.15))

$$\begin{aligned} \boldsymbol{W}(:,i)^T \boldsymbol{p}^{(k+1)} &= \boldsymbol{W}(:,i)^T (\boldsymbol{p}^{(k)} + \alpha^{(k)} \boldsymbol{q}^{(k)}) \\ &= \underbrace{\boldsymbol{W}(:,i)^T \boldsymbol{p}^{(k)}}_{=\boldsymbol{w}_i} + \underbrace{\alpha^{(k)}}_{>0} \underbrace{\boldsymbol{W}(:,i)^T \boldsymbol{q}^{(k)}}_{<0} \\ &< \boldsymbol{w}_i. \end{aligned} \tag{3.26}$$

For an infeasible direction concerning the constraint $i$

$$\boldsymbol{W}(:,i)^T \boldsymbol{q}^{(k)} > 0 \tag{3.27}$$

holds. The feasible region would be left through a step of arbitrary step size in this direction (cf. Eq. (3.16))

$$\begin{aligned} \boldsymbol{W}(:,i)^T \boldsymbol{p}^{(k+1)} &= \boldsymbol{W}(:,i)^T (\boldsymbol{p}^{(k)} + \alpha^{(k)} \boldsymbol{q}^{(k)}) \\ &= \underbrace{\boldsymbol{W}(:,i)^T \boldsymbol{p}^{(k)}}_{=\boldsymbol{w}_i} + \underbrace{\alpha^{(k)}}_{>0} \underbrace{\boldsymbol{W}(:,i)^T \boldsymbol{q}^{(k)}}_{>0} \\ &> \boldsymbol{w}_i. \end{aligned} \tag{3.28}$$

This must not happen throughout the algorithm.

#### 3.4.2.3 Outline of the Binding-Direction Primal Active-Set Method

Having the concepts described in the last two paragraphs at hand an outline of the Binding-Direction Primal Active-Set algorithm can be stated, which is given in Alg. 3.1.

**Algorithm 3.1:** Binding-Direction Primal Active-Set Algorithm

**Input:**
$\boldsymbol{Q}_{[u\times u]}$,$\mathbf{c}_{[u\times 1]}$ $\cdots$ Matrix and vector with the coefficients of the objective function
$\boldsymbol{H}_{[u\times p]}$,$\overline{\boldsymbol{h}}_{[p\times 1]}$ $\cdots$ Matrix and corresponding right-hand side of the inequality constraints
$\boldsymbol{G}_{[u\times \overline{p}]}$,$\overline{\boldsymbol{g}}_{[\overline{p}\times 1]}$ $\cdots$ Matrix and corresponding right-hand side of the equality constraints
$\boldsymbol{p}^{(0)}_{[u\times 1]}$ $\cdots$ feasible initial solution vector
**Output:**
$\boldsymbol{p}_{[u\times 1]}$ $\cdots$ Vector containing the solution of the QP
$\boldsymbol{k}_{w[p_w\times 1]}$ $\cdots$ Vector containing the Lagrange multipliers of the active constraints

```
1 [W, w, V, v] = findActiveConstraints(H, h̄, G, ḡ, p)
2 for i = 1 : i_max do
3     [q, k_w] = computeSearchDirection(Q, c, W, p)
4     α = computeStepLength(V, v, p, q)
5     p = p + αq
6     [W, w, V, v] = updateActiveSet(W, w, V, v, p, α)
7     if min(k_w ≥ 0) then
8         break
9 return p,k_w
```

Starting at a feasible point $\boldsymbol{p}^{(k)}$ an initial active set can be obtained by evaluating the constraints at $\boldsymbol{p}^{(k)}$ (line 1 of Alg. 3.1). It has to be taken care of that the initial matrix of active constraints $\boldsymbol{W}$ is of full column rank (cf. Wong [2011]). Therefore, it might be necessary to eliminate dependent columns from $\boldsymbol{W}$ and the corresponding values in $\boldsymbol{w}$. Afterwards a search direction $\boldsymbol{p}$, the Lagrange multipliers of the active constraints $\boldsymbol{k}_w$ (line 3), and a step length $\alpha$ (line 4) are computed, and an update of the solution is performed (line 5). Subsequently, a decision has to be made, which constraints stay in the active set and which are dropped (line 6). The last

four steps are repeated iteratively until all Lagrange multipliers associated with inequalities are non-negative. Starting with the search direction all steps are explained in more detail in the following.

#### 3.4.2.4 Search Direction q

The search direction $\boldsymbol{q}^{(k)}$ shall be computed in a way that all active constraints

$$\boldsymbol{W}^{(k)^T}\boldsymbol{p}^{(k)} = \boldsymbol{w}^{(k)}$$

are kept active after the update

$$\boldsymbol{p}^{(k+1)} = \boldsymbol{p}^{(k)} + \boldsymbol{q}^{(k)} \tag{3.29}$$

and such that the value of the objective function of problem 3.10 becomes minimal in the current subspace (i.e. the null-space of the matrix of active constraints). This ensures that we move along the boundary of the feasible region towards the minimum. We have intentionally omitted the step length $\alpha$ here, as we want to derive a search direction that directly points to the minimum of the current subspace without the need for a "scaling". The step length will come into play again in the next paragraph, as it is essential to not violate constraints inactive at the point of the current solution. Therefore, the following subproblem has to be solved. It should be noted that in this case, $\boldsymbol{q}$ is the optimization variable and $\boldsymbol{p}$ is fixed (iteration indices were omitted)

$$\min_{\boldsymbol{q}\in\mathbb{R}^u} \quad \frac{1}{2}(\boldsymbol{p}+\boldsymbol{q})^T\boldsymbol{Q}(\boldsymbol{p}+\boldsymbol{q}) + \boldsymbol{c}^T(\boldsymbol{p}+\boldsymbol{q}) \tag{3.30a}$$

$$s.t. \quad \boldsymbol{W}^T\boldsymbol{q} = \boldsymbol{0} \tag{3.30b}$$

$$\tag{3.30c}$$

The search direction can either be obtained by computing a direction to the unconstrained minimum and using projectors to map it to the null-space of the matrix of active constraints $\boldsymbol{W}$ or by an approach using the Lagrangian of the problem (cf. Best [1984]). The Lagrangian of the problem (3.30) reads

$$\begin{aligned} L(\boldsymbol{q},\boldsymbol{k}_w) &= \frac{1}{2}(\boldsymbol{p}+\boldsymbol{q})^T\boldsymbol{Q}(\boldsymbol{p}+\boldsymbol{q}) + \boldsymbol{c}^T(\boldsymbol{p}+\boldsymbol{q}) + \boldsymbol{k}_w(\boldsymbol{W}^T\boldsymbol{q}) \\ &= \frac{1}{2}\boldsymbol{q}^T\boldsymbol{Q}\boldsymbol{q} + \boldsymbol{p}^T\boldsymbol{Q}\boldsymbol{q} + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{Q}\boldsymbol{p} + \boldsymbol{c}^T\boldsymbol{p} + \boldsymbol{c}^T\boldsymbol{q} + \boldsymbol{k}_w\boldsymbol{W}^T\boldsymbol{q} \end{aligned} \tag{3.31}$$

$\boldsymbol{k}_w$ are the Lagrange multipliers linked with the active constraints. The gradients of (3.31) with respect to $\boldsymbol{q}$ and $\boldsymbol{k}_w$ read

$$\begin{aligned} \nabla_{\boldsymbol{q}} L(\boldsymbol{q},\boldsymbol{k}_w) &= \boldsymbol{Q}\boldsymbol{q} + \boldsymbol{Q}\boldsymbol{p} + \boldsymbol{c} + \boldsymbol{W}\boldsymbol{k}_w \\ &= \boldsymbol{Q}\boldsymbol{q} + \boldsymbol{g}_L + \boldsymbol{W}\boldsymbol{k}_w \\ \nabla_{\boldsymbol{k}_w} L(\boldsymbol{q},\boldsymbol{k}_w) &= \boldsymbol{W}^T\boldsymbol{q}. \end{aligned} \tag{3.32}$$

The gradient

$$\boldsymbol{g}_L = \boldsymbol{g}_L(\boldsymbol{p}) = \boldsymbol{Q}\boldsymbol{p} + \boldsymbol{c}, \tag{3.33}$$

was used as substitution. Setting the derivatives equal to zero results in

$$\begin{bmatrix} \boldsymbol{Q} & \boldsymbol{W} \\ \boldsymbol{w}^T & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{q} \\ \boldsymbol{k}_w \end{bmatrix} = \begin{bmatrix} -\boldsymbol{g}_L \\ \boldsymbol{0} \end{bmatrix} \tag{3.34}$$

As a consequence, if we solve the above equations, a search direction $\boldsymbol{q}$ can be computed which will not violate any active constraints. Furthermore, we can also obtain the Lagrange multipliers of all active constraints. They can be used to determine which active constraints should be deactivated in the later step. However, so far we have never made any statement about inactive constraints. In fact, they are used when dealing with the step length $\alpha$.

#### 3.4.2.5 Step Length $\alpha$

When we computed the search direction $\boldsymbol{q}$, it should always follow that no active constraint will be violated after the parameter update. In this section, the determination of a maximal feasible step length $\alpha \in (0, 1]$ is introduced, to ensure that also no inactive constraint will be violated. Here we need to clarify that this step length $\alpha$ is different from $\alpha_{it}$ in Eq. (3.12). The latter will be introduced in section 3.4.3. In the beginning, it shall be identified that those constraints could possibly be violated through a step in direction $\boldsymbol{q}$. Those inactive constraints for which

$$\boldsymbol{V}(:,i)^T \boldsymbol{q}^{(k)} \leq 0 \tag{3.35}$$

holds, cannot be violated through a step in direction $\boldsymbol{q}$. This can be verified by examining the product

$$\begin{aligned} \boldsymbol{V}(:,i)^T \boldsymbol{p}^{(k+1)} &= \boldsymbol{V}(:,i)^T (\boldsymbol{p}^{(k)} + \alpha^{(k)} \boldsymbol{q}^{(k)}) \\ &= \underbrace{\boldsymbol{V}(:,i)^T \boldsymbol{p}^{(k)}}_{<\boldsymbol{v}_i} + \underbrace{\alpha^{(k)}}_{>0} \underbrace{\boldsymbol{V}(:,i)^T \boldsymbol{p}^{(k)}}_{\leq 0} \\ &< \boldsymbol{v}_i. \end{aligned} \tag{3.36}$$

This is equivalent to the statement that the constraint $i$ will stay inactive in the next iteration step. Therefore, only constraints for which

$$\boldsymbol{V}(:,i)^T \boldsymbol{q}^{(k)} > 0 \tag{3.37}$$

holds, can become active or violated in the next step. Those constraints shall be called potentially violated constraints. If

$$\boldsymbol{V}(:,i)^T \boldsymbol{p}^{(k+1)} = \boldsymbol{V}(:,i)^T (\boldsymbol{p}^{(k)} + \alpha_i^{(k)} \boldsymbol{q}^{(k)}) = \boldsymbol{v}_i \tag{3.38}$$

holds for a certain $\alpha_i$, the constraint $i$ will become active in the next iteration step. Reformulating Eq. (3.38) with respect to $\alpha_i$ can be used to determine the distance to the constraint $i$ in the direction $\boldsymbol{q}$

$$\alpha_i^k = \frac{\boldsymbol{v}_i - \boldsymbol{V}(:,i)^T \boldsymbol{p}^{(k)}}{\boldsymbol{V}(:,i)^T \boldsymbol{q}^{(k)}} \tag{3.39}$$

Therefore, the maximal feasible step length $\alpha_{max}$ is restrained by the constraint $i$ that would first be violated through a step in direction $\boldsymbol{q}$. As described above, the best achievable step length would be one, as in this case, the new point $\boldsymbol{p}^{(k+1)}$ will be the minimum of the current subspace. As a consequence, the optimal step length is defined as

$$\alpha = \min(\alpha_i^{(k)}, 1), \quad \forall \left\{ i | \boldsymbol{V}(:,i)^T \boldsymbol{q}^{(i)} > 0 \right\} \tag{3.40}$$

#### 3.4.2.6 Update of the Active Set

Until now, we have explained how to obtain an optimal solution in the current subspace. In this section, the "core" of the algorithm will be introduced. That is, drop constraints from the set or include new ones, which is the strategy to update the set of active constraints. This strategy determines the subspace and therefore the sequence of equality constrained subproblems to be solved. Depending on the chosen step length $\alpha$, two cases have to be distinguished.

If $\alpha < 1$, the decision of taking a "full step" is prevented by the hitherto inactive constraint $i$. In this case, the constraint $i$ is removed from the inactive set $\boldsymbol{V}$,$\boldsymbol{v}$ and included in the active set $\boldsymbol{W}$,$\boldsymbol{w}$. As a consequence, in the next iteration step, a new equality constrained optimization problem has to be solved.

If $\alpha = 1$ holds, due to its design, the point $\boldsymbol{p}^{(k+1)}$ is the minimum of the current subspace. This means it is the point that has the minimal value of the objective function that keeps all active constraints of iteration step $(k)$ also active after step $(k+1)$. Therefore, to further minimize the objective function without removing a constraint from the active set is no longer possible. It is mandatory to identify those active constraints that prevent a decrease in the objective function. It can be shown that all inequality constraints which are associated with a negative Lagrange multiplier are the ones to deactivate. We should notice that equality constraints should never be removed from the active set. As a result, if there are no negative Lagrange multipliers, it can be proven that $\boldsymbol{p}^{(k+1)}$ is the optimal solution and the algorithm terminates since it is not possible to further reduce the value of the objective function.

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

As explained in the last sections, the search direction $\boldsymbol{p}_{it}$ in Eq. (3.12) is the solution of the QP subproblem(3.10). Here, we will introduce how to determine the step length parameter $\alpha_{it}$. However, the choice of step length is complicated by the fact that, not only do we wish to reduce the objective function, but we have to satisfy the constraints. Therefore, 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 [More and Wright, 1993]. The merit function in an objective function form is expressed as

$$\Phi(\boldsymbol{\beta}, \boldsymbol{v}) = S(\boldsymbol{\beta}) + \sum_{j=1}^{j=p} v_j \cdot h_j(\boldsymbol{\beta}) + \sum_{j=p+1}^{j=p+\overline{p}} v_j \cdot \max\left[0, g_{j-p}(\boldsymbol{\beta})\right], \tag{3.41}$$

requiring that $\alpha_{it}$ should satisfy the condition

$$\Phi(\boldsymbol{\beta}_{it} + \alpha_{it}\boldsymbol{p}_{it}, \boldsymbol{v}) < \Phi(\boldsymbol{\beta}_{it}, \boldsymbol{v}), \tag{3.42}$$

where $v_j$ are penalty parameters. Powell [1978] gives the recommend value. On the first iteration, we let $v_j = K_j = [\boldsymbol{k}^T, \overline{\boldsymbol{k}}^T]_j (j = 1, ..., p, ..., p+\overline{p})$. On the other iterations, we apply the formula

$$v_j^{(it)} = \max\left[K_j, \frac{v_j^{(it-1)} + K_j}{2}\right]. \tag{3.43}$$

Condition 3.42 can be obtained if the function

$$\phi(\alpha) = \Phi(\boldsymbol{\beta} + \alpha\boldsymbol{p}, \boldsymbol{v}) \tag{3.44}$$

decreases initially when $\alpha$ is set positive. The procedure for choosing $\alpha$ is as follows. It depends on a number $\delta$ that is usually the derivative $\phi'(0)$. We then build a sequence $\alpha_k (k = 0, 1, ...)$ until it gives a suitable value of $\alpha$. $\alpha_0$ is defined as 1 and, for $k \geq 1$, the value of $\alpha_k$ depends on $\phi_k(\alpha)$, which is the quadratic approximation of $\phi(\alpha)$ and is defined by

$$\begin{aligned} &\phi_k(0) = \phi(0) \\ &\phi_k'(0) = \delta \\ &\phi_k(\alpha_{k-1}) = \phi(\alpha_{k-1}). \end{aligned} \tag{3.45}$$

We let $\alpha_k$ be the greater of $0.1\alpha_{k-1}$ and the value of $\alpha$ that minimizes $\phi_k(\alpha)$. For each term in the sequence, we test the condition

$$\phi(\alpha_k) \leq \phi(0) + 0.1\alpha_k\delta \tag{3.46}$$

and we set the step length to $\alpha_k$ as soon as this inequality is satisfied.

# 4 Electron Density Modeling

As already described in Chapter 2, the ionosphere can be subdivided into four layers, which are the $D$, $E$, $F_1$, $F_2$ layer along height. If we take the plasmasphere into account, the electron density can be considered as the sum of all four layers in the ionosphere as well as the plasmaphere, where we get the so-called multi-layer approach.

$$\begin{aligned} N_e(h) &= N_e^D(h) + N_e^E(h) + N_e^{F_1}(h) + N_e^{F_2}(h) + N_e^P(h) \\ &= \sum_{Q=1}^{4} N_e^Q(h) + N_e^P(h) \end{aligned} \tag{4.1}$$

with $Q \in \{D, E, F_1, F_2\}$ is a summation index for layers in ionosphere and $P$ means plasmasphere.

## 4.1 General modeling of the vertical electron density distribution

For the four layers, we adopt the Chapman function as described in Eq. (2.13). While for the plasmasphere, we apply the exponential function,

$$N_e^P(h) = N_0^P \exp(-\frac{h}{H^P}) \quad (h > h_m^{F_2}) \tag{4.2}$$

where $N_0^P$ presents the basic electron density whereas $H^P$ means the scale height for the plasmasphere.

Here, it can be summarized that, if we want to model the distribution of electron density along the height, in total there are 14 key parameters

$$\mathcal{K} = \{N_m^D, h_m^D, H^D, N_m^E, h_m^E, H^E, N_m^{F_1}, h_m^{F_1}, H^{F_1}, N_m^{F_2}, h_m^{F_2}, H^{F_2}, N_0^P, H^P\}. \tag{4.3}$$

Based on $\mathcal{K}$, the MLCM can is defined. However, because of the strong correlations between the key parameters, if we estimate all of them using the classical least squares method, it will usually lead to physically unrealistic results. To aviod this, the inequality constrained optimization which is already explained in Chapter 3 is applied, and we select appropriate and realistic subsets $\mathcal{K}_1$ in which parameters are constrained by inequality constraints and then put the rest into an equality constrained subset $\mathcal{K}_2$.

## 4.2 Parametrization for key parameters - B-splines

Since we want to calculate a global 3D model of the electron density and the Chapman function is introduced in the last section, here B-spline presentations of key parameters will be introduced and explained. Other types of representations, Spherical Harmonics (SH) and Empirical Orthogonal Function (EOF) are introduced and compared by Liang [2017].

The approach based on B-splines has been adopted in this work. The main characteristic of this approach is: The basis functions are different from zero in a local environment on the sphere to allow for the modification of present data or incorporation of new data into the model without causing a global effect. Polynomial and trigonometric B-splines with local support have been selected as appropriate basis function candidates for representing the ionospheric information derived from space-geodetic observation techniques in the model space.

For the case of a 1D representation, the approximation function $s(x)$ with $x \in [x_{min}, x_{max}]$ is expressed as

$$s(x) = \sum_{k=1}^{K} d_k^J \, \phi_k^J(x) \tag{4.4}$$

where $\phi$ identifies a linearly independent set of 1D scaling functions $\phi_1^J, ..., \phi_k^J$ of level $J$ and $d_k^J$ are the associated series coefficients.

In the following, two different kinds of B-spline basis functions are introduced, namely the normalized quadratic polynomial B-splines and normalized trigonometric B-splines. Both offer excellent features for ionosphere modeling in the regional and global domains.

### 4.2.1 Normalized quadratic polynomial B-splines

For regional modeling applications, the normalized quadratic polynomial B-splines denoted by $\phi_k^{J_\Phi}(x) = N_{J_\Phi,k}^m(x)$ is considered as 1D basis function for representing the signal within a bounded interval. $N_{J_\Phi,k}^m(x)$ are usual normalized B-splines of degree $m = 2$ with

$$\sum_{k=0}^{K_\Phi - 1} N_{J_\Phi,k}^2(x) = 1 \quad x \in [0,1] \tag{4.5}$$

where $J_\Phi \in \mathbf{N}_0$ defines the B-spline resolution level and $k \in \{0, 1, ..., K_\Phi - 1\}$ identifies a specific spline function; $k$ is denoted as shift [Schmidt et al., 2015]. The model resolution is controlled by the level, i.e., the higher $J_\Phi$ is chosen, the finer the signal structures that can be resolved. The total number of B-splines is computed from $K_\Phi = 2^{J_\Phi} + 2$. The basis is deployed by an increasing sequence of $K_\Phi + 3$ so-called knot points $v_k^{J_\Phi} \in \{v_0^{J_\Phi}, v_1^{J_\Phi}, ..., v_{K_\Phi+2}^{J_\Phi}\}$ where, at the boundaries, multiple knots may be linked to a specific coordinate point. The knot interval $v_{k+1}^{J_\Phi} - v_k^{J_\Phi}$ must not mandatory be constant. The basis for normalized quadratic polynomial B-splines is defined recursively [Schumaker and Traas, 1991; Stollnitz et al., 1995] with

$$N_{J_\Phi,k}^0(x) = \begin{cases} 1 & v_k^{J_\Phi} \le x < v_{k+1}^{J_\Phi} \\ 0 & \text{otherwise} \end{cases}, \quad k = 0, ..., K_\Phi - 1 \tag{4.6}$$

$$N_{J_\Phi,k}^m(x) = \frac{x - v_k^{J_\Phi}}{v_{k+m}^{J_\Phi} - v_k^{J_\Phi}} N_{J_\Phi,k}^{m-1}(x) + \frac{v_{k+m+1}^{J_\Phi} - x}{v_{k+m+1}^{J_\Phi} - v_{k+1}^{J_\Phi}} N_{J_\Phi,k+1}^{m-1}(x), \quad m \ge 1. \tag{4.7}$$

A uniform knot sequence is established as

$$\underbrace{0 = v_0^{J_\Phi} = \cdots = v_2^{J_\Phi}}_{\text{Boundary multiplicity}} \underbrace{< v_3^{J_\Phi} < \cdots <}_{\text{Internal sequence}} \underbrace{v_{K_\Phi}^{J_\Phi} = \cdots = v_{K_\Phi+2}^{J_\Phi} = 1}_{\text{Boundary multiplicity}} \tag{4.8}$$

where $K_\Phi$ distinct knots are taken into account. At the boundaries, there is a multiplicity of $m$ knots each that allow for the endpoint-interpolation. The knot distance $v_{k+1}^{J_\Phi} - v_k^{J_\Phi}$ yields consequently $(K_\Phi - 2)^{-1}$ on the unit interval.

Figure 4.1 provides examples of polynomial B-spline functions of $m = 2$ regarding different levels $J_\Phi = 0, 1, 2, 3$. The number of splines varies with $J_\Phi$ where the subplots are related to $J_\Phi = 0 \rightarrow K_\Phi = 3$ (top left), $J_\Phi = 1 \rightarrow K_\Phi = 4$ (top right), $J_\Phi = 2 \rightarrow K_\Phi = 6$ (bottom left) and $J_\Phi = 3 \rightarrow K_\Phi = 10$ (bottom right). Special features of polynomial B-splines are in particular given by the endpoint-interpolation, i.e., adaptation of the splines to a bounded interval, and localization, i.e., compact support only within a restricted interval.

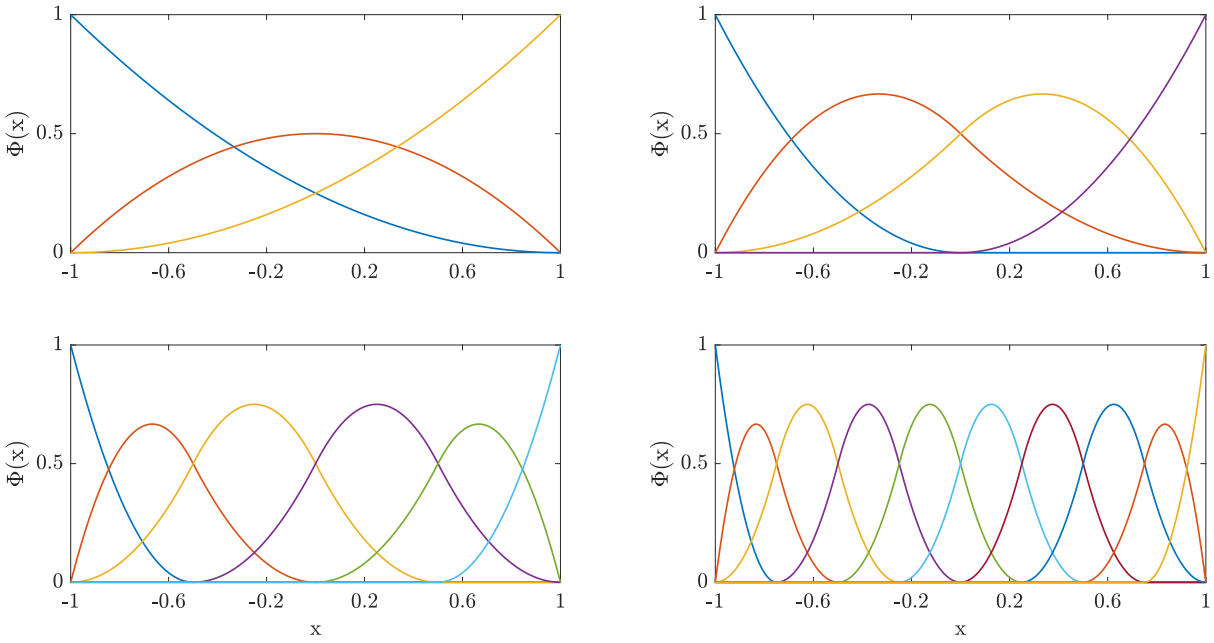


Figure 4.1: Normalized quadratic polynomial B-splines with different levels $J_\Phi = 0, 1, 2, 3$ and accordingly different number of B-splines $K_\Phi = 3, 4, 6, 10$.

It can be clearly seen from the plots, that the first two and last two splines are different with respect to the interior spline functions contributing to the endpoint-interpolation in the bounded interval. Each spline differs from zero only within a certain interval according to the level $J_\Phi$. In addition, it becomes visible that always three adjacent splines are overlapping, i.e., a single data point contributes to the determination of exactly three B-spline coefficients. The model resolution defined by $J_\Phi$ should be adapted to the input data density to overcome data gaps. For the case of a homogeneous data sampling, Schmidt et al. [2011] derived the relation

$$\Delta si = \frac{si_{max} - si_{min}}{K_\Phi - 1} \tag{4.9}$$

with the sampling $si$ on the interval $[si_{min}, si_{max}]$. This formulation can be transformed to

$$J_\Phi < \log_2\left(\frac{si_{max} - si_{min}}{\Delta si} - 1\right). \tag{4.10}$$

### 4.2.2 Normalized trigonometric B-splines

As a second type of basis function, normalized trigonometric B-splines $T^m_{J_T,k}(x)$ of order $m = 3$ are introduced in this section [Schumaker, 1981; Schumaker and Traas, 1991].

Trigonometric B-splines with resolution level $J_T \in \mathbf{N}_0$ are defined on a circle in the closed interval $[0, 2\pi)$ and have no knot multiplicity but comply with the constraint of $s(0) = s(2\pi)$. Periodic trigonometric B-splines are particularly suitable for global modeling applications due to special properties in the definition on the $[0, 2\pi)$ interval with "wrapping-around" condition and localization, i.e., compact support only within a restricted interval as also provided by the polynomial B-splines. The number of spline functions is computed from $K_T = 3 \cdot 2^{J_T}$ distributed on the basis interval, meaning that the interrupted boundary splines are completed by the corresponding opposing sub-spline to enable periodicity. Similar to the polynomial B-splines, the distance between two consecutive knots $v^{J_T}_k$ and $v^{J_T}_{k+1}$ for $k = 0, 1, ..., K_T - 1$ reads

$$h_{J_T} = \frac{360^\circ}{K_T} = \frac{120^\circ}{2^{J_T}}. \tag{4.11}$$

with the non-decreasing sequence of distinct knots

$$0 = v^{J_T}_0 < v^{J_T}_1 < \cdots < v^{J_T}_{K_T} < v^{J_T}_{K_T-1} < 2\pi. \tag{4.12}$$

and additional knots

$$v^{J_T}_{K_T+i} = v^{J_T}_i + 360^\circ \quad \text{for} \quad i = 0, 1, 2. \tag{4.13}$$

Following Lyche and Schumaker [2000], the functions are defined as

$$M_{J_T,k_2}(\lambda) = T^3_{J_T,k}(\lambda) = T^3_{h_{J_T}}(v - v^{J_T}_k) \tag{4.14}$$

with setting $h_{J_T} = h$ and $v - v^{J_T}_k = \Theta$ for the sake of simplification, thus, the function $T^3_{h_{J_T}}(v - v^{J_T}_k) = T^3_h(\Theta)$ can be computed via

$$T^3_h(\Theta) = \begin{cases} \dfrac{\sin^2(\Theta/2)}{\sin(h/2)\sin(h)} & \text{for} \quad 0 \le \Theta < h \\ \dfrac{1}{\cos(h/2)} - \dfrac{\sin^2((\Theta - h)/2) + \sin^2((2h - \Theta)/2)}{\sin(h/2)\sin(h)} & \text{for} \quad h \le \Theta < 2h \\ \dfrac{\sin^2((3h - \Theta)/2)}{\sin(h/2)\sin(h)} & \text{for} \quad 2h \le \Theta < 3h \\ 0 \quad \text{others.} \end{cases} \tag{4.15}$$

Finally, the basis functions are obtained as

$$T^3_{J_T,k}(\lambda) = \begin{cases} M_{J_T,k}(\lambda) & \text{for} \quad k = 0, ..., K_T - 3 \\ M_{J_T,k}(\lambda) + M_{J_T,k}(\lambda - 360^\circ) & \text{for} \quad k = K_T - 2, K_T - 1. \end{cases} \tag{4.16}$$

A set of trigonometric B-splines of $m = 3$ for different levels $J_T = 0 \rightarrow K_T = 3$ (top left), $J_T = 1 \rightarrow K_T = 6$ (top right), $J_T = 2 \rightarrow K_T = 12$ (bottom left) and $J_T = 3 \rightarrow K_T = 24$ (bottom right) is provided by Fig. 4.2. Similarities to the polynomial B-spline functions can be found in the local support and the overlapping of three splines in each point along $x$. The main difference is related to the boundary splines where, in contrast to the endpoint-interpolation

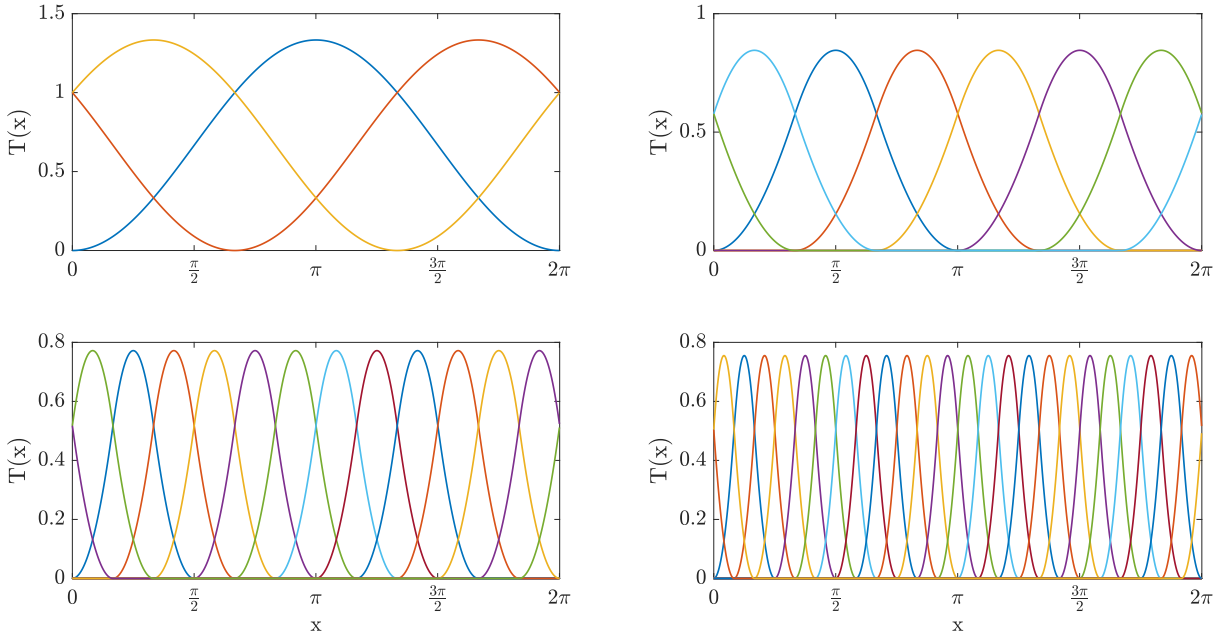


Figure 4.2: Trigonometric B-splines with different levels $J_T = 0, 1, 2, 3$ and accordingly different number of B-splines $K_T = 3, 6, 12, 24$.

of polynomial splines, the periodicity is visible. Although different colors have been chosen for each function, it can be clearly seen that every sub-spline at one boundary is continued at the opposite boundary.

Similarly to Eq. (4.10), the relation of the data density to the B-spline level can be derived from [Schmidt et al., 2011]

$$J_T < \log_2\left(\frac{si_{max} - si_{min}}{3\Delta si}\right). \tag{4.17}$$

### 4.2.3 B-spline tensor products

$S_\omega$ shall identify a unit sphere as

$$R_{S_\omega} := \{(\varphi, \lambda) : -\frac{\pi}{2} \leq \varphi \leq \frac{\pi}{2} \quad \text{and} \quad 0 \leq \lambda < 2\pi\} \tag{4.18}$$

with polar coordinates $\varphi$ and $\lambda$ in an angular system mapped on a rectangle $R_{S_\omega}$ in a 2D space $\mathbb{R}^2$. The representation of a subsurface of $S'_\omega$ defined within $\varphi \in [\varphi_{min}, \varphi_{max}]$ and $\lambda \in [\lambda_{min}, \lambda_{max}]$ in the rectangular domain can be obtained with

$$R_{S_\omega} := \{(\varphi, \lambda) : \varphi_{min} \leq \varphi_{max} \leq \frac{\pi}{2} \quad \text{and} \quad \lambda_{min} \leq \lambda \leq \lambda_{max}\}. \tag{4.19}$$

In order to represent multidimensional information on the rectangular modeling surfaces $R_{S_\omega}$ or $R'_{S_\omega}$ as introduced by Eqs. (4.18) and (4.19), tensor products of B-spline base functions shall be applied in an orthogonal coordinate system.

For a 2D case representation, the approximation function $s$ can be constructed from

$$s(x_1, x_2) = \sum_{k_1=0}^{K_1-1} \sum_{k_2=0}^{K_2-1} d_{k_1,k_2}^{J_1,J_2} \phi_{k_1}^{J_1}(x_1) \phi_{k_2}^{J_2}(x_2). \tag{4.20}$$

Here, tensor products of two linearly independent 1-D B-spline functions $\phi_{k_1}^{J_1}$ and $\phi_{k_2}^{J_2}$ have been introduced together with the corresponding series coefficients $d$. It should be noticed that $\phi_1$ and $\phi_2$ may differ but can also be of the same type.

So far, the B-spline levels and numbers were expressed in relation to the B-spline type $(J_T, K_T, J_\Phi, K_\Phi)$. From now on, the identification will be based on indices $(J_1, K_1, J_2, K_2, ...)$ to distinguish between B-splines of the same kind in the tensor product notation. Furthermore, the degree of polynomial B-splines will permanently be considered as $m = 2$.

At first, polynomial B-splines are chosen on both, the $x_1$ and $x_2$ interval, with

$$x_1 \rightarrow \phi_1 = \Phi_{k_1}^3 \quad \text{and} \quad x_2 \rightarrow \phi_2 = \Phi_{k_2}^2 \tag{4.21}$$

for the levels $J_1 = 3$ and $J_2 = 2$. The corresponding basis is depicted in Fig. 4.3. Both plots show the support area spanned by the tensor product of two polynomial B-splines which are emphasized by thick lines. As can be clearly seen from the left illustration of Fig. 4.3, an ellipse shaped support area is spanned by $\phi_4^3(x_1)$ (green) and $\phi_2^2(x_2)$ (orange). Choosing the same level in both directions naturally would result in a circle shaped area. The subfigure on the right, exemplarily depicts the support area at the boundary for $\phi_4^3(x_1)$ (green) and $\phi_0^2(x_2)$ (blue) constraint by endpoint-interpolation on the $x_2$ axis.

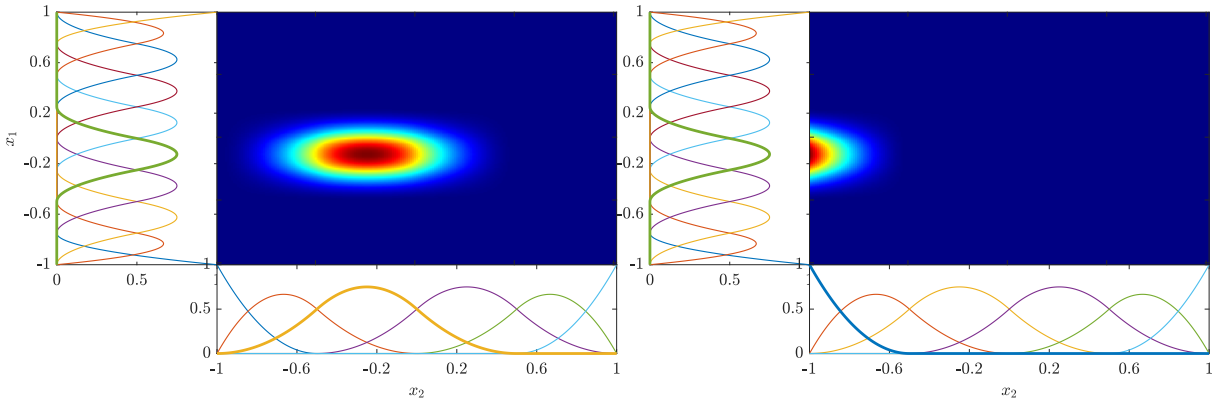


Figure 4.3: Polynomial B-splines $\Phi_{k_1}^3(x_1)$ and $\Phi_{k_2}^2(x_2)$ of degree $m = 2$ with different levels $J_1 = 3$ and $J_2 = 2$. $\phi_2^2(x_2)$ (left), $\phi_0^2(x_2)$ (right) and $\phi_4^3(x_1)$ are emphasized to show the support area.

In the next step, the 2D basis is generated from the combination of polynomial and trigonometric B-spline functions. The basis is defined as

$$x_1 \rightarrow \phi_1 = \Phi_{k_1}^3 \quad \text{and} \quad x_2 \rightarrow \phi_2 = T_{2,k_2}^3 \tag{4.22}$$

for different levels $J_1 = 3$ and $J_2 = 2$ with endpoint-interpolation in the $x_1$ and continuity in the $x_2$ direction. Figure 4.4 shows the support area for $\phi_4^3(x_1)$ and $T_7^2(x_2)$ in this spline constellation. The right subplot additionally shows the 3D shape of a 2D tensor B-spline product basis related to the emphasized splines of the left graph.

Compared to the first two parametrization approaches - SH and EOF, B-spline functions can allow for an appropriate handling of the heterogeneous data distribution, including data gaps [Schmidt et al., 2011]. Therefore, B-splines are used as basis functions to represent ionospheric key parameters in this study.

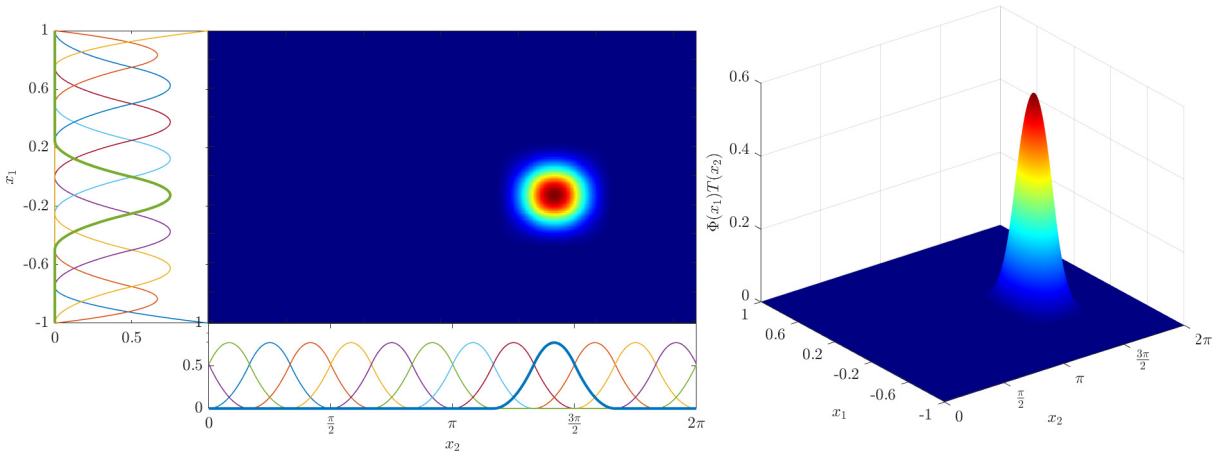

Figure 4.4: Combination of trigonometric B-splines with level $J_1 = 3$ and polynomial B-splines with level $J_2 = 2$. A specific spline combination identified by $k_1 = 4$ and $k_2 = 7$ has been highlighted and plotted in the center part of the left subplot. Accordingly, a 3D representation of the tensor product is given on the right hand side.

## 4.3 Linearized observation equation system

Since now we already introduced our modeling equation, it is worth substituting the Chapman function into two observation Eqs. (3.2) and (3.4). In other words, the inequality constrained optimization of the nonlinear problem will be applied in this section for MLCM.

In the following, we will only consider the $F_2$ layer, and thus only the key parameters $N_m^{F_2}$, $h_m^{F_2}$ and $H^{F_2}$ are included in the system in order to avoid too long equations and expressions. Taking Eq. (3.2) into account, we obtain

$$N_e^* - N_e(\boldsymbol{x}_0) + \boldsymbol{e} = \left[\frac{\partial N_e(\boldsymbol{x}_0)}{\partial N_m^{F_2}}\Big|_{\boldsymbol{x}_0}\right] \Delta N_m^{F_2} + \left[\frac{\partial N_e(\boldsymbol{x}_0)}{\partial h_m^{F_2}}\Big|_{\boldsymbol{x}_0}\right] \Delta h_m^{F_2} + \left[\frac{\partial N_e(\boldsymbol{x}_0)}{\partial H^{F_2}}\Big|_{\boldsymbol{x}_0}\right] \Delta H^{F_2} \tag{4.23}$$

where $N_e^*$ is the observation; $\boldsymbol{x}_0 = \left[N_m^{F_2}|_0, h_m^{F_2}|_0, H^{F_2}|_0\right]$ represents the linearization point, while $N_e(\boldsymbol{x}_0)$ denotes the approximate value of the electron density at that point.

Since the key parameters shall be presented using B-splines, we have

$$\kappa_i(\varphi, \lambda) + v(\varphi, \lambda) = \sum_{k_1=0}^{K_1-1} \sum_{k_2=0}^{K_2-1} (d_{k_1,k_2}^{J_1,J_2})_i \; \phi_{k_1}^{J_1}(\varphi) \; \phi_{k_2}^{J_2}(\lambda) \tag{4.24}$$

with $v$ denotes the truncation error. According to the equation above, Eq. (4.23) can be adapted in order to let the B-spline coefficients of the key parameters being the unknown parameters, which leads to:

$$N_e^* - N_e(\boldsymbol{\kappa}(\boldsymbol{x}_0)) + \boldsymbol{e} = \sum_{i=1}^{I} \left( \left[\frac{\partial N_e}{\partial \kappa_i}\Big|_{\boldsymbol{\kappa}(\boldsymbol{x}_0)}\right] \left[\frac{\partial \kappa_i}{\partial \boldsymbol{d}_i}\Big|_{\boldsymbol{x}_0}\right] \right)^T \Delta d_i \tag{4.25}$$

where $I$ is the number of key parameters. Here $I = 3$ for only 3 key parameters in the $F_2$ layer as an example while in the next chapter, $I = 14$ for all Chapman key parameters are chosen. $\kappa$ denotes the key parameters whereas $d$ gives the B-spline coefficients.

In the following, the partial derivatives of the key parameters and B-spline coefficient with respect to the Chapman function will be derived, which are in fact the most important part on the right side of Eq. (4.25).

1. Partial derivatives of the Chapman function for $N_m^{F_2}$ and the B-spline coefficients:

$$\frac{\partial N_e}{\partial N_m^{F_2}} = \exp(\frac{1}{2}(1 - \frac{h - h_m^{F_2}}{H^{F_2}} - \exp(-\frac{h - h_m^{F_2}}{H^{F_2}}))) \tag{4.26}$$

$$\frac{\partial N_e}{\partial N_m^{F_2}} \frac{\partial N_m^{F_2}}{\partial (d_{k_1,k_2}^{J_1,J_2})_{N_m^{F_2}}} = \frac{\partial N_e}{\partial N_m^{F_2}} \phi_{k_1}^{J_1}(\varphi)\, \phi_{k_2}^{J_2}(\lambda) \tag{4.27}$$

2. Partial derivatives of the Chapman function for $h_m^{F_2}$ and the B-spline coefficients:

$$\frac{\partial N_e}{\partial h_m^{F_2}} = \begin{cases} \frac{N_m^{F_2}}{2H^{F_2}} \frac{\partial N_e}{\partial N_m^{F_2}} (1 - \exp(-\frac{h - h_m^{F_2}}{h^{F_2}})) + \frac{N_0^P}{H^P} \frac{\partial N_e}{\partial N_0^P} & \text{if} \quad h \geq h_m^{F_2} \\ \frac{N_m^{F_2}}{2H^{F_2}} \frac{\partial N_e}{\partial N_m^{F_2}} (1 - \exp(-\frac{h - h_m^{F_2}}{h^{F_2}})) - \frac{N_0^P}{H^P} \frac{\partial N_e}{\partial N_0^P} & \text{else} \end{cases} \tag{4.28}$$

$$\frac{\partial N_e}{\partial h_m^{F_2}} \frac{\partial h_m^{F_2}}{\partial (d_{k_1,k_2}^{J_1,J_2})_{h_m^{F_2}}} = \frac{\partial N_e}{\partial h_m^{F_2}} \phi_{k_1}^{J_1}(\varphi)\, \phi_{k_2}^{J_2}(\lambda) \tag{4.29}$$

3. Partial derivatives of the Chapman function for $H^{F_2}$ and the B-spline coefficients:

$$\frac{\partial N_e}{\partial H^{F_2}} = N_m^{F_2} \frac{\partial N_e}{\partial N_m^{F_2}} \frac{h - h_m^{F_2}}{2(H^{F_2})^2} (1 - \exp(-\frac{h - h_m^{F_2}}{h^{F_2}})) \tag{4.30}$$

$$\frac{\partial N_e}{\partial H^{F_2}} \frac{\partial H^{F_2}}{\partial (d_{k_1,k_2}^{J_1,J_2})_{H^{F_2}}} = \frac{\partial N_e}{\partial H^{F_2}} \phi_{k_1}^{J_1}(\varphi)\, \phi_{k_2}^{J_2}(\lambda) \tag{4.31}$$

For other key parameters, the partial derivatives can be derived similarly according to the equations above.

## 4.4 Procedure of modeling

The main steps of the computation procedure for the discussed electron density model are shown in Fig. 4.5. The observations can be taken from satellite measurements, simulation data and their combinations. And the parameters need to be divided into two subsets with equality and inequality constraints, respectively. After setting the initial values of unknown parameters, the observation equation system is established including the computation of electron density differences and the Jacobian matrix. The corrections of the B-spline coefficients for key parameters $\Delta d$ can be obtained after the estimation based on Inequality Constrained Optimization (ICO), which can be used for correcting the unknown parameters. Here, a check for $\Delta d$ is performed to decide whether to finish the procedure. If it is smaller than the threshold, the B-spline coefficients are transformed into key parameters and the validation is performed including the comparisons of key parameters and electron densities before and after estimation, as well as using external validation data. Or the observation equation system is updated and the computation is repeated until the correction is smaller than the threshold.

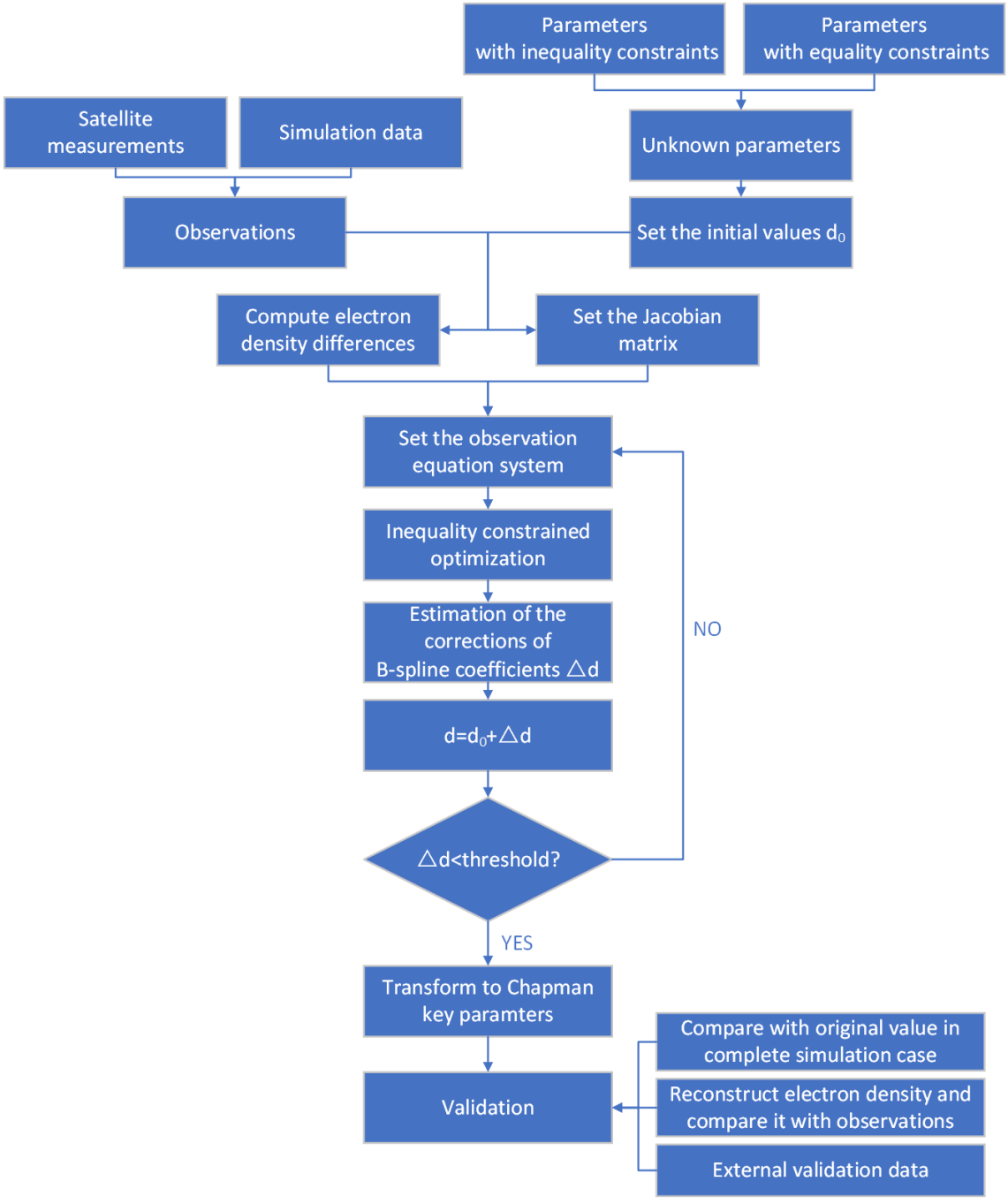


Figure 4.5: Procedure developed within this thesis

# 5 Numerical Analysis

The main contributions of this study are described in this chapter and can be subdivided into four scenarios corresponding to four sections, respectively, which are different in terms of the number of key parameters constrained by inequality constraints or the source of input data, see Table 5.1. However, the coverage, the spatial resolution in latitude, longitude and height, and the levels of B-splines are the same for different scenarios. A more detailed description of these configurations will be given in Section 5.1.

## 5.1 One parameter with inequality constraints

In this section, we adopt a scenario that only $N_m^{F_2}$ is unknown and, thus, constrained by inequality constraints while the other parameters are given with equality constraints.

The input data is the electron density computed based on the key parameters provided via IRI2012 and Chapman function. Specifically speaking, according to Eq. (4.24) the key parameters provided by IRI2012 are transformed into sets of B-spline coefficients, which are reused to compute the new key parameters. Then, with the use of the Chapman function, the new key parameters are involved to derive the electron density which is the input observation in the experiment. It is worth mentioning that $H^{F_2}$ is computed as:

$$H^{F_2} = \frac{VTEC}{4.13 \cdot N_m^{F_2}} \tag{5.1}$$

where the VTEC values are taken from VTEC models for example 'othg'. In this and the next two sections, the main purpose is to achieve a so-called closed loop simulation, which algorithm is shown in Fig. 5.1. In this validation, step 1 means the key parameters for the Chapman profile function need to be presented by B-splines as much as possible. Therefore, an inverse modeling and a forward modeling are performed in order to drop the truncation error. Then in the second step, the key parameters are estimated using the electron density computed from $\hat{\kappa}$ and we get $\widetilde{\kappa}$. Finally, the evaluation of the closed loop validation can be performed according to the differences between the $\hat{\kappa}$ (which is called 'original value' later) and $\widetilde{\kappa}$ (which is called 'estimated value' later), as well as the differences between original electron density $N_e$ and reconstructed electron density $\widetilde{N_e}$.

In this thesis, we choose the key parameters provided by IRI2012 with a spatial resolution of $5^\circ \times 5^\circ$ with respect to latitude $\varphi$ and longitude $\lambda$ at 12:00, March 20, 2015. In the vertical profile, we apply an irregular sampling strategy, which is 5 km between 50 km and 540 km and 20 km between 540 km and 1000 km. In total there are 122 samples along one vertical profile. This strategy is adopted because below 540 km, there are $D$, $E$, $F_1$ and part of $F_2$ layers, which leads to much more variations and needs more samples than the ionosphere above 540 km.

Table 5.1: Contents of the scenarios

| Scenario | Number of key parameters with inequality constraints | Activeness of constraints | Coverage | Constraints for each gird point | Observations | Noise |
|---|---|---|---|---|---|---|
| 1-1 | One | Inactive | Global | Same | Simulation | 0 |
| 1-2 | One | Active | Global | Same | Simulation | 0 |
| 2-1 | Three | Inactive | Single profile | \ | Simulation | 0 |
| 2-2 | Three | Active | Single profile | \ | Simulation | 0 |
| 2-3 | Three | Inactive | Global | Same | Simulation | 0 |
| 2-4 | Three | Active | Global | Same | Simulation | 0 |
| 2-5 | Three | Inactive | Global | Different | Simulation | 5% |
| 3-1 | Nine | Inactive | Single profile | \ | Simulation | 0 |
| 3-2 | Nine | Active | Single profile | \ | Simulation | 0 |
| 3-3 | Nine | Inactive | Single profile | \ | Simulation | 5% |
| 3-4 | Nine | Active | Single profile | \ | Simulation | 5% |
| 3-5 | Nine | Inactive | Global | Different | Simulation | 0 |
| 3-6 | Nine | Inactive | Global | Different | Simulation | 5% |
| 4-1 | Nine | \ | Global | Different | Separability approach | \ |

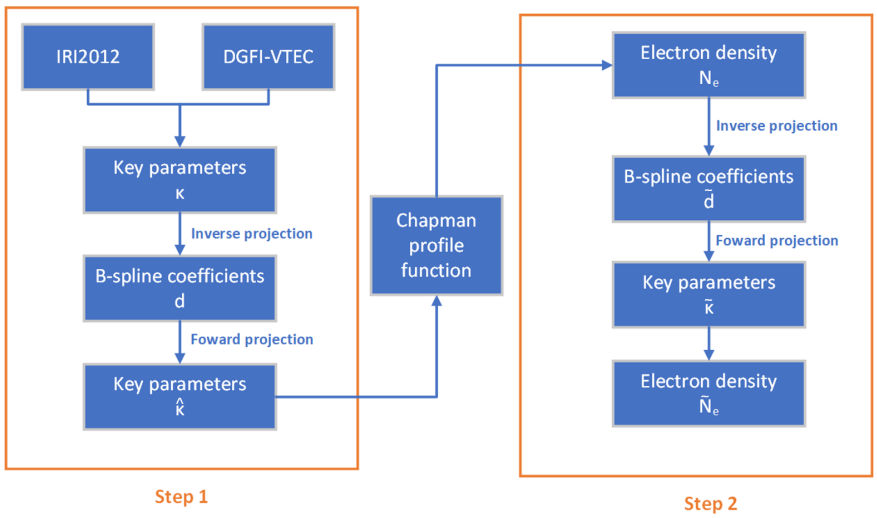


Figure 5.1: Flow chart of closed loop validation

For the key parameters, their maps are characterized by a spatial resolution of $5^\circ \times 5^\circ$ with respect to latitude $\varphi$ and longitude $\lambda$ and the B-spline coefficients are represented by polynomial B-splines of level 4 and trigonometric B-splines of level 3 for latitude and longitude, respectively.

### Scenario 1-1

Since now only one key parameter is constrained with inequality constraints, the classical least squares method can also be applied as a reference strategy if we now consider $N_m^{F_2}$ as the only unknown parameter. The results are plotted in Fig. 5.2 where Electron Density Unit (EDU) is defined as 1 EDU $= 10^{-12}$ el/m$^3$.

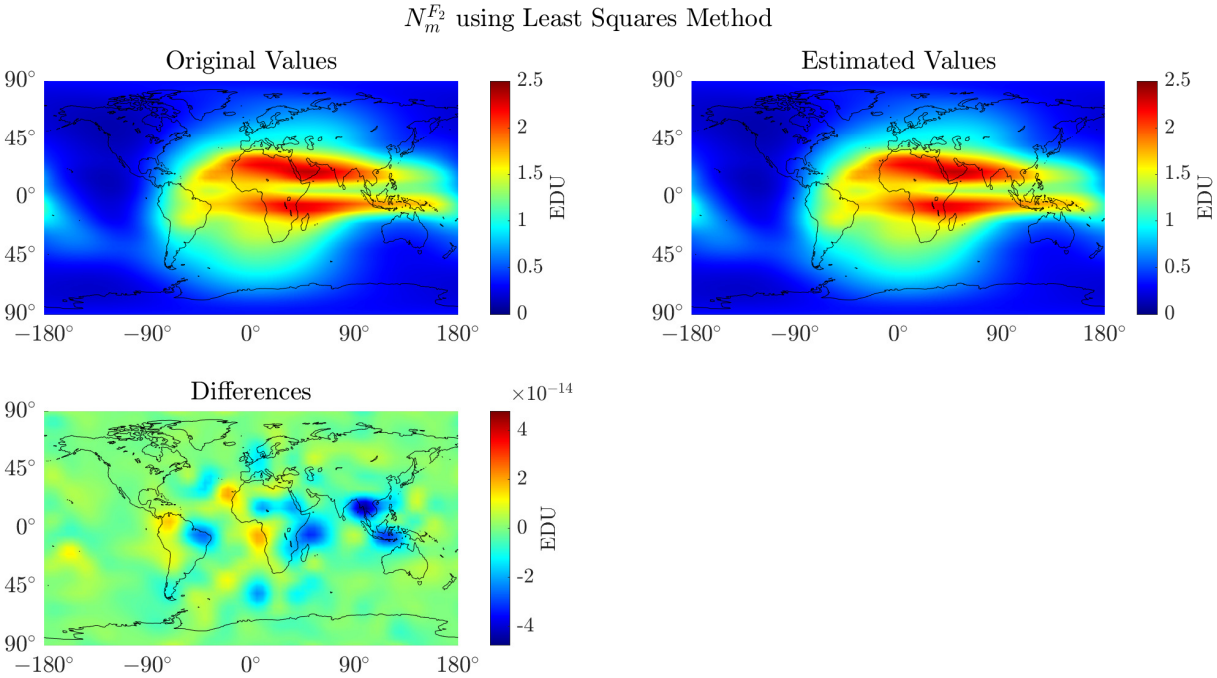


Figure 5.2: Maps of the original key parameter $N_m^{F_2}$, the estimated key parameter from the least squares method and the differences between them.

The top left panel shows the original values for electron density which are calculated from the key parameter provided by IRI2012. Since the $F_2$ layer contains around 70% to 80% of the electron density in the ionosphere [Liang, 2017], $N_m^{F_2}$ can roughly show the distribution of electron density horizontally. It can be found that there are obviously two crests on both sides of the equator, which is the so-called equatorial anomaly described in Section 2.1.2. Furthermore, according to the distribution, we can conclude that the activeness of the ionosphere is highly dependent on solar activity. In Africa, since it is noon, the electron density is high while in night regions (e.g. the Pacific), the electron density is low.

The estimated results of the least squares method are shown in the top right panel. Compared with the original values, the estimated values follow the same distribution, and their differences are given in the bottom left panel. The differences are all smaller than $10^{-13}$ EDU and they mainly come from the computation error during the matrix inversion process. The maximum difference occurs at the peak of the original value. Besides, the least squares method will produce large differences at the region where the original values have large variations,

e.g. Africa and South Asia shown in Fig. 5.2. In contrast, in stable areas, the estimated values are closer to the original values.

Now, the estimated results of ICO are presented in Fig. 5.3. In the current configuration, we set the lower bound to be equal to 0.1 EDU and use 2.5 EDU as the upper bound, which means

$$0.1 \text{ EDU} \leq N_m^{F_2} \leq 2.5 \text{ EDU} \tag{5.2}$$

holds for every $N_m^{F_2}$. However, the minimum and maximum of the original values are 0.116 EDU and 2.353 EDU, respectively. This means, both the lower bound and upper bound will stay inactive during the estimation process. The differences between the original value and the ICO results are illustrated in the bottom left panels in Fig. 5.3. It can be found that using the algorithm of ICO comes to differences lower than $10^{-14}$ EDU, which is better than the results using Least Squares Method (LSM). For the distribution of the differences, a similar conclusion can be found, which is in the areas containing large variations of the original value, the difference is also large.

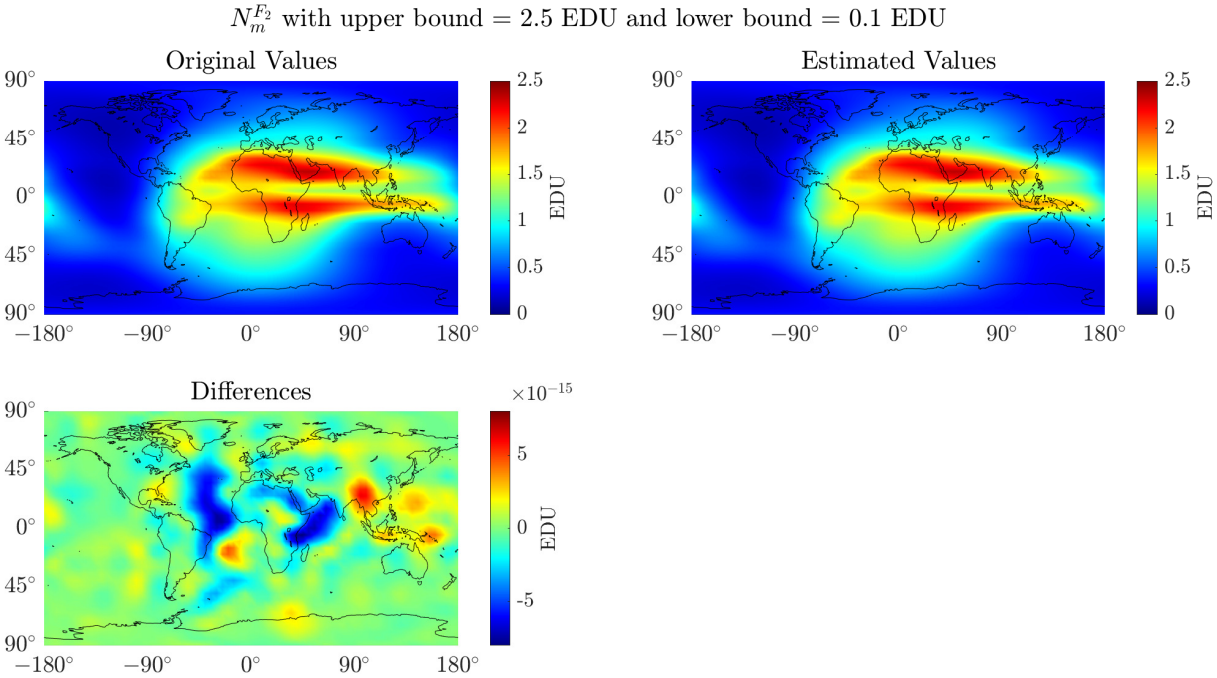


Figure 5.3: Maps of the original key parameter $N_m^{F_2}$, the estimated key parameter in scenario 1-1 and the differences between them.

## Scenario 1-2

Afterwards, we now give another lower bound and upper bound for the $N_m^{F_2}$ estimation:

$$0.12 \text{ EDU} \leq N_m^{F_2} \leq 2.3 \text{ EDU} \tag{5.3}$$

should hold for every $N_m^{F_2}$ in the grid. Now, the minimum value of the original values is smaller than the lower bound whereas the maximum value of the original values is larger

than the upper bound. This means either the lower bound or the upper bound will be active in certain areas.

The results using the above settings are plotted in Fig. 5.4, and we could clearly observe the influences of the active bounds. In the Mideast area, where the original values are larger than 2.3 EDU, the estimated results are limited to 2.3 EDU while in the Northern area of North America, the estimated results are limited to 0.12 EDU due to the lower constraints. However, for other regions, since they are in the interval between the lower bound and upper bound, the estimated results are close to the original values. In other words, the differences in those areas are close to 0. Furthermore, it is worth mentioning that, there are oscillations only around the areas where the constraints are active. This is because the representation of B-spline functions is characterized by a compact support.

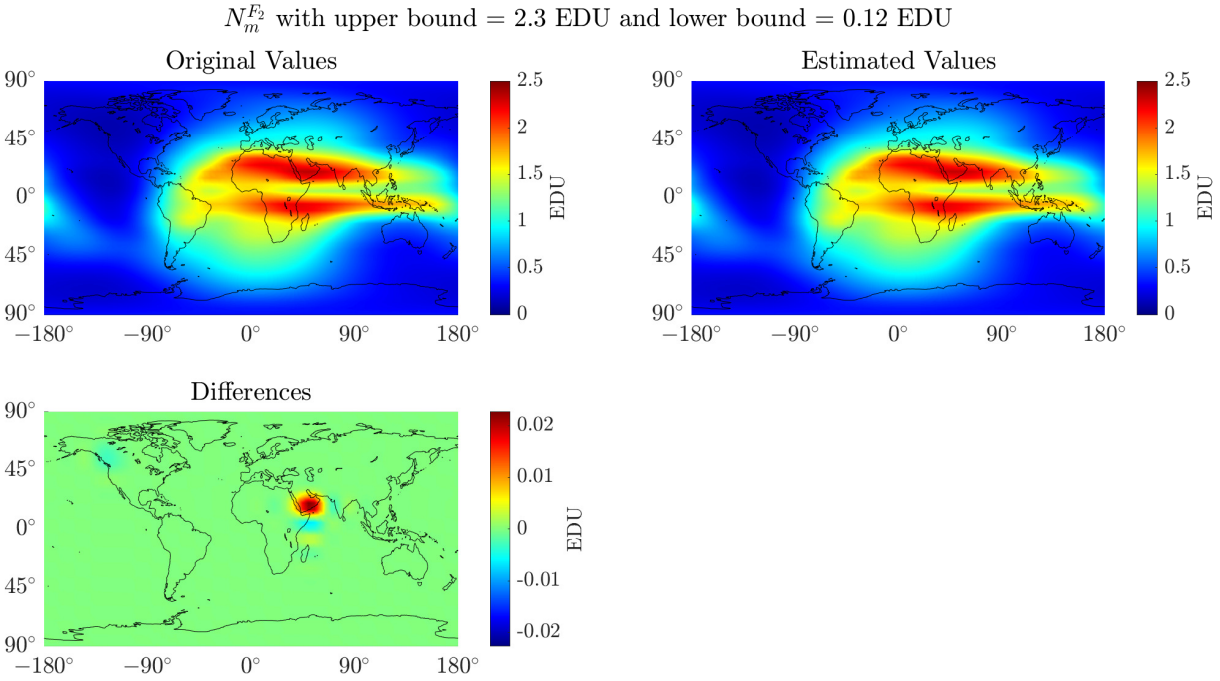


Figure 5.4: Maps of the original key parameter $N_m^{F_2}$, the estimated key parameter in scenario 1-2 and the differences between them

## 5.2 Three parameters with inequality constraints

In this section, we start another scenario, in which three parameters are estimated with inequality constraints and they are $N_m^{F_2}$, $h_m^{F_2}$ and $H^{F_2}$. Since they have significant correlations with each other, the classical LSM can lead to physically unrealistic results [Lalgudi Gopalakrishnan and Schmidt, 2022]. However, besides the comparisons of key parameters and electron density before and after the estimation, the Monte Carlo method is adopted to simulate the standard deviation of the estimated parameters in case 5% noise is added during the procedure.

Table 5.2: Values of the key parameters

| Key parameter | Value | Key parameter | Value |
|---|---|---|---|
| $N_0^P$ | 0.025 EDU | $H^{F_1}$ | 40 km |
| $H^P$ | 80 km | $N_m^E$ | 0.1 EDU |
| $N_m^{F_2}$ | 2.5 EDU | $h_m^E$ | 100 km |
| $h_m^{F_2}$ | 480 km | $H^E$ | 20 km |
| $H^{F_2}$ | 80 km | $N_m^D$ | 0.05 EDU |
| $N_m^{F_1}$ | 0.2 EDU | $h_m^D$ | 80 km |
| $h_m^{F_1}$ | 250 km | $H^D$ | 10 km |

Table 5.3: Relevant values in scenario 2-1

| Key parameter | Original value | Lower bound | Upper bound | Estimated value | Difference |
|---|---|---|---|---|---|
| $N_m^{F_2}$ | 2.5 EDU | 2.2 EDU | 2.8 EDU | 2.5 EDU | $1.8 \times 10^{-11}$ EDU |
| $h_m^{F_2}$ | 480 km | 250 km | 500 km | 480 km | $-5.4 \times 10^{-9}$ km |
| $H^{F_2}$ | 80 km | 75 km | 120 km | 80 km | $-3.7 \times 10^{-9}$ km |

### Scenario 2-1

We first start with one vertical profile estimation, which means only one grid point of the global grid will be considered and the electron density distribution along the height is illustrated. In this step, we drop the data from IRI2012 and Table 5.2 gives the values of the key parameters defined by ourselves, which will be used in the current scenario using the single vertical profile estimation.

In the following, the original values, lower and upper bounds, the estimated values as well as the differences between the original values and the estimated values are given in Table 5.3. It can be found that in the current step, all of the original values are in the region between lower bound and upper bound, and the results are nearly the same as the original values.

Fig. 5.5 describes the original electron density computed with the original key parameters and the reconstructed electron density computed with the estimated key parameters. And their differences are plotted in Fig. 5.6. The average absolute difference is $2.34 \cdot 10^{-11}$ EDU. According to the distribution of the differences, we can also conclude that the ICO will generate large deviations when the original values are large.

### Scenario 2-2

Compare to the constraints given in Table 5.3, we now set other constraints for the three key parameters, which are presented in Table 5.4 in accompany with its corresponding estimated results. Here, we believe that the peak height of the $F_2$ layer should not be higher than 450km. This means the upper bound of $h_m^{F_2}$ will become active during the estimation while the other five keep staying inactive.

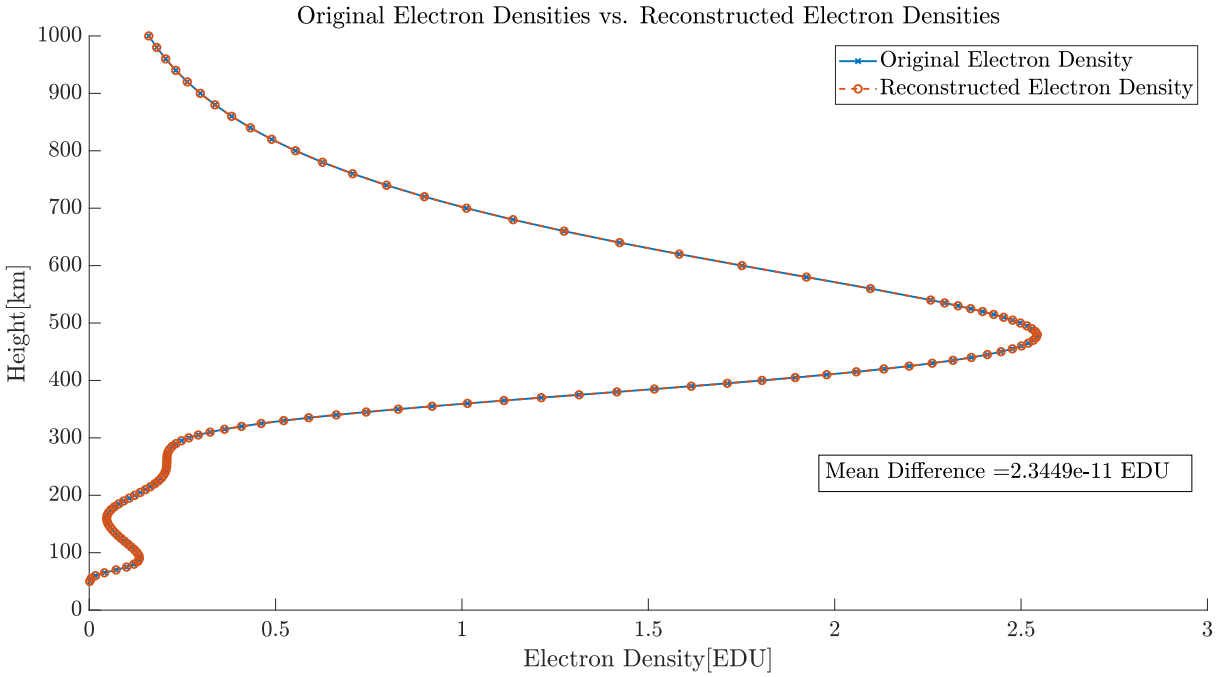


Figure 5.5: Original electron densities vs. reconstructed electron densities in scenario 2-1

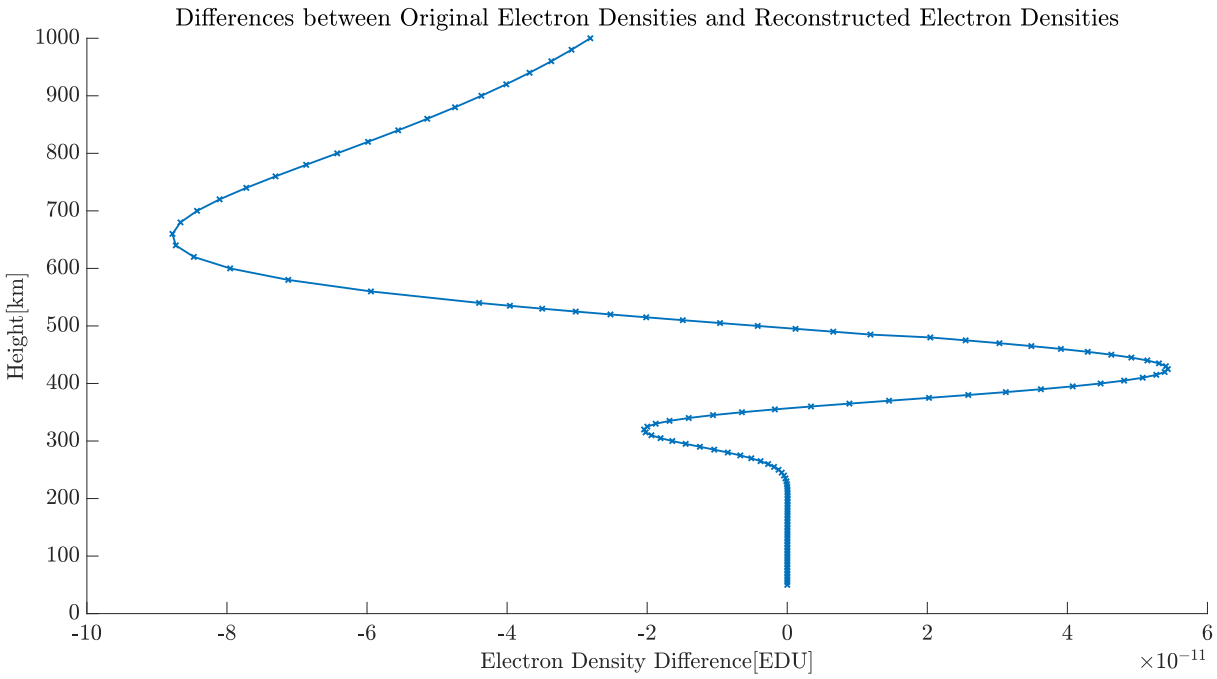


Figure 5.6: Differences between original electron densities and reconstructed electron densities in scenario 2-1

Table 5.4: Relevant values in scenario 2-2

| Key parameter | Original Value | Lower bound | Upper bound | Estimated value | Difference |
|---|---|---|---|---|---|
| $N_m^{F_2}$ | 2.5 EDU | 2.2 EDU | 2.8 EDU | 2.401 EDU | 0.09EDU |
| $h_m^{F_2}$ | 480 km | 250 km | 450 km | 450 km | 30 km |
| $N_m^{F_2}$ | 80 km | 75 km | 120 km | 75 km | 5 km |

As for the results shown in the last two columns, $h_m^{F_2}$ comes to 450 km which is the same as its upper bound. Besides, due to the correlations between the three parameters, the results of the other two are affected by $h_m^{F_2}$. However, there is no violation happening. In other words, the estimated results are still between or at the bounds. The comparison between the original electron density and the estimated electron density is plotted in Fig. 5.7. The red curve is the best fitting curve to the blue curve under the constraints in Table 5.4. And we can clearly distinguish that, due to the upper bound of $h_m^{F_2}$, the peak height for the $F_2$ layer of the reconstructed electron density is lower by 30 km than that of the original electron density.

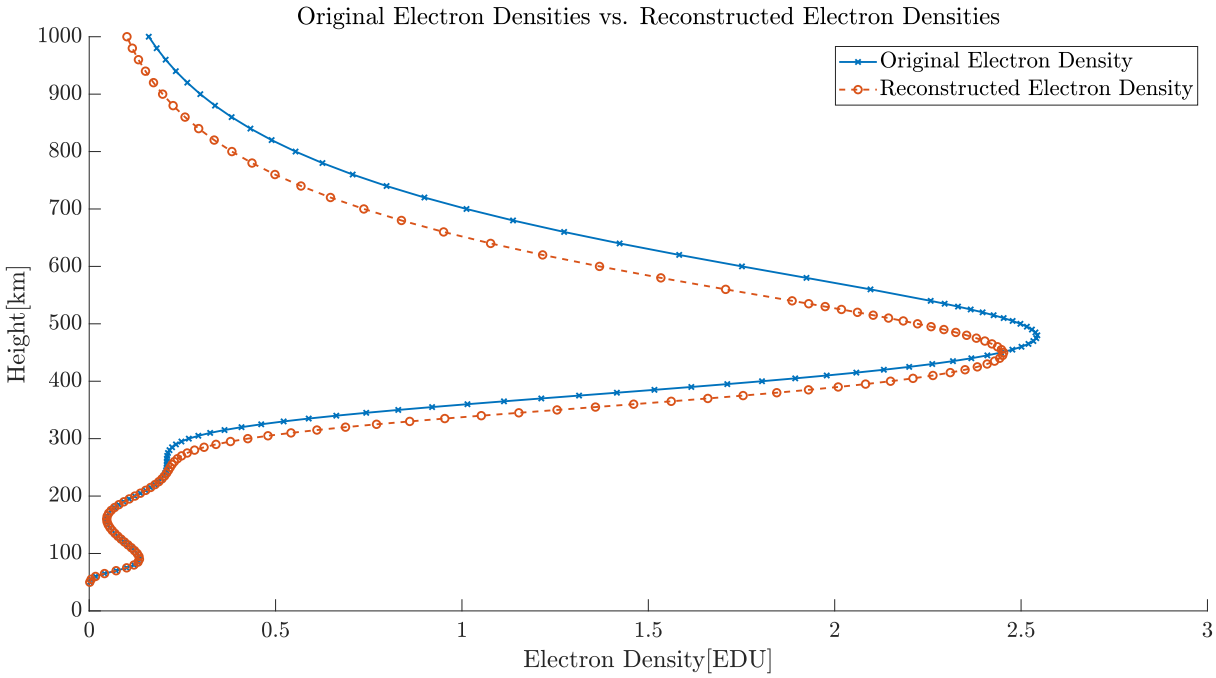


Figure 5.7: Differences between original electron densities and reconstructed electron densities in scenario 2-2.

## Scenario 2-3

After the estimation of a single vertical profile, we now come to the global estimation. At first, we set the constraints for the three parameters described in Table 5.5 to all grid points where the minimum and maximum values are taken from the original values. The constraints are set according to the boundary values of the key parameters. In this step, all the constraints are inactive in the estimation procedure. The results are given in the following.

Table 5.5: Constraints of the key parameters in scenario 2-3

| Key parameter | Minimum value | Lower bound | Maximum value | Upper bound |
|---|---|---|---|---|
| $N_m^{F_2}$ | 0.113 EDU | 0.1 EDU | 2.353 EDU | 2.5 EDU |
| $h_m^{F_2}$ | 242.7 km | 240 km | 410.6 km | 450 km |
| $N_m^{F_2}$ | 80.4 km | 80 km | 117.5 km | 120 km |

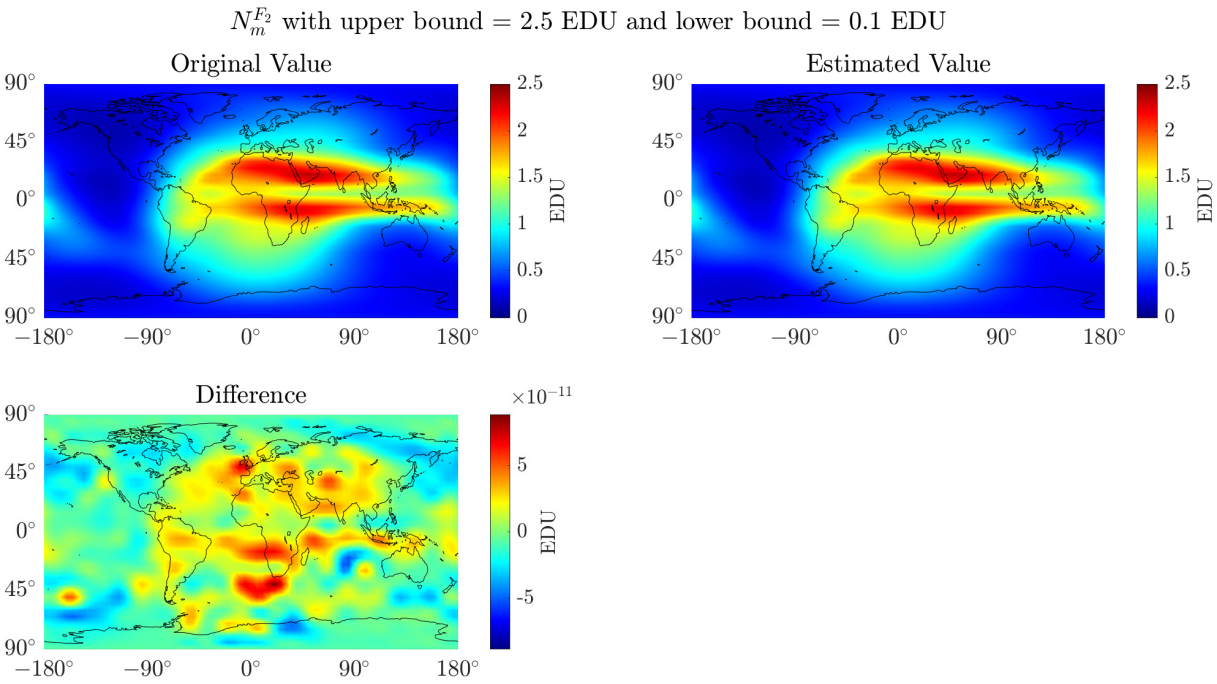


Figure 5.8: Maps of the original key parameter $N_m^{F_2}$, the estimated key parameter in scenario 2-3 and the differences between them.

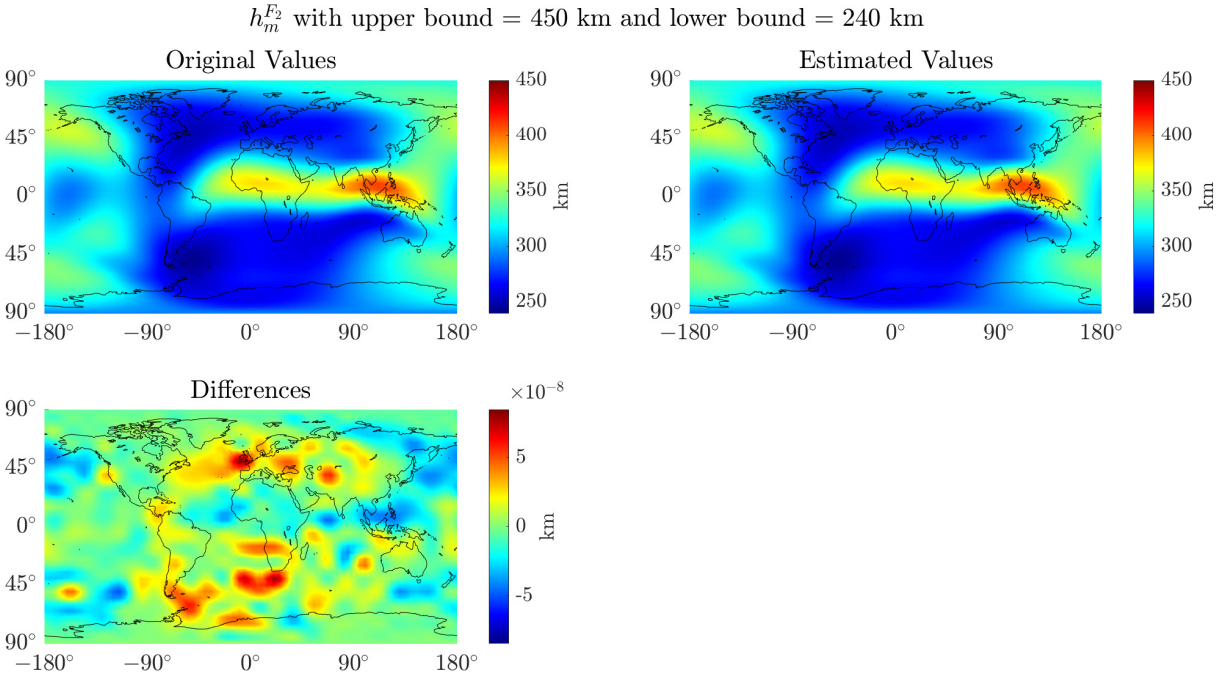


Figure 5.9: Maps of the original key parameter $h_m^{F_2}$, the estimated key parameter in scenario 2-3 and the differences between them.

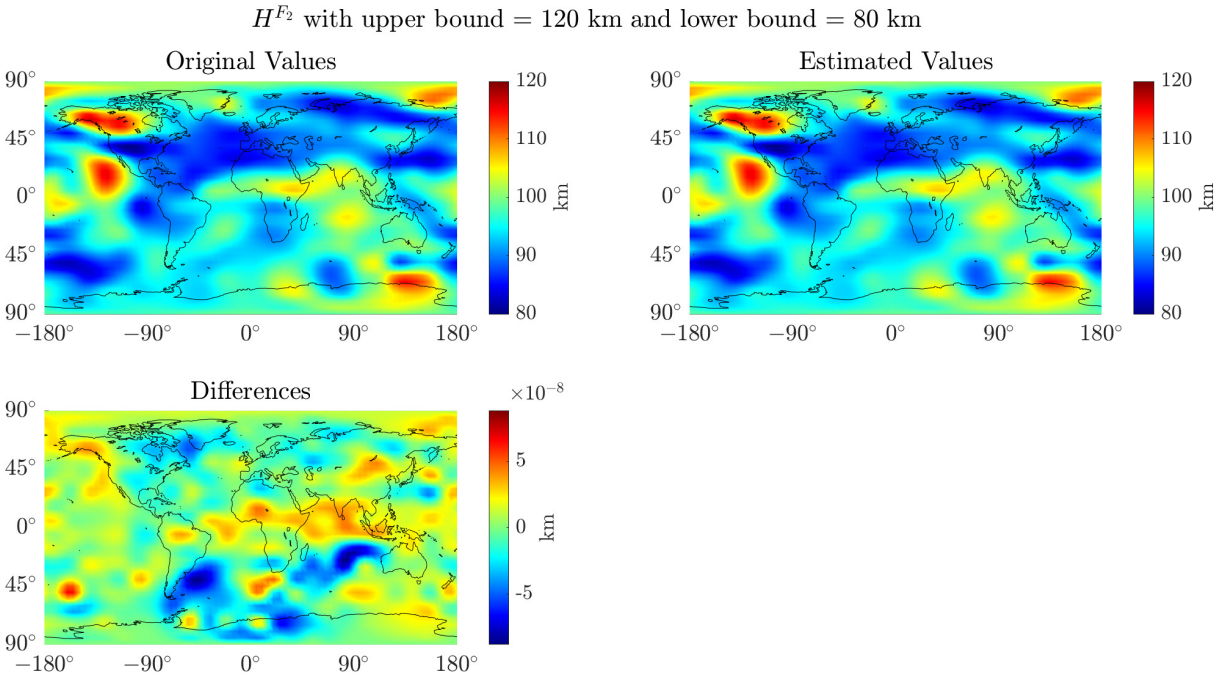


Figure 5.10: Maps of the original key parameter $H^{F_2}$, the estimated key parameter in scenario 2-3 and the differences between them.

According to the Figs. 5.8, 5.9 and 5.10 which show the estimated results for $N_m^{F_2}$, $h_m^{F_2}$ and $H^{F_2}$, respectively. We found that the estimated results are similar to the original values. The differences for $N_m^{F_2}$ is around $10^{-11}$ EDU whereas the differences for the other two parameters are at the level of $10^{-8}$ km. However, their relative differences are at the same level. Compared to the results in scenario 1-1, the results in the current scenario cannot reach the same accuracy as before due to the correlations among the three parameters. In the following, comparisons between the original electron density and the reconstructed electron density are given in Figs. 5.11 and 5.12.

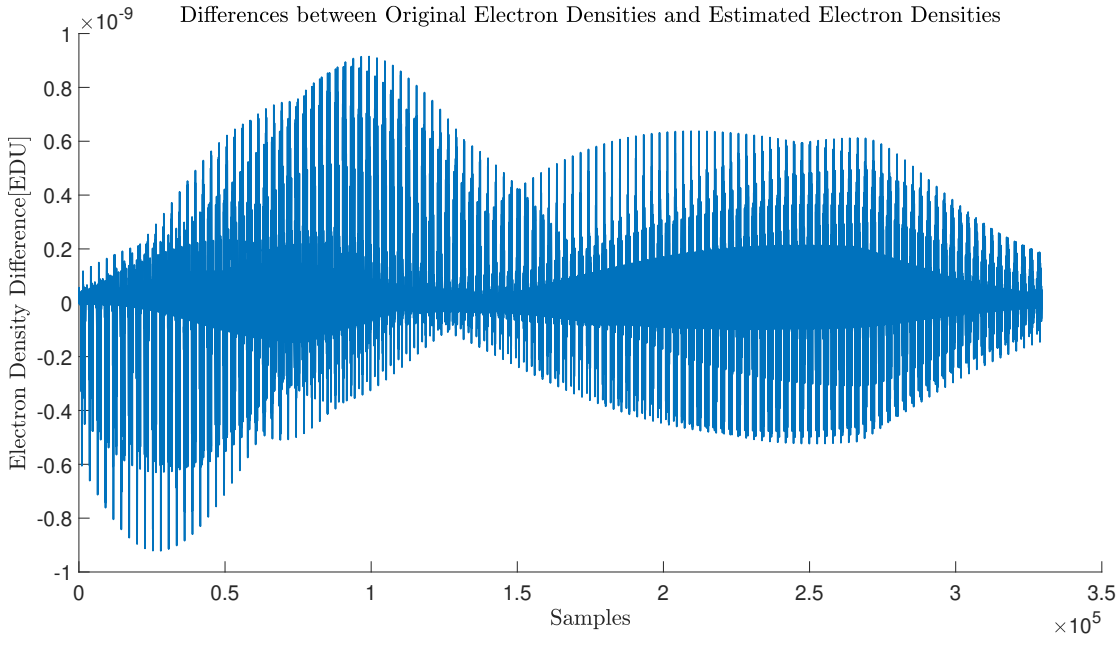


Figure 5.11: Differences between original electron densities and reconstructed electron densities in scenario 2-3

## Scenario 2-4

Another estimation will be performed in the following where some of the constraints are active during the procedure. In this step, the new constraints for $N_m^{F_2}$ and $h_m^{F_2}$ are defined as:

$$\begin{aligned} 0.12 \text{ EDU} \leq & N_m^{F_2} \leq 2.3 \text{ EDU} \\ 250 \text{ km} \leq & h_m^{F_2} \leq 450 \text{ km} \end{aligned} \tag{5.4}$$

for all grid points. The constraints for $H^{F_2}$ remain the same as given in Table 5.5. The comparisons between the original and estimated key parameters are presented in Figs. 5.13, 5.14 and 5.15. Due to the constraints for $N_m^{F_2}$, there are two strong deviations located in Midasia and Canada, which corresponds to the regions where original values are larger than 2.3 EDU or lower than 0.12 EDU. Besides, the blue area around southern Africa in Fig. 5.12 means, the $h_m^{F_2}$ value smaller than 250 km is changed to 250 km. In other words, the estimation procedure obeys all the constraints. Furthermore, all three panels that plot the differences show similar distributions. This means the active constraints will have impacts on all correlated parameters with inequality constraints.

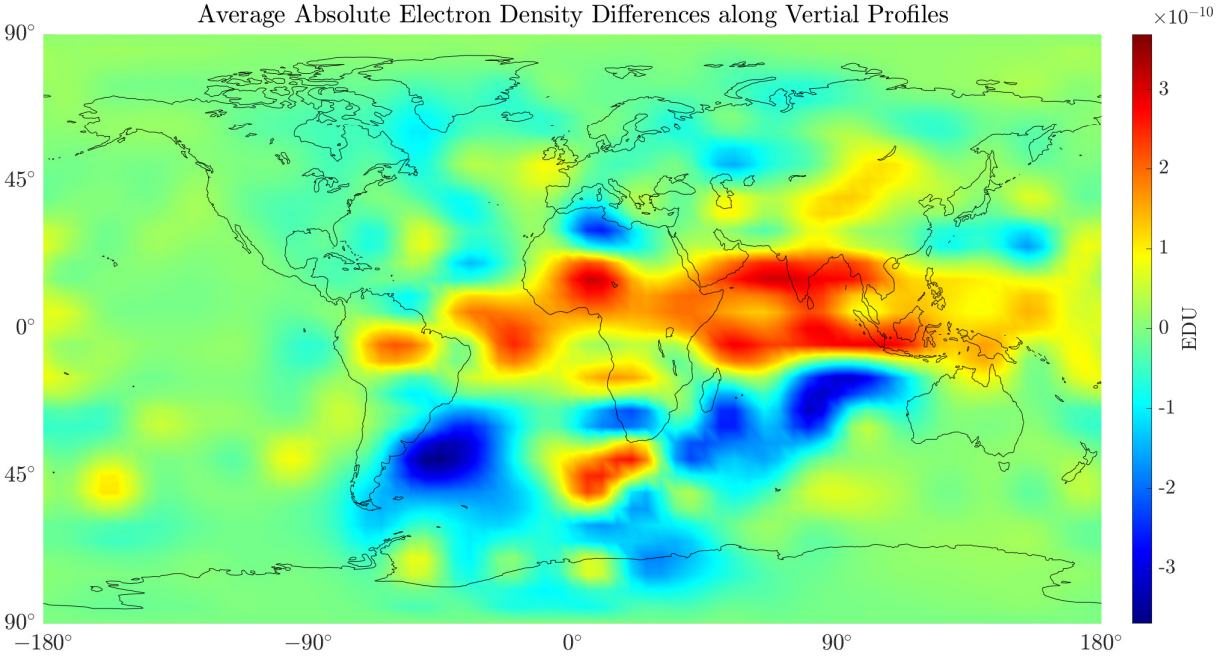


Figure 5.12: Map of average absolute electron density differences along vertical profiles in scenario 2-3

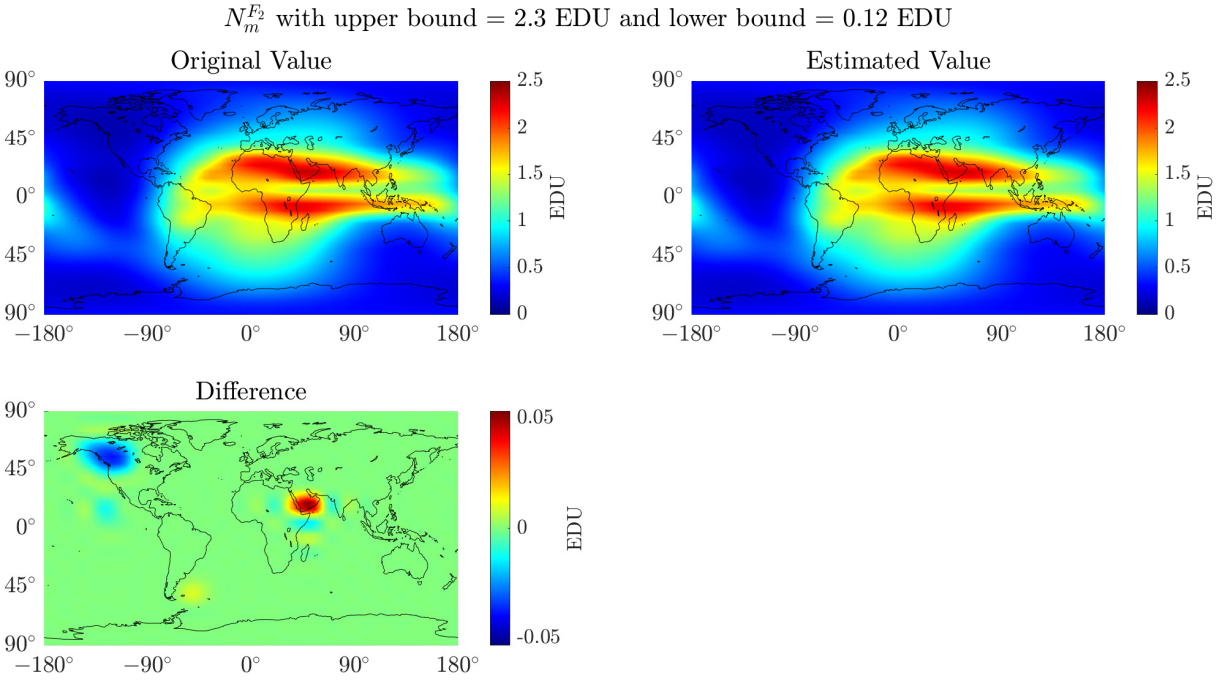


Figure 5.13: Maps of the original key parameter $N_m^{F_2}$, the estimated key parameter in scenario 2-4 and the differences between them.

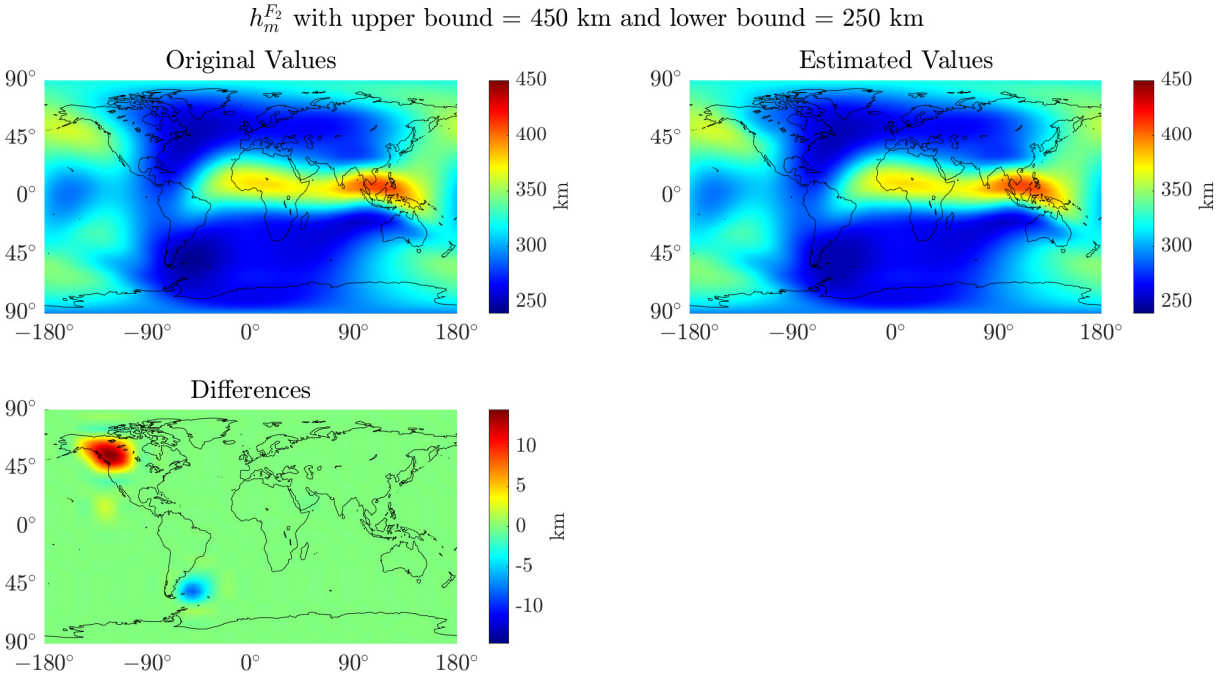


Figure 5.14: Maps of the original key parameter $h_m^{F_2}$, the estimated key parameter in scenario 2-4 and the differences between them.

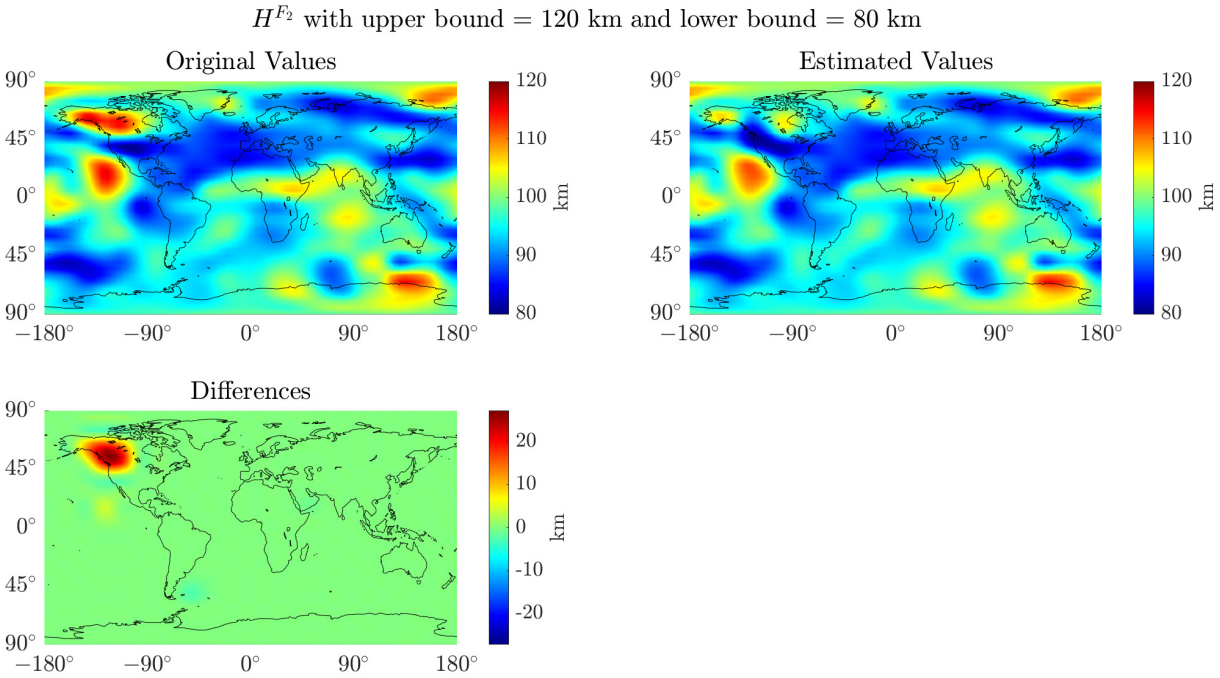


Figure 5.15: Maps of the original key parameter $H^{F_2}$, the estimated key parameter in scenario 2-4 and the differences between them.

### Scenario 2-5

Until now, we only perform ICO based on simulated data without noise. However, in real cases, any observation will contain noise. Therefore, in this step, 5% noise will be added to the electron density computed with simulated key parameters as observations. At the same time, we set different constraints for each grid point, see Figs. 5.16 and 5.17. Since now we are performing the closed loop simulation, the constraints are taken as 80% and 120% of the original values as lower and upper constraints, respectively, which can be written as,

$$0.8\kappa \leq \hat{\kappa} \leq 1.2\kappa \tag{5.5}$$

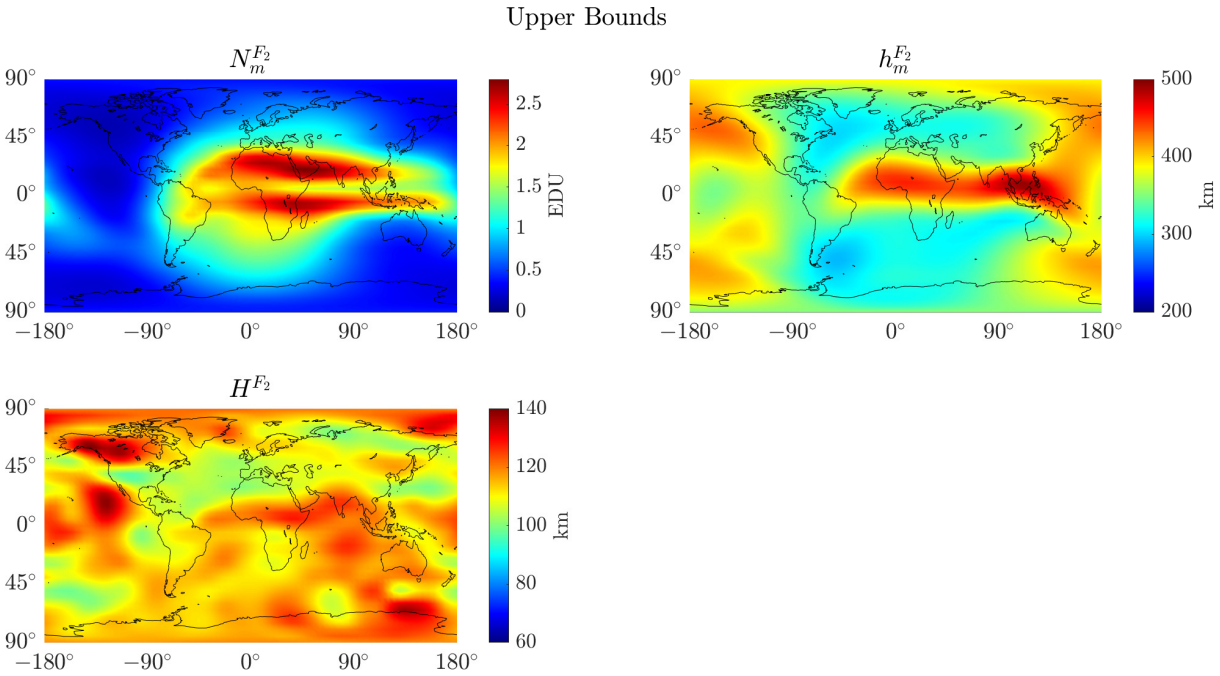


Figure 5.16: Maps of upper bounds of the key parameters in scenario 2-5

The estimated results are given in Fig. 5.18. It can be found that the estimated results still follow similar distributions to the original values shown in the figures before.

It is of great importance and necessity to compare the relations between the bounds and the results to see if there is any violation. The differences between upper bounds and results are presented in Fig. 5.19 while Fig. 5.20 is given for the differences with lower bounds.

It can be seen that all values in Fig. 5.19 are larger than 0 which means the estimated results are below the upper bounds whereas according to Fig. 5.20, all estimated values are above the lower bounds. This means there is no violation in the current estimation.

After the check of violation, it can be concluded that the algorithm of ICO can effectively make the estimated results located within the design bounds. However, it is also important to evaluate the accuracy of the algorithm. Two strategies are adopted. The first one is to compare the differences between the noise-free electron density (the truth) and the reconstructed electron density, see Figs. 5.21 and 5.22.

The electron density differences are at the level of $10^{-3}$ EDU, because 5% noise is added to the observations but B-spline expansion can only reconstruct the part without noise. However,

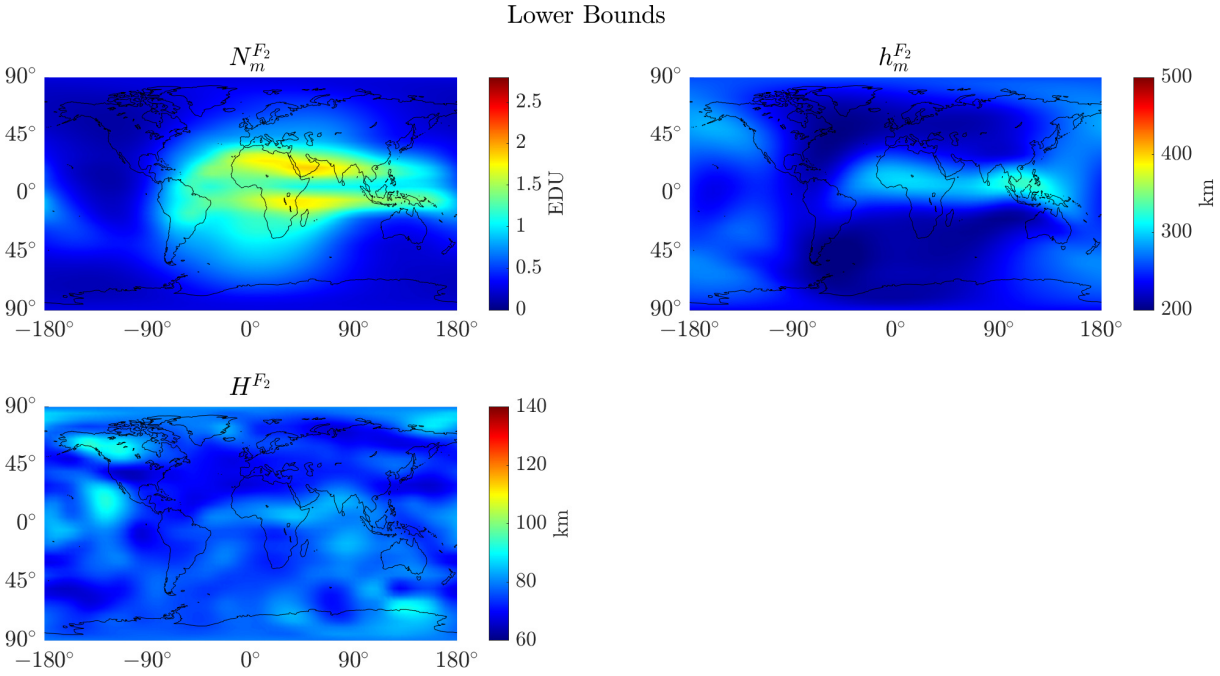


Figure 5.17: Maps of lower bounds of the key parameters in scenario 2-5

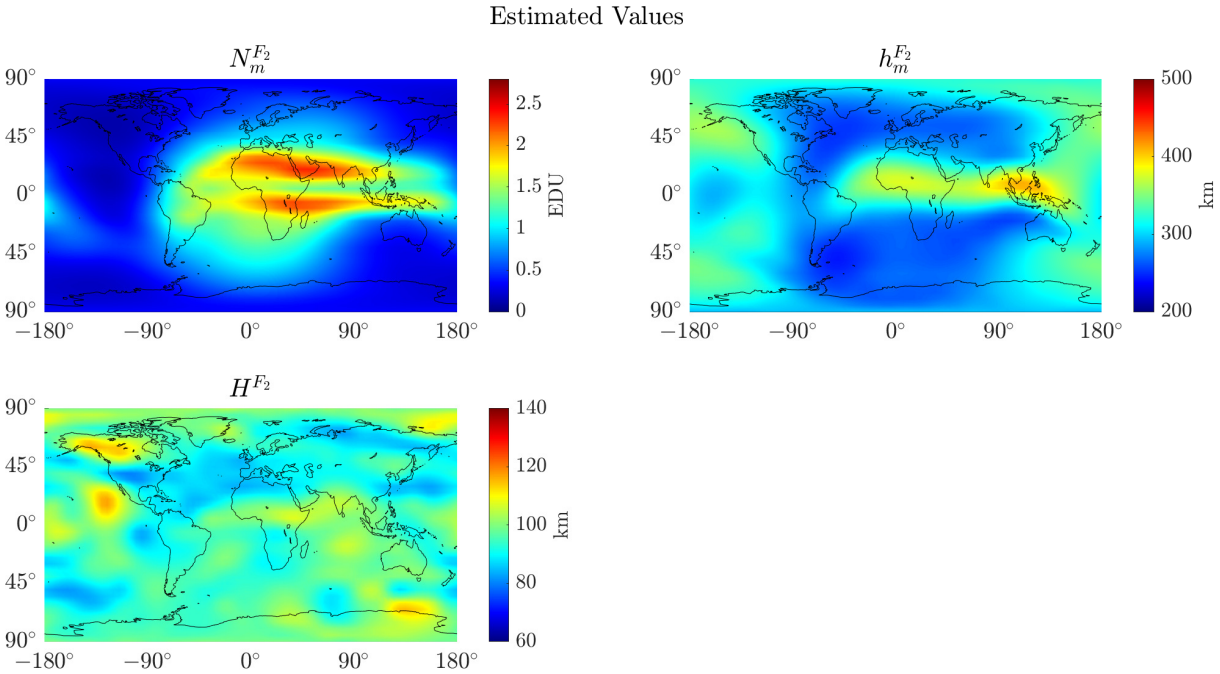


Figure 5.18: Maps of estimated results of the key parameters in scenario 2-5

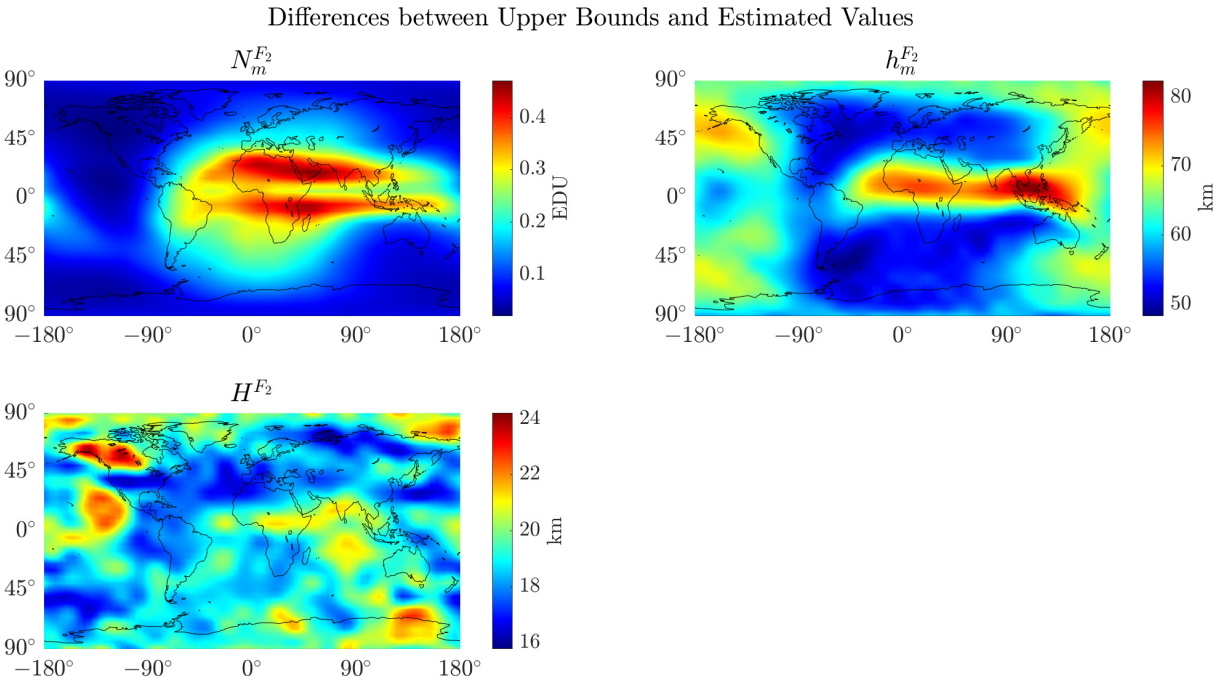


Figure 5.19: Differences between upper bounds and estimated values in scenario 2-5

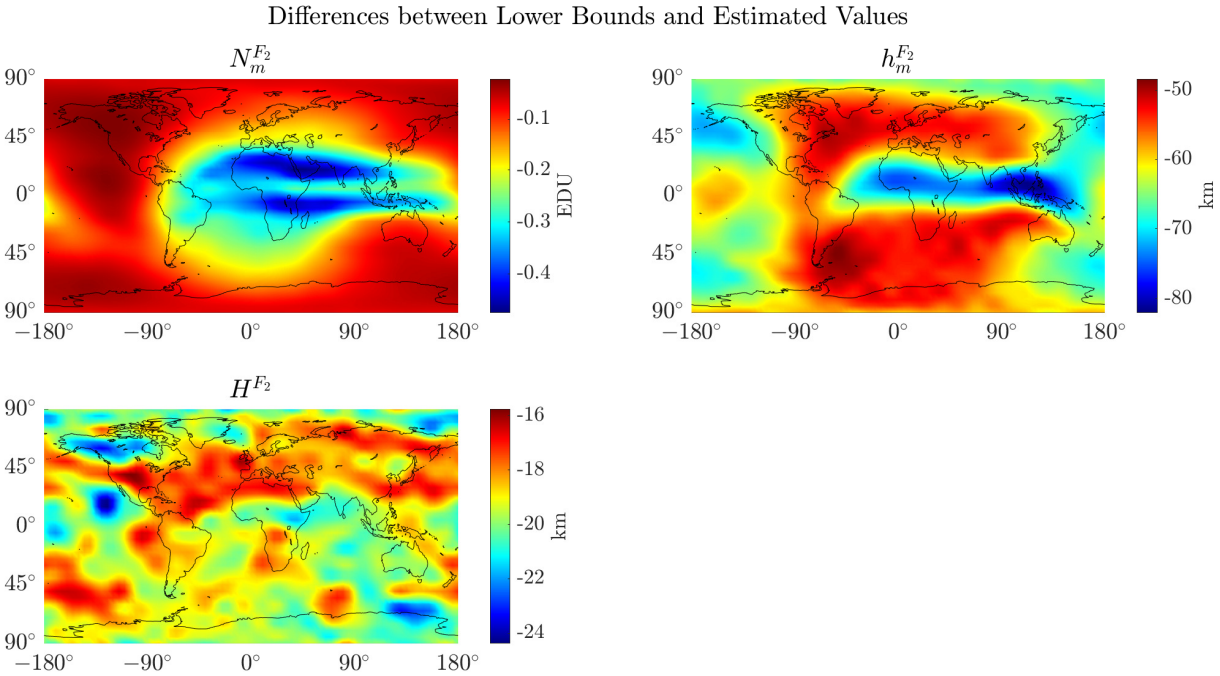


Figure 5.20: Differences between lower bounds and estimated values in scenario 2-5

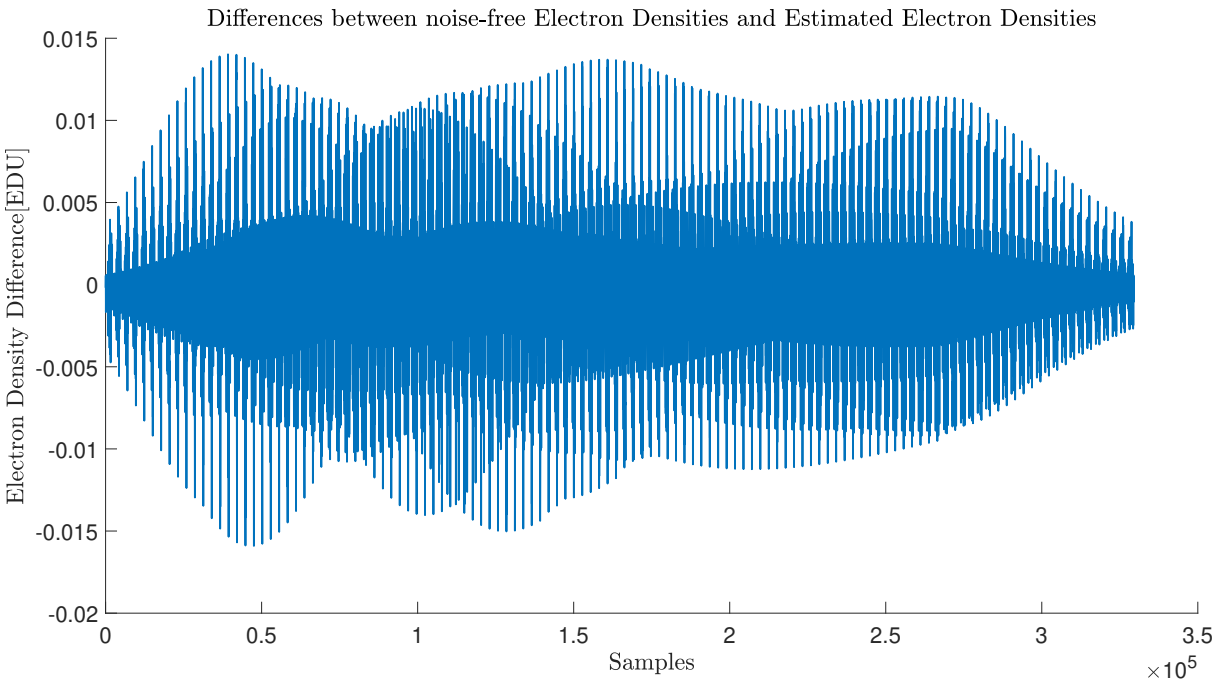


Figure 5.21: Differences between noise-free electron densities and reconstructed electron densities in scenario 2-5

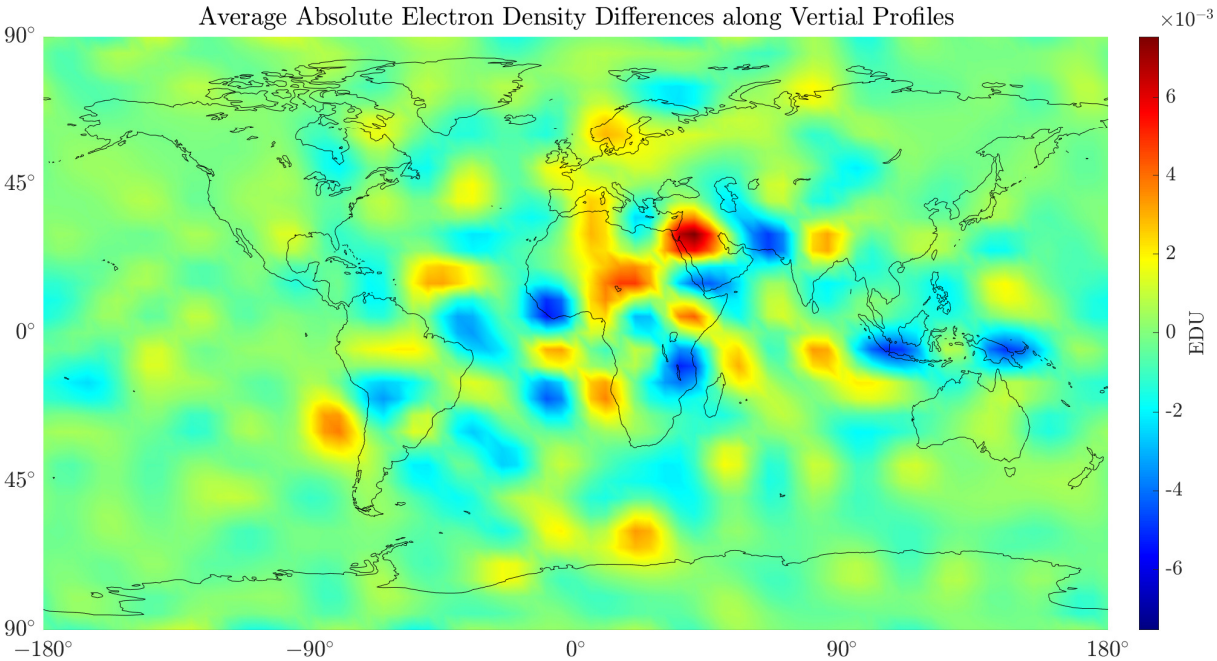


Figure 5.22: Map of average absolute electron density differences along vertical profiles in scenario 2-5

such differences are already smaller than the noise, which means the system is convergent and the ICO has the ability for resisting the noise.

Another strategy for evaluating our algorithm is to compute the standard deviation based on the Monte Carlo method. The standard deviation of the current scenario is given in Fig. 5.23.

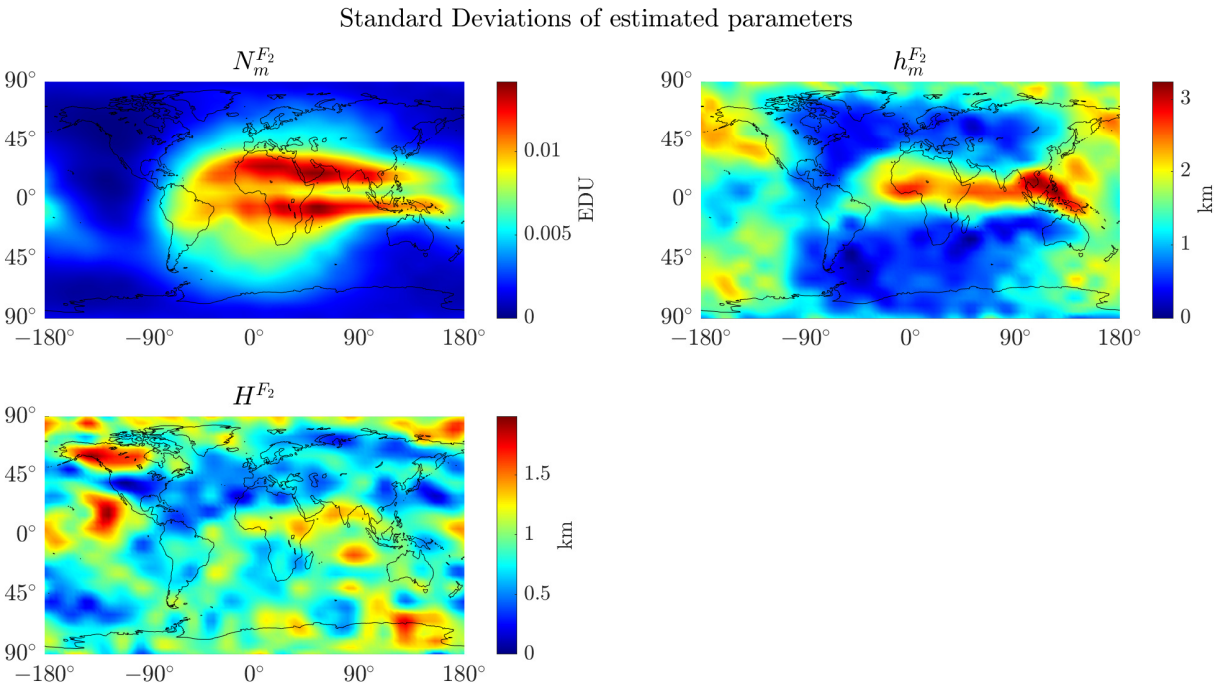


Figure 5.23: Maps of standard deviations of estimated parameters in scenario 2-5

It can be found that the standard deviations have similar structures as the original values but obviously with some vibrations, especially for the nonlinear parameter $h_m^{F_2}$ and $H^{F_2}$. However, the standard deviations are smaller than the noise added to the observations.

## 5.3 Nine parameters with inequality constraints

In this scenario, the estimations are performed under the condition that nine parameters are constrained with inequality constraints, which can be presented as:

$$\begin{aligned} \boldsymbol{\kappa}_1 &= \left[N_0^P, H^P, N_m^{F_2}, h_m^{F_2}, H^{F_2}, N_m^{F_1}, h_m^{F_1}, N_m^E, N_m^D\right] \\ \boldsymbol{\kappa}_2 &= \left[H^{F_1}, h_m^E, H^E, h_m^D, H^D\right] \end{aligned} \tag{5.6}$$

### Scenario 3-1

Firstly, one vertical profile estimation is performed, which is similar to the experiment in scenario 2-1. The values of relevant parameters are given in Table 5.6. According to the last column, all estimated parameters are close to their original values and come to similar relative differences. For other key parameters, they are adopted with the values shown in Table 5.2.

Table 5.6: Relevant values in scenario 3-1

| Key parameter | Original value | Lower bound | Upper bound | Estimated value | Difference |
|---|---|---|---|---|---|
| $N_0^P$ | 0.025 EDU | 0.02 EDU | 0.03 EDU | 0.025 EDU | $5.1 \times 10^{-10}$ EDU |
| $H^P$ | 80 km | 75 km | 120 km | 80 km | $8.1 \times 10^{-7}$ km |
| $N_m^{F_2}$ | 2.5 EDU | 2.2 EDU | 2.6 EDU | 2.5 EDU | $-4.3 \times 10^{-10}$ EDU |
| $h_m^{F_2}$ | 480 km | 220 km | 500 km | 480 km | $8.4 \times 10^{-9}$ km |
| $H^{F_2}$ | 80 km | 75 km | 120 km | 80 km | $6.6 \times 10^{-9}$ km |
| $N_m^{F_1}$ | 0.2 EDU | 0.1 EDU | 0.5 EDU | 0.2 EDU | $-1.1 \times 10^{-10}$ EDU |
| $h_m^{F_1}$ | 250 km | 200 km | 300 km | 250 km | $-9.1 \times 10^{-9}$ km |
| $N_m^E$ | 0.1 EDU | 0.01 EDU | 0.5 EDU | 0.1 EDU | $-3.8 \times 10^{-11}$ EDU |
| $N_m^D$ | 0.05 EDU | 0.01 EDU | 0.2 EDU | 0.05 EDU | $2.1 \times 10^{-11}$ EDU |

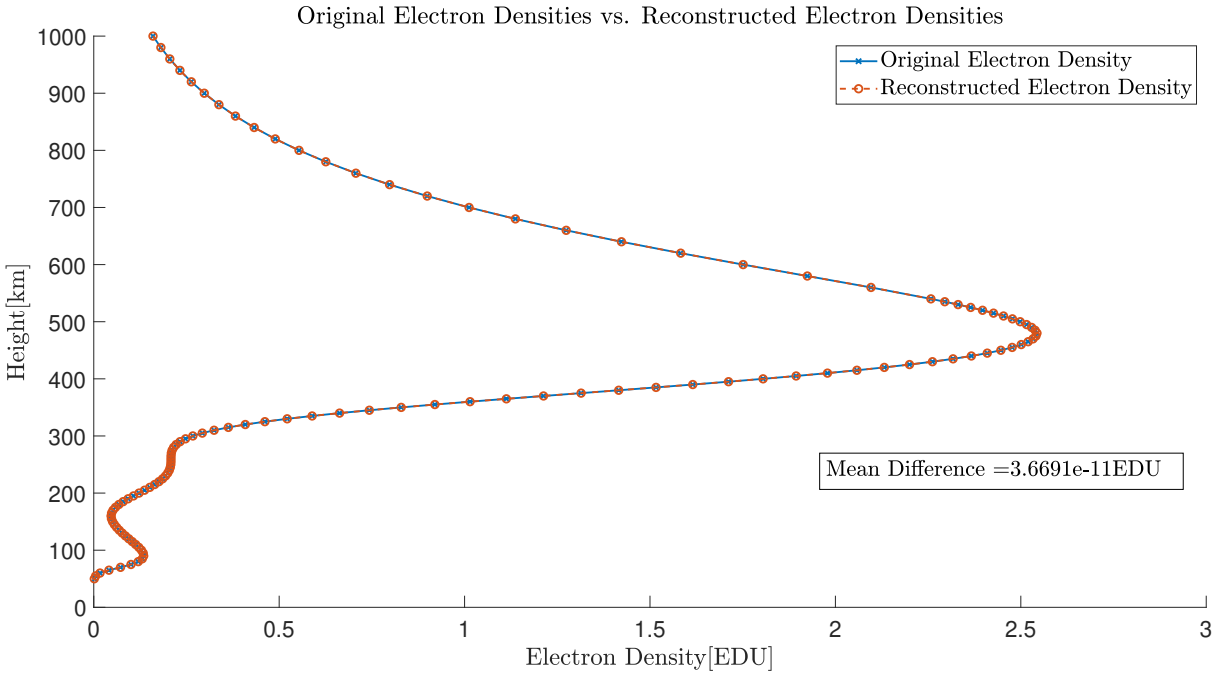


Figure 5.24: Original electron densities vs. reconstructed electron densities in scenario 3-1

Fig. 5.24 describes the reconstructed electron density with respect to the original electron density. It can be found that the two curves still fit well with each other and the average absolute difference is $3.67 \cdot 10^{-11}$ EDU. The differences between the original and reconstructed electron density are plotted in Fig. 5.25. Since in this scenario, key parameters in the $D$, $E$ and $F_1$ layer are taken into consideration, there are also deviations below 250 km. Furthermore, the largest deviations are still located in the $F_2$ layer.

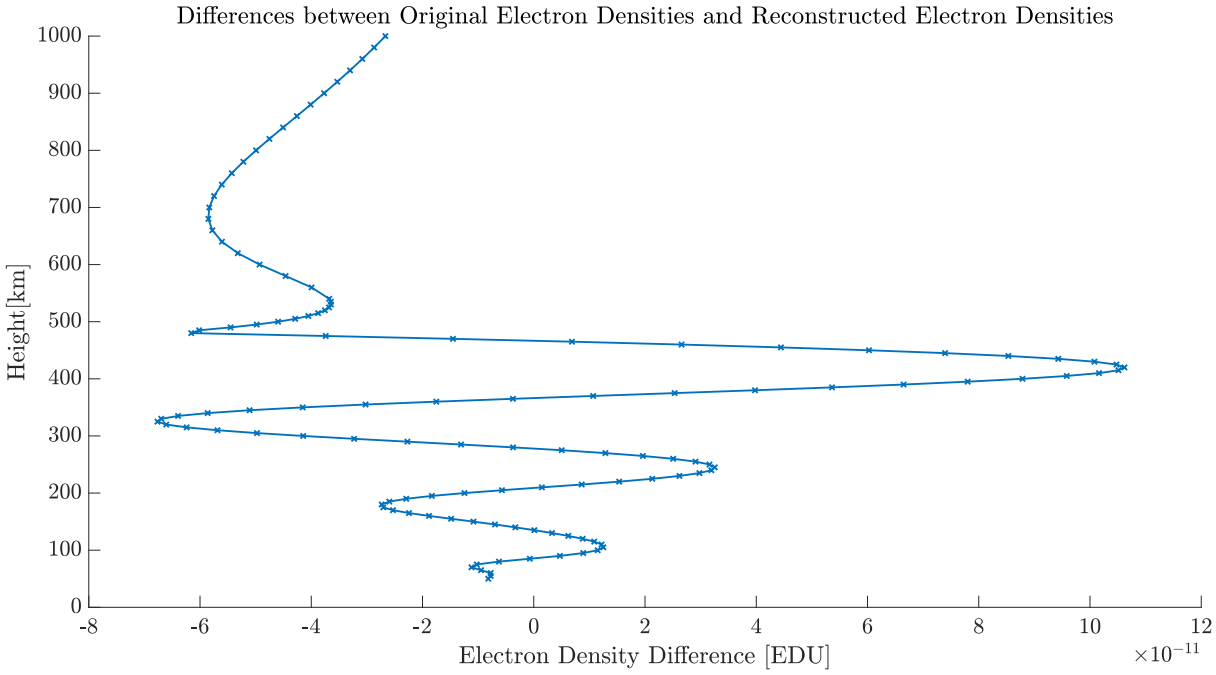


Figure 5.25: Differences between original electron densities and reconstructed electron densities in scenario 3-1

### Scenario 3-2

In the following, we change several constraints in order to make them active during the estimation. We believe the peak height of the $F_2$ layer should not be higher than 450 km while the scale height of the $F_2$ layer cannot be smaller than 85 km, and the maximum electron density of the $E$ layer must be larger than 0.15 EDU. The updated constraints, as well as the corresponding estimated results, are shown in Table 5.7.

According to the table above, there are no violations in this estimation and all results are located between the given bounds. However, since the original values of $h_m^{F_2}$, $H^{F_2}$ and $N_m^E$ are out of their bounds, all the other parameters are affected by them due to the correlations. The original and reconstructed electron densities are illustrated in Fig. 5.26.

### Scenarios 3-3 and 3-4

In the next step, 5% noise is added to the observations in one vertical profile. The estimations are performed again using the constraints in Tables 5.6 and 5.7, which results are presented in Tables 5.8 and 5.9, Figs. 5.27 and 5.28.

Table 5.7: Relevant values in scenario 3-2

| Key parameter | Original value | Lower bound | Upper bound | Estimated value | Difference |
|---|---|---|---|---|---|
| $N_0^P$ | 0.025 EDU | 0.02 EDU | 0.03 EDU | 0.02 EDU | −0.005 EDU |
| $H^P$ | 80 km | 75 km | 120 km | 75 km | -5 km |
| $N_m^{F_2}$ | 2.5 EDU | 2.2 EDU | 2.6 EDU | 2.31 EDU | -0.19 EDU |
| $h_m^{F_2}$ | 480 km | 220 km | 450 km | 450 km | -30 km |
| $H^{F_2}$ | 80 km | 85 km | 120 km | 85 km | 5 km |
| $N_m^{F_1}$ | 0.2 EDU | 0.1 EDU | 0.5 EDU | 0.1 EDU | -0.1 EDU |
| $h_m^{F_1}$ | 250 km | 200 km | 300 km | 200 km | -50 km |
| $N_m^E$ | 0.1 EDU | 0.15 EDU | 0.5 EDU | 0.15 EDU | 0.05 EDU |
| $N_m^D$ | 0.05 EDU | 0.01 EDU | 0.2 EDU | 0.01 EDU | -0.04 EDU |

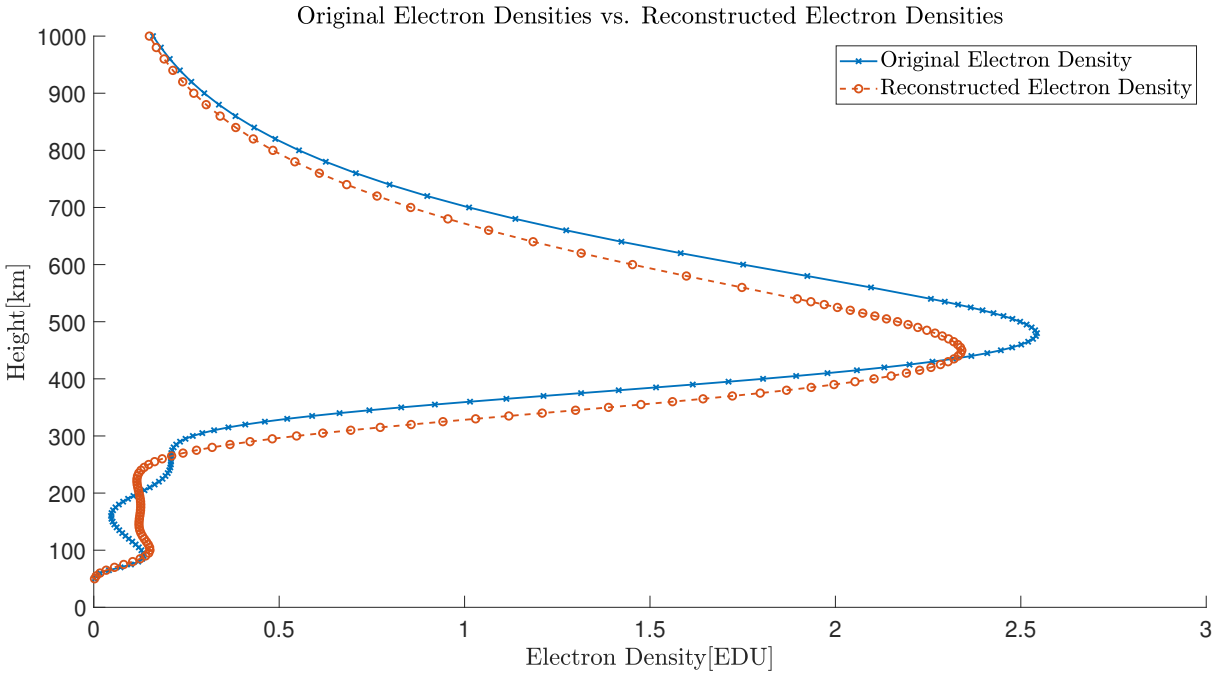


Figure 5.26: Original electron densities vs. reconstructed electron densities in scenario 3-2

Table 5.8: Relevant values in scenario 3-3

| Key parameter | Original value | Lower bound | Upper bound | Estimated value | Difference |
|---|---|---|---|---|---|
| $N_0^P$ | 0.025 EDU | 0.02 EDU | 0.03 EDU | 0.027 EDU | 0.002 EDU |
| $H^P$ | 80 km | 75 km | 120 km | 75 km | -5 km |
| $N_m^{F_2}$ | 2.5 EDU | 2.2 EDU | 2.6 EDU | 2.49 EDU | -0.01 EDU |
| $h_m^{F_2}$ | 480 km | 220 km | 500 km | 480.57 km | 0.57 km |
| $H^{F_2}$ | 80 km | 75 km | 120 km | 80.35 km | 0.35 km |
| $N_m^{F_1}$ | 0.2 EDU | 0.1 EDU | 0.5 EDU | 0.21 EDU | 0.01 EDU |
| $h_m^{F_1}$ | 250 km | 200 km | 300 km | 248.99 km | -1.01 km |
| $N_m^E$ | 0.1 EDU | 0 EDU | 0.5 EDU | 0.096 EDU | -0.004 EDU |
| $N_m^D$ | 0.05 EDU | 0.01 EDU | 0.2 EDU | 0.057 EDU | 0.007 EDU |

Table 5.9: Relevant values in scenario 3-4

| Key parameter | Original value | Lower bound | Upper bound | Estimated value | Difference |
|---|---|---|---|---|---|
| $N_0^P$ | 0.025 EDU | 0.02 EDU | 0.03 EDU | 0.02 EDU | $-0.005$ EDU |
| $H^P$ | 80 km | 75 km | 120 km | 75 km | -5 km |
| $N_m^{F_2}$ | 2.5 EDU | 2.2 EDU | 2.6 EDU | 2.32 EDU | -0.18 EDU |
| $h_m^{F_2}$ | 480 km | 220 km | 450 km | 450 km | -30 km |
| $H^{F_2}$ | 80 km | 85 km | 120 km | 85 km | 5 km |
| $N_m^{F_1}$ | 0.2 EDU | 0.1 EDU | 0.5 EDU | 0.1 EDU | -0.1 EDU |
| $h_m^{F_1}$ | 250 km | 200 km | 300 km | 200 km | -50 km |
| $N_m^E$ | 0.1 EDU | 0.15 EDU | 0.5 EDU | 0.15 EDU | 0.05 EDU |
| $N_m^D$ | 0.05 EDU | 0.01 EDU | 0.2 EDU | 0.01 EDU | -0.04 EDU |

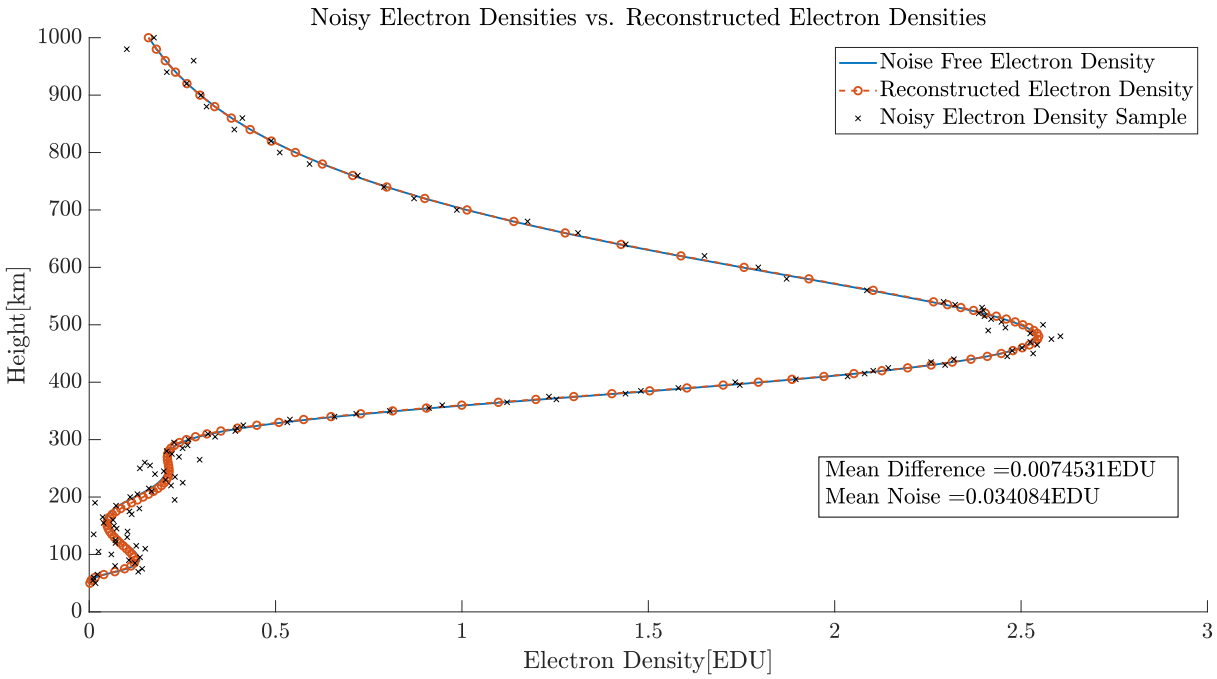


Figure 5.27: Original electron densities vs. reconstructed electron densities in scenario 3-3

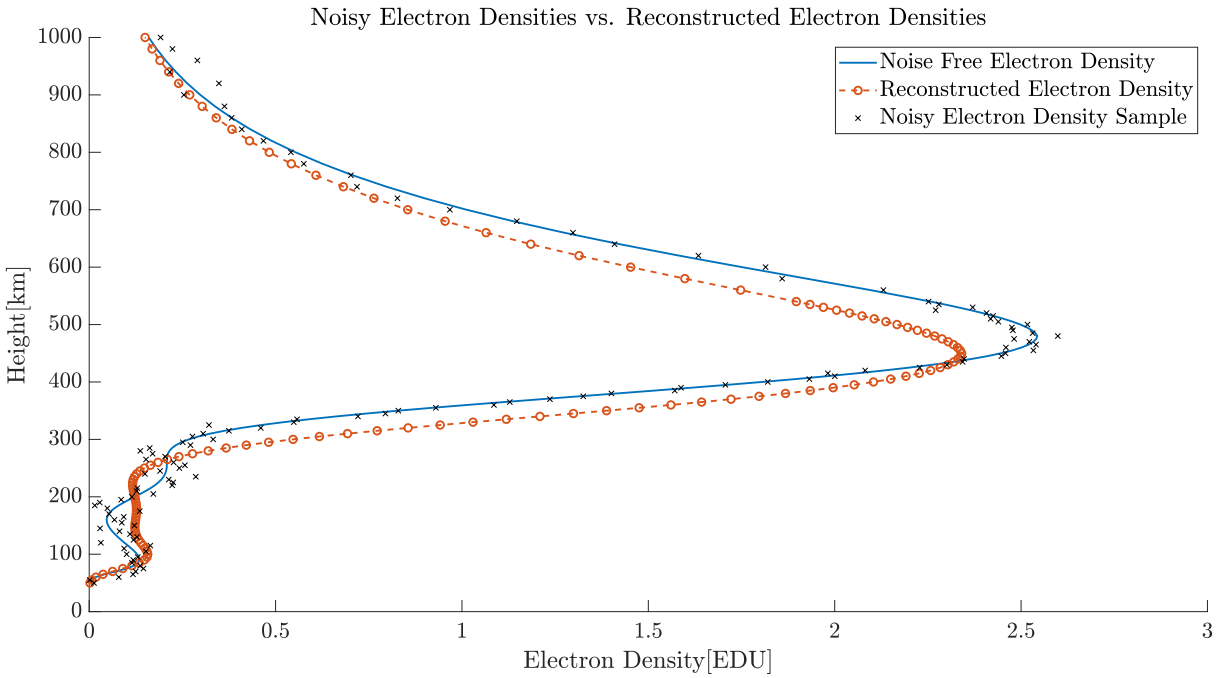


Figure 5.28: Original electron densities vs. reconstructed electron densities in scenario 3-4

In the two figures, the blue curve corresponds to the truth while the red curve describes the reconstructed electron density and the black crosses are the electron density observations with noise. Fig. 5.27 shows that, when nine parameters are estimated with inequality constraints under 5% noise, the average absolute difference comes to 0.0078 EDU which is smaller than the average of noise (0.037 EDU). As shown in Fig. 5.28, scenario 3-4 has similar results with the results in scenario 3-2.

### Scenario 3-5

In this section, the nine parameters are estimated with inequality constraints globally. Since we now have considered many parameters from multiple layers, it is no longer possible to set the same constraints in all grid points for each parameter. Therefore, the strategy given in Eq. (5.5) is adopted for this scenario and the lower bounds and upper bounds are plotted in Figs. 5.29 and 5.30.

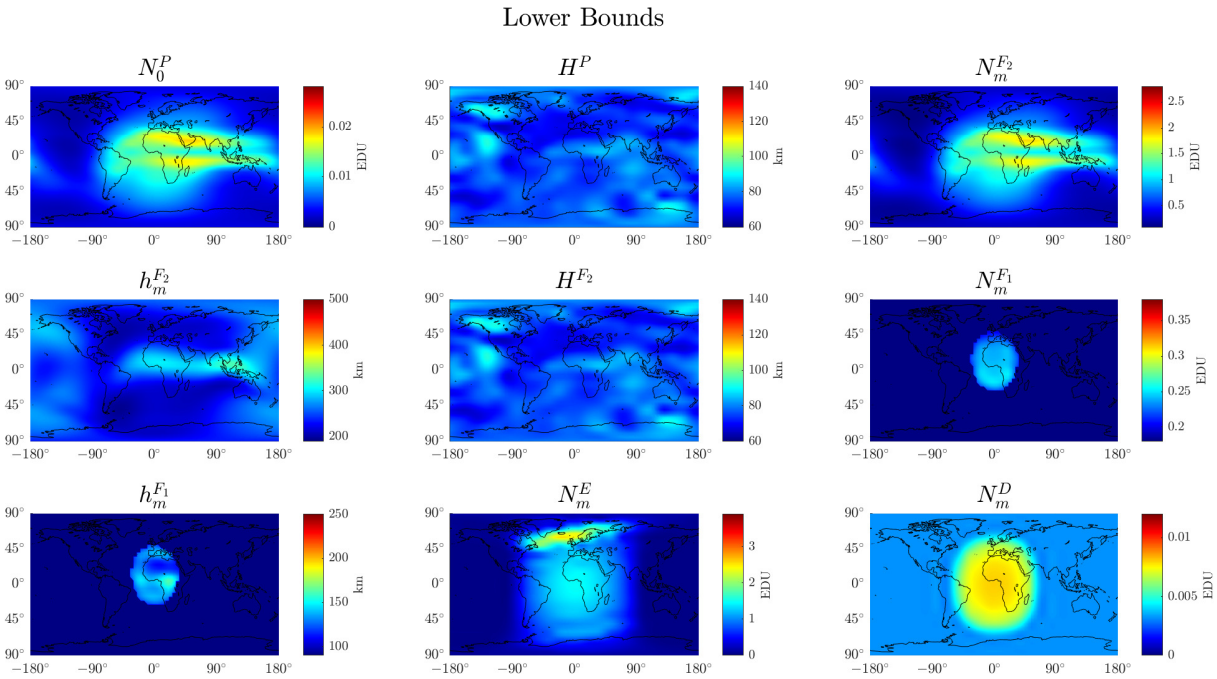


Figure 5.29: Maps of lower bounds of key parameters in scenario 3-5

It is worth mentioning that the $F_1$ layer will merge into the $F_2$ layer after sunset. Therefore, in mid right and bottom left panels in Figs. 5.29 and 5.30, the surrounding region has both lower bounds and upper bounds equal to zero because no $F_1$ layer exists there.

The estimated results are presented in Fig. 5.31 which still follow similar distributions as the original values. And Figs. 5.32 and 5.33 show their relations with the upper and lower bounds, respectively. It can be seen that all the results are located in the feasible regions.

Now, we reconstruct the electron density based on the estimated key parameters and perform a comparison between the original and reconstructed electron densities, see Fig. 5.34. The differences reach a level of $10^{-6}$ EDU. Besides, the average absolute difference along one vertical profile is presented in Fig. 5.35. It can be seen there are larger variations in North America, which corresponds to the large values of $H^P$ and $H^{F_2}$.

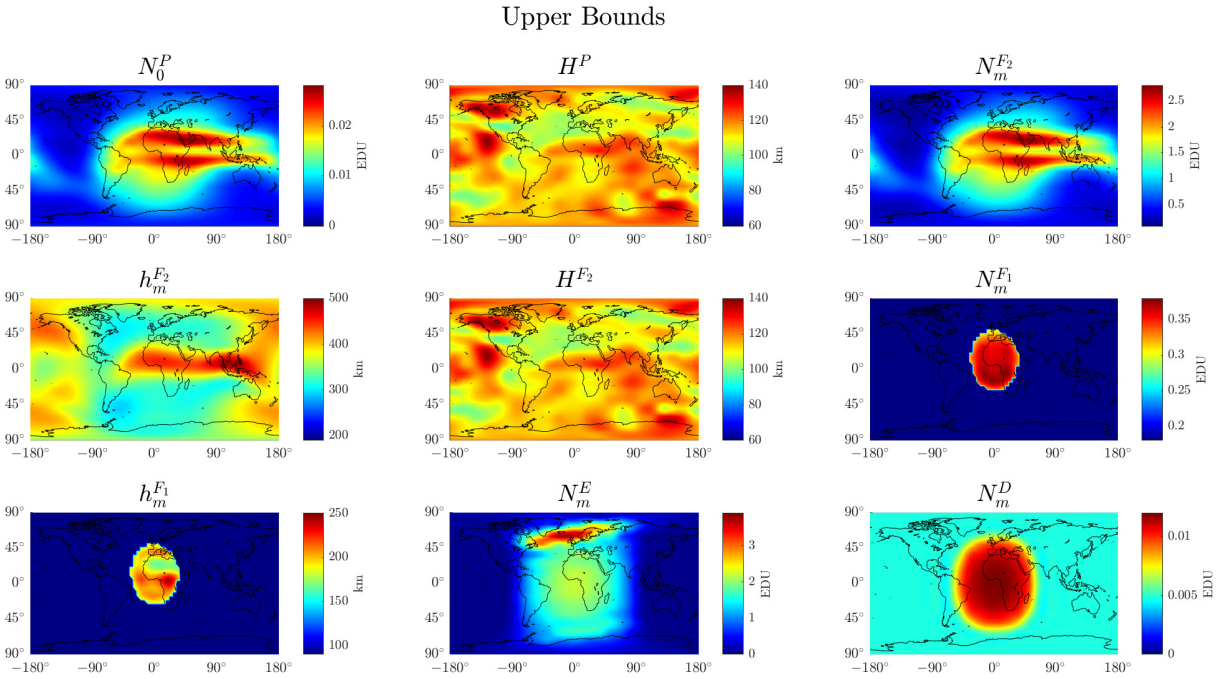


Figure 5.30: Maps of upper bounds of key parameters in scenario 3-5

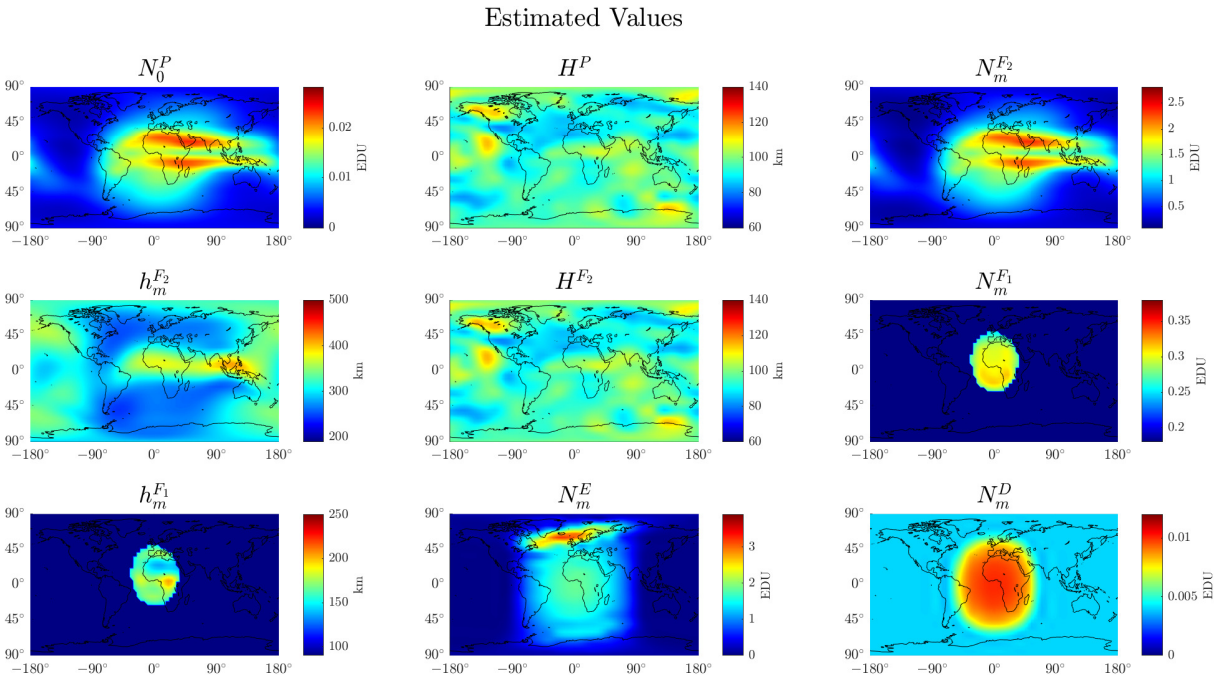


Figure 5.31: Maps of estimated results of the key parameters in scenario 3-5

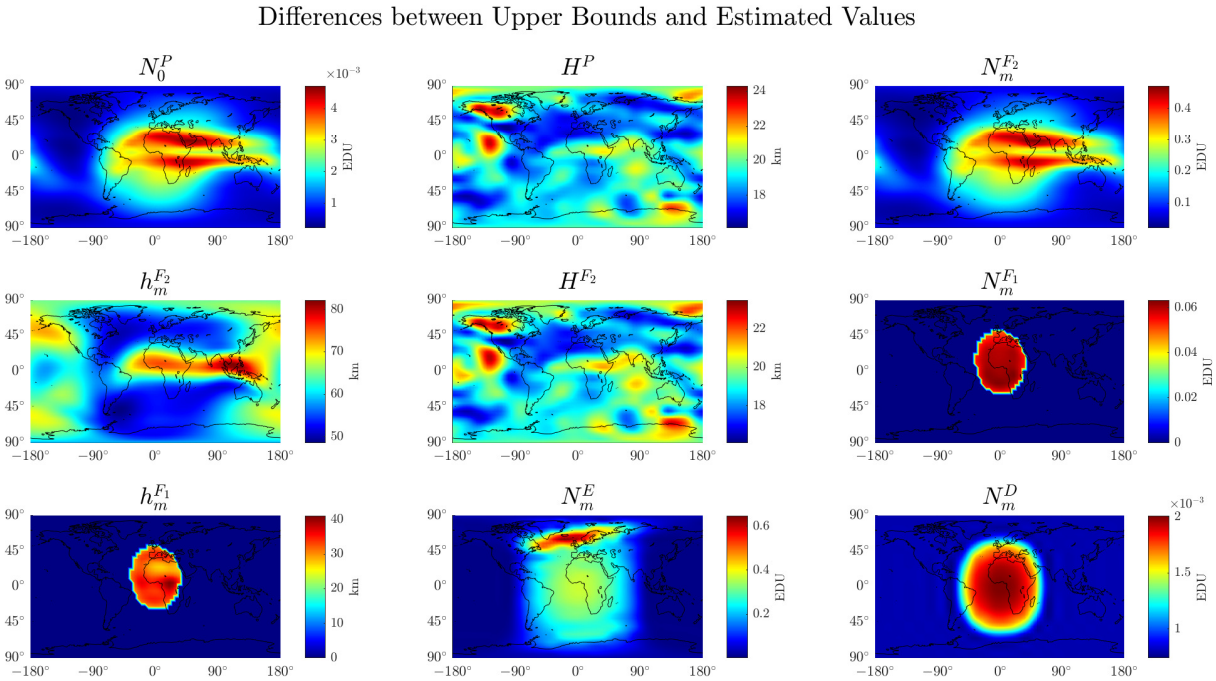


Figure 5.32: Differences between upper bounds and estimated values in scenario 3-5

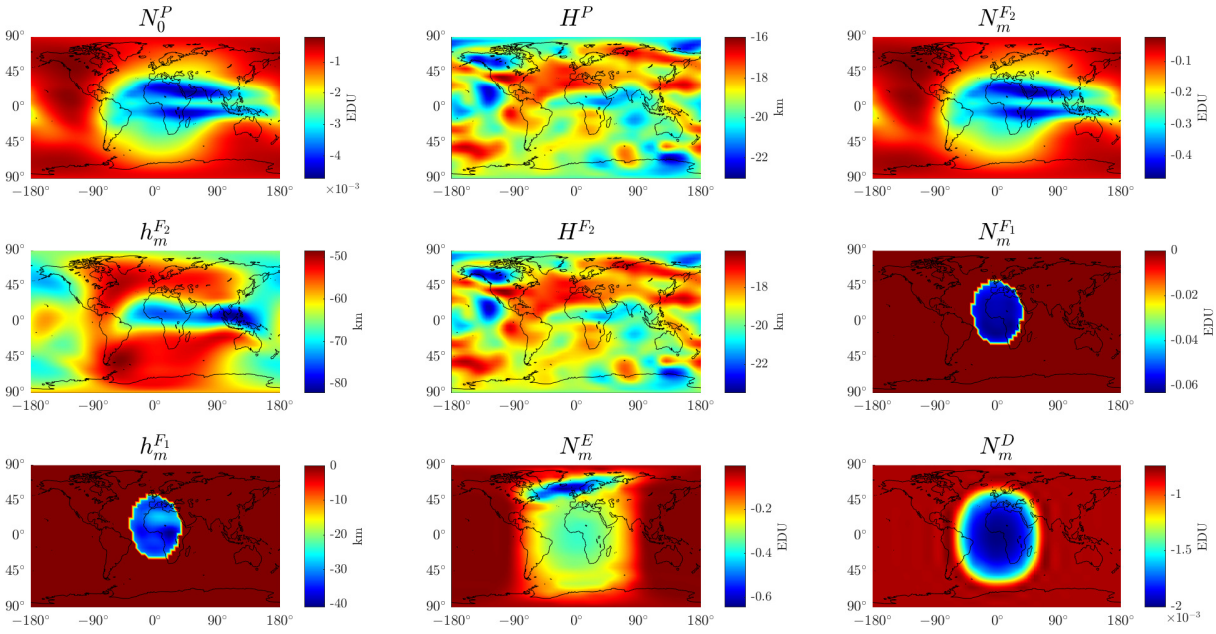


Figure 5.33: Differences between lower bounds and estimated values in scenario 3-5

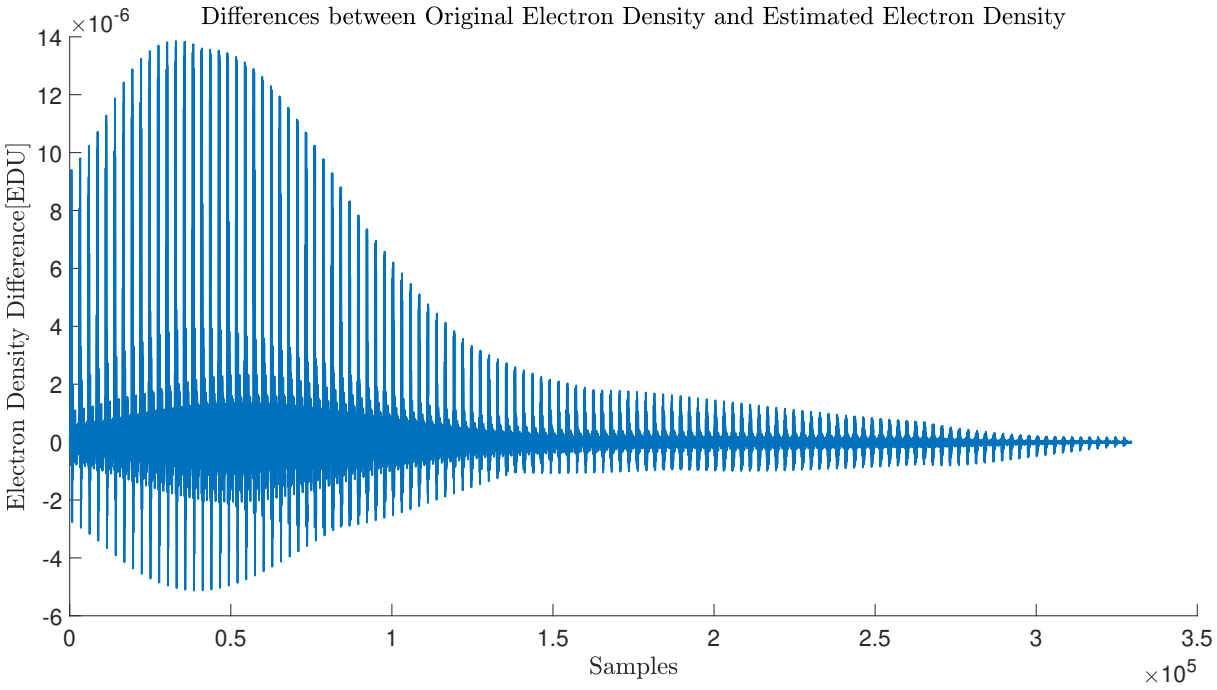


Figure 5.34: Differences between original electron densities and reconstructed electron densities in scenario 3-5

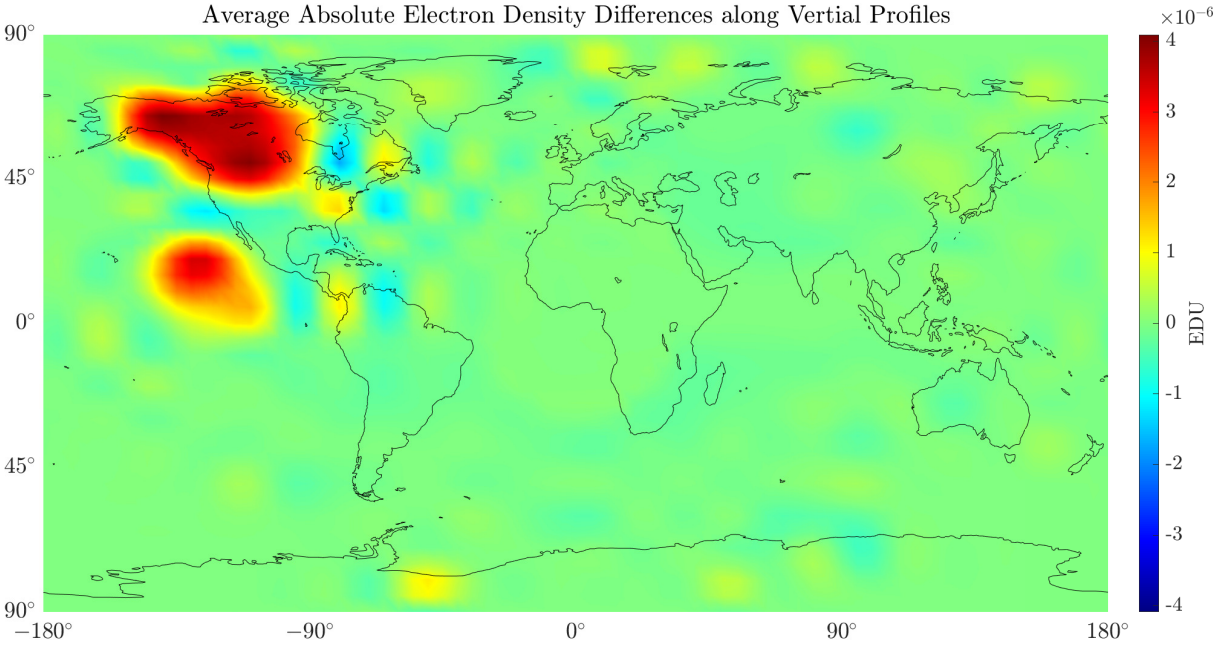


Figure 5.35: Map of average absolute electron density difference along vertical profiles in scenario 3-5

### Scenario 3-6

Similar to the experiments in scenario 2-5, we now add 5% noise to the electron densities computed from the original key parameters. The constraints are the same as given in scenario 3-5, see Figs. 5.29 and 5.30. The estimated results, as well as their relations with bounds, are shown in the following figures.

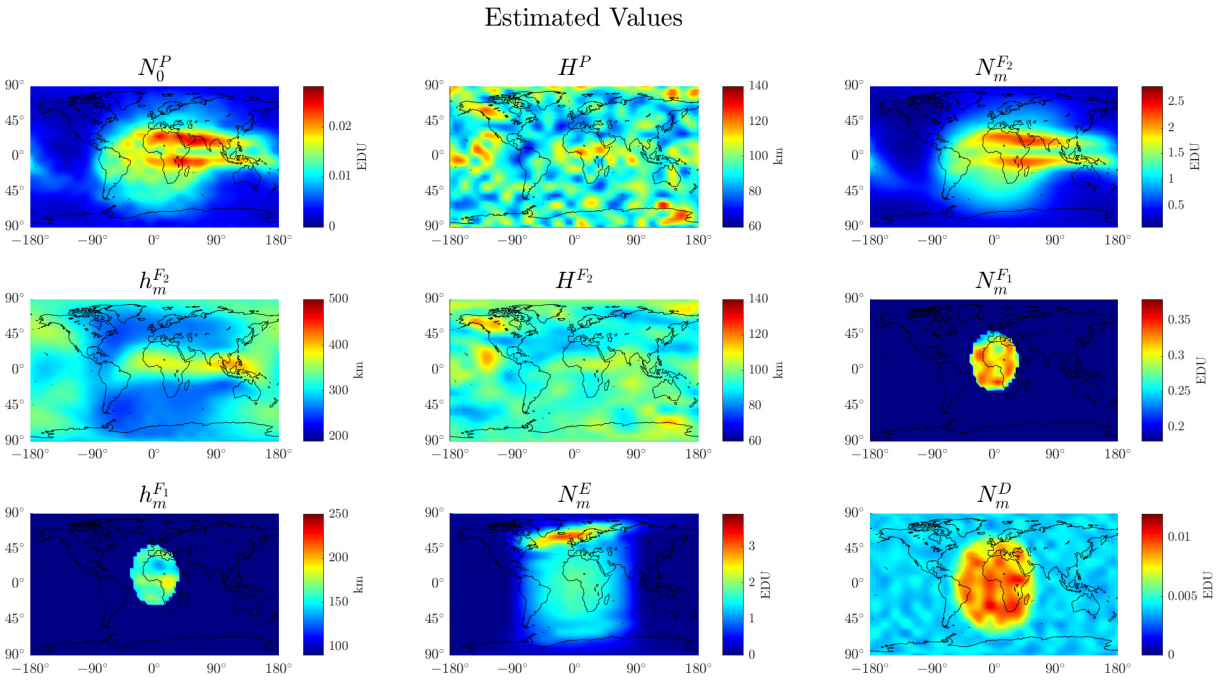


Figure 5.36: Maps of estimated results of the key parameters in scenario 3-6

Although the estimated results contain oscillations because of the noise, they are still located between the lower and upper bounds and no violation happens. The comparison between the truth and reconstructed electron densities is given in Figs. 5.39 and 5.40. The differences are around 0.03 EDU and the average electron density difference along one vertical profile shows the comparably random distribution but is larger in the regions where the ionosphere is more active.

## 5.4 Solution with separability approach

### Scenario 4-1

In this section, we stop the tests based on complete simulated data and start the estimation procedure based on a combination of real observations and simulation. The so-called separability approach is applied for the current scenario [Limberger, 2015]. In this approach, the electron density can be presented as

$$N_e(\varphi, \lambda, h) = VTEC(\varphi, \lambda) \cdot \overline{p}(h) \tag{5.7}$$

where $\varphi$ and $\lambda$ correspond to latitude and longitude whereas $h$ is the height. The VTEC values are taken from VTEC models. In this thesis, we select DFGI-TUM's high-resolution

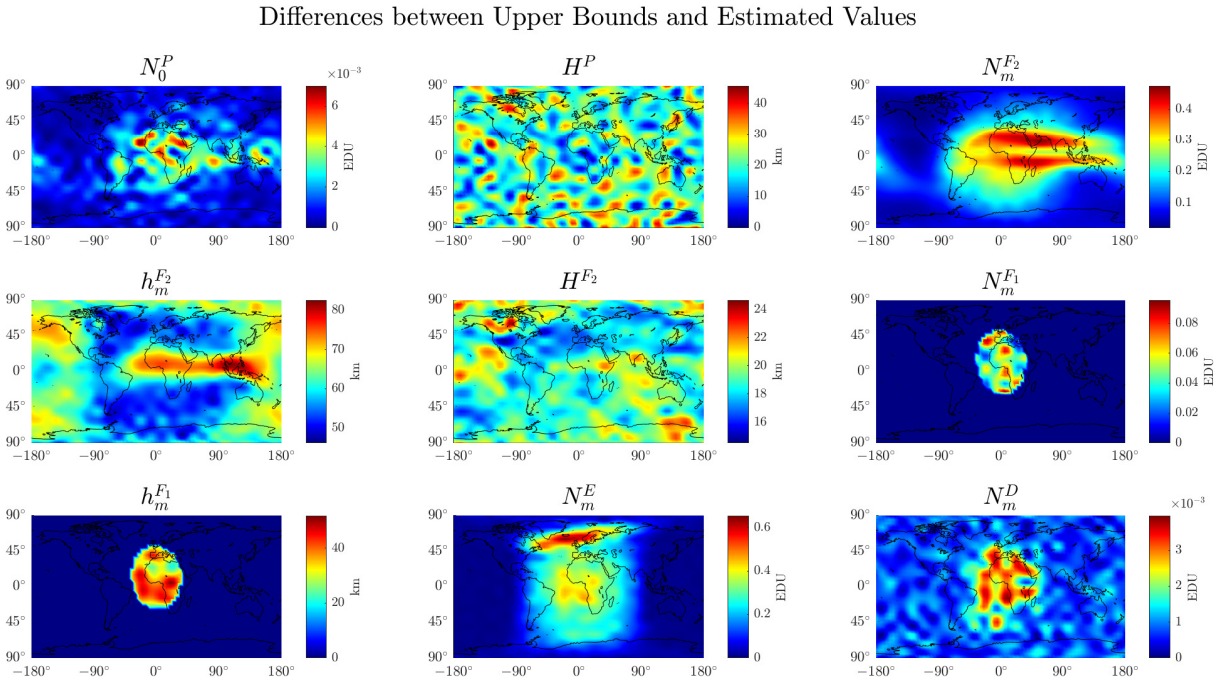


Figure 5.37: Differences between upper bounds and estimated values in scenario 3-6

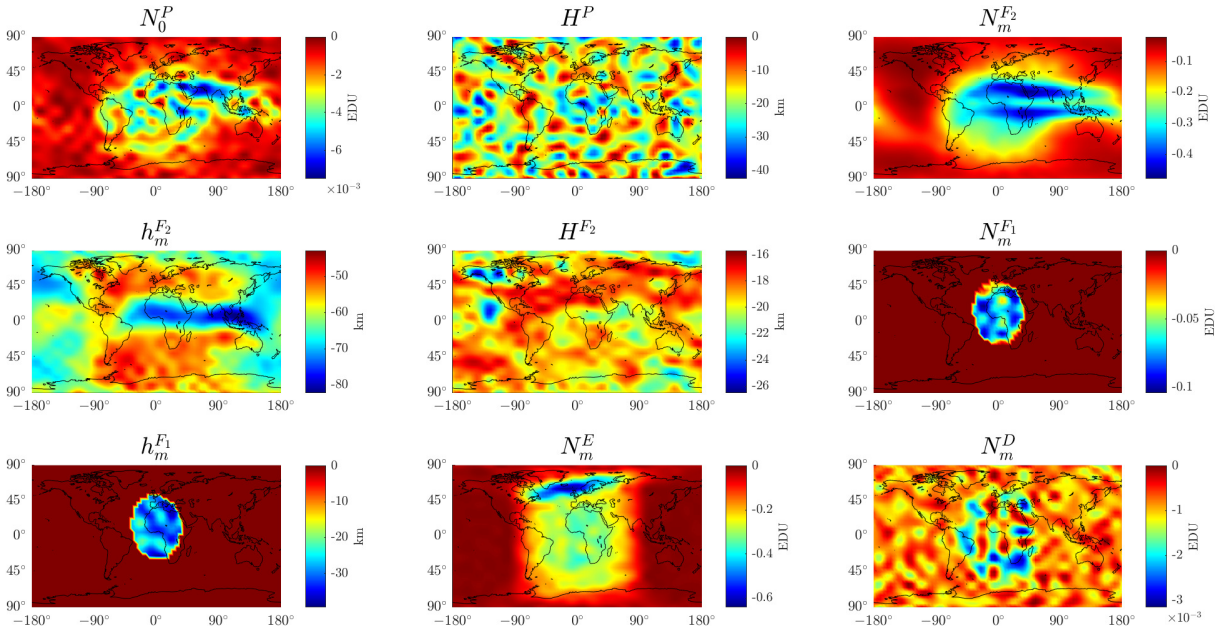


Figure 5.38: Differences between lower bounds and estimated values in scenario 3-6

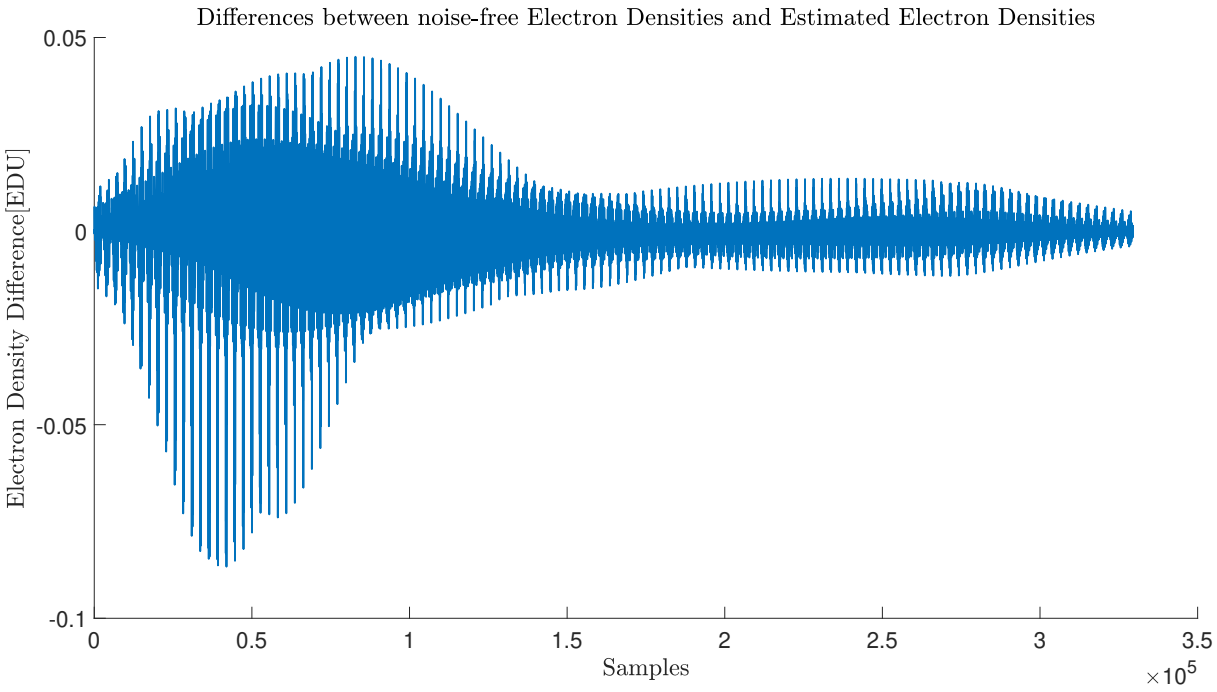


Figure 5.39: 4: Differences between noise-free electron densities and reconstructed electron densities in scenario 3-6

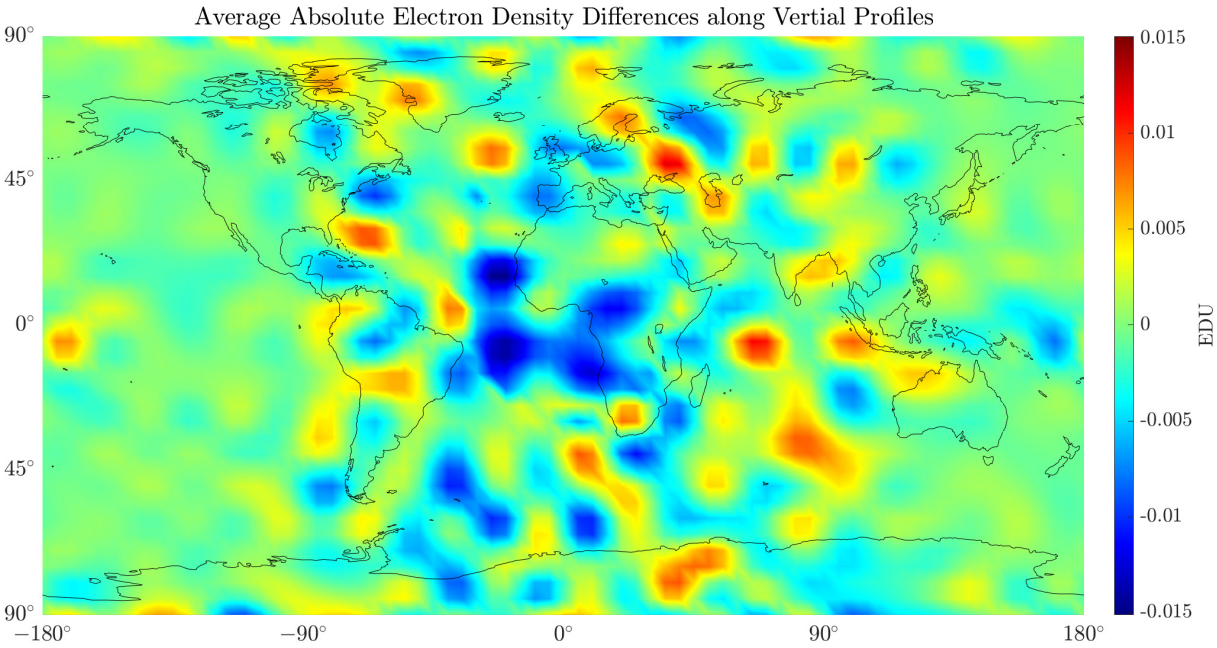


Figure 5.40: Map of average absolute electron density differences along vertical profiles in scenario 3-6

Global Ionosphere Map (GIM) 'othg' which comprises a much higher spectral content (up to spherical harmonic degree $n$=33) as the GIMs from the Ionospheric Associated Analysis Centers (IAAC) of International GNSS Service (IAG) [Lalgudi Gopalakrishnan and Schmidt, 2022]. As stated in the IGS Technical Report 2018 [Villiger and Dach, 2018], 'othg' is one of the worldwide best GIMs. $\overline{p}(h)$ means vertical profile function at the horizontal grid point $P(\varphi, \lambda)$ evaluated at height $h$ fulfilling the normalization condition where $p(h)$ is the electron density at certain height.

$$\overline{p}(h) = \frac{p(h)}{\int_{h_{min}}^{h_{max}} p(h)dh}. \tag{5.8}$$

As constraints, we still use the constraints from scenario 3-5, see Figs. 5.29 and 5.30. The estimated results are presented in Fig. 5.41.

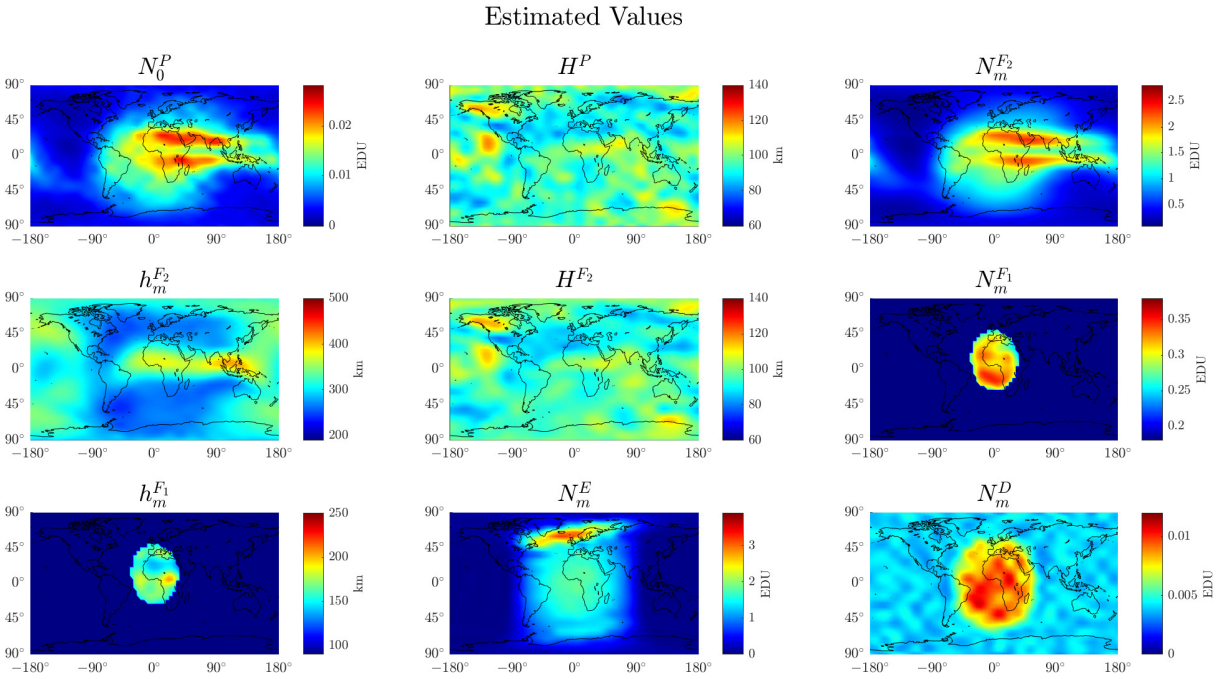


Figure 5.41: Maps of estimated results of the key parameters in scenario 4-1

The results contain more oscillations than the results in scenario 3-6, which means the noise in the separability approach is larger than 5%, around 8%. Figs. 5.42 and 5.43 describe the difference between results and bounds. Since all values in Fig. 5.42 are larger than 0 while all values in Fig. 5.43 are negative, there is no violation in this estimation.

Similar to before, the original and reconstructed electron densities are compared in the following, see Figs. 5.44 and 5.45. It can be found that, the estimation procedure with separability approach comes to similar but a bit worse accuracy than the simulation scenario with 5% noise.

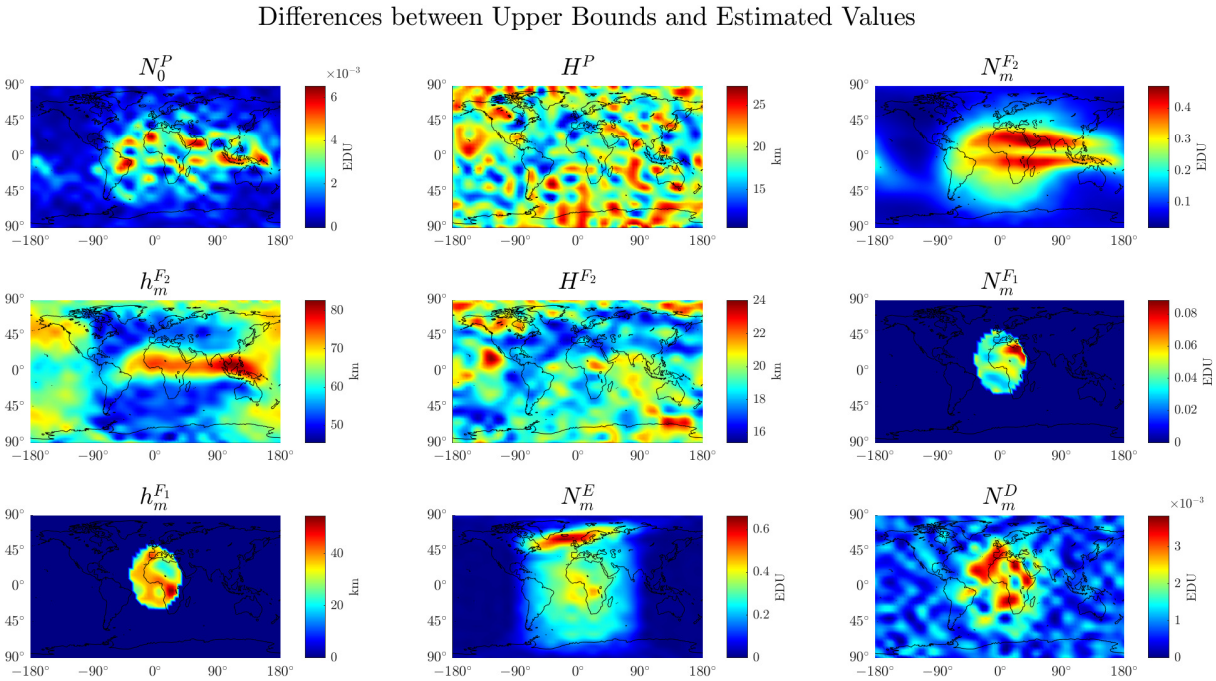


Figure 5.42: Differences between upper bounds and estimated values in scenario 4-1

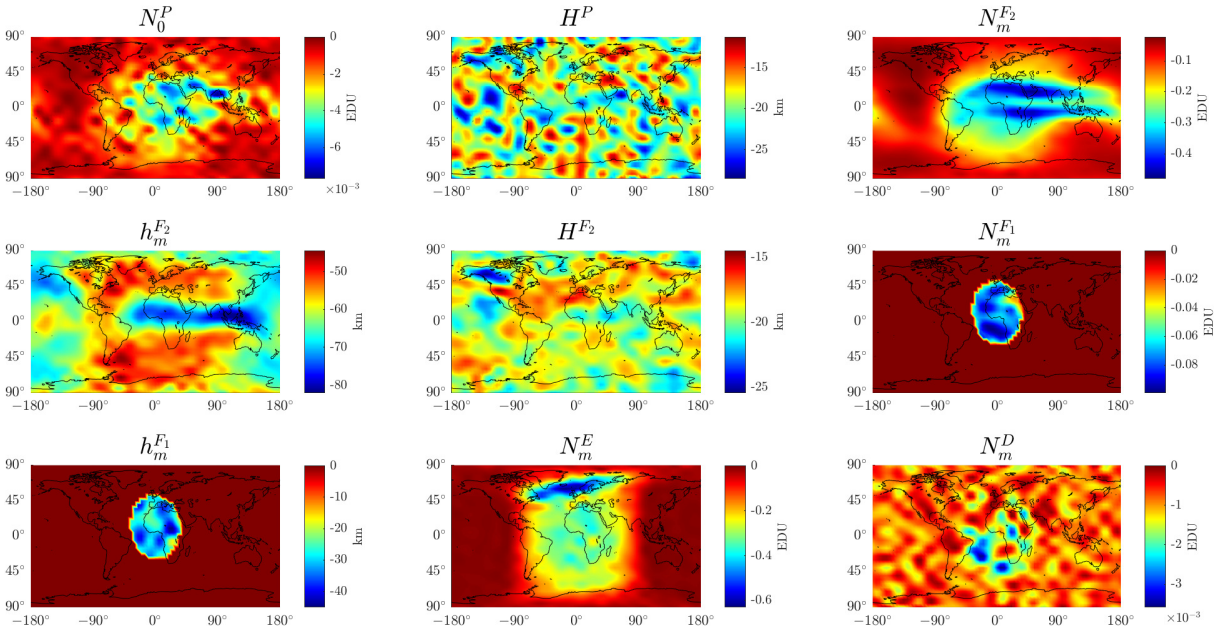


Figure 5.43: Differences between lower bounds and estimated values in scenario 4-1

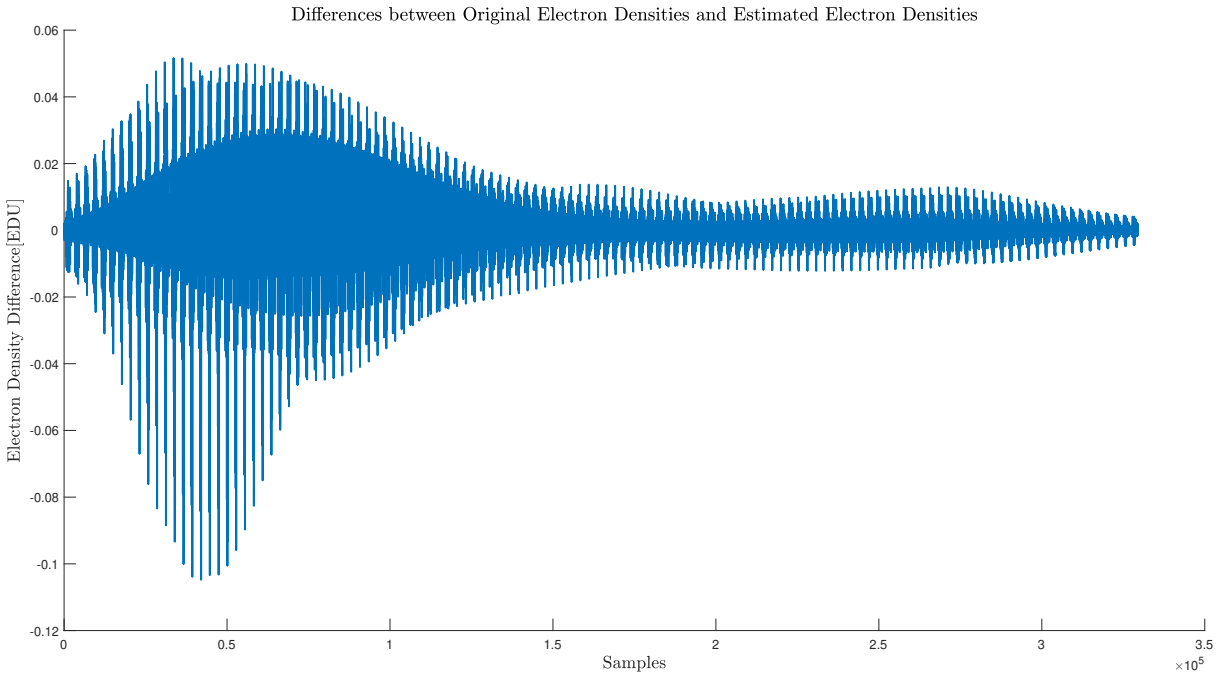


Figure 5.44: Differences between original electron densities and reconstructed electron densities in scenario 4-1

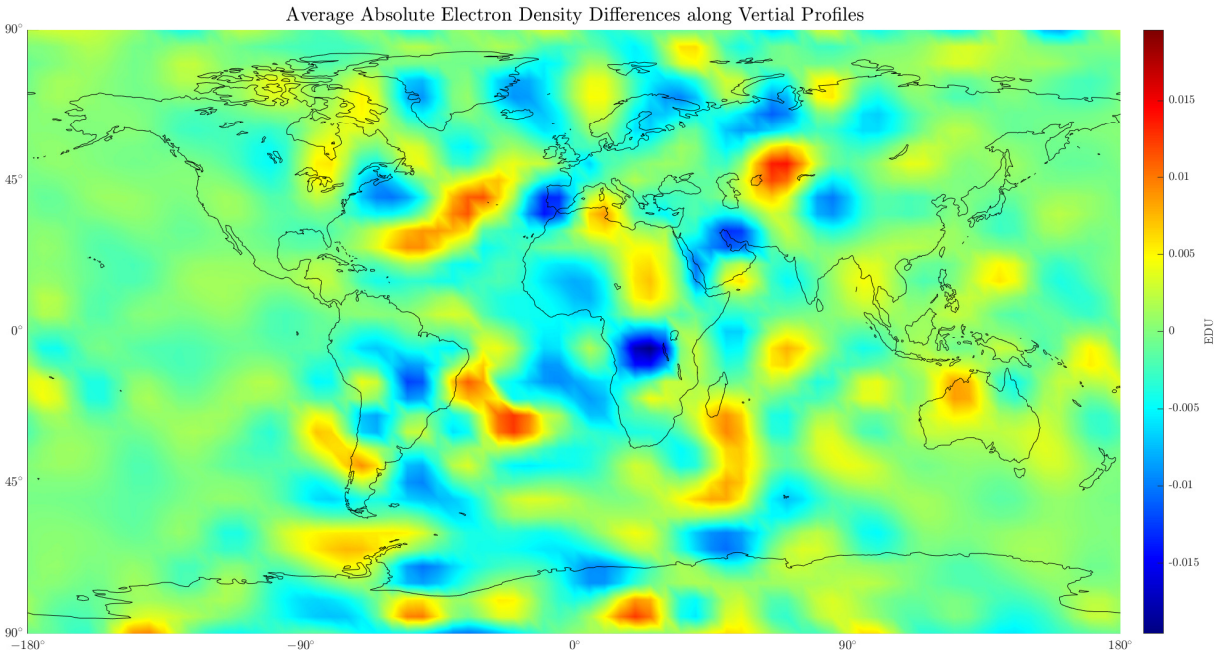


Figure 5.45: Map of average absolute electron density differences along vertical profiles in scenario 4-1

# 6 Conclusions and Outlook

Based on the numerical examples presented in this work, the following conclusions have been drawn and the direction of future works has been given.

## 6.1 Conclusions

Electron density is one of the most important geodetic parameters in modeling the upper atmosphere. Therefore, its precise modeling is both necessary and challenging. In this thesis, the Chapman function is applied for modeling the electron density distribution along vertical profiles. In total 14 key parameters of the MLCM approach are estimated, and each key parameter is presented via a series expansion in terms of 2D tensor products of B-spline functions horizontally. Furthermore, the algorithm of inequality constraint optimization is adopted in order to estimate the key parameters. With the help of inequality constraints, physically unrealistic results such as negative peak density values can be avoided and the estimated results can be located between given bounds. In Chapter 5, different scenarios are selected and estimated based on simulated data or the combination of simulated data and real observations. Some important findings from our investigations can be summarized as follows:

For the scenario in which only one parameter is constrained with inequalities, the result can reach better accuracy than that with the classical least squares method because the step size of the inequality constraint optimization can be set smaller than the level of computational error.

For the scenario with three parameters with inequality constraints, the estimated results cannot reach the same accuracy as in the scenario with one parameter with inequality constraints. In the case of noise, it is no longer possible to set the same constraints for each key parameter in every grid point because this will lead to unrealistic results. The estimated results show that the inequality constraint optimization has the property of resisting noise.

For the nine parameters with inequality constraints scenario, it has the worst accuracy. However, all the estimated results are located in the feasible regions and there is no violation.

For the separability approach solution, here, we apply a combination of real observations (VTEC model values) and vertical profile functions as input data. This means, we now do not rely on IRI. Although there is more noise in this scenario than before, still no violation occurs during the estimation. And it leads to a less than 1% deviation in the comparison of the original and reconstructed electron density.

In our study, the results are quite sensitive to multiple configurations. The initial values of the unknown parameters and the bounds are the most important ones. Therefore, we should be very careful when setting the initial values and bounds.

## 6.2 Outlook

In the future, there is still work for going deeper into this study. Firstly, more real observations from multiple techniques such as radio occultation measurements, e.g. from the Constellation Observing System for Meteorology Ionosphere and Climate (COSMIC) mission need to be combined. Here, we only use the combination of VTEC data and vertical profile functions. Although this is independent from IRI, it is worthy to perform the estimated procedure with pure real observations if more observations are available. The Chapman function, which is used as the main function for modeling the ionosphere vertically, can also be modified or other profile functions can be selected according to the local time and space weather, etc. In the future, if we have a better understanding of the electron density distribution along the height with a variety of observations, it is also essential to modify the profile functions without the current assumptions in the Chapman function. Besides, at the parameter level, although now we already worked with nine inequality constrained parameters, it is still necessary to give some tests for different combinations of nine parameters, and it is even more challenging to work with more than nine inequality constrained parameters. Different B-spline levels for the individual key parameters are also valuable strategies. Furthermore, for the evaluation of the estimated results, external validation data like ground-based observations should be also taken into consideration.

# Bibliography


Amiri-Simkooei, A. and Asgari, J. (2012). Harmonic analysis of total electron contents time series: methodology and results. *GPS solutions*, 16(1):77–88.

Aragon-Angel, A., Zürn, M., and Rovira-Garcia, A. (2019). Galileo ionospheric correction algorithm: An optimization study of nequick-g. *Radio Science*, 54(11):1156–1169.

Balan, N., Otsuka, Y., Bailey, G., and Fukao, S. (1998). Equinoctial asymmetries in the ionosphere and thermosphere observed by the mu radar. *Journal of Geophysical Research: Space Physics*, 103(A5):9481–9495.

Barab, J. (1962). National aeronautics and space administration.

Bauer, S. J. (1957). A possible troposphere-ionosphere relationship. *Journal of Geophysical Research*, 62(3):425–430.

Bellchambers, W. and Piggott, W. (1958). Ionospheric measurements made at halley bay. *Nature*, 182(4649):1596–1597.

Best, M. J. (1984). Equivalence of some quadratic programming algorithms. *Mathematical Programming*, 30(1):71–87.

Bilitza, D., Altadill, D., Zhang, Y., Mertens, C., Truhlik, V., Richards, P., McKinnell, L.-A., and Reinisch, B. (2014). The international reference ionosphere 2012–a model of international collaboration. *Journal of Space Weather and Space Climate*, 4:A07.

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

Bilitza, D., Rawer, K., Bossy, L., Kutiev, I., Oyama, K.-I., Leitinger, R., and Kazimirovsky, E. (1990). International reference ionosphere 1990. Technical report.

Bilitza, D. and Reinisch, B. W. (2008). International reference ionosphere 2007: Improvements and new parameters. *Advances in space research*, 42(4):599–609.

Bishop, G., Secan, J., and Delay, S. (2009). Gps tec and the plasmasphere: Some observations and uncertainties. *Radio Science*, 44(01):1–11.

Boggs, P. T. and Tolle, J. W. (1995). Sequential quadratic programming. *Acta numerica*, 4:1–51.

Bust, G. S. and Mitchell, C. N. (2008). History, current state, and future directions of ionospheric imaging. *Reviews of Geophysics*, 46(1).

Cander, L. R. (2019). *Ionospheric space weather*. Springer.

Cander, L. R., Leitinger, R., and Levy, M. (1999). Ionospheric models including the auroral environment. In *Workshop on Space Weather, Report WPP-155, Noordwijk, The Netherlands, European Space Agency (ISSN 1022-6656)*, pages 135–142.

Chapman, S. (1931a). The absorption and dissociative or ionizing effect of monochromatic radiation in an atmosphere on a rotating earth. *Proceedings of the Physical Society (1926-1948)*, 43(1):26.

Chapman, S. (1931b). The absorption and dissociative or ionizing effect of monochromatic radiation in an atmosphere on a rotating earth part ii. grazing incidence. *Proceedings of the Physical Society (1926-1948)*, 43(5):483.

Chapman, S. and Mian, A. M. (1942a). The rate of ion-production at any height in the earth's atmosphere: Ii—the fourier expression for its daily variation. *Terrestrial Magnetism and Atmospheric Electricity*, 47(1):38–44.

Chapman, S. and Mian, A. M. (1942b). The rate of ion-production at any height in the earth's atmosphere: I—the spherical harmonic representation of its world-wide distribution. *Terrestrial Magnetism and Atmospheric Electricity*, 47(1):31–38.

Ching, B. K. and Chiu, Y. T. (1973). A phenomenological model of global ionospheric electron density in the e-, f1-and f2-regions. *Journal of Atmospheric and Terrestrial Physics*, 35(9):1615–1630.

Chiu, Y. T. (1975). An improved phenomenological model of ionospheric density. *Journal of atmospheric and terrestrial physics*, 37(12):1563–1570.

Choy, S., Zhang, K., Silcock, D., et al. (2008). An evaluation of various ionospheric error mitigation methods used in single frequency ppp. *Positioning*, 1(13).

Coleman, T. F. and Li, Y. (1996). A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM Journal on Optimization*, 6(4):1040–1058.

Cueto, M., Coisson, P., Radicella, S., Herraiz, M., Ciraolo, L., and Brunini, C. (2007). Topside ionosphere and plasmasphere: Use of nequick in connection with gallagher plasmasphere model. *Advances in Space Research*, 39(5):739–743.

Da Rosa, A., Waldman, H., Bendito, J., and Garriott, O. (1973). Response of the ionospheric electron content to fluctuations in solar activity. *Journal of Atmospheric and Terrestrial Physics*, 35(8):1429–1442.

Davies, K. (1990). *Ionospheric radio*. Number 31. IET.

Davies, K. and Baker, D. M. (1965). Ionospheric effects observed around the time of the alaskan earthquake of march 28, 1964. *Journal of Geophysical Research*, 70(9):2251–2253.

Feltens, J., Angling, M., Jackson-Booth, N., Jakowski, N., Hoque, M., Hernández-Pajares, M., Aragón-Àngel, A., Orús, R., and Zandbergen, R. (2011). Comparative testing of four ionospheric models driven with gps measurements. *Radio Science*, 46(06):1–11.

Fiacco, A. V. and McCormick, G. P. (1990). *Nonlinear programming: sequential unconstrained minimization techniques*. SIAM.

Forsgren, A., Gill, P. E., and Wright, M. H. (2002). Interior methods for nonlinear optimization. *SIAM review*, 44(4):525–597.

Fritsch, D. (1982). Second order design of geodetic networks: Problems and examples. In *Proceedings of the International Symposium on Geodetic Networks and Computations of the International Association of Geodesy*.

Fritsch, D. (1983a). Optimal design of two-dimensioal fir-filters. In *ICASSP*.

Fritsch, D. (1983b). Optimal design of two-dimensioal fir-filters. In *ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 8, pages 383–386. IEEE.

Fritsch, D. (1985). Some additional informations on the capacity of the linear complementarity algorithm. In *Optimization and design of geodetic networks*, pages 169–184. Springer.

Gallagher, D. L., Craven, P. D., and Comfort, R. H. (2000). Global core plasma model. *Journal of Geophysical Research: Space Physics*, 105(A8):18819–18833.

García-Rigo, A., Monte, E., Hernández-Pajares, M., Juan, J., Sanz, J., Aragón-Angel, A., and Salazar, D. (2011). Global prediction of the vertical total electron content of the ionosphere based on gps data. *Radio science*, 46(06):1–3.

Gill, P., Murray, W., and Wright, M. (1981). Practical optimization.

Gill, P. E., Murray, W., Saunders, M. A., and Wright, M. H. (1984). Procedures for optimization problems with a mixture of bounds and general linear constraints. *ACM Transactions on Mathematical Software (TOMS)*, 10(3):282–298.

Goldsmith, M. J. (1999). *Sequential quadratic programming methods based on indefinite Hessian approximations*. stanford university.

Goto, Y., Kasahara, Y., and Ide, T. (2012). Improvement of equatorial density distribution of the global core plasma model using gps-derived tec. *Radio Science*, 47(05):1–9.

Gould, N. (2003). Some reflections on the current state of active-set and interior-point methods for constrained optimization.

Gulyaeva, T. (2002). The ionosphere-plasmasphere model software for iso. *Acta Geod. Geophys. Hung.*, 37(3):143–152.

Gulyaeva, T. L., Huang, X., and Reinisch, B. W. (2002). Plasmaspheric extension of topside electron density profiles. *Advances in Space Research*, 29(6):825–831.

Haldoupis, C. (2011). A tutorial review on sporadic e layers. *Aeronomy of the Earth's Atmosphere and Ionosphere*, pages 381–394.

Han, S.-P. (1977). A globally convergent method for nonlinear programming. *Journal of optimization theory and applications*, 22(3):297–309.

Hargreaves, J. K. (1992). *The solar-terrestrial environment: an introduction to geospace-the science of the terrestrial upper atmosphere, ionosphere, and magnetosphere*. Cambridge university press.

Holzworth, R., Kelley, M., Siefring, C., Hale, L., and Mitchell, J. (1985). Electrical measurements in the atmosphere and the ionosphere over an active thunderstorm: 2. direct current electric fields and conductivity. *Journal of Geophysical Research: Space Physics*, 90(A10):9824–9830.

Hoque, M. M., Jakowski, N., and Prol, F. S. (2022). A new climatological electron density model for supporting space weather services. *Journal of Space Weather and Space Climate*, 12:1.

Hu, W., Zheng, D., and Nie, W. (2014). Research on methods of regional ionospheric delay correction based on neural network technology. *Survey Review*, 46(336):167–174.

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. *Advances in Space Research*, 33(6):829–832.

Kelley, M., Siefring, C., Pfaff, R., Kintner, P., Larsen, M., Green, R., Holzworth, R., Hale, L., Mitchell, J., and Le Vine, D. (1985). Electrical measurements in the atmosphere and the ionosphere over an active thunderstorm: 1. campaign overview and initial ionospheric results. *Journal of Geophysical Research: Space Physics*, 90(A10):9815–9823.

Kelley, M. C. (2009). *The Earth's ionosphere: Plasma physics and electrodynamics*. Academic press.

Koch (1981). Hypothesis testing with inequalities.

Koch, K. R. (1985). First order design: Optimization of the configuration of a network by introducing small position changes. In *Optimization and design of geodetic networks*, pages 56–73. Springer.

Kuhn, H. W. and Tucker, A. W. (2014). Nonlinear programming. In *Traces and emergence of nonlinear programming*, pages 247–258. Springer.

Lalgudi Gopalakrishnan, G. and Schmidt, M. (2022). Ionospheric electron density modelling using b-splines and constraint optimization. *Earth, Planets and Space*, 74(1):1–23.

Liang, W. (2017). *A regional physics-motivated electron density model of the ionosphere*. PhD thesis, Technische Universität München.

Limberger, M. (2015). *Ionosphere modeling from GPS radio occultations and complementary data based on B-splines*. PhD thesis, Technische Universität München.

Llewellyn, S. K. and Bent, R. B. (1973). Documentation and description of the bent ionospheric model. Technical report, ATLANTIC SCIENCE CORP INDIAN HARBOUR BEACH FL.

Lötstedt, P. (1984). Solving the minimal least squares problem subject to bounds on the variables. *BIT Numerical Mathematics*, 24(2):205–224.

Luenberger, D. G., Ye, Y., et al. (1984). *Linear and nonlinear programming*, volume 2. Springer.

Lunt, N., Kersley, L., Bishop, G., Mazzella, A., and Bailey, G. (1999). The effect of the protonosphere on the estimation of gps total electron content: Validation using model simulations. *Radio Science*, 34(5):1261–1271.

Lyche, T. and Schumaker, L. L. (2000). A multiresolution tensor spline method for fitting functions on the sphere. *SIAM Journal on Scientific Computing*, 22(2):724–746.

Macalalad, E. P., Tsai, L.-C., Wu, J., and Liu, C.-H. (2013). Application of the taiwan ionospheric model to single-frequency ionospheric delay corrections for gps positioning. *GPS solutions*, 17(3):337–346.

Maeda, K. and Badillo, V. L. (1966). Equatorial spread-f and tropospheric tropical disturbances. *Journal of the Atmospheric Sciences*, 23(6):812–815.

Maes, C. M. (2010). *A regularized active-set method for sparse convex quadratic programming*. Stanford University.

Mead, J. L. and Renaut, R. A. (2010). Least squares problems with inequality constraints as quadratic constraints. *Linear Algebra and its Applications*, 432(8):1936–1949.

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(5):463–478.

More, J. J. and Wright, S. J. (1993). *Optimization software guide*. SIAM.

Nava, B., Coisson, P., and Radicella, S. (2008). A new version of the nequick ionosphere electron density model. *Journal of atmospheric and solar-terrestrial physics*, 70(15):1856–1862.

Nishida, A. (1968). Coherence of geomagnetic dp 2 fluctuations with interplanetary magnetic variations. *Journal of Geophysical Research*, 73(17):5549–5559.

Nocedal, J. and Wright, S. J. (1999). *Numerical optimization*. Springer.

Nsumei, P., Huang, X., Reinisch, B., Song, P., Vasyliunas, V., Green, J., Fung, S., Benson, R., and Gallagher, D. (2003). Electron density distribution over the northern polar region deduced from image/radio plasma imager sounding. *Journal of Geophysical Research: Space Physics*, 108(A2).

Øvstedal, O. (2002). Absolute positioning with single-frequency gps receivers. *GPS Solutions*, 5(4):33–44.

Park, J., von Frese, R. R., Grejner-Brzezinska, D. A., Morton, Y., and Gaya-Pique, L. R. (2011). Ionospheric detection of the 25 may 2009 north korean underground nuclear test. *Geophysical Research Letters*, 38(22).

Peng, J., Zhang, H., Shong, S., and Guo, C. (2006). An aggregate constraint method for inequality-constrained least squares problems. *Journal of geodesy*, 79(12):705–713.

Penndorf, R. (1965). The average ionospheric conditions over the antarctic. *Geomagnetism and Aeronomy: studies in the ionosphere, geomagnetism and atmospheric radio noise*, 4:1–45.

Powell, M. J. (1978). A fast algorithm for nonlinearly constrained optimization calculations. In *Numerical analysis*, pages 144–157. Springer.

Reinisch, B., Huang, X., Song, P., Green, J., Fung, S., Vasyliunas, V., Gallagher, D., and Sandel, B. (2004). Plasmaspheric mass loss and refilling as a result of a magnetic storm. *Journal of Geophysical Research: Space Physics*, 109(A1).

Reinisch, B. W., Haines, D., Bibl, K., Cheney, G., Galkin, I., Huang, X., Myers, S., Sales, G., Benson, R., Fung, S., et al. (2000). The radio plasma imager investigation on the image spacecraft. In *The IMAGE Mission*, pages 319–359. Springer.

Rich, F. J., Sultan, P. J., and Burke, W. J. (2003). The 27-day variations of plasma densities and temperatures in the topside ionosphere. *Journal of Geophysical Research: Space Physics*, 108(A7).

Rockafellar, R. T. (1973). The multiplier method of hestenes and powell applied to convex programming. *Journal of Optimization Theory and applications*, 12(6):555–562.

Roese-Koerner, L. R. (2015). *Convex optimization for inequality constrained adjustment problems*. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, Landwirtschaftliche Fakultät ....

Rush, C. (1989). Ionospheric mapping-an update of fof2 coefficients. *Telecomm. J.*, 56:179–182.

Schaer (1999). *Mapping and predicting the Earth's ionosphere using the Global Positioning System*, volume 59. Institut für Geodäsie und Photogrammetrie, Eidg. Technische Hochschule ....

Schittkowski, K. (1986). Nlpql: A fortran subroutine solving constrained nonlinear programming problems. *Annals of operations research*, 5(2):485–500.

Schmidt, M., Dettmering, D., Mößmer, M., Wang, Y., and Zhang, J. (2011). Comparison of spherical harmonic and b spline models for the vertical total electron content. *Radio Science*, 46(06):1–8.

Schmidt, M., Dettmering, D., and Seitz, F. (2015). Using b-spline expansions for ionosphere modeling. In *Handbook of geomathematics*, pages 939–983. Springer.

Schumaker, L. (1981). Spline functions: Basic theory, john willey and sons. *Inc., New York*.

Schumaker, L. L. and Traas, C. (1991). Fitting scattered data on spherelike surfaces using tensor products of trigonometric and polynomial splines. *Numerische Mathematik*, 60(1):133–144.

Schunk, R. and Nagy, A. (2009). *Ionospheres: physics, plasma physics, and chemistry*. Cambridge university press.

Stollnitz, E. J., DeRose, A., and Salesin, D. H. (1995). Wavelets for computer graphics: a primer. 2. *IEEE computer graphics and applications*, 15(3):76–84.

Tang, J., Cheng, H., and Liu, L. (2012). Using nonlinear programming to correct leakage and estimate mass change from grace observation and its application to antarctica. *Journal of Geophysical Research: Solid Earth*, 117(B11).

Tang, L., Li, Z., and Zhou, B. (2018). Large-area tsunami signatures in ionosphere observed by gps tec after the 2011 tohoku earthquake. *GPS Solutions*, 22(4):1–8.

Teunissen, P. J. and Montenbruck, O. (2017). *Springer handbook of global navigation satellite systems*, volume 10. Springer.

Titheridge, J. (1973). The electron content of the southern mid-latitude ionosphere, 1965–1971. *Journal of Atmospheric and Terrestrial Physics*, 35(5):981–1001.

Venter, G. (2010). Review of optimization techniques.

Villiger, A. and Dach, R. (2018). International gnss service: Technical report 2017.

Webb, P. and Essex, E. (2003). Modifications to the titheridge upper ionosphere and plasmasphere temperature model. *Journal of Geophysical Research: Space Physics*, 108(A10).

Webb, P. A. and Essex, E. A. (2001). A dynamic diffusive equilibrium model of the ion densities along plasmaspheric magnetic flux tubes. *Journal of Atmospheric and Solar-Terrestrial Physics*, 63(11):1249–1260.

Wong, E. (2011). *Active-set methods for quadratic programming*. University of California, San Diego.

Yizengaw, E., Moldwin, M., Galvan, D., Iijima, B., Komjathy, A., and Mannucci, A. (2008). Global plasmaspheric tec and its relative contribution to gps tec. *Journal of Atmospheric and Solar-Terrestrial Physics*, 70(11-12):1541–1548.

Zhang, X. and Tang, L. (2015). Traveling ionospheric disturbances triggered by the 2009 north korean underground nuclear explosion. In *Annales Geophysicae*, volume 33, pages 137–142. Copernicus GmbH.

Zolesi, B. and Cander, L. R. (2014). *Ionospheric prediction and forecasting*. Springer.