SPATIAL STATISTICS AND ANALYSIS OF EARTH'S IONOSPHERE

THOMAS W. BUTLER

Dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

SPATIAL STATISTICS AND ANALYSIS OF EARTH'S IONOSPHERE

by

THOMAS W. BUTLER

B.S., Mississippi State University, 2004

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2013

Approved by

First Reader

Joshua L. Semeter, PhD
Professor of Electrical and Computer Engineering, Boston University

Second Reader

W. Clem Karl
Professor of Electrical Engineering, Boston University

Third Reader

David Castañón
Professor of Electrical and Computer Engineering, Boston University

Fourth Reader

Philip Erickson
Research Scientist, MIT Haystack Observatory

To my family and friends

for their loving support.

To those who never gave up.

Thank you.

The creative potential, the capacity to solve problems, changes in a man in ebbs and flows, and over this he has little control. I had learned to apply a kind of test. I would read my own articles, those I considered the best. If I noticed in them lapses, gaps, if I saw that the thing could have been done better, my experiment was successful. If, however, I found myself reading with admiration, that meant I was in trouble.

His Master's Voice
Stanisław Lem

Acknowledgments

I would like to thank all the researchers and organizations who have provided either insight or data for use in this dissertation. In particular, I would like to thank SRI International, and the Center for Geospace Studies, especially John Kelly, Craig Heinselman, Mike Nicolls, and Mary McCready. My internship at SRI in the summer of 2007 was a very positive and encouraging experience. The staff have such an enthusiastic devotion to introducing students from a wide range of disciplines to the near-Earth space environment, and making sense of such esoteric and acronymic things as ISR, ACF's, IMF, ...

The University of Alaska's Geophysical Institute also deserves credit for operating the Poker Flat Research Range. Besides PFISR, the facilities there have provided invaluable additional context for studying the events described herein. Don Hampton provided the optical data used in Chapters 3 and 4. Bill Bristow and Mark Conde have also been very helpful and gracious.

The staff of the MIT Haystack Observatory has also been dedicated to student outreach. Phil Erickson, Anthea Coster, John Foster, John Holt, Frank Lind.

This material is based upon work supported by the National Science Foundation under Grant Nos. DGE-0221680, ATM-0538868, & ATM-0547934.

Finally, I wish to express my deepest thanks to those patient souls who supported me throughout this long process.

SPATIAL STATISTICS AND ANALYSIS OF EARTH'S IONOSPHERE

(Order No.)

THOMAS W. BUTLER

Boston University, College of Engineering, 2013

Major Professor: Joshua L. Semeter, PhD,
Department of Electrical and Computer Engineering

ABSTRACT

The ionosphere, a layer of Earths upper atmosphere characterized by energetic charged particles, serves as a natural plasma laboratory and supplies proxy diagnostics of space weather drivers in the magnetosphere and the solar wind. The ionosphere is a highly dynamic medium, and the spatial structure of observed features (such as auroral light emissions, charge density, temperature, etc.) is rich with information when analyzed in the context of fluid, electromagnetic, and chemical models.

Obtaining measurements with higher spatial and temporal resolution is clearly advantageous. For instance, measurements obtained with a new electronically-steerable incoherent scatter radar (ISR) present a unique space-time perspective compared to those of a dish-based ISR. However, there are unique ambiguities for this modality which must be carefully considered. The ISR target is stochastic, and the fidelity of fitted parameters (ionospheric densities and temperatures) requires integrated sampling, creating a tradeoff between measurement uncertainty and spatio-temporal resolution.

Spatial statistics formalizes the relationship between spatially dispersed observations and the underlying process(es) they represent. A spatial process is regarded as a random field with its distribution structured (e.g., through a correlation function) such that data, sampled over a spatial domain, support inference or prediction of the process. Quantification of uncertainty, an important component of scientific data analysis, is a core value of spatial statistics.

This research applies the formalism of spatial statistics to the analysis of Earths ionosphere using remote sensing diagnostics. In the first part, we consider the problem of volumetric imaging using phased-array ISR based on optimal spatial prediction ("kriging"). In the second part, we develop a technique for reconstructing two-dimensional ion flow fields from line-of-sight projec-

tions using Tikhonov regularization. In the third part, we adapt our spatial statistical approach to global ionospheric imaging using total electron content (TEC) measurements derived from navigation satellite signals.

Contents

List of Tables

List of Figures

List of Abbreviations

Instruments

AMISR *Advanced Modular Incoherent Scatter Radar* (p.2)

CCD *charge-coupled device* (p.3)

DASC *digital all-sky camera* (p.71)

FPI *Fabry-Pérot interferometer* (p.95)

MSP *meridian-scanning photometer* (p.95)

PFISR *Poker Flat Incoherent Scatter Radar*, an AMISR installation erected at the Poker Flat Research Range in Alaska (p.3)

RISR-N *Resolute Bay* ISR – *north face* (p.3)

Radar terminology

FOV *field-of-view* (p.51)

IPP *interpulse period* (p.41)

IQ *in-phase/quadrature*, a method of representing complex-valued time signals (p.50)

LOS *line-of-sight* (p.4)

r.c.s. *radar cross-section* (p.40)

RTI *range-time-intensity* (p.51)

Geophysics

GMF *geomagnetic field* (p.2)

Global radar imaging

GNSS *Global Navigation Satellite System* (p.4)

IPP *ionospheric pierce point* (p.109)

TEC *total electron content* (p.xiv)

WAAS *Wide-Area Augmentation System* (p.4)

Probability and estimation theory

BLUP *best linear unbiased predictor* (p.15)

cdf *cumulative density function*, e.g. $F_X(x)$ where $F_X(x) = \Pr(X \leq x)$ (p.7)

FRK *fixed-rank kriging* (p.4)

GLS *generalized least squares* (p.18)

$\mathcal{GP}$ *Gaussian process*, e.g. $X(t) \sim \mathcal{N}(\mu, \Sigma)$ (p.25)

i.i.d. *independent and identically distributed* (p.16)

LLSE *linear least squares estimator* (p.16)

MCMC *Markov chain Monte Carlo* (p.136)

MLE *maximum likelihood estimator* (p.16)

MSPE *mean-square prediction error* (p.12)

OLS *ordinary least squares* (p.17)

pdf *probability density function*, e.g. $f_X(x)$ where $F_X(x) = \int_{-\inf}^{x} f_X(t)\,dt$ (p.8)

r.p. *random process*, e.g. $X(t) \sim f_{X(t)}$ (p.9)

r.v. *random variable*, e.g. $X \sim f_X$ (p.7)

r.v. *random vector*, e.g. $\underline{X} \sim f_{\underline{X}}$ (p.8)

SNR *signal-to-noise ratio*, a ratio of powers (p.41)

SSS *strict-sense stationary*, a.k.a. strictly stationary (p.10)

WSS *wide-sense stationary*, a.k.a. weakly stationary (p.10)

On notation

Since this thesis makes use of both multidimensional quantities (e.g. sets of parameters) and spatial coordinates, it is helpful to distinguish these notationally. For this purpose, a *vector* is a physical quantity (such as position or velocity) having both magnitude and direction and is typeset thus: $\mathbf{v} = \begin{bmatrix} v_x & v_y & v_z \end{bmatrix}^\mathsf{T}$. An *array* is an ordered n-tuple used in computations and is typeset so: $\underline{e} = \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix}^\mathsf{T}$. Formulas with matrix-vector products, for instance, become matrix-array products, so to speak (though a position vector may still appear as an argument, as in $\underline{x}(\mathbf{s})$). Whenever an ordered sequence of vectors (e.g. a vector-valued field) is involved in computations with a matrix, the sequence is decomposed into its individual components and the components are stacked. Such an array signified by combining the notation for arrays and vectors:

$$\begin{aligned} \underline{\mathbf{v}} &= \begin{bmatrix} \underline{v}_x^\mathsf{T} & | & \underline{v}_y^\mathsf{T} & | & \underline{v}_z^\mathsf{T} \end{bmatrix}^\mathsf{T} \\ &= \begin{bmatrix} v_x(\mathbf{s}_1) & \cdots v_x(\mathbf{s}_n) & | & v_y(\mathbf{s}_1) & \cdots v_y(\mathbf{s}_n) & | & v_z(\mathbf{s}_1) & \cdots v_z(\mathbf{s}_n) \end{bmatrix}^\mathsf{T}. \end{aligned}$$

Matrices are set as boldface, usually majuscule, letters. Both vectors and arrays are regarded as columns. When it is important to regard matrices as groups of augmented arrays, the construction proceeds column-wise.

Also, matrix products and *particularly matrix inverses* represent a rather general form of shorthand. Their purpose is pedagogical, to emphasize the algebraic patterns of their derivation. In practice, *these terms should seldom be computed directly!* Instead, many subroutines exist to solve the linear system $\mathbf{A}\underline{x} = \underline{b}$ which exploit the structure of the matrix $\mathbf{A}$, and which are both faster and more numerically stable than direct computation of $\mathbf{A}^{-1}$. Additionally, it is not advisable to form the normal equations matrix $\mathbf{A}^\mathsf{T}\mathbf{A}$, since (among other things) the matrix product is usually performed in the native numerical precision of $\mathbf{A}$, and the truncation error of this operation carries through to all subsequent operations (Golub and Van Loan, 1996).

Chapter 1

Introduction

The material for this thesis was prompted by an interest in remote sensing of the Earth's ionosphere. I was told about a new radar. It had the unusual ability to steer its beam in many directions during the time a dish antenna would dwell in one direction, transmitting and receiving enough pulses to develop reliable statistics. This seemed interesting, if perhaps a little haphazard (petulant, even!). Then I was told the advantage of such a mode: direct three-dimensional imaging of the ionosphere! And I could be one of the first to try it out! Now that was a project!

I spent the next several years digging through data, learning about the ionosphere itself, looking for a project. Along the way, I developed a few visualizations I was quite proud of. One of the challenges in making a graphical representation of this data is interpolating it to a regular rectangular grid. Sometimes the results would look great, sometimes not so great. Since acquiring a more detailed image meant sacrificing integration time, we knew we were dealing with quite a noisy signal. So how much faith should we put in an interpolation based on the fact that the shapes it reveals seem "coherent?"

In other words, how do we distinguish spatial structure from spatial randomness? Do we at all?[1] It's an issue that seems to be ignored at least as often as it's encountered.

And so I encountered kriging. At first it was simply a useful way of getting data lined up on a display without the artifacts of linear or cubic interpolation. But it also came with a "kriging variance," which seemed to be a way of describing just what I had been trying to articulate: the idea of spatial uncertainty. And the opportunity of making use of that uncertainty if you have a model to describe it accurately. This led me to the realm of spatial statistics, where I stand now. This document bridges two periods of my career, as I hope my work will help unite the communities of researchers involved in these fields.

[1] This question echos the old objectivist/subjectivist divide in statistics. I maintain that we *should* acknowledge the distinction, as well as the ambiguity in decoupling the two (God *does* play dice!). We should state our assumptions about how the two factor, and **always include error bars!**

1.1 The ionosphere

The material for this thesis has evolved from work concerning the optimal analysis of remote sensing observations of the Earth's *ionosphere*. The ionosphere (altitude 90 km to 1000 km) is the layer of Earth's atmosphere defined with respect to the behavior of charged particles. Namely, the kinetic energy of charge carriers within the ionosphere is comparable to the energy of their Coulomb attractions. On the smallest scale, positive and negative charges continually oscillate while the aggregate gas appears neutral on the whole. This *quasineutral* state is called a *plasma*,[2] and because it is governed by a combination of fluid-mechanical and electromagnetic laws, plasma physics remains an active field of research. Although the laws of plasma physics are well-established from first principles, their behavior is often complex.

Earth's ionosphere is of particular interest for its availability and its size. But beyond these, the ionosphere is part of an active geospace environment. The response of the ionosphere to the many environmental drivers (among them Earth's gravity, the *geomagnetic field* (GMF), and the solar wind) provides proxy diagnostics for those same drivers.

Measurements of the ionosphere are thus characteristically complex: dynamic activity with rich (and often surprising) spatial and temporal patterns. Nevertheless, studying these patterns and structures has led to many discoveries, despite the complexity of the underlying processes. Indeed, as such structures are observed with finer resolution, ever more unexpected processes continue to be discovered.

1.2 Spatial statistics & measurement

While statistical analysis is nothing new to the ionospheric science community, instruments are becoming available with the bandwidth, throughput, and resolution to present a uniquely spatial context.[3] For example, *Advanced Modular Incoherent Scatter Radar* (AMISR), an instrument for studying the ionosphere, uses a phased antenna array to rapidly acquire measurements from many directions.

One might presume to use image analysis techniques or some ad hoc spatial extension of the usual statistical tools, which were developed for radar in 1D. But this raises the question of whether the information contained in the data has been used optimally. This, in turn, is subject to how one

[2] Plasma, the so-called "fourth phase" of matter seems quite exotic to us Earthlings, sheltered as we are by Earth's magnetic field. In fact, some 99 % of the matter in the universe is in the plasma state.

[3] This is part of a larger trend toward *doing* more with greater *numbers* of smaller devices. Consider also small satellites, sensor networks, and aggregators of Big Data.

understands the terms "information" and "optimally."

This work does not attempt to quantify or derive theoretical bounds for the information within observations of process *A* by instrument *B*. Nor does it claim to present the "best" method of analysis. The techniques presented herein are justified within the context of spatial statistics, but represent only a limited subset of possible choices. As always, the practitioner's decision relies on a combination of practical constraints, experience, and personal preference.

The approach presented here is rooted in a branch of statistics tailored to the case of data and random processes whose relative positions in space strongly influence their distributional properties. Spatial statistics is the subject of Chapter 2.

While an electronically-steerable instrument like AMISR is quite valuable in single-beam experiments (e.g. Varney et al., 2009), it is notable for its ability to repoint the beam on a pulse-by-pulse basis. This enables experimenters to capture the complex spatial structure of the ionospheric plasma in a "snapshot" mode.

In the case of a dish radar, the only reasonable strategy for accumulating statistics is to dwell in a given direction long enough to gather a statistically significant sample. Steer the antenna, and repeat. On the other hand, an electronically-steerable phased array can send a single pulse, receive its return signal, then adjust its phase table to steer the beam. Cycling in this way through a pre-defined table of look directions, each returned power signal is registered in 3D space. In this way it constructs a 3D image through a kind of time-domain multiplexing. The beam steering is fast enough that each sweep is essentially simultaneous, so that the resolved image truly represents the average activity within the region of interest. This can be likened to the scanning of a *charge-coupled device* (CCD), in that the radar acquires the measurements needed to form an *autocorrelation function* (ACF) essentially simultaneously in all directions through a raster scan (or similar) of the sky. Extending that analogy, a scanning-mode dish antenna would be a slit-scan camera.

There are presently two AMISR installations: *Poker Flat Incoherent Scatter Radar* (PFISR) in Alaska and the *Resolute Bay* ISR – *north face* (RISR-N) in Nunavut, Canada. PFISR will be relocated to Argentina in 2013, plans are underway to construct RISR's south-facing companion, and an Antarctic AMISR mission has long been talked about. The EISCAT Scientific Association's EISCAT 3D project, also based on a phased-array platform, promises even greater flexibility. We expect more researchers to take advantage of these tools as they become available.

1.3 Flow field estimation

The spectrum of incoherent scatter returns is very accurately modeled (Farley, 1960; Evans, 1969), given perfect knowledge of a small set of state parameters. Recovery of these state parameters from data is a problem of inverse theory. For instance, the plasma drift velocity results in a bulk Doppler shift of the return spectrum. Estimating this drift is equivalent to estimating the projection of plasma drift onto the direction of the radar *line-of-sight* (LOS).

A monostatic radar (single transmitter/receiver) can measure only this one component. However, using neighboring measurements, it is possible to reconstruct vector velocities. Indeed, using constraints motivated by physical properties of the random process, it is possible to reconstruct a flow-field from a spatially distributed set of monostatic measurements. Chapter 4 discusses one such method of reconstruction as an example.

1.4 Total electron content

The 3D *incoherent scatter radar* (ISR) imaging application above is an example of optimal interpolation or spatial prediction. Spatial statistics can also be applied to the problem of global mapping of satellite observations. Nychka et al. (2002), Berliner et al. (2003), Stein (2008), and Kang et al. (2010) all demonstrate Bayesian methods for efficient global-scale prediction from sparse satellite measurements. Wikle et al. (2001) and Cressie and Johannesson (2006) demonstrate multi-resolution approaches. Cressie and Johannesson (2008) introduce *fixed-rank kriging* (FRK), a reduced-dimensional method also suitable for global prediction.

A related and independent diagnostic of the ionosphere is *total electron content* (TEC), or electron density integrated along a column. TEC is related to the total ionization encountered on the ray path of a satellite-to-ground signal, e.g. a *Global Navigation Satellite System* (GNSS) signal. The ionosphere's effect on the navigational accuracy of GNSS signals is significant enough to warrant the development of augmentation systems (such as *Wide-Area Augmentation System* (WAAS), (see Blanch, 2004; Sparks et al., 2011a)), which use TEC to correct for these effects. Dense, global coverage of TEC estimates is limited by the orbital path of the satellites and the availability of ground receivers). Hence the need for spatial prediction if no data are available at a requested position.

While GNSS augmentation systems typically use regional results, global TEC mapping is also of interest to atmospheric physicists, since (in well-covered regions) this provides a high spatial and temporal resolution glimpse of ionospheric events. This is especially interesting in conjuction

with electron density imaging from ISR in chapter 3, which (after integrating) provides a direct comparison with TEC. Global prediction of TEC is the subject of chapter 5.

1.5 Major contributions of this dissertation

The sections above outline the specific topics comprising the chapters to follow. The projects described therein may appear disjointed. In fact, a few threads tie the subjects of chapters 3 to 5 together. The overarching themes constitute the major contributions of this dissertation:

- A framework for remote sensing, drawing from spatial statistics and described in chapter 2.

Example applications of this framework to

- ISR imaging (direct 3D imaging of the ionospheric state parameters, chapter 3),
- plasma flow field reconstruction (higher-level analysis of ISR state parameters, chapter 4),
- regional and global mapping of TEC (spatial statistics applied to satellite measurements, chapter 5).

In addition, some practical matters are addressed and implemented in the accompanying Python and MATLAB codes:

- tips for efficient (i.e. vectorized) techniques for implementing spatial statistical data analysis on medium- to large-size data sets
- suggestions for Bayesian implementations accommodating non-linear models and non-Gaussian distributions

Not limited to ionospheric science

Although this thesis presents spatial statistics entirely within the context of remote sensing of the ionosphere, the techniques are applicable to any framework in which observations possess a dependency structure relative to their location in space.

Spatial statistics can aid in the efficient deployment of sensor networks. Le and Zidek (2006) discuss geostatistical data analysis for environmental monitoring networks: estimation of structural parameters for the purpose of designing efficient data gathering networks (see also Zidek et al., 2012).

Even more generally, spatial statistics can be used to describe and analyze properties related by their proximity in non-spatial senses of distance (e.g. feature space, or discrete / cabdriver distance, constrained distances, river crossing, etc.). Kuzma (2004) expounds on the connections between spatial statistics, direct inversion, (e.g. Tikhonov regularization, least squares, etc.), and support vector machines, which "can be applied to data whose axes are any form of data," not only spatial coordinates.

Chapter 2

Mathematical Preliminaries

The aim of science is not to open the door to infinite wisdom but to set a limit to infinite error.

Life of Galileo
Bertolt Brecht

Modern remote sensing platforms provide an abundance of spatial (and spatiotemporal) data. The sheer volume of which data could overwhelm computing resources if handled naïvely. Furthermore, such data presents other challenges relating to their uniquely spatial nature. For instance, measurements are often collected at arbitrary or random positions, either punctually or integrated over a region, or in tandem with some other (non-coincident) dataset. Whereas the objective is often to infer some properties over a continuous spatial domain, data are necessarily finite and discrete (so that any dataset, however large, is necessarily incomplete). To make inferences at any point within the domain, the first step is to "fill in the gaps" not covered by the data. Hence the emphasis in this chapter on optimal spatial prediction.

2.1 Probability and Statistics

It is assumed the reader is familiar with probability and statistics at the undergraduate level. Any of the standard texts will provide the necessary background. However, for reference and consistency of notation, the most commonly used elements are defined here.

2.1.1 Random variables

A *random variable* (r.v.) X maps a random event to the space of real numbers. X is associated with a function $P_X : \mathbb{R} \mapsto [0,1]$, called the *cumulative density function* (cdf):

$$P_X(a) = \Pr(X \leq a).$$

Alternatively, X is characterized by the *probability density function* (pdf):[1]

$$p_X(a) = \frac{d}{da} P_X(a) \qquad \text{with} \quad p_X(x) \geq 0 \text{ and } \int p_X(x)\, dx = 1.$$

Moments

Whenever attention to pdfs is restricted to summaries, only the first two moments are considered. The *mean* of X is

$$\mu_X = \mathrm{E}[X] = \int s\, p_X(s)\, ds,$$

and the *variance* is

$$\sigma_X^2 = \mathrm{E}\left[(X - \mu_X)^2\right] = \mathrm{E}\left[X^2\right] - \left(\mathrm{E}[X]\right)^2.$$

For two r.v.s X and Y, the *covariance* is

$$\sigma_{XY} = \mathrm{E}\left[(X - \mu_X)(Y - \mu_Y)\right] = \mathrm{E}\left[XY\right] - \mu_X \mu_Y.$$

It is often convenient to begin by restricting our attention to the first two moments of a pdf. (This implies the r.v. X has a Gaussian distribution, which is completely specified by its first two moments.) The methods developed in classical geostatistics invoke this approximation. When interpreting results, it is important to keep in mind that this approximation is quite strong and might be unjustified.

2.1.2 Random vectors

A principal theme of spatial statistics is the dependency of neighboring samples. It will be necessary to consider the interrelationship between many *random variables* at once. These can be stacked into columns to form the more convenient *random vector* (r.v.) notation. For instance, N variables make up the r.v.

$$\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}.$$

The expectation operator works element-wise. So the *mean vector* is just the stacked vector of means:

$$\underline{\mu}_X = \mathrm{E}\left[\underline{X}\right] = \begin{pmatrix} \mathrm{E}\, X_1 \\ \vdots \\ \mathrm{E}\, X_N \end{pmatrix},$$

[1] This thesis deals only with continuous r.v.s, so the notation for discrete r.v.s is ignored here.

and the second moments are matrices: the *covariance matrix*

$$\Sigma_X = \mathrm{E}\left[\left(\underline{X} - \underline{\mu}_X\right)\left(\underline{X} - \underline{\mu}_X\right)^{\mathsf{T}}\right] = \mathrm{E}\left[\underline{X}\,\underline{X}^{\mathsf{T}}\right] - \underline{\mu}_X\underline{\mu}_X^{\mathsf{T}}$$

and the *cross-covariance matrix*

$$\Sigma_{XY} = \mathrm{E}\left[\left(\underline{X} - \underline{\mu}_X\right)\left(\underline{Y} - \underline{\mu}_Y\right)^{\mathsf{T}}\right] = \mathrm{E}\left[\underline{X}\,\underline{Y}^{\mathsf{T}}\right] - \underline{\mu}_X\underline{\mu}_Y^{\mathsf{T}}.$$

2.1.3 Random processes

A *random process* (r.p.) is a generalization of a *random vector* to functions with continuous arguments. It may be either scalar- or vector-valued. A realization of $X(t)$, the sample path $x(t)$, is a deterministic function of t. For any argument t, $X(t) = X$, a random variable.

The *mean process* is also a function of time or space:

$$\mu_X(t) = \mathrm{E}\left[X(t)\right] = \int x\, p_{X(t)}(x;t)\,dx.$$

The second-order moments are functions of two variables. Following typical conventions, denote the *autocorrelation function*

$$R_X(u,v) = \mathrm{E}\left[X(u)\,X(v)\right]$$

and the *autocovariance function*

$$\begin{aligned} C_X(u,v) &= \mathrm{E}\left[\left(X(u) - \mu_X(u)\right)\left(X(v) - \mu_X(v)\right)\right] \\ &= R_X(u,v) - \mu_X(u)\,\mu_X(v). \end{aligned}$$

The cross-correlation $R_{XY}(\cdot,\cdot)$ and cross-covariance $C_{XY}(\cdot,\cdot)$ functions (respectively) are defined similarly:

$$\begin{aligned} R_{XY}(u,v) &= \mathrm{E}\left[X(u)\,Y(v)\right], \\ C_{XY}(u,v) &= \mathrm{E}\left[\left(X(u) - \mu_X(u)\right)\left(Y(v) - \mu_Y(v)\right)\right] \\ &= R_{XY}(u,v) - \mu_X(u)\,\mu_Y(v). \end{aligned}$$

2.1.4 Stationarity

Of prime importance in optimal prediction of a spatial random process is the characterization of its distributional properties for all $\mathbf{s} \in D_s \subset \mathbb{R}^d$. Stationarity simplifies this. Consider the r.p. $X(\mathbf{s})$

sampled at positions $\mathbf{s}_i \in D_s,\ i = 1,\ldots,k$, so that the joint cumulative distribution of the samples is $P_X(x(\mathbf{s}_1),\ldots,x(\mathbf{s}_k))$. Then $X(\mathbf{s})$ is *strict-sense stationary* (SSS) (or strongly stationary) if, for all k, for all $\mathbf{u}$, and for all $x(\mathbf{s}_i),\ i = 1,\ldots,k$,

$$P_X\left(x(\mathbf{s}_1),\ldots,x(\mathbf{s}_k)\right) = P_X\left(x(\mathbf{s}_1+\mathbf{u}),\ldots,x(\mathbf{s}_k+\mathbf{u})\right). \tag{2.1}$$

In particular, strict stationarity implies the mean does not vary with $\mathbf{s}$ and, if the correlation function exists, it depends only on the distance between two samples:

$$\mu_X(\mathbf{s}) \equiv \mu_X\ ; \qquad C_X\left(\mathbf{s}_1,\mathbf{s}_2\right) \equiv C_X\left(\mathbf{s}_2-\mathbf{s}_1\right). \tag{2.2}$$

A process that satisfies just (2.2) is called *wide-sense stationary* (WSS) (or weakly stationary). A WSS process need not satisfy the conditions (2.1), but it has useful properties. The larger class of *intrinsically stationary* processes includes all WSS processes. These have a spatially invariant mean *and* satisfy the slightly weaker second-order condition

$$E\Big[\left(X(\mathbf{s}_1)-X(\mathbf{s}_2)\right)^2\Big] = 2C_X(\mathbf{0}) - 2C_X(s_2-s_1),$$

i.e. their increments are *wide-sense stationary*.

2.2 Optimal Spatial Prediction

Prediction versus estimation

The ultimate goal of geostatistical data analysis is usually *prediction* of the numerical value of a function at an unmeasured location. Optimal spatial prediction typically takes the form of a (linear) predictor that incorporates models describing the data and the underlying process. But, unlike time series, spatial data lack the following simplifying properties:

- one-dimensionality,
- prescribed directionality (i.e. the "arrow of time" delineating cause and effect),
- (often) uniform sampling.

So, for example, there exists no analogous notion of causality in 2D or 3D; the predicted value at any point is influenced by its neighbors in all directions.

Optimal prediction differs from deterministic interpolation by incorporating a specialized statistical model of the r.p.. In the time domain, optimal interpolation (a.k.a. *filtering* or *smoothing*)

refers to inference on the state of a system based on a time series of observations. It is common to refer to the output of, say, a Wiener filter as an "estimate" of the process at that time. In this context, "estimation" is synonymous with "prediction" in its intuitive sense: a guess of the upcoming state, informed by the immediate history of the system and its typical behavior in response to causal stimuli.

On the other hand, the following distinction (due to Cressie (1993, pp.105–106)) is illustrative of the stages of spatial statistical analysis. In geostatistics, spatial prediction is often preceded by a separate stage of *structural analysis*, which entails selecting a model under which predictions possess minimal uncertainty. Often this model belongs to a parameterized class, and structural analysis involves specifying the model parameters, which are either assigned based on prior knowledge or estimated via the usual statistical methods (likelihood, method of moments, etc.). Hence, **estimation** refers to inference on fixed but unknown parameters, while **prediction** refers to inference on the random process.

2.2.1 Geostatistics and spatial statistics

The problem of optimal prediction can be described generally as follows: given a set of observations $\{y_i \mid i = 1, \ldots, n\}$, determine the value $\widehat{y}_{n+1}$ that minimizes a certain objective function. The data $\{y_i\}$ are ultimately sampled from the r.p. Y with joint distribution $P(Y_1, \ldots, Y_n)$.

Spatial data are distinguished by their dependence of neighboring measurements. This is expressed differently for continuous or discrete spatial domains, and for zero-volume (punctual) or finite-volume (regional) measurements, but all are variations of the First Law of Geography, articulated by Waldo R. Tobler: "[E]verything is related to everything else, but near things are more related than distant things" (Tobler, 1970).

In other words, spatial random processes exhibit a distance decay relationship. This highlights the role of the structure function in spatial data analysis: individual observations are *not* independent. Optimal prediction exploits the redundancy of such measurements within the context of such a structure model.

The theory of optimal prediction was applied to spatial data beginning in the early 1960's. It is generally attributed to the geologist Georges Matheron and the meteorologist Pierre Gandin who, building upon the seminal works of Wiener, Kolmogorov, and others, independently developed the optimal predictor for this problem. (See Cressie (1990) for a more complete early history.) Naturally, the optimal predictor of the spatial random process relies on suitable distributional

assumptions regarding the underlying r.p.. These assumptions are commonly expressed in a statistical model, whose parameters are estimated from the data themselves. Probabilistic models acknowledge spatial uncertainty, expressed as either imperfect or incomplete data about the some quantity, or as the degree of variability inherent to that quantity within the domain. This methodology implicitly assumes some degree of regularity within the region of interest (the ergodic principle). Using the weaker assumption of (second-order) stationarity, Matheron and Gandin each derived linear unbiased predictors of Y at $\mathbf{s}_0$ based on data at $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_m)^\mathsf{T}$ that minimize the *mean-square prediction error* (MSPE)

$$\text{MSPE} = \mathrm{E}\left[Y(\mathbf{s}) - \left(\widehat{Y}(\mathbf{s}_0) \right)^2 \right]. \tag{2.3}$$

Matheron called this method "kriging" after D.G. Krige, a South African mining geologist whose work preceded Matheron's development.

Kriging became the basis for a branch of study called *geostatistics*, a philosophy for applying probabilistic methods of inference on random variables (such as mining or oil deposits) over a continuous domain. These regionalized variables exhibit both spatial correlation and high irregularity of detail.

For instance, large-scale patterns induces a spatial trend on data, perhaps due to mineral deposit patterns. Other processes introduce small-scale variability. These qualitative properties comprise a spectrum from long-range spatial dependence ("smoothness") to short-range detail ("roughness"). It is this balance of factors, often (following Tobler) expressed as a function of the distance between samples, which allows practitioners to quantify *spatial uncertainty*, a major goal of geostatistics. (Chilès and Delfiner, 2012, p. 2).

As a method of data analysis, kriging demands a detailed model of the underlying processes at each data position, including the dependence structure of observations over the entire domain. With a view to model selection and estimation, the underlying r.p. is often assumed to be (either second-order or implicitly) stationary and to have sufficient sample coverage that the process is separable into low-frequency and high-frequency components. The low-frequency (large-scale) part can be fitted to a trend model, leaving only the high-frequency (residual) component to fit to the (stationary) model.

That is, spatial prediction is closely analogous to linear filtering, and optimal spatial prediction of data involves designing a filter which best represents the continuity properties of the underlying r.p.. Stein (1999, ch. 3) provides theoretical justification for this analogy, which relies on the

assumption of wide-sense stationarity (of the process) and interpolation (rather than extrapolation) being the analyst's purpose for predicting from data.

Although geostatistics is historically linked to the earth sciences, spatial analysis is relevant to many other fields such as ecology, public health, computational geometry, image processing, control theory, sensor networks, machine learning, and complex systems. The more inclusive term *spatial statistics* reflects this broader scope. Under this rubric, Cressie (1993) identifies three subdomains for formalizing spatial data analysis: (1) geostatistics for continuous space r.p.s, based largely on the development from Matheron's school and incorporating classical statistical formulations; (2) data on a fixed lattice, using Markov random fields and the Hammersley-Clifford theorem to express the relative effects of connections between nodes; and (3) spatial point processes, in which the positions of data are randomized. Moore (2001) provides several examples. Gelfand et al. (2010) provide an up-to-date review. This dissertation is based on the broader sense of geostatistics as defined by Cressie, analysis of continuous-space processes.

2.2.2 Simple kriging

To illustrate, we derive the "simple kriging" predictor. We adopt the notation of Cressie and Wikle (2011) and describe kriging in general terms. In later chapters, we apply similar techniques to atmospheric/aeronomic data. We wish to predict the value of an unobserved random variable $Y(\cdot)$ at location $\mathbf{s}_0$ based on observations in the region $D_s \subset \mathbb{R}^d$.

Let $Y(\cdot) \triangleq \{Y(\mathbf{s}) \mid \mathbf{s} \in D_s\}$ be a zero-mean[2], second-order stationary r.p. with (known) covariance function $C_Y(\mathbf{u},\mathbf{v}) = \mathrm{Cov}(Y(\mathbf{u}), Y(\mathbf{v})), \forall \mathbf{u},\ \mathbf{v} \in D_s$. Let us also assume an additive noise model to represent the *observations* (with measurement error) at locations $\{\mathbf{s}_i \mid i = 1,\ldots,m\}$

$$Z_i = Z(\mathbf{s}_i) = Y(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \tag{2.4}$$

where $\epsilon(\cdot) \perp Y(\cdot)$ and $\epsilon(\cdot) \triangleq \{\epsilon(\mathbf{s}_i) \mid \mathbf{s}_i \in D_s\}$ is a zero-mean white noise process with finite variance $\sigma^2_\epsilon > 0$. We wish to predict $Y(\mathbf{s}_0)$, based on observations $\underline{Z} = [Z(\mathbf{s}_1),\ldots,Z(\mathbf{s}_m)]^\mathsf{T}$.

From these (noise-corrupted) measurements, we now derive the predictor of $Y(\mathbf{s}_0)$ that is optimal in the mean-square sense, i.e. that minimizes the MSPE given by (2.3).[3] We restrict ourselves to

[2] Equivalently, the mean process $\mu_Y(\mathbf{s})$ is known for all $\mathbf{s}$.

[3] An alternative, geometrical derivation is given by Zimmerman and Stein (2010).

the class of predictors that are affine functions of the data $\underline{Z}$:

$$\widehat{Y}(\mathbf{s}_0;\underline{\lambda},\kappa) = \sum_{i=1}^{m} \lambda_i Z(\mathbf{s}_i) + \kappa = \underline{\lambda}^{\mathsf{T}}\underline{Z} + \kappa, \tag{2.5}$$

That is, $\widehat{Y}$ is a weighted sum of the data, where the weights $\underline{\lambda} \in \mathbb{R}^m$ are determined based on the regularity conditions imposed on the process (e.g. in this case, exactly specified mean and covariance functions)[4], and $\kappa \in \mathbb{R}$ can be viewed as a Lagrange multiplier limiting the "size" of the solution

Combining (2.3) and (2.5), the (mean-square) optimal predictor then satisfies the optimality condition

$$\begin{aligned}
\underline{\lambda}^*, \kappa^* &= \underset{\underline{\lambda},\kappa}{\operatorname{argmin}}\ \textsc{mspe}\,(\underline{\lambda},\kappa) \\
&= \underset{\underline{\lambda},\kappa}{\operatorname{argmin}}\ \mathrm{E}\left[\left(Y(\mathbf{s}_0) - \left(\underline{\lambda}^{\mathsf{T}}\underline{Z} + \kappa\right)\right)^2\right] \\
&= \underset{\underline{\lambda},\kappa}{\operatorname{argmin}}\ \mathrm{Var}\left[Y(\mathbf{s}_0) - \left(\underline{\lambda}^{\mathsf{T}}\underline{Z} + \kappa\right)\right] + \left\{\mathrm{E}\left[Y(\mathbf{s}_0) - \left(\underline{\lambda}^{\mathsf{T}}\underline{Z} + \kappa\right)\right]\right\}^2 \\
&= \underset{\underline{\lambda},\kappa}{\operatorname{argmin}}\ \mathrm{Var}\left[Y(\mathbf{s}_0) - \underline{\lambda}^{\mathsf{T}}\underline{Z}\right] \quad + \quad \left\{\mathrm{E}\left[Y(\mathbf{s}_0)\right] - \mathrm{E}\left[\underline{\lambda}^{\mathsf{T}}\underline{Z}\right] - \kappa\right\}^2
\end{aligned} \tag{2.6}$$

(The final equality is because the variance term is invariant to the scalar shift κ.) The second term, the bias term, is minimized (is indeed exactly zero) by selecting $\kappa = \mathrm{E}[Y(\mathbf{s}_0)] - \underline{\lambda}^{\mathsf{T}}\mathrm{E}[\underline{Z}]$. And since the measurement error is zero-mean, $\mathrm{E}[\underline{Z}] = \mathrm{E}[\underline{Y}] = \underline{\mu}_Y = (\mu_Y(\mathbf{s}_1),\ldots,\mu_Y(\mathbf{s}_m))^{\mathsf{T}}$, so that the required $\kappa = \mu_Y(\mathbf{s}_0) - \underline{\lambda}^{\mathsf{T}}\underline{\mu}_Y$.

With that, (2.6) becomes

$$\begin{aligned}
\underline{\lambda}^* &= \underset{\underline{\lambda}}{\operatorname{argmin}}\ \mathrm{Var}\left[Y(\mathbf{s}_0) - \left(\underline{\lambda}^{\mathsf{T}}\underline{Z}\right)\right] \\
&= \underset{\underline{\lambda}}{\operatorname{argmin}}\ C_Y(\mathbf{s}_0,\mathbf{s}_0) - 2\sum_i \lambda_i \mathrm{Cov}\left(Y(\mathbf{s}_0), Y(\mathbf{s}_i)\right) + \sum_i \sum_j \lambda_i \lambda_j (\mathbf{C}_Z)_{ij},
\end{aligned} \tag{2.7}$$

where $C_Y(\mathbf{s}_0,\mathbf{s}_0)$ is the *process variance*, the matrix $(\mathbf{C}_Z)_{ij}$ is given by the *data covariance function*,

$$\begin{aligned}
\mathbf{C}_Z(\mathbf{s}_i,\mathbf{s}_j) &\triangleq \mathrm{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) \\
&= \begin{cases} C_Y(\mathbf{s}_i,\mathbf{s}_j) + \sigma_\epsilon^2 & \mathbf{s}_i = \mathbf{s}_j \\ C_Y(\mathbf{s}_i,\mathbf{s}_j) & \mathbf{s}_j \neq \mathbf{s}_j, \end{cases}
\end{aligned}$$

[4] Stationarity, though required for the simple kriging predictor, is not a formal requirement for kriging in general.

and, from the data model (2.4), the middle term reduces to

$$2\sum_i \lambda_i \operatorname{Cov}(Y(\mathbf{s}_0), Y(\mathbf{s}_i)) = 2\underline{\lambda}^{\mathsf{T}}\underline{c}_Y(\mathbf{s}_0),$$

where $\underline{c}_Y(\mathbf{s}_0) = (C_Y(\mathbf{s}_0,\mathbf{s}_1),\ldots,C_Y(\mathbf{s}_0,\mathbf{s}_m))^{\mathsf{T}}$.

Because expression (2.7) is quadratic in $\underline{\lambda}$, and $\mathbf{C}_Z$ is positive definite, the unique minimum satisfies $\mathbf{C}_Z\underline{\lambda} = \underline{c}_Y(\mathbf{s}_0)$. Denote the solution to that system $\underline{\lambda}^* = \mathbf{C}_Z^{-1}\underline{c}_Y(\mathbf{s}_0)$. Then the constant scalar term is $\kappa^* = \mu_Y(\mathbf{s}_0) - \underline{c}_Y^{\mathsf{T}}(\mathbf{s}_0)\mathbf{C}_Z^{-1}\underline{\mu}_Y$. Substituting these into (2.3), the *simple kriging predictor* at point $\mathbf{s}_0$ from data at points $\{\mathbf{s}_i \mid i = 1,\ldots,m\}$ is

$$\boxed{\widehat{Y}_{\text{sk}}(\mathbf{s}_0) = \underline{c}_Y^{\mathsf{T}}(\mathbf{s}_0)\mathbf{C}_Z^{-1}\left[\underline{Z} - \underline{\mu}_Y\right] + \mu_Y(\mathbf{s}_0).} \tag{2.8}$$

The minimized MSPE, called the *simple kriging variance*, by substitution of $\underline{\lambda}^*$ and κ^* into (2.7), is

$$\boxed{\widehat{\sigma}^2_{\text{sk}}(\mathbf{s}_0) \triangleq C_Y(0) - \underline{c}_Y(\mathbf{s}_0)^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{c}_Y(\mathbf{s}_0).} \tag{2.9}$$

2.2.3 Some properties of the simple kriging predictor

Exact interpolator. In the absence of measurement error (i.e. $Z(\mathbf{s}_i) = Y(\mathbf{s}_i)$), $\widehat{Y}_{\text{sk}}(\mathbf{s})$ is an exact interpolator; it "honors the data" at their sampled positions $\{\mathbf{s}_i\}$. Consequently, at these same positions, the kriging variance is zero, indicating absolute certainty (since the measurement was without error). If the data model includes measurement error, the kriging variance is bounded below by σ_ϵ^2, the measurement noise.

Best linear unbiased predictor. $\widehat{Y}_{\text{sk}}(\cdot)$ is unbiased, since $\mathrm{E}\left[\widehat{Y}_{\text{sk}}(\cdot)\right] = \mu_Y(\cdot) = \mathrm{E}[Y(\cdot)]$. Kriging belongs to the class of *best linear unbiased predictors* (BLUPs). The Kalman filter is also a BLUP, as (2.4), (2.8), and (2.9) suggest. For the Kalman filter, $Y(\cdot)$ is assumed to be a first-order autoregressive process.

Kriging variance is data-independent. Both $\widehat{Y}_{\text{sk}}(\cdot)$ and $\widehat{\sigma}^2_{\text{sk}}(\cdot)$ can be evaluated for any $\mathbf{s} \in D_s$. But note from (2.9) that $\widehat{\sigma}^2_{\text{sk}}$ does not depend on the data $\underline{Z}$. It is completely determined by the geometry of the problem (position and dispersion of the sample points, and clustering if sampling is nonuniform) and the regularity/continuity constraints on the process implicit in the covariance

function $C_Y(\cdot)$.

Map of uncertainty. It is tempting to interpret $\widehat{\sigma}^2_{\text{sk}}$ as a descriptor of local roughness of the data (Papritz and Stein, 2002). But since $\widehat{\sigma}^2_{\text{sk}}$ does not depend on the data, two independent realizations of $Y(\mathbf{s})$, sampled at the same set of points, will share identical maps of kriging variance, though their predicted values may differ dramatically.[5] Rather, $\widehat{\sigma}^2_{\text{sk}}(\cdot) = \mathrm{E}\left[(Y(\cdot) - \widehat{Y}(\cdot\,; \underline{\lambda}^*, \kappa^*))^2\right]$ is an ensemble average over all possible realizations $Y(\cdot)$. It reflects the "density of information" around each prediction point $\mathbf{s}_0$ provided by the samples; i.e. the availability of information and the relative importance of data sampled at a given set of positions (Wackernagel, 2003). The kriging variance $\widehat{\sigma}^2_{\text{sk}}(\cdot)$ should always be displayed alongside the corresponding prediction $\widehat{Y}_{\text{sk}}(\cdot)$.

Kriging variance for sample design. Both because $\widehat{\sigma}^2_{\text{sk}}$ is both a spatial map of uncertainty, and because it is independent of particular data, it is often used for sample design. That is, given a statistical characterization of a site and the physics of phenomena expected to be studied there, kriging variance is a tool for assessing the (expected) quality of an experiment. Given the scale and dynamics of a nonstationary process, what is the optimal sample design? How many sensors are needed to achieve a level of precision?

LLSE and MLE of a Gaussian process. The form of (2.8) is a familiar result from estimation theory. The *linear least squares estimator* (LLSE), or equivalently, *maximum likelihood estimator* (MLE) of $Y(\mathbf{s}_0)$ if $Y(\cdot)$ is a Gaussian process and each $\epsilon(\mathbf{s}_i)$ is an *independent and identically distributed* (i.i.d.) Gaussian random variable.

Recall the assumptions made in order to use simple kriging: (1) the mean process is exactly specified, and (2) the process $Y(\cdot)$ is stationary. These are rather strong conditions (particularly (1)) and may not apply to natural processes. In the next section, we briefly discuss some extensions to this method and their properties.

2.3 Other kriging predictors

As the name suggests, simple kriging is a (rather limited) member of a family of kriging predictors. Detailed derivations of these predictors can be found in most geostatistics texts (e.g. Cressie (1993); Chilès and Delfiner (2012)).

[5] Chilès and Delfiner (see 2012, p. 178) for a thorough discussion of this phenomenon.

2.3.1 Kriging with unknown mean

Kriging predictors come in a variety of "flavors," each corresponding to a different set of assumptions. For instance, simple kriging assumes the process $Y(\cdot)$ is at least wide-sense stationary with zero mean. (Equivalently, $\mu(\cdot)$ is known exactly.) If $\mu_Y(\cdot)$ is *not* known, it can be modeled as a linear combination of predictive variables (covariates).

Ordinary and universal kriging (see below) assume a *mixed effects model* for the process $Y(\cdot)$:

$$Y(\mathbf{s}) = \underline{x}(\mathbf{s})^\mathsf{T}\underline{\beta} + \delta(\mathbf{s}), \qquad \mathbf{s} \in D_s, \tag{2.10}$$

where $\underline{x}(\mathbf{s}) \triangleq \left(x_1(\mathbf{s}),...,x_p(\mathbf{s})\right)^\mathsf{T}$ is a vector of covariates (e.g. spatial basis functions, or other explanatory variables, such as elevation or temperature, which are known for many positions, and which can justifiably be included in a linear model for the mean process $\mathrm{E}Y(\mathbf{s}) = \underline{x}(\mathbf{s})^\mathsf{T}\underline{\beta}$ (see Zimmerman and Stein, 2010, p. 32)), $\underline{\beta} \triangleq \left(\beta_1,...,\beta_p\right)^\mathsf{T}$ is a vector of unknown *fixed effect* parameters, and $\delta(\mathbf{s})$ is the *random effect*, a zero-mean random process with covariance function $C_Y(\mathbf{u},\mathbf{v})$.

The vector $\underline{\beta}$ can be seen as an unknown trend parameter and fit to the data, for instance by *ordinary least squares* (OLS): $\widehat{\underline{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\underline{Z}$, where $\mathbf{X} = (\underline{x}(\mathbf{s}_1),\underline{x}(\mathbf{s}_2),\ldots,\underline{x}(\mathbf{s}_m))^\mathsf{T}$ is an $m \times p$ matrix of covariates at $\mathbf{s}$. The delineation between fixed effect and random effect is ambiguous. The random effect is usually interpreted as covering small-scale variation while the fixed effect represents large-scale trends.

Ordinary kriging

Suppose the mean is an unknown constant μ. The ordinary kriging predictor (OK) is the BLUP that minimizes mean-square prediction error (2.3). Its derivation is very similar to simple kriging (SK), except that the unbiased constraint must be explicitly included. That is,

$$\mathrm{E}\left[\widehat{Y}_{\mathrm{ok}} - Y\right] = \mathrm{E}\left[\underline{\lambda}^\mathsf{T}\underline{Z} - \mu_Y\right] = \sum_{i=1}^{m}\lambda_i\mu - \mu = 0$$

$$\sum_{i=1}^{m}\lambda_i = 1.$$

The *ordinary kriging predictor* is given by

$$\widehat{Y}_{\mathrm{ok}}(\mathbf{s}_0) = \widehat{\mu}_{\mathrm{GLS}} + \underline{c}_Y(\mathbf{s}_0)^\mathsf{T}\mathbf{C}_Z^{-1}\left(\underline{Z} - \widehat{\mu}_{\mathrm{GLS}}\underline{1}\right), \tag{2.11}$$

where $\widehat{\mu}_{\text{GLS}} = (\underline{1}^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{Z})/(\underline{1}^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{1})$ is the *generalized least squares* (GLS) estimator of μ, and $\underline{1}$ is a vector of ones. The associated *ordinary kriging variance*

$$\sigma^2_{\text{ok}}(\mathbf{s}_0) = C_Y(\mathbf{s}_0,\mathbf{s}_0) - \underline{c}_Y(\mathbf{s}_0)^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{c}_Y(\mathbf{s}_0) + \left(1 - \underline{1}^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{c}_Y(\mathbf{s}_0)\right)^2/(\underline{1}^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{1}) \tag{2.12}$$

has an additional term compared to (2.9), reflecting additional uncertainty after estimating the mean.

Universal kriging

Universal kriging generalizes both simple and ordinary kriging. In terms of the mixed effect model (2.10), it fits a higher-order trend to the data. In general, the *universal kriging predictor* is

$$\widehat{Y}_{\text{uk}}(\mathbf{s}_0) = \underline{x}(\mathbf{s}_0)^{\mathsf{T}}\hat{\underline{\beta}}_{\text{GLS}} + \underline{c}_Y(\mathbf{s}_0)^{\mathsf{T}}\mathbf{C}_Z^{-1}\left(\underline{Z} - \mathbf{X}\hat{\underline{\beta}}_{\text{GLS}}\right), \tag{2.13}$$

where $\widehat{\underline{\beta}}_{\text{GLS}} \triangleq \left(\mathbf{X}^{\mathsf{T}}\mathbf{C}_Z^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{Z}$ is the generalized least-squares estimator of $\underline{\beta}$, $\underline{x}(\mathbf{s})$ is a $p \times 1$ vector of covariates for position $\mathbf{s}$, and $\mathbf{X} = (\underline{x}(\mathbf{s}_1),\ldots,\underline{x}(\mathbf{s}_m))^{\mathsf{T}}$ is an $m \times p$ matrix of covariates at the data positions. For instance, using coordinates $\{\mathbf{s}_i\}$ of the data, or polynomials of those coordinates. If $\mathbf{s}_i = (\mathsf{x}_i, \mathsf{y}_i)$,

$$\mathbf{X} = \begin{bmatrix} 1 & \mathsf{x}_1 & \mathsf{y}_1 & \mathsf{x}_1^2 & \mathsf{x}_1\mathsf{y}_1 & \mathsf{y}_1^2 \\ 1 & \mathsf{x}_2 & \mathsf{y}_2 & \mathsf{x}_2^2 & \mathsf{x}_2\mathsf{y}_2 & \mathsf{y}_2^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \mathsf{x}_m & \mathsf{y}_m & \mathsf{x}_m^2 & \mathsf{x}_m\mathsf{y}_m & \mathsf{y}_m^2 \end{bmatrix},$$

and $p = 6$. The associated *universal kriging variance* is

$$\begin{aligned}\sigma^2_{\text{uk}}(\mathbf{s}_0) = {} & C_Y(\mathbf{s}_0,\mathbf{s}_0) - \underline{c}_Y(\mathbf{s}_0)^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{c}_Y(\mathbf{s}_0) \\ & + \left(\underline{x}(\mathbf{s}_0) - \mathbf{X}^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{c}_Y(\mathbf{s}_0)\right)^{\mathsf{T}}\left(\mathbf{X}^{\mathsf{T}}\mathbf{C}_Z^{-1}\mathbf{X}\right)^{-1}\left(\underline{x}(\mathbf{s}_0) - \mathbf{X}^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{c}_Y(\mathbf{s}_0)\right).\end{aligned} \tag{2.14}$$

Again, the final term reflects additional uncertainty due to having estimated the trend parameters from the same data.

2.4 Geostatistical Model Selection and Parameter Estimation

Classical geostatistical modeling involves generating summary statistics of the sample to assess spatial uncertainty. This is commonly expressed in the form of a structure function, or variogram, a function describing the decorrelation of the process $Y(\cdot)$ with distance. This is often an exploratory process, in which the analyst makes decisions regarding shape, scale, and complexity that affect the

predictive power of the model, perhaps also bringing a priori considerations to bear in interpreting the model.

2.4.1 Semivariogram

The spatial dependence of a *wide-sense stationary* (WSS) random process $Y(\cdot)$ is summarized by a (constant) mean function $\mathrm{E}[Y(\mathbf{s})] = \mu_Y \ \forall \mathbf{s} \in D_s$ and a covariance function which depends only on lag $\mathbf{h}$

$$C_Y(\mathbf{h}) = \mathrm{Cov}\left(Y(\mathbf{s}+\mathbf{h}),\ Y(\mathbf{s})\right) \qquad \forall \mathbf{s},\ \mathbf{s}+\mathbf{h} \in D_s,$$

which typically characterizes the distance decay expressed in the First Law of Geography. A large class of spatial processes (at least approximately) satisfy these properties. Additionally, a constant-mean r.p. is said to be *intrinsically stationary* if its *increments* are WSS, that is, the difference of $\mathbf{h}$-displaced variables varies in a way that depends only on $\mathbf{h}$:

$$\boxed{\mathrm{Var}\left(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})\right) \triangleq 2\gamma_Y(\mathbf{h}),} \qquad \forall\ \mathbf{s},\ \mathbf{s}+\mathbf{h} \in D_s. \tag{2.15}$$

The quantity $2\gamma_Y$ is called the *variogram* of Y; (γ_Y is the *semivariogram*). If $C_Y(\mathbf{h})$ describes the distance decay of correlation, $2\gamma_Y(\mathbf{h})$ typically embodies the *de*correlation of Y with itself as a function of distance between points. Compare the two panels of Figure 2·1. Both plots convey the same information about some r.p. $Y(\cdot)$. But the semivariogram increases to a maximum decorrelation as distance h increases, while the covariance function diminishes to zero as $h \to \infty$. (Note that, by definition, $\gamma(\mathbf{0}) \equiv 0$.)

The set of intrinsically stationary processes can be shown to include the set of WSS processes (e.g. Cressie and Wikle, 2011, p. 130). Indeed, any process with a stationary covariance function also has a stationary semivariogram through the identity

$$\gamma_Y(\mathbf{h}) = C_Y(\mathbf{0}) - C_Y(\mathbf{h}), \qquad \forall\ \mathbf{h} \in \mathbb{R}^d. \tag{2.16}$$

The reverse is not true. Many intrinsically stationary processes have no WSS counterpart.[6] The semivariogram is therefore the more general second-order moment.

[6] For example, the Wiener process $\{W(s) : s \geq 0\}$ has variogram $2\gamma(h) = -|h|$, but $\mathrm{Cov}(W(s), W(u)) = \min(s, u)$, which is not a function of $|s-u|$ (Cressie and Wikle, 2011).

Nonstationary processes

So far, we have mostly considered stationary processes. Such models have the advantage that data can be reasonably gathered from throughout the domain D_s and combined. This is important for parameter fitting, but it is not a requirement for analysis or prediction. Kriging can also be performed with nonstationary processes. The nonstationary forms of of the covariance function and variogram are

$$C_Y(\mathbf{u},\mathbf{v}) \triangleq \mathrm{Cov}(Y(\mathbf{u}, Y(\mathbf{v})), \qquad \mathbf{u},\ \mathbf{v} \in D_s$$

and

$$2\gamma_Y(\mathbf{u},\mathbf{v}) \triangleq \mathrm{Var}(Y(\mathbf{u}) - Y(\mathbf{v})), \qquad \mathbf{u},\ \mathbf{v} \in D_s.$$

A more general form of identity (2.16) also exists:

$$2\gamma_Y(\mathbf{u},\mathbf{v}) = C_Y(\mathbf{u},\mathbf{u}) + C_Y(\mathbf{v},\mathbf{v}) - 2C_Y(\mathbf{u},\mathbf{v}), \qquad \mathbf{u},\ \mathbf{v} \in D_s.$$

Isotropy

Isotropy is another simplifying assumption that hardly occurs in the real world. Isotropic processes are invariant to rotation about the origin. An isotropic semivariogram or covariance function can be expressed as $2\gamma(h)$ or $C_Y(h)$ where $h = \|\mathbf{h}\|$. Many anisotropic processes can be accommodated by transforming the geometry so that

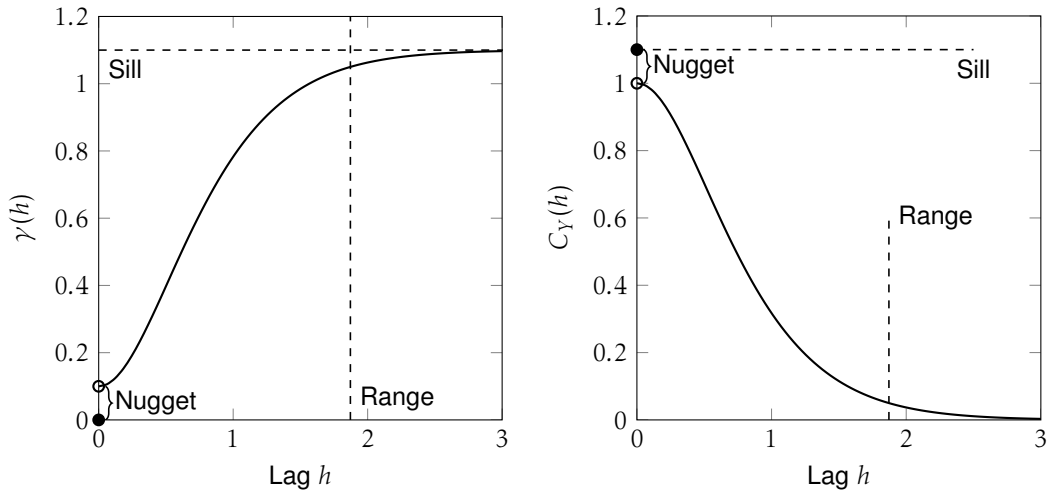
$$\gamma(\mathbf{s}+\mathbf{h},\mathbf{s}) = \gamma_{\mathrm{iso}}(\|\mathbf{A}\mathbf{h}\|),$$

where $\mathbf{A}$ is a transformation matrix. Such a process is said to be *geometrically anisotropic*. (See also Chilès and Delfiner, 2012, pp.98–99)

2.4.2 Fitting variogram parameters

Kriging is widely used for predicting natural processes. Its performance depends on the quality of the process model. Simple kriging is the optimal predictor when the mean and covariance are perfectly known. Other kriging predictors relax this requirement, permitting drift/trend regression, lending flexibility to the solution at the expense of increased uncertainty. Still, an appropriate model is needed.

In classical geostatistics, modeling is carried out for a given data set through a combination of exploratory data analysis and automatic parameter estimation, collectively called *variography*.



(a) Semivariogram

(b) Covariance function

Figure 2·1: Stationary semivariogram and covariance function with canonical geostatistical parameters labeled. Both functions describe how a random process decorrelates with distance. The parameter names reflect their provenance in the mining literature.

The process may involve a human modeler iterating through the following stages:

1. Identify the locations of all data.
2. Compute all pairwise square-differences $\big(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\big)^2$ and plot a cloud of points versus distance $\|\mathbf{s}_i - \mathbf{s}_j\|$.
3. Determine appropriate binning and compute a method-of-moments estimator, e.g.

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} \big(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\big)^2, \qquad \mathbf{h} \in \mathbb{R}^d, \tag{2.17}$$

   where $N(h)$ is a bin near $\mathbf{h}$ and $|N(\mathbf{h})|$ is the number of elements in the bin.
4. Fit a variogram model.

The modeler selects a variogram model to fit. The selection may be based on the shape of the point cloud in stage 3 above, or it may be motivated by the physics of the underlying process, or it may be especially favored by the modeler. Variography thus involves a combination of objective and subjective justification.

Spatial statistics inherits classes of variogram models from classical geostatistics, with parameters relating to the shape, size, and magnitude of a random field. Some generic parameters, found in most models, are described below. (See Figure 2·1 for an illustration.)

Sill

Total process variance. The *sill* is the maximum level of decorrelation between any two points $Y(\mathbf{s}_1)$ and $Y(\mathbf{s}_2)$. This is the asymptote in Figure 2·1(a). That is, the sill is $\lim_{h\to\infty} \gamma_Y(h)$ or $C_Y(0)$.

The variogram of an intrinsically stationary process need not be finite. In that case, identity (2.16) indicates that such a process cannot be described by a covariance function. Every WSS process has a bounded variogram, i.e. a finite process variance.

Range

The variogram describes the rate of decorrelation with distance. If $Y(\cdot)$ has a finite variance (sill), this rate eventually levels off so that $\mathrm{Var}[Y(\mathbf{s}_2) - Y(\mathbf{s}_1)] = \mathrm{Var}[Y(\mathbf{s}_3) - Y(\mathbf{s}_1)]$ if $\|(\mathbf{s}_2 - \mathbf{s}_1)\| > a$ and $\|(\mathbf{s}_3 - \mathbf{s}_1)\| > a$. In other words, there is some distance a, the *range*, beyond which samples of $Y(\cdot)$ are maximally decorrelated.

If the variogram reaches its maximum exactly, the range a is the distance at which this happens. If it approaches the maximum asymptotically, then a is defined as the point at which the variogram reaches some fraction of the sill, typically 95 %.

Nugget effect

While sill and range quantify long-range behavior, the *nugget effect* describes microscale variability (i.e. smaller than the shortest intersample distance). The name reflects its origin in mining, since the discovery of a "nugget" (an atypical sample) is not predicted by any neighboring measurements.

By definition, $\gamma_Y(0) = 0$. So if a nugget is present, it represents a discontinuity at $h = 0$. The nugget effect is practically indistinguishable from white measurement noise, since both produce measurements with no detectable small-scale correlation structure.

Differentiability

Stein (1999) demonstrates the connection of short-lag behavior of $\gamma_Y(h)$ to mean-square differentiability of $Y(\mathbf{s})$. He argues that, among all variogram properties, optimal interpolation is most sensitive to the behavior near the origin. While most geostatistical variogram models have this property fixed implicitly a priori, the Matérn model includes a parameter (ν) that explicitly affects mean-square differentiability. For this reason, Stein (1999) also advocates using the Matérn model exclusively, allowing the data to influence the smoothness of the prediction during the structural

analysis phase.

2.4.3 Which function should be fitted?

In principle, a method-of-moments covarinace function estimator $\hat{C}(\mathbf{h})$ could be constructed and fit in much the same way as $2\hat{\gamma}(h)$ in (2.17). However, Cressie (1993, pp. 70–73) shows that the variogram estimator is (1) unbiased when μ_Y is constant, and (2) less biased than $\hat{C}(\mathbf{h})$ when $Y(\cdot)$ posesses a trend. The crux of the covariance estimator is the necessity to first "plug in" an empirical mean. The variogram estimator has no such requirement.

2.4.4 Sensitivity of kriging to semivariogram misspecification

Assume that $Y(\cdot)$ is stationary, and its semivariogram is known exactly. Then (2.8), (2.11), and (2.13) are unbiased, minimum-MSPE predictors.

In practice, the spatial structure is not known exactly, but is estimated from (usually imperfect) observations. What effect does misspecification of the semivariogram have on the kriging predictions?

Cressie (1993, pp. 289–299) discusses this at some length. He concludes (1) that estimates of the above parameters are biased, and (2) that the estimated semivariogram is more stable than the estimated covariance function. Furthermore, (3) the kriging predictor is stable under misspecification, but (4) the kriging variance suffers more dramatically, especially when biased parameter estimates are used to compute it. He recommends various robust methods for both parameter estimation and prediction.

Chilès and Delfiner (2012, pp.176–177) also acknowledge substituting an estimated variogram (assumed known without error) into (2.12) or (2.14) fails to account for the total error. Diggle and Ribeiro Jr. (2007) motivate their Bayesian prediction by examining the sub-optimal performance of such "plug-in" predictors.

Stein (1999), emphasizing the fact that short-range behavior has the most dramatic impact on the predictor, recommends the Matèrn model exclusively, arguing that the flexibility provided by the smoothness parameter ν likely conveys as much benefit as either multi-model trial-and-error or sticking to a few pet models, especially if ν is estimated from the data. The author also shows that, at least in the case of interpolation, as data become more closely-spaced, the prediction depends less on the particular choice of semivariogram.

The consensus is apparent. At least in the case of interpolation, kriging is fairly robust to the

choice of predictive semivariogram. As long as the data can support it, even fairly large departures from the truth may have little effect on the quality of the prediction. The corresponding uncertainty estimate, on the other hand, is likely to be overly optimistic.

2.5 Simple Kriging and Conditional Simulation

The simple kriging predictor (2.8) is a weighted sum, a predictor at unobserved locations conditioned on observations $\underline{Z}$. Since it is a linear combination of data, it tends to exhibit less spatial variability than a typical realization of $Y(\cdot)$. Rather, $\widehat{Y}_{\text{sk}}$ summarizes the behavior of random processes matching the first- and second-order distributional properties of $Y(\cdot)$.

The kriging variance (2.9) is a non-conditional statistic (not a function of the data) and represents the minimized MSPE of $\widehat{Y}_{\text{sk}}$.

Smoothness and differentiability

For stationary processes, the behavior of the semivariogram near the origin is affects the high-frequency part of the spectrum (power spectral density). Naturally then, the element of semivariogram structure that most directly influences the smoothness or roughness of the random process is that near the origin. The Matérn function takes four parameters $\underline{\theta} = (\sigma_0^2, \phi, a, \nu)$, the first three correspond to the nugget, sill, and range (respectively); the fourth is a "smoothness" parameter. A process with this type of covariance functions is $\lfloor \nu - 1/2 \rfloor$-times differentiable (in the mean-square sense).

Mean-square differentiability does not guarantee smooth realizations, though. The nugget effect parameter σ_0^2 also influences short-lag behavior of the semivariogram, namely introducing a discontinuity at the origin (Papritz and Stein, 2002). Thus, even very closely-spaced values $y(\mathbf{s})$ and $y(\mathbf{s} + \delta\mathbf{s})$ are not guaranteed to be strongly correlated, resulting in "rough" realizations $y(\cdot)$.

As for the kriging predictor, the short-lag behavior of the semivariogram determines the smoothness of the predicted surface, particularly at the data sites. Three cases for the behavior of the semivariogram near the origin:

Discontinuous $\widehat{Y}(\cdot)$ is discontinuous at the data points.

Linear $\widehat{Y}(\cdot)$ is continuous everywhere but not everywhere differentiable.

Parabolic $\widehat{Y}(\cdot)$ is both continuous and differentiable everywere.

It is often a goal of spatial prediction to "smooth" the measured surface. Some problems, though, call for a process that reflects the spatial uncertainty of the process itself. In that case, it is possible to simulate a realization of $Y(\cdot)$ that passes through the data points.

2.5.1 Conditional simulation

In the absense of noise, the kriging predictor is an exact interpolator: the predicted function passes through all measured points. But the predicted function does not represent a realization of the random process. In image processing terms, it lacks the texture of a realization of $Y(\cdot)$. In some situations (for instance, estimating the length of a curve on the predicted surface), it is better to generate one or more *conditional simulations*, realizations of a random process that both a) exhibit the statistical properties assumed by the predictor and b) honor the data.

To generate such a simulation from known data and a known semivariogram, begin with the kriging predictor,

$$\widehat{y}_{\text{sk}}(\mathbf{s};\underline{Z}) = \mu_Y(\mathbf{s}) + \underline{c}_Y(\mathbf{s})^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{Z}. \tag{2.18}$$

Let us regard the data $\underline{Z}$ as a random vector. Then (2.18) is a random function, so for now let us use the upper-case convention and drop the hat indicating a predictor. Now consider the following decomposition:

$$Y(\mathbf{s}) \quad = \quad \underbrace{Y_{sk}(\mathbf{s};\underline{Z})}_{\text{kriging prediction}} \quad + \quad \underbrace{Y(\mathbf{s}) - Y_{sk}(\mathbf{s};\underline{Z})}_{\text{kriging residual}} \qquad \mathbf{s} \in D_s. \tag{2.19}$$

The components on the right-hand side of (2.19) are two independent *Gaussian processes* ($\mathcal{GP}$s). Once the data are specified to be $\underline{Z} = \underline{z}$, the kriging prediction is known and given by (2.8). The kriging residual is unknown because $Y(\mathbf{s})$ is unknown. But it can be simulated since $\mu_Y(\mathbf{s})$ and $C_Y(\mathbf{h})$ are known. Consider a similar decomposition:

$$Y^{us}(\mathbf{s}) = \widehat{Y}^{us}_{\text{sk}}(\mathbf{s}) \quad + \quad Y^{us}(\mathbf{s}) - \left(\mu_Y(\mathbf{s}) + \underline{c}_Y(\mathbf{s})^{\mathsf{T}}\mathbf{C}_Y^{-1} \begin{bmatrix} Y^{us}(\mathbf{s}_1) \\ \vdots \\ Y^{us}(\mathbf{s}_m) \end{bmatrix} \right), \tag{2.20}$$

where $Y^{us}(\cdot)$ is an *unconditional* simulation of $Y(\mathbf{s})$, which is simulated both at $\mathbf{s}$ and at the sample points $\{\mathbf{s}_i,\ i = 1,\ldots,m\}$. Then, after generating a realization of $y^{us}(\mathbf{s})$ and computing $\widehat{y}^{us}_{\text{sk}}(\mathbf{s})$, we can simulate the kriging residual and substitute it in (2.19):

$$y^{cs}(\mathbf{s};\underline{z}) \quad = \quad \widehat{y}_{sk}(\mathbf{s};\underline{z}) + y^{us}(\mathbf{s}) - \widehat{y}^{us}_{\text{sk}}(\mathbf{s}) \tag{2.21}$$

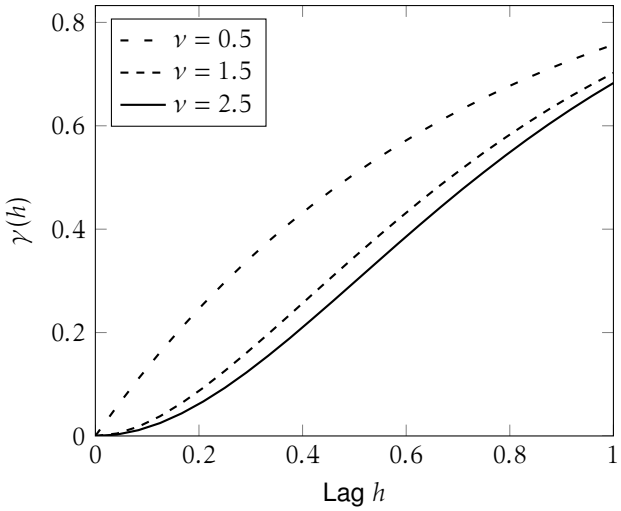


Figure 2·2: Behavior near the origin of the Matérn semivariogram for different values of ν. A processes with any of these covariance functions has mean-square differentiability $\lfloor \nu - 1/2 \rfloor$.

Note that only the first term, the kriging predictor, actually depends on the data $\underline{z}$. The procedure above generates a conditional simulation. Since $\widehat{Y}_{\text{sk}}$ is an exact interpolator, at any sample point $\mathbf{s}_\alpha$ we have $\widehat{y}^{us}_{\text{sk}}(\mathbf{s}_\alpha) = y^{us}(\mathbf{s}_\alpha)$. Also, $\widehat{y}_{sk}(\mathbf{s}_\alpha; \underline{z}) + y^{us}(\mathbf{s}) = y(\mathbf{s}_\alpha$, the actual measurement at $\mathbf{s}_\alpha$. Therefore, $y^{cs}(\mathbf{s}_\alpha; \underline{z}) = y(\mathbf{s}_\alpha$.

2.5.2 Exploration of variogram parameters on kriging prediction and simulations

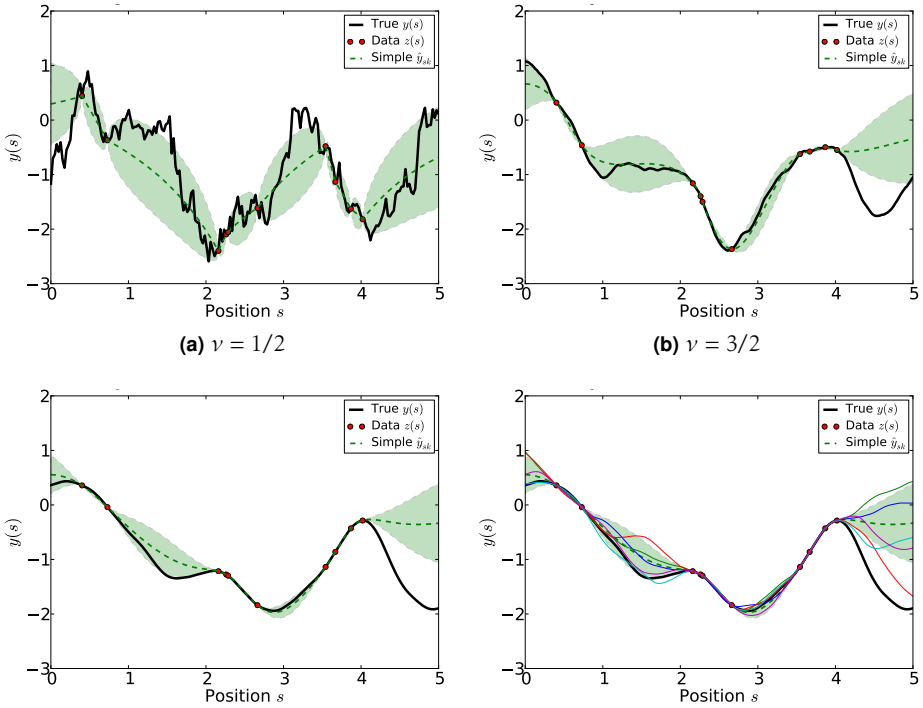
The following examples begin with a known variogram, the Matérn model:

$$\gamma(h) = \sigma_0^2 I(h \neq 0) + (\sigma_\eta^2 - \sigma_0^2)\left(1 - \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{2\sqrt{\nu}h}{a}\right)^{\nu} K_\nu\left(\frac{2\sqrt{\nu}h}{a}\right)\right), \tag{2.22}$$

where $\Gamma(\cdot)$ is the Gamma function and $K_\nu(\cdot)$ is a modified Bessel function of the first kind of order ν. In the following examples, the Matérn model (2.22) is used to generate realizations of $Y(s)$ on a dense grid (`nmesh`=200 points (1-D), 75×75 (2-D)). Each process is randomly point-sampled using an additive white noise model. We now examine the effects of some of the variogram parameters on both the realizations and the simple kriging predictions and variances.

Effect of differentiability parameter ν

The Matérn class of covariance functions is very flexible in that the parameter ν directly affects the differentiability of realizations generated using the model. As Stein (1999) shows, a process with one of this family of covariance functions is $\lfloor \nu - 1/2 \rfloor$-times differentiable (in the mean-square



(a) $\nu = 1/2$ **(b)** $\nu = 3/2$ **(c)** $\nu = 5/2$ **(d)** Same as (c), with conditional simulations

Figure 2·3: Effect of Matérn "smoothness" (differentiability) parameter ν. Realizations of a zero-mean Gaussian process with Matérn covariance for various values of the smoothness parameter ν. Solid line: a sample path of the process $Y(s)$. Dashed line: simple kriging predictor with 1σ (68%) confidence intervals (shaded region). Thin lines (d only): conditional simulations better representing the behavior of individual realizations of $Y(s)$ (for $\nu = 5/2$ only).

sense).

Figure 2·3 compares typical examples of random paths generated by a Gaussian process with zero mean and Matérn covariance function (simulated on a 200-point grid). (The random seed used to generate these paths was consistent between runs, and all other covariance parameters were held constant; hence the general similarity between sample paths.)

The mean-square differentiability of the process $Y(\mathbf{s})$ is governed by the parameter ν, and the prediction surface $\widehat{Y}_{\text{sk}}(\mathbf{s})$ is differentiable only for processes with variograms that are parabolic near the origin (Papritz and Stein, 2002) (which is also a function of ν). Note the (non-differentiable) cusp-like features at the sample points in Figure 2·3a ($\nu = \frac{1}{2}$). By contrast, the predictors in Figures 2·3b&c ($\nu = 3/2$ and $5/2$, respectively) are clearly at least once differentiable.

Figure 2·3d illustrates a few conditional simulations y^{cs}. It is less important whether each sim-

ulation approximates the underlying process more closely than the kriging predictor $\widehat{Y}_{\text{sk}}$ at a given position than to observe the behavior of both $\widehat{Y}_{\text{sk}}(\cdot)$ and $y^{cs}(\cdot)$ between sample points (e.g. note the behavior of simulations for $s \in (1,2)$ and $s \in (4,5)$). The simulations capture the smoothness properties of the underlying process. This is examined in more detail below.

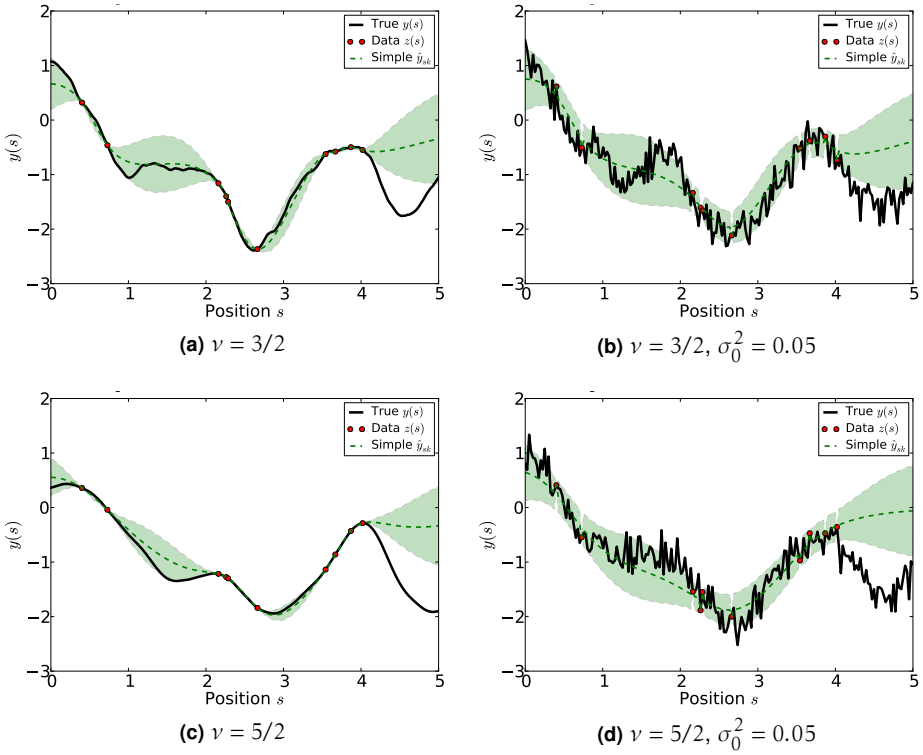
Another feature to notice in all the examples to follow is the behavior of the MSPE $\widehat{\sigma}^2_{\text{sk}}(\mathbf{s})$ in relation to the sample points $\mathbf{s}_i, i = 1,\ldots,m$. The kriging variance at position $\mathbf{s}$—plotted here as $\pm 1\sigma$ confidence intervals about $\widehat{Y}_{\text{sk}}$—is a function of the distances $\{|\mathbf{s} - \mathbf{s}_i|, i = 1,\ldots,m\}$. At each sample point $\mathbf{s}_i$, the prediction error reaches its minimum value. Between sample points, $\widehat{\sigma}^2_{\text{sk}}$ increases relative to (1) the distance between adjacent points, (2) the clustering of nearby points, and (3) the shape and size properties of $\gamma_Y(\cdot)$.

Outside the sample domain (here, for $s \gtrapprox 4$), $\widehat{\sigma}^2_{\text{sk}}$ increases up to the sill while $\widehat{Y}_{\text{sk}}$ shrinks to the mean path (zero in this case). This is a feature of WSS processes: beyond some distance a from the last observation, $Y(\cdot)$ (and indeed $Y(\cdot)|\underline{Z}$) is maximally decorrelated. This is an important theoretical distinction between interpolation and extrapolation. It is also an important practical difference between prediction in the spatial sense (non-causal interpolation using samples within an d-dimensional region of interest) and the temporal sense (extrapolation from an ordered, one-dimensional sequence with little or no foreknowledge of what lies beyond the present), since the latter involves extrapolation while the former is interpolation. Stein (1999) argues in detail why interpolation can usually be assumed the goal of spatial prediction.

Nugget effect

The nugget effect is so named because it models the effect of atypical deposits within a mining survey. Such "nuggets" of high-grade ore are decorrelated from their neighbors (even at arbitrarily small distances), so the nugget effect is represented as a delta function at the origin of the correlation function, or for the semivariogram a constant function at all lags (except at the origin, since by definition $\gamma(0) = 0$). In either case, a discontinuity at the origin models total decorrelation at the micro-scale (i.e. below the shortest intersample distance). (See Figure 2·1.)

A "nugget" could turn up at any position within the prediction domain, thereby increasing the prediction error at that point. As a constant component of decorrelation (except at the origin), the nugget effect increases the kriging variance for all $s \in D_s \setminus \{s_i \mid i = 1,\ldots,m\}$ (i.e. all non-sampled points). Note that the expected prediction error equals zero where $s = s_i$, $i = 1\ldots,m$. From the viewpoint of classical geostatistics, which did not always account for measurement error, kriging



(a) $\nu = 3/2$ **(b)** $\nu = 3/2$, $\sigma_0^2 = 0.05$

(c) $\nu = 5/2$ **(d)** $\nu = 5/2$, $\sigma_0^2 = 0.05$

Figure 2·4: Nugget effect parameter σ_0^2. Realizations of zero-mean Gaussian processes with Matérn covariance. Top: $\nu = 3/2$. Bottom: $\nu = 5/2$. Left: no nugget effect. Right: nugget $\sigma_0^2 = 0.05$.

(in the noiseless case) is an exact interpolator of the data (i.e. $\widehat{Y}_{\text{sk}}(s_i) \equiv Z(s_i)$), so the "prediction" error is nil wherever the process has been sampled directly.

Effect of measurement error

Measurement error has a similar effect but operates by a different mechanism. If data are corrupted with white noise, then neighboring measurements are decorrelated, even arbitrary close ones. So, as with the nugget effect, this is expressed as an added component of uncertainty in the kriging prediction error. However, this decorrelation "at arbitrarily close" distances extends also to the zero distance. I.e., noisy measurements are (so to speak) decorrelated with themselves. More precisely, since measurement error comprises a component of $\underline{Z}$ which is independent of $Y(\cdot)$ and i.i.d. with variance σ_ϵ^2, this is reflected in the kriging variance by a non-zero minimum at any sample point: $\widehat{\sigma}_{\text{sk}}^2 \geq \sigma_\epsilon^2$ for all $\mathbf{s} \in D_s$ *including* sampled locations. Consequently, the predictor $\widehat{Y}_{\text{sk}}(\cdot)$ is not an exact interpolator. It does not "honor the data." Indeed, if the noise variance is appropriately specified, it honors the uncertainty inherent in the data. This added level of uncertainty gives

$\widehat{Y}_{\text{sk}}(\cdot)$ more flexibility in fitting optimal kriging weights to the data. In the presence of noise, this leads to a closer approximation of the particular sample path $y(s)$ (see later section). In Figure 2·8, this becomes apparent.

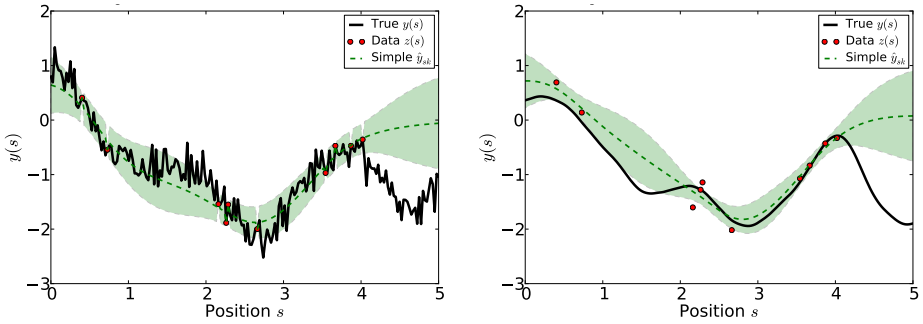
Conflation of nugget and noise

In the past, the distinction of nugget effect and measurement error has been a point of contention between geostatisticians and mathematical statisticians. The difference between the two can be subtle. Quite some effort has been spent clarifying both the theoretical and practical implications of this difference. Matheron's original kriging formulation does not account for noise. A common workaround is to substitute a nugget effect of size σ_ϵ^2, since their effects are similar. While Matheron's orginal formulation attempts to predict unobserved data from observed data, Cressie (1993) argues that the underlying process $Y(\cdot)$ is usually the more scientifically relevant quantity to predict. Diggle and Ribeiro Jr. (2007, p. 139) argue that, while the effect is similar, the difference matters in practice since, if a large isolated datum is encountered, the choice determines whether it should be interpolated (nugget) or not (noise).

Indeed, the uncorrelated nature of most models suggests that $Y(\cdot)$ should be smoother than $\underline{Z}$ $(= Y + \eta_\epsilon)$. Depending on the scale of the covariance model and the fidelity of the data, the interpolating property of kriging explains why *actual* prediction error (i.e. not in the mean-square sense) propagates over a wider area as the predictor attempts to overfit noisy data. (See figure.)

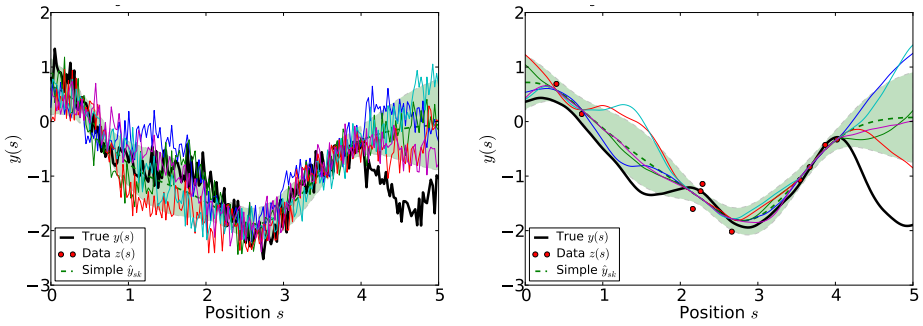
Furthermore, the function mapping process $Y(\cdot)$ to data $\underline{Z}$ is often more complicated than point or block sampling (e.g. nonlinear, $Z = g(Y)$). Inverting these transformations is a discipline unto itself, with solutions that are unstable, non-unique, or nonexistant! In general, such problems are very sensitive to noise. Predicting data *before* inverting is a poor strategy. Instead of kriging, which is encumbered with linearity and Gaussianity assumptions, it is better in such cases to incorporate the full measurement model (including both noise and the (possibly nonlinear) observation function $f(Y)$) into a predictor that operates directly on the data. (Kriging is a special case of such predictors, in which Y is a Gaussian process, with identity $g(\cdot)$, and ϵ AWGN.)

The nugget effect is a property of the process itself, whereas measurement error is a function of the instrument and other environmental factors. Assuming no noise, repeated samples of a nugget at exactly the same position would yield identical measurements (being a realization of a random process, and thus a deterministic function of position). Noise adds an inherent level of uncertainty that can only be estimated through repeated sampling. (Perhaps the confusion stems from the fact



(a) $\nu = 5/2$, $\sigma_0^2 = 0.05$ **(b)** $\nu = 5/2$, $\sigma_n^2 = 0.25$

Figure 2·5: Nugget effect (left) versus measurement error (right).



(a) $\nu = 5/2$, $\sigma_0^2 = 0.05$ **(b)** $\nu = 5/2$, $\sigma_n^2 = 0.25$

Figure 2·6: Nugget effect (left) versus measurement error (right).

that mining samples are destructive and impossible to truly repeat.)

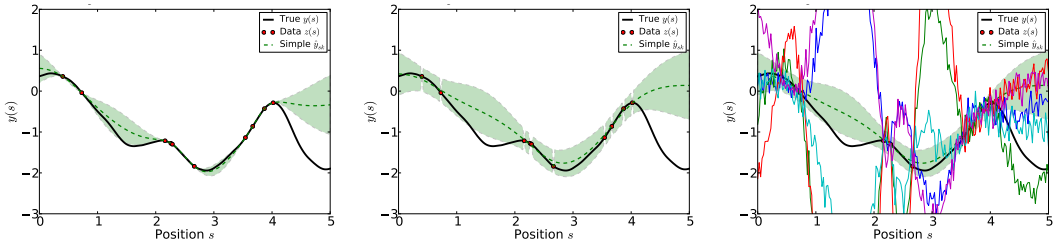
Figure 2·5 compares the results of modeling a nugget effect and additive measurement error. The difference is made obvious here since the measurements (dots) are so much more sparse than the simulated true process (bold line). Note that the underlying sample path is "rougher" when a nugget effect is included. (These are the "nuggets!").

Figure 2·6 also compares nugget effect to measurement error in the case of conditional simulations.

Effect of semivariogram misspecification

Simple kriging relies on very strong assumptions regarding the mean and covariance functions. Figures 2·7 and 2·8, respectively, exemplify the kinds of prediction errors that occur when a nugget effect or noise is misspecified in the semivariogram model.

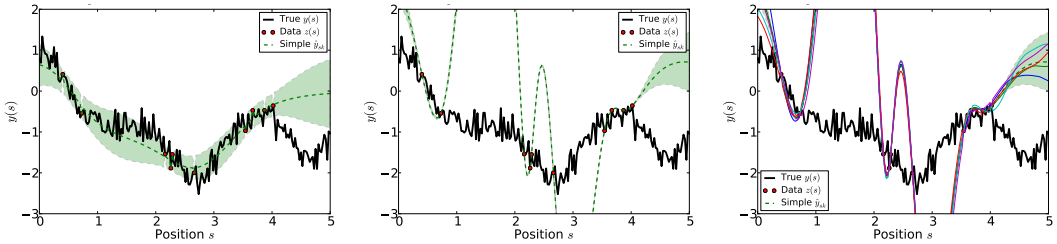
From Figures 2·7b and 2·8b, it would seem wiser to overestimate either noise or nugget effect.



(a) No nugget

(b) Predictor assumes nugget $\sigma_0^2 = 0.05$.

(c) Same as (b).



(d) Nugget $\sigma_0^2 = 0.05$

(e) Predictor assumes no nugget.

(f) Same as (e).

Figure 2·7: Nugget mismatch. Noiseless examples. Kriging with correct model (left) versus incorrect model (right).

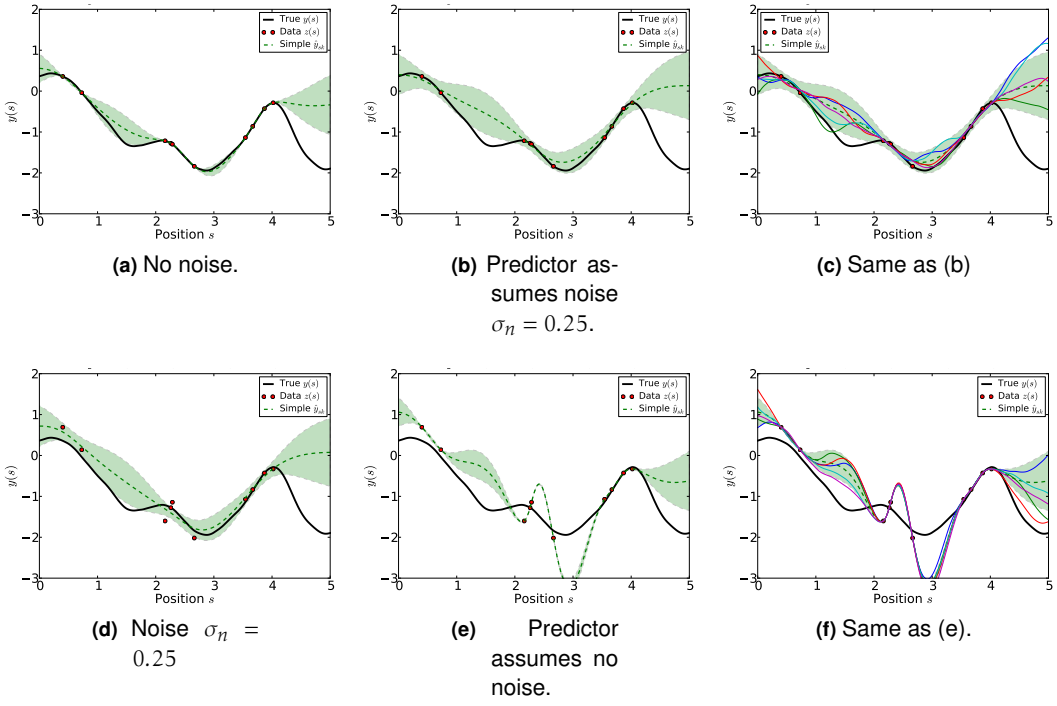
This is reflected in formula (2.9), through $\mathbf{C}_Z$ and in the figures by a widening of the confidence intervals. (This is formally equivalent to shrinkage estimation, damped least squares, and Tikhonov regularization.) Whereas, underestimating either noise or nugget based on sparse data causes the predictor to overfit and oscillate far outside the process' range (Figures 2·7e and 2·8e).

Overestimating the nugget effect results in conditional simulations which may be exceedingly rough (though perfectly plausible, especially if the value of σ_0^2 can be justified by external considerations).

Which parameter to place more emphasis on depends on the goals of the prediction. All that can be recommended generally are rules of thumb. These are conjecture based on this limited set of tests, but they agree with the general recommendations laid out by Cressie and Wikle (2011) and Stein (1999).

- If a smooth representation of the underlying process is desired, model all micro-scale variability as noise and use a kriging predictor.

- Use conditional simulations to get a better impression of the process' spatial variability.



(a) No noise.

(b) Predictor assumes noise $\sigma_n = 0.25$.

(c) Same as (b)

(d) Noise $\sigma_n = 0.25$

(e) Predictor assumes no noise.

(f) Same as (e).

Figure 2·8: Noise model mismatch. In each panel, nugget effect $\sigma_0^2 = 0.0$. Correct noise model (left) versus incorrect model (right).

- Include a nugget effect only if it can be estimated accurately.
- For meaningful representations (i.e. smooth simulations) of $Y(\cdot)$, it is better to overestimate noise than nugget.
- Be careful not to underestimate either parameter.

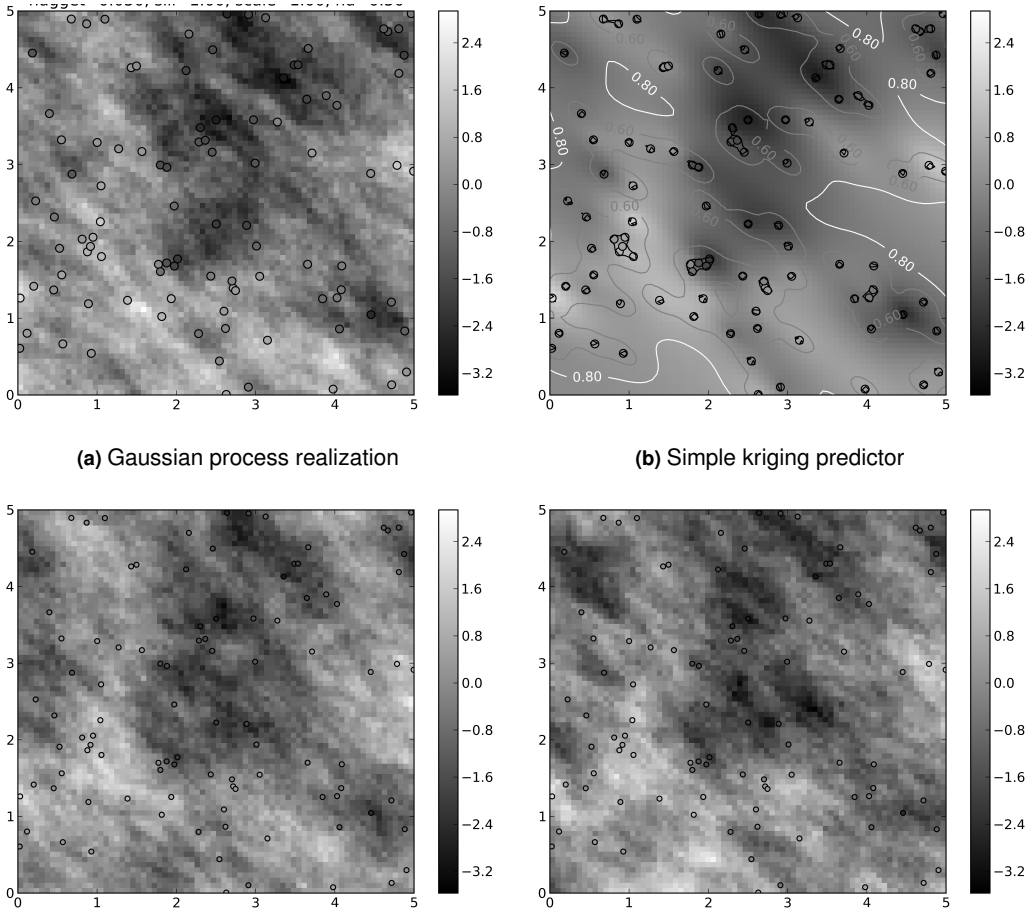
2.5.3 2D kriging example

The relevant statistics of $Y(\mathbf{s})$ are assumed to vary depending only on the spatial coordinate $\mathbf{s}$. This includes but is not limited to the stationary case. And, when $\mathbf{s}$ has dimensionality greater than 1, it includes both isotropic and anisotropic cases. However, we limit our scope to anisotropic processes with $\gamma_Y(h) = \gamma_Y(\|\mathbf{Ah}\|)$, i.e. geometric anisotropy.

In Figure 2·9 a 2D Gaussian process is simulated with anisotropy matrix

$$\mathbf{A} = \begin{bmatrix} 2.0 & 0 \\ 0 & 1.0 \end{bmatrix} \begin{bmatrix} \cos 50 & \sin 50 \\ -\sin 50 & \cos 50 \end{bmatrix}.$$

The covariance function is Matérn type. (See figure caption for parameters.) One hundred obser-



(a) Gaussian process realization

(b) Simple kriging predictor

(c) Conditional simulation #1

(d) Conditional simulation #2

Figure 2·9: Kriging analysis of a Gaussian random process: Matérn covariance function parameters: Nugget: $c_0^2 = 0.05$, Sill $\sigma_0^2 = 1.00$, Scale $\theta = 1.00$, Differentiability $\nu = 0.5$, Noise $\sigma_\epsilon^2 = 0.00$, Anisotropy ratio: $2.0{:}1.0$, Angle $50°$.

vations are randomly sampled from this process (data represented by the color of dots in panel a). The simple kriging predictor (b) is accompanied by the kriging variance. To draw an analogy to the preceding 1D examples, the variance reaches its minimum at the sample positions. It grows with distance from data points. Consequently, the arrangement of samples in space influences the shapes of contours in (b).

Two conditional simulations (c) & (d) were generated independently. The data are also plotted on these graphs for comparison. No noise was assumed in either the generative or predictive models, so $y^{\text{cs}}(\mathbf{s}_i) = y(\mathbf{s}_i)$ for each simulation. The poorly-sampled regions, top center for instance, exhibit the most variability among these simulations (and others not shown). Using (2.21) and the

Table 2.1: Error statistics for the simulations and predictors in Figure 2·11.

Method	RMSE	MAE
Linear	0.362	0.251
Natural neighbor	0.338	0.241
Simple kriging	0.214	0.155
Cond. sim.	0.288	0.212

unbiasedness of $\widehat{y}_{\text{sk}}$, the mean of a very large number of these simulations converges to (b).

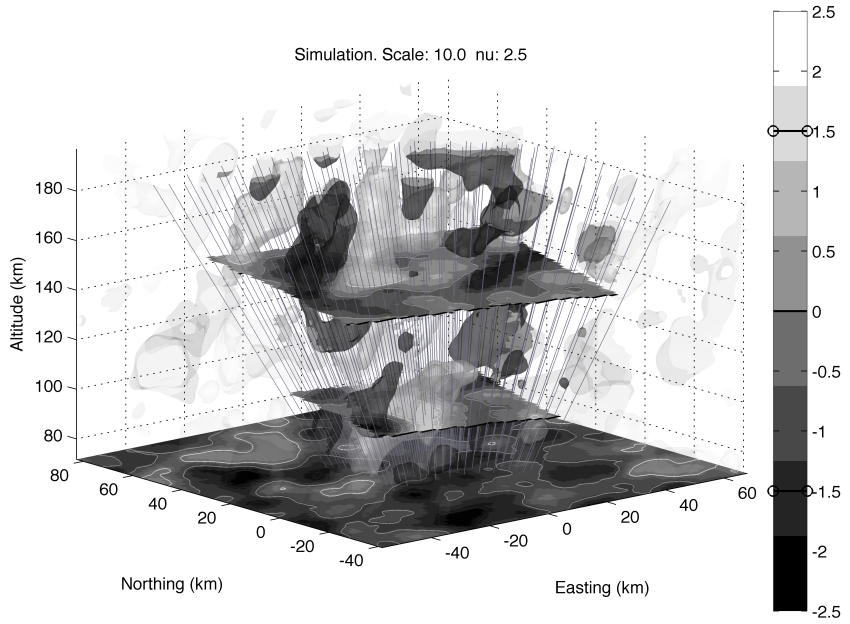
2.5.4 3D kriging example

Finally, in Figure 2·11 mimics a 3D AMISR example by sampling a 3D Gaussian process over an 11×11 grid of beams radiating from the origin (not shown). The sampling interval along each beam is 4.5 km.

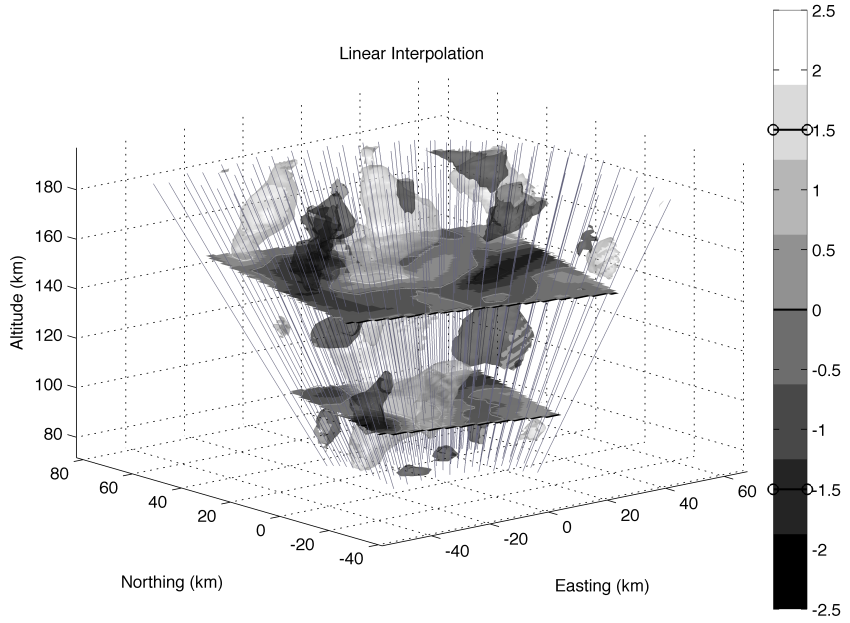
The linear interpolator (b) is continuous but not differentiable, owing to sharp changes of slope at the the edges of the Delaunay triangulation (consistent with Cressie, 1993, p. 374). The natural neighbor interpolator (c), another Delaunay-based method, is visually very similar to simple kriging (d).

A simulation (e) of the process, using the same mean and covariance parameters (assumed known), is conditioned by (d) to match the data.

The kriging variance (f) is minimized near the sample points, with contour lines following the outline of the beams. The variance generally remains low within the sampled region (where prediction corresponds to interpolation), and grows monotonically outside.

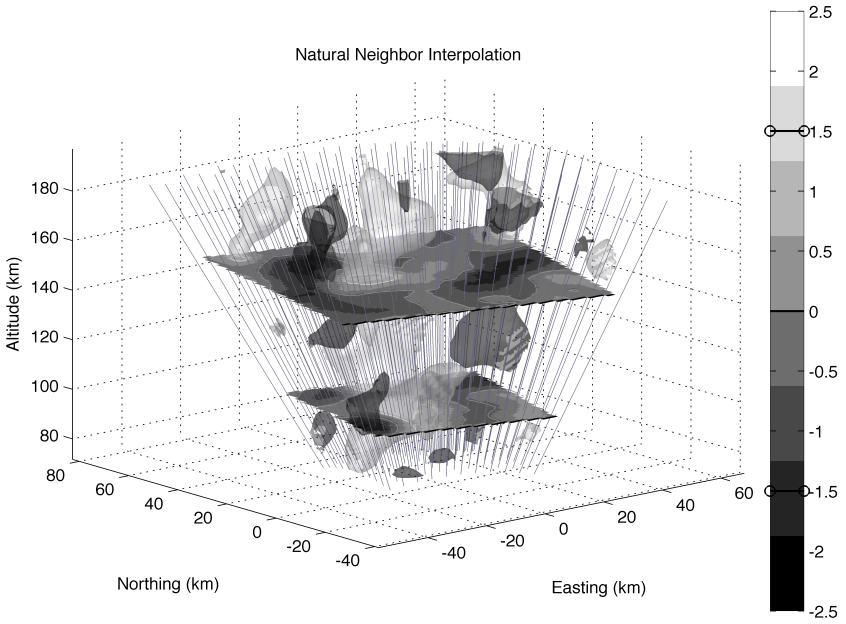


(a) Gaussian process simulated on a $100 \times 100 \times 100$ grid.

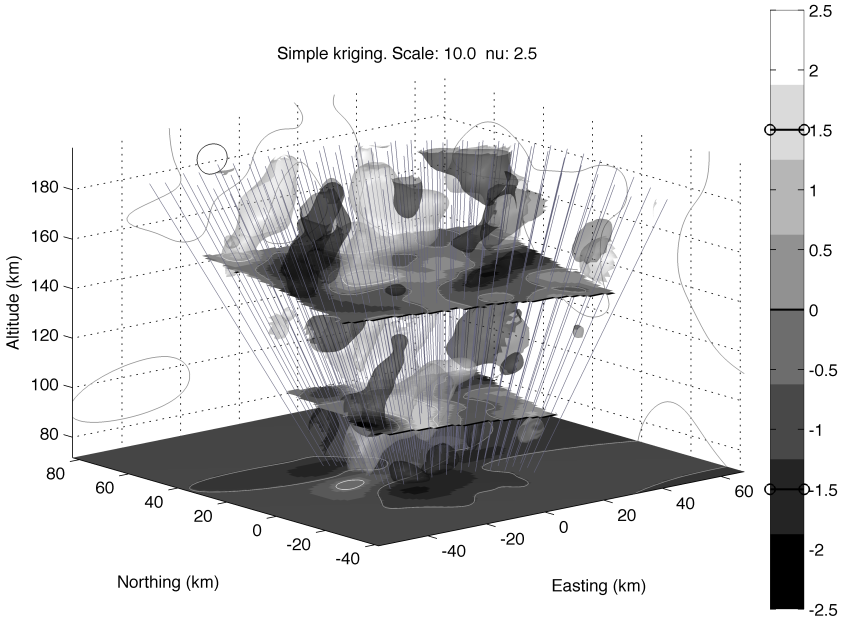


(a) Linear interpolation

Figure 2·11: Analysis of a random process in 3D. Matérn covariance parameters: Sill $\sigma_0^2 = 1.0$, Nugget $c_0^2 = 0.0$, Differentiability $\nu = 1.5$, Scale $\theta = 10$, Anisotropy ratio $1.0\,\mathrm{N/S} : 1.0\,\mathrm{E/W} : 0.8\,\mathrm{z}$ [i.e. slightly elongated vertically], 30 turning bands.The thin lines are sample points along radar beams in an 11×11 angular grid.

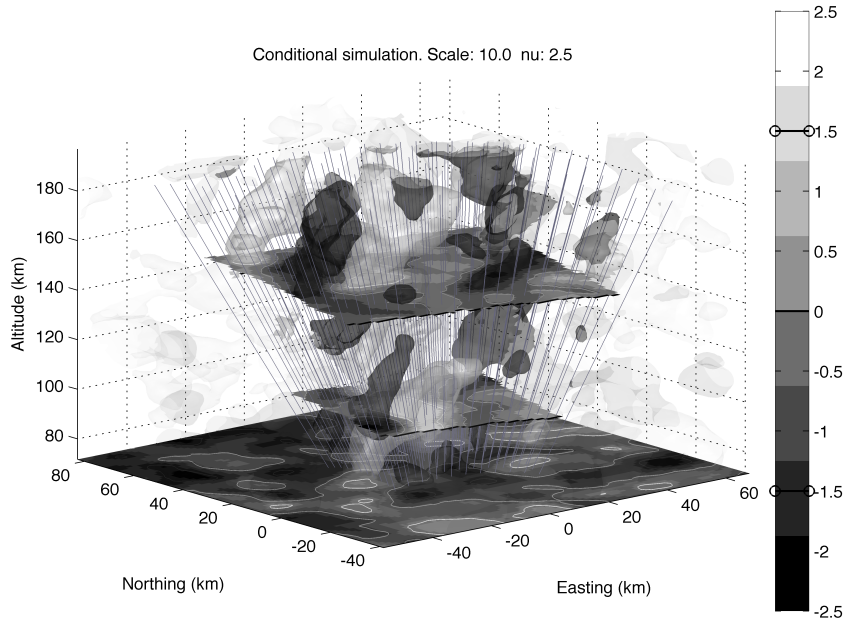


(b) Natural neighbor interpolation

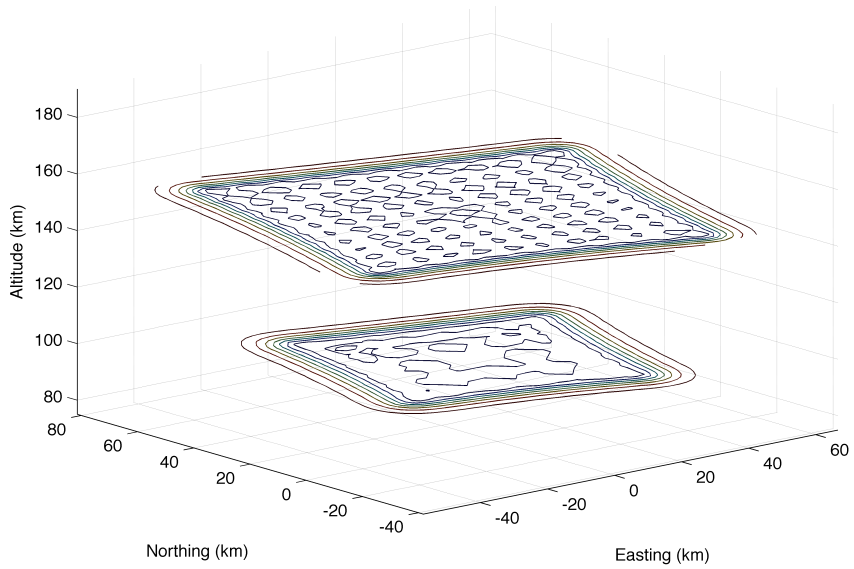


(c) Simple kriging

Figure 2·11: (continued) Natural neighbor interpolation and simple kriging.



(d) Conditional simulation



(e) Kriging standard deviation

Figure 2·11: (continued) A simulation conditioned by (d). Within the sampling region, the simple kriging variance forms contours around the samples (here an 11×11 grid of beams), with its minimum value at the sample location. Outside the sample grid, the variance increases monotonically to its maximum (the sill, if it exists).

Chapter 3

Three-dimensional ISR Imaging

> *It is beyond the tool, and by virtue of it, that we rediscover nature, an experience that we share with gardeners, sailors, or poets.*
>
> Terre des Hommes
> Antoine de Saint-Exupéry
> (Paraphrased from a translation by Chilès and Delfiner (2012).)

The ionosphere is a highly dynamic system that exhibits a wide variety of structured features in response to forces from the magnetosphere. (At high latitudes, the striking patterns and colors of the aurorae are perhaps the most immediate and familiar effects.) As technology advances and new instruments come online, it is worth considering the relative advantages these convey and additional context they provide. Observations that resolve both the structure and dynamics of the ionosphere form a key component in understanding our space environment.

In this chapter, we explore the problem of direct volumetric imaging of the ionosphere via densely-sampled multi-beam *incoherent scatter radar* (ISR) imaging. We begin with a visualization application, applying first linear interpolation and then optimal spatial prediction (or "kriging") to determine the values between measured points (also known as interpolation). Interpolation is crucial for visualization, which is in turn an important component of modern science.

Visualization, though, is only one end to which spatial prediction is a means. Consider also the problem of comparing data from different instruments. If two instruments measure a common region, each with its own sample pattern, their measurements may need to be aligned before analyzing jointly. One solution is to "edit" the samples of one instrument and assume the positions are coincident. Depending on the process and measurement properties, and on the type of analysis, the error incurred by doing so may be acceptable. Another solution is interpolation, which can also be a source of error.

A more subtle problem occurs, for instance, when samples Y_1 and Y_2 are measured with different supports. That is, Y_1 represents an aggregate measure over some region D_1, and Y_2 is an

aggregate measure over some other region D_2. If the regions overlap, there will be a statistical correlation between the two measurements. Furthermore, the correlation may behave unexpectedly, depending on the individual distributions underlying the aggregated regions![1] Any inference involving both data sets must model that correlation.

In a case study, we demonstrate how spatially distributed radar measurements can be compared directly to optical measurements obtained during an active aurora. We also begin to examine the radar's time-resolution by measuring the response of the ionosphere to an enhanced ionization feature.

3.1 Incoherent scatter radar

3.1.1 Radar

Radar is a remote sensing technique used to observe targets that can reflect electromagnetic radiation. Using a carrier wave frequency anywhere from $\sim 5\,\mathrm{MHz}$ to several GHz, a radar transmitter sends a pulse toward a target, and a receiver observes the reflected signal. If the transmitter Tx and receiver Rx are collocated, the radar is *monostatic*. Otherwise, it is *bistatic* or *multistatic*.

The idealized assumption of hard-target radar (such as aircraft tracking) is that the target occupies the space of a point (so that its radar characteristic is isotropic), but is the only body reflecting the incident radar pulse. From the receiver's point of view, a target with nonzero volume can be modeled as the composition of many point targets. The radar pulse signal, as it travels from Tx to the target to Rx, loses power (through propagation, line loss, etc.). The various sources of loss are summarized in the *radar equation*:

$$P_r = P_t \left(\frac{G}{4\pi r^2}\right)\left(\frac{\sigma}{4\pi r^2}\right)\left(\frac{G\lambda^2}{4\pi}\right)\left(t_{\mathrm{ot}}\right)\left(\frac{1}{L}\right), \tag{3.1}$$

where r is range, P_r is the received power, P_t is the average transmitted power, G is the antenna gain (counted twice for transmission and receiving), $\lambda^2/4\pi$ accounts for the effective area of the receiving antenna, L accounts for various losses, t_{ot} is the dwell time, and σ is the *radar cross-section* (r.c.s.) of the target, measured in units of area, and representing the ability of the target to redirect power toward the receiving antenna. Assuming white thermal noise, which is limited by

[1] Depending on the context, this is known as Simpson's paradox, the ecological fallacy, or the Modifiable Areal Unit Problem (MAUP).

the receiver bandwidth B, the *signal-to-noise ratio* (SNR) out of the receiver

$$\text{SNR} = \frac{P_r}{P_n} \propto \frac{P_t G^2 \lambda^2 \sigma}{(4\pi)^3 k_B T_n B L r^4}, \tag{3.2}$$

where T_n is the noise temperature and k_B is Boltzmann's constant.

When the target *fills the volume* of the beam (as in plasma and ionospheric experiments), this is called *soft-target radar*. Here, the data represent an aggregate of many differential volume elements within the beam. This aggregate behavior is described in terms of the effect of the plasma spectrum on the received power signal. After computing a spectrum of the returned pulses, it is possible to infer several properties of the plasma within the scattering volume.

Since the target fills the beam, its volume (and hence the total number of scatterers) *expands* with the beam (at the same rate, $\propto 4\pi r^2$) as it propagates. So for incoherent scatter, the radar equation is

$$P_r \propto \frac{P_t \tau_p}{4\pi r^2 L} \frac{N_e(r)\sigma_e}{(1 + k^2\lambda_D^2)(1 + k^2\lambda_D^2 + T_e/T_i)}, \tag{3.3}$$

where τ_p is the duration of the transmitted pulse, σ_e is the radar cross section of a single electron. The salient factors which differ between equations (3.1) and (3.3) is (1) n_e, the electron density in the numerator, and (2) r^2 in the denominator rather than r^4. Likewise the SNR, which, considering mainly thermal white noise $P_n = k_B T_n B$, becomes

$$\text{SNR} \propto \frac{P_t \tau_p}{4\pi r^2 L k_B T_n B} \frac{N_e(r)\sigma_e}{(1 + k^2\lambda_D^2.)(1 + k^2\lambda_D^2 + T_e/T_i)}. \tag{3.4}$$

Radar resolution considerations

There are four resolution requirements that determine the parameters of an ionospheric radar experiment. The following are adapted from Lehtinen (1986):

Spatial resolution Determined by the width of a radar pulse, since a pulse occupies the space of $\frac{c\tau}{2}$ m, where τ is the pulse length (in seconds), and c is the speed of light (in m s^{-1}). This must be sufficient to capture the spatial features of the target.

Lag (or frequency) resolution Higher resolution requires higher bandwidth. Limited by the correlation time of the target. For overspread targets like the ionosphere, the time during which the scattered signal does not change significantly.

Time resolution The *interpulse period* (IPP) is the shortest interval over which independent measurements are recorded. Additional integration multiplies this interval. The target should

not be expected to change significantly during this time.

Accuracy Accuracy is balanced with the rest of these requirements. For instance, a longer integration time improves data fidelity at the expense of time resolution. Spatial resolution is improved with a shorter pulse, but this requires a wider bandwidth receiver filter, which admits more noise.

There are also two extent requirements. The range extent should be long enough (without ambiguous reflections) to resolve the range of interest. This is governed by the IPP via $T \geq 2L/c$. The lag extent must be broad enough to estimate the relevant Doppler characteristics of the target. Thus the sample time $T \leq$ target bandwidth.

3.1.2 Incoherent Scatter Radar

Incoherent scatter radar is used to measure the ionosphere. This is accomplished by transmitting a pulse with frequency well beyond the plasma frequency $\omega_p \triangleq (Ne^2/\epsilon_0 m)^{-1/2}$ through the ionosphere. This excites the electrons along the beam path, which begin oscillating with the frequency of the radar signal (a phenomenon called Thomson scatter). Each electron acts as a dipole radiator. Randomly oriented and under random thermal motion, the backscattered signal arrives at the receiver.[2]

The radar signal, after backscatter, obtains the temporally correlated signature of the plasma (or, equivalently, its power spectrum). The plasma ACF embodies many properties of the scattering volume. Soon, a series of papers appeared describing accurately and in detail the spectrum obtained in ISR measurements.

The ionosphere is an overspread target, i.e. delay and Doppler cannot simultaneously be resolved (the correlation time of the plasma is much shorter than the IPP. So, rather than standard Doppler analysis, with similar bandwidths of the transmitter and receiver, it is necessary to oversample the received signal and construct a correlation function within a single IPP.

The simplest way to do this is to transmit a single pulse (long pulse, LP) while the receiver oversamples the return signal $z(t)$, then compute lag products $z(t)z^*(t-\tau)$, forming an estimate of the plasma ACF. This can be done by sending more than one short pulse and receiving at corresponding lags $z(t)$, $z(t-\tau)$, $z(t-2\tau)$,... (Farley, 1972), or by using modulating a long pulse: coded pulses

[2] In the original formulation, Gordon (1958) expected a broad Gaussian spectrum, corresponding to free thermal motion. The first experiments by Bowles (1958) showed the spectrum was orders of magnitude narrower. The electrons were indeed the scattering species (the magnitude of the spectrum confirmed this), but they were bound to the heavier ions, the Debye shielding effect limiting their reaction to the radar E-field, and damping the hypothesized Doppler spreading.

(Gray and Farley, 1973) or an alternating sequence of coded pulses (Sulzer, 1993; Lehtinen et al., 1997). The goal, in any case, is to concentrate the signal power over a short region (to improve range resolution) while retaining the energy of a long pulse (proportional to the pulse length τ).

The plasma spectrum

ISR theory was originally developed independently by Fejer (1960), Salpeter (1960), Dougherty and Farley (1960, 1963), Farley et al. (1961), and Hagfors (1961). Reviews of ionospheric scatter methods have been presented by Evans (1969), Farley (1970), Beynon and Williams (1978), and Hagfors (2003). The exposition of Kudeki and Milla (2011) may be helpful to engineers new to this field.

The goal of incoherent scatter radar is to infer the quantitative characteristics of the ionosphere by studying the power spectum of the scattered signal. This is obtained by sampling the returned signal, forming an empirical autocorrelation function, and fitting to an analytic function. The shape of this function is largely controlled by ion dynamics, even though it is the electrons that scatter the radar pulse.

The driving phenomenon of ISR is Thomson scattering. The radar electric field incident on an electron at position $\mathbf{r}$ is (in phasor form)

$$\mathbf{E}_i = E_o(\mathbf{r})e^{-jk_0 r}\hat{r},$$

where $k_o = \omega_o/c$ is the wavenumber of the radar operating at frequency ω_o. $E_o(\mathbf{r})$ is a slowly-varying function of $\mathbf{r}$. This electron is accelerated by the force $-qE_i$ and begins to reradiate at the operating frequency of the radar, essentially acting as a Hertzian dipole. The re-radiated electric field phasor is

$$\begin{aligned}\mathbf{E}_s &= -\frac{q^2\mu_0 \sin\delta}{4\pi r m_e}\mathbf{E}_i e^{-jk_o r} \\ &= -\frac{r_e}{r}\sin\delta\mathbf{E}_i e^{-jk_o r},\end{aligned}$$

where $r_e = \frac{q^2\mu_0}{4\pi m_e} = 2.82\times10^{-15}$ m is the classical electron radius, and δ is the polarization angle (which for linear polarization is the angle between $\mathbf{E}_i$ and $\mathbf{r}_s$. The magnitude of $\mathbf{E}_i$ can be considered approximately constant at E_0 throughout the scattering volume.

Consider a monostatic radar (single antenna), i.e. the backscatter case such that $\delta = \pi/2$, $\sin(\delta) =$

1, and the Bragg wave vector is $\mathbf{k} = -2k_o\hat{r}$. The backscattered field due to a single electron is

$$E_s = -\frac{r_e}{r}E_i e^{-jk_o r} = -\frac{r_e}{r}E_o(\mathbf{r})e^{-j2k_o r}.$$

The total field is the superposition of contributions from individual electons over the subvolume ΔV:

$$E_s = -\sum_{p=1}^{N_0\Delta V}\frac{r_e}{r_p}E_{o,p}e^{-j2k_o r_p} \approx -\frac{r_e}{r}E_o\sum_{p=1}^{N_0\Delta V}e^{j\mathbf{k}\cdot\mathbf{r}_p}.$$

Note that the approximation states $r_p \approx r$ in the fraction, but not in the exponent.

The scattered wave phasor becomes

$$E_s(t) = -\frac{r_e}{r}E_i\sum_{p=1}^{N_0\Delta V}e^{j\mathbf{k}\cdot\mathbf{r}_p(t-r/c)}, \tag{3.5}$$

where the r/c term accounts for the propagation delay from the radar to $\mathbf{r}$. Now the trajectories of individual particles $\mathbf{r}_p(t)$ come into play. The *autocorrelation function* (ACF) of the scattered field is

$$\langle E_s^*(t)E_s(t+\tau)\rangle = \frac{r_e^2}{r^2}|E_i|^2\sum_{p=1}^{N_0\Delta V}\sum_{q=1}^{N_0\Delta V}\left\langle e^{j\mathbf{k}\cdot\left[\mathbf{r}_q(t+\tau-r/c)-\mathbf{r}_p(t-r/c)\right]}\right\rangle. \tag{3.6}$$

If we can regard all the electrons as statistically independent ($p \neq q$), then the ACF reduces to

$$\langle E_s^*(t)E_s(t+\tau)\rangle = \frac{r_e^2}{r^2}|E_i|^2 N_0\Delta V\left\langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}}\right\rangle, \tag{3.7}$$

where $\Delta\mathbf{r} = \mathbf{r}_q(t+\tau-r/c) - \mathbf{r}_q(t-r/c)$ represent particle displacements over τ. This leads to the broadband result originally expected by Gordon. Electrons are not independent, and this form does not account for macroscopic effects.

At this point it is useful to recognize that the summation that appears in (3.5) can be rewritten

$$n_e(\mathbf{k},t) = \sum_{p=1}^{N_0\Delta V}e^{j\mathbf{k}\cdot\mathbf{r}_p(t)},$$

which, in this form, is meant to evoke a 3D spatial Fourier transform of

$$n_e(\mathbf{r},t) = \sum_{p=1}^{N_0\Delta V}\delta(\mathbf{r}_p(t).$$

Taking the Fourier transform of (3.6),

$$\begin{aligned}\left\langle |E_s(\omega)|^2 \right\rangle &= \int d\tau e^{-j\omega\tau} \left\langle E_s^*(t) E_s(t+\tau) \right\rangle \\ &= \frac{r_e^2}{r^2} |E_i|^2 \left\langle |n_e(\mathbf{k},\omega)|^2 \right\rangle \Delta V,\end{aligned}$$

where (3.5) implies

$$\left\langle |n_e(\mathbf{k},\omega)|^2 \right\rangle = \int d\tau e^{-j\omega\tau} \frac{1}{\Delta V} \sum_{p=1}^{N_0\Delta V} \sum_{q=1}^{N_0\Delta V} \left\langle e^{-j\mathbf{k}\cdot\mathbf{r}_p(t-r/c)} e^{j\mathbf{k}\cdot\mathbf{r}_p(t-r/c+\tau)} \right\rangle. \tag{3.8}$$

For independent electrons, this simplifies to

$$\left\langle |n_{te}(\mathbf{k},\omega)|^2 \right\rangle \triangleq N_0 \int d\tau e^{-j\omega\tau} \left\langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}} \right\rangle. \tag{3.9}$$

Although it is not a complete description of the plasma, the correct plasma spectrum is a linear combination of (3.9) and a similar expression for the ions.

Collective effects in a plasma are governed by quasi-static macroscopic currents, forced by polarization fields produced by the mismatch of thermally-driven fluctuations $nte(\mathbf{k},t)$ and $nti(\mathbf{k},t)$. Kudeki and Milla (2011) draw analogies to Kirkhoff's current law and dissipation-fluctuation within an equivalent electric circuit. This leads to the following system of equations, which comprise a general framework for ionospheric ISR, all in terms of $\left\langle e^{j\mathbf{k}\cdot\mathbf{r}_s} \right\rangle$, the single-species ACF.

- Plasma ACF:

$$\left\langle |n_e(\mathbf{k},\omega)|^2 \right\rangle = \frac{|j\omega\epsilon_0 + \sigma_i|^2 \left\langle |n_{te}(\mathbf{k},\omega)|^2 \right\rangle}{|j\omega\epsilon_0 + \sigma_e + \sigma_i|^2} + \frac{|\sigma_e|^2 \left\langle |n_{ti}(\mathbf{k},\omega)|^2 \right\rangle}{|j\omega\epsilon_0 + \sigma_e + \sigma_i|^2} \tag{3.10}$$

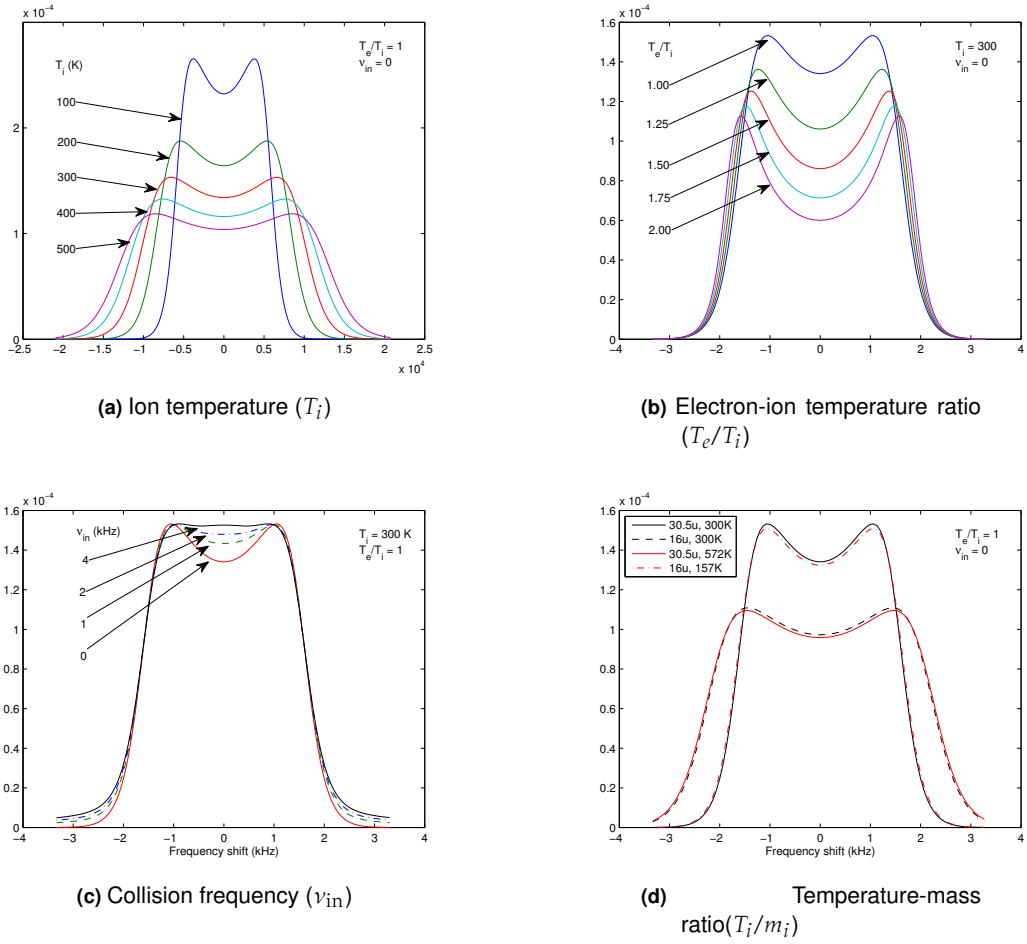
- Constraint on thermal single-species ACFs

$$\frac{\left\langle |n_{ts}(\mathbf{k},\omega)|^2 \right\rangle}{N_0} = 2Re\{J_s(\omega_s)\}$$

- Constraint on conductivities for each species

$$\frac{\sigma_s(\mathbf{k},\omega)}{j\omega\epsilon_0} = \frac{1 - j\omega_s J_s(\omega_s)}{k^2 D_s^2},$$

  where $\omega_s \triangleq \omega - \mathbf{k}\cdot\mathbf{V}_s$ a Doppler-shifted frequency due to the mean velocity $\mathbf{V}_s$ of species s and $d_s \triangleq (\epsilon_0 k_B T_s / N_0 q^2)^{1/2}$ is the Debye length.



(a) Ion temperature (T_i)

(b) Electron-ion temperature ratio (T_e/T_i)

(c) Collision frequency (ν_{in})

(d) Temperature-mass ratio(T_i/m_i)

Figure 3·1: Effects of plasma parameters on ISR spectrum.

Also, in the last two equations, the Gordeyev integral is

$$J_s(\omega) \triangleq \int_0^\infty d\tau e^{-j\omega\tau} \left\langle e^{j\mathbf{k}\cdot\Delta\mathbf{r}_s} \right\rangle.$$

The formula (3.10) is quite general[3], so long as the appropriate single particle ACF is determined. For instance, in a collisionless and nonmagnetized plasma, the Maxwellian pdf is appropriate. The spectrum then has the double-humped form of Figure 3·1.

To demonstrate the effect of various plasma parameters on the spectrum shape, Figure 3·1 consists of spectra evaluated with a variety of plasma parameters. In these plots, the radar frequecy is 931.5 MHz and (except where indicated) the ion mass is 30.5 amu (a mixture of O_2^+ and NO^+

[3]It is also assumed that the medium is stationary and uniform over the scattering volume

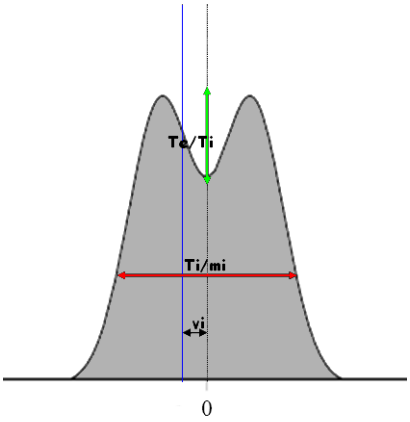


Figure 3·2: Summary of effects of plasma parameters on ISR spectrum.

ions). Figure 3·1a shows the effect of ion temperature T_i with a constant ratio $T_e/T_i = 1$ and zero frequency of ion-neutral collisions (ν_{in}). With increasing thermal excitation, the ion displacements grow larger, broadening the spectrum.

Figure 3·1b shows the effect of varying the electron-ion temperature ratio while keeping the ion temperature constant. As above, an increase in electron temperature broadens the spectrum. However as each ion line moves outward, its broadening is less indicating weaker attenuation of the ion-acoustic wave. Thus the "humps" are narrower and the minimum deeper than in the spectrum with the same width in the top panel.

Figure 3·1c shows the effect of ion-neutral collisions at fixed temperatures. While the width remains fixed, the minimum becomes shallower with increasing collision frequency $v\nu_{\text{in}}$ until it completely disappears and the spectrum becomes Lorentzian[4]. This change is due to further damping of ion acoustic waves and is characteristic of the denser D-region, where the ion-neutral collision frequency is high.

Temperature and mass affect the spectrum in similar ways, and their influence can be ambiguous (Figure 3·1d). Four different spectra are shown, calculated for heavy (mixture of O_2^+ and NO^+, mass 30.5 amu) and light (O^+, mass 16 amu) ions. The narrow spectrum plotted with a continuous line corresponds to heavy ions at $T_i = 300$ K, and the wide spectrum plotted with a dashed line corresponds to light ions at the same temperature. The wide, continuous spectrum is for heavy ions at a temperature of $30 \times 300/16 = 572$ K; the narrow, dashed spectrum is for light ions at a temperature of $16 \times 300/30.5 = 157$ K. The result shows that the two spectrum pairs nearly overlap

[4]The Lorentzian shape is $a/(1 + b\omega^2)$, where a and b are constants.

and it is the ratio T_i/m_i that determines the spectrum width. Figure 3·2 gives a summary of the effects described above.

In principle, it should be possible to determine all the above parameters (ion temperature, temperature ratio, collision frequency, and the concentration ratio of ions with different masses) from the shape of the observed spectrum. However, ambiguities such as those in Figure 3·1d make the task more difficult. In particular, the effects of *ion temperature* and *mass* are difficult to determine simultaneously. This problem arises in the F-region, where a transition occurs from heavy molecular ions to light atomic ions. In practice, the concentration ratios come from a model so that mass is essentially removed from the inversion.

A second difficulty is associated with collision frequency. When the spectrum is double-humped (*i.e.*, in the E- and F-regions), it is difficult to distinguish between the effects of *temperature ratio* and *collision frequency* (Figures 3·1b & 3·1c). The usual solution is to assume $T_e/T_i = 1$ in the lower E-region (below 110 km), and to set $\nu_{\text{in}} = 0$ at greater heights.

In the D-region, the rate of collisions increases dramatically, and the spectrum approaches a Lorentzian shape (i.e. single peak). This shape, being governed by two parameters, only allows the determination of *temperature ratio* and *collision frequency*.

Two other parameters can be determined from the ion lines: electron density and bulk ion velocity. The electron density n_e affects the magnitude of the backscattered signal. Thus power measurements can be converted to n_e using the radar equation.

If the ionospheric plasma is in motion, the total spectrum is shifted and a single component of the plasma bulk velocity can be determined from this Doppler shift. In the case of backscatter, the frequency shift gives the line-of-sight ion velocity. In the case of a bistatic configuration, it resolves the component along the bisector.

The ISR signal

Because of the random thermal motion of the electrons, the scattered signal is a random variable. At the receiver, the scattered signal is also corrupted by sky noise and thermal noise within the instrument:

$$P_{S+N} = S^2 + N_{\text{sky}}^2 + N_{\text{sys}}^2.$$

To reduce the variance of the random fluctuations, K pulses are integrated, so that $P_{S+N} = \frac{1}{K}\sum_k P_k = \frac{1}{K}\sum_k I_k^2 + Q_k^2$.

To distinguish the scattered signal from noise, one set of observations must be devoted to mea-

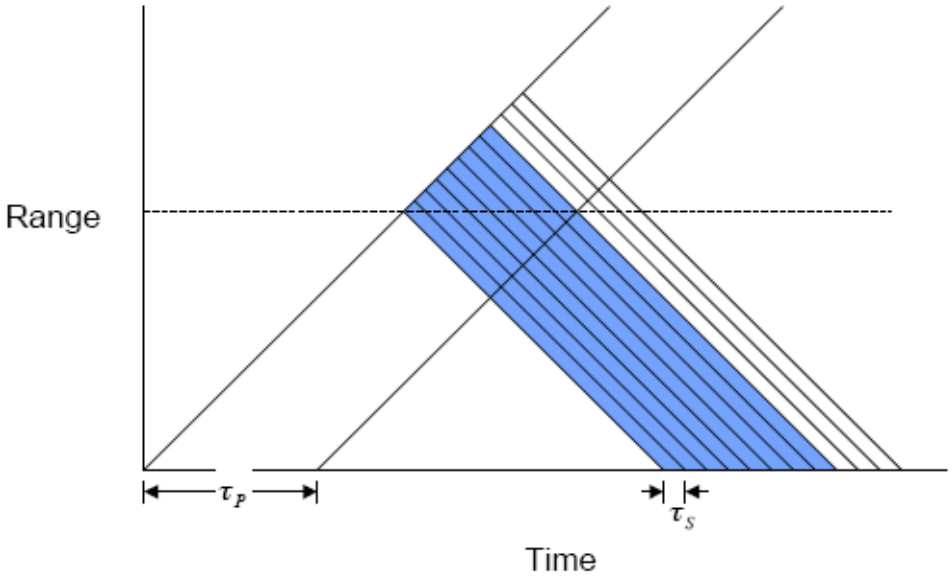


Figure 3·3: The ISR measurement process (monostatic, long pulse). A pulse is transmitted of duration τ, and the receiver oversamples returns in order to estimate the acf. Image credit: Phil Erickson.

suring only the sky noise, which of is stationary (along the radar beam) and independent of S^2. It is not constant, though. So it must be continually compensated by sampling beyond the plasma, where the received signal consists only of the noise components. The total noise is estimated by averaging these long-range samples: $P_N = \frac{1}{K_N}\sum_k N^2_{k,\text{sky}} + N^2_{\text{sys}}$. The system noise N^2_{sys} is estimated by injecting noise of a known temperature into the receiver during an idle period. Then the noise-compensated power signal is

$$\widehat{P} = P_{S+N} - P_N.$$

The normalized variance of of $\widehat{P}$ is

$$\text{Var}\left(\frac{\widehat{P}}{P}\right) \propto \frac{1}{K}\left(\frac{S^2+N^2}{S^2}\right) = \frac{1}{K}\left(1+\frac{1}{\text{SNR}}\right). \tag{3.11}$$

(See Farley (1969) and Lehtinen (1986) for more details.)

Estimating the plasma ACF from lagged products

The ISR measurement process involves sending a pulse, then sampling and storing lag products. (See Figure 3·3.) For instance, after sampling $\{Z_0, Z_1, Z_2, \ldots\}$ (where $Z_i = I_i + jQ_i$ is a complex signal

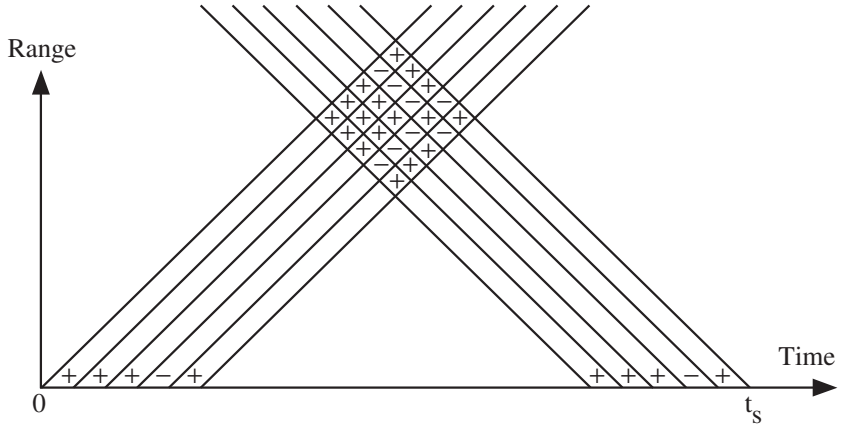


Figure 3·4: The ISR measurement process (monostatic, Barker coded pulse). A 5-baud phase-coded pulse. In the receiver's matched filter, the decoded components add in-phase only from a narrow range interval. Image credit: (Farley, 2008).)

resulting from *in-phase/quadrature* (IQ) modulation), the lag products are computed:

$$\langle Z_0 Z_0^* \rangle = I_0^2 + Q_0^2$$
$$\left\langle Z_i Z_j^* \right\rangle = (I_i^2 + Q_j^2) + j(I_j Q_i - I_i Q_j)$$

These products are the starting point for a nonlinear least squares fitting procedure. The ACF encodes information about the plasma such as electron density, electron and ion temperatures, ion composition (by mass), and bulk $E \times B$ drift.

Figure 3·3 shows the range-time diagram for a straightforward measurement method. A single, long pulse is transmitted. Then the receiver samples the backscattered signal at a faster rate. The shaded regions of overlapping ranges depict those regions of the scattering volume that—for those lags—are correlated, and which therefore contribute to the estimate of plasma ACF. Conversely, the unshaded regions contribute noise to the measurements in the form of uncorrelated clutter. Although the volume of the correlated region decreases with increasing lag, the range resolution is governed by the pulse length τ.

To improve range resolution, we could use a shorter pulse. But that would (1) transmit less power and (2) require an increase in the receiver bandwidth. The combination of reduced signal power and increased noise is not a wise strategy. Especially, considering (3.11), if it leads to a very low SNR.

On the other hand, there are a number of clever methods of improving range resolution. These include transmitting multiple short pulses and receiving at corresponding lags $z(t)$, $z(t-\tau)$, $z(t-$

2τ),... (Farley, 1972). An alternating sequence of coded pulses is another strategy (Sulzer, 1993; Lehtinen et al., 1997). The goal, in any case, is to concentrate the signal power over a short region (to improve range resolution) while retaining the energy of a long pulse (proportional to the pulse length τ).

Because we wish to observe high variability in both space and time, we focus on the region below ~ 300 km and use Barker coded pulses (Gray and Farley, 1973) to probe the region with high spatial and temporal resolution. A Barker code is a specific type of binary phase code (with phase indicated by '+'/'-') with the property that, after matched filtering, its sidelobes have magnitude no greater than one. An M-length Barker code results in an a main lobe with magnitude M. Barker coded pulses yield measurements with high range resolution ($c\tau/2M$), at the expense of spectral information due to the sidelobe clutter.

Propagation of uncertainty

The estimation variance associated with a Barker code is on the same order as (3.11). In a given beam direction, AMISR observes range-resolved power. Each measurement is accompanied by a measurement variance $\left\langle \widehat{P} \right\rangle / \langle P \rangle$, which is then propagated through the kriging variance via (2.9). This generates the spatial map of uncertainty as discussed in Chapter 2.

The enabling technology for this thesis is a relatively new class of instrument in ionospheric study. The electronically-steerable ISR platform known as AMISR is capable of repointing its beam while gathering data, providing a level of spatial context not previously afforded to ISR. As the beam of AMISR sweeps across the sky, it registers the returned pulses in angular direction as well as range. Depending on the dynamics of the process under observation, measurement over the entire *field-of-view* (fov) can be regarded as simultaneous. This is the essential difference of an electronically steerable beam: data registered simultaneously in both azimuth and angle as well as range constitute a 3D "snapshot" of the target volume.

3.2 Exploring volumetric ISR data

A single, stationary beam yields observations of how reactive the target is to radar pulses as well as the distance (range) to the target along the beam. A *range-time-intensity* (RTI) plot provides a quick visual summary of the evolution of features within a given beam. (See Figure 3·5.) In a range-time-intensity (RTI) plot, the color axis depicts return signal power P_r. (For the plots in this chapter have transformed P_r to an approximation of N_e, the electron density, because this is the quantity of

scientific interest.)

Figure 3·5a is simply computed from the RTI plot corresponding to one beam direction. In fact, this is the beam closest to the zenith (actually 88° elevation). The four plots of Figure 3·5c are its neighbors to the NE, SE, SW, and NW, each at 86° elevation. The RTI plot in panel b is a complete fiction. It represents a "virtual beam" in the zenith direction, emerging from the center of the radar. (Any orientation can be selected, of course, but since most atmospheric properties vary with height, and their vertical profiles will ultimately be compared, it is reasonable to choose the zenith or the magnetic zenith.)

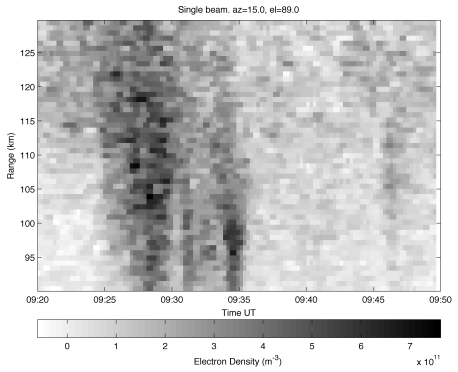
The purpose of such a mapping is to align measurements for comparison. RTI plots are commonly used as summaries, displayed one beside another. Patterns can be obscured this way since a beam pointing 60° off the horizon will highlight, not merely lower-altitude features than a zenith beam, but possibly different types altogether in an aniostropic atmosphere. One could imaging simply mapping range to height, which may be suitable for summary purposes, but it again neglects the possible horizontal features as above.

This zenith beam is not a simulation of what the radar would "see" were it operating in single-beam mode. It's orientation and location can be chosen anywhere within the observing region, with the full resolution of the data available. Two adjacent virtual zenith beams could be predicted and, while their profiles will be similar owing to the spatial dependence of the random field they would be unique. (A sequence of such profile predictions is, of course, the basis of volumetric imaging.)

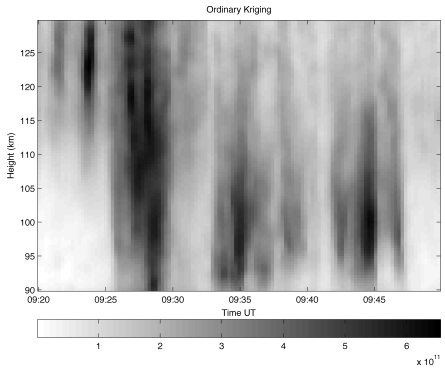
The image of Figure 3·5b is necessarily smoother than the others, since it is constructed using a kriging predictor. The spatial variability of kriging predictors is less than the processes they predict.

A detail of this event is depicted in profiles of electron density versus height (Figure 3·6). Each row depicts the same five-minute period, but different (offline) integration times. Beginning at the top, panel a collapses all five minutes into a single profile, panel b is integrated approximately half as long, panel c half again, and panel d represents the finest time resolution available for these data, $\sim 15\,\mathrm{s}$. Naturally, the course of propagation is more easily traced in the higher-resolution profiles.

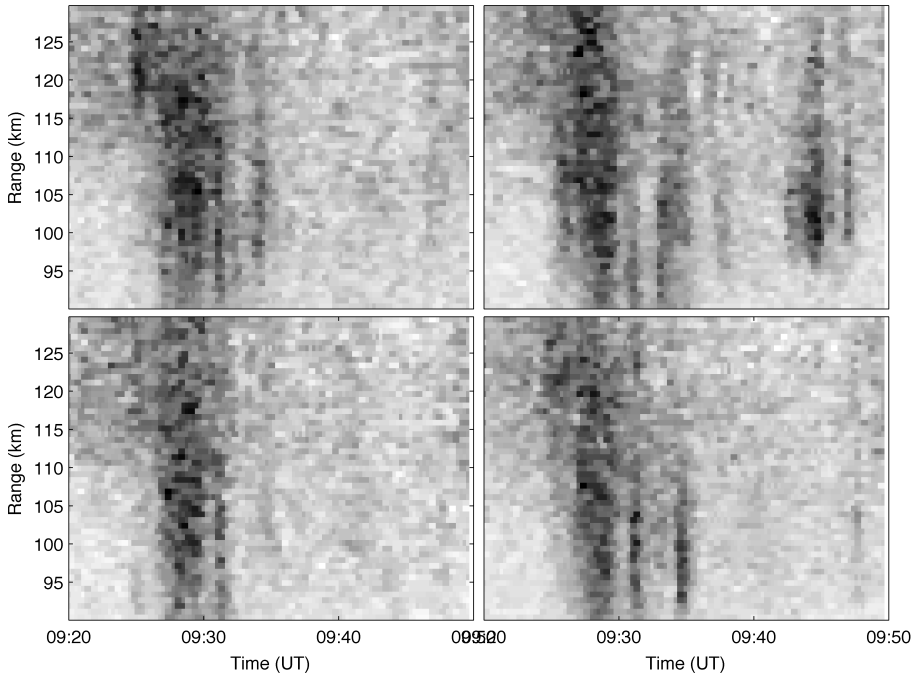
Figure 3·6 also compares the kriging predictor with two popular interpolators: trilinear and natural neighbor. Both are based on Delaunay triangulation. Linear interpolation, despite its popularity, is known to be non-differentiable in three dimensions. Natural neighbor interpolation, which weights the influence of nearby data based on the volumes of their respective Voronoi poly-



(a) RTI from power returns in one beam only

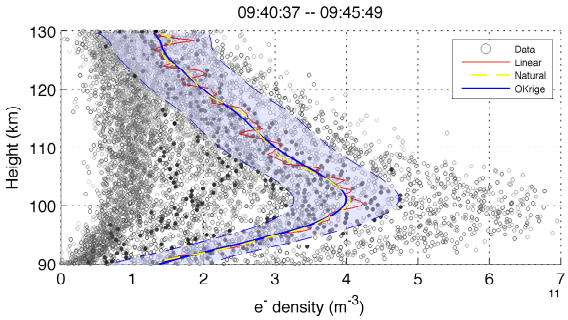


(b) "Virtual" RTI in zenith direction, near center of fov.

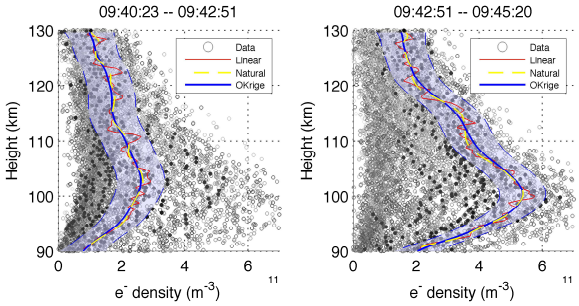


(c) Four beams surrounding the on in (a).

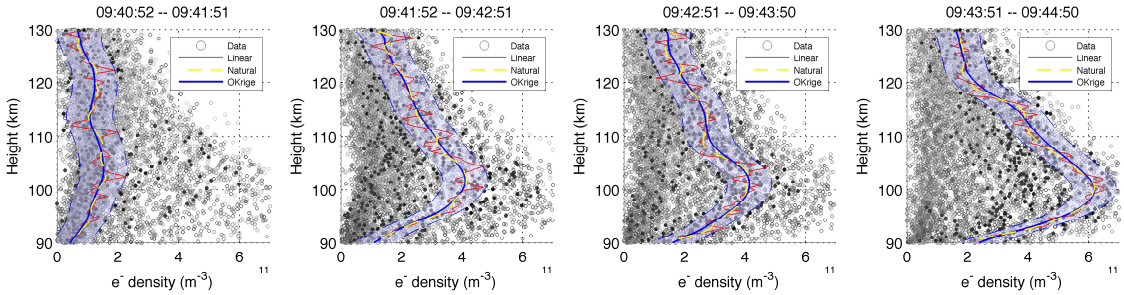
Figure 3·5: Range-time-intensity (RTI) plots showing the ionospheric response to an auroral ionization structure. The downward approach of (a) The view along a single beam. N_e is computed from the instantaneous power returns at each range gate. (b) A "virtual" RTI oriented along the zenith.



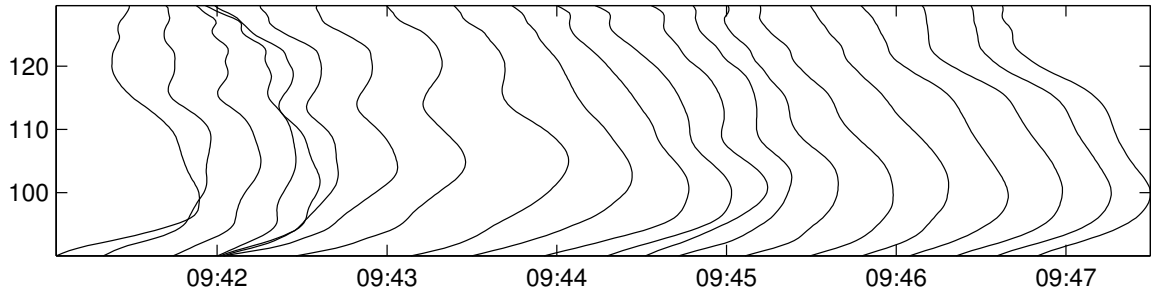
(a) 5 minutes (960 pulses)



(b) 2 minutes (480 pulses)



(c) 1 minute (192 pulses)



(d) 15 seconds (48 pulses)

Figure 3·6: Vertical profiles of electron density, predicted along a virtual beam near the center of the fov. Kriging versus interpolation. Each successive row is a time sequence of increasing time resolution (achieved by post-integrating data with native resolution $14.6\,\mathrm{s}$). Kriging uses all the data (values depicted on the axes as dots), after weighting by distance (here, shading). Row (d) shows kriging only.

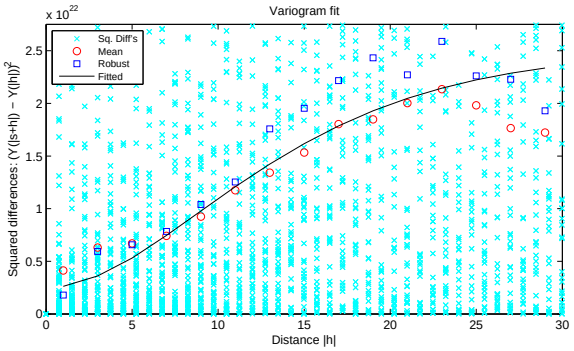


Figure 3·7: Fitting the variogram of electron density data versus altitude.

hedra (Sibson, 1981; Cueto et al., 2003), is very similar to kriging. Cressie (1993, pp. 373–376) compares both of these Delaunay methods to kriging.

Also plotted along with the profiles in Figure 3·6 are the data they interpolate. Each data point is shaded according to its horizontal distance from the "virtual beam" of the prediction (darker=closer). The kriging predictor uses all the data, but weights it according to distance, clustering, and the covariance function $C_Y(\mathbf{s})$.

Variography

The form of kriging used in this example is Ordinary Kriging (see Section 2.3). This method assumes a known covariance function $C_Y(\mathbf{s})$ and, from data $\underline{Z}$, determines both the unknown, constant mean μ_Y and the minimum-MSPE predictor $\widehat{Y}_{OK}(\underline{Z})$. Of course, $C_Y(\mathbf{s})$ is not known. Rather, in this case, it is estimated by examining the *empirical semivariogram*. Letting $\underline{Z}$ represent a set of electron densities along single vertical line between 130 km to 200 km the empirical semivariogram is given by (2.17):

$$\left(Z_i - Z_j\right)^2 \qquad \forall i, j \in \{1, \ldots, m\}.$$

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} \left(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right)^2, \qquad \mathbf{h} \in \mathbb{R}^d. \tag{3.12}$$

The plot in Figure 3·7 shows the semivariogram cloud, the squared-differences versus their respective vertical distances $s_{z,i} - s_{z,j}$. To estimate the actual semivariogram, the cloud is binned and averages computed. The standard estimator is simply the mean value of points in each bin. Cressie and Hawkins (1980) suggests an alternative to mitigate the effects of extreme outliers:

$$2\bar{\gamma}(h) = \left\{ \frac{1}{N} \left| Z_i - Z_j \right|^{1/2} \right\}^4 / (0.457 + 0.494/N).$$

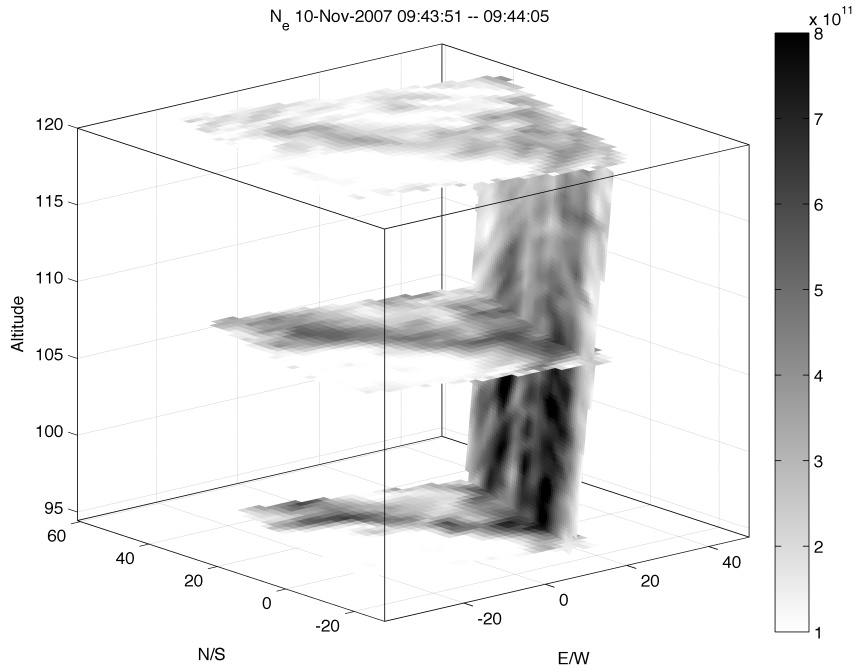
Both estimators are plotted in Figure 3·7 along with an estimated variogram of Matérn type ($\nu = 5/2$) with nugget $c_0 = 2.5 \times 10^{21}$, sill $\sigma_0^2 = 2.5 \times 10^{22}$, and scale parameter 17 km.

Finally, Figure 3·8 also shows these data in 3D, first using trilinear interpolation, then kriging using the parameters derived above.

3.3 Experiment: Direct volumetric imaging of ISR electron densities

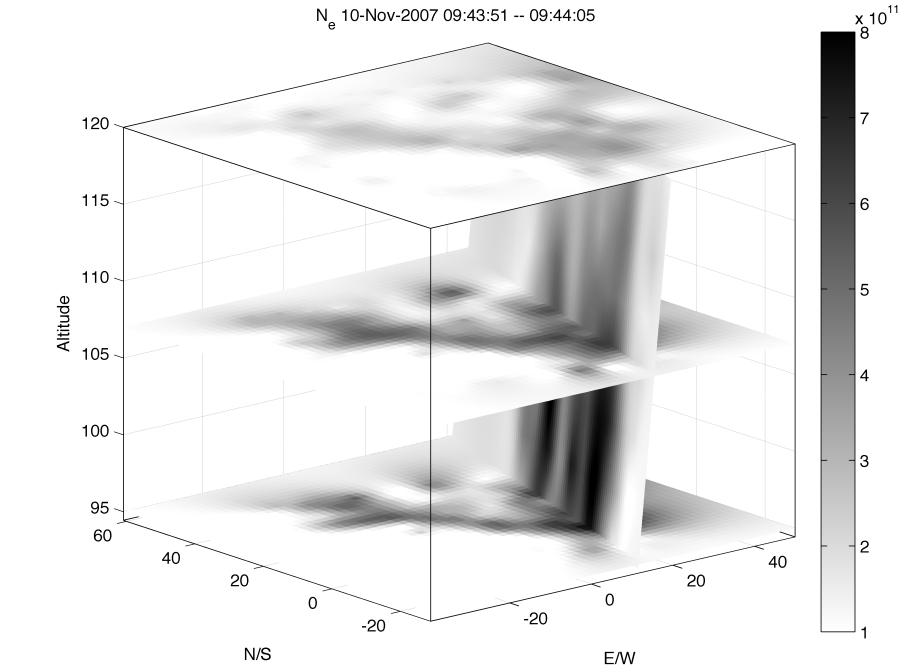
Data were collected on 10 Nov, 2007 with PFISR cycling through an 11×11 grid of beam positions. This is an extremely dense sampling mode, with 3° separation between adjacent beams in each orthogonal direction. At 100 km altitude, the sampled region is approximately rectangular with sides ~ 65 km $\times$ 60 km. At the same height, the horizontal spacing between beam centers is ~ 5.2 km to 6.2 km. At that time, only 96 panels were installed, and the $1° \times 1.5°$ beamwidth at 100 km altitude was ~ 1.7 km $\times$ 2.6 km to 2.1 km $\times$ 3.1 km.

PFISR is capable of running three channels simultaneously, and for this experiment we used data from two channels, each operating with a 13-baud Barker coded pulse. With 10 µs baud

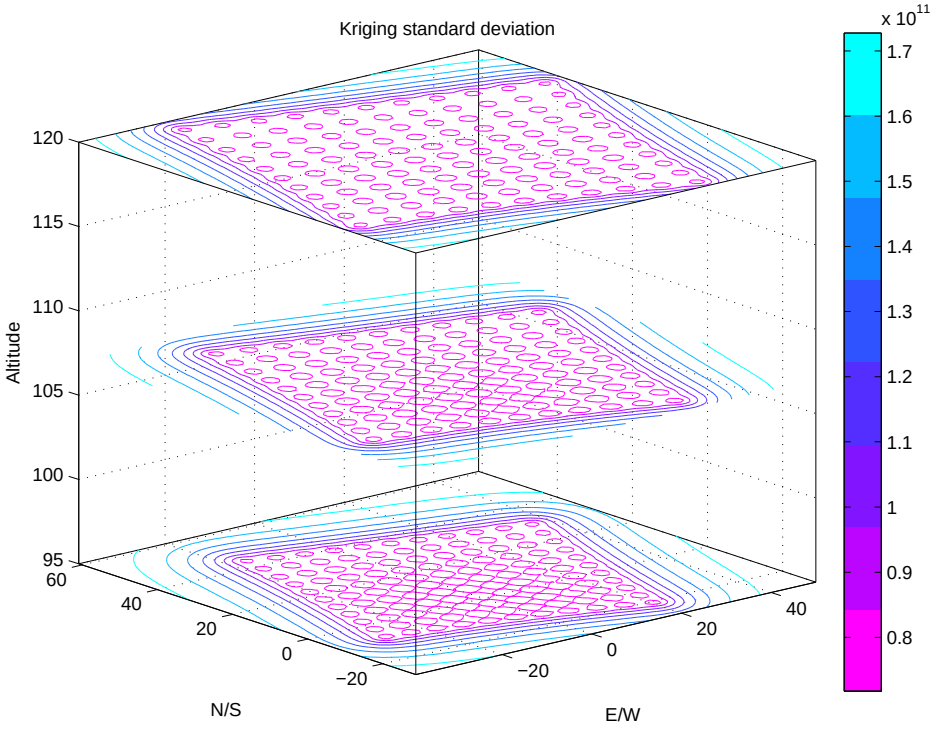


(a) Linear interpolation

Figure 3·8: Trilinear interpolation of electron density derived from backscatter power. 11 Nov, 2007. Integration time: 15 s.



(b) Ordinary kriging



(c) Kriging standard deviation

Figure 3·8: (continued) Ordinary kriging prediction of electron density from backscatter power. 11 Nov, 2007. Integration time: 15 s.

lengths, this results in a range resolution of ~ 1.5 km.

This setup was used to monitor E-region electron density in each of the 121 beams up to about 150 km altitude. Electron density was computed from received power. In this region, neutral particle collisions place electrons and ions in (approximate) thermal equilibrium. Substituting $T_e = T_i$, that is the effective r.c.s. for this scattering volume is

$$\sigma = \frac{\sigma_e}{(1 + k^2\lambda_D^2)(2 + k^2\lambda_D^2)},$$

where σ_e is the r.c.s. of a single electron (a known constant), k is the radar signal wavenumber, and the Debye length $\lambda_D = (\epsilon_0 k_B T_e / n_e q^2)^{1/2}$ is a fundamental scale length in plasma physics.

Using the radar equation for volume scatter, the electron density is directly proportional to received power

$$N_{e,\mathrm{raw}}(r,\theta,\phi) = \frac{2C_s r^2 P_{rx}(r,\theta,\phi)}{P_{tx}\tau} \tag{3.13}$$

The system calibration constants $C_s(\theta,\phi)$, encapsulating various losses, are provided with the data.

The data products in these experiments are spatio-temporally–resolved physical parameters. As discussed in Section 3.1, the range resolution, temporal resolution, and cross-range resolution are determined by a tradeoff between the pulse width, IPP, and number of beams, respectively. (The investigator determines which combination is most appropriate based on the resolution and extent requirements of a particular experiment.)

To cycle through this grid of 121 beams takes PFISR 0.61 s. To reduce the amount of data for storage, 24 pulses were integrated for each beam, giving a temporal resolution of ~ 15 s. Since the SNR is very high in this experiment, the expected error for point measurements (3.11) reduces to $1/\sqrt{K}$. For $K = 2\times 24 = 48$ (observing on two independent channels), the uncertainty is 14.4%. This is quite high, but it can be reduced by post-integrating the 15-second samples (thereby lowering the time resolution).

3.3.1 3D imaging

With such a small number of pulses, only the most energetic events are likely to be resolved above the level of noise. One such event, an auroral arc activation, is also characterized by dynamics that make a short-cadence instrument attractive for studying it. With a cadence of 15 s, though not comparable to the speed of an optical camera, PFISR was able to capture the response of the ionosphere to an individual arc activation. The following events can be observed in the 3D images:

09:23:30–09:24:59 UT An auroral ionization structure at 120 km altitude extending down to nearly 107 km. (This is observed in roughly the same horizontal location directly below, where we would expect it, so we must be fully resolving this event in time at this cadence.) Figure 3·9.

09:33:26–09:35:24 UT An annular structure at 100 km. Compare this with the reconstructions at 5-min cadence. Figure 3·11.

09:37:39–09:39:37 UT West-to-east apparent motion. Recombination time on the order of seconds. Apparent motion actually due to motion of ionizing sources. Figure 3·13.

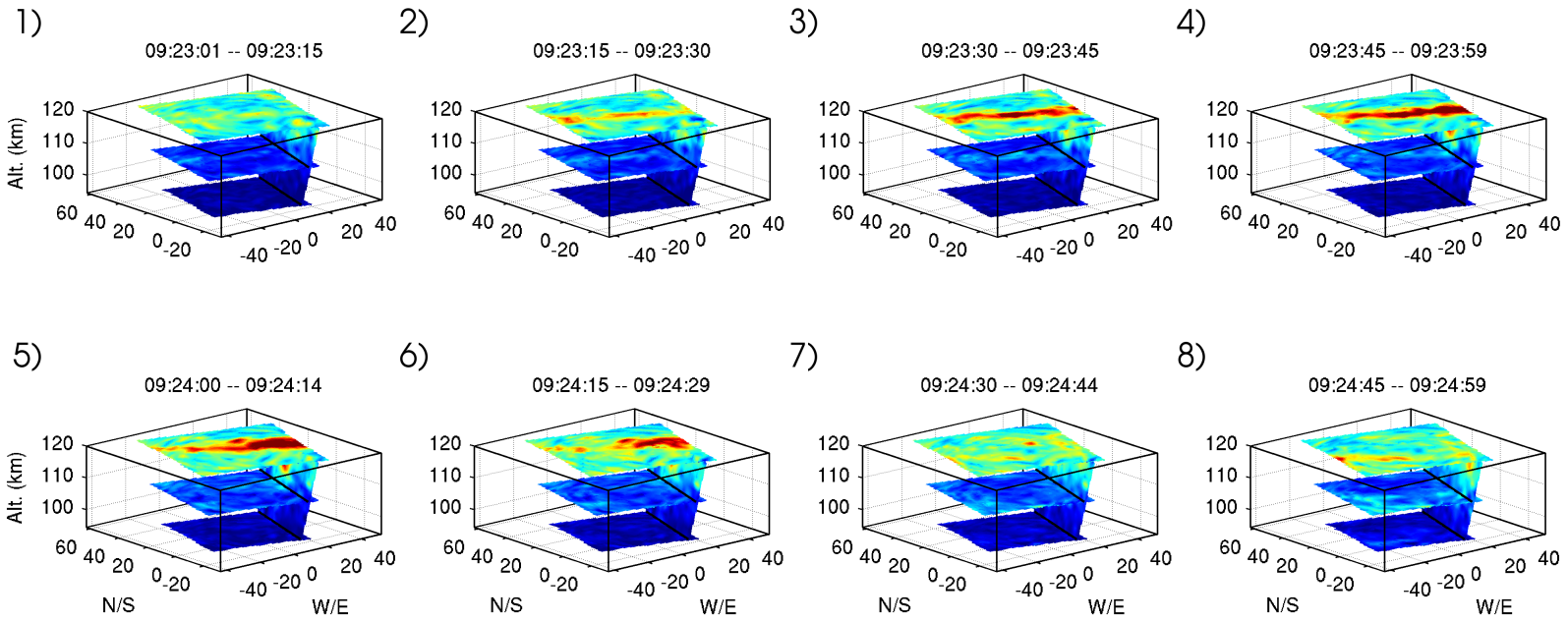


Figure 3·9: 10 November, 2007. 15-second reconstructions. Trilinear interpolation.

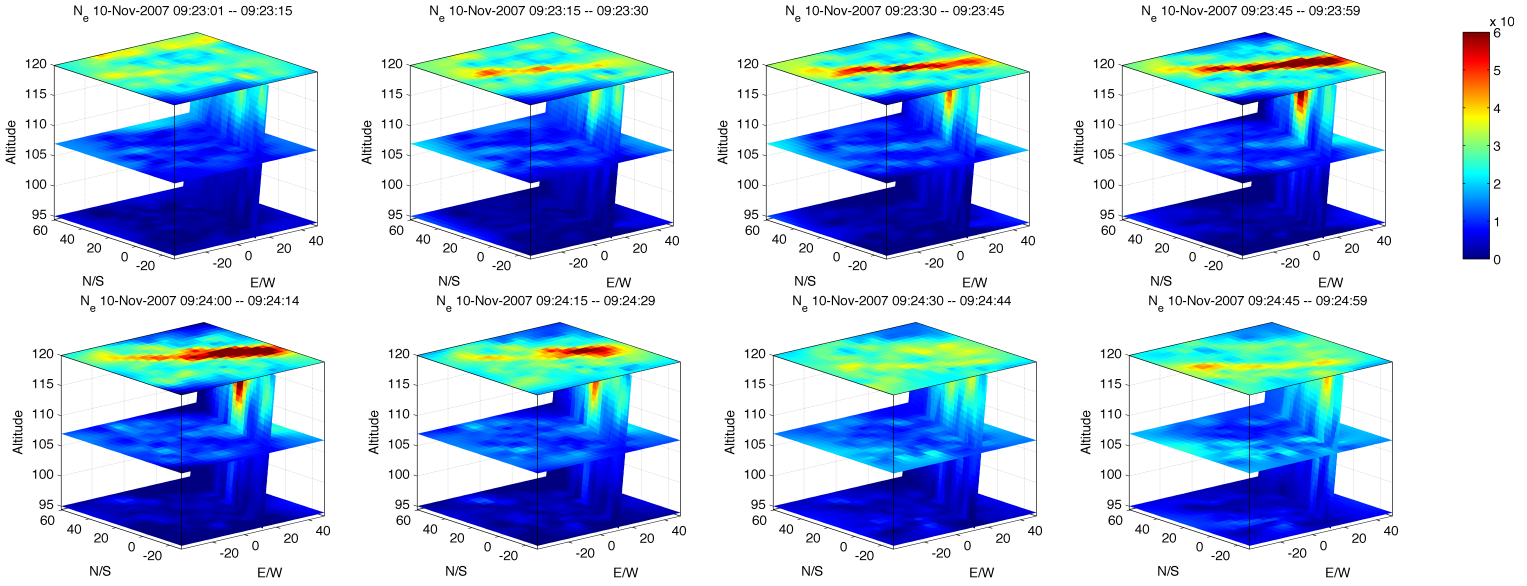


Figure 3·10: 10 November, 2007. 15-second reconstructions. Universal kriging.

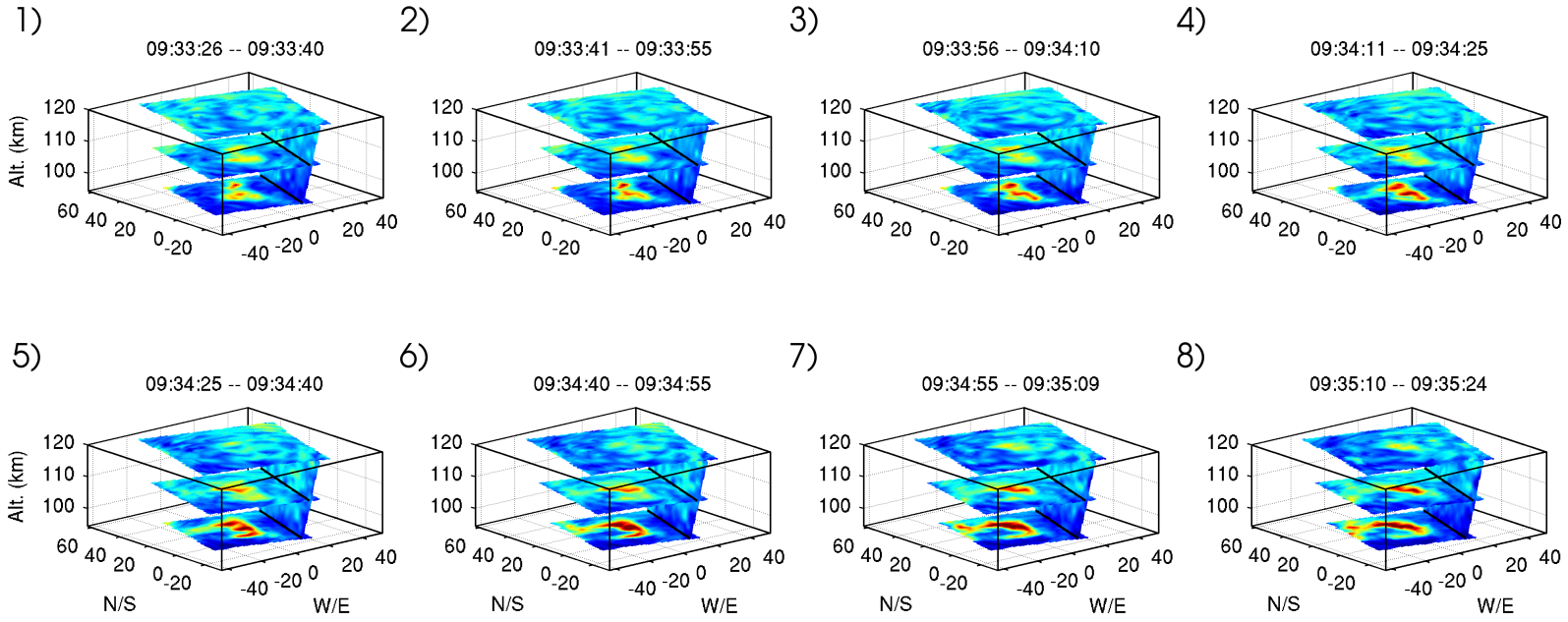


Figure 3·11: 10 November, 2007. 15-second reconstructions. Trilinear interpolation.

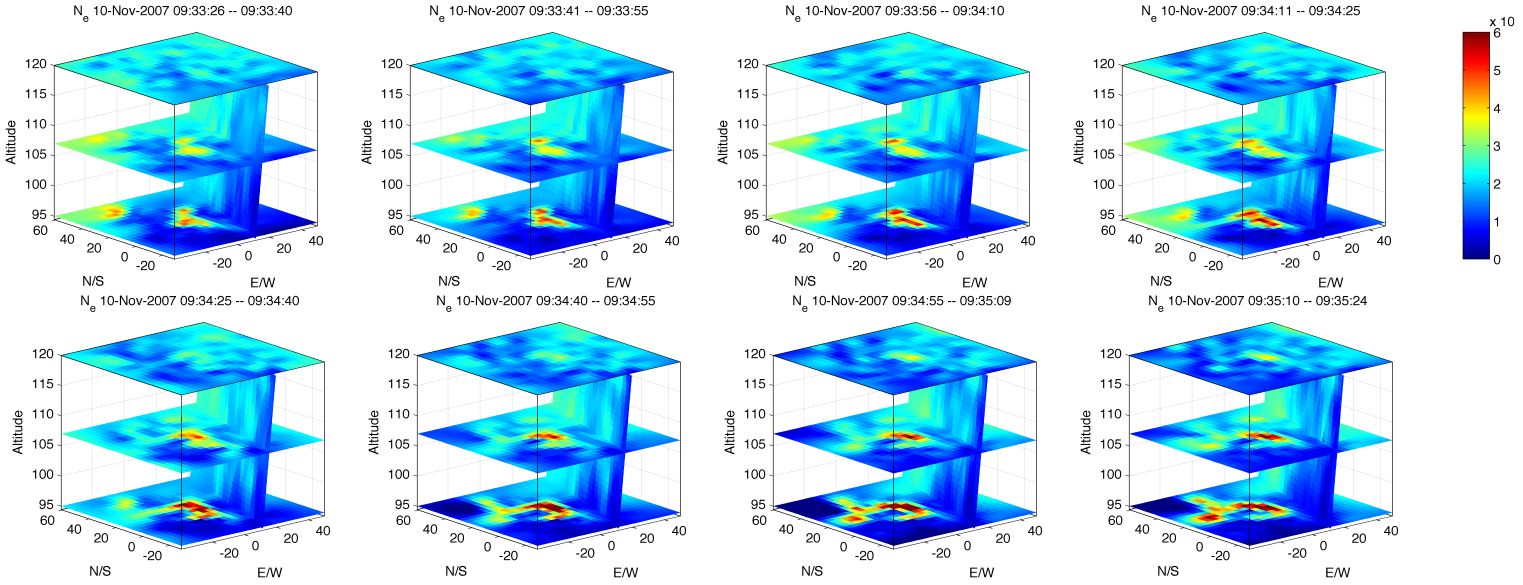


Figure 3·12: 10 November, 2007. 15-second reconstructions. Universal kriging.

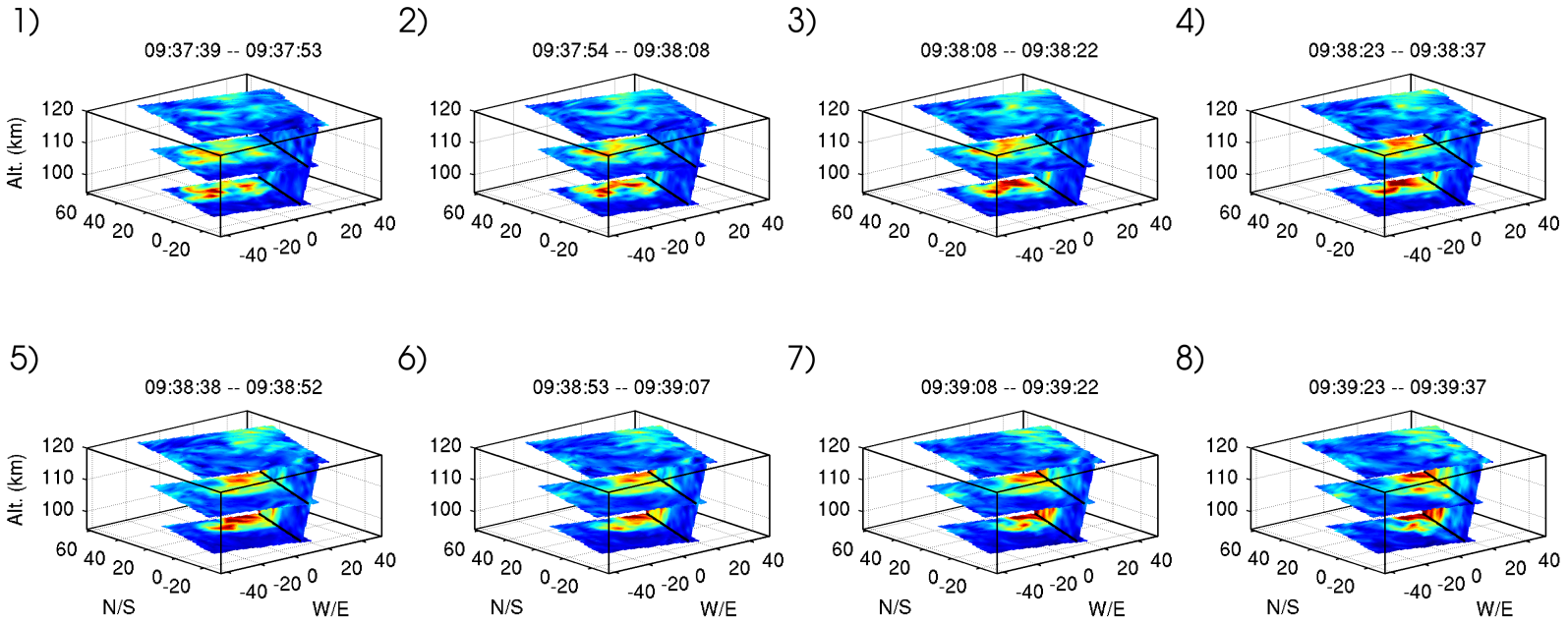


Figure 3·13: 10 November, 2007. 15-second reconstructions. Trilinear interpolation.

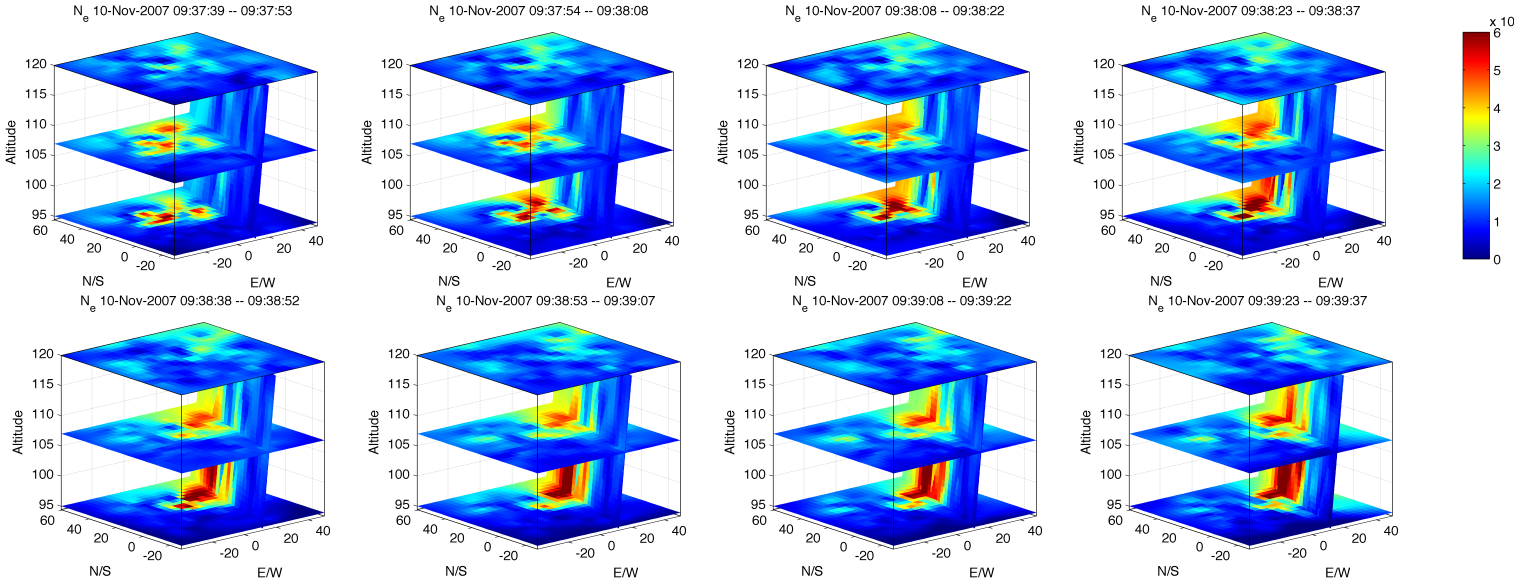


Figure 3·14: 10 November, 2007. 15-second reconstructions. Universal kriging.

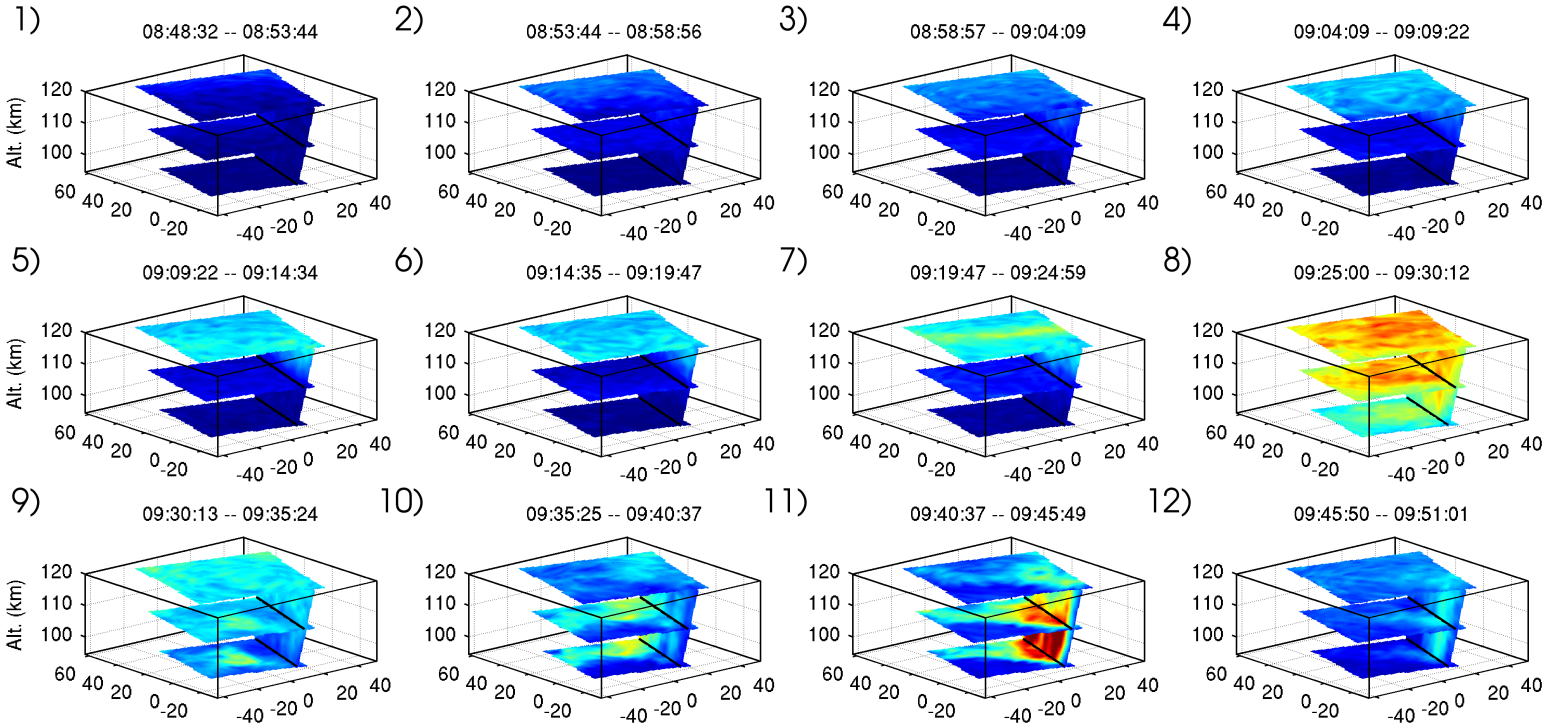
Switching from high time resolution to high data fidelity, Figure 3·14 shows the results after integrating for five minutes:

08:48–09:30 A steady increase in N_e at 120 km altitude. (Primary electron energies are ~ 2 keV. The increase appears to occur simultaneously over a large region.

09:30–09:45 The ionization at 120 km decreases. Meanwhile, a 100 km to 110 km, at structured enhancement appears.

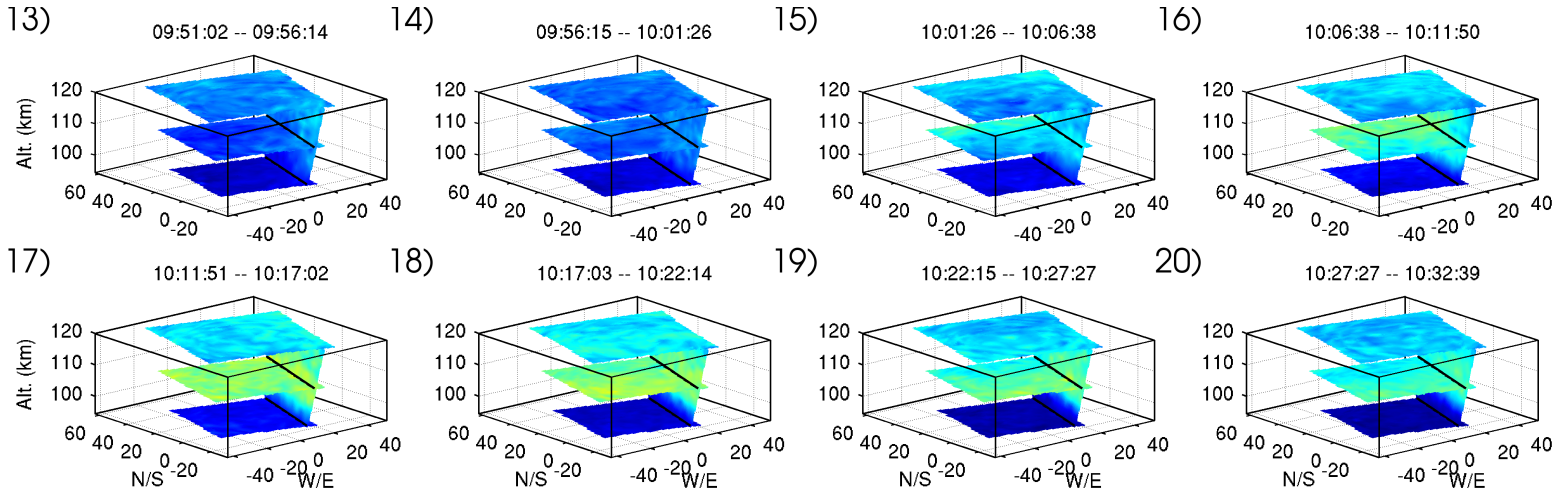
09:45–09:51 The lower structure abates. (Electrons precipitating below 100 km primarily have energies > 20 keV.)

10:01–10:32 Electron density peaks around 107 km. (Electrons ~ 10 keV.)



(a) Electron density. 10 November, 2007. Integration time $5\,\mathrm{min}$.

Figure 3·15: Trilinear interpolation



(b) Electron density. 10 November, 2007. Integration time $5\,\mathrm{min}$.

Figure 3·14: (continued) Trilinear interpolation

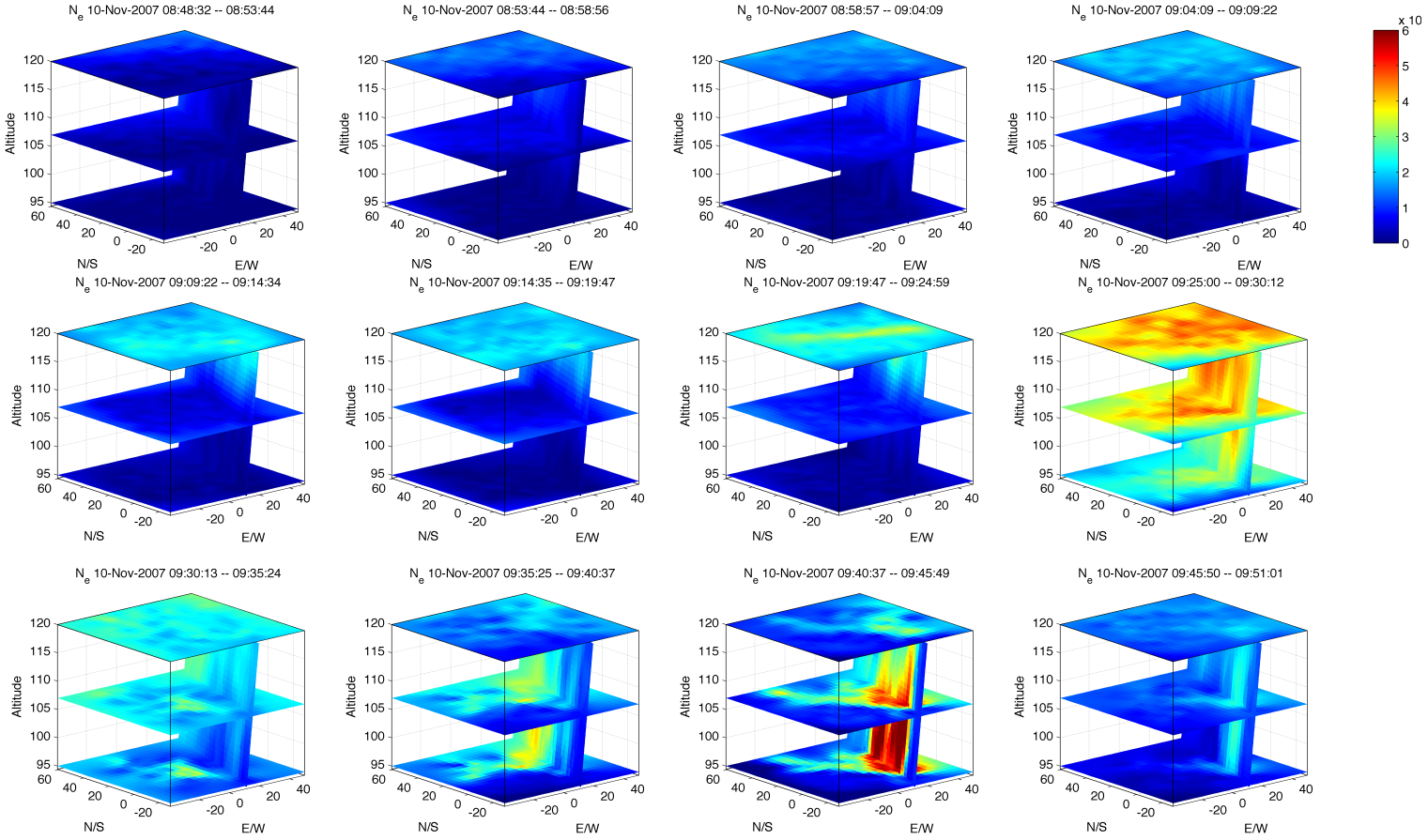


Figure 3·15: 10 November, 2007. Integration time: 5 minutes. Universal kriging.

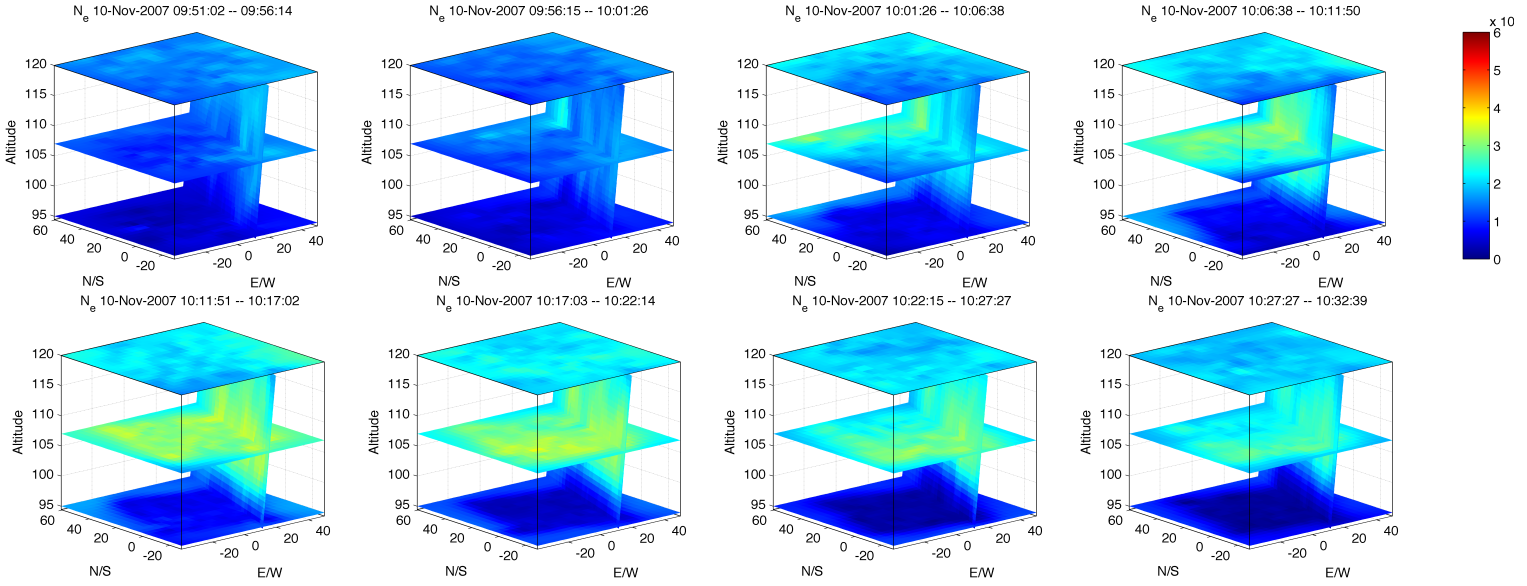
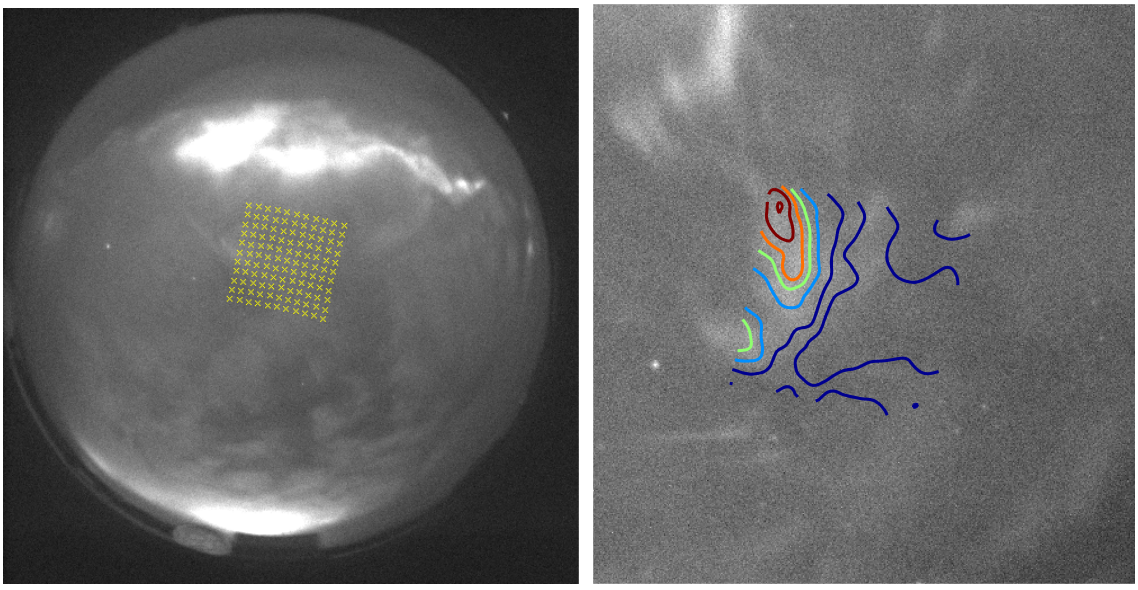


Figure 3·16: (Continued) 10 November, 2007. Integration time: 5 minutes. Universal kriging.



(a) PFISR field of view.

(b) Predicted luminence.

Figure 3·17: Coregistered data from PFISR and a digital all-sky camera (DASC). (a) A frame from the DASC with an auroral arc (top) traveling south during a substorm. The radar's 11×11 beam grid is projected onto the dome of the sky. (b) Detail during a substorm. The lines represent the integrated ion production along the line of sight of both instruments. Auroral luminence is also proportional to this quantity. The correspondence of the two instruments.

3.3.2 Radar-optical comparison

This experiment is also supported by optical data from a nearby *digital all-sky camera* (DASC) recording white light with a 10 s cadence. The camera and radar are essentially collocated. In Figure 3·17a, the grid of beam directions is projected onto the field of view of the all-sky camera. An auroral arc can be seen moving equatorward from the top of the frame, just outside of PFISR's field of view.

The two instruments can measure the same quantity. The DASC detects the aggregate of photon emissions falling upon its sensor. This is proportional to the rate of ion production integrated along the line of sight (Semeter and Doe, 2002). Due to the high rate of collisions in the E-region, the plasma continuity equation is kept in an approximate steady state, such that production equals loss. Plasma loss occurs through chemical recombination at the rate $\alpha N_i N_e = \alpha n_e^2$. In the auroral ionosphere, Vickrey et al. (1982) gives this recombination coefficient as $\alpha = 2.5 \times 10^{-6} \exp(-z/51.2) \mathrm{cm}^3\,\mathrm{s}^{-1}$, and (Semeter and Kamalabadi, 2005) explore the range of validity for this approximation.

Assuming the camera and radar are in the same polar coordinate system, such that a single PFISR beam is approximately equivalent to a group of optical pixels, then (to within a constant K) the optical brightness ϵ can be estimated from N_e by integrating over range:

$$\hat{\epsilon} \sim \int_0^\infty \alpha N_e^2 dr \tag{3.14}$$

Figure 3·17b shows the correspondence of these two measurements.

3.4 Exploiting spatial redundancy

Because a monostatic radar relies on a time sequence of 1D point measurements, spatial information is acutally inferred from temporal properties of the received signal. In radar, space = time. Namely, the *range* of the target is $r = \frac{c\tau}{2}$, where τ is the *time delay* of a pulse from the transmitter to receiver. Designing a radar system, the following factors represent design constraints: Maximum range = $\frac{cIPP}{2}$, Range resolution = $\frac{cT}{2}$. IPP is the inter-pulse period, the time between which a (monostatic) radar waits before sending its next pulse. T is the length (in μsec of the pulse, and c is the speed of light.

In ISR, there is a four-way tradeoff between (1) spatial context or resolution, (2) temporal resolution, (3) spectral resolution or lag extent, and (4) data fidelity. But the rapid-scanning multi-beam experiment is not simply a multiplicity of single-beam experiments. There may be a gestalt advantage to gathering data in this way, rather than as a scan. Whether it outweighs the reduction in temporal resolution or fidelity will depend on the user's needs.

Space-time ambiguity and spatial context

Broadly speaking, precision and uncertainty are reciprocal concepts, and there generally exists a tradeoff between pairs of properties that can be said to be duals in this sense. For instance, pulse shaping is a strategy for enhancing range resolution, but at the expense of bandwidth (and thus noise power). This precision/uncertainty tradeoff is characteristic of choosing a resolution in experiment design (see, e.g. Menke, 1989).

Re-pointing the radar beam presents the experimenter with one or two additional spatial dimensions and improved *spatial context* (read: "resolution," in that the instrument now resolves more than a single point). From the point of view of the monostatic radar, however, the data are a

1D time series. Spatial context can be *gained*, but only at the *expense* of temporal resolution. [5]

Consider scanning a dish antenna. There is a practical limit to how quickly the dish can be repointed, owing to its inertia. While scanning, it continues to operate in its usual mode, accumulating measurements to form a reliable estimate of the plasma ACF. This internal accumulation amounts to destructive (non-reversible) error, merging (at least for the duration of the accumulation) distinct targets, i.e. smearing.

Of course, performing such a scan conveys a benefit over single-direction ISR: the scanning dish can distinguish (1) a process that is dynamic (in intensity) but stationary (in position) and filling the beam from (2) a localized structure that is static (in intensity) but crossing the beam. To the rigid antenna, both appear as processes in time. This is the classic *space-time ambiguity* problem.

However, the speed of the scan is limited by the inertia of the dish antenna. As the beam traverses the region of interest (at such a speed and integration time to minimize smear), an image emerges of the observed process along the path of the scan. However, as **Figure 2** demonstrates, there is now also a different type of ambiguity, though its impact is mitigated by its low likelihood. Figure 2 depicts the pathological scenario of a localized, beam-filling structure convecting at the same angular speed as the radar's scan. To the radar, this scenario is indistinguishable from an unmoving, broadened feature. Though we gain greater spatial context through scanning, we haven't escaped spatio-temporal ambiguity.

Now consider a radar capable of pulse-by-pulse steering. Rather than dwell in one position and gather a statistically significant sample, the beam loops through a pre-programmed sequence of directions. Although certainly not free of similar ambiguities, the greater spatial context (and now "simultaneously" measured) at least leaves open the possibility of resolving such. Of course, the cost for spatial context is temporal resolution! If the beam cycles through N positions per frame, the process is observed with only $1/N$ times the sampling frequency, and dynamics may not be adequately captured. Alternatively, if the experimenter can afford to "dial down" the integration time, precision can be traded for resolution in both space and time. Indeed, Semeter et al. (2008) analyzed 3D structures at a cadence of 14.6 s, corresponding to an extremely low 48 pulses/integration/beam. And yet the structures were consistent with their longer-integration counterparts! Chapter 3 discusses that example, among others.

[5] This tradeoff of spatial and temporal resolution is analogous to the interlaced sampling in television systems, except that the experimenter has some control of the sampling pattern. Jain (1989) describes interlaced sampling.

3.5 Conclusions

PFISR is the first electronically steerable ISR dedicated to ionospheric study. This chapter demonstrates the capabilities of PFISR for producing three-dimensional volumetric images of the ionospheric *E*-region during auroral activity. The volumetric data were acquired using a square array of 11×11 beams. A phase-coded pulse was used which provided ~ 1.5 km range resolution. The output from the demodulator was converted from backscattered power to electron density. The resulting 3D images were quantitatively compared with all-sky white-light camera observations through an ion continuity equation, demonstrating good agreement.

The time taken to cycle through beam pattern places a practical limit on the temporal resolution. In this arrangement, PFISR can capture ~ 1.6 frames/s, which corresponds to 48 pulses/angle, yielding uncertainties of $\sim 14\,\%$. The efficacy of this mode for addressing time-dependent studies of magnetosphere-ionosphere interactions is discussed.

Chapter 4

Velocity field imaging: F-region bulk plasma drift

A thousand pictures can be drawn from one word
Only who is the artist
We gotta agree

"I'm Just a Singer (in a Rock and Roll Band)"
The Moody Blues

Perhaps for the 1960s rock group The Moody Blues, it was simply a clever inversion of a familiar adage, but the above quote aptly describes the role of modeling in inverse theory. An underdetermined problem is one for which, roughly speaking, the data space is smaller than the model (or parameter) space. That is, a single "word" of data corresponds to innumerable possible pictures of the reality captured by the observation. Selecting from among these possibilities is the art of solving inverse problems. The mathematics (ordinary least squares, maximum likelihood, etc.) are only part of the answer. The solution may still be meaningless without a conscious effort to identify the the "artist," that is the model that generated that picture.

Pulse-by-pulse beam steering provides experimenters additional flexibility in the form of an additional dimension in which to trade off temporal resolution versus spatial context, making observable small-scale spatial variability in ionospheric structure, while also capturing the dynamics of ionospheric processes. In this chapter, we investigate an inverse-theoretic approach to predicting *F*-region flow fields from a monostatic electronically-steerable ISR. First, we compare two predictors of velocity field. Then we explore two case studies.

The principal application is the study of substorms through two concomitant phenomena: dynamic auroral activity and local variations in ionospheric flow. Although the basic plasma physics describing these effects is well-developed, their relation to one another is poorly understood (namely, their causitive order and their linkage through the greater near-Earth space environment, including the magnetosphere and the solar wind). The techniques developed here may help clarify that connection.

Background

In the previous chapter, we examined the capability of a monostatic radar to resolve (in space and time) various plasma parameters derived from the ISR spectrum. This chapter focuses on the bulk Doppler shift of the spectrum, which provides a measure of the bulk ion flow (more precisely, that component of the ion flow lying along the line-of-sight of the radar beam). Strictly speaking, a monostatic radar can only observe the component of velocity lying along its *line-of-sight* (LOS). (A multistatic configuration, by which the target is observed by more than a single transmitter/receiver simultaneously, is needed to resolve more than one component in truth.) But rapid electronic beam steering has a slight advantage: each pre-programmed beam direction has a slightly different view of the target. By combining these views, and by making some spatio-temporal regularity assumptions, it is possible to reconstruct the underlying vector velocity field. A similar trick is demonstrated by Hagfors and Behnke (1974) at Arecibo Observatory, recovering a three-dimensional velocity vectors by continuously scanning the beam in azimuth for twenty minutes. Doupnik et al. (1977) include a physical model of ionospheric velocity to estimate the electric field vector. Sulzer et al. (2005) introduced linear regularization to deal with rapid variations within the scanning time of the antenna.

Despite these advances in processing, spatial and temporal resolution are ultimately limited by the hardware. In the time required to steer a heavy dish antenna, details of the most dynamic events in the ionosphere will have been smeared across its scanning region. In this chapter, we use the 3D "snapshot" mode of PFISR to investigate its ability to resolve localized flow variations. We now focus our attention on ionospheric events associated with such flow variations.

Substorms

Magnetospheric *substorms* are regularly occurring, often violent, disturbances of Earth's magnetosphere that frequently affect plasma convection patterns in the ionosphere. Though originally defined and classified by their more readily visible effects (the expansion of the auroral oval, followed by spectacular discrete auroral arcs) (Akasofu, 1964), substorms are now understood to be caused by the impulsive dissipation of free energy from the magnetosphere to the ionosphere (Rostoker et al., 1980; Rostoker, 1999). Although the exact triggering mechanism is not clearly understood (Zhu et al., 2009; Lyons et al., 2009), tremendous effort is spent studying the flow of energy in and from the magnetosphere (Angelopoulos et al., 2008).

A simple begins with the magnetic field of the Sun, bound by the solar wind and traveling

Earthward, which merges with the Earth's own magnetic field (GMF) on the day side and convects these now-open field lines to the night side. Thus, the magnetosphere is compressed on the day side and elongated on the night side into a magnetotail. Under increasing magnetic stress deep in the magnetotail, the field lines reconnect, resulting in sudden particle acceleration toward Earth, hot plasma injection into the ionosphere, and the well-documented auroral oval expansion (Schunk and Nagy, 2009, Chapter 12).

In addition to enhanced auroral emissions and particle precipitation, substorms are associated with global-scale electrical currents and localized regions of enhanced electric field. Convective disturbances can be a direct result of these electric field enhancements (Bristow and Jensen, 2007). Optical aurorae, then, are a secondary effect, a response of the ionospheric plasma to currents arising from these flow disturbances. And yet it is the auroral morphologies that define the canonical substorm phases. The physical processes connecting these two phenomena remain poorly understood, due partly to inadequate observation.

This chapter explores the imaging of local flow disturbances in the high-latitude *F*-region using ISR-derived measurements of LOS velocity. The image reconstruction is based on linear inverse theory. We analyze the limitations of this type of reconstruction and present two case studies. In Section 4.1, we describe the measurement process, observation geometry, and important assumptions. In Sections 4.2 & 4.3 we detail how we exploit the rapid scanning capability of PFISR to generate a two-dimensional "snapshot" of ion flow patterns. The accuracy of these techniques is evaluated in Section 4.4. We then present case studies (Section 4.5) showing some of the features demonstrated in the preceding analysis. Section 4.6 presents a summary of findings and suggested extensions.

4.1 Methodology

We frame the problem of velocity field prediction as a linear discrete inverse problem. That is, given a forward model mapping the underlying field $\mathbf{v}(x,y)$ to a set of independent *line-of-sight* (LOS) measurements $\underline{v}_{\text{los}}$, we develop an inverse model to evaluate $\widehat{\mathbf{v}}(x,y)$, a predictor of the original field $\mathbf{v}(x,y)$.

Phased-array radar experiments generally involve an arbitrary number and arrangement of beams. Thus the problem of predicting velocity components from projections may be overdetermined (equations outnumber the unknowns). Heinselman and Nicolls (2008) predict using linear least squares, which handles the overdetermined problem gracefully, making use of appropriate

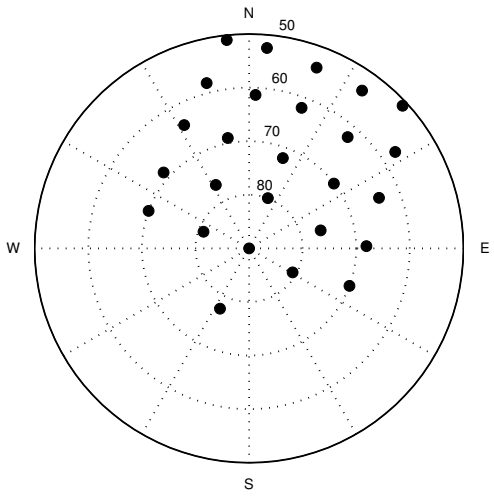


Figure 4·1: The 26-beam configuration used in the experiment of Section 4.5.

data while respecting the limited rank of the forward operator. The authors generate time sequences of velocity vectors resolved along magnetic latitude, under the assumption that flows are ordered in that dimension. (This is a standard assumption in ISR analysis of high-latitude convection, and is useful for scanning dish antennas.) Instead, we envision the radar measurement process as a direct three-dimensional acquisition, generating a "snapshot" of the entire fov near-simultaneously. Our measurement model is overdetermined, and we use regularization to impose physical constraints on the solution. The technique described here is general in that it can be applied to any beam sequence and the prediction evaluated on an arbitrary grid.

Observation geometry

To illustrate, let us focus on a beam sequence particular to the PFISR experiments in this chapter. Volumetric data were acquired using a grid of 26 beam positions (in a 5×5 grid with one addtional beam in the direction of the magnetic fieldline. At 350 km altitude, the sampled angular space is approximately rectangular and subtends a 300 km $\times$ 250 km region.

Each data point in Figure 4·2 is the midpoint of a range gate. Assume that a nonlinear fitting procedure has assigned estimates of ISR plasma parameters to each point. Among these is v_{los}, the *line-of-sight* component of bulk ion drift. We select those data ranging in altitude from 200 km to

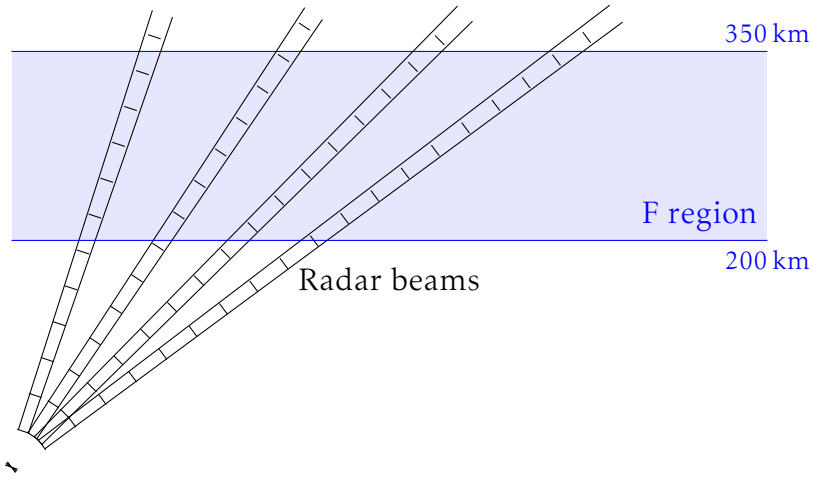
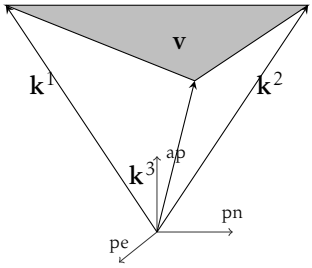


Figure 4·2: Range gates along beams. (Side view.)



$$
\begin{aligned}
v^1_{los} &= k^1_{pe} v_{pe} + k^1_{pn} v_{pn} + k^1_{ap} v_{ap} \\
v^2_{los} &= k^2_{pe} v_{pe} + k^2_{pn} v_{pn} + k^2_{ap} v_{ap} \\
v^3_{los} &= k^3_{pe} v_{pe} + k^3_{pn} v_{pn} + k^3_{ap} v_{ap}
\end{aligned}
$$

Figure 4·3: A simple example. A uniform velocity vector is projected onto three lines-of-sight. In this full-rank system, as long as $\mathbf{k}^1 \ldots \mathbf{k}^3$ are unique, we can recover all three components of $\mathbf{v}$ from the measurements $v^1_{\text{los}} \ldots v^3_{\text{los}}$.

350 km, where the vertical component of $\mathbf{v}$ is considered negligible.[1] Then the entire set of sample points is collapsed onto a horizontal plane. In this range of altitudes, $\mathbf{E}$ (and thus $\mathbf{v}$) maps directly up the field line, so nothing is lost.

Forward model

Each range-gated ACF yields an independent measurement of v_{los}, the bulk ion velocity projected along the direction of the beam, given by

$$v_{\text{los}} = \tilde{k} \cdot \tilde{v}, \tag{4.1}$$

[1] Although significant ion upwelling may occur, field-aligned velocities in this range are $< 200\,\text{m}\,\text{s}^{-1}$ (Wahlund et al., 1992; Semeter et al., 2003; Zettergren et al., 2007), while convective flows are typically in the $\text{km}\,\text{s}^{-1}$ range (Whalen et al., 1974; Fujii et al., 2002).

$\tilde{v} = \begin{bmatrix} v_e, & v_n, & v_z \end{bmatrix}^T$ is the bulk ion velocity within the measurement volume (the tildes signify a radar-centered geodetic coordinate system). $\tilde{k}$ is a unit vector defined in terms of direction cosines

$$\tilde{k} = \begin{bmatrix} k_e \\ k_n \\ k_z \end{bmatrix} = \begin{bmatrix} \cos\alpha \\ \cos\beta \\ \cos\gamma \end{bmatrix} = \begin{bmatrix} x/R \\ y/R \\ z/R \end{bmatrix},$$

where x, y, & z are the distances east, north, and vertically from the radar, respectively. and $R = \sqrt{x^2 + y^2 + z^2}$ is the range to a given measurement point from the radar. For high elevation angles, the Earth's curvature is negligible and

$$\tilde{k} = \begin{bmatrix} k_e \\ k_n \\ k_z \end{bmatrix} = \begin{bmatrix} \cos\theta \sin\phi \\ \cos\theta \cos\phi \\ \sin\theta \end{bmatrix},$$

where θ is elevation and ϕ is azimuth (measured east from north).

In the F-region above ~ 150 km, neutral collisions have less influence, and an ion encountering electric field **E** and magnetic field **B** experiences guiding center drift velocity

$$\mathbf{v} = \frac{\mathbf{E} \times \mathbf{B}}{B^2}. \tag{4.2}$$

Assume both **E** and **B** are constant along a magnetic field line from 150 km to the maximum range of the radar (~ 400 km). The natural geometry for this problem, then, is the geomagnetic reference frame defined by **B**. For our purposes, this is a simple rotation from radar-centered geographic coordinates according to local magnetic inclination (or dip) I and declination δ. I.e., a rotation matrix is applied to $\tilde{k}$,

$$\begin{aligned} \mathbf{k} = \begin{bmatrix} k_{pe} \\ k_{pn} \\ k_{ap} \end{bmatrix} &= \begin{bmatrix} \cos\delta & -\sin\delta & 0 \\ \sin I \sin\delta & \sin I \cos\delta & \cos I \\ -\cos I \sin\delta & -\cos I \cos\delta & \sin I \end{bmatrix} \begin{bmatrix} k_e \\ k_n \\ k_z \end{bmatrix} \\ &= \mathbf{R}_{\text{geo}\to\text{gmag}} \tilde{k}, \end{aligned}$$

and its inverse (transverse) will be applied later to plot the solution in geographic coordinates. The subscripts pe, pn, and ap stand for perpendicular-east, perpendicular-north, and anti-parallel (since **B** is directed downward in the northern hemisphere).

A single LOS projection, as in equation (4.1) does not provide enough information to resolve the vector velocity; more measurements are needed. Consider the simple example in Figure 4·3. The three measurements $\left\{v_{los}^i\right\}$ are projections of a uniform velocity **v** onto three unique beam

directions. This threefold projection can be expressed in matrix form as

$$\underline{v}_{\text{los}} = \begin{bmatrix} v^1_{\text{los}} \\ v^2_{\text{los}} \\ v^3_{\text{los}} \end{bmatrix} = \begin{bmatrix} k^1_{\text{pe}} & k^1_{\text{pn}} & k^1_{\text{ap}} \\ k^2_{\text{pe}} & k^2_{\text{pn}} & k^2_{\text{ap}} \\ k^3_{\text{pe}} & k^3_{\text{pn}} & k^3_{\text{ap}} \end{bmatrix} \begin{bmatrix} v_{\text{pe}} \\ v_{\text{pn}} \\ v_{\text{ap}} \end{bmatrix} = \mathbf{Av}, \tag{4.3}$$

i.e. by stacking the corresponding projection vectors into a projection matrix **A**. If **A** is not singular, this could be solved by direct matrix inversion. However, for any pair of k^i's sufficiently similar, **A** becomes nearly singular. This especially becomes a problem in the presence of noise, as propagation of error is compounded. Secondly, (4.3) does not allow the inclusion of even one additional measurement. The problem becomes overdetermined and in the worst case a solution does not exist.

In this simplified example, all three beams measure a uniform velocity **v**. Since we are interested in resolving the spatial variability of the velocity field $\mathbf{v}(x,y)$, we do not assume uniformity among all the measurements in a given frame. Instead, the matrix **A** must be expanded to include multiple, spatially distributed vectors $\mathbf{v}(x,y)$. Already, we run up against a limitation of algebraically inverting the projection operation, since this expanded **A** is not necessarily a full-rank matrix. The inversion may be either overdetermined or underdetermined, and the technique we apply must handle either case.

4.2 Inversion 1—Overlapping pixels

The first such expansion of **A** is formulated by repeating a sequence of discrete inversions in different bins. This method is described by Semeter et al. (2010), and is a two-dimensional extension of that described by Heinselman and Nicolls (2008). Since the magnetic field acts as a perfect conductor, we may assume constant horizontal velocity along the magnetic field line. Furthermore, the flow field should be somewhat smooth, so neighboring measurements represent somewhat similar velocities. So we collapse the flow field to a horizontal plane in geomagnetic coordinates.

Measurement samples were selected in the altitude range from 150 km to 400 km and binned into a 4×4 grid of pixels (see Figure 4.4). Pixel boundaries are defined by considering the total horizontal extent of the data points. Each is approximately $100\,\text{km}^2 \times 100\,\text{km}^2$. Each pixel shares 50% of its area in either direction with its nearest neighbors. This imposes correlation between neighboring pixels, and is equivalent to a spatial smoothness constraint.

The 4×4 pixelization satisfies a trade-off between spatial resolution and the amount of independent information contained in each pixel. Although we could choose a finer sampling, each

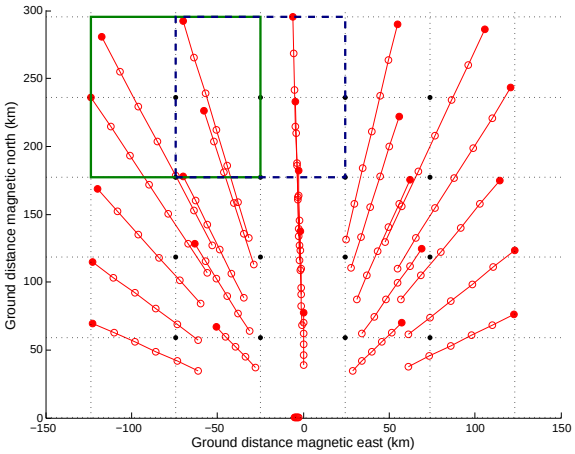


Figure 4·4: The "overlapping pixels" predictor described by Semeter et al. (2010) uses overlapping pixels. Circles represent range gates on each beam, and the filled circle is the last gate selected for that beam. The colored boxes identify two neighboring pixels. Black dots indicate pixel centers. cf. Figure ??.

pixel should contain data from approximately three beams.

For each pixel, we solve a separate discrete inverse problem. Assuming uniform flow $\mathbf{v}$ within each pixel, the forward model describing the projection of $\mathbf{v}$ onto M lines-of-sight is formed, not unlike equation (4.3), by stacking the corresponding projection vectors:

$$\underline{v}_{\text{los}} = \begin{bmatrix} v^1_{\text{los}} \\ v^2_{\text{los}} \\ \vdots \\ v^M_{\text{los}} \end{bmatrix} = \begin{bmatrix} k^{1\mathsf{T}} \\ k^{2\mathsf{T}} \\ \vdots \\ k^{M\mathsf{T}} \end{bmatrix} \mathbf{v} + \begin{bmatrix} e^1_{\text{los}} \\ e^2_{\text{los}} \\ \vdots \\ e^N_{\text{los}} \end{bmatrix} \tag{4.4}$$

$$= \mathbf{A}_{\text{pixel}}\mathbf{v} + \underline{e}_{\text{los}}, \tag{4.5}$$

where now $\underline{e}_{\text{los}}$ represents the random perturbations inherent in the measurement process. We will assume this is a zero-mean Gaussian with covariance matrix Σ_e, the diagonal elements of which are provided by the ISR fitter.

Note from the figure that most pixels include multiple measurements from a given beam. Although in the absence of noise, these projections onto the same $\mathbf{k}$ would provide no additional information over a single measurement[2], here they all contribute to the solution, serving to reduce statistical uncertainties. This is important given the ill-conditioned nature of the inversion.

[2] Equivalently, the projection matrix in equation (4.5) contains linearly dependent rows, i.e. identical $\mathbf{k}$ vectors.

The least squares solution, for each pixel, is a well-known result (Tarantola, 2005, e.g.):

$$\hat{\mathbf{v}} = \Sigma_{\mathbf{v}}\mathbf{A}^{\mathsf{T}}\left(\mathbf{A}\Sigma_{\mathbf{v}}\mathbf{A}^{\mathsf{T}} + \Sigma_{e}\right)^{-1}\underline{v}_{\text{los}}, \tag{4.6}$$

with assosciated error covariance

$$\Sigma_{\hat{\mathbf{v}}} = \Sigma_{\mathbf{v}}\mathbf{A}^{\mathsf{T}}\left(\mathbf{A}\Sigma_{\mathbf{v}}\mathbf{A}^{\mathsf{T}}\right)^{-1}\mathbf{A}\Sigma_{\mathbf{v}}. \tag{4.7}$$

where $\Sigma_{\mathbf{v}}$ can be interpreted as a prior constraint: that $\mathbf{v}$ is a zero-mean Gaussian r.v. with covariance matrix $\Sigma_{\mathbf{v}}$. This is the same predictor used by Heinselman and Nicolls (2008) for F-region drifts.

The overlapping pixels predictor is an ad hoc extension of of Heinselman and Nicolls (2008) to two dimensions. Although it illustrates the capability of resolving vector velocities from LOS projections, it has limited flexibility regarding recovery regions (or pixels). This method depends on two stages of smoothness assumptions: first that the velocity is uniform within a given pixel, and second that the measurements in overlapping regions are reasonably consistent. Formula (4.6) is evaluated in each pixel independently, with the understanding that the second smoothness assumption will likely be violated. This method is explored in more detail in Section 4.4

4.3 Inversion 2—Tikhonov regularization

To take better advantage of the correlations between neighboring measurements, this method is framed more rigorously in the context of inverse theory. Rather than solving independent problems in each pixel, we consider the unknown velocity field $\mathbf{v}(x,y)$ a latent process, and construct a forward model mapping the $\mathbf{v}$ to the measurements (the set of LOS projections $\underline{v}_{\text{los}}$). Discretization is handled explicitly and separate from the inversion, providing greater flexibility by offering a choice of reconstruction basis functions. A spatial smoothness constraint is physically justified and implemented in a classic Tikhonov regularization framework.

Discretization

Although the LOS measurements are inherently discrete, we assume an underlying continuous velocity field $\mathbf{v}(x,y)$. For implementation in a computer, this can be discretized spatially and regarded as a column vector, i.e.

$$\mathbf{v}(x,y) = \sum_{j=1}^{N} \mathbf{v}_j b_j(x,y),$$

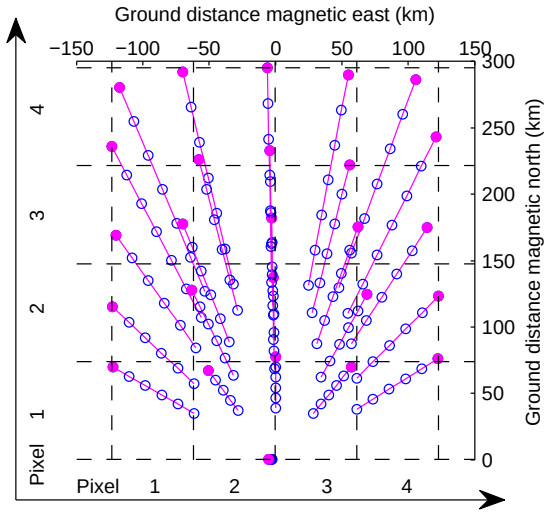


Figure 4·5: Pixelization for Tikhonov-regularized predictor.

where $\left\{b_j(x,y)\right\}_{j=1}^{N}$ is some basis spanning the region of interest. This can be simple rectangular pixels, or something more elaborate such as a multiscale or sparse basis. In practice, each component of the vector $\mathbf{v}$ is discretized independently, and the resulting column vectors stacked so that

$$\underline{\mathbf{v}} = \left[v_{pe}^1,\ldots,v_{\mathrm{pe}}^N,v_{\mathrm{pn}}^1,\ldots,v_{\mathrm{pn}}^N,v_{\mathrm{ap}}^1,\ldots,v_{\mathrm{ap}}^N\right]^{\mathsf{T}}.$$

For illustration, we use the intuitive rectangular pixel basis. Because the beams in Figure 4·1 are roughly aligned with the magnetic meridian, the data points are first rotated to geomagnetic coordinates so that they align with pixels. Once again, the 4×4 grid of Figure 4·5 is an attempt to balance the compromise between spatial resolution and information content within each pixel. Although we could choose a finer sampling, or a non-uniform one, the goal is for each pixel to contain data from approximately three beams in order to approach observability.

In general, the velocity field $\mathbf{v}(\mathsf{x},\mathsf{y})$ is divided into N pixels and we observe M LOS projections. This set of projections is expressed in the $M\times3N$ matrix

$$\mathbf{A}_{\mathrm{Tik}} = \left[\begin{array}{ccc|ccc|ccc} k_{pe}^1 & \cdots & \cdots & k_{\mathrm{pn}}^1 & \cdots & \cdots & k_{\mathrm{ap}}^1 & \cdots & \cdots \\ \cdots & k_{\mathrm{pe}}^2 & \cdots & \cdots & k_{\mathrm{pn}}^2 & \cdots & \cdots & k_{\mathrm{ap}}^2 & \cdots \\ & \vdots & & & \vdots & & & \vdots & \\ \cdots & \cdots & k_{\mathrm{pe}}^M & \cdots & \cdots & k_{\mathrm{pn}}^M & \cdots & \cdots & k_{\mathrm{ap}}^M \end{array}\right].$$

That is, each row contains exactly three nonzero elements mapping the velocity vector in the jth

pixel to the ith LOS measurement such that (by analogy to (4.3))

$$\underline{v}_{\text{los}} = \mathbf{A}_{\text{Tik}}\underline{\mathbf{v}} + \underline{e}. \tag{4.8}$$

Clearly, for an arbitrary choice of beams and reconstruction grid, this matrix is not directly invertible. This constraint will be included as a part of the inverse model.

Inversion

The forward model (4.8) consists of four terms: the LOS observations ($\underline{v}_{\text{los}}$), a velocity field ($\underline{\mathbf{v}}$), the operator ($\mathbf{A}$) mapping one to the other, and random additive noise $\underline{e}$. An inverse model is constructed from these elements and then applied to the observations to recover the underlying field. Several approaches have been developed, and we proceed with the classic method of Tikhonov regularization.

Using results developed elsewhere, the generalized Tikhonov predictor that solves equation (4.8) (and introduces a side constraint $\mathbf{L}$) is

$$\hat{\underline{\mathbf{v}}} = \left(\mathbf{A}^{\mathsf{T}}\Sigma_e^{-1}\mathbf{A} + \alpha^{-2}\left(\mathbf{L}^{\mathsf{T}}\Sigma_{\mathbf{v}}^{-1}\mathbf{L}\right)^{-1}\right)^{-1}\mathbf{A}^{\mathsf{T}}\Sigma_e^{-1}\underline{v}_{\text{los}} \tag{4.9}$$

with error covariance

$$\widehat{\Sigma}_{\mathbf{v}} = \left(\mathbf{A}^{\mathsf{T}}\Sigma_e^{-1}\mathbf{A} + \mathbf{L}^{\mathsf{T}}\Sigma_{\mathbf{v}}^{-1}\mathbf{L}\right)^{-1}, \tag{4.10}$$

where Σ_e is the covariance matrix of $\underline{e}$, and $\Sigma_{\mathbf{v}}$ is the covariance of a zero-mean Gaussian random vector $\underline{\mathbf{v}}$, representing exogenous information, in this case the a priori probabilistic characterization of the quantity we wish to predict.

The side constraint (encoded in the matrix $\mathbf{L}$) is the other component of prior information. Common choices are $\mathbf{L} = \mathbf{I}$ (equivalent to penalizing large-norm solutions, or $\mathbf{L}$ = a first derivative, to enforce smoothness. To choose a constraint for ionospheric drift, the divergence operator seems a natural fit. I.e., the ionosphere is incompressible ($\nabla \cdot \mathbf{v} = 0$), and this constraint can be expressed in the Tikhonov formulation (4.9) and (4.10) through the matrix

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_{\text{pe}} & 0 & 0 \\ 0 & \mathbf{L}_{\text{pn}} & 0 \\ 0 & 0 & \mathbf{L}_{\text{ap}} \end{bmatrix}, \tag{4.11}$$

where the submatrices encode discrete approximations of the first derivatives, e.g.

$$\mathbf{L}_{\text{pn}} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & 0 & \dots & 0 \\ & & & \ddots & & & \\ 0 & \dots & 0 & 0 & -1 & 1 & 0 \\ 0 & \dots & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \end{bmatrix}.$$

Because the forward difference generates a shorter vector than its input, each submatrix has some all-zero rows. This makes the matrix $\mathbf{L}^T \Sigma_{\mathbf{v}}^{-1} \mathbf{L}$ degenerate; it has some zero eigenvalues corresponding to the boundaries. The boundary conditions need not be included in (4.9) since the data-fit term will select values from the observations. The smoothness constraint is only enforced for the perpendicular components. Since the field-aligned component is very small, we impose a constraint on the *magnitude* of this component rather than its smoothness, i.e. $\mathbf{L}_{\text{ap}} = \mathbf{I}$.

4.4 Simulation

The regularization parameter α is a non-negative factor that controls the relative influence of measured data and a priori information. Before applying our Tikhonov predictor to experimental data, it is important to evaluate how α affects the result. Since selecting the regularization parameter is typically a subjective process, an "optimal" value of α can be difficult to define. A single value is not likely to yield subjectively "optimal" results for all measurements. Nevertheless, in a controlled simulation, iterative methods of selection—whether semi-objective or utterly subjective (e.g. visual inspection)—can aid in finding a useful practical range of α.

In this section, a simulated flow-field is predicted using (4.9). We examine the effect of the incompressible flow constraint (4.11) and compare it to the norm-conserving $\mathbf{L} = \mathbf{I}$ predictor.

The side constraint $\mathbf{L}$ is absorbed into the precision matrix $\Sigma_{\mathbf{v}}^{-1}$. When $\mathbf{L} = \mathbf{I}$, the prior weighting term $\alpha \mathbf{L}^{\mathsf{T}} \Sigma_{\mathbf{v}}^{-1} \mathbf{L} = \alpha \Sigma_{\mathbf{v}}^{-1}$, i.e. the parameters of the prior model come down to choosing variances for each component of $\mathbf{v}$. Assuming the horizontal components are independent, let $\sigma_{\text{pe}} = \sigma_{\text{pn}} = 500\,\text{m/s}$, $\sigma_{\text{ap}} = 15\,\text{m/s}$. Also let Σ_e, the error covariance, be a diagonal matrix with variances inversely proportional to range squared, and with a scaling factor chosen so that the standard deviation is 10 m/s at 100 km.

The results in this section are specific to the phantom flow field and therefore do not comprise a general analysis of the Tikhonov prediction. Instead, they are meant to motivate the use of one Tikhonov constraint matrix $\mathbf{L}$ over the other for a class of process typically encountered in high-

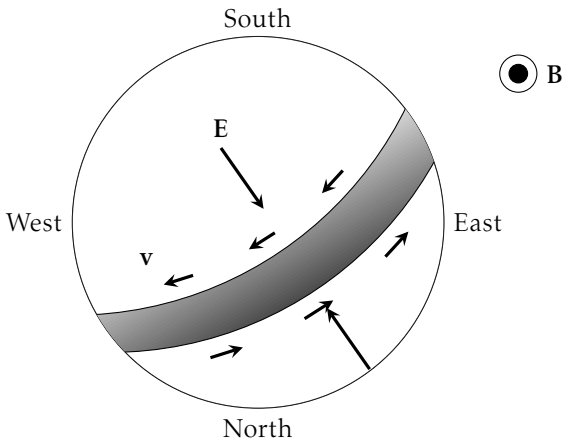


Figure 4·6: A model of plasma drift surrounding an ionization enhancement (e.g. an auroral arc). Within the enhanced region, the increased conductivity reduces electric field magnitude E (hence $\mathbf{v}$, the $\mathbf{E}\times\mathbf{B}$ drift). Meanwhile, $\mathbf{E}$ drives a polarization current within the arc, forcing charges to accumulate along the boundaries, and establishing a polarization field. The polarization field, in turn, results in $\mathbf{E}\times\mathbf{B}$ drift tangential to the arc boundary .
The diagram mimics the view upward from the ground facing north in the northern hemisphere, with $\mathbf{B}$ directed downward (out of the page).

latitude ISR research.

Flow shear simulation

We begin by simulating a velocity field morphology that is common during a substorm—namely, a flow shear along an active auroral boundary. (See Figure 4·6.) This pattern is motivated by auroral observations (de la Beaujardière et al., 1977; de la Beaujardière and Vondrak, 1982; Weber et al., 1991; Bahcivan et al., 2006). The auroral arc is a region of enhanced plasma density, and thus conductivity. Strong currents originating in the magnetosphere dominate the effects of the ambient electric field $\mathbf{E}$. It is possible, however, for $\mathbf{E}$ to drive a Hall current across the thin boundary of the arc. The rule of current continuity causes charges to accumulate on opposite sides of the arc, establishing a polarization field. The net effect in this case is a reduction of $\mathbf{E}$, thus a reduction of $\mathbf{E}\times\mathbf{B}$ drift within the arc. Likewise, the potential gradient *across* the arc boundary produces a secondary electric field such that plasma drift is parallel to the boundary.

Comparison of predictors

The light-colored arrows in Figures 4·7 and 4·8 represent the "ground truth" horizontal flows. Altogether, $\mathbf{v}(x_i, y_i)$, is divided into two regions: zero flow (within the arc) and uniform flow parallel to the arc boundary. The thick diagonal line signifies the arc boundary separating the two regions.

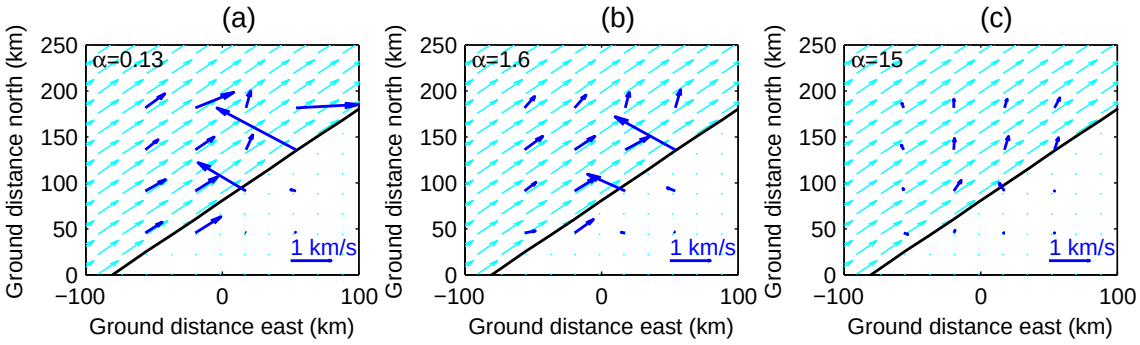


Figure 4·7: Method A (Field magnitude constraint). Simulated velocity field and predictions for three values of α. Light arrows represent the simulated velocity field. Dark arrows indicate the predicted field for each pixel.

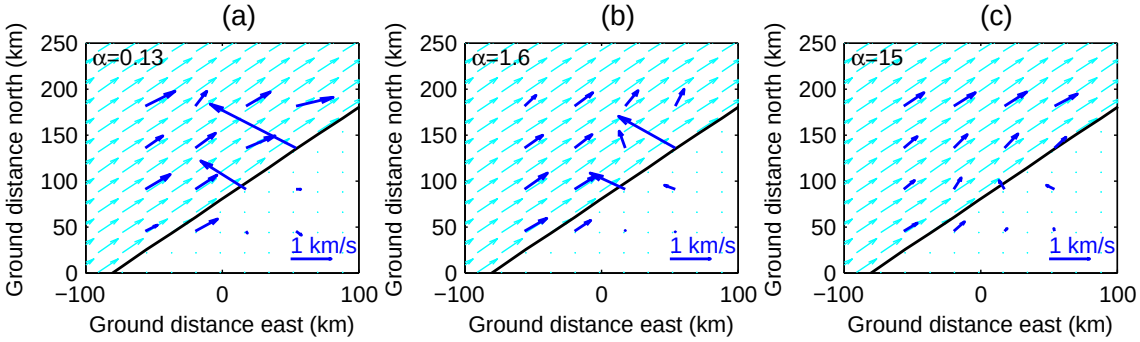


Figure 4·8: Same as Figure 4·7, but for Method B (Incompressible flow constraint).

The drop in the electric field is quite abrupt, and for the resolution considered here, the step function between the two regions is a valid approximation.

In the simulation, LOS measurements are generated by discretizing $\mathbf{v}(x_i, y_i)$ according to the 4×4 grid of Figure 4·5, projecting via (4.8), and perturbing by a zero-mean Gaussian noise vector $\underline{e}_{\text{los}}$ with covariance Σ_e as described above. In Figure 4·7, a velocity field is predicted using Method A with three values of the regularization parameter α. Figure 4·8 shows the corresponding predictions for Method B. In general, both predictors have difficulty resolving the discontinuity (a violation of the assumption of uniformity within each pixel). For small α, both produce very similar solutions (after all, as α approaches zero, the predictors are equivalent). As α increases, the respective side constraints come into play. In Figure 4·7, the preferred solution is the minimum-l^2norm , while in Figure 4·8, the solution exhibits smooth transitions between neighboring pixels.

The performance of the inversion is highly dependent on the geometry (i.e., the LOS direction vectors in row (pixel) j of $\mathbf{A}$). Inversion demands that the direction cosines be sufficiently

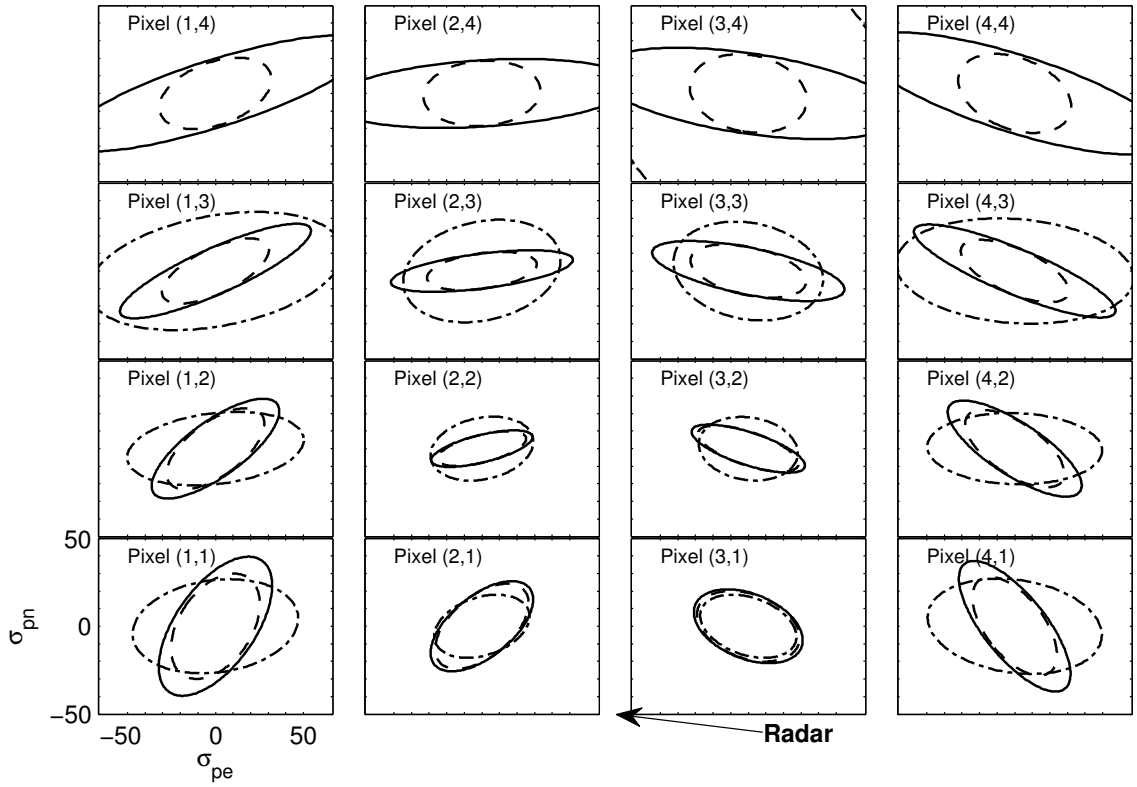


Figure 4·9: 1σ error ellipses for each of the predictors: Method A (- - -), Method B (——), and Overlapping pixels (- - -·). The units are m/s.

dissimilar. Otherwise, not enough independent information is present to recover the cross-range component.

Method A resolves the zero region better because it favors a zero solution. In the non-zero region, Method A's solution is very poor (see Figure 4·7c).

Prediction error

The data $\underline{v}_{\text{los}}$ for these predictions originates from an ISR parameter fitter. The fitter also provides an error estimate $\widehat{\sigma}_{v_{\text{los}}}$. Let these be the diagonal elements of Σ_e. Propagating this matrix through equation 4.10 for each pixel j results in a 3×3 covariance matrix $\widehat{\Sigma}_{\mathbf{v},j}$, quantifying the uncertainty in the predicted $\widehat{\mathbf{v}}_j$. These uncertainties are plotted in Figure 4·9 (for Methods A and B and regularization parameter $\alpha = 15$) in the form of error ellipses.[3] Each ellipse corresponds (in a sense) to a *confidence interval* (CI) of one standard deviation, i.e., there is a 39.4% chance that $\widehat{\mathbf{v}}_j$ lies on or

[3]Only the horizontal components pe and pn are shown.

within the ellipse. Thus a wider radius indicates a greater uncertainty in that direction.[4]

The angle of an ellipse indicates cross-correlation of the components of $\widehat{vecv}_j$ (in the geomagnetic coordinate system in which it is evaluated). In Figure 4·9 every ellipse is oriented with respect to the line-of-sight/transverse direction, suggesting strong dependence on the observation geometry.

The eccentricity of an ellipse indicates whether the predictor has a directional preference. In Figure 4·9, the semiminor axis generally points toward the radar. That is, equation (4.10) is assigning lower uncertainty to the LOS component, or equivalent, the predictor can infer the LOS component with greater confidence than the transverse component. Conversely, the semimajor axis reflects the poor observability of the transverse component, which is inherent to the problem. The semimajor axis is wider for Method B (solid lines) because of its wider support that $\mathbf{L}$ introduces (compared to the point support of Method A). The error covariance is greater because each $\widehat{\mathbf{v}}_j$ is constrained to agree with neighboring $\widehat{\mathbf{v}}_{J\setminus j}$'s by the prior term in addition to fitting its own local data.

The geometry of the problem strongly influences the uncertainty: those pixels farthest from the radar (top row of Figure 4·9) incur the largest errors for two reasons. First, the measurement error increases with the square of range. Second, because fewer samples fall within the pixels, which have uniform volume w.r.t. ground distance (see Figures 4·5 and 4·2).

A third set of ellipses in Figure 4·9 represents the "overlapping pixels" predictor described in Section 4.2. In a few pixels (the closest to the radar), that predictor matches or surpasses the error performance of Methods A and B. However, the uncertainty grows much faster with range than the other two.

Figure 4·10 shows the L-curves for this simulation. The L-curve is a semi-quantitative strategy for selecting an optimal level of regularization. The vertical axis measures total divergence; the horizontal axis measures how well the prediction fits the data. The curve is plotted for a range of α to characterize the tradeoff between smoothness (vertical axis) and data fit (horizontal). Closer to the origin is better. This curve typically takes the shape of the letter "L." The vertical segment corresponds to low α, where data fit takes priority over smoothness. In this regime, increasing α results in a smoother prediction that is still consistent with the data. In the horizontal segment, α has less effect on the smoothness of the solution but results in an ever more inconsistent prediction.

[4] In 1D, a 1σ confidence interval corresponds to a 68.4% certainty level. Let $\sigma_{2D} = (\sigma_{\text{pe}}^2 + \sigma_{\text{pn}}^2)^{1/2}$. Then the 1σ error ellipse is the locus of points $(\sigma_{\text{pe}}, \sigma_{\text{pn}})$ corresponding to the $1\sigma_{2D}$ confidence interval.

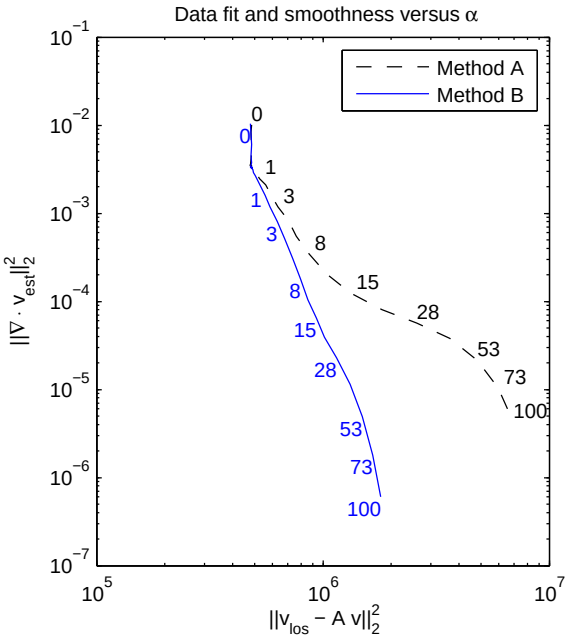


Figure 4·10: L-curves for Methods A and B for the ground truth (cyan) velocity field pattern in Figures 4·7 and 4·8. Data fit metric is on the horizontal axis. Roughness metric is on the vertical axis.

The "optimal" α lies between these two extremes, at the knee of the curve if such a point can be identified.

The roughness metric in Figure 4·10 for a given α is lower (i.e. better) for Method B while the data fit is consistently better. This is not surprising, since Method B is designed to minimize both metrics, while the side constraint in Method A, with its preference to shrink toward zero, is anathema to the goal of accurately predicting the field! The knee of the curve is more easily identifiable for Method A at $\alpha \approx 15$. We will now use this value in comparisons to qualitatively assess the performance of the predictors versus α.

Other simulation cases

In the discussion above we justified using the field shown in Figures 4·7 and 4·8. This was motivated by a particular phenomenon that is expected to occur in the ionosphere during substorms. The advantage of spatial regularization is that it provides robustness in the presence of spatial variation. Hence we now consider two variations of the earlier pattern: a uniform field $\mathbf{v}(x, y) \equiv \mathbf{v}$ (see Figure 4·11) and a very thin enhancement with oppositely directed velocity on the other side of the arc (see Figure 4·12). The top (bottom) row shows a pair of predicted fields using Method A

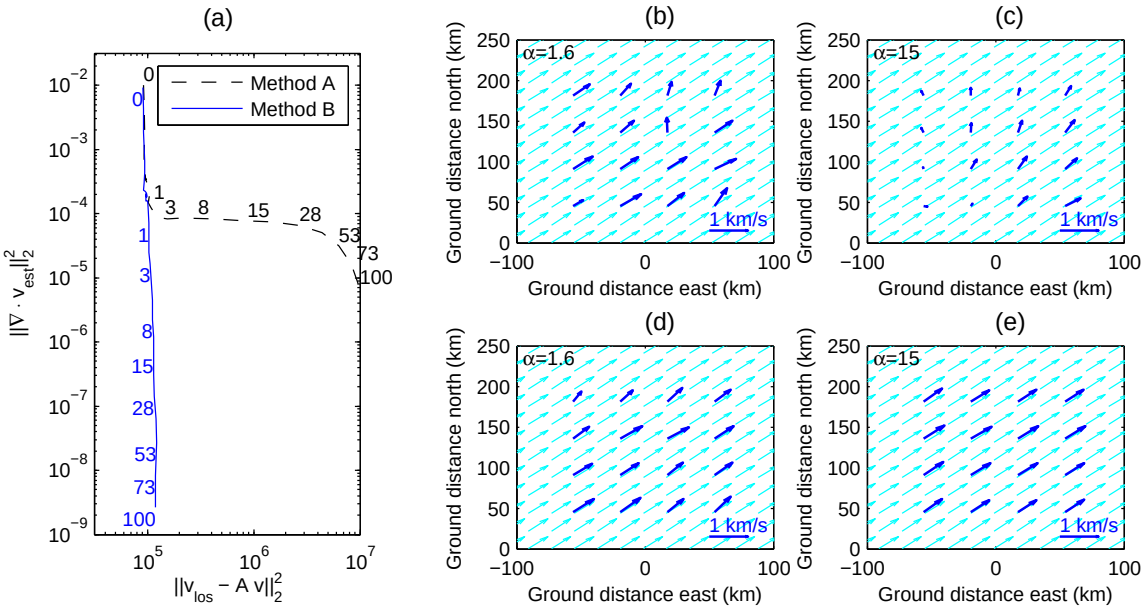


Figure 4·11: Uniform flow field. (a) L curve for Methods A and B. Sample reconstructions for both methods are shown: (b) & (c) Method A, (d) & (e) Method B.

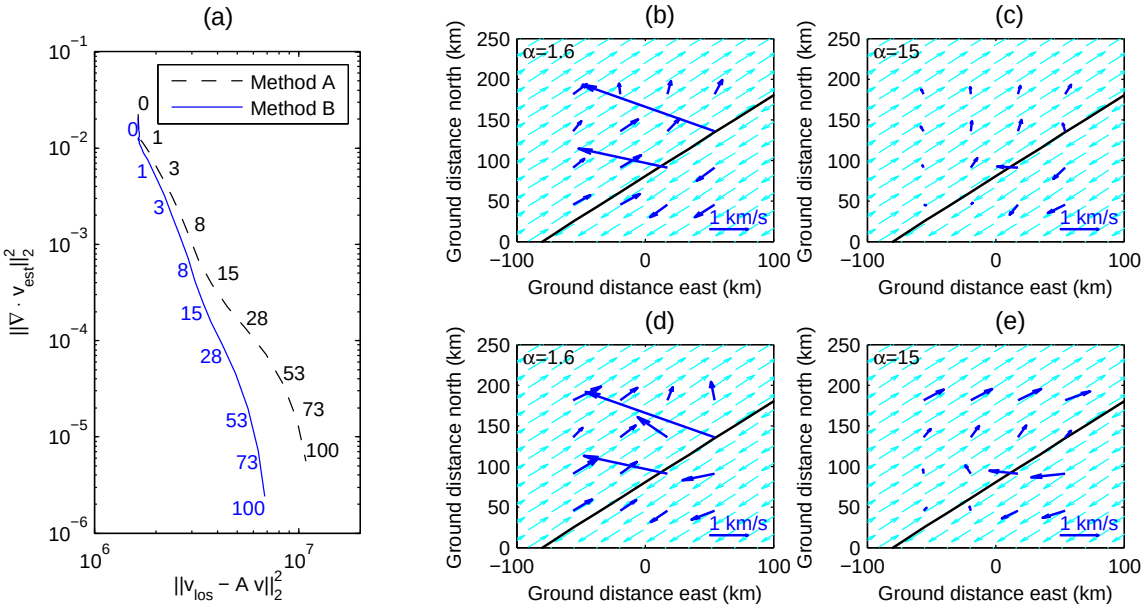


Figure 4·12: Shear with field reversal inside the arc. (a) L curve for Methods A and B. Sample reconstructions for both methods are shown: (b) & (c) Method A, (d) & (e) Method B.

(Method B) for both low and high values of α.

Figures 4·11a and 4·12a show the L-curves for each of these two variations. Some general observations can be made pertaining to both. As before, Method A reaches a point of diminishing returns when it begins to emphasize smallness over data fit. The L-curve levels to horizontal and the predicted field approaches zero. By comparison, the L-curve for Method B is closer to the origin for all values of α.

The uniform field (Figure 4·11) presents no challenge for either predictor, since uniformity is an important assumption in its design. But when the regularization "kicks in," Method A defeats itself by approaching the zero field. More importantly, the "shrinkage" in Method A dramatically alters the direction and overall shape of the predicted flow pattern. It is this overemphasis of the side constraint that leads to the horizontal segment of the L-curve. By comparison, Method B preserves the uniform direction of $\hat{\mathbf{v}}$ for all α, and the L-curve is practically vertical.

Turning now to the shear flow case (Figure 4·12), the discontinuity is even more difficult to resolve than the step function considered previously. Though reconstruction errors do extend beyond the position of the discontinuity (again due to the relatively wide support of the divergence operator), both methods perform best where the underlying field matches the assumption of uniformity. In particular, the top row of predictions is nearly perfect. Around $\alpha = 15$ (Figure 4·12e), Method B comes closest to the true field. For higher values (not shown), the solution begins to approach something like solenoidal flow (i.e., the ideal solution if $\nabla \cdot \mathbf{v} = 0$ exactly). Hence the knee (albeit slight) located around $\alpha = 15$ in panel a. The differences in the L-curves for this case are not as dramatic, but the same general observations apply: Method B performs better, i.e., it is consistently both smoother and a better fit to the observed data. Furthermore, the constraint of Method B (approximately divergence-free flow) constitutes a prior model informed by knowledge of the physics of the process.

4.5 Case studies

The following examples cover the canonical substorm phases: following a rapid onset, there is a period of growth, then expansion of large auroral structures, and finally a long recovery phase in which the flow slowly returns to a steady background field as the auroral activity diminishes. In many cases, the results indicate a coincidence of flow shears with auroral boundaries, consistent with theory.

We validate our estimates by generating composite images of velocity with other observations.

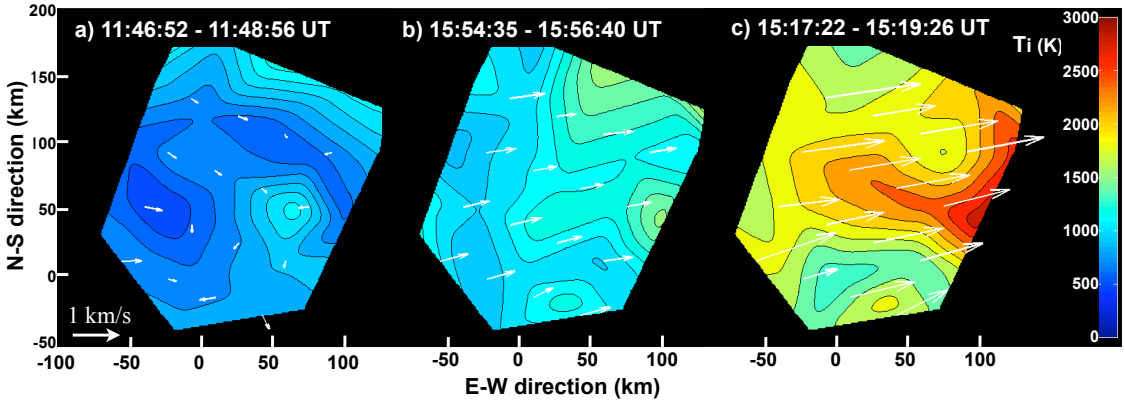


Figure 4·13: Examples of observed correlations between $|\mathbf{v}_i|$ and T_i in agreement with equation (4.12).

For instance, since in the F-region altitudes from 140–300 km the ion energy equation is dominated by frictional heating and collisional cooling, there is a direct relation between the ion temperature T_i and speed $v = |\mathbf{v}|$ (St.-Maurice et al., 1996):

$$T_i = \frac{v^2 M_n}{3k_B}\left[\left(\frac{\nu_{\text{in}}}{\Omega_i}\right)+1\right]^{-1} + T_n, \tag{4.12}$$

where M_n is neutral mass, k_B is Boltzmann's constant, ν_{in} is ion-neutral collision frequency, Ω_i is ion gyration frequency, and T_n is neutral temperature. This relationship was frequently observed during experiments, for instance in Figure 4·13. The velocity vectors (arrows) were computed from LOS measurements as described in Section 4.2. Ion temperatures were extracted (one per beam) at an altitude of ~ 240 km and interpolated to form contour plots. These plots generally agree with equation (??), with hotter regions corresponding to faster flows and cooler regions having lower velocities.

Figure 4·14 is a scatter plot of ion temperature versus speed. While there is some spread, the parabolic trend suggests a relationship much like equation (??). The red line plots equation ?? directly using neutral parameters computed using the NRL-MSISE-00 empirical model (Picone et al., 2002).

This will be a common theme throughout the following case studies. Several examples are chosen that are representative of the expected behavior of aurora and ion flow fields during substorms. Occasionally the estimator(s) generate what can plainly be judged are artifacts, resulting from some violated assumption either in the physics of the process (e.g. very small v_{ap}) or in the

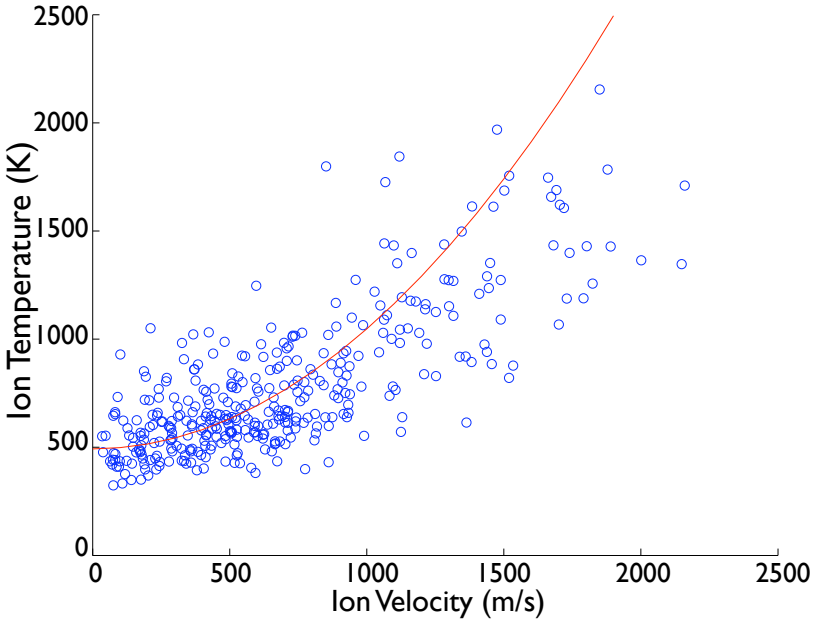


Figure 4·14: Observed ion temperatures versus speed. The red line plots equation ?? using neutral parameters obtained from the MSIS model.

discretization or inversion (i.e. inadequate spatial/temporal resolution or over-/under-smoothing). In such cases, we offer alternative hypotheses and suggestions for revising the technique to resolve such pathological cases.

In addition to ion temperature, we also generate composite images of aurorae measured by a DASC at Poker Flat. Other remote sensing diagnostics do not yield themselves to composite imaging, but these data (including *meridian-scanning photometer* (MSP), magnetometer, *Fabry-Pérot interferometer* (FPI) data) provide further context for studying the events captured in our experiments. The optical portions of these images are generated by projecting the portion of the all-sky images that intersects with the PFISR fov and converting to cartesian coordinates assuming a fixed emission height of 110 km.

4.5.1 26 March 2008

In an experiment run 24 March 2009, PFISR was operated in the 26-beam mode of Figure **??**. PFISR sampled the full array of 26 beams ("frame") every 5 s, on two interleaved frequency channels: (1) an uncoded 480 μs pulse (to probe the F-region), and (2) an alternating code (to probe the E-region). Since this study is restricted to F-region convective flow, only measurements from the uncoded channel are used. Treating each beam direction separately, AMISR forms the raw IQ (volt-

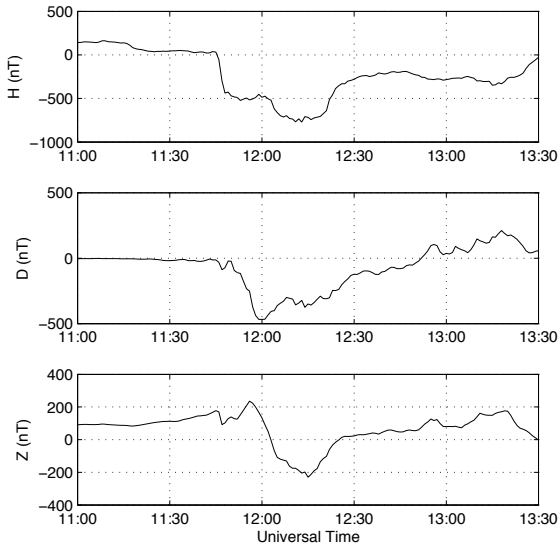


Figure 4·15: Magnetometer traces for 26 March 2008, from Poker Flat.

age) signals then integrates these to form range-gated ACFs (complex power signals). A nonlinear ISR fitter then generates estimates at each sample point of the plasma parameters N_e (electron density), T_e (electron temperature), T_i (ion temperature), and v_{los} (*line-of-sight* projection of ion drift velocity). (Other parameters affecting the theoretical ACF are modeled rather than estimated.)

Example #1: The three canonical substorm phases

The optical readings from this night bear the signature of a classic substorm. Figure 4·16 features keograms from the MSP located at Poker Flat, tracking brightness versus elevation on four bands: 557.7 nm and 427.8 nm (corresponding to emissions resulting from precipitating electrons), 486.1 nm (corresponding to proton precipitation), and 630.0 nm (corresponding to ionization and photochemical emissions). A large structure is rapidly propagating southward from 1115–1145 UT (growth phase) before a sudden burst of brightness, particularly in the 427.8 nm band (expansion). The subsequent recovery phase is a lengthy return to normalcy.

We compare our derived ion drift velocities to MSP data in Figure 4·17. Plasma drift accelerates rapidly leading up to the start of the growth phase. The drift speed then drops suddenly just as the luminous region comes into view of this "pixel," followed by a return to the prior speed with the exit of the luminous region. At the start of the expansion phase, the brightness and velocity once again have an inverse relationship. During the recovery phase, however, there may be a weak direct correlation.

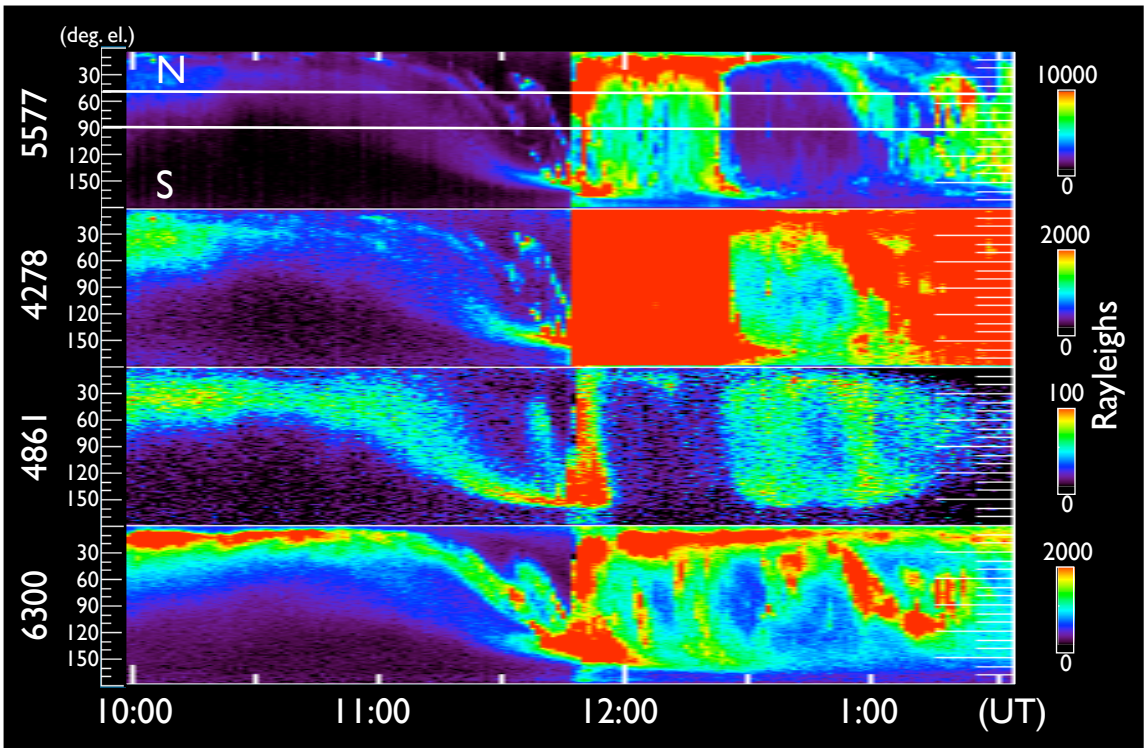


Figure 4·16: MSP data from four bands for 26 March 2008. The substorm begins around 1145 UT.

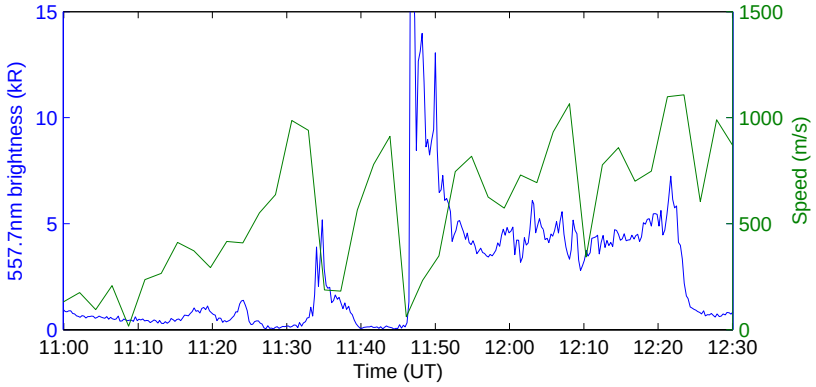


Figure 4·17: MSP data (blue) and predicted ion speed (green, using overlapping pixel method) close to the time of substorm onset, 26 March 2008.

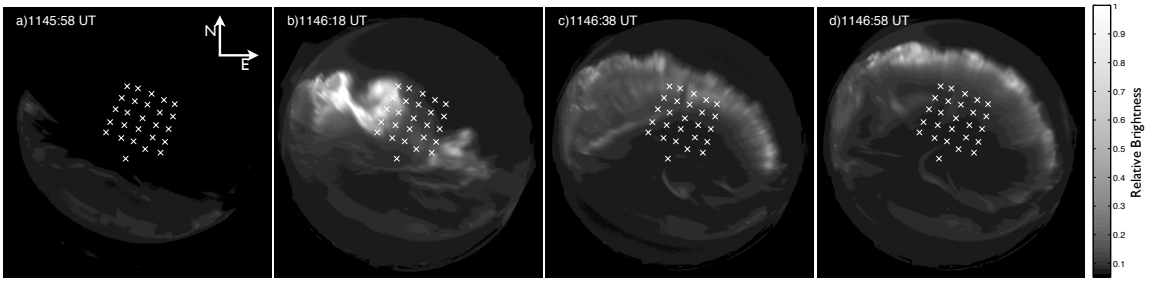


Figure 4·18: All-sky images from Poker Flat showing southward-moving substorm onset activity. The crosses represent $5 \times 5 + 1$ PFISR beam directions for this experiment.

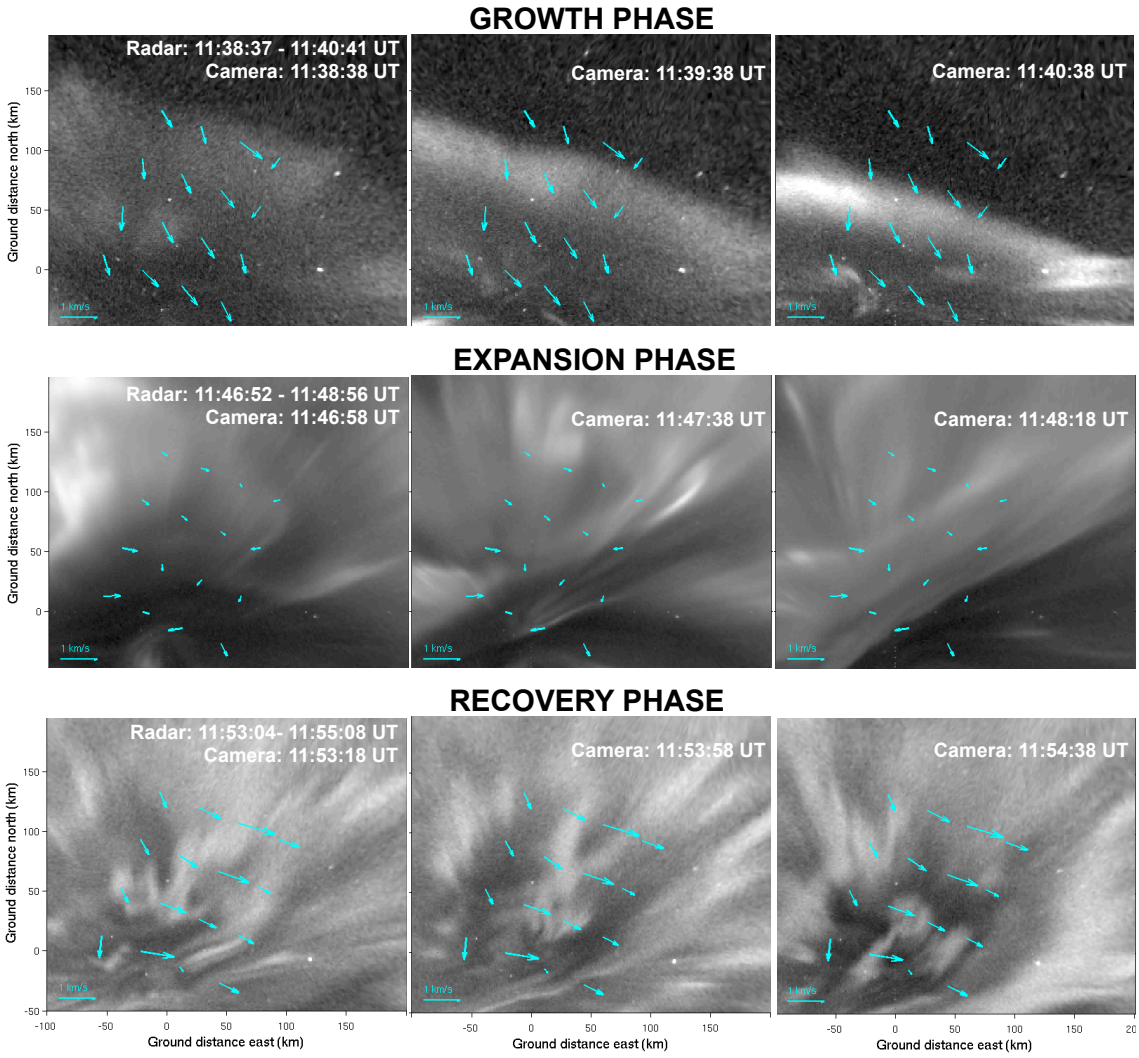


Figure 4·19: Composite images of substorm auroral activity and PFISR-derived ion flow fields for the three substorm phases.

Finally, we examine the composite optical/radar wide-field images. Figure 4·18 shows the field of view for this experiment (white crosses represent PFISR beams) vis-a-vis the nearby DASC. This instrument operated at a cadence of 20 s in its white light (unfiltered) mode. The image sequence depicts the rapid onset of an expansion phase: a sudden and bright burst lasting one to two minutes.

Expanding our scope, Figure 4·19 depicts isolated examples from each of the three substorm phases: growth, expansion, and recovery. PFISR's integration time was two minutes, compared to the DASC cadence of 20 s, so each flow field estimate corresponds to multiple DASC frames. We get some information here about

- Fine-scale spatial relationships between brightness and electric field,
- How the "average" dynamic behavior within a pixel may affect prediction,
- Once PFISR estimates of N_e and FPI data are accounted for, taking out neutral wind, separating out electric field effects from convective disturbances.

Example #2: Arc activation

Figure 4·20 shows the activation of an auroral arc about 12 minutes before substorm onset. When the arc passes through the radar fov (white crosses), ion temperature and drift velocity are superimposed on a magnified portion of the all-sky images (Figure 4·21). Again brightness and velocity appear anticorrelated. In panels a & b, the high-temperature regions correspond to low brightness at the altitudes shown (110 km optical, 240 km temperature). In panel a, there are two distinct arcs, between which the velocities are generally tangentially aligned. When the arc begins to diminish in panel c, both the temperature and velocity drop rapidly in that region. This again follows the relationship given by (4.12) and is also consistent with the polarization effect described in Section 4.4.

Example #3: A westward-traveling arc

Finally, on the same night, a wide, north–south-aligned arc traveled westward through the radar fov (Figure 4·22). In panels a–c, as the arc moves into the rightmost edge of the fov, the velocities subside and dramatically reverse direction. The direction of the flow parallel to the the eastern arc boundary also suggests a polarization effect directed east.

Following that, the velocity field appears to rotate south-west in sync with the progression of

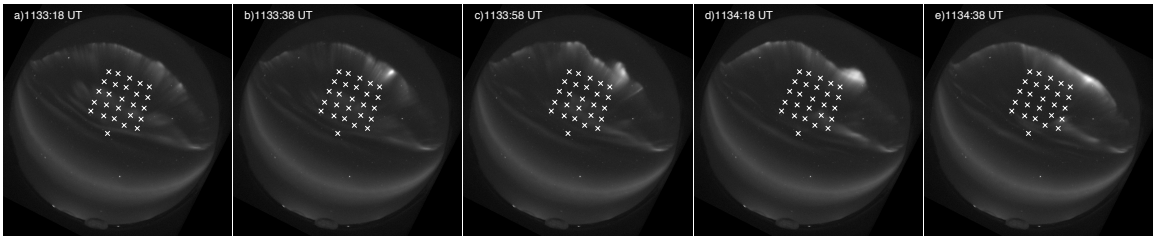


Figure 4·20: Example #2. DASC images for an arc activation occurring – – Example #2. All-sky images from Poker Flat of the auroral activation event (26 March 2008) described in Section 4.5.1. White crosses represent PFISR beams.

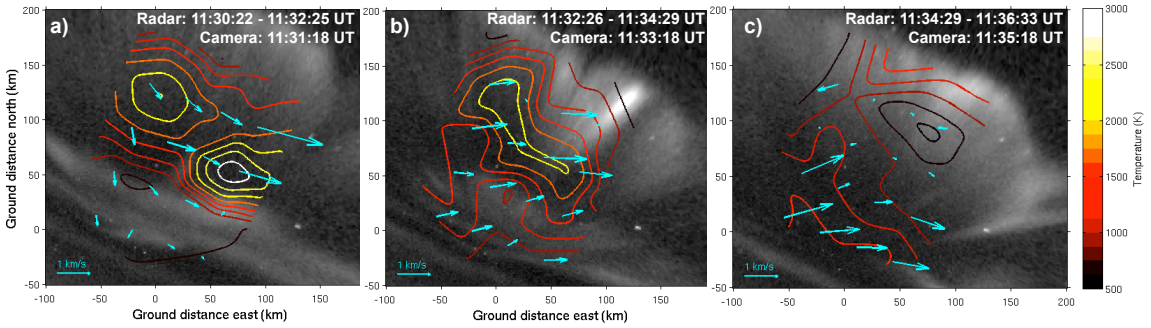


Figure 4·21: Example #2. Detail of Figure 4·20. Recovered flows and ion temperatures for the arc activation of Section 4.5.1 are superimposed on DASC images.

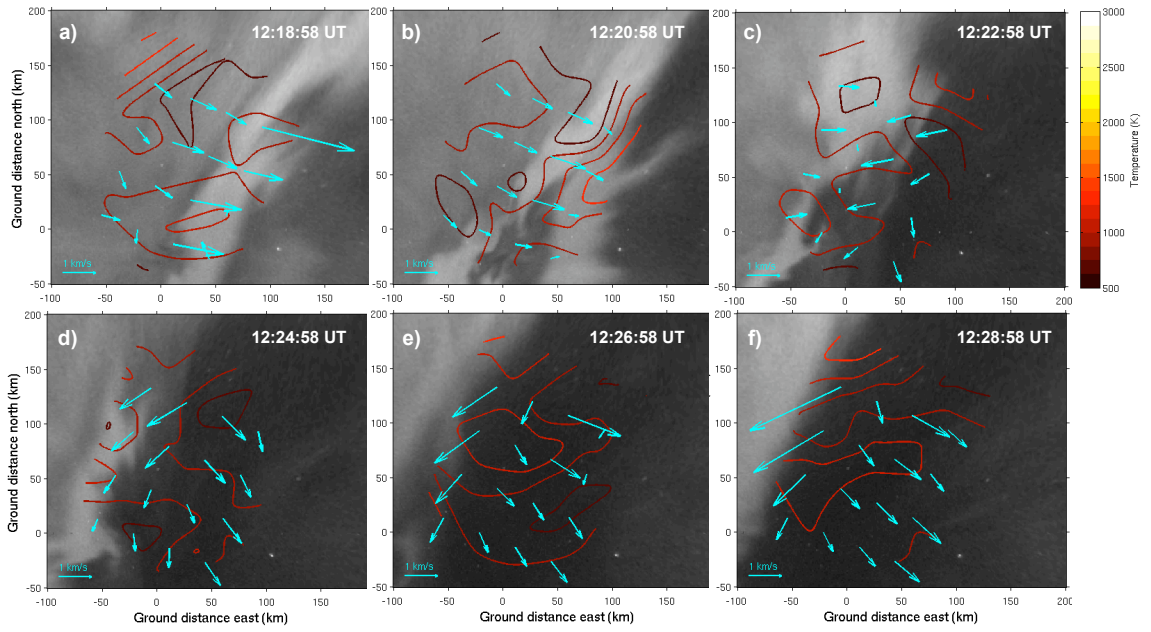


Figure 4·22: Example #3. A westward-traveling north-south arc and the associated ion temperature and flow fields.

Table 4.1: Parameters for the PFISR experiments conducted 26 Mar 2008 and 24 Mar 2009.

	26 Mar 2008	24 Mar 2009
Beam positions	$5 \times 5 + 1$	$5 \times 5 + 1$
Pulse type	Long pulse + coded (unused)	Long pulse
Pulse length	480 μs	480 μs
# channels used	1	2
Integration time	2 min	30 s
All-sky wavelengths	unfiltered	5577, 6300 Å

the arc. Once the arc is finally clear of the fov, the predicted velocity field appears to return to it's previous state, drift under the influence of an ambient electric field.

4.5.2 24 March 2009

In an experiment run 24 March 2009, PFISR was operated in the same 26-beam mode. The experiment from above was altered to improve the statistics of the readings. Two frequency channels were used, each now using an uncoded 480 µs pulse. Every 5.5 s, returns were sampled from 14 pulses on each channel in each direction, The auroral activity of this night generated returns with high SNR, allowing estimates of LOS velocities from relatively few samples. Table 4.1 summarizes the setup for the two experiments.

The results described in this section were obtained during a 1 h period using a radar integration time of 30 s, corresponding to ~140 pulses-per-beam. In addition to LOS estimates, the ISR fitter also supplies error covariances. These values provide the diagonal elements of Σ_e. Using Method B (divergence-constrained regularization), the velocity fields were reconstructed with regularization parameter $\alpha = 5$. (This value was chosen based on trial and error.)

The predicted fields are superimposed onto optical images captured by the nearby DASC. The camera captured both 557.7 nm and 630 nm wavelengths, but only the 557.7 nm data is displayed in the following figures. At a cadence of 20 seconds, the all-sky imager captures dynamics with timescales comparable to those captured by the radar reconstructions. The velocity fields and optical data are mapped to a common plane on the page by assuming an auroral emission altitude of 120 km.

Figure 4·23 shows four contiguous 30 s flow field predictions in the vicinity of a stable east-west aligned arc of ~ 50 km width. Panel a shows a relatively uniform flow in the magnetic westward direction, tangential to the arc boundary. The bulk drift is slightly slower within the arc, consistent with a reduced electric field within the region of increased conductivity. Panels b and c depict the development of a flow reversal near the poleward boundary of the arc. The circulatory appearance

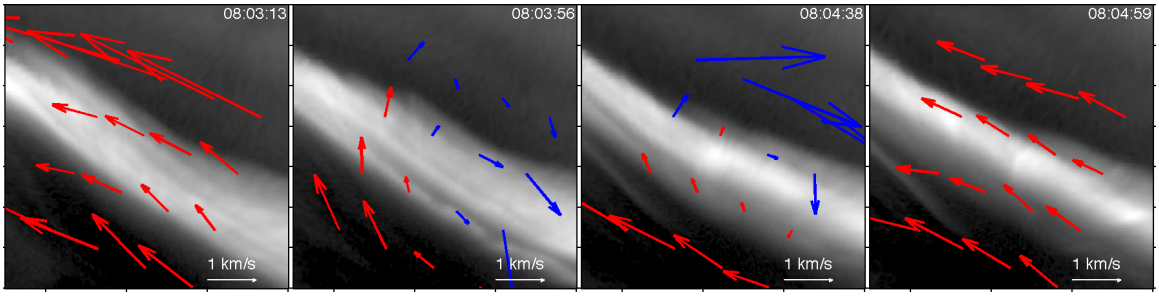


Figure 4·23: Co-registered ion convective flow fields and auroral forms constructed at 30 s cadence. Panels b and c illustrate the formation of a transient region of reversed flow near the poleward boundary of the arc.

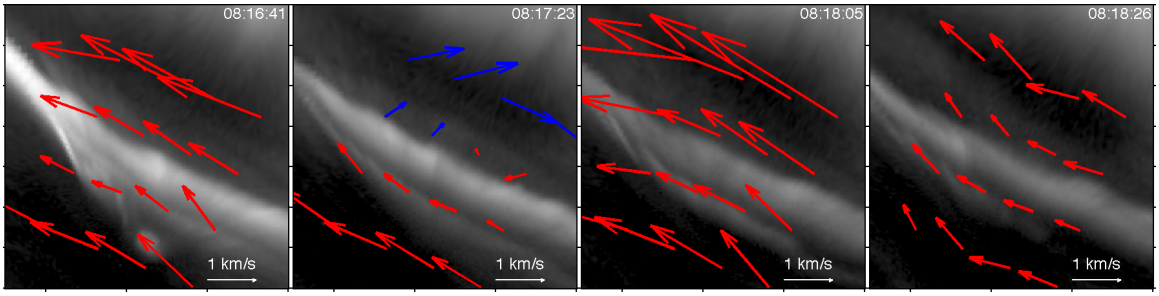


Figure 4·24: Another example similar to Figure 4·23.

of the flow field is reminiscent of Figure 4·12e. In Panel d, the flow resumes its uniform westward course.

Figure 4·24 is a second example of a sharp flow shear apparently developing very rapidly (30 s) in the vicinity of a pre-existing auroral form. Rapid localized fluctuations in convective flow have previously been identified by Bristow (2008) using the SuperDARN HF radar network. Their cause remains unclear. If this is the case, the morphology is clearly under-sampled in time, as the reversal appears only in one frame. Such rapid fluctuations are common throughout this experiment. Figure 4·25 shows a longer sequence of flow field predictions during a period of dynamic auroral activity. Although the correlation with auroral boundaries is less clear, we again see large fluctuations in both magnitude and direction of flow, as well as the ephemeral appearance of strong flow shears, throughout.

4.6 Discussion / General observations

We have demonstrated the capability of an electronically steerable ISR to predict F-region flow fields. In order to achieve robustness in the presence of spatial variation, we chose to implement regularization in the solution. In our analysis we compared the performance of two regularization

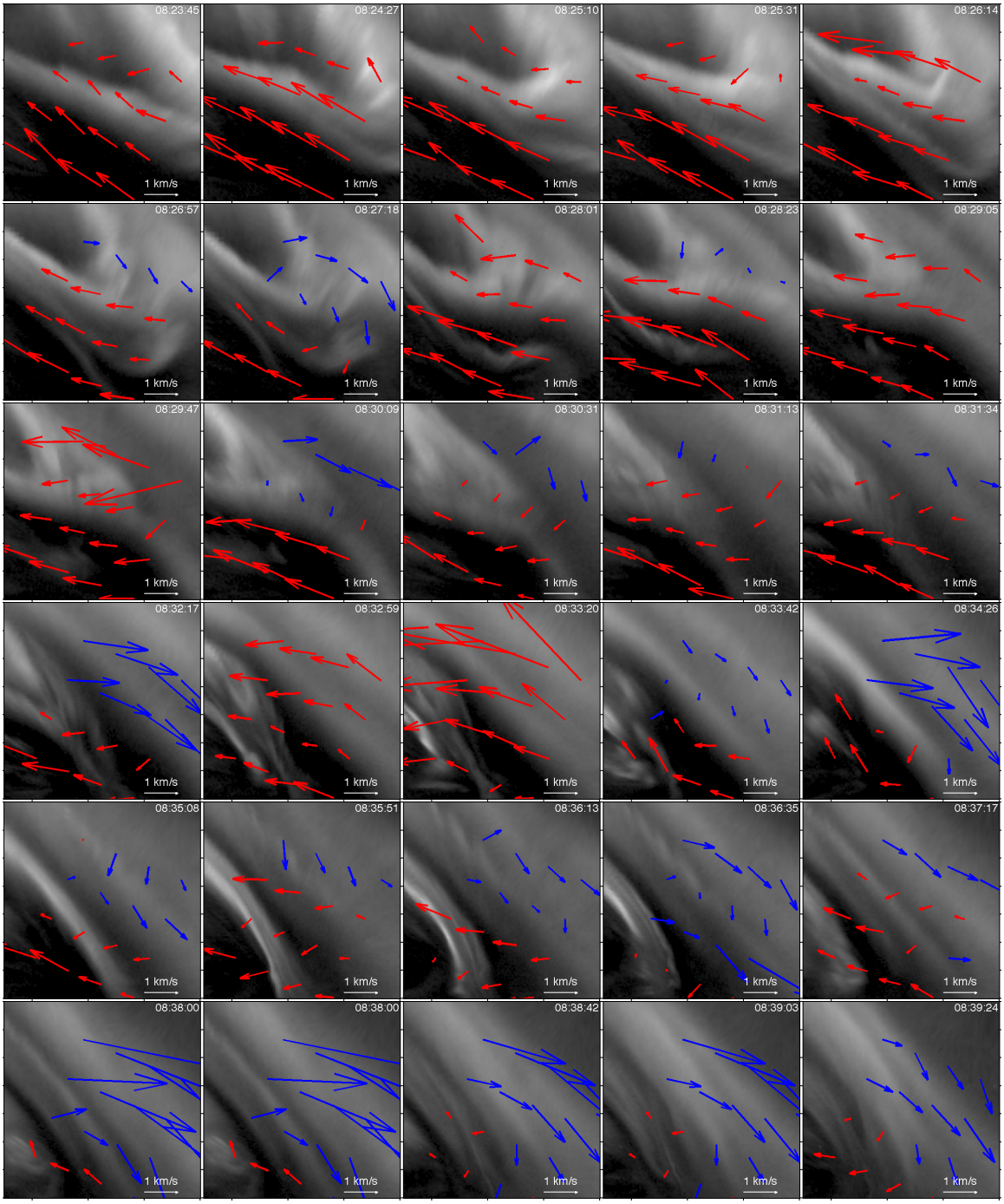


Figure 4·25: A longer sequence illustrating the relationship between flows and auroral forms.

functionals: Method A, with a penalty on large-magnitude solutions, and Method B, the "incompressible flow" estimator with a penalty on local spatial variability.

Both estimators have trouble resolving a sharp discontinuity, such as that seen in the simulation of Section 4.4 (Figures 4·7 & 4·8). Both perform well in regions of uniform flow. For large values of the regularization parameter α, Method A shrinks to the zero field and in the process dramatically alters the morphology of the solution. Method B enforces uniformity (locally) or approaches the solenoidal solution (globally). Whether or not these solutions are realistic depends on the spatial variability of the process under observation. It is therefore crucial to consider the effect of the regularization parameter α on the analysis, whether it causes oversmoothing (Figure 4·12e), or whether undersampling causes artifacts to appear as a result of the violated assumption of uniformity (Figure 4·12, panels b and d).

The accuracy of the velocity reconstruction depends heavily on the geometry of the problem. Hence each pixel is characterized by a unique error profile (Figure 4·9).

In applying the estimation technique to PFISR measurements, we validated our findings by comparing to a sequence of co-registered all-sky optical images from the same night. The optical data were captured at a time resolution similar to the radar integration time, so that dynamics of similar time-scale could could be compared. The salient features of these data are (1) a reduction of convective flow within an auroral enhancement (de la Beaujardière and Vondrak, 1982) and (2) the generally parallel direction of the ion drift at arc boundaries, consistent with a polarization effect within the arc (e.g. Lanchester et al., 1996).

The Tikhonov formulation adopted here is advantageous for a variety of reasons. It is capable of handling the overdetermined problem. A smoothness constraint is straightforward to introduce by penalizing large local differences in the solution. Through the data covariance matrix Σ_e, the estimator accounts for the uncertainty inherent to all practical measurements. The second-order statistics of the prior model are encoded in Σ_v, and the theory provides a measure of estimator uncertainty via equation 4.10.

Heinselman and Nicolls (2008) develop an method of estimating velocity vectors from LOS projections. However, their approach relies on a particular beam arrangement, with the goal of determining velocities (equivalently electric fields) as a function of magnetic latitude. The result is a time-sequence of latitudinally distributed vectors. The technique described here is somewhat agnostic of beam arrangement, meaning that velocity fields can be obtained in experiments not necessarily designed for that purpose (for instance, high-resolution ionospheric imaging, as in

Chapter 3). The reconstruction grid can be somewhat arbitrary too. The result is a time-sequence of vector fields distributed in both latitude and longitude. The Heinselman/Nicolls approach is analogous to slit-scan photography, if ours is compared to video imaging.

To emphasize the novelty of this approach, it is worth comparing the acquisition and estimation procedure presented here to another method capable of estimating three dimensional F-region ion flow. The tristatic EISCAT system receives three independent LOS projections of ion flow velocity within a common volume. This allows unambiguous recovery of all three velocity components within the volume. PFISR is a monostatic radar, and the recovery of vector velocity requires the combination of neighboring measurements as described above. Although EISCAT is routinely operated in a meridional scanning mode that provides estimates along latitude, PFISR's electronic steerability allows acquisition of a "snapshot" as described in Section 4. The monostatic arrangement is inherently unable to resolve the full flow vector due to the limited amount of independent information provided by neighboring measurements. The only way to resolve this ambiguity is to introduce outside information. The Tikhonov method of regularization is a natural way to incorporate such information.

Exogenous information may come in the form of a physical model, e.g. a statistical model (Sulzer et al., 2005; Hysell et al., 2009). It may also include ancillary data from separate instruments (i.e. sensor fusion). For instance, if there is a reason to believe the direction of ion flow is dominated by large-scale convection (e.g., if SuperDARN measurements indicate such a large-scale flow), the solution can be "steered" to a preferred direction to make use of this assumption. The solution is encoded with a directional preference by designing the a priori covariance matrix $\mathbf{Q}$ such that the horizontal variabilities σ^2_{pe} and σ^2_{pn} reflect confidence in the estimate of the respective components.

In this work, we have used coregistered optical images to provide a context for interpreting the results. The optical brightness serves as a proxy for conductivity. Wherever an auroral arc occurs, the conductivity is higher. In order to maintain current continuity, the electric field in this region (and thus the drift velocity) is reduced. After identifying the arc boundary in the optical data, this can be used by the estimator to segment the solution into regions with different prior constraints. For instance, since we expect the plasma flow at the boundary of an auroral arc to be parallel to the arc, we may tune the prior model to steer the solution in the appropriate direction. Rather than to perform this tuning by hand for each image, such contextual information could be provided to the predictor and automatically applied to its results.

A notable feature of Figure 4·9 is the spatial heteroskedasticity of the error covariances. This is due to the irregular sampling of the pixelization in Figure 4·5. That is, the measurement sample points are determined by the radar geometry, and we have laid a uniform 4×4 grid over these sample points. As a result, some pixels contain more measurements (i.e. better statistics) than others. The velocity estimates in those pixels are more reliable than the poorly-sampled top row of pixels. A sampling strategy based on homogenizing or reducing spatial uncertainty may help in this case. The kriging variance is often used for this purpose.

Even when the pixelization accommodates the radar geometry, the course discretization coupled with an implicit assumption of uniformity within each pixel is in direct opposition to the goal of resolving spatial variability. Following the example of geostatistics in earlier chapters, a large-scale trend with a small-scale random effect is a sensible approach. In the auroral zone, there is often a uniform background convection superimposed with variations from ionospheric phenomena. Like Tikhonov regularization, this also provides a natural way to incorporate prior information in the form of statistical parameters.

Machine-learning and classification approaches also come to mind. For instance, in the simulation in Section 4.4, the optimal pixelization would be two triangular pixels separated by the boundary indicated in Figures 4·7 and 4·8. As few as one velocity estimate might be recovered per segment.

Chapter 5

Global data: Mapping total electron content

The preceding chapters have focused on mapping observations and their uncertainties in space. Chapter 2 presented the statistical theory of optimal prediction within the framework of spatial statistics. Chapter 3 applied optimal spatial prediction to map spatially-dispersed radar observations to unobserved locations. Chapter 4 demonstrated, using a link to inverse theory, the use of neighboring data to constrain the recovery of unobserved vector components.

This chapter serves as a "jumping-off" point for future studies. These suggestions are, to varying degrees, developed conceptually, awaiting implementation. The first suggestion, especially. Like the previous two chapters, it focuses on an application of ionospheric aeronomy.

5.1 Total electron content

Total electron content (TEC) is an important characteristic of the ionosphere, with interest stemming mainly from the role of the ionosphere in degrading the radio signals between satellite and ground-based transceivers. Where precision is a critical requirement (e.g. in geolocation, satellite tracking, instrument calibration), it is desirable to map, monitor, and mitigate for the effects of TEC disturbances.

Because radio signals propagate more slowly through the ionosphere, it is necessary to correct for this delay at the receiver end. Particularly challenging is the nonuniform and complex spatiotemporal behavior of the ionospheric plasma. A message-bearing radio wave encounters innumerable pockets of concentrated and rarefied plasma, each differential of conductivity contributing to the total index of refraction and thus delay. Hence the need for an independent estimate of TEC, whether or not the goal is to study ionospheric density in particular.

For example, the network of GNSS relies on accurate timing, and atmospheric effects are a major source of performance loss. Many receivers are equipped with augmentation systems for detecting or correcting such problems. The Jet Propulsion Laboratory maintains the WAAS, including a network of static, ground-based GNSS receivers, which provide coverage across the United States.

Users of WAAS-enabled devices can compute a real-time local TEC estimate from a grid of zenith-mapped estimates. Correction for atmospheric effects ultimately leads to more precise location services. WAAS also monitors the quality of estimates on the network's receivers, detects irregular ionospheric conditions, and warns users when these effects compromise the reliability of the system (Sparks et al., 2011a). The latest version (WAAS Follow-On Release 3) uses kriging to assign zenith-mapped TECs to each gridpoint, and kriging variances are used to assess the system's reliability (Sparks et al., 2011b). Because navigation is often a safety-critical application, WAAS is intentionally conservative in its error estimates.

Systems like WAAS use GNSS receivers to diagnose the system they augment. Similar receivers are often used to study the ionosphere. MIT Haystack Observatory hosts globally referenced TEC maps drawn from observations on a large network of GNSS receivers (Rideout and Coster, 2006). Quite the opposite of WAAS, these maps are intended to be analyzed for the study of large-scale, regional, and global geophysical events. The associated software, MAPGPS, was developed with the goal in mind of detailed mapping, and less emphasis was placed on detecting and describing error.

MAPGPS results are available through the Madrigal Database `http://cedar.openmadrigal.org/` as vertical TEC estimates registered to a regular latitude/longitude grid wherever GPS measurements are available. Summary plots are also provided. Although these high-level data are well-suited for visual inspection, the intervening subsampling and truncation constitute a non-trivial destructive transformation. Lower level data are available upon request. These are not directly from the sensors; rather they represent an intermediate stage of MAPGPS immediately before mapping to a regular grid.

5.2 Description of the data

The basic data product at this level is an estimate of slant TEC, defined as the electron density within a 1 m^2 cylinder, integrated along the line-of-sight from receiver r to satellite s:

$$\text{slTEC} \triangleq \int_{l_r}^{l_s} n_e(\mathbf{x}(l))dl \qquad [\text{m}^2], \tag{5.1}$$

where l_r and l_s are the positions of receiver and satellite, respectively, n_e is electron density in electrons/m^2, and $\mathbf{x}(l)$ indicates the the line-of-sight. It is usually reported in TECu, where 1 TECu = 10^{16} electrons/m^2.

Since (5.1) depicts a projection, it is natural to consider tomographic reconstruction as an ionospheric diagnostic. This would, after all, recover the three-dimensional structure of the ionosphere. And while a 3D representation would presumably lead to more accurate navigational corrections (by directly computing a discrete approximation of (5.1)), the information needed to describe such a model in sufficient detail could easily exceed the storage or bandwidth limitations of a real-time, auxiliary network like WAAS. Hansen (2002) discusses this scenario quite thoroughly. It is worth adding that GPS satellites provide irregular surface coverage, and their continually shifting ray paths as they orbit Earth are susceptible to subtle effects not modeled in the receiver's internal ephemeris, making ionospheric tomography a challenging, though not insurmountable, problem. Although a 3D tomographic reconstruction is undoubtedly an asset to aeronomic study, it is also considerably complex on a large scale. Also, it is not obvious whether a practical system would benefit from a tomographic approach over a simpler model.

Indeed the prevailing approach involves collapsing the influence of the entire ionosphere to a limited region and assuming a 2D ionosphere. There are a variety of such models (e.g., see Coster et al., 1992). The simplest and most common is the *thin shell model*. The ionosphere is collapsed to an infinitesimal spherical shell, concentric with Earth and having radius $R_e + h_m$, where R_e is Earth's radius, and h_m is the altitude of the shell (roughly coinciding with the ionospheric peak, ~ 350 km to 450 km for high latitudes).

Figure 5·1 illustrates the geometry of this system. When receiver r receives a signal from satellite s, the point along $\mathbf{x}(l)$ where the signal's path intersects the thin shell is an *ionospheric pierce point* (IPP). The cumulative effect of the wave's path through the ionosphere is reduced to an instantaneous effect at the IPP. This allows a common mapping of TEC to the zenith (vertical TEC, or vTEC), independent of a receiver's location.

For each receiver-satellite pair, (5.1) can be transformed to an integral with respect to height:

$$\text{slTEC} = \int_{h_r}^{h_s} \left[1 - \left(\frac{R_e \cos(\text{el})}{R_e + h_m} \right)^2 \right]^{-1/2} n_e(\mathbf{x}_{rs}(h))\, dh, \tag{5.2}$$

where el is the elevation angle of the receiver. This function integrates n_e along the line-of-sight r to s, but $\mathbf{x}_{rs}$ is a function of height. The *obliquity factor* $\left[1 - \left(\frac{R_e \cos(\text{el})}{R_e + h_m} \right)^2 \right]^{-1/2}$ accounts for the reparameterization.

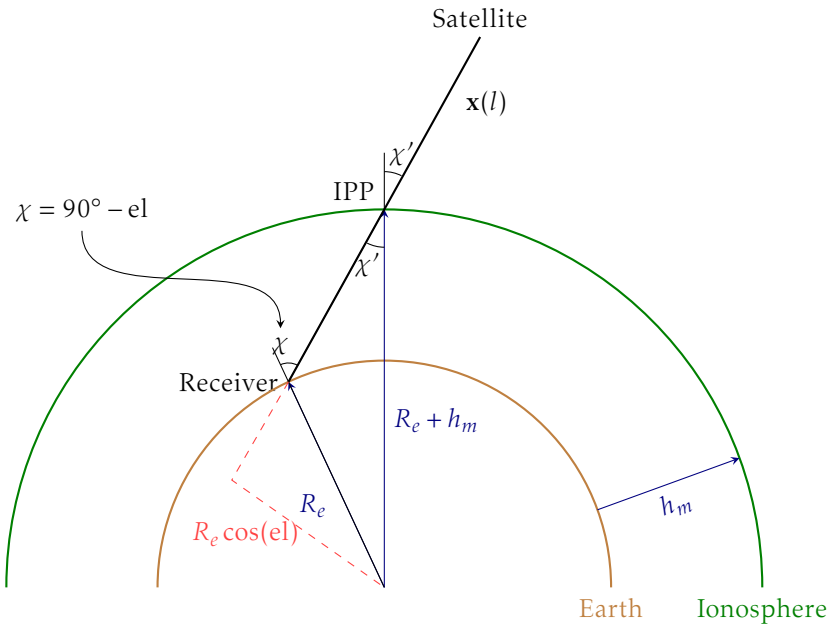


Figure 5·1: Geometry of TEC observations by ground-based GNSS receivers.

Signifying the vertical TEC as the special case when el = 90°, i.e.

$$\mathrm{vTEC} \triangleq \int_{h_r}^{h_s} n_e(h)dh.$$

the slant TEC at some other position $\mathbf{s}_0$ is approximated by mapping vTEC to $\mathbf{s}_0$ in a similar way to the Earthbound case:

$$\widehat{\mathrm{slTEC}} = \left[1 - \left(\frac{R_e \cos(\mathrm{el})}{R_e + h_m}\right)^2\right]^{-1/2} \mathrm{vTEC}. \tag{5.3}$$

The data obtained from MIT Haystack contains estimates of both slant and vertical TEC. They also include latitude and longitude coordinates for each receiver and IPP, for a thin shell at altitude 335 km.

5.3 Global Prediction of TEC from GNSS measurements

The data consist of slTEC estimates, error estimates, and the latitude and longitude positions of the corresponding IPP's. Details of that stage of estimation are discussed in Rideout and Coster (2006). These data are registered in 30 s intervals, consisting of typically $N \sim 15\,000$ TEC estimates. covering the globe. The goal is

- to generate predictions based on these data,

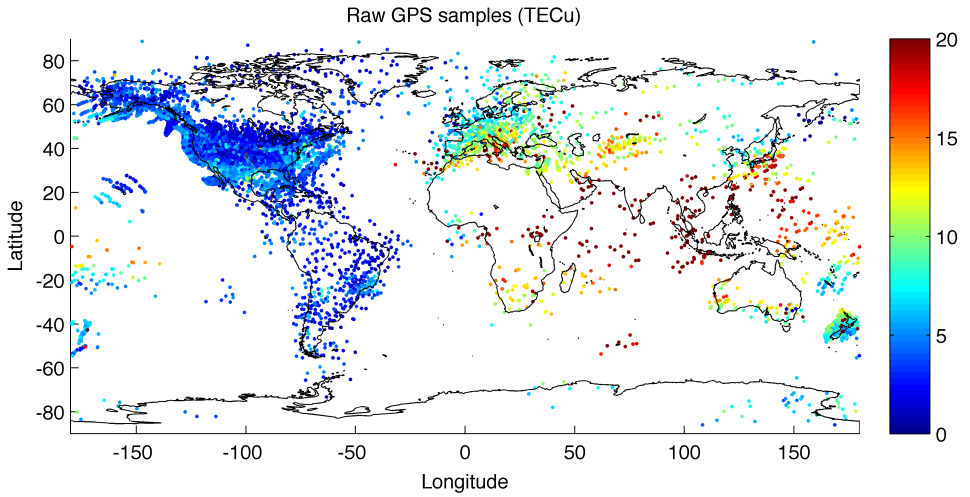


Figure 5·2: Estimated zenith-aligned total electric content (vTEC) from 24 March, 2009. Each point is a receiver-satellite pair in the network, plotted at the corresponding ionospheric pierce point (IPP). North America is covered with receivers and very densely sampled, while the oceans points above the ocean are few and associated with islands.

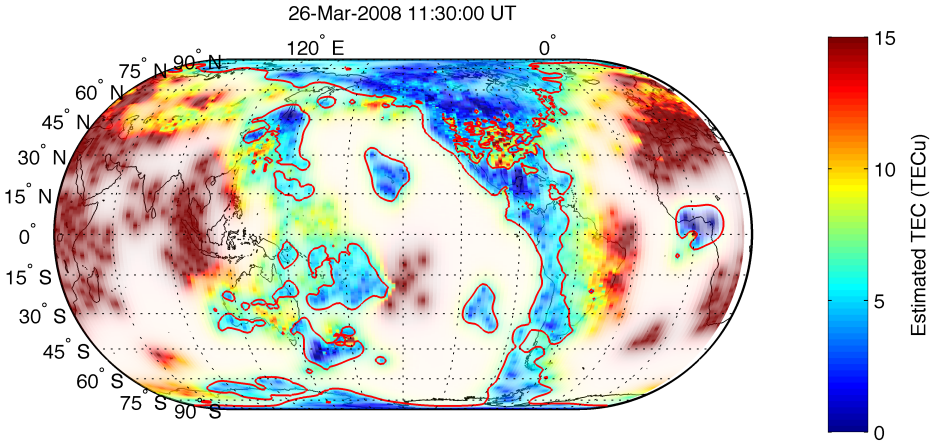


Figure 5·3: Predicted vTEC with a transparency mask mapped to $\hat{\sigma}^2_{OK}$.

- to map the predictions, and
- to provide an intuitive visual cue of the uncertainties.

As in previous chapters, this last objective sets kriging apart from deterministic interpolation. The "variance" component is an important distinguishing feature of kriging, and of spatial statistics in general, to quantify uncertainty and how it relates, for example, to the the sample coverage of a region.

For instance, consider the data set of Figure 5·2, where each vTEC is plotted versus position. In this case, much of the Earth is completely unsampled. In the absence of data, the ordinary kriging predictor shrinks to the estimated mean $\hat{\mu}$ and kriging variance reaches its maximum. To relay this level of uncertainty to the viewer in an intuitive way, we use transparency as a raster image equivalent of an error bar. Transparency is an intuitive option (Wilkinson, 2005). In Figure 5·3, each pixel is assigned a color ($\widehat{y}_{ok}(\cdot)$) and a transparency ($\propto \widehat{\sigma}^2_{ok}(\cdot)$). This preserves detail where it is available without unduly suggesting a trend not supported by the data.

Finally, the shortest distance between points on a sphere is the great-circle arc between them. (Euclidean distance can be an approximation for small distances and near the equator.) That is for standpoint $s = (\phi_s, \lambda_s)$ and forepoint $f = (\phi_f, \lambda_f)$ on a sphere, where ϕ and λ are latitude and longitude, respectively, the proper distance metric in the thin shell model is

$$d_{\mathrm{GC}}(f, s) = (R_e + h_m) \arctan\left(\frac{\mathbf{n}_f \times \mathbf{n}_s}{\mathbf{n}_f \cdot \mathbf{n}_s}\right), \tag{5.4}$$

where $R_e + h_m$ is the radius of the thin shell (R_e = Earth radius, and h_m is the height of the sphere above Earth's surface) as above, and $\mathbf{n}_{f,s}$ are unit normal vectors at the corresponding (ϕ, λ) coordinates. The cross product and dot product can be evaluated in cartesian coordinates following the usual transformation from the unit sphere:

$$\mathbf{n} = \begin{bmatrix} \cos\phi \cos\lambda \\ \cos\phi \sin\lambda \\ \sin\phi \end{bmatrix}. \tag{5.5}$$

5.4 Modeling the thin-shell ionosphere

Process model

The vTEC on the ionospheric thin shell is modeled with a mixed effect model

$$\underline{Y}(s) = \mathbf{X}^{\mathsf{T}}(s)\underline{\beta} + \underline{\delta}(s) \tag{5.6}$$

with a linear mean effect $\mathbf{X}^\mathsf{T}\underline{\beta}$ and random spatial effect $\delta \sim \mathcal{GP}(0, C_Y)$, a Gaussian process specified by a covariance function C_Y. The fixed effect must be periodic on the sphere in order to avoid introducing discontinuities. In cartesian coordinates, $\mathbf{X}$ is often a polynomial of the coordinates, The only polynomial satisfying the periodicity requirement is the constant function. Ordinary kriging will do. Other explanatory variables could also be considered, in which case $X(s)$ is a more complicated function, and universal kriging is needed. Finally, the covariance function should be one of those identified as valid on the sphere. Jun and Stein (2007) identify some isotropic covariance functions that are also valid on the sphere. The exponential covariance is one of these, and it is used in this example.

Data model

The vTEC estimates provided within MAPGPS at this stage are accompanied by the slant TEC estimates from which they were derived, along with the associated error. Propagating the slant error through (5.3), the corresponding vTEC error is

$$\sigma_v^2 = \left[1 - \left(\frac{R_e \cos(\text{el})}{R_e + h_m}\right)^2\right] \sigma_{sl}^2. \tag{5.7}$$

Gathering these values into a diagonal matrix Σ_e provides the covariance function for our assumed Gaussian data model:

$$\underline{Z}|\underline{Y} \sim \mathcal{N}(\underline{0}, \Sigma_e). \tag{5.8}$$

5.5 Prediction

The model above describes a Gaussian process. Stack vTEC into a vector $\underline{Z}$ and error variances into the diagonal matrix Σ_e. Prediction can then be carried out at an arbitrary reconstruction point $\mathbf{s}_0$ by following the universal kriging (UK) procedure:

1. Since error variances are known, we form a preliminary (weighted least squares) estimate of the fixed effect coefficients: $\underline{\hat{\beta}}_{\text{WLS}} = \left(\mathbf{X}^\mathsf{T}\Sigma_e^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\Sigma_e^{-1}\,\underline{Z}$,
2. Fit the variogram of the residuals: $\bar{\gamma}\left(\underline{Z} - \mathbf{X}^\mathsf{T}\underline{\hat{\beta}}_{\text{WLS}}\,;\,\underline{\theta}^*\right)$.
3. Revise the mean estimate via GLS using $(\mathbf{C}_Z)_{ij} = C_Y(\mathbf{s}_i, \mathbf{s}_j\,;\,\underline{\theta}^*) + \Sigma_e$:
4. Evaluate the UK predictor and variance at each $\mathbf{s}_0$:

$$\widehat{y}(\mathbf{s}_0) = \underline{\hat{\beta}}_{\text{GLS}} + \underline{c}_Y^\mathsf{T}(\mathbf{s}_0)\mathbf{C}_Z^{-1}\left(\underline{Z} - \underline{\hat{\beta}}_{\text{GLS}}\mathbf{X}\right)$$

$$\widehat{\sigma}^2_{\text{uk}}(\mathbf{s}_0) = C_Y(\mathbf{s}_0,\mathbf{s}_0) - \underline{c}_Y(\mathbf{s}_0)^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{c}_Y(\mathbf{s}_0) + \left(\underline{x}(\mathbf{s}_0) - \mathbf{X}^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{c}_Y(\mathbf{s}_0)\right)^{\mathsf{T}}\left(\mathbf{X}^{\mathsf{T}}\mathbf{C}_Z^{-1}\mathbf{X}\right)^{-1}\left(\underline{x}(\mathbf{s}_0) - \mathbf{X}^{\mathsf{T}}\mathbf{C}_Z^{-1}\underline{c}_Y(\mathbf{s}_0)\right)$$

5.6 Reassessment of an earier case study

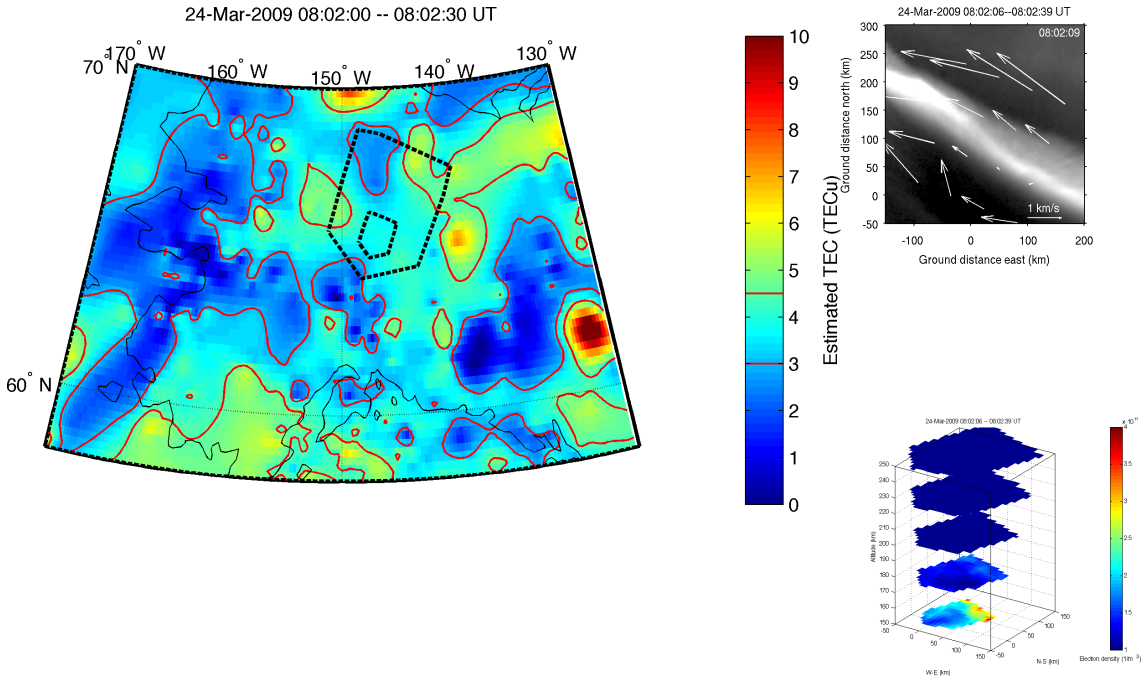
Having available a global dataset directly analogous to Chapter 3's maps of radar-derived electron densities presents an irresistible opportunity to revisit those results and discover how they fit within a global context.

In Chapter 4 we noticed some congruencies between radar and optical data. Although regions of enhanced ionization are often found near auroral arcs (when the latter are present), and although a reasonable case can be made for the association of certain flow fields with simple auroral morphologies, these relations need not always hold. As seen in previous chapters, such models often do bear reliable predictive and explanatory value. However, the ionosphere is frequently driven to an excited state (e.g. during a substorm) such that idealized assumptions are invalid. Optical forms need not map directly to ionization.

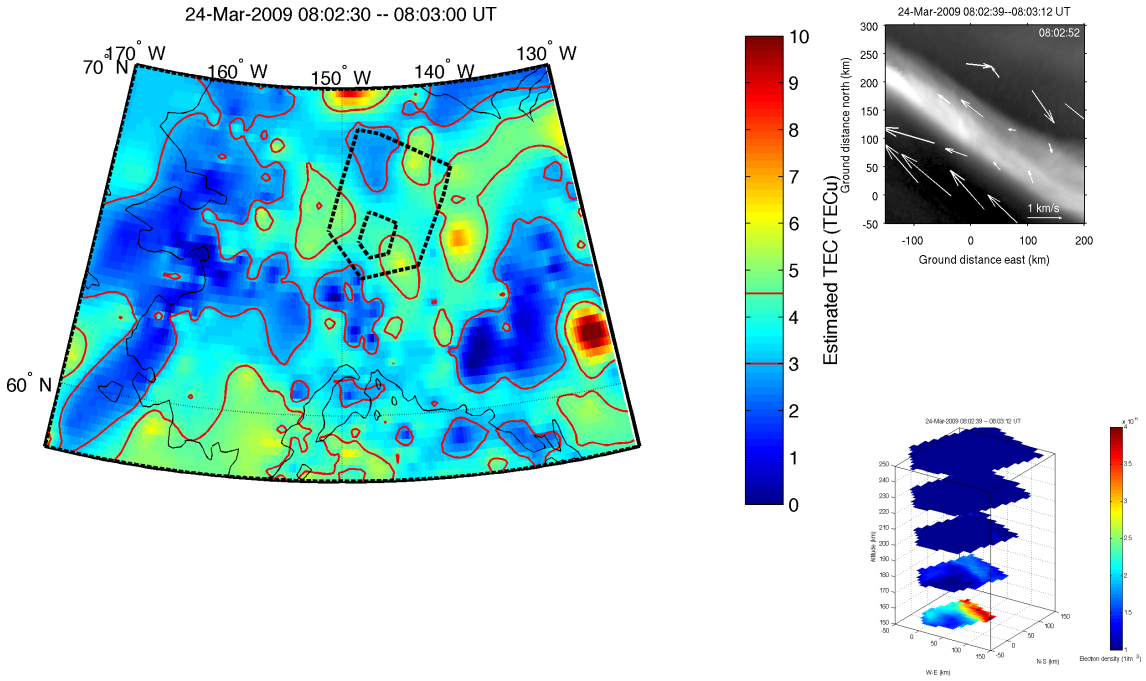
On the other hand, TEC is explicitly related to electron density via the integral $\int dh\, n_e(h)$. There should be a close correspondence between the two. In particular, by forming a discrete approximation of the above integral, we can compare directly the estimate of vertical TEC by two instruments: PFISR measuring electron density versus height, and the GPS receiver network with TEC mapped to zenith. Further, whenever a GPS raypath passes through the radar f.o.v., we can approximate the slant TEC integral (5.1).

24 March 2009: Ionospheric structure around an auroral arc during growth and expansion phases of a magnetic substorm

On this night, a stable auroral arc held its position within PFISR's field of view (f.o.v..). Around 0804 UT, the spatial configuration of Doppler velocities led the fitter to suggest a sudden, transient reversal of plasma flow.

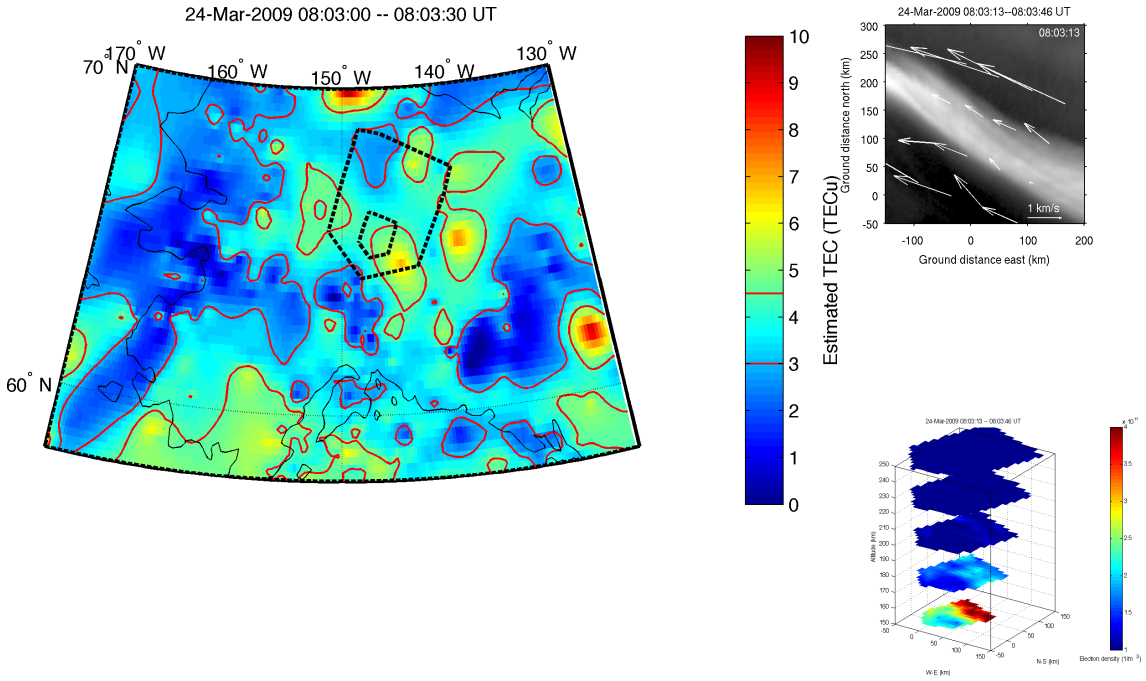


(a) 08:02:00 UT

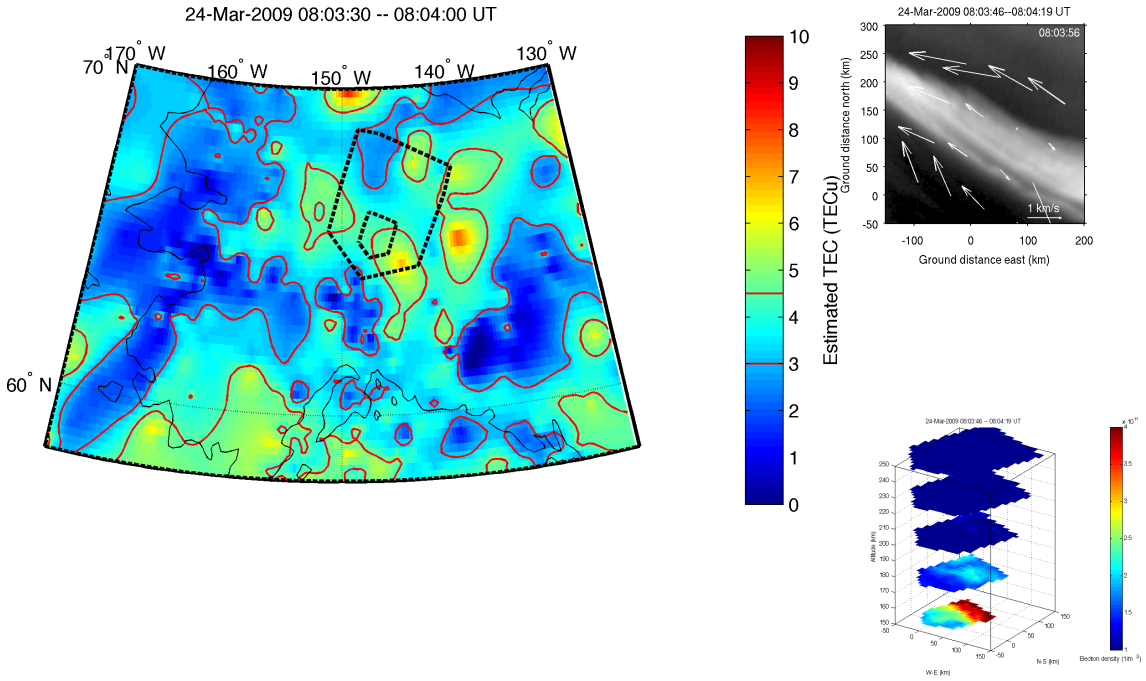


(a) 08:02:30 UT

Figure 5·5: Comparisons of GNSS TEC (ordinary kriging), optical data, radar-derived flow field, and radar N_e. Time sequence from 24 March, 2009. Time resolution: $\sim 30\,\mathrm{s}$

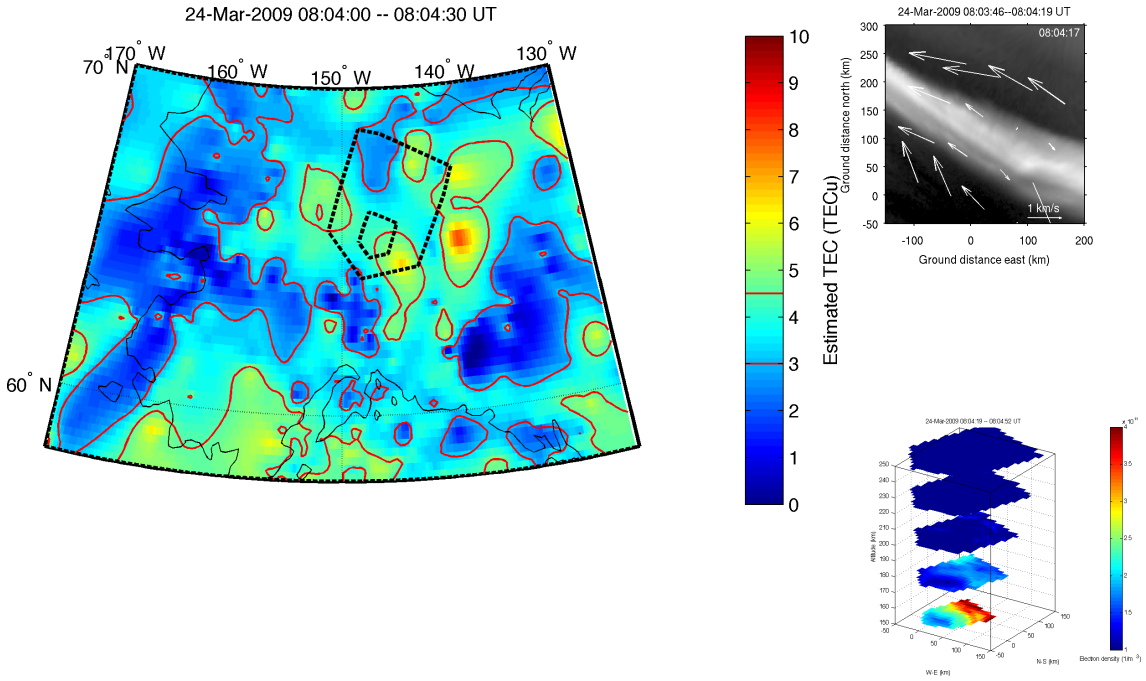


(b) 08:03:00 UT

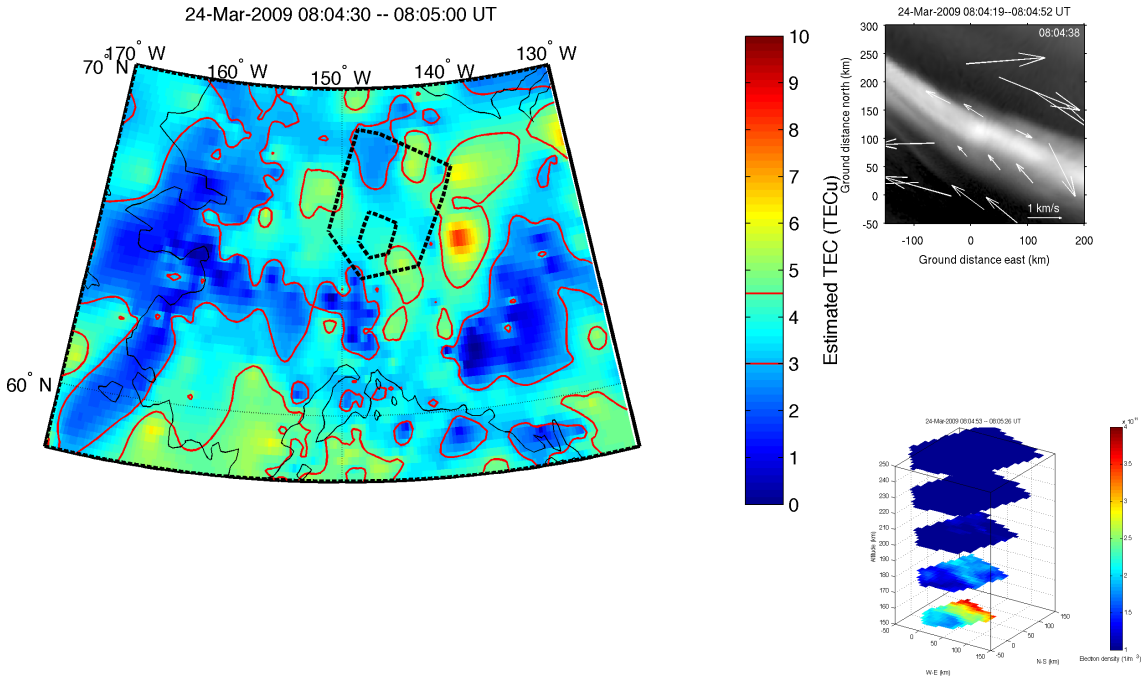


(c) 08:03:30 UT

Figure 5·5: Comparisons of GNSS TEC (ordinary kriging), optical data, radar-derived flow field, and radar N_e. Time sequence from 24 March, 2009. Time resolution: $\sim 30\,\mathrm{s}$



(d) 08:04:00 UT



(e) 08:04:30 UT

Figure 5·5: Comparisons of GNSS TEC (simple kriging), optical data, radar-derived flow field, and radar N_e. Time sequence from 24 March, 2009. Time resolution: $\sim 30\,\mathrm{s}$

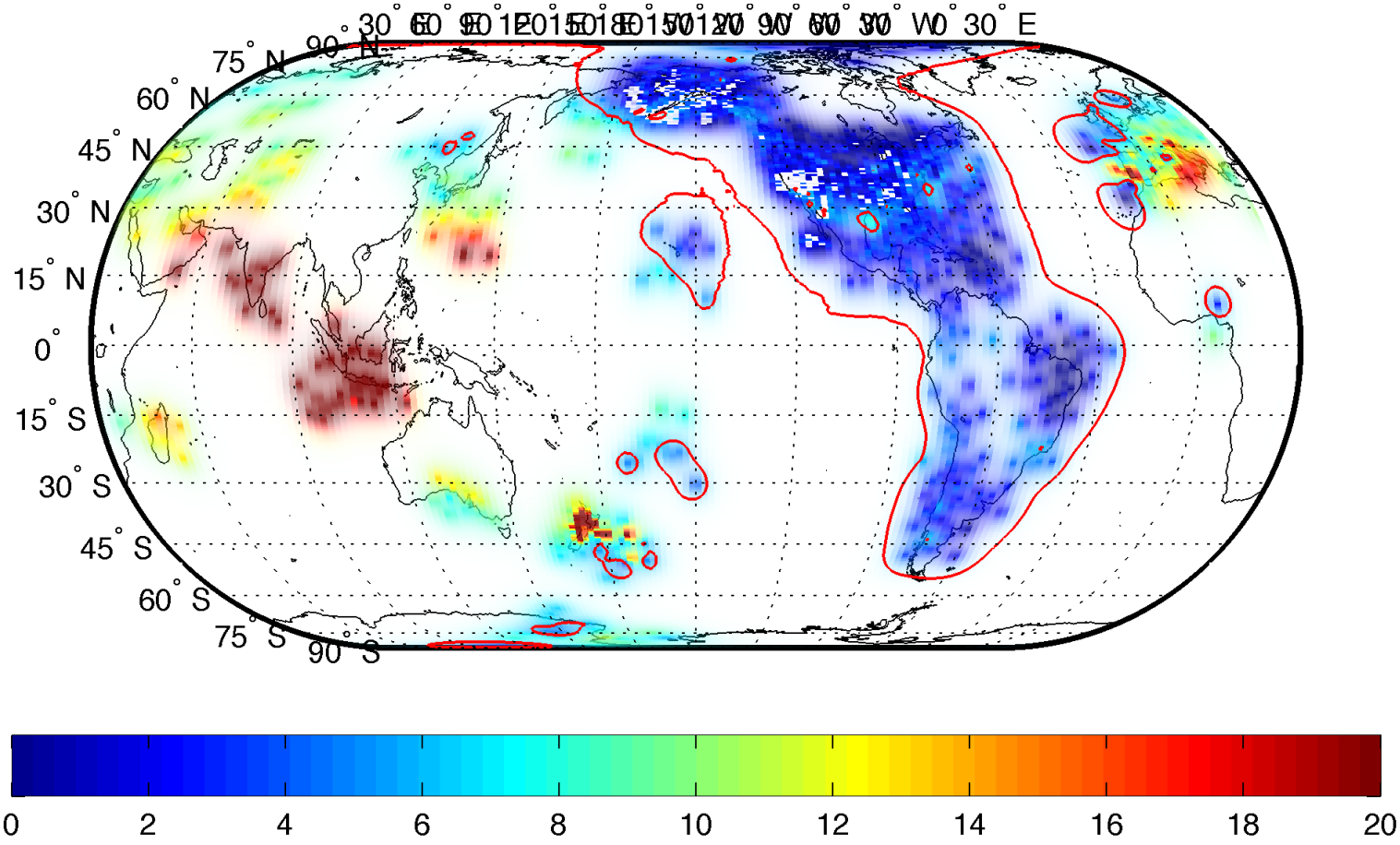


Figure 5·6: High-resolution GNSS TEC (ordinary kriging) for the same approximate time period as the previous figure: 24 March, 2009. Time resolution: 5 min.

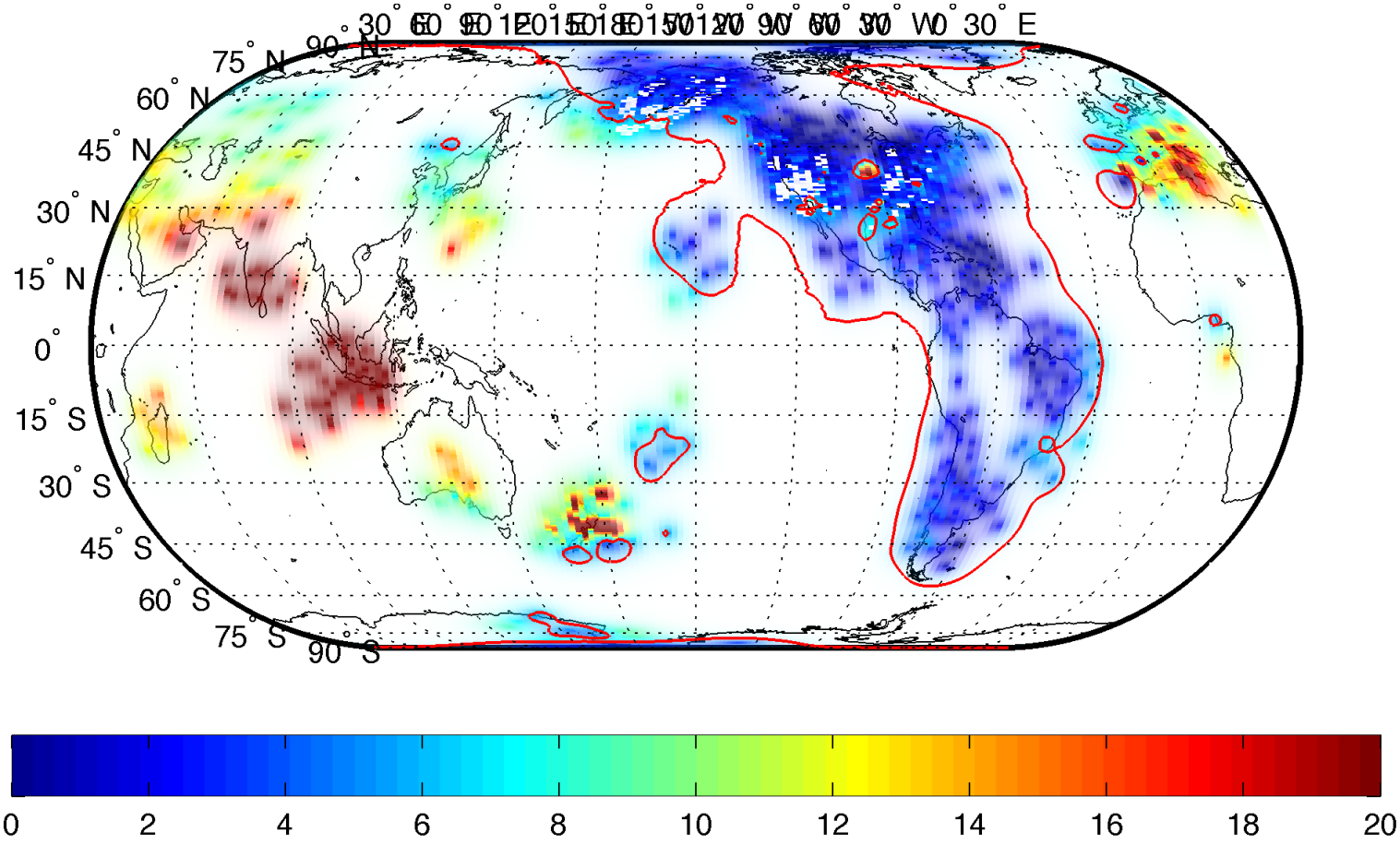


Figure 5·7: High-resolution GNSS TEC (ordinary kriging) for the same approximate time period as the previous figure: 24 March, 2009. Time resolution: 5 min.

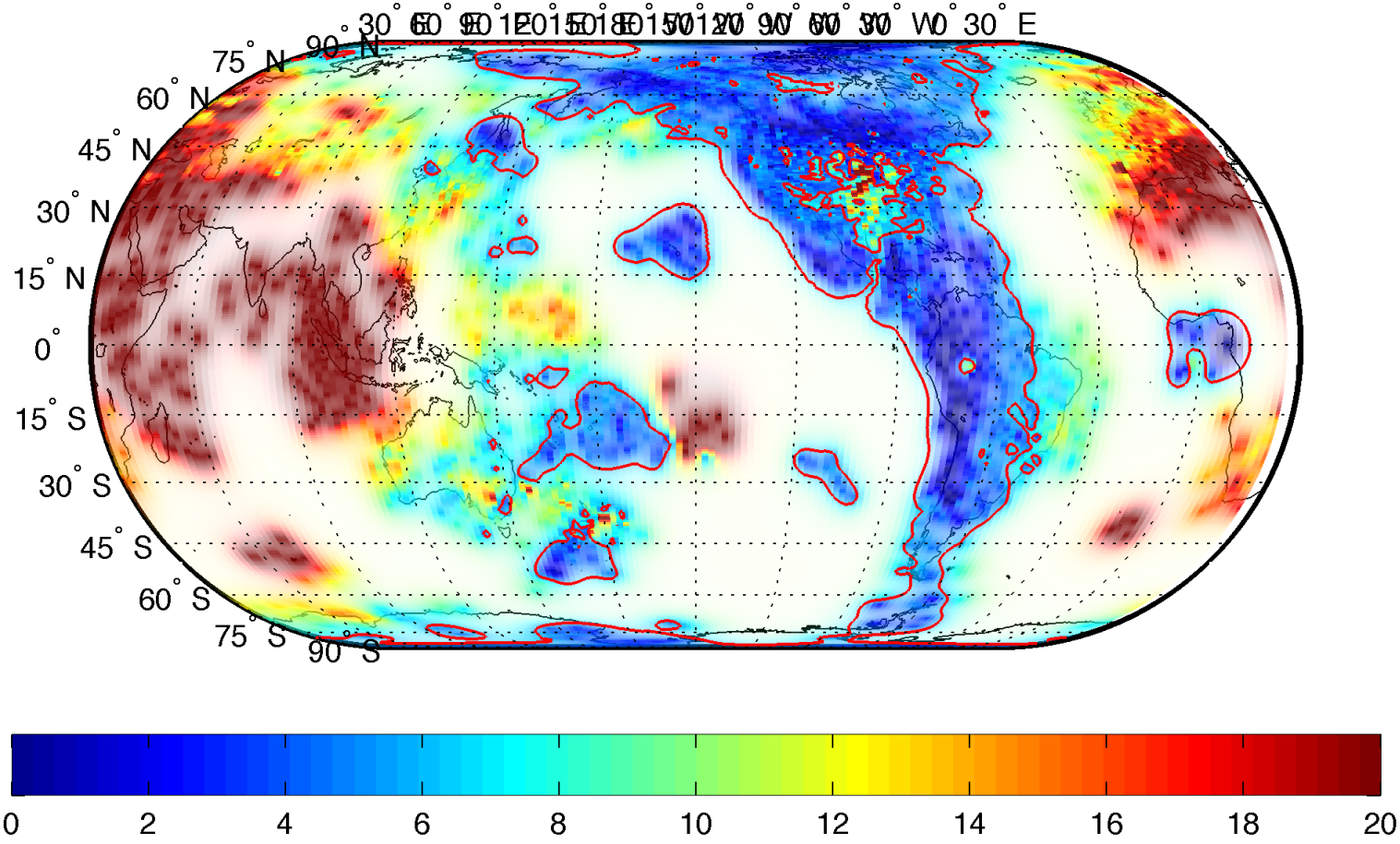


Figure 5·8: High-resolution GNSS TEC (ordinary kriging). Different date: 26 March, 2008. Time resolution: $10\,\mathrm{min}$.

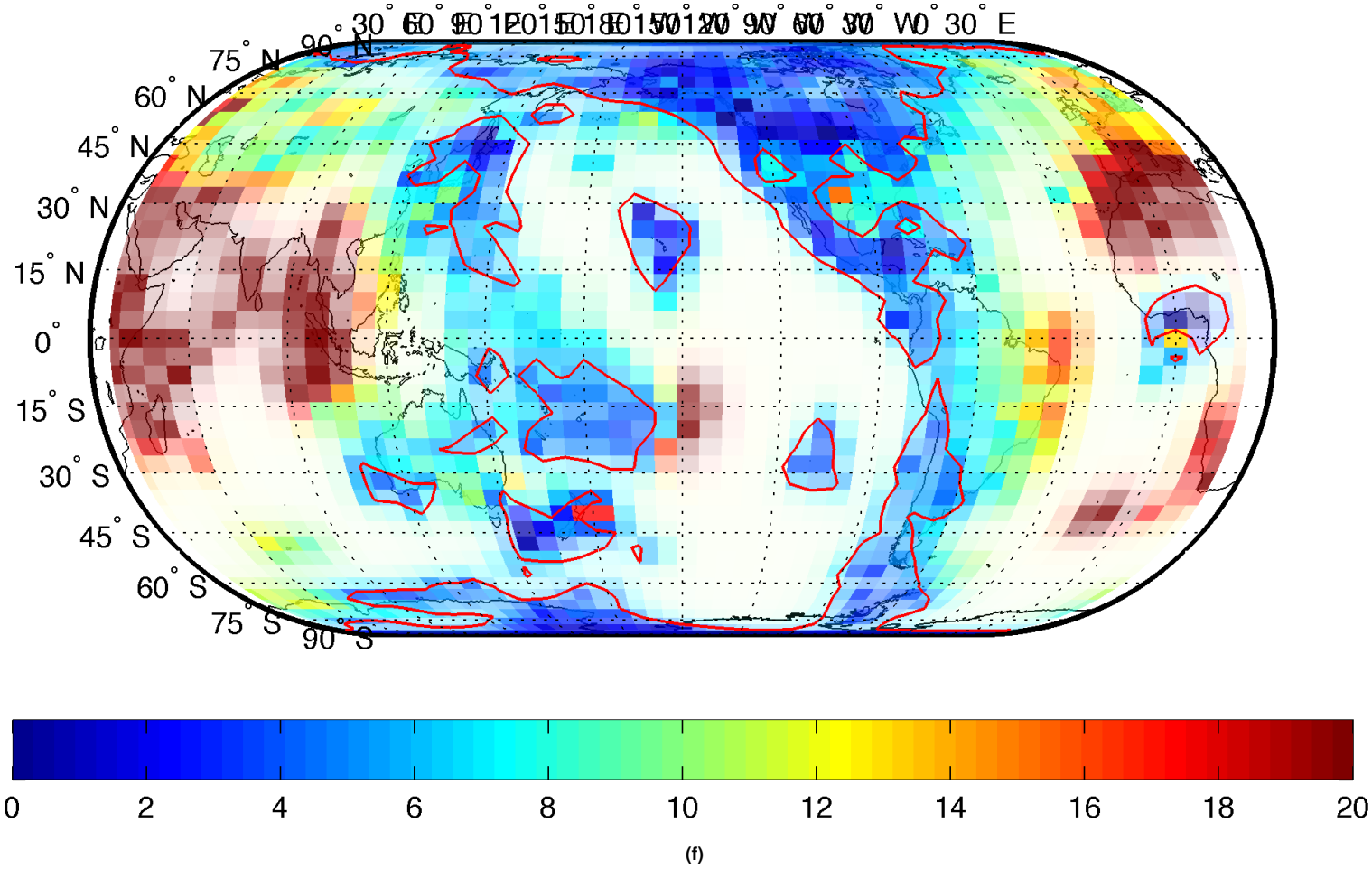


Figure 5·7: Low-resolution GNSS TEC (ordinary kriging). Different date: 26 March, 2008. Time resolution: $10\,\mathrm{min}$. This image uses data from all receivers, but maps to low resolution.

5.7 Challenges particular to global prediction

In some respects, global prediction is similar to spatial prediction on the small, flat domain, differing (rather importantly!) in two fundamental attributes of Earth: size and shape. The diagram of Figure 5·8 summarizes the effect on geostatistical modeling of global data. The size of the domain D_s need not be a fundamental difference, if the measurements' support scales proportionally with the domain size. Instead, our geospatial measurements tend to be fixed at human-order scale, even as they spread to global coverage. resolve processes on a scale similar to those of a smaller domain may require very high spatial resolution. Problems of shape emerge from the spherical topology of global data, which places restrictions on what models are valid.

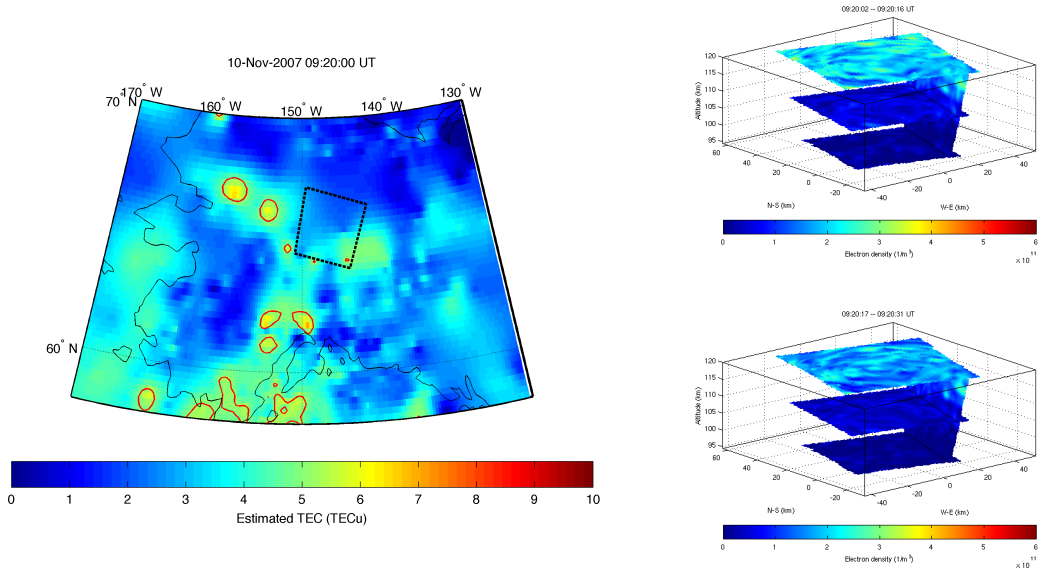
Size

Latent processes. Previous chapters assumed that distance was sufficient to predict natural processes over a given domain. But the dependence among sites may depend more on environmental factors (such as temperature or elevation) that are not well-predicted by distance (Le and Zidek, 2006, p.70). On the global scale, these factors may interact in complex ways. If these can be reasonably incorporated, the process model (5.6) should be augmented to reflect this dependence.

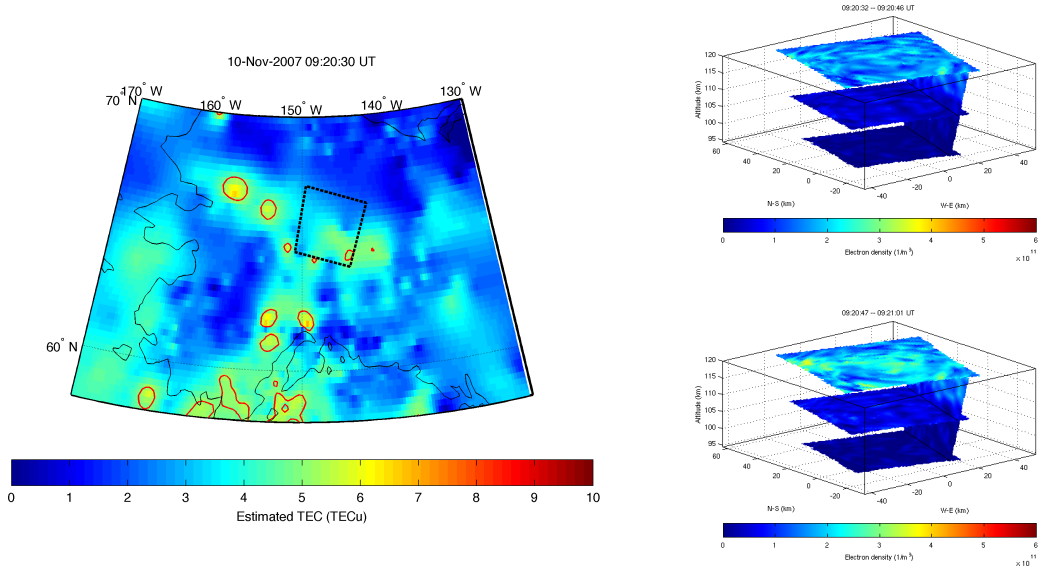
Data aggregation. It may be possible to directly reduce the number of data needed. This may involve random or systematic subsampling. Or nearby data can be combined into aggregate measurements. Of course, this reduces the effective spatial resolution.

Data reduction. Kriging on limited neighborhoods[1] is a divide-and-conquer technique, essentially restricting the size of the problem by throwing away or reducing the influence of data outside a local neighborhood, then building up the global prediction from the predictions on the sub-regions. Neighborhood selection is a problem-specific challenge: depending on the process model and the dispersion of data, prediction restricted to the smaller domain is not guaranteed to be consistent with respect to the full problem. If the dimensions of the neighborhood fall well below the process scale, the neighborhood's covariance matrix $\mathbf{C}_Z$ may be singular to within machine precision. Finally, even if every subproblem on every neighborhood is valid and soluble, the

[1] Kriging on a neighborhood is distinguished from kriging locally or regionally. In the latter case, the goal is only to interpolate within the region. The neighborhood method assembles a prediction from smaller sub-region predictions, explicitly limiting the number of data used in each sub-region

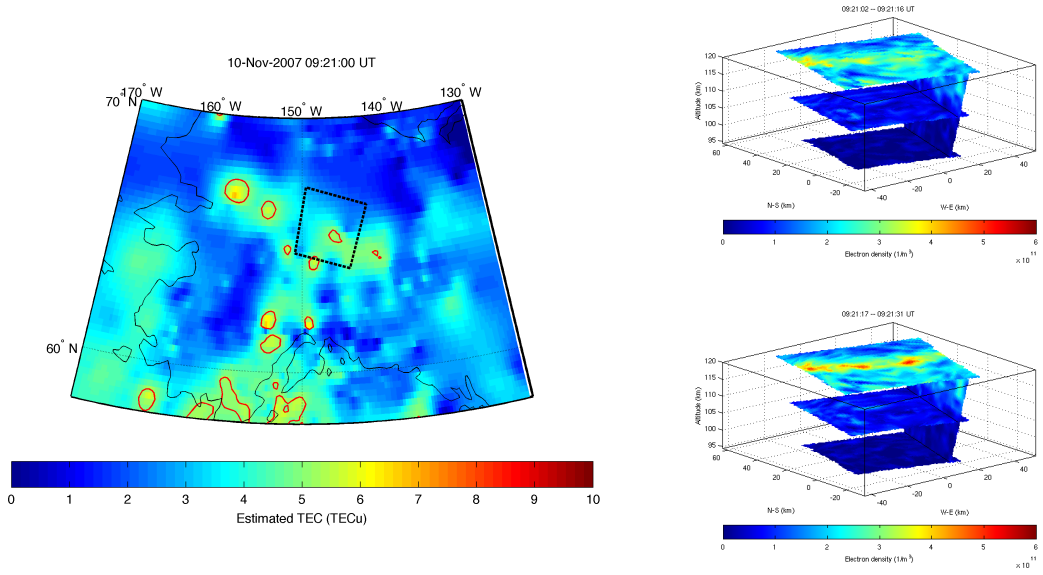


(a) 09:20:00 UT

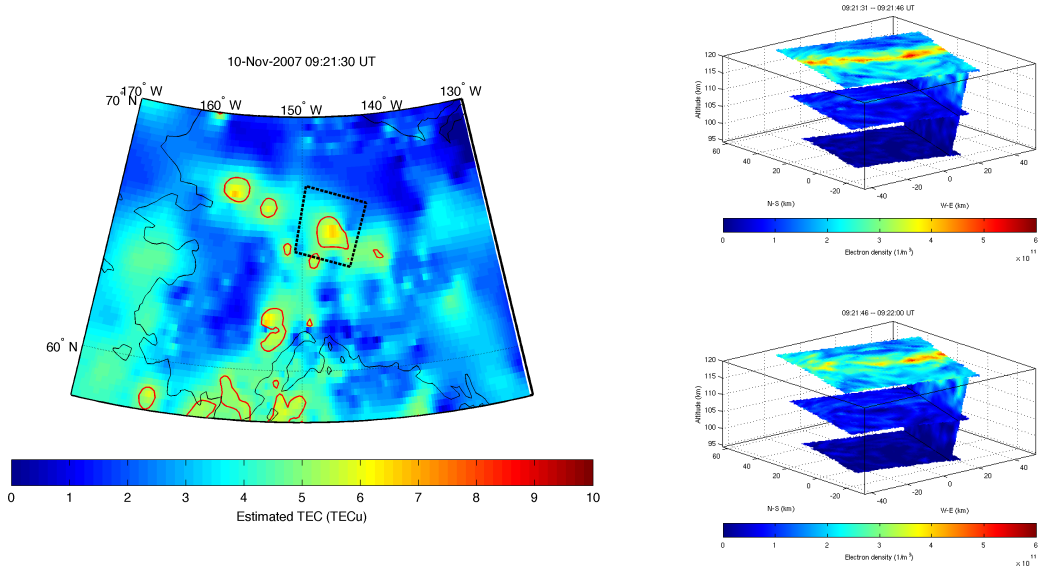


(a) 09:20:30 UT

Figure 5·9: (Left) GNSS-TEC. (Right) ISR. 10 November, 2007.

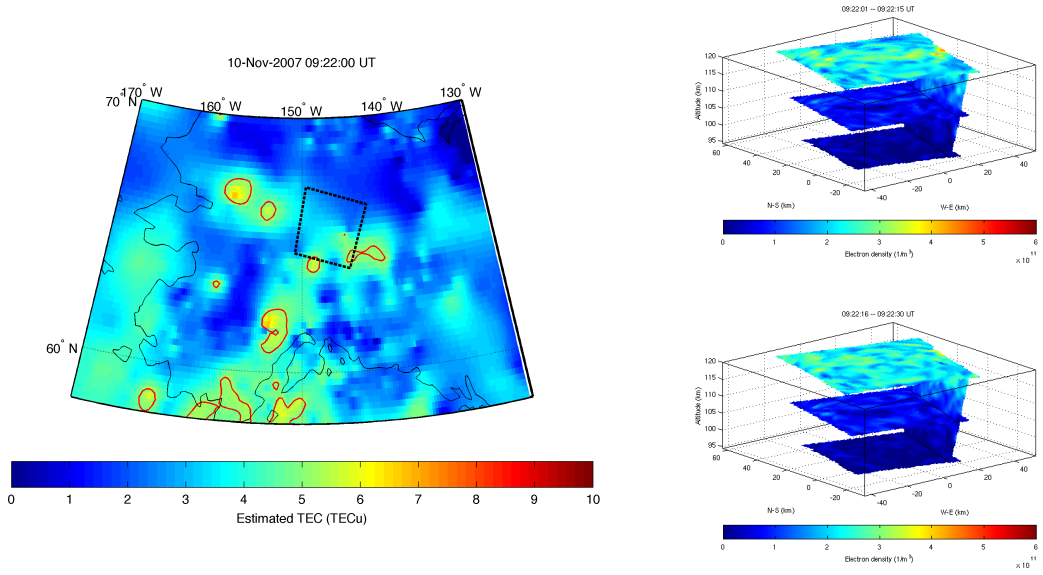


(b) 09:21:00 UT

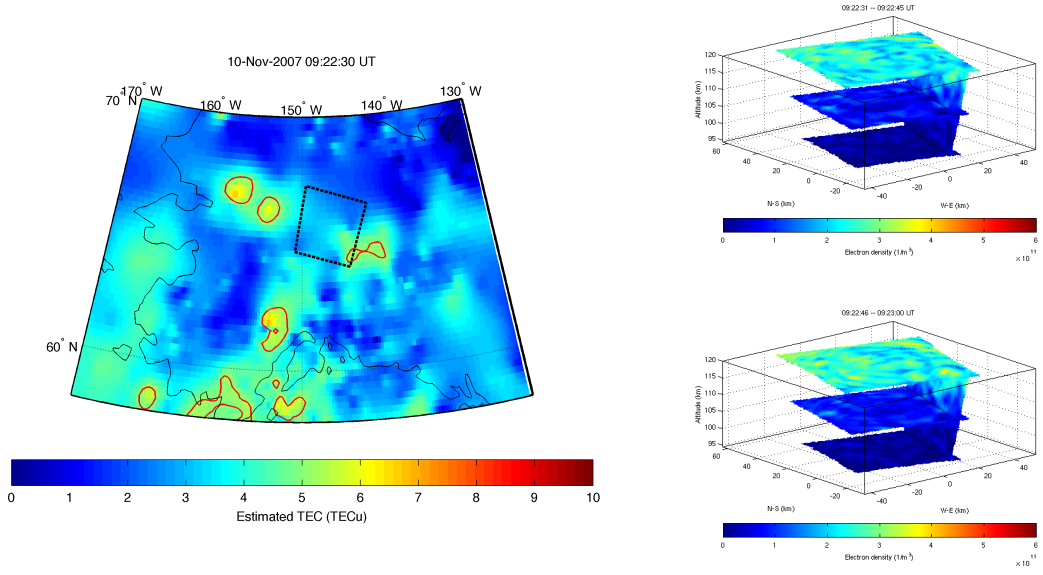


(c) 09:21:30 UT

Figure 5·9: (Left) GNSS-TEC. (Right) ISR. 10 November, 2007.

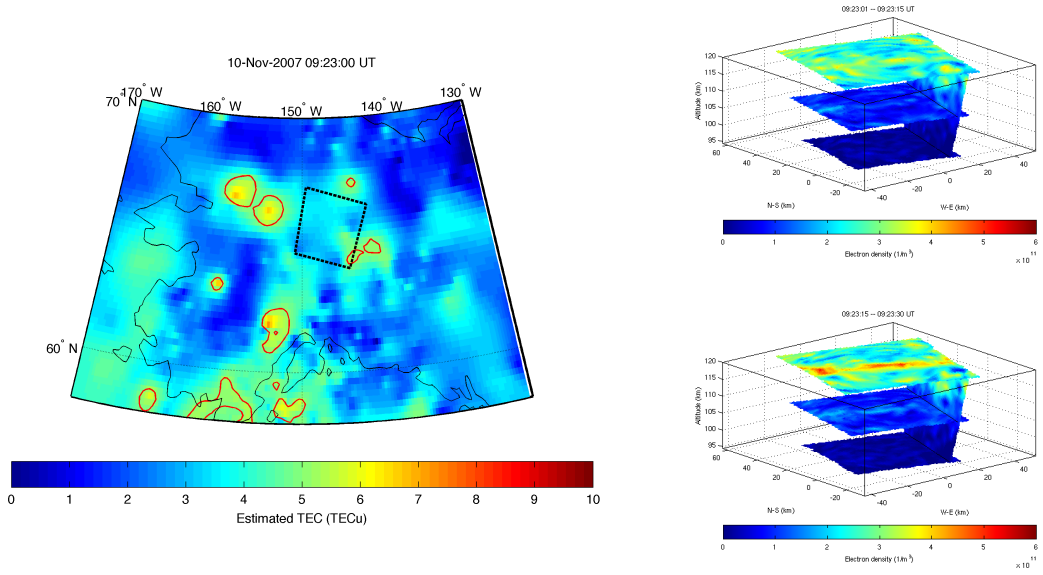


(d) 09:22:00 UT

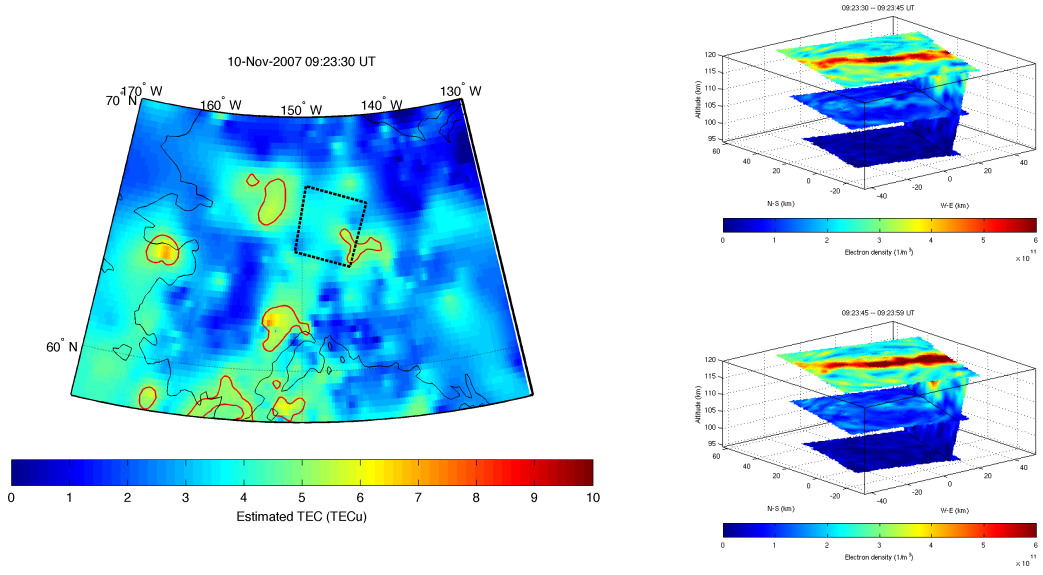


(e) 09:22:30 UT

Figure 5·9: (Left) GNSS-TEC. (Right) ISR. 10 November, 2007.

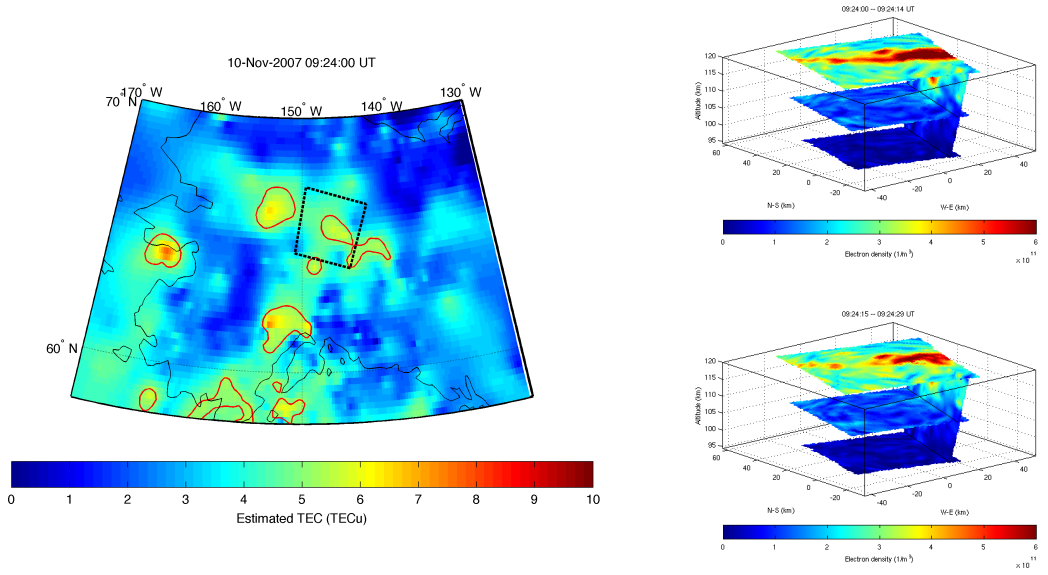


(f) 09:23:00 UT

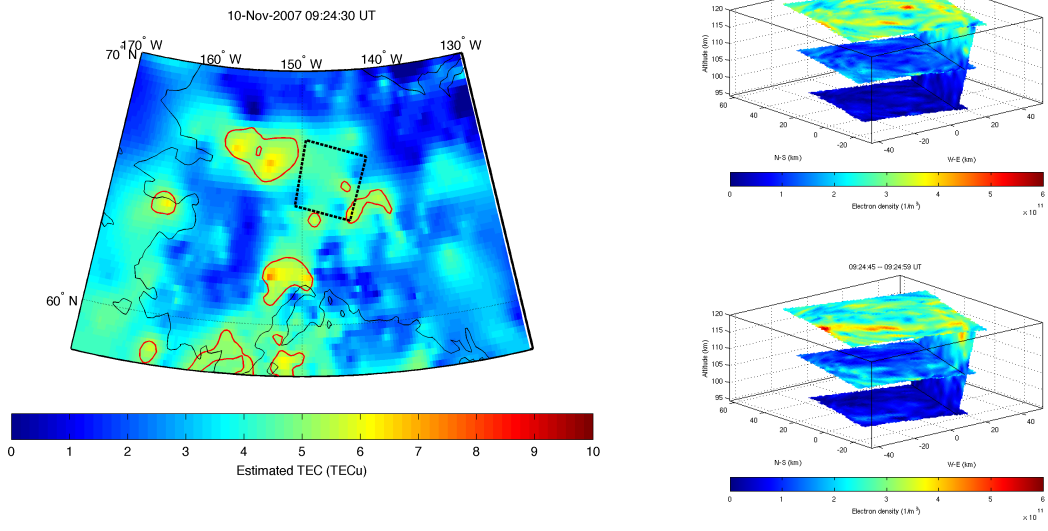


(g) 09:23:30 UT

Figure 5·9: (Left) GNSS-TEC. (Right) ISR. 10 November, 2007.



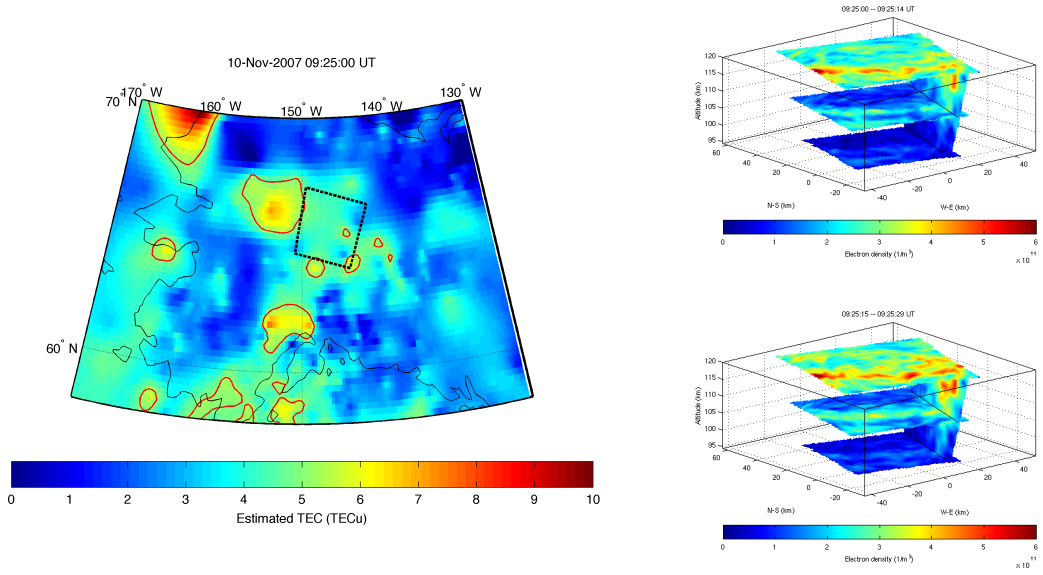
(h) 09:24:00 UT



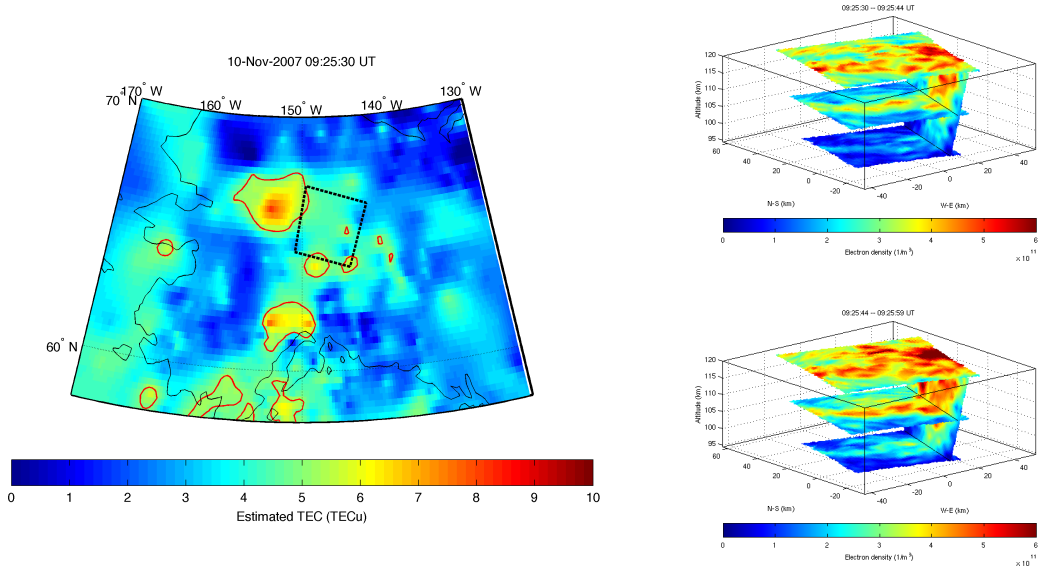
(i) 09:24:30 UT

(j) ISR

Figure 5·9: (Left) GNSS-TEC. (Right) ISR. 10 November, 2007.

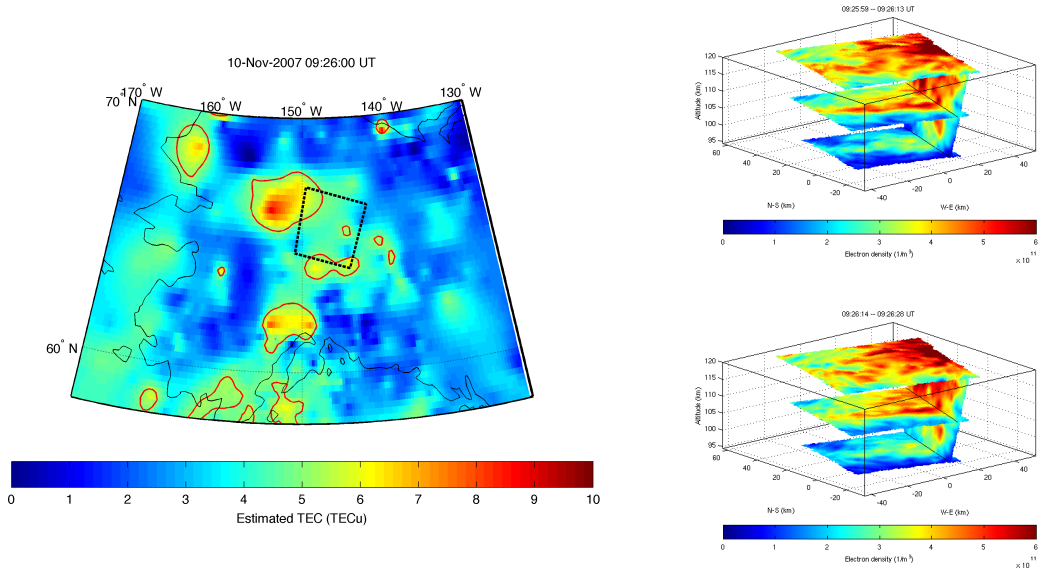


(k) 09:25:00 UT

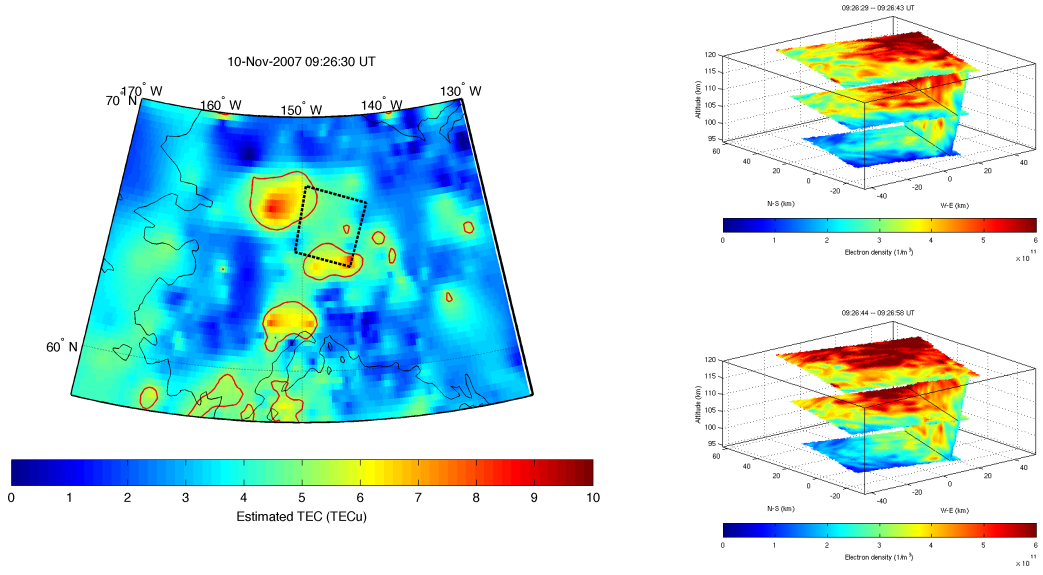


(l) 09:25:30 UT

Figure 5·9: (Left) GNSS-TEC. (Right) ISR. 10 November, 2007.

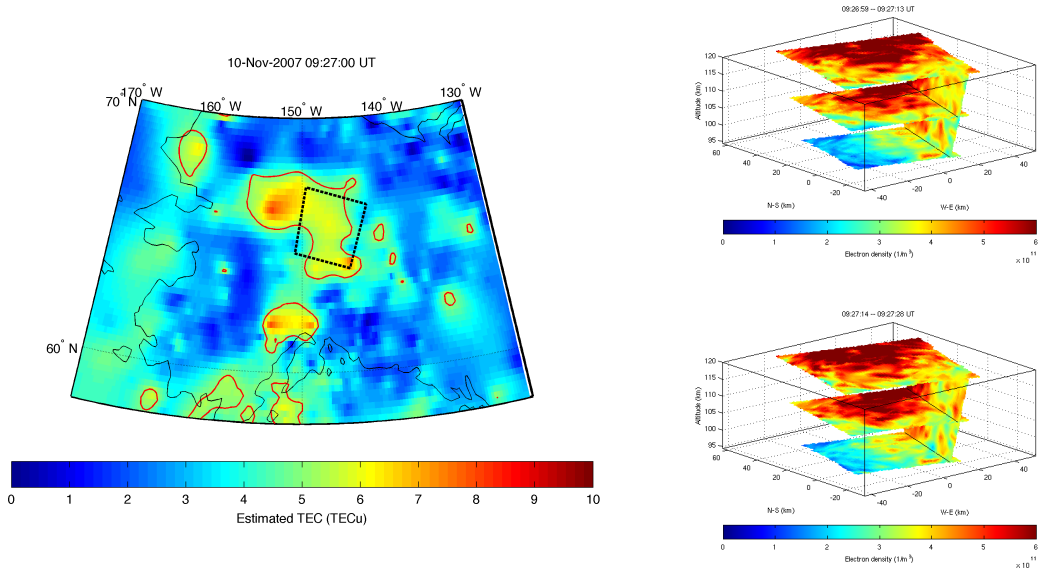


(m) 09:26:00 UT

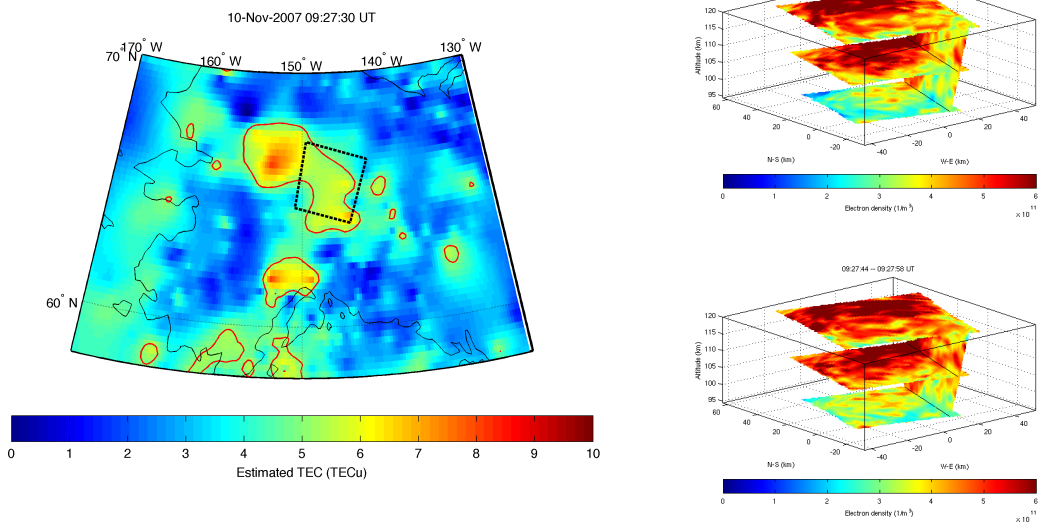


(n) 09:26:30 UT

Figure 5·9: (Left) GNSS-TEC. (Right) ISR. 10 November, 2007.



(o) 09:27:00 UT



(p) 09:27:30 UT

Figure 5·9: (Left) GNSS-TEC. (Right) ISR. 10 November, 2007.

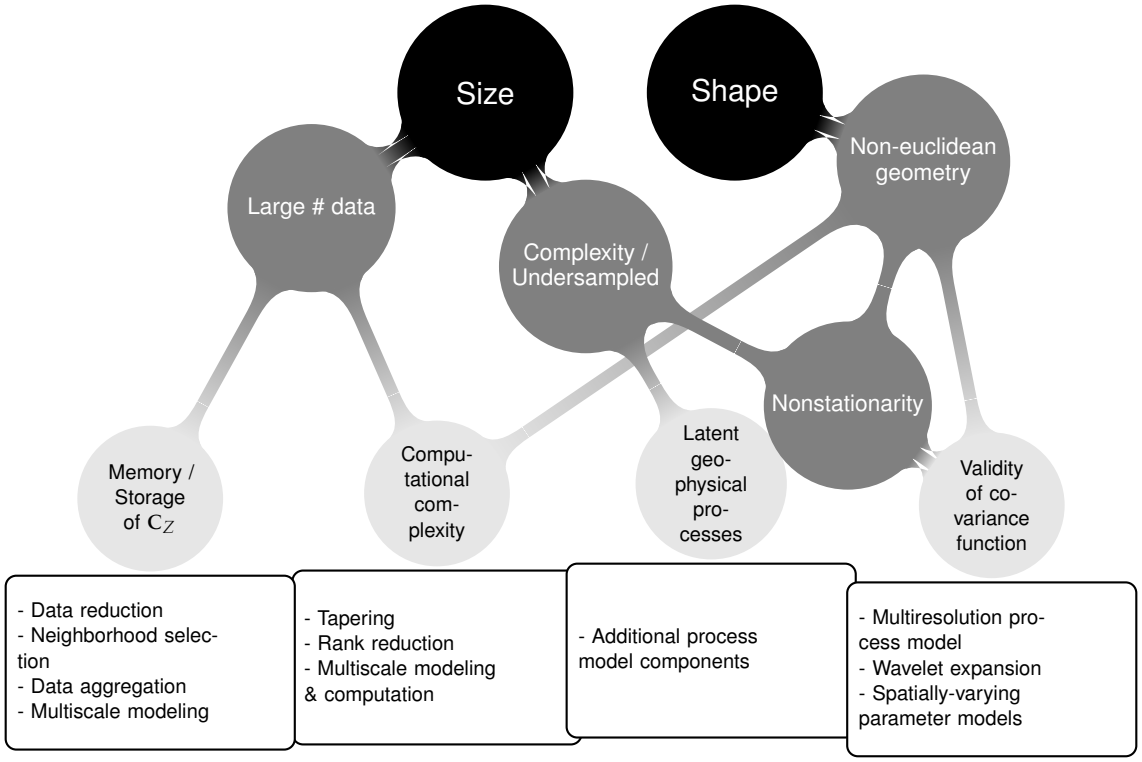


Figure 5·10: Geostatistical modeling of global data. Compared to local domains, estimation and prediction on the globe are complicated by (1) a) very large-scale spatial processes rendered unobservable by individual instruments, b) smaller-scale processes are not captured by aggregated data (effective spatial resolution), and (2) the topology of the globe presents further modeling constraints.

aggregated global solution does not necessarily have a valid covariance matrix.[2] This can introduce non-physical artifacts in $\widehat{y}(\cdot)$.

Approximating $\mathbf{C}_Z$. The dominant factor of memory and computational requirements of kriging is storage and inversion of the covariance matrix. For m scalar measurements, a naïve implementation of the kriging predictor stores the $m \times m$ covariance matrix $\mathbf{C}_Z$. Computing the inverse requires $O(m^3)$ operations. Strategies to reduce the computational complexity of kriging follow a common theme: reduce the rank of $\mathbf{C}_Z$. For example, covariance tapering replaces $\mathbf{C}_Z$ with a sparse approximation so that $\mathbf{C}_{ij} = 0$ if the distance between points i and j is sufficiently large (Furrer et al., 2006).

[2] The *global* covariance matrix may not be valid, even if those of the neighborhoods are. (Paciorek and Schervish, 2004) describes how to construct a valid, nonstationary covariance function from linear combinations of stationary covariance functions.

Low-rank representation. Often, however, $\mathbf{C}_Z$ is inherently structured such that it possesses an equivalent, low-rank representation. For example, Cressie and Johannesson (2006, 2008) develop a fixed-rank version of kriging, relying on the well-known Woodbury matrix identity to reduce the operation count of the inversion to $O(Nk^2)$, where $k \leq N$. Wikle (2010) discusses choices of basis functions and the advantages of choosing a low-rank representation.

Aside from these kriging-specific solutions, general recommendations for efficient computing may also improve performance. All the kriging implementations in this work use in-place evaluation to avoid making multiple copies of $\mathbf{C}_Z$ in RAM. Also, large data structures can be thrown away when they are no longer needed. A library of efficient subroutines is invaluable, particularly for linear algebra and optimization. Modern programming environments (such as MATLAB and the Python package NumPy) often use these to facilitate cache-optimized vector operations (such as element-wise arithmetic primitives), and it is worth familiarizing oneself with those features as well.

Non-euclidean distance

Model validity. Kriging on the sphere is different from kriging in cartesian coordinates. First, the topology requires the prediction to be periodic. The (deterministic) trend model must be periodic to satisfy this requirement. Huang et al. (2011); Curriero (2006) also show that many covariance functions that are widely used on the plane fail to generate positive definite matrices, as required for kriging, on the sphere.

Computation. Since distance is measured along the great circle arc between points, computing distances on the sphere is slightly more expensive. Although modern computers can efficiently process vector operations, the additional cost could be significant for very large data.

Nonstationarity. A process with covariance function specified in spherical coordinates is inherently nonstationary (Jun and Stein, 2007). Note the strong latitude (ϕ) dependence in (5.5). The distance between two meridians varies with latitude. This makes fitting scale parameters more difficult, for instance. One approach to this form of nonstationarity is to develop a spatial model for the parameters (scale, shape, etc.) of the process model, and allow these to be estimated, say, at different latitudes.

Multiresolution modeling. The structure of the ionosphere is indeed nonstationary. This is especially so on the global scale, with regions and events characteristic of particular latitudes, altitudes, local time, and season. Yue et al. (2007) characterize the statistics of spatiotemporal dependencies in the ionosphere at various scales, confirming the nonstationarity of global geophysical processes. Nychka et al. (2002) discuss the implications of this on prediction and propose multiresolution methods to model global nonstationarity. (See also Ferreira and Lee (2007).) Multiresolution process models incorporate

5.8 Suggestions for improvement

In Section 5.4, several model assumptions were made. The mean effect $X^{\mathsf{T}}\underline{\beta}$ was assumed constant, but the ionosphere is quite often (and successfully) modeled by expanding in terms of spherical harmonics (e.g., Venkata Ratnam and Sarma, 2012). Replacing the columns of $\mathbf{X}$ with basis functions of the type

$$\sum_n \sum_m P_{nm}(\sin\phi)(a_{nm}\cos m\lambda + b_{nm}\sin\lambda$$

may better represent the trend (large-scale variability) when evaluating $\underline{\beta}_{\text{GLS}}$.

The ionosphere is strongly influenced by the position of the sun, making time an important explanatory variable. The process model should also include a component based on time.

Additionally, as demonstrated in Rideout and Coster (2006), unmodeled temperature dependence in each receiver may introduce systematic bias in the vTEC estimates that comprise the "data" $\underline{Z}$ in (5.8). As described in Chapter 2, predicting from estimates is a sub-optimal approach for precisely that reason: estimates are subject to both random errors and systematic biases, which may (or may not) be adequately compensated, and which may (or may not) be represented in the accompanying error estimate. The predictor should include (1) a more comprehensive model incorporating the stages of processing described by Rideout and Coster (2006) and (2) an allowance of leeway in estimating the parameters of that conversion. Namely, a Bayesian hierarchical model (Banerjee et al., 2004). Examples in similar applications include Wikle et al. (2003), Cressie et al. (2009), Kang and Cressie (2011), and Zidek et al. (2012).

Chapter 6

Suggestions for Further Study

In Chapter 1, we examined the capabilities of an electronically steerable ISR to resolve both fine and dynamic features of the ionosphere. Using insight gleaned from spatial statistics, we demonstrated a linear filter approach to predict electron density at unmeasured locations.

One powerful application of this method of direct imaging is that we can change the integration time (hence the dwell time) offline, during the analysis stage. That is, the experimenter is free to adapt the sample rate to the dynamics of the measurements. So, for instance, during periods of low activity (low SNR), we can crank up the integration time for more accurate estimates without suffering the loss of (effective) spatial resolution due to temporal blurring (since little of interest has occurred within the frame, presuming low activity is associated with low-sped processes and so slower sampling is sufficient). Conversely, during periods of high activity (high electron density and thus high SNR), we can reduce the integration time and still obtain relatively accurate estimates while resolving the spatial structure and dynamics of the event.

6.1 Suitability and limitations of the geostatistical model

> *"For us, such and such a planet is as arid as the Sahara, another as frozen as the North Pole, yet another as lush as the Amazon basin. . . . We have no need of other worlds."*
>
> *Solaris*
> Stanisław Lem

In Lem's novel *Solaris*, the character Snau condemns mankind for its lack of imagination. Because our conceptions of the unfamiliar are limited by the language used to express them, and because our language stems from a need to describe the familiar, we cannot truly comprehend what we lack the language to describe. Fortunately, the language of mathematics is robust and continually evolving.

Model-based statistical inference faces a similar dilemma. Bayesian inference attempts to make sense of data within the context of a prior model. But the act of assuming a prior model necessar-

ily limits the space of inferences (indeed that is its purpose!). Futhermore, classical geostatistics assumes that a covariance model is either known or can be estimated from the data *and* that this second-order model is sufficient to perform inference. In the absense of perfect knowledge of the underlying data

Assumption of Stationarity

For instance, underlying the covariance function (a.k.a. structure function or variogram) so widely used in geostatistics/spatial statistics (to describe the dependence structure of data and their generative processes) is the assumption of *wide-sense stationarity*. As a second-order moment of a much richer pdf, the variance/covariance is necessarily limited in its descriptive capability. Yet, because it is easy to implement and usually good enough in practice, it serves as a convenient tool for linking the pdf of the data to that of the unknown, predicted point. Richer tools exist for mapping full marginal distributions to a joint cumulative distribution. The copula is one example gaining popularity in the statistical literature. It could prove useful in spatial statistics as Bayesian methods become more prevalent.

Then again, perhaps describing the full joint distribution of the r.p. will prove to be an unwise strategy. Much of the literature modeling nonstationary processes so far has been focused on describing subsets of the domain of interest in terms of stationary processes; that is, forming nonstationary covariance models from mixtures of stationary models (Paciorek and Schervish, 2004).

Another strategy is to exploit the structure of data at different spatial scales. The shared strategy of these approaches is to transform or reorder the data so that the covariance matrix possesses a structure that can be easily factorized. Multiresolution models (Nychka et al., 2002; Berliner et al., 2003) are particularly useful for very large data sets, such as global satellite data. Fixed-rank kriging (Cressie and Johannesson, 2008) encodes the variability of the the r.p. at different scales directly into the covariance matrix.

For very large datasets, the covariance matrix can be approximated by a hierarchically semiseparable matrix (Martinsson, 2011). The matrix is factorized with a tree structure such that fine-scale correlations inherit information from the courser levels. The resulting covariance matrices are frequently sparse, but more generally possess a computationally advantageous structure.

Assumption of Gaussian processes

The classical predictors of geostatistics are linear functions of the data. It can be shown (e.g. Cressie, 1993) that this predictor is equivalent to assuming the r.p. is Gaussian. This is a convenient assumption that is often sufficient, but it also presents limitations, for instance, when the process is known to take only positive values or discrete values. Such processes can be accommodated with classical linear predictors following a transformation of the data. But even that approach implies a strong distributional assumption (the transformed data must have a Gaussian distribution).

Rather than make distributional assumptions and shoehorning data into a particular prior, a Bayesian approach requires all pdfs to be discovered from the data. In particular, a hierarchical Bayesian model may specify a non-Gaussian pdf for the process $Y(\mathbf{s})$, for instance Laplacian:

$$f(y;\mu,\tau) = \frac{\tau}{2}\exp\left(-\tau\left|y-\mu\right|\right).$$

The parameters μ and τ must then be estimated from the data. Alternatively, these too are fitted to pdfs (e.g. uniform for the location parameter μ and gamma for the scale parameter τ). *Markov chain Monte Carlo* (MCMC) techniques are used to estimate the hyperparameters of these distributions.

Bayesian view of simple kriging

In contrast to the derivation in Section 2.2.2, the Bayesian approach views both the data and the process model parameters as random variables. Let $Y(\cdot)$ be given by the mixed-effects model

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \tag{6.1}$$

where $\mu(\cdot)$ is the non-random large-scale trend and $\delta(\cdot)$ is the small-scale random component. Let $\delta(\cdot)$ be a zero-mean *Gaussian process* ($\delta(\cdot) \sim \mathcal{N}(0, C_Y(\underline{\theta}))$, and $\underline{\theta} = \left(\sigma_0^2, \sigma_1^2, a, \nu\right)$ is a vector of process parameters defining the covariance function C_Y. The data are point-sampled from $Y(\cdot)$ with additive white Gaussian noise:

$$Z(\mathbf{s}_i) = Y(\mathbf{s}_i) + \epsilon_i, \tag{6.2}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

Hierarchical Bayesian modeling has emerged as a successful approach for analyzing and predicting spatial and spatiotemporal data (Banerjee et al., 2004; Cressie and Wikle, 2011; Wikle et al., 2001; Kang and Cressie, 2011; Berliner et al., 2003, e.g.). Key to its success is the parsimonious expression of interdependencies among variables. Establishing conditional independence is impor-

tant for that reason, but also because conditional distributions tend to be easier to to model than full joint distributions. A general strategy for spatial processes, from Cressie and Wikle (2011), is to decompose the joint density of unknowns into distinct stages:

$$[\text{Data} \mid \text{Process}][\text{Process} \mid \text{Parameters}][\text{Parameters}].$$

So we begin translating the components of the simple kriging system to probability distributions. The data model (6.2) becomes

$$\left[\underline{Z} \mid Y(\cdot),\ \sigma_\epsilon^2\right] = \mathcal{N}(Y(\cdot),\ \sigma_\epsilon^2 \mathbf{I}).$$

Similarly, $Y(\cdot)$ is a Gaussian process with mean $\mu_Y(\cdot)$ and $\text{Cov}(Y(\mathbf{u},\mathbf{v}\ ;\ \underline{\theta}) \triangleq C_Y(\mathbf{u},\mathbf{v}\ ;\ \underline{\theta})$.

For prediction and analysis of the process, start with the posterior distribution including all parameters (and hyperparameters) among them the semivariogram γ_Y (or covariance function C_Y), parameterized by $\underline{\theta}$. Bayesian prediction involves finding the predictive distribution, i.e. the expected value of [Y | all parameters and data], with the expectation taken over the posterior distribution:

$$p(Y|\underline{Z}, \sigma_\epsilon^2) = \text{E}_{\underline{\theta}|\underline{Z},\sigma_\epsilon^2}\left[p_Y(Y|\underline{\theta}\right]. \tag{6.3}$$

The posterior distribution is

$$[Y, \sigma_\epsilon^2, \underline{\theta} \mid \underline{Z}] \propto [\underline{Z} \mid \sigma_\epsilon^2, \delta, \underline{\theta}\ ][\delta \mid \underline{\theta}][\sigma_\epsilon^2, \underline{\theta}]. \tag{6.4}$$

In general, this distribution is difficult to obtain from data and requires simulation (via MCMC, for example) in order to marginalize over the parameters.

However, for simple kriging the variogram parameters are known, and it is not necessary to marginalize. (Cressie and Wikle (2011) show how the posterior distribution is Gaussian such that

$$Y(\mathbf{s}) \mid \underline{Z} \quad \sim \quad \mathcal{N}\left(\widehat{Y}_{\text{sk}}(\mathbf{s}), \widehat{\sigma}_{\text{sk}}^2\right). \tag{6.5}$$

The simple kriging can be solved in closed form, but this Bayesian framework for prediction is much more flexible than classical geostatistical prediction. It provides a rigorously justifiable way of accounting for transformations in any stage of the measurement processes, including nonlinearities. It permits the use of any distribution function, not just the Gaussian, since it must be approximated by simulation in order to carry out the prediction. Finally, this solution through simulation uses the data to compute a full posterior distribution (6.4), i.e. approximating the joint

distribution of both the process *and* the parameters conditioned on the data. This obviates the need for manual variogram fitting, although this doesn't mean process modeling is suddenly simplified. In addition to being more computationally complex, the "art" involved in variography is transferred to the nuance of applying MCMC and other Bayesian methods (Brooks et al., 2011; Cressie et al., 2009).

The Bayesian framework illuminates the connection between estimation (of parameters) and prediction (of $Y(\cdot)$). There are two options for accounting for random parameters in a Bayesian model: (1) evaluate (numerically) the full posterior $[Y(\cdot), \text{all parameters}|\underline{Z}]$ and then marginalize over the parameter distributions, or (2) estimate the parameters first (through curve-fitting, variography, etc.), then substitute the estimates into the predictive distribution $[Y|\underline{Z}]$. Cressie and Wikle (2011) call the former Bayesian hierarchical modeling (BHM) and the latter estimated hierarchical modeling (EHM), or plug-in prediction. Other authors argue that plug-in prediction results in an overly-optimistic uncertainty estimate (Diggle and Ribeiro Jr., 2007; Chilès and Delfiner, 2012). Goel and Degroot (1981) show that accounting for uncertainty in parameters presents valuable information in the prediction model. (But also that the regression should not be followed to hyper-hyperparameters, etc.)

6.2 Temporal component and data fusion

Throughout this work, spatial analysis and processing has occurred at each instant of time. Traditionally, the temporal component of spatial analysis has often been neglected. This is especially the case in geostatistics for mining surveys, in which the random field is quite stationary in time. The ionosphere, on the other hand, is highly dynamic. As shown in Chapter 3, density structures may last on the order seconds (below the integration time needed for high-resolution ISR imaging), or they may pass through the radar fov and evolve in complex ways over minutes. Clearly, the temporal component of data processing is important here.

Extensions of kriging directly from m spatial observations to mT space-time observations ultimately face the challenge of inverting an $mT \times mT$ covariance matrix. The space-time semivariogram can also be problematic to model without assuming some form of separability. Instead, Cressie and Wikle (2011) recommends incorporating a dynamical model into the spatial prediction, such that the sequential nature of the time axis plays a part. For instance, Kerwin and Prince (1999) incorporate a kriging predictor in the update step of a Kalman filter. Kang et al. (2010) use a temporal update model to improve the spatial mapping of satellite data.

In the case of radar measurements, and particularly ISR, the dynamical model is afforded a natural parameter in the form of a Doppler velocity estimate. If a vector flow field can be recovered, the upcoming state of the density can be predicted through a relatively simple model. More complex models may incorporate ion and electron temperatures (also estimated from the radar backscatter spectrum), and atmospheric and geomagnetic models. In a more general form, the Bayesian approach to spatial(-temporal) analysis introduced above provides a natural way of incorporating ancillary models or measurements.

References

...Thus I rediscovered what writers have always known (and have told us again and again): books always speak of other books, and every story tells a story that has already been told.

Postscript to The Name of the Rose
Umberto Eco

Akasofu, S.-I. (1964). The development of the auroral substorm. *Planetary and Space Science*, 12(4):273 – 282.

Angelopoulos, V., Sibeck, D., Carlson, C., McFadden, J., Larson, D., Lin, R., Bonnell, J., Mozer, F., Ergun, R., Cully, C., Glassmeier, K., Auster, U., Roux, A., LeContel, O., Frey, S., Phan, T., Mende, S., Frey, H., Donovan, E., Russell, C., Strangeway, R., Liu, J., Mann, I., Rae, J., Raeder, J., Li, X., Liu, W., Singer, H., Sergeev, V., Apatenkov, S., Parks, G., Fillingim, M., and Sigwarth, J. (2008). First results from the themis mission. *Space Science Reviews*, 141:453–476. 10.1007/s11214-008-9378-4.

Bahcivan, H., Hysell, D., Lummerzheim, D., Larsen, M., and Pfaff, R. (2006). Observations of colocated optical and radar aurora. *J. Geophys. Res.*, 111:A12308.

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Number 101 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton.

Berliner, L. M., Milliff, R. F., and Wikle, C. K. (2003). Bayesian hierarchical modeling of air-sea interaction. *J. Geophys. Res.*, 108(C4).

Beynon, W. and Williams, P. (1978). Incoherent scatter of radio waves from the ionosphere. *Reports of Progress in Physics*, 41:909–956.

Blanch, J. (2004). *Using kriging to bound satellite ranging errors due to the ionosphere*. PhD thesis, Stanford University.

Bowles, K. (1958). Observations of vertical incidence scatter from the ionosphere at 41 mc/sec. *Phys. Rev. Lett.*, (1):454.

Bristow, W. (2008). Statistics of velocity fluctuations observed by SuperDARN under steady interplanetary magnetic field conditions. *J. Geophys. Res.*, 113(A12):11202–+.

Bristow, W. and Jensen, P. (2007). A superposed epoch study of SuperDARN convection observations during substorms. *Journal of Geophysical Research (Space Physics)*, 112(A11):A06232.

Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L., editors (2011). *Handbook of Markov Chain Monte Carlo*. Handbooks of Modern Statistical Methods. Chapman & Hall/CRC Press, Boca Raton.

Chilès, J.-P. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. Wiley Series in Probability and Statistics (Applied Probability and Statistics Section). John Wiley & Sons, New York, 2nd edition.

Coster, A., Gaposchkin, E., and Thornton, L. (1992). Real-time ionospheric modeling using the gps. *Navigation*, 39(2):191–204.

Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, 22(3):239–252.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 2nd edition.

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, 19(3):553–570.

Cressie, N. and Hawkins, D. M. (1980). Robust estimation of the variogram: I. *Mathematical geology*, 12(2):115–125

Cressie, N. and Johannesson, G. (2006). Spatial prediction of massive datasets. In *Proceedings of the Australian Academy of Science Elizabeth and Frederick White Conference*, pages 1–11.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-temporal Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken.

Cueto, E., Sukumar, N., Calvo, B., Martínez, M. A., Cegoñino, J., and Doblaré, M. (2003). Overview and recent advances in natural neighbour galerkin methods. 10(4):307–384.

Curriero, F. (2006). On the use of non-euclidean distance measures in geostatistics. *Mathematical Geology*, 38:907–926. 10.1007/s11004-006-9055-7.

de la Beaujardière, O. and Vondrak, R. (1982). Chatanika radar observations of the electrostatic potential distribution of an auroral arc. *J. Geophys. Res.*, 87(A2):797–809.

de la Beaujardière, O., Vondrak, R., and Baron, M. (1977). Radar observations of electric fields and currents associated with auroral arcs. *J. Geophys. Res.*, 82:5051.

Diggle, P. and Ribeiro Jr., P. (2007). *Model-based Geostatistics*. Springer Series in Statistics. Springer, New York.

Dougherty, J. and Farley, D. (1960). A theory of incoherent scattering of radio waves by a plasma. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 259(1296):79–99.

Dougherty, J. and Farley, D. (1963). A Theory of Incoherent Scattering of Radio Waves by a Plasma, 3 Scattering in a Partly Ionized Gas. *J. Geophys. Res.*, 68:5473–5486.

Doupnik, J., Brekke, A., and Banks, P. (1977). Incoherent scatter radar observations during three sudden commencements and a Pc 5 event on august 4, 1972. *J. Geophys. Res.*, 82(4):499–514.

Evans, J. (1969). Theory and practice of ionosphere study by thomson scatter radar. *Proceedings of the IEEE*, 57(4):496–530.

Farley, D. (1960). A theory of electrostatic fields in the ionosphere at nonpolar geomagnetic latitudes. *J. Geophys. Res.*, 65(3):869–877.

Farley, D. (1969). Incoherent Scatter Correlation Function Measurements. *Radio Science*, 4:935–953.

Farley, D. (1970). Incoherent scattering at radio frequencies. *J. Atmosph. Terr. Phys.*, 32:693–704.

Farley, D. (2008). *Untitled textbook*. Unpublished.

Farley, D., Dougherty, J., and Barron, D. (1961). A theory of incoherent scattering of radio waves by a plasma ii. scattering in a magnetic field. *Proc. Roy. Soc.*, A263(1313):238–258.

Farley, D. T. (1972). Multiple-pulse incoherent-scatter correlation function measurements. *Radio Science*, 7(6):661–666.

Fejer, J. (1960). Scattering of radio waves by an ionized gas in thermal equilibrium. *Can. J.Phys.*, 38:1114–+.

Ferreira, M. A. and Lee, H. K. (2007). *Multiscale Modeling: A Bayesian Perspective*. Springer Series in Statistics. Springer, New York.

Fujii, R., Oyama, S., Buchert, S., Nozawa, S., and Matuura, N. (2002). Field-aligned ion motions in the E and F regions. *J. Geophys. Res.*, 107(A5):1049.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P., editors (2010). *Handbook of Spatial Statistics*. Handbooks of Modern Statistical Methods. Chapman & Hall/CRC Press, Boca Raton.

Goel, P. K. and Degroot, M. H. (1981). Information about hyperparamters in hierarchical models. *Journal of the American Statistical Association*, 76(373):140–147.

Golub, G. and Van Loan, C. (1996). *Matrix Computations*. Johns Hopkins Studies in Mathematical Sciences. The Johns Hopkins University Press, 3rd edition.

Gordon, W. (1958). Incoherent scattering of radio waves by free electrons with application to space exploration by radar. *Proc. TRE*, 46:1824.

Gray, R. and Farley, D. (1973). Theory of incoherent-scatter measurements using compressed pulses. *Radio Science*, 8:123–+.

Hagfors, T. (1961). Density Fluctuations in a Plasma in a Magnetic Field, with Applications to the Ionosphere. *J. Geophys. Res.*, 66:1699–1712.

Hagfors, T. (2003). Basic physics of incoherent scatter. EISCAT Summer School 2003 lecture.

Hagfors, T. and Behnke, R. (1974). Measurement of three dimensional plasma velocities at the arecibo observatory. *Rad. Sci.*, 9(2):89–93.

Hansen, A. (2002). *Tomographic Estimation of the Ionosphere Using Terrestrial GPS Sensors*. PhD thesis, Stanford University.

Heinselman, C. and Nicolls, M. (2008). A bayesian approach to electric field and *e*-region neutral wind estimation with the poker flat advanced modular incoherent scatter radar. *Radio Science*, 43:RS5013.

Huang, C., Zhang, H., and Robeson, S. (2011). On the validity of commonly used covariance and variogram functions on the sphere. *Mathematical Geosciences*, 43(6):721–733.

Hysell, D., Michhue, G., Nicolls, M., Heinselman, C., and Larsen, M. (2009). Assessing auroral electric field variance with coherent and incoherent scatter radar. *Journal of Atmospheric and Solar-Terrestrial Physics*, 71(6-7):697 – 707.

Jain, A. K. (1989). *Fundamentals of digital image processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Jun, M. and Stein, M. L. (2007). An approach to producing space–time covariance functions on spheres. *Technometrics*, 49(4):468–479.

Kang, E. L. and Cressie, N. (2011). Bayesian inference for the spatial random effects model. *Journal of the American Statistical Association*, 106(495):972–983.

Kang, E. L., Cressie, N., and Shi, T. (2010). Using temporal variability to improve spatial mapping with application to satellite data. *Canadian Journal of Statistics*, 38(2):271–289.

Kerwin, W. and Prince, J. (1999). The kriging update model and recursive space-time function estimation. *IEEE Transactions on Signal Processing*, 47(11):2942–2952.

Kudeki, E. and Milla, M. (2011). Incoherent scatter spectral theories—Part i: A general framework and results for small magnetic aspect angles. *Geoscience and Remote Sensing, IEEE Transactions on*, 49(1):315–328.

Kuzma, H. (2004). *Support Vector Machines for Geophysical Inversion*. PhD thesis, University of California, Berkley.

Lanchester, B., Kaila, K., and McCrea, I. (1996). Relationship between large horizontal electric fields and auroral arc elements. *J. Geophys. Res.*, 101(A3):5075–5084.

Le, N. D. and Zidek, J. V. (2006). *Statistical Analysis of Environmental Space-Time Processes*. Springer Series in Statistics. Springer New York. 10.1007/0-387-35429-8_8.

Lehtinen, M., Huuskonen, A., and Markkanen, M. (1997). Randomization of alternating codes: Improving incoherent scatter measurements by reducing correlations of gated autocorrelation function estimates. *Radio Science*, 32(6):2271–2282.

Lehtinen, M. S. (1986). *Statistical theory of incoherent scatter radar measurements*. PhD thesis, University of Helsinki.

Lyons, L., Wang, C.-P., Gkioulidou, M., and Zou, S. (2009). Connections between plasma sheet transport, Region 2 currents, and entropy changes associated with convection, steady magnetospheric convection periods, and substorms. *Journal of Geophysical Research (Space Physics)*, 114(A13):A00D01.

Martinsson, P. (2011). A fast randomized algorithm for computing a hierarchically semiseparable representation of a matrix. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1251–1274.

Menke, W. (1989). *Geophysical Data Analysis: Discrete Inverse Theory*. International Geophysics Series, New York: Academic Press, 1989, Rev.ed.

Moore, M., editor (2001). *Spatial Statistics: Methodological Aspects and Applications*, volume 159 of *Lecture Notes in Statistics*. Springer, New York.

Nychka, D., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2(4):315–331.

Paciorek, C. J. and Schervish, M. J. (2004). Nonstationary covariance functions for gaussian process regression.

Papritz, A. and Stein, A. (2002). Spatial prediction by linear kriging. In Stein, A., Meer, F., Gorte, B., Meer, F. D., and Marçal, A., editors, *Spatial Statistics for Remote Sensing*, volume 1 of *Remote Sensing and Digital Image Processing*, pages 83–113. Springer Netherlands. 10.1007/0-306-47647-9_6.

Picone, J., Hedin, A., Drob, D., and Aikin, A. (2002). NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues. *Journal of Geophysical Research (Space Physics)*, 107:1468.

Rideout, W. and Coster, A. (2006). Automated gps processing for global total electron content data. *GPS Solutions*, 10:219–228. 10.1007/s10291-006-0029-5.

Rostoker, G. (1999). The evolving concept of a magnetospheric substorm. *Journal of Atmospheric and Solar-Terrestrial Physics*, 61(1-2):85 – 100.

Rostoker, G., Akasofu, S.-I., Foster, J., Greenwald, R., Lui, A., Kamide, Y., Kawasaki, K., McPherron, R., and Russell, C. (1980). Magnetospheric substorms - Definition and signatures. *J. Geophys. Res.*, 85:1663–1668.

Salpeter, E. (1960). Electron Density Fluctuations in a Plasma. *Phys. Rev.*, 120:1528–1535.

Schunk, R. and Nagy, A. (2009). *Ionospheres: physics, plasma physics, and chemistry*. Cambridge Atmospheric and Space Science Series. Cambridge University Press, 2nd edition.

Semeter, J., Butler, T., Heinselman, C., Nicolls, M., Kelly, J., and Hampton, D. (2008). Volumetric imaging of the auroral ionosphere: Initial results from PFISR. *J. Atmosph. Sol.-Terr. Phys.*

Semeter, J., Butler, T., Zettergren, M., Heinselman, C., and Nicolls, M. (2010). Composite imaging of auroral forms and convective flows during a substorm cycle. *J. Geophys. Res.*, 115(A8).

Semeter, J. and Doe, R. (2002). On the proper interpretation of ionospheric conductance estimated through satellite photometry. *J. Geophys. Res.*, 107:1200.

Semeter, J., Heinselman, C., Thayer, J., Doe, R., and Frey, H. (2003). Ion upflow enhanced by drifting F-region plasma structure along the nightside polar cap boundary. *Geophys. Res. Lett.*, 30(22):2139.

Semeter, J. and Kamalabadi, F. (2005). Determination of primary electron spectra from incoherent scatter radar measurements of the auroral *E*-region. *Radio Science*, 40:RS2006.

Sibson, R. (1981). A brief description of natural neighbour interpolation. *Interpreting multivariate data*, 21.

Sparks, L., Blanch, J., and Pandya, N. (2011a). Estimating ionospheric delay using kriging: 1. methodology. *Radio Sci.*, 46.

Sparks, L., Blanch, J., and Pandya, N. (2011b). Estimating ionospheric delay using kriging: 2. impact on satellite-based augmentation system availability. *Radio Sci.*, 46.

St.-Maurice, J.-P., Kofman, W., and James, D. (1996). In situ generation of intense parallel electric fields in the lower ionosphere. *J. Geophys. Res.*, 101:335–356.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer, New York.

Stein, M. L. (2008). A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society*, 37(1):3–10.

Sulzer, M., Aponte, N., and González, S. (2005). Application of linear regularization methods to Arecibo vector velocities. *J. Geophys. Res.*, 110(A9):A10305.

Sulzer, M. P. (1993). A new type of alternating code for incoherent scatter measurements. *Radio Science*, 28(6):995–1001.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Soc. for Ind. and App. Math., Philadelphia, PA.

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240.

Varney, R. H., Nicolls, M. J., Heinselman, C. J., and Kelley, M. C. (2009). Observations of polar mesospheric summer echoes using pfisr during the summer of 2007. *Journal of Atmospheric and Solar-Terrestrial Physics*, 71(3–4):470–476. ¡ce:title¿Global Perspectives on the Aeronomy of the Summer Mesopause Region¡/ce:title¿ ¡xocs:full-name¿Eighth International Workshop on Layered Phenomena in the Mesopause Region¡/xocs:full-name¿.

Venkata Ratnam, D. and Sarma, A. (2012). Modeling of low-latitude ionosphere using gps data with shf model. *Geoscience and Remote Sensing, IEEE Transactions on*, 50(3):972 –980.

Vickrey, J., Vondrak, R., and Matthews, S. (1982). Energy deposition by precipitating particles and Joule dissipation in the auroral ionosphere. *J. Geophys. Res.*, 87:5184–5196.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer-Verlag, Berlin, 3rd edition.

Wahlund, J.-E., Opgenoorth, H., Haggstrom, I., Winser, K., and Jones, G. (1992). Eiscat observations of topside ionospheric ion outflows during auroral activity: Revisited. *J. Geophys. Res.*, 97(A3):3019–3037.

Weber, E., Vickrey, J., Heinselman, C., Gallagher, H., Weiss, L., Heelis, R., and Kelley, M. (1991). Coordinated radar and optical measurements of stable auroral arcs at the polar cap boundary. *J. Geophys. Res.*, 96:17847–17863.

Whalen, B., Green, D., and McDiarmid, I. (1974). Observations of ionospheric ion flow and related convective electric fields in and near an auroral arc. *J. Geophys. Res.*, 79(19):2835–2842.

Wikle, C. K. (2010). *Low-Rank Representations for Spatial Processes*, volume Handbook of Spatial Statistics of *Handbooks of Modern Statistical Methods*, chapter 8. Chapman & Hall/CRC Press, Boca Raton.

Wikle, C. K., Berliner, L. M., and Milliff, R. F. (2003). Hierarchical bayesian approach to boundary value problems with stochastic boundary conditions. *Monthly Weather Review*, 131(6):1051–1062.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001). Spatiotemporal hierarchical bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association*, 96(454):382–397.

Wilkinson, L. (2005). *The Grammar of Graphics*. Statistics and Computing. Springer, 2nd edition.

Yue, X., Wan, W., Liu, L., and Mao, T. (2007). Statistical analysis on spatial correlation of ionospheric day-to-day variability by using gps and incoherent scatter radar observations. *Annales Geophysicae*, 25(8):1815–1825.

Zettergren, M., Semeter, J., Blelly, P.-L., and Diaz, M. (2007). Optical estimation of auroral ion upflow: Theory. *J. Geophys. Res.*, 112(A11):A12310.

Zhu, P., Raeder, J., Germaschewski, K., and Hegna, C. (2009). Initiation of ballooning instability in the near-Earth plasma sheet prior to the 23 March 2007 THEMIS substorm expansion onset. *Annales Geophysicae*, 27:1129–1138.

Zidek, J., Le, N., and Liu, Z. (2012). Combining data and simulated data for space–time fields: application to ozone. *Environmental and Ecological Statistics*, 19:37–56. 10.1007/s10651-011-0172-1.

Zimmerman, D. and Stein, M. (2010). *Classical Geostatistical Methods*, volume Handbook of Spatial Statistics of *Handbooks of Modern Statistical Methods*, chapter 3. Chapman & Hall/CRC Press, Boca Raton.

Index

CURRICULUM VITAE

Joe Graduate

Basically, this needs to be worked out by each individual. Perhaps here one can forgive the hapless PhD candidate for jamming in a PDF CV that was originally generated from some other source. This could be accomplished with Acrobat or something, but the candidate should make sure that the pages appear in the table of contents.