Hierarchical Spatial and Spatio-Temporal Modeling of Massive Datasets, with Application to Global Mapping of CO_2

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Matthias Katzfuss, M.S.

Graduate Program in Statistics

The Ohio State University

2011

Dissertation Committee:

Noel Cressie, Advisor

Peter Craigmile

Tao Shi

Abstract

This dissertation is comprised of an introductory chapter and three stand-alone chapters, tied together by a unifying theme: the statistical analysis of very large spatial and spatio-temporal datasets. These datasets now arise in many fields, but our focus here is on environmental remote-sensing data. Due to sparseness of daily datasets, there is a need to fill spatial gaps and to borrow strength from adjacent days. Nonetheless, many satellite instruments are capable of conducting on the order of 100,000 retrievals per day, which makes it computationally challenging to apply traditional spatial and spatio-temporal statistical methods, even in supercomputing environments. In addition, the datasets are often observed on the entire globe. For such large domains, spatial stationarity assumptions are typically unrealistic.

We address these challenges using dimension-reduction techniques based on a flexible spatial random effects (SRE) model, where dimension reduction is achieved by projecting the process onto a basis-function space of low dimension. The spatio-temporal random effects (STRE) model extends the SRE model to the spatio-temporal case by modeling the temporal evolution, on the reduced space, using a dynamical autoregressive model in time.

Another focus of this work is the modeling of fine-scale variation. Such variability is typically not part of the reduced space spanned by the basis functions, and one needs to account for a component of variability at a fine scale. We address this issue throughout the

dissertation with increasingly complex and realistic models for a component of fine-scale variation.

After a general introductory chapter, the subsequent two chapters focus on estimation of the reduced-dimensional parameters in the STRE model from both an empirical-Bayes and a fully Bayesian perspective, respectively. In Chapter 2, we develop maximum likelihood estimation via an expectation-maximization (EM) algorithm, which offers stable computation of valid estimators and makes efficient use of spatial and temporal dependence in the data, assuming a multivariate Gaussian model. In Chapter 3, we develop a multiresolutional prior for the propagator matrix on the reduced-dimensional space that allows for unknown (random) sparsity and shrinkage, and we describe how sampling from the posterior distribution can be achieved in a feasible way, even if this matrix is very large.

Finally, in Chapter 4, we return to the spatial-only case. We generalize the standard SRE model and provide informative prior distributions for the parameters of the generalized SRE model based on a nonstationary covariance model in physical space. We propose a comprehensive model that takes account of all scales of variation, in particular by allowing for the fine-scale-variation component to exhibit spatial dependence. We make inference on the number, locations, and shapes of the basis functions. Computational feasibility is maintained by assuming that the fine-spatial-scale covariance is compactly supported, resulting in a very sparse covariance matrix for the fine-scale-variation component.

All methodological results are illustrated and compared using simulation studies and a dataset of global satellite measurements of CO_2, which came from the Atmospheric InfraRed Sounder (AIRS) instrument on NASA's Aqua satellite.

Acknowledgments

First and foremost, I would like to thank my advisor, Dr. Noel Cressie, for his generous support, guidance, and mentorship.

I would also like to thank: Mike Turmon for originally suggesting that EM estimation might work in the context of the spatial random effects model; Kevin Sahr for providing the DGGRID software and for advice on how to shift the basis-function centers for Chapters 2 and 3; the AIRS Project CO_2 team, particularly Dr. Moustafa T. Chahine, Dr. Edward T. Olsen, and Mr. Luke L. Chen for their helpful input on the analysis of the AIRS data; Amy Braverman, Dorit Hammerling, Anna Michalak, and Hai Nguyen for their comments on various aspects of the research in Chapters 2 and 3; Scott Holan, an anonymous referee, and especially Chris Wikle for their excellent suggestions regarding a manuscript version of Chapter 2; and Wenceslao González Manteiga and two anonymous referees for their helpful reviews of a manuscript version of Chapter 3.

This research was partially supported by NASA under grant NNX08AJ92G issued through the ROSES Carbon Cycle Science Program and grant NNH08ZDA001N issued through the Advanced Information Systems Technology ROSES 2008 Solicitation, and by the Office of Naval Research Grant N00014-08-1-0464.

Vita

April 2005 - August 2007	Undergraduate Studies in Statistics, University of Munich
December 2008	M.S. in Statistics, The Ohio State University
September 2007 - March 2009	Graduate Teaching Associate, Department of Statistics, The Ohio State University
October 2008 - present	Graduate Research Associate, Department of Statistics, The Ohio State University

Publications

Research Publications

Katzfuss, M., and Cressie, N. 2009. Maximum likelihood estimation of covariance parameters in the spatial random effects model. *2009 Proceedings of the Joint Statistical Meetings*, American Statistical Association, Alexandria, VA.

Heaton, M.J., Katzfuss, M., Ramachandar, S., Pedings, K., Gilleland, E., Mannshardt-Shamseldin, E., and Smith, R. 2011. Spatio-temporal models for large-scale indicators of extreme weather. *Environmetrics*, 22, 294–303.

Katzfuss, M., and Cressie, N. 2011. Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets. *Journal of Time Series Analysis*, 32, 43–446.

Katzfuss, M., and Cressie, N. 2011. Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics*, under revision.

Fields of Study

Major Field: Statistics

Table of Contents

List of Tables

List of Figures

Figure **Page**

Chapter 1: Introduction

As the title of this dissertation suggests, I am concerned with the intersection of three broad statistical topics: hierarchical modeling, spatial and spatio-temporal statistics, and the analysis of very-large-to-massive datasets. In this chapter, I shall attempt to give a short overview of each, before giving the specifics associated with the intersection of the three.

Hierarchical modeling has become increasingly prevalent in the statistics literature. In this framework, model specification is comparatively easy, since it is done conditionally for each component of the model. In addition, the advent of simulation-based Bayesian inference over the past two decades is particularly well suited to fitting hierarchical models: the Gibbs sampler exploits very naturally the (often) simple conditional model structure. Berliner (1996) formulated an attractive general hierarchical model for time series, which consists of three conditionally specificied "stages": the data model ($[\text{data}|\text{process}, \text{parameters}]$), the process model ($[\text{process}|\text{parameters}]$), and the parameter (or prior) model ($[\text{parameters}]$). The product of the three results in the generic joint distribution:

$$[\text{data}, \text{process}, \text{parameters}] = [\text{data}|\text{process}, \text{parameters}]\,[\text{process}|\text{parameters}]\,[\text{parameters}].$$

This type of modeling is especially relevant to environmental statistics, where measurements of a process of interest are virtually always incomplete and noisy.

Spatial statistics deals with modeling data that are spatially referenced. It follows the principle that observations that are close in space are typically more closely related than observations that are far apart. For an overview of the field of spatial statistics, see Cressie (1993). Depending on whether the spatial domain of interest is fixed and continuous, or fixed and countable, or random, spatial processes are often referred to as geostatistical processes, or lattice processes, or spatial point processes, respectively. This dissertation deals with geostatistical processes observed on continuous, fixed domains (such as the globe). Spatio-temporal statistics is, depending on one's perspective, a generalization of a spatial process to a spatial process evolving in time, or a generalization of a time series to multiple, spatially referenced observations at each time point. Here, I view spatio-temporal data as a realization of a spatial process evolving in discrete time. A recent overview of spatio-temporal statistics, with an emphasis on hierarchical-modeling approaches, is given by Cressie and Wikle (2011).

Automation of measurement procedures, and an increasing desire to measure the status and monitor the performance of systems, has led to an explosion in the amount of data being collected in all fields of science and in all aspects of society. This phenomenon has had, and continues to have, a profound impact on the field of statistics. Statistical techniques of the last century were not designed to deal with the large number of variables and/or the massive amount of observations prevalent in most modern datasets. Thus, new approaches and models have been developed to deal with datasets containing many variables and/or many observations. I focus here on one variable (or process), and my particular concern is with computational feasibility of the statistical methodology when the number of observations is very large.

At the intersection of these three topics has arisen a very active field of research dealing with statistical analysis of very large spatial and spatio-temporal datasets. The analysis of massive datasets is of concern in spatial statistics, which is particularly prone to the curse of dimensionality. Traditional geostatistics requires the evaluation and inversion of the covariance matrix of the data. If there are n measurements available, this is an $n \times n$ matrix, and its inversion requires on the order of n^3 computations. Thus, direct inversion is clearly not feasible for very-large-to-massive n, and dimension-reduced models, computational speed-ups, and/or approximations are required.

The research in this dissertation focuses on one such dimension-reduction technique, in which the process is projected onto a low-dimensional space spanned by the linear combination of a set of basis functions. The resulting model is called a spatial random effects (SRE) model. The SRE model is extended to a spatio-temporal random effects (STRE) model by modeling the temporal evolution of the low-dimensional process using a dynamical autoregressive model in time. An important research topic is the treatment of the part of the process that is not part of the reduced-dimensional space of basis functions, here called the fine-scale variation.

In Chapter 2, I present an empirical-Bayes approach to estimating the reduced-dimensional parameters in an STRE model. An expectation-maximization (EM) algorithm is presented, which can be used to find maximum likelihood estimators of the parameters. Properties and extensions of the EM algorithm are given. I also compare the estimators to the previously used binned-method-of-moments estimators, in a simulation study and on a real-world dataset of global CO_2 measurements.

Chapter 3 presents a fully Bayesian approach to the same problem (inference on the STRE parameters). I develop a prior distribution for the propagator matrix on the reduced

space, which allows for random sparsity and shrinkage of the elements as a function of distance of the corresponding pairs of basis functions. I also generalize the second-moment assumptions on the fine-scale variation to allow for spatial heterogeneity, and I compare the fully Bayesian approach to the EM-estimation approach of Chapter 2 in a similar simulation study and on the same CO_2 dataset as in Chapter 2. Some of the more technical aspects of the posterior inference via Markov chain Monte Carlo (MCMC) can be found in Appendix A.

In Chapter 4, I return to the spatial-only case, where I generalize the SRE model by allowing the set of basis functions to be random. I make inference on the number, locations, and shapes of the basis functions. I develop priors that are motivated by the predictive process (Banerjee et al., 2008), and I generalize further the distribution of the fine-scale variation by allowing for (local) spatial dependence. This allows modeling of both long-range dependence (through the SRE component) and short-range dependence (through the fine-scale variation), while still allowing for feasible computation times for large datasets.

The research in Chapters 2 and 3 was carried out jointly with Dr. Noel Cressie. He gave me general ideas to work on, and I was responsible for the technical derivations, the computer code, and writing the initial drafts. Dr. Cressie contributed general comments and discussed with me the results presented in Chapter 4.

Chapter 2: Spatio-Temporal Smoothing and EM Estimation for Massive Remote-Sensing Datasets

This chapter is published as: Katzfuss, M., and Cressie, N. 2011. Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets. *Journal of Time Series Analysis*, 32, 43–446.

2.1 Introduction

Many datasets have spatial and temporal information attached to the attribute information, and nearer observations in space or time generally result in higher statistical correlation. This dependence can be *described* through specification of a spatio-temporal covariance function, or it can be *explained* through a dynamical model that gives either a probabilistic or a statistical-physical mechanism for the evolution of the "present" from the "past." It is the dynamical-modeling approach that we take in this chapter.

The spatial domain is discretized, and so it can be thought of as a (generally) large m-dimensional vector, where m denotes the number of pixels in the discretization. Should the spatial domain evolve also, there might be m_t pixels at time t. We define $\mathbf{Y}_t$ to be the m_t-dimensional vector of the true spatial process at time t. In this chapter, time is discrete and, hence, the true spatio-temporal process is a *vector-valued time series*,

$$\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_t, \mathbf{Y}_{t+1}, \ldots .$$

Observations on $\mathbf{Y}_t$ result in its degradation; here we are concerned with "missingness" and "noise" (measurement error). That is, $\mathbf{Z}_t$ is an n_t-dimensional vector ($n_t < m_t$) of observations at time t given by,

$$\mathbf{Z}_t = O_t \mathbf{Y}_t + \boldsymbol{\epsilon}_t \,;\; t = 1, 2, \ldots \,, \tag{2.1}$$

where $\{\boldsymbol{\epsilon}_t\}$ are independent $N_{n_t}(\mathbf{0}, \sigma^2_{\epsilon,t} V_{\epsilon,t})$, respectively, and O_t is an $n_t \times m_t$ incidence matrix of mostly 0s and a 1 in each row. In (2.1), O_t captures the missingness and $\boldsymbol{\epsilon}_t$ captures the measurement error (assumed to be independent in both space and time); we call (2.1) the *data model*, following Berliner (1996). Modeling the temporal evolution of $\{\mathbf{Y}_t \colon\; t = 1, 2, \ldots\}$ is discussed at length in Cressie and Wikle (2011); in this chapter, we choose a vector-autoregressive process.

This chapter is concerned with applications of spatio-temporal statistics to global remote sensing. Here, n_t and m_t can be very large, on the order of tens or hundreds of thousands, and the tendency is towards massive (gigabytes and beyond). Therefore, while the model above can in principal lead to inference using the Kalman filter (Kalman, 1960), there are severe computational problems that require some form of dimension reduction. Furthermore, the Kalman filter requires parameters like the propagator matrix and the innovation covariance matrix (see (2.4) below) to be specified, which in practice usually means they must be *estimated* (or a prior could be put on them). It is at the confluence of dimension reduction and parameter estimation that this chapter takes its place.

Denote as $Y_t(\mathbf{s})$ the element of $\mathbf{Y}_t$ that corresponds to spatial location $\mathbf{s}$. In this chapter, we assume a *spatio-temporal mixed effects* (STME) model,

$$Y_t(\mathbf{s}) = \mathbf{x}_t(\mathbf{s})' \boldsymbol{\beta}_t + \nu_t(\mathbf{s}) \,;\; \mathbf{s} \in D_s,\, t = 1, 2, \ldots, \tag{2.2}$$

where D_s is the discretized spatial domain, $\mathbf{x}_t(\cdot)$ is a known p-dimensional vector of covariates, $\boldsymbol{\beta}_t$ is a vector of fixed but unknown trend coefficients, and $\nu_t(\cdot)$ (and its vector $\boldsymbol{\nu}_t$) captures the spatio-temporal dependence. We impose dimension reduction on $\{\boldsymbol{\nu}_t\}$ by modeling it as a *spatio-temporal random effects* (STRE) process,

$$\nu_t(\mathbf{s}) = \mathbf{b}_t(\mathbf{s})'\boldsymbol{\eta}_t + \delta_t(\mathbf{s})\,;\ \mathbf{s} \in D_s,\ t = 1, 2, \ldots, \tag{2.3}$$

where $\mathbf{b}_t(\cdot) := [b_{1,t}(\cdot), \ldots, b_{r,t}(\cdot)]'$ is a vector of r (known) spatial basis functions. The coefficient vectors $\{\boldsymbol{\eta}_t\}$ are assumed to follow a vector-autoregressive process of order one,

$$\boldsymbol{\eta}_t | \boldsymbol{\eta}_{t-1}, \ldots, \boldsymbol{\eta}_1 \sim N_r(H_t\boldsymbol{\eta}_{t-1}, U_t),\ t = 1, 2, \ldots, \tag{2.4}$$

with initial state $\boldsymbol{\eta}_0 \sim N_r(\mathbf{0}, K_0)$. The $r \times r$ matrices H_t and U_t are often referred to as propagator and innovation matrices, respectively. The fine-scale-variation component $\delta_t(\cdot)$ in (2.3) is assumed to be uncorrelated across time and space and independent of $\{\boldsymbol{\eta}_t\}$, with $\delta_t(\mathbf{s}) \sim N(0, \sigma^2_{\delta,t} v_{\delta,t}(\mathbf{s}))$. Here, $v_{\delta,t}(\cdot)$ is typically considered known, although this assumption can be weakened considerably (see Section 2.3.4). The component $\delta_t(\cdot)$ is an important part of the model, as it is an attempt to account for the error that is introduced by the dimension reduction in replacing $\nu_t(\cdot)$ by $\mathbf{b}_t(\cdot)'\boldsymbol{\eta}_t$.

There are numerous examples of dimension-reduction models similar to (2.3) in the literature on spatial-only modeling (e.g., Wikle, 2010). Outside of the massive-data setting, so-called spatial linear mixed models have received a great deal of attention (e.g., Christensen and Waagepetersen, 2002; Zhang, 2002). Nychka et al. (2002) consider a wavelet basis and enforce sparsity on the covariance matrix $K := var(\boldsymbol{\eta})$, via thresholding in the wavelet-transformed space; Stein (2008) parameterizes K with only a handful of parameters by assuming axial symmetry for total column ozone on the globe, but he allows the

fine-scale variation to exhibit spatial dependence; Jun and Stein (2008) apply the discrete Fourier transform to data on a regular grid to achieve fast computation times. Banerjee et al. (2008) take a Bayesian approach, replacing the data locations with a smaller set of space-filling locations and approximating the original process with a predictive process depending on a fixed number of knots. Furrer et al. (2007) and Lopes et al. (2008) also take a Bayesian perspective, but they assume K to be diagonal. There has also been work on such models for Markov random fields (see Zhu et al., 2007, and references therein).

Equations (2.1)–(2.4) above describe what is referred to as a standard state-space model in the time-series literature (see, e.g., Hamilton, 1994, Chap. 13; Shumway and Stoffer, 2006, Chap. 6). The idea of extending the state-space model to the spatio-temporal case by using spatial basis functions in the vector $\mathbf{b}_t(\cdot)$ goes back at least as far as Smith et al. (1996) and Kaplan et al. (1998).

A key feature of our model is the dimension reduction that makes it possible to deal with a very large number of observations at each time point. The use of a vector-autoregressive (VAR) model of order one allows for sequential processing of subsequent time points via the Kalman filter and smoother (Kalman, 1960; Shumway and Stoffer, 2006). Examples of the use of Kalman filters for reduced-dimension spatio-temporal models can be found in Mardia et al. (1998), Wikle and Cressie (1999), Farrell and Ioannou (2001), Cressie and Wikle (2002), Wikle and Hooten (2006), and Voutilainen et al. (2007). A component similar to our fine-scale-variation term $\delta_t(\cdot)$ has been included in Wikle and Cressie (1999), Berliner et al. (2000), Wikle et al. (2001), and Cressie et al. (2010).

A different approach to the analysis of very large (spatio-)temporal datasets is to assume a multi-resolutional tree structure to describe the spatial-dependence structure (Cressie and Wikle, 2002; Johannesson and Cressie, 2004; Tzeng et al., 2005; Johannesson et al., 2007).

These models also offer rapid computation via Kalman-filter-type algorithms. Apart from Tzeng et al. (2005), this approach results in covariance functions and predictions that tend to be "blocky," and the models have some arbitrariness in specifying which pixels are "close" at smaller scales.

In this chapter, we assume a fixed-rank STME model (2.2), which was proposed by Cressie et al. (2010) and Kang et al. (2010), motivated by the spatial-only fixed-rank model of Cressie and Johannesson (2008). In this fixed-rank framework, the spatial basis functions, $b_1(\cdot), \ldots, b_r(\cdot)$, are not necessarily orthogonal. Shi and Cressie (2007) also considered a spatial-only version of the model, and they proposed the use of W-wavelet basis functions instead of the bisquare functions employed by Cressie and Johannesson (2008). In fixed-rank models, because r is typically much smaller than the number of observations, optimal predictors can be calculated exactly, even in large-data settings (see the end of Section 2.2.2).

In many of these articles on the fixed-rank approach, covariance parameters are estimated using a binned-method-of-moments (MM) technique that was first described in Cressie and Johannesson (2008). Katzfuss and Cressie (2009) proposed maximum likelihood (ML) estimation via an expectation-maximization (EM) algorithm for the spatial-only case, and they showed this approach to be more unsupervised and, in some aspects, more efficient.

The EM algorithm is very well suited for ML estimation of parameters in STRE models. This idea is usually attributed to Shumway and Stoffer (1982). Mardia et al. (1998) take this EM-estimation approach while reducing dimensionality by projecting the state-process on a set of spectral basis functions. Xu and Wikle (2007) consider different parameterizations for the matrix parameters. They show how an advection-diffusion equation can be

used to derive a parameterization of the propagator matrix H, and they allow for spatial dependence in the error component. For some of these parameterizations, generalizations or modifications of the ordinary EM algorithm have to be employed.

A very similar approach to ours is taken by Fassò and Cameletti (2009a,b). They also make use of the EM algorithm to estimate parameters in their state-space model. In addition, they quantify estimation uncertainty by performing a parametric bootstrap procedure. However, their model is not feasible for the remote-sensing data considered here. They assume that measurement locations are identical at each time point. Additionally, their dense fine-scale-variation covariance matrices result in a computational complexity of $\mathcal{O}(Tn^3)$ for their procedure, where n is their number of measurement locations (identical for each time point); this prohibits the analysis of very large datasets.

With continuing increases in computing power, it has become feasible to fit fully Bayesian STRE models. Often, assumptions are made to reduce the number of parameters in the covariance matrices and the propagator matrix. Zhao et al. (2006) assume K to be diagonal. Stroud et al. (2001) introduce a STRE model with weighting kernels on linear basis functions in a Bayesian framework with simple random-walk dynamics. The Bayesian spatial dynamic factor-analysis model of Lopes et al. (2008) assumes K to be diagonal. There has also been some work on incorporating physical or biological models directly into the parameterization (e.g., Wikle, 2003; Wikle and Hooten, 2006). In the spatial-only fixed-rank setting, Kang and Cressie (2011) describe a fully Bayesian approach, and they develop a multi-resolution prior for the (non-diagonal) covariance matrix K.

In light of this extensive body of literature on parameter estimation for reduced-dimension state-space models, the main contributions of this chapter are the following: We have a strong focus on very large datasets as obtained by remote-sensing platforms, which take

measurements at arbitrary, non-gridded locations (or even areal footprints) that can differ over time and can be sparse with respect to the spatial domain of interest. In addition to the standard EM algorithm for state-space models, we incorporate the estimation of trend coefficients and fine-scale-variation parameters into the algorithm. We also give some possible extensions of our EM algorithm (Section 2.3.4). Finally, we see this chapter as a valuable resource for practitioners who analyze data using the fixed-rank models described in earlier papers by Cressie and coauthors, referred to above. We show that the previously used MM estimation can be improved upon by the EM approach described here, and we give further practical insights on the analysis of global remote-sensing data with areal footprints.

The rest of this chapter is organized as follows. Section 2.2 introduces Fixed Rank Smoothing (FRS) for the STME model, where the parameters are assumed known. Section 2.3 describes how the STME-model parameters can be estimated using the EM algorithm. A simulation study comparing FRS based on the EM estimators and FRS based on the MM estimators is given in Section 2.4. Section 2.5 contains an application of our methodology to global retrievals of mid-tropospheric CO_2 from NASA's AIRS instrument, and discussion and conclusions are given in Section 2.6.

2.2 Fixed Rank Smoothing in the Spatio-Temporal Mixed-Effects Model

2.2.1 The Spatio-Temporal Mixed-Effects Model

As established in Section 2.1, our interest is in a spatio-temporal process $\{Y_t(\mathbf{s}) : \mathbf{s} \in D_s,\, t = 1, 2, \ldots\}$, which is modeled as a STME process as in (2.2)–(2.3).

For the purpose of this chapter, we shall take a *smoothing* perspective. That is, we are interested in predicting $Y_t(\mathbf{s}_0)$ at locations $\mathbf{s}_0 \in D_s$ for any $t \in \{1, \ldots, T\}$, from a

number of measurements taken at spatial locations $\{\mathbf{s}_{i,t}\}$ and time points $t = 1, \ldots, T$. As described in (2.1), we assume that the measurements $Z_t(\mathbf{s}_{i,t})$ of the process are degraded by additive measurement error:

$$Z_t(\mathbf{s}_{i,t}) = Y_t(\mathbf{s}_{i,t}) + \epsilon_t(\mathbf{s}_{i,t});\ i = 1, \ldots, n_t,\ t = 1, \ldots, T, \tag{2.5}$$

where $\epsilon_t(\mathbf{s}_{i,t}) \sim N(0, \sigma^2_{\epsilon,t} v_{\epsilon,t}(\mathbf{s}_{i,t}))$ is assumed to be independent across time and space and independent of $Y_t(\cdot)$. Throughout this chapter, we assume both $\sigma^2_{\epsilon,t}$ and the variance-modification function $v_{\epsilon,t}(\cdot)$ to be known. If there is no independent information on $\sigma^2_{\epsilon,t}$ (e.g., from prior experiments), this variance term can be estimated from the data via an estimation technique based on extrapolating the variogram to the origin (for more information, see Kang et al., 2009). Unless further information is available, $v_{\epsilon,t}(\cdot) \equiv 1$ is a default choice for the variance-modification function.

The main application of the STME model (2.2) is to massive spatial or spatio-temporal datasets, for which traditional statistical approaches are infeasible due to the large number of observations. The key to this approach is the dimension reduction that is achieved because r, the dimension of $\boldsymbol{\eta}_t$, is typically much smaller than n_t, the number of observations at time t. This point will be further elaborated upon at the end of Section 2.2.2 below.

Let $\mathcal{S}^O_t := \{\mathbf{s}_{1,t}, \ldots, \mathbf{s}_{n_t,t}\}$ be the set of locations at which there are observations at time t. Evaluating all model components at these sets of locations, stacking the resulting scalars into column vectors, and stacking row vectors into matrices, we can write $\mathbf{Z}_t := [Z_t(\mathbf{s}_{1,t}), \ldots, Z_t(\mathbf{s}_{n_t,t})]'$, the vector of measurements at time t, as,

$$\mathbf{Z}_t = X_t\boldsymbol{\beta}_t + B_t\boldsymbol{\eta}_t + \boldsymbol{\delta}_t + \boldsymbol{\epsilon}_t;\ t = 1, 2, \ldots. \tag{2.6}$$

Here, B_t is an $n_t \times r$ matrix with i-th row given by $\mathbf{b}_t(\mathbf{s}_{i,t})'$, $\boldsymbol{\delta}_t := [\delta_t(\mathbf{s}_{1,t}), \ldots, \delta_t(\mathbf{s}_{n_t,t})]'$, and the other matrices and vectors are defined analogously. Corresponding covariance

matrices are $K_t := var(\boldsymbol{\eta}_t)$, and

$$D_t := var(\boldsymbol{\delta}_t + \boldsymbol{\epsilon}_t) = \sigma^2_{\delta,t} V_{\delta,t} + \sigma^2_{\epsilon,t} V_{\epsilon,t}\,, \tag{2.7}$$

which is diagonal with $V_{\delta,t} := diag(v_{\delta,t}(\mathbf{s}_{1,t}), \ldots, v_{\delta,t}(\mathbf{s}_{n_t,t}))$ and
$V_{\epsilon,t} := diag(v_{\epsilon,t}(\mathbf{s}_{1,t}), \ldots, v_{\epsilon,t}(\mathbf{s}_{n_t,t}))$.

The model, in its most general form as described above, depends on several parameters, many of them matrix-valued. The vector of the unknown parameters, denoted by $\boldsymbol{\theta}$, consists of the trend coefficients $\{\boldsymbol{\beta}_t : t = 1, \ldots, T\}$, the fine-scale-variation variances $\{\sigma^2_{\delta,t} : t = 1, \ldots, T\}$, and the elements defining the matrices that describe the VAR process, K_0, $\{H_t : t = 1, \ldots, T\}$, and $\{U_t : t = 1, \ldots, T\}$.

2.2.2 Fixed Rank Smoothing

As mentioned above, our main focus in this chapter is on smoothing. That is, after having observed $\mathbf{Z}_1 = \mathbf{z}_1, \ldots, \mathbf{Z}_T = \mathbf{z}_T$, we are interested in inference on the hidden process $\{Y_t(\mathbf{s}_0)\}$ at any spatial location $\mathbf{s}_0 \in D_s$ and for any time $t = 1, \ldots, T$. Let the set $\mathcal{S}^P_t$ consist of all m_t (observed or not observed) spatial locations at which we want to predict the hidden process $Y_t(\cdot)$ at time t. Often, $\mathcal{S}^P_t$ is a grid over the spatial domain of interest that does not depend on t.

To make clear the distinction between observed quantities and predicted quantities, for the remainder of this chapter we shall use a superscript P for vectors and matrices that were derived by evaluation of model components at the set of prediction locations. The process vector of interest is therefore,

$$\mathbf{Y}^P_t := X^P_t \boldsymbol{\beta}_t + B^P_t \boldsymbol{\eta}_t + \boldsymbol{\delta}^P_t;\ t = 1, \ldots, T, \tag{2.8}$$

where now evaluation is at the set of m_t prediction locations, $\mathcal{S}^P_t$. The diagonal variance matrix, $var(\boldsymbol{\delta}^P_t) := \sigma^2_{\delta,t} V^P_{\delta,t}$, is defined correspondingly.

From (2.8), we see that inference on $\mathbf{Y}_t^P$ essentially consists of predicting $\boldsymbol{\eta}_t$ and $\boldsymbol{\delta}_t^P$ (and estimating the trend coefficients, $\boldsymbol{\beta}_t$). For now we assume that $\boldsymbol{\beta}_t$ is known, along with all other parameters in $\boldsymbol{\theta}$. Parameter estimation will be addressed separately in Section 2.3.

Let $\mathbf{z}_{1:\tilde{t}} := [\mathbf{z}_1', \ldots, \mathbf{z}_{\tilde{t}}']'$, for any time point $\tilde{t} > 0$. For the conditional expectations of $\boldsymbol{\eta}_t$ and $\boldsymbol{\delta}_t$ based on data $\mathbf{z}_{1:\tilde{t}}$, we will use the notation $\boldsymbol{\eta}_{t|\tilde{t}} := E(\boldsymbol{\eta}_t|\mathbf{z}_{1:\tilde{t}})$ and $\boldsymbol{\delta}_{t|\tilde{t}}^P := E(\boldsymbol{\delta}_t^P|\mathbf{z}_{1:\tilde{t}})$. The conditional covariance matrix of $\boldsymbol{\eta}_t$ will be denoted as $P_{t|\tilde{t}} := var(\boldsymbol{\eta}_t|\mathbf{z}_{1:\tilde{t}})$, and $R_{t|\tilde{t}}^P := var(\boldsymbol{\delta}_t^P|\mathbf{z}_{1:\tilde{t}})$ is the conditional covariance matrix of $\boldsymbol{\delta}_t^P$.

To obtain the smoothing distributions of $\{\boldsymbol{\eta}_t\}$ and $\{\boldsymbol{\delta}_t^P\}$, we make use of a technique called Fixed Rank Smoothing (FRS), which is an extension of the Kalman smoother (see Cressie et al., 2010, for more details). Similar to the original Kalman smoother, this technique consists of two parts: forward-filtering and backward-smoothing.

The filtering algorithm is initialized by setting $\boldsymbol{\eta}_{0|0} = \mathbf{0}$ and $P_{0|0} = K_0$. Then, for $t = 1, \ldots, T$, the filtering quantities are calculated sequentially as,

$$\begin{aligned}
\boldsymbol{\eta}_{t|t} &= \boldsymbol{\eta}_{t|t-1} + P_{t|t-1}B_t'[B_tP_{t|t-1}B_t' + D_t]^{-1}(\mathbf{z}_t - X_t\boldsymbol{\beta}_t - B_t\boldsymbol{\eta}_{t|t-1}) \\
\boldsymbol{\delta}_{t|t}^P &= \sigma_{\delta,t}^2V_{\delta,t}^PO_t'[B_tP_{t|t-1}B_t' + D_t]^{-1}(\mathbf{z}_t - X_t\boldsymbol{\beta}_t - B_t\boldsymbol{\eta}_{t|t-1}) \\
P_{t|t} &= P_{t|t-1} - P_{t|t-1}B_t'[B_tP_{t|t-1}B_t' + D_t]^{-1}B_tP_{t|t-1} \\
R_{t|t}^P &= \sigma_{\delta,t}^2V_{\delta,t}^P - \sigma_{\delta,t}^2V_{\delta,t}^PO_t'[B_tP_{t|t-1}B_t' + D_t]^{-1}O_tV_{\delta,t}^P\sigma_{\delta,t}^2,
\end{aligned} \tag{2.9}$$

where O_t is defined in (2.1), B_t and X_t are defined in (2.6), D_t is defined in (2.7), and the one-step-ahead forecasts are,

$$\begin{aligned}
\boldsymbol{\eta}_{t|t-1} &= H_t\boldsymbol{\eta}_{t-1|t-1} \\
P_{t|t-1} &= H_tP_{t-1|t-1}H_t' + U_t.
\end{aligned}$$

The smoothing quantities are then obtained by updating "backwards" in time. The smoothing expectations and covariances for the last time point $t = T$ are already calculated

as the last step in (2.9). For $t = T-1, T-2, \ldots, 1$, we calculate,

$$\begin{aligned}
\boldsymbol{\eta}_{t|T} &= \boldsymbol{\eta}_{t|t} + J_t(\boldsymbol{\eta}_{t+1|T} - \boldsymbol{\eta}_{t+1|t}) \\
\boldsymbol{\delta}^P_{t|T} &= \boldsymbol{\delta}^P_{t|t} - M_t(\boldsymbol{\eta}_{t+1|T} - \boldsymbol{\eta}_{t+1|t}) \\
P_{t|T} &= P_{t|t} + J_t(P_{t+1|T} - P_{t+1|t})J_t' \\
R^P_{t|T} &= R^P_{t|t} + M_t(P_{t+1|T} - P_{t+1|t})M_t',
\end{aligned} \tag{2.10}$$

where

$$\begin{aligned}
J_t &:= P_{t|t}H_{t+1}'P^{-1}_{t+1|t} \\
M_t &:= \sigma^2_{\delta,t}V^P_{\delta,t}O_t'P_{t|t-1}B_t'[B_tP_{t|t-1}B_t' + D_t]^{-1}H_{t+1}'P^{-1}_{t+1|t}.
\end{aligned}$$

For the EM algorithm in Section 2.3, we also need the smoothing distribution of the initial state $\boldsymbol{\eta}_0$, specified by

$$\begin{aligned}
\boldsymbol{\eta}_{0|T} &= \boldsymbol{\eta}_{0|0} + J_0(\boldsymbol{\eta}_{1|T} - \boldsymbol{\eta}_{1|t}) \\
P_{0|T} &= P_{0|0} + J_0(P_{1|T} - P_{1|t})J_0'\,.
\end{aligned} \tag{2.11}$$

The cross-covariance term, $P_{t,t-1|T} := cov(\boldsymbol{\eta}_t, \boldsymbol{\eta}_{t-1}|\mathbf{z}_{1:T})$, is given in Shumway and Stoffer (2006, pp. 337-338):

$$\begin{aligned}
P_{T,T-1|T} &= (I_r - P_{T|T-1}B_T'[B_TP_{T|T-1}B_T' + D_T]^{-1}B_T)H_TP_{T-1|T-1} \\
P_{t,t-1|T} &= P_{t|t}J_{t-1}' + J_t(P_{t+1,t|T} - H_{t+1}P_{t|t})J_{t-1}';\ t = T-1, T-2, \ldots, 1.
\end{aligned} \tag{2.12}$$

Clearly, all the filtering and smoothing distributions of $\boldsymbol{\eta}_t$ and $\boldsymbol{\delta}^P_t$ are normal distributions, because they are linear combinations of $\mathbf{z}_{1:T}$, which we defined to be normal in Section 2.2.1. The filtering and smoothing distributions are therefore fully determined by their first two moments given in (2.9) and (2.10).

Finally, we obtain the FRS prediction vectors as,

$$\mathbf{Y}^P_{t|T} := E(\mathbf{Y}^P_t|\mathbf{z}_{1:T}) = X^P_t\boldsymbol{\beta}_t + B^P_t\boldsymbol{\eta}_{t|T} + \boldsymbol{\delta}^P_{t|T};\ t = 1, \ldots, T. \tag{2.13}$$

An important advantage of statistical approaches to prediction problems, over *ad hoc* engineering solutions, is the possibility of uncertainty evaluation. For $t = 1, \ldots, T$, the vectors

of mean squared prediction errors (MSPEs) corresponding to $\mathbf{Y}^P_{t|T}$ can be calculated as,

$$\begin{aligned}\boldsymbol{\sigma}^2_{t|T} &:= diag\{E([\mathbf{Y}^P_t - \mathbf{Y}^P_{t|T}][\mathbf{Y}^P_t - \mathbf{Y}^P_{t|T}]')\} \\ &= diag\{B^P_t P_{t|T} B^{P\prime}_t + R^P_{t|T} - 2\sigma^2_{\delta,t} B^P_t P_{t|t-1} B'_t [B_t P_{t|t-1} B'_t + D_t]^{-1} O_t V^P_{\delta,t}\},\end{aligned} \tag{2.14}$$

where here $diag\{A\}$ denotes the vector of diagonal elements of a matrix A.

The great strength of the STME model (2.2) is in reducing the computational cost of deriving these distributions. The Kalman filtering/smoothing technique employs sequential updating, and so we need only consider the n_t observations taken at a particular time point t at each step (instead of all $\sum_{t=1}^{T} n_t$ observations simultaneously). The inversion of the $n_t \times n_t$ matrix required at each "naive" Kalman filter update can still prove to be prohibitively expensive in some massive-data situations, such as for remote-sensing applications. However, due to the dimension reduction achieved by the basis functions in the vector $\mathbf{b}(\cdot)$ in (2.3), direct inversion of $var(\mathbf{z}_t|\mathbf{z}_{1:t-1}) = [B_t P_{t|t-1} B'_t + D_t]$ in (2.9) can be averted by making use of a Sherman-Morrison-Woodbury formula (Sherman and Morrison, 1950; Woodbury, 1950; Henderson and Searle, 1981),

$$[B_t P_{t|t-1} B'_t + D_t]^{-1} = D_t^{-1} - D_t^{-1} B_t [P^{-1}_{t|t-1} + B'_t D_t^{-1} B_t]^{-1} B'_t D_t^{-1}. \tag{2.15}$$

From (2.15), we only have to invert $r \times r$ matrices and diagonal $n_t \times n_t$ matrices. The number of basis functions, r, is typically chosen to be much smaller than n_t. At each filtering step, the computational complexity is therefore reduced from $\mathcal{O}(n_t^3)$ to $\mathcal{O}(r^3 n_t)$; that is, the required number of operations now increases linearly in n_t (instead of cubic in n_t). This ensures computational feasibility, even for very large or massive datasets.

2.3 Maximum Likelihood Estimation Via an EM Algorithm

The inference described in Section 2.2.2 assumes that the entire vector of parameters, $\boldsymbol{\theta}$, is known. Of course, this will typically not be the case in practice. Here we describe how

to obtain maximum likelihood estimates (MLEs) of the parameters of our spatio-temporal model via an expectation-maximization (EM) algorithm. These estimates can then be substituted into the filtering or smoothing equations to obtain empirical spatio-temporal predictions.

MLEs of variance terms in mixed models are known to be biased downward, since the estimation does not take into account the uncertainty in estimating the fixed effects (McLachlan and Krishnan, 2008, p. 187). Restricted-maximum-likelihood (REML) approaches can remedy this problem. However, here we are interested in data-rich applications. From large datasets, a small number of fixed effects can usually be estimated with extremely high precision, and hence the difference between REML and ML estimates will be negligible. We therefore focus our attention on maximizing the likelihood function.

2.3.1 The Likelihood Function

The likelihood is the probability density function of the data as a function of the unknown parameter vector $\boldsymbol{\theta}$. An MLE of $\boldsymbol{\theta}$ is a value of the vector that maximizes the likelihood. For the STME model (2.2), this likelihood is rather complicated; see the end of Section 2.2.1 where all the components of $\boldsymbol{\theta}$ are given. To derive the likelihood functions, it is helpful to define so-called innovations, $\boldsymbol{\alpha}_t := \mathbf{z}_t - X_t\boldsymbol{\beta}_t - B_t\boldsymbol{\eta}_{t|t-1}$, $t = 1, \ldots, T$, as in Shumway and Stoffer (2006, p. 312-313). These innovations are independent normal random vectors with mean zero and covariance matrix, $\Sigma_{\boldsymbol{\alpha}_t} := B_t P_{t|t-1} B_t' + D_t$, $t = 1, \ldots, T$, respectively. Then,

$$-2\log L(\boldsymbol{\theta}) := -2f(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T|\boldsymbol{\theta}) = \sum_{t=1}^{T} \log|\Sigma_{\boldsymbol{\alpha}_t}(\boldsymbol{\theta})| + \sum_{t=1}^{T} \boldsymbol{\alpha}_t(\boldsymbol{\theta})' \Sigma_{\boldsymbol{\alpha}_t}(\boldsymbol{\theta})^{-1} \boldsymbol{\alpha}_t(\boldsymbol{\theta}) + \text{const.}, \tag{2.16}$$

where henceforth "const." denotes a constant that does not depend on $\boldsymbol{\theta}$, and we have emphasized that both the innovations and their covariance matrices are functions of the parameter vector.

Clearly, analytical optimization of this function with respect to the parameters is highly problematic. Indeed, Katzfuss and Cressie (2009) gave a simple example in the spatial-only case ($T = 1$) that indicates the impossibility of finding MLEs analytically.

Note also that this function depends on H_t and U_t only through $\boldsymbol{\eta}_{t|t-1} := H_t\boldsymbol{\eta}_{t-1|t-1}$ and $P_{t|t-1} := H_tP_{t-1|t-1}H_t' + U_t$. Since we place no constraints on H_t and U_t (other than U_t being a valid covariance matrix), it is clear that there cannot be a unique MLE if H_t and U_t are both allowed to vary freely with t. To achieve identifiability of these parameters, we set $H := H_1 = \ldots = H_T$ and $U := U_1 = \ldots = U_T$ for the remainder of this chapter.

2.3.2 EM Estimation

The EM algorithm (Dempster et al., 1977) has traditionally been employed for ML estimation in STRE models (see, e.g., Shumway and Stoffer, 1982, 2006; Xu and Wikle, 2007). We shall extend this approach here to an STME model of the form given by (2.2). This allows estimation of K_0, H and U, along with $\{\sigma^2_{\delta,t}\}$ and $\{\boldsymbol{\beta}_t\}$, where the index $t = 1, \ldots, T$, all in a single algorithm. A related EM-estimation approach of both random and fixed effects has been considered in Shumway and Stoffer (2006) in a time series context.

We begin with a short review of the EM algorithm. It is an iterative computational technique that can be used to find MLEs in situations where knowledge of some unobserved random variable (called the "missing data"), in addition to the observed data, would make the complete-data likelihood (i.e., the joint distribution function of the observed and the

missing data) much easier to evaluate and maximize than the observed likelihood. At each iteration, the algorithm consists of an expectation step and a maximization step. In our case, if we were able to observe $\boldsymbol{\eta}_t$ and $\boldsymbol{\delta}_t$ for all t directly, then the corresponding terms in the resulting log-likelihood become additive, and we can treat the terms containing $\boldsymbol{\eta}_t$ and $\boldsymbol{\delta}_t$ separately when maximizing the function with respect to the parameters. Thus, in the EM context, we define the missing data to be $\boldsymbol{\eta}_{1:T} := [\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_t']'$ and $\boldsymbol{\delta}_{1:T} := [\boldsymbol{\delta}_1', \ldots, \boldsymbol{\delta}_T']'$. Then, the so-called complete-data log-likelihood is given by,

$$\begin{aligned} -2\log L_c(\boldsymbol{\theta}) &:= -2\log f(\mathbf{z}_{1:T}, \boldsymbol{\eta}_{1:T}, \boldsymbol{\delta}_{1:T}|\boldsymbol{\theta}) \\ &= \textstyle\sum_{t=1}^{T} tr(V_{\epsilon,t}^{-1}[\mathbf{z}_t - X_t\boldsymbol{\beta}_t - B_t\boldsymbol{\eta}_t - \boldsymbol{\delta}_t][\mathbf{z}_t - X_t\boldsymbol{\beta}_t - B_t\boldsymbol{\eta}_t - \boldsymbol{\delta}_t]')/\sigma_{\epsilon,t}^2 \\ &\quad + \textstyle\sum_{t=1}^{T} n_t \log \sigma_{\delta,t}^2 + \sum_{t=1}^{T} tr(V_{\delta,t}^{-1}\boldsymbol{\delta}_t\boldsymbol{\delta}_t')/\sigma_{\delta,t}^2 + \log|K_0| + tr(K_0^{-1}\boldsymbol{\eta}_0\boldsymbol{\eta}_0') \\ &\quad + T\log|U| + \textstyle\sum_{t=1}^{T} tr(U^{-1}[\boldsymbol{\eta}_t - H\boldsymbol{\eta}_{t-1}][\boldsymbol{\eta}_t - H\boldsymbol{\eta}_{t-1}]') + \text{const.} \end{aligned} \tag{2.17}$$

Assume we are at iteration $l+1$ of the EM algorithm. The expectation step of the algorithm consists of finding $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[l]}) := E_{\boldsymbol{\theta}^{[l]}}\{-2\log L_c(\boldsymbol{\theta})|\mathbf{z}_{1:T}\}$, essentially the conditional expectation of the complete-data log-likelihood for $\boldsymbol{\theta} = \boldsymbol{\theta}^{[l]}$ (given the observed data) with respect to the missing data. From a smoothing perspective, the observed data vector is $\mathbf{z}_{1:T}$. Using the current value of the parameter vector, $\boldsymbol{\theta} = \boldsymbol{\theta}^{[l]}$, we apply the FRS equations (2.9)–(2.12) to obtain the conditional expectations and covariance matrices of the "missing data," $\boldsymbol{\eta}_{t|T}^{[l]} := E_{\boldsymbol{\theta}^{[l]}}(\boldsymbol{\eta}_t|\mathbf{z}_{1:T})$, $\boldsymbol{\delta}_{t|T}^{[l]} := E_{\boldsymbol{\theta}^{[l]}}(\boldsymbol{\delta}_t|\mathbf{z}_{1:T})$, along with $P_{t|T}^{[l]} := var_{\boldsymbol{\theta}^{[l]}}(\boldsymbol{\eta}_t|\mathbf{z}_{1:T})$, $R_{t|T}^{[l]} := var_{\boldsymbol{\theta}^{[l]}}(\boldsymbol{\delta}_t|\mathbf{z}_{1:T})$, and $P_{t,t-1|T}^{[l]} := cov_{\boldsymbol{\theta}^{[l]}}(\boldsymbol{\eta}_t, \boldsymbol{\eta}_{t-1}|\mathbf{z}_{1:T})$, for all t. For the FRS procedure applied inside of the EM algorithm, the set of prediction locations is temporarily equal to the observed locations, $\mathcal{S}_t^P = \mathcal{S}_t^O$, for all $t = 1, \ldots, T$. (After EM estimates are obtained, the actual prediction locations $\mathcal{S}_t^P$ are used in the FRS equations to obtain smoothing predictions and MSPEs.)

Defining the quantities $K_t^{[l+1]} := P_{t|T}^{[l]} + \boldsymbol{\eta}_{t|T}^{[l]}\boldsymbol{\eta}_{t|T}^{[l]\prime}$ and $L_t^{[l+1]} := P_{t,t-1|T}^{[l]} + \boldsymbol{\eta}_{t|T}^{[l]}\boldsymbol{\eta}_{t-1|T}^{[l]\prime}$, the expectation step is given by,

$$\begin{aligned}
Q(\boldsymbol{\theta};\boldsymbol{\theta}^{[l]}) &:= E_{\boldsymbol{\theta}^{[l]}}\{-2\log L_c(\boldsymbol{\theta})|\mathbf{z}_{1:T}\} \\
&= \textstyle\sum_{t=1}^{T} \frac{1}{\sigma_{\epsilon,t}^2} tr(V_{\epsilon,t}^{-1}[\mathbf{z}_t - X_t\boldsymbol{\beta}_t - B_t\boldsymbol{\eta}_{t|T}^{[l]} - \boldsymbol{\delta}_{t|T}^{[l]}][\mathbf{z}_t - X_t\boldsymbol{\beta}_t - B_t\boldsymbol{\eta}_{t|T}^{[l]} - \boldsymbol{\delta}_{t|T}^{[l]}]') \\
&\quad + \textstyle\sum_{t=1}^{T} n_t \log\sigma_{\delta,t}^2 + \sum_{t=1}^{T} tr(V_{\delta,t}^{-1}[R_{t|T}^{[l]} + \boldsymbol{\delta}_{t|T}^{[l]}\boldsymbol{\delta}_{t|T}^{[l]\prime}])/\sigma_{\delta,t}^2 + \log|K_0| + tr(K_0^{-1}K_0^{[l+1]}) \\
&\quad + T\log|U| + \textstyle\sum_{t=1}^{T} tr(U^{-1}[K_t^{[l+1]} - HL_t^{[l+1]\prime} - L_t^{[l+1]}H' + HK_{t-1}^{[l+1]}H']) + \text{const.}
\end{aligned} \tag{2.18}$$

In the maximization step, because we are considering the negative log-likelihood, we minimize (2.18) with respect to each of the parameters. This is fairly easy to do, in that (with the exception of H and U) each term of the summation in (2.18) only contains one of the parameters, so most other terms can be disregarded when taking the derivative with respect to that parameter. The minima of (2.18) then define the new value of the parameter vector, $\boldsymbol{\theta}^{[l+1]}$. The updates are,

$$\begin{aligned}
\boldsymbol{\beta}_t^{[l+1]} &= (X_t'V_{\epsilon,t}^{-1}X_t)^{-1}X_t'V_{\epsilon,t}^{-1}[\mathbf{z}_t - B_t\boldsymbol{\eta}_{t|T}^{[l]} - \boldsymbol{\delta}_{t|T}^{[l]}] \\
\sigma_{\delta,t}^{2\ [l+1]} &= tr(V_{\delta,t}^{-1}[R_{t|T}^{[l]} + \boldsymbol{\delta}_{t|T}^{[l]}\boldsymbol{\delta}_{t|T}^{[l]\prime}])/n_t \\
K_0^{[l+1]} &= P_{0|T}^{[l]} + \boldsymbol{\eta}_{0|T}^{[l]}\boldsymbol{\eta}_{0|T}^{[l]\prime} \\
H^{[l+1]} &= (\textstyle\sum_{t=1}^{T} L_t^{[l+1]})(\sum_{t=0}^{T-1} K_t^{[l+1]})^{-1} \\
U^{[l+1]} &= (\textstyle\sum_{t=1}^{T} K_t^{[l+1]} - H^{[l+1]}\sum_{t=1}^{T} L_t^{[l+1]\prime})/T.
\end{aligned} \tag{2.19}$$

Summary of the EM Algorithm for the STME Model

1. Choose initial values $\boldsymbol{\theta}^{[0]}$ in the parameter space $\boldsymbol{\Theta}$ for the parameters.
2. For $l = 0, 1, 2, \ldots$ (until convergence):
   (a) Carry out FRS (with $\mathcal{S}_t^P = \mathcal{S}_t^O$) using $\boldsymbol{\theta} = \boldsymbol{\theta}^{[l]}$ to obtain the smoothing quantities as described in (2.9)–(2.12).
   (b) Obtain $\boldsymbol{\theta}^{[l+1]}$ by calculating the updates given in (2.19).

Due to the large number of parameters, we have found it most convenient to monitor the convergence of the algorithm on the basis of the sequence of likelihood values, obtained by evaluating the (observed negative log-) likelihood (2.16) at the current value of the parameter vector at each iteration of the EM algorithm. Note that to evaluate the log-likelihood, we must find the determinants and the inverses of the $n_t \times n_t$ covariance matrices, $\Sigma_{\boldsymbol{\alpha}_t} = B_t P_{t|t-1} B_t' + D_t$, for $t = 1, \ldots, T$. The inversions can be obtained as described in (2.15). A computationally convenient formula (e.g., Cressie and Johannesson, 2008) for determinants yields,

$$|\Sigma_{\boldsymbol{\alpha}_t}| = |D_t||P_{t|t-1}||P_{t|t-1}^{-1} + B_t' D_t^{-1} B_t| \,. \tag{2.20}$$

Calculating these determinants directly leads to serious numerical instabilities for massive datasets. Fortunately, for the purpose of monitoring convergence using the log-likelihood, we are really interested in the sum of the *logarithm* of these quantities. All three quantities on the right-hand side of (2.20) can be calculated by making use of the fact that the log-determinant of a generic $N \times N$ matrix A is given by $\log|A| = \sum_{i=1}^{N} \log \lambda_i$, where λ_i are the eigenvalues of A.

2.3.3 Properties of the EM Estimator

Let the EM estimator be $\hat{\boldsymbol{\theta}}_{EM}$, the value of the parameter vector (or one of them, if there are multiple) that is obtained from the EM algorithm when the sequence of likelihood values has converged. Because $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[l]})$ given by (2.18) is continuous in both arguments (i.e., $\boldsymbol{\theta}, \boldsymbol{\theta}^{[l]} \in \boldsymbol{\Theta}$), this convergence is guaranteed for our algorithm by a theorem in Wu (1983). According to the same theorem, $\hat{\boldsymbol{\theta}}_{EM}$ must be a solution to the likelihood equations (the set of equations obtained by setting the partial derivatives of (2.16) equal to zero). This does

not mean that $\hat{\boldsymbol{\theta}}_{EM}$ is, in fact, the unique MLE, as there could be several (or even infinitely many) solutions to the likelihood equations. Nevertheless, under some general conditions, $\hat{\boldsymbol{\theta}}_{EM}$ is both consistent and asymptotically normal (see Hannan and Deistler, 1988, Ch. 4, for more details). Unfortunately, asymptotic results are often of little use when the number of unknown parameters is large relative to the number of time points T. For time series of short-to-moderate length, Shumway (2006) discusses a bootstrapping approach to assess the uncertainty in the parameter estimates. Our emphasis in this chapter is instead on smoothing the unknown process, for which parameter estimates of $\boldsymbol{\theta}$ are needed.

If the algorithm is initialized with parameters in the parameter space (i.e., $\boldsymbol{\theta}^{[0]} \in \boldsymbol{\Theta}$), then we can see from (2.19) that $\boldsymbol{\theta}^{[l]} \in \boldsymbol{\Theta}$, $l = 1, 2, \ldots$. Specifically, this means that if the initial values for the covariance-matrix parameters K_0 and U are proper covariance matrices, then the EM updates, $K_0^{[l+1]} = P_{0|T}^{[l]} + \boldsymbol{\eta}_{0|T}^{[l]} \boldsymbol{\eta}_{0|T}^{[l]\prime}$ and $U^{[l+1]} = \sum_{t=1}^{T} E_{\boldsymbol{\theta}^{[l]}}([\boldsymbol{\eta}_t - H\boldsymbol{\eta}_{t-1}][\boldsymbol{\eta}_t - H\boldsymbol{\eta}_{t-1}]'|\mathbf{z}_{1:T})/T$, and hence the EM estimators of these matrix parameters, are also symmetric and at least nonnegative-definite. Similarly, if we choose $\sigma_{\delta,t}^{2\,[0]} > 0$, then it is guaranteed that $\hat{\sigma}_{\delta,t;EM}^2 \geq 0$. This is a very desirable property, as the constraints on a positive-definite matrix can be very hard to handle when optimizing a function with respect to that matrix (see, e.g., Katzfuss and Cressie, 2009). Here, these constraints are satisfied automatically.

Since the research in this chapter is geared toward computational efficiency for massive datasets, the computational cost of the algorithm is highly relevant. As we have described at the end of Section 2.2.2, the computational complexity of the FRS that needs to be carried out at each iteration of the EM algorithm is linear in each n_t. The fact that this smoothing procedure has to be carried out several times for the EM algorithm (until convergence is reached) does not change its theoretical computational complexity as a function of the

number of observations, and so the algorithm as a whole is scalable. It should be noted though that the stability and other desirable features of the EM algorithm come at the cost of a rather slow convergence (McLachlan and Krishnan, 2008), so if a high precision in estimating the parameters is required, the number of iterations needed until convergence might be quite large.

2.3.4 Possible Extensions

The EM algorithm described above can be easily modified to the spatial-only case. If all measurements are regarded to be from the same time point (say time $t = 1$), interest is in estimating the unknown parameters $\boldsymbol{\beta}_1$, $\sigma^2_{\delta,1}$, and $K_1 = HK_0H' + U$, from data $\mathbf{z}_1$. Thus, if we put $T = 1$, the EM algorithm becomes much simpler. The FRS algorithm reduces to a spatial-only Fixed Rank Kriging procedure (Cressie and Johannesson, 2008). For details of the EM algorithm in the spatial-only case, see Katzfuss and Cressie (2009).

Note that we might also want to force the trend coefficients, $\{\boldsymbol{\beta}_t\}$, and/or the fine-scale-variation variances, $\{\sigma^2_{\delta,t}\}$, to be constant for a certain number of time steps, in order to achieve greater stability for estimation of these parameters when the data are sparse. In addition, if the total number of time steps T is large, we might not want the propagator matrices $\{H_t\}$ and the innovation matrices $\{U_t\}$ to be constant for all time steps, but they might be allowed to change every few time steps. The EM algorithm can be easily modified to accommodate such modeling assumptions. All that one needs to change is the form of the updates in (2.19), by summing (or not summing) quantities on the right-hand side of (2.18) over appropriate time steps. For example, if we assume that $H_1 = \ldots = H_{T_1} \neq H_{T_1+1} = \ldots = H_T$ and $U_1 = \ldots = U_{T_1} \neq U_{T_1+1} = \ldots = U_T$, the EM updates for H_1 and

U_1 become,

$$H_1^{[l+1]} = (\sum_{t=1}^{T_1} L_t^{[l+1]})(\sum_{t=0}^{T_1-1} K_t^{[l+1]})^{-1}$$
$$U_1^{[l+1]} = (\sum_{t=1}^{T_1} K_t^{[l+1]} - H_1^{[l+1]} \sum_{t=1}^{T_1} L_t^{[l+1]\prime})/T_1.$$

Similarly, if we assume $\sigma^2_{\delta,t}$ to be constant for all $t = 1, \ldots, T$, then the EM update for this parameter becomes,

$$\sigma_\delta^{2[l+1]} = \sum_{t=1}^{T} tr(V_{\delta,t}^{-1}[R_{t|T}^{[l]} + \boldsymbol{\delta}_{t|T}^{[l]}\boldsymbol{\delta}_{t|T}^{[l]\prime}])/\sum_{t=1}^{T} n_t. \tag{2.21}$$

As we briefly mentioned in Section 2.1, one might want to avoid having to make the assumption that the function $v_{\delta,t}(\cdot)$, which describes the variance heterogeneity of the fine-scale variation, is known. While one could, for simplicity, always set $v_{\delta,t}(\cdot) \equiv 1$, this might not be appropriate in some situations where the true process exhibits different variability and/or smoothness in different parts of the spatial domain. If no expert knowledge about the form of this variance function is available, the function could be modeled as $v_{\delta,t}(\cdot) = \exp\{\mathbf{b}_\delta(\cdot)'\boldsymbol{\eta}_\delta\}$, where $\mathbf{b}_\delta(\cdot)$ is a vector of r_δ basis functions, with typically $r_\delta < r$. The vector $\boldsymbol{\eta}_\delta$ is not identifiable if both σ^2_δ and $\boldsymbol{\eta}_\delta$ are allowed to vary freely, and so we assume $\boldsymbol{\eta}_\delta \sim N_{r_\delta}(\mathbf{0}, \sigma^2_{\eta_\delta} I_{r_\delta})$, for some fixed $\sigma^2_{\eta_\delta}$. This parameter can be chosen according to prior beliefs via a calibration exercise. For example, if one chooses $1/2$ and 2 as the lower and upper endpoints of a 95% credible interval for the ratio of two fine-scale-variation variances at two distant locations in the spatial domain D_s, we obtain $\sigma^2_{\eta_\delta} \approx 0.25^2$ (for more details, see Chapter 3). We can find the posterior mode as an estimate of $\boldsymbol{\eta}_\delta$ by augmenting the EM updates (2.19) by one Newton-Raphson step,

$$\boldsymbol{\eta}_\delta^{[l+1]} = \boldsymbol{\eta}_\delta^{[l]} - \left(\frac{\partial Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[l]})}{\partial \boldsymbol{\eta}_\delta' \partial \boldsymbol{\eta}_\delta}\Big|_{\boldsymbol{\eta}_\delta = \boldsymbol{\eta}_\delta^{[l]}}\right)^{-1} \left(\frac{\partial Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{[l]})}{\partial \boldsymbol{\eta}_\delta}\Big|_{\boldsymbol{\eta}_\delta = \boldsymbol{\eta}_\delta^{[l]}}\right),$$

where

$$\frac{\partial Q(\boldsymbol{\theta};\boldsymbol{\theta}^{[l]})}{\partial \boldsymbol{\eta}_\delta} = \sum_{t=1}^{T} B'_{\delta,t}(I_{n_t} - \Lambda_t/\sigma^2_{\delta,t})\mathbf{1}_{n_t} + 2\boldsymbol{\eta}_\delta/\sigma^2_{\eta_\delta}$$
$$\frac{\partial Q(\boldsymbol{\theta};\boldsymbol{\theta}^{[l]})}{\partial \boldsymbol{\eta}'_\delta \partial \boldsymbol{\eta}_\delta} = \sum_{t=1}^{T} B'_{\delta,t}\Lambda_t B_{\delta,t}/\sigma^2_{\delta,t} + 2I_{r_\delta}/\sigma^2_{\eta_\delta},$$

and $\Lambda_t := diag(\{E(\delta(\mathbf{s}_{i,t})^2|\mathbf{z}_{1:T})\exp(-\mathbf{b}_\delta(\mathbf{s}_{i,t})'\boldsymbol{\eta}_\delta) : i = 1,\ldots,n_t\})$. The resulting algorithm is now a Generalized EM algorithm (see McLachlan and Krishnan, 2008, for details).

The algorithm can also be modified to take a filtering perspective. In this context, for the conditional expectations in (2.18), the conditioning would be on $\mathbf{z}_{1:t}$ (instead of $\mathbf{z}_{1:T}$). The FRS component of the EM algorithm can be easily modified to a Fixed Rank Filtering (FRF) approach, by carrying out only (2.9) and (2.12), and leaving out (2.10) and (2.11). The smoothing quantities in (2.19) would be replaced by the corresponding FRF quantities. At each time point t, we obtain a new set of data, $\mathbf{z}_t$, and a new EM algorithm can be run to estimate the time-dependent quantities $\boldsymbol{\beta}_t$ and $\sigma^2_{\delta,t}$. Unfortunately, it is not so clear how H_t and U_t would be handled. As mentioned above, it is not sensible to allow them to vary freely for each time point. On the other hand, making them constant for all time points requires running the EM algorithm again for all measurements $\mathbf{z}_{1:t}$ observed so far, every time a new set of data $\mathbf{z}_t$ is obtained. This would quickly become infeasible as t increases. More practically, one could hold H_t and U_t constant only for a certain number of time units (eight, say), as described in the previous paragraph. This way, one has to "go back" only eight time units when estimating the current H_t and U_t.

2.4 Simulation Study: Comparison of EM Estimation to Binned Method-of-Moments Estimation

Previously, frequentist estimation of the STME model (2.2) was carried out using a binned method-of-moments (MM) technique (Cressie and Johannesson, 2008; Kang et al., 2010). As such, the MM estimation is the natural candidate for a comparison to our proposed EM estimation.

2.4.1 Simulation Setup

Our simulated data are meant to resemble a very simplistic version of satellite data in only one spatial dimension. The spatial domain consists of the locations $\mathcal{S}^P = \{1, \ldots, 256\}$, and there are $T = 16$ time points. The "satellite" here has a repeat cycle of two time units. The two tracks of the satellite have a width of 64: For odd time points, the tracks are $\{1, \ldots, 64\}$ and $\{129, \ldots, 192\}$; for even time points, the tracks are $\{65, \ldots, 128\}$ and $\{193, \ldots, 256\}$. Within each track at each time point, 50% of the values are declared missing at random. This is meant to simulate non-retrieval due to cloud cover and other problems. This setup leaves us with 64 observations at each time point.

As basis functions, we use bisquare functions,

$$g_{bi}(\mathbf{s}) := \{1 - (\|\mathbf{s} - \mathbf{c}\|/w)^2\}^2 I(\|\mathbf{s} - \mathbf{c}\| < w), \tag{2.22}$$

where $\mathbf{c}$ is the center point, w is the range, and $I(\cdot)$ is an indicator function. In this simulation study, we have $r = 5$ bisquare basis functions, all with a range of $w = 96$, and centered at 0.5, 64.5, 128.5, 192.5, and 256.5, respectively. An example of the data simulated from the STME model (2.2)–(2.5) is shown in Figure 2.1. (We only show the first four time points; the setup for the rest of the time points is analogous.)

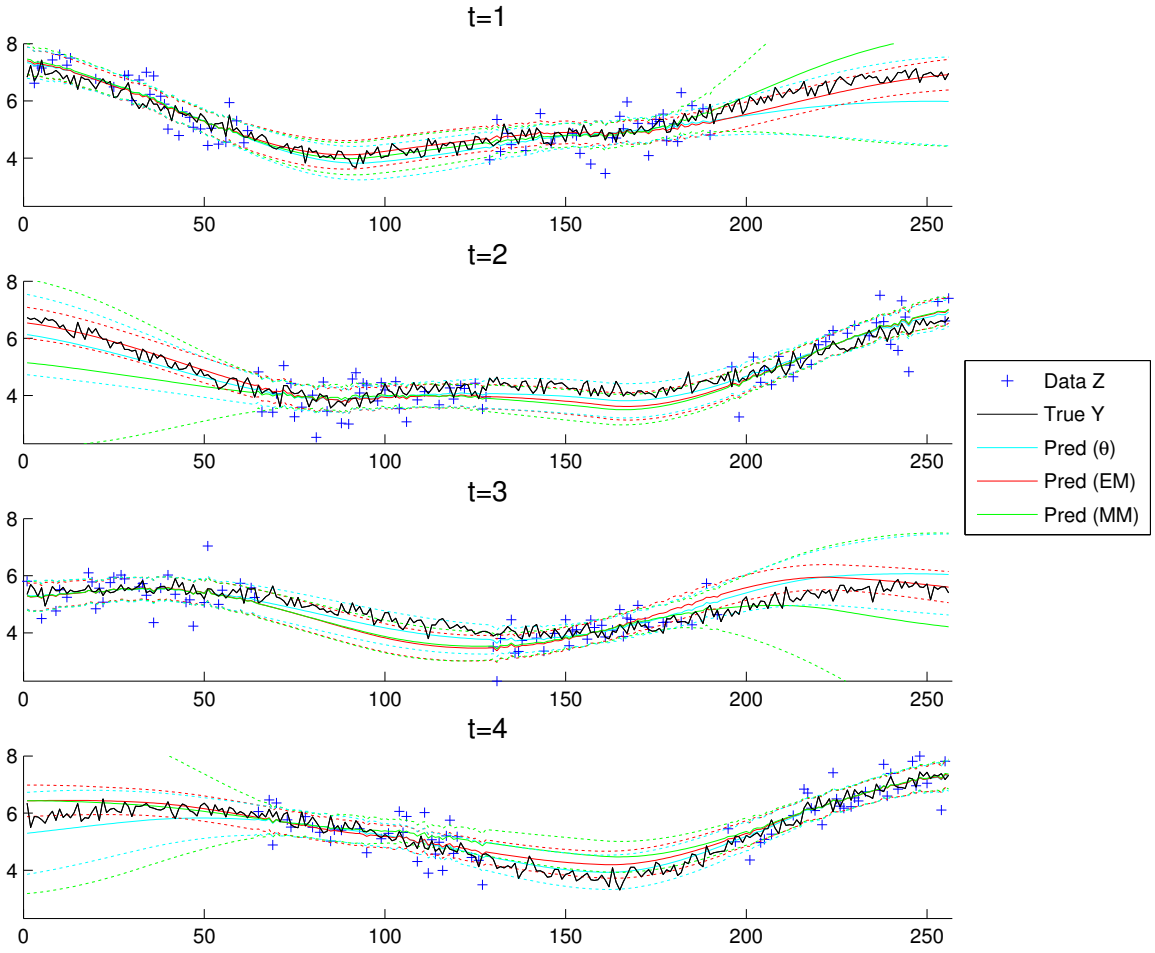

Figure 2.1: Example of the data observed at the first four time points in our simulation study for SNR=2. Also shown are the FRS predictions using the true parameter values, the EM parameter estimates, and the MM parameter estimates, respectively, as solid lines; dotted lines are the respective 95% confidence intervals. These should be compared to the true-process values in black.

The true parameters are calibrated in a manner similar to that of the simulation study described in Cressie et al. (2010). The covariance matrix of $\boldsymbol{\eta}_t$, namely K_t, is taken to be stationary. The matrix parameters K_0, H, and U are chosen to match (as measured by the Frobenius norm) an exponential-covariance matrix with (i, j)-th entry, $\exp(-|i - j|/25)$, and a lag-1 temporal correlation of 0.8. Choosing a fine-scale-variation proportion of 0.05 results in $\sigma^2_\delta = 0.0321$, which is held constant over time. The measurement-error variance is assumed known and also held constant over time. It is determined by the signal-to-noise ratio (SNR), for which we chose two levels: SNR=2, resulting in $\sigma^2_\epsilon = 0.3206$, and SNR=5, resulting in $\sigma^2_\epsilon = 0.1282$. Finally, we chose a constant mean of $\mu = 5$ (i.e., $\mathbf{x}_t(\cdot) \equiv 1$ and $\boldsymbol{\beta}_t \equiv 5$ for all t).

The EM estimation is described in Section 2.3, but due to the constant fine-scale variance σ^2_δ, the update for this term in the algorithm will be of the form (2.21). As there is some sensitivity to the initial values in the algorithm (which is evidence for multiple local maxima of the likelihood, at least for some of the simulated datasets), we initialized the algorithm at the true parameter values in this simulation study.

The MM estimation is largely as described in Kang et al. (2010). We begin by obtaining an ordinary-least-squares estimate of the trend at each time point, which in this case is simply the data mean for each t. We also obtain a pooled estimate of σ^2_δ by using the variogram-extrapolation technique from Kang et al. (2009), but accounting for the fact that σ^2_ϵ is already known, and using all pairs of adjacent data points from all times t. To estimate the matrix parameters, we divided the spatial domain into 16 bins of size 16 each. Due to the missing data, many bins at each time point will have very few or even no observations. To avoid unnecessary instability in the estimates, we discard all bins containing less than five observations. Kang et al. (2010) let both H and U vary with time. To allow for

estimation of a constant propagator matrix and innovation matrix, we have to modify the MM approach slightly. After having obtained MM estimates $\{\hat{K}_{MM,1}, \ldots, \hat{K}_{MM,T}\}$ and $\{\hat{L}_{MM,2}, \ldots, \hat{L}_{MM,T}\}$ as described in Kang et al. (2010), we combine these estimates in a similar fashion as in the EM algorithm (2.19), which results in:

$$\begin{aligned}
\hat{H}_{MM} &= (\textstyle\sum_{t=2}^{T} \hat{L}_{MM,t})(\sum_{t=1}^{T-1} \hat{K}_{MM,t})^{-1} \\
\hat{U}_{MM} &= (\textstyle\sum_{t=2}^{T} \hat{K}_{MM,t} - \hat{H}_{MM} \sum_{t=2}^{T} \hat{L}'_{MM,t})/(T-1).
\end{aligned}$$

Finally, we substitute both sets of parameters into the FRS equations, to obtain the smoothed predictions at all 256 spatial locations at all T=16 time points. As a reference, we also obtain predictions using the true parameters $\boldsymbol{\theta}$.

For the MM procedure, there is no estimate of K_0, since there are no data at $t = 0$. To obtain predictions using the MM estimates, we modify the filtering updates (2.9) for time point $t = 1$ as follows. The conditional expectations and variance-covariance matrices of $\boldsymbol{\eta}_1$ and $\boldsymbol{\delta}_1^P$ are,

$$\begin{aligned}
\boldsymbol{\eta}_{1|1} &= K_1 B_1' [B_1 K_1 B_1' + D_1]^{-1} (\mathbf{z}_1 - X_1 \boldsymbol{\beta}_1) \\
\boldsymbol{\delta}_{1|1}^P &= \sigma_{\delta,1}^2 V_{\delta,1}^P O_1' [B_1 K_1 B_1' + D_1]^{-1} (\mathbf{z}_1 - X_1 \boldsymbol{\beta}_1) \\
P_{1|1} &= K_1 - K_1 B_1' \Sigma_1^{-1} B_1 K_1 \\
R_{1|1}^P &= \sigma_{\delta,1}^2 V_{\delta,1}^P - \sigma_{\delta,1}^2 V_{\delta,1}^P O_1' [B_1 K_1 B_1' + D_1]^{-1} O_1 V_{\delta,1}^P \sigma_{\delta,1}^2.
\end{aligned} \tag{2.23}$$

Then the filtering steps for $t = 2, \ldots, T$ are the same as before. So is the smoothing procedure, but now $P_{1|0} = K_1$, and we stop the backward smoothing at $t = 1$ (instead of $t = 0$).

2.4.2 Simulation Results

Using the setup described in Section 2.4.1, we generated 2,000 datasets for both levels of the SNR. For each dataset, we carried out both EM and MM estimation, and we obtained

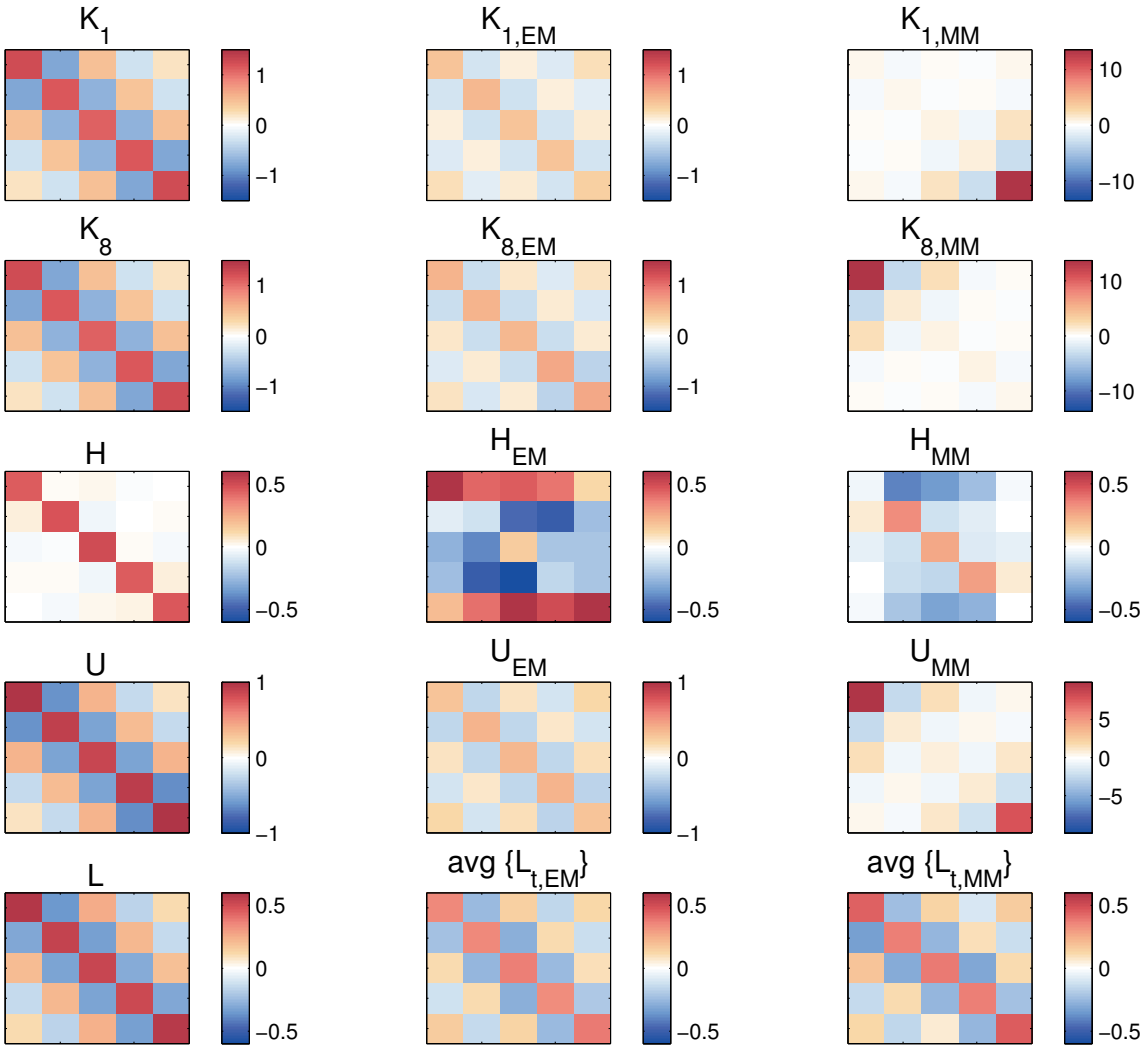

Figure 2.2: Medians (elementwise) of the parameter estimates for SNR=2 in the simulation study. The left, middle, and right columns contain the true parameters, the EM estimates, and the MM estimates, respectively. Note that the scale for the MM estimates is not always the same as for the true parameters.

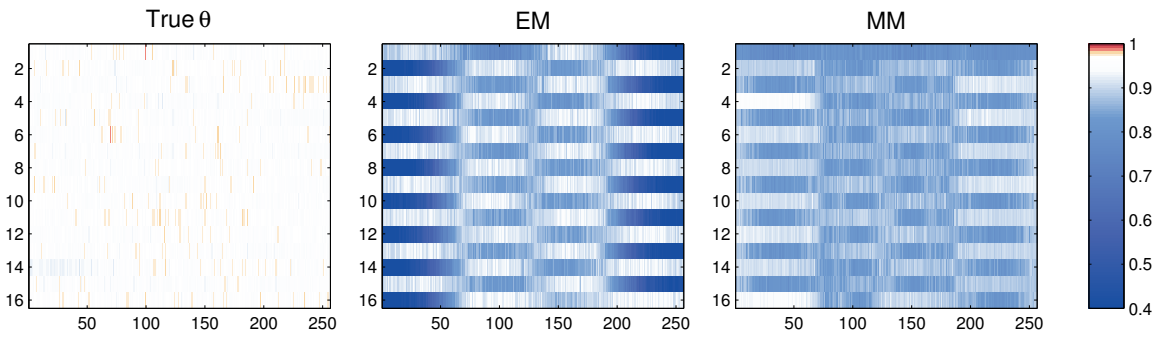

Figure 2.3: Proportion of times the 95% prediction interval covered the true Y at each spatial and temporal location (for SNR=2) in the simulation study, for the predictions using the true parameters (left), the EM estimates (middle), and the MM estimates (right). The rows in each panel correspond to the 16 time units, and the x-axis corresponds to the (one-dimensional) space.

predictions and standard errors via FRS using (i) the known true parameters $\boldsymbol{\theta}$, (ii) the EM parameter estimates, and (iii) the MM parameter estimates. One such spatio-temporal dataset, along with the true process, predictions, and prediction intervals (PIs) is shown in Figure 2.1.

Due to the large amount of missing data, and the medium-to-small SNR, this is a fairly hard problem to handle. In many cases, it was not possible to preserve the total variability when lifting eigenvalues in the MM procedure (Kang et al., 2010). In addition, even in cases where K_t and L_t were estimated successfully, the MM estimate of U was not always positive-definite. Sometimes the EM algorithm did not converge in (the aribitrarily set maximum number of) 200 iterations. The proportion of successful estimations for both procedures are summarized in the first line of Table 2.1. All results in this section are based on only the datasets for which both estimation procedures produced valid estimates.

Elementwise medians of the valid EM and MM matrix-parameter estimates, along with the true parameters, are shown in Figure 2.2 for the SNR=2 scenario. Even in this relatively

Table 2.1: Results of the simulation experiment, with the following acronyms: MSPE = (Empirical) Mean Squared Prediction Error, PIC = (Empirical) Prediction Interval Coverage (95%), MSEE = (Empirical) Mean Squared Estimation Error

	SNR=2				SNR=5			
	True $\boldsymbol{\theta}$	EM	MM	MM/EM	True $\boldsymbol{\theta}$	EM	MM	MM/EM
Success rate	—	0.9775	0.2615	—	—	0.9495	0.5965	—
MSPE	0.1151	0.2028	0.4440	2.1899	0.0920	0.1589	0.2358	1.4840
MSPE - on track	0.0503	0.0556	0.0577	1.0376	0.0375	0.0394	0.0402	1.0215
MSPE - off track	0.1798	0.3499	0.8303	2.3732	0.1464	0.2785	0.4314	1.5493
PIC (t=8, s=96)	0.9511	0.9159	0.8317	—	0.9615	0.9453	0.9444	—
PIC (t=7, s=96)	0.9511	0.8102	0.8650	—	0.9552	0.8737	0.9194	—
PIC (t=2, s=32)	0.9550	0.4442	0.8924	—	0.9489	0.4633	0.9203	—
MSEE (σ_δ^2) $\times 100$	—	0.0058	0.0614	10.501	—	0.0026	0.0121	4.6345
MSEE (μ_t)	—	0.2345	0.2379	1.0145	—	0.2333	0.2403	1.0297

high-noise situation, the EM parameter estimates are acceptable. Due to the small sample sizes at each time unit, the typical negative bias of maximum likelihood estimation (here, EM estimation) of variance terms is apparent, which leads to an overfitting of the elements of H. The MM estimation is heavily affected by the presence or absence of data. The MM estimates of variances that correspond to basis functions in off-track regions are much bigger than the true values.

From Table 2.1, spatio-temporal prediction using the EM estimates is much more accurate than the prediction using MM estimates; this effect is especially strong at off-track locations, which are locations in a swath where no data were observed (at every other time point). For SNR=2, the empirical mean squared prediction error (MSPE) is almost 2.5 times as big for MM than for EM at these off-track locations.

Table 2.1 also shows some other summaries of the experiment. It is clear that parameter estimation is more efficient using the EM algorithm, and the difference between the results

based on EM and MM is more pronounced when the SNR is lower (i.e., when the estimation is more challenging). The estimation of the mean is fairly easy, and so the difference between EM and MM in terms of the (empirical) mean squared estimation error (MSEE) is relatively small there. When estimating the fine-scale variance, σ_δ^2, EM estimation is far superior to MM estimation.

We also examined the validity of 95% prediction intervals (PIs) produced by FRS based on the true parameters and both parameter-estimation methods. Figure 2.3 shows the proportion of times (out of the 2000 simulated datasets) that these prediction intervals covered the true process value $Y_t(i)$ at each location $i = 1, \ldots, 256$ and at each time point $t = 1, \ldots, 16$. Reassuringly, using the true parameters in the FRS procedure leads to very precise assessment of the variability associated with the FRS predictions; the PI coverage (PIC) shown in the left panel is very close to 95% for all $\mathbf{s}$ and all t. The empirical-Bayes approach (i.e., "plugging in" of parameter estimates) taken in this chapter does not account for uncertainty in the estimation of parameters. This becomes plainly apparent in the middle and right panels of Figure 2.3. The PIs produced with the MM estimates are generally too liberal, but the coverage is never too bad, due to the severe overestimation of variance terms corresponding to off-track locations, especially at the edges of the spatial domain (see again Figure 2.2). The EM estimates of the covariance matrices, $\{K_t\}$, were closer to the truth, and so not accounting for the uncertainty in parameter estimation results in severe undercoverage of the PIs in off-track regions at the edge of the spatial domain ($\mathbf{s} \leq 32$ for t even, $\mathbf{s} \geq 193$ for t odd), despite the higher prediction accuracy. This edge effect is likely to be much less problematic for an analysis of real data on the globe, which does not have edges.

2.5 Application: Analysis of Global Satellite CO_2 Data

In this section, we shall demonstrate an application of the proposed FRS prediction and EM parameter estimation to real data. Despite a high-dimensional parameter space, smoothing a very large, spatio-temporal dataset of global CO_2 measurements is fast and produces reasonable results.

2.5.1 Mid-tropospheric CO_2 Measurements by AIRS

We consider a spatio-temporal dataset consisting of 16 days of measurements of global mid-tropospheric CO_2, available from `http://airs.jpl.nasa.gov/AIRS_CO2_Data/`. The data were recorded by the Atmospheric InfraRed Sounder (AIRS) (Chahine et al., 2006) on board the Aqua satellite, which is part of the "A-train" of Earth-observing satellites. Remote sensing via satellites provides information on the Earth's land, atmosphere, and ocean that would otherwise be too costly or too dangerous to obtain. Measurements from these platforms can be actual images, but spectra are more useful for many purposes (for an introduction to remote sensing, see, e.g., Landgrebe, 2003). Satellite measurements of CO_2 are based on the latter approach: Using an elaborate retrieval algorithm, the spectral radiances obtained by the AIRS instrument are converted to CO_2 concentrations (Chahine et al., 2006) in the mid-troposphere (roughly, between 2km and 8km in altitude).

The data represent measurements of mid-tropospheric CO_2 between -60° and 90° latitude at roughly 1:30pm local time on May 1 through May 16, 2003, from now on referred to as days 1 through 16, respectively; data at latitudes south of -60° have not been released yet. The unit of measurement is parts per million (ppm).

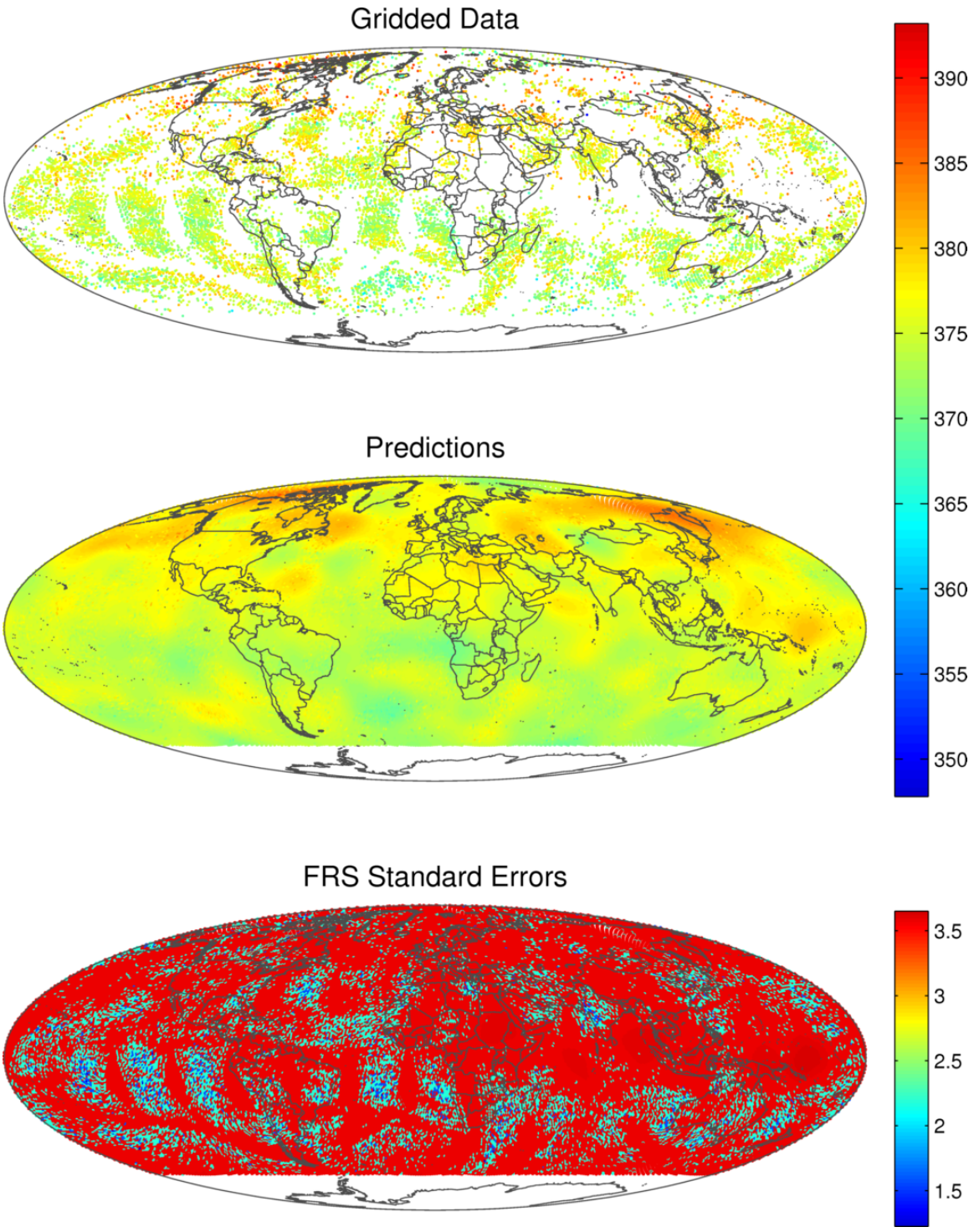

Figure 2.4: Mid-tropospheric CO_2 as measured by AIRS on May 1, 2003, with corresponding FRS predictions and standard errors, both obtained using EM parameter estimates. Units are ppm.

While our methodology does not require gridded data, higher-level data products derived from remote-sensing measurements are often provided in more or less aggregated and gridded form, and hence we moved the original data to a regular grid. The gridding is purely for illustration, as concepts such as the proportion of missing data are clearer in a gridded setting, and we want to emphasize that our methodology can deal with arbitrary and changing (over time) measurement locations. Treating the grid locations as the true measurement locations results in a slight location error, but this error tends to have a very small effect on the prediction error (Cressie and Kornak, 2003). For the remainder of this section, we regard the process evaluated at the grid to be the quantity of interest.

Using DGGRID software (Sahr, 2003), we generated a hexagonal grid across the globe, with the goal of matching the area of the roughly circular AIRS footprint of about 45km radius as closely as possible (ISEA Aperture 3 Hexagons at resolution 8). Discarding all grid cells with centers south of -60° latitude, we obtain $m = 61,236$ grid cells at which the process is to be predicted. If a particular grid cell contains several of the original measurements on a particular day, the data value at that grid cell was taken to be the average of those measurements and the measurement-error covariance matrix, $V_{\epsilon,t}$, was appropriately modified (e.g., Cressie and Johannesson, 2008). The resulting gridded data for the first day is shown in the top panel of Figure 2.4. The number of observed grid cells per day, $\{n_t\}$, was fairly stable over time, between 11,862 and 12,971. This puts the proportion of grid cells with missing data at each time point at around 80%.

As mentioned before in Section (2.2.1), our EM-estimation scheme assumes that the measurement-error variances, $\sigma^2_{\epsilon,t}$, are known for all time points $t = 1, \ldots, T$. For this analysis of AIRS data, we obtained estimates of these variances using the variogram-extrapolation technique described in Kang et al. (2009), accounting for the number of

measurements that went into each grid cell average. As $\hat{\sigma}^2_{\epsilon,t}$ was fairly constant over time (ranging roughly between 5.2 and 6), we used a pooled estimate, $\hat{\sigma}^2_{\epsilon} = 5.6062$. To account for the averaging that occurred in obtaining the gridded data from the original measurements, we set $v_{\epsilon,t}(S_{i,t}) = 1/N_t(S_{i,t})$, where $N_t(S_{i,t})$ is the number of measurements going into grid cell $S_{i,t}$ at time t; see equation (2.7).

For simplicity, we assumed that the variance of the fine-scale variation term is constant over both time and space: $\sigma^2_{\delta} := \sigma^2_{\delta,1} = \ldots = \sigma^2_{\delta,16}$ and $v_{\delta,t}(S_{i,t}) \equiv 1$ for all i and t. Based on exploratory data analysis and consultation with carbon-cycle experts, the large-scale spatial trend in CO_2 was modeled by an intercept and a latitudinal gradient; that is, we set $\mathbf{x}_t(\cdot) = [1 \;\; lat(\cdot)]'$, independent of t. However, the vector of trend coefficients $\boldsymbol{\beta}_t$ is assumed to vary with time index t.

The model (2.2) assumes that observations were made at the point level. Here, our gridded data has areal support in the shape of a hexagon. Denoting the set of prediction grid cells by $\mathcal{S}^P = \{S_1, \ldots, S_m\}$ (i.e., they are the same for all t), we modify the process model to be,

$$Y_t(S_i) = \mathbf{x}_t(S_i)'\boldsymbol{\beta}_t + \mathbf{b}_t(S_i)'\boldsymbol{\eta}_t + \delta_t(S_i);\; i = 1, \ldots, m,\; t = 1, 2, \ldots,$$

where $g(S) := \int_S g(\mathbf{s})d\mathbf{s}/|S|$ for a generic function $g(\cdot)$, and $|S|$ is the area of S. The integrated fine-scale variation has the same distributional assumptions as it had at the point level, and the variance parameters are described in the previous paragraph. The integration over the trend $\mathbf{x}_t(\cdot)$ and the basis-function vector $\mathbf{b}_t(\cdot)$ is practically more challenging. For the purpose of this data analysis, we approximated the terms $\mathbf{x}_t(S_i)$ and $\mathbf{b}_t(S_i)$ by a Monte Carlo integration, namely by drawing a uniform random sample of 40 points within each grid hexagon, evaluating the functions at each point, and then averaging over the results. Then one can simply replace $\mathbf{x}_t(\mathbf{s}_i)$ and $\mathbf{b}_t(\mathbf{s}_i)$ in the FRS equations (2.9)–(2.12) by $\mathbf{x}_t(S_i)$

and $\mathbf{b}_t(S_i)$, respectively, to model observations (and obtain predictions) at the hexagonal grid support (instead of at the point level).

2.5.2 Bisquare Basis Functions on the Globe

It is a somewhat open problem how many and what type of basis functions to choose for the STRE component (2.3) of the STME model (2.2). In this application, the spatial domain of interest is the globe, which we assume to be a perfect sphere. W-wavelets, which have been successfully used in the STRE context (see, e.g., Kang et al., 2010), cannot be evaluated on the sphere. Instead, we follow Cressie and Johannesson (2008) in using bisquare basis functions of the form (2.22). This choice is only for illustration; our methodology is purposely left very general to allow for any choice of basis functions (for a review of some common choices, see Wikle, 2010). However, bisquare functions do have a number of desirable properties, in that they have a clear center and range, they can be defined on any spatial domain for which a measure of distance is available, they can be evaluated at any point in space (without interpolation), and they can be interpreted as convolution kernels.

It is generally recommended to employ several resolutions of basis functions, to capture different scales of spatial variation in the underlying process. Using the same settings as for the prediction grid above (ISEA Aperture 3 Hexagons), the DGGRID software provides us with a suitable set of basis-function centers at the first three resolutions. Some previous simulation experiments, similar to the one described in Section 2.4, have provided us with strong evidence that prediction results using our model are best when basis functions of different resolutions do not share the same centers. When using the DGGRID program with the default orientation, centers of coarser resolutions will always coalesce with centers of

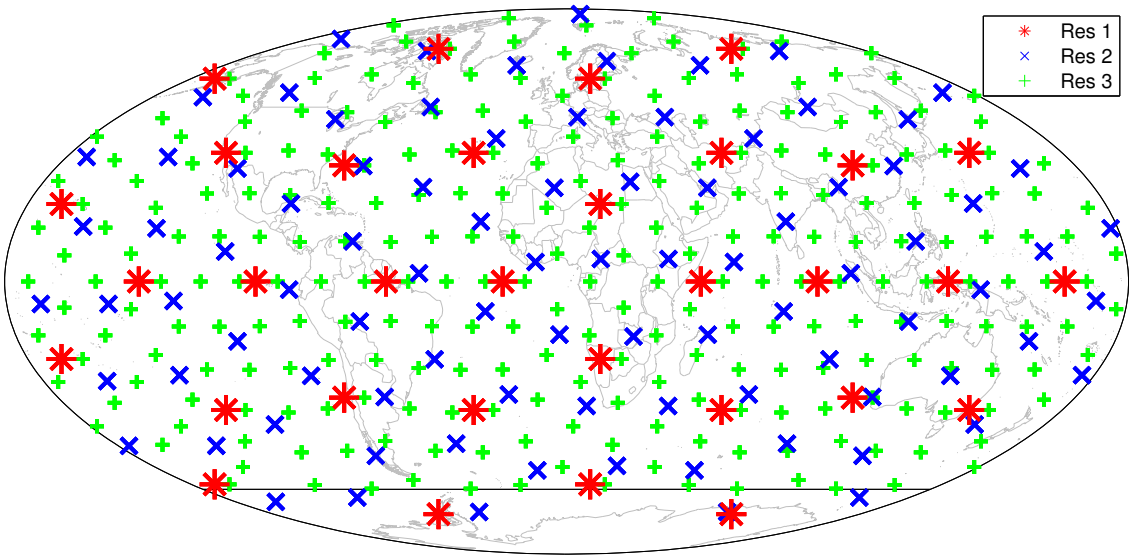

Figure 2.5: Locations of the basis function centers of all three resolutions on the globe.

finer resolutions. To off-set the centers from different resolutions, we shifted the orientation point for the second and third resolution.

Due to the lack of data south of -60° latitude, some of the basis functions near the south pole are not included in the model; we deleted two centers of the second resolution and 14 centers of the third resolution. Following the recommendation of Cressie and Johannesson (2008) to set the range w of the bisquare functions equal to 1.5 times the distance of two adjacent centers at the same resolution, we finally obtained 32 bisquare functions with range $w = 6241$km in spherical distance, 90 bisquare functions with range $w = 3491$km, and 258 bisquare functions with range $w = 2048$km, for the three resolutions, respectively. The resulting $r = 380$ basis function centers are shown in Figure 2.5.

2.5.3 Parameter Estimates and FRS Results

For the AIRS dataset, the EM algorithm is sensitive to the choice of initial values. Initializing K_0 based on no data is problematic, and hence we used the modification described around (2.23) at the end of Section 2.4.1. This allowed us instead to specify an initial value for the covariance matrix K_1. Then the spatial-only EM algorithm of Katzfuss and Cressie (2009), which is based on the FRK procedure (2.23), yielded estimates of K_1 and $\sigma^2_{\delta,1}$. We used *these* estimates as the initial values of K_1 and σ^2_δ, respectively, in the spatio-temporal EM algorithm.

The resulting EM algorithm took 23 iterations until convergence was obtained. Each iteration took about 2.5min to complete on an eight-core server, resulting in a total time for estimation of about 58min. The estimates of the matrix parameters are shown in Figure 2.6. Clearly, there is some overfitting with regard to H again, causing the estimate of U to be damped somewhat. The EM estimate for σ^2_δ is 3.5650. The estimates of the intercept and the slope of the latitudinal gradient are very stable over time, being around 375.3 and 0.05, respectively.

We obtain predictions and corresponding standard errors at all $m = 61,236$ hexagons for all 16 days, by substituting the EM estimates into the FRS equations (2.9)–(2.12), with the modification given by (2.23). The FRS predictions and standard errors for day 1 are shown in the middle and lower panel of Figure 2.4, respectively. We can see that the standard-error map is dominated by the fine-scale variance. At hexagons without data, the standard error is much higher than at hexagons with data. If the number of original retrievals that went into a particular hexagon is two or more, the standard error for that cell is even lower. Figure 2.7 shows the FRS predictions based on EM estimates of parameters, for all even-numbered days (on a scale different to that of Figure 2.4); accompanying plots

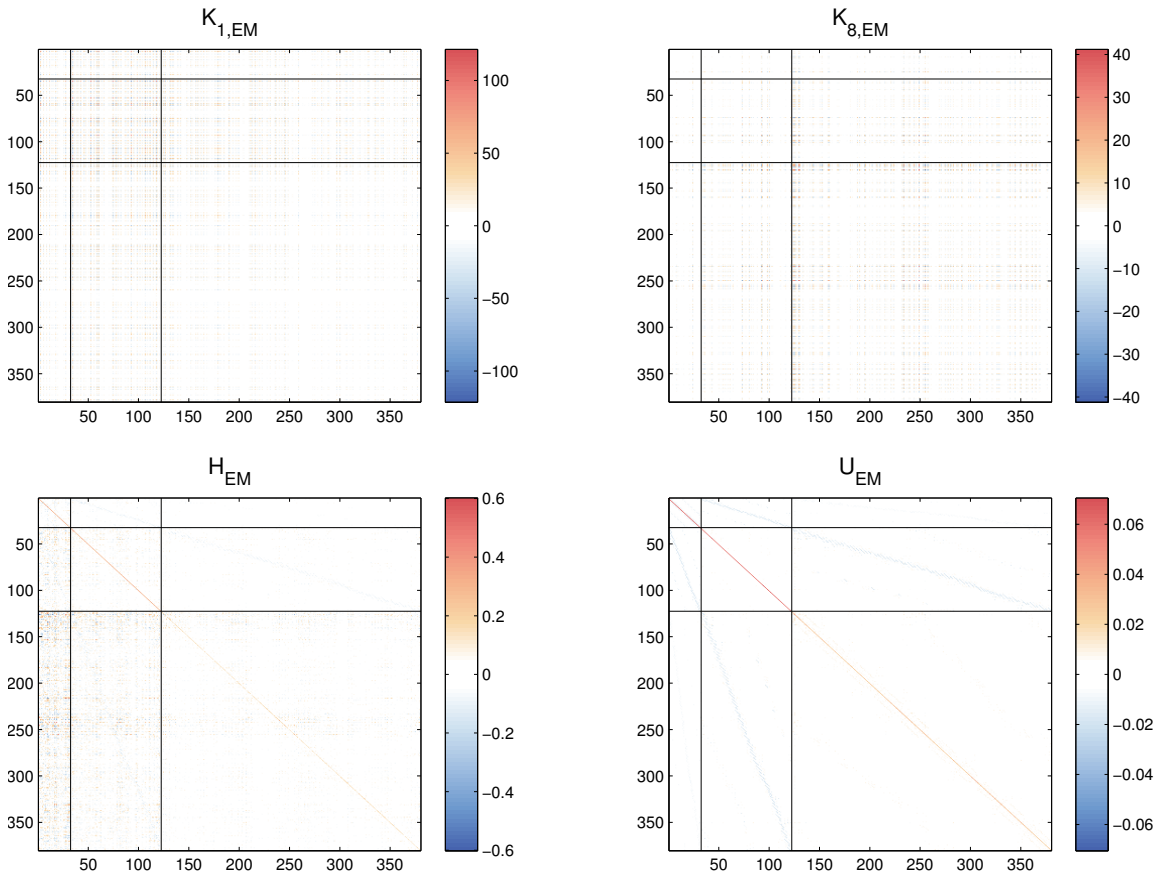

Figure 2.6: EM estimates of matrix parameters from 16 days of AIRS data. The black lines divide the matrices into parts corresponding to the three resolutions of basis functions. For example, for the plot on the bottom left, the top-right region in the plot corresponds to the elements in the estimated propagator matrix H that describe how the basis-function coefficients of the first resolution on day $t + 1$ are generated by the basis-function coefficients of the third resolution on day t.

of FRS standard errors, like that shown in Figure 2.4, are not included here but could be. From the predictions, the AIRS measurements clearly indicate that mid-tropospheric CO_2 concentrations were much higher in the Northern Hemisphere than in the Southern Hemisphere during the time period under consideration (May 1–16, 2003). Accordingly, most of the temporal evolution takes place in the Northern part of the globe.

2.6 Discussion and Conclusions

This chapter develops maximum likelihood estimation via an EM algorithm for the STME model. Once parameters are estimated, FRS can be used for spatio-temporal smoothing of the data. Due to the scalability of FRS with regard to the number of observations at each time point, this methodology is suitable even for massive data sets. An important application of the methodology is to remote sensing, where the number of observations is typically large, big gaps between satellite tracks make exploitation of spatial and temporal correlations in the underlying process imperative, and the nonstationarity of any process over the globe requires a flexible spatial covariance model that does not rely on stationarity assumptions.

One issue with our STME model is the simple structure imposed on the fine-scale variation. Its covariance matrix, $V_{\delta,t}$, is diagonal and, in our AIRS example, its diagonal elements are constant for each time point. This can lead to a map of FRS standard errors that does not account for fine-scale heterogeneity (see bottom panel of Figure 2.4). To remedy this problem, we can estimate or model the diagonal elements of $V_{\delta,t}$. For example, in Section 2.3.4, we show how spatial structure in $\delta_t(\cdot)$ could be accounted for by modeling the variance heterogeneity through $v_{\delta,t}(\cdot)$.

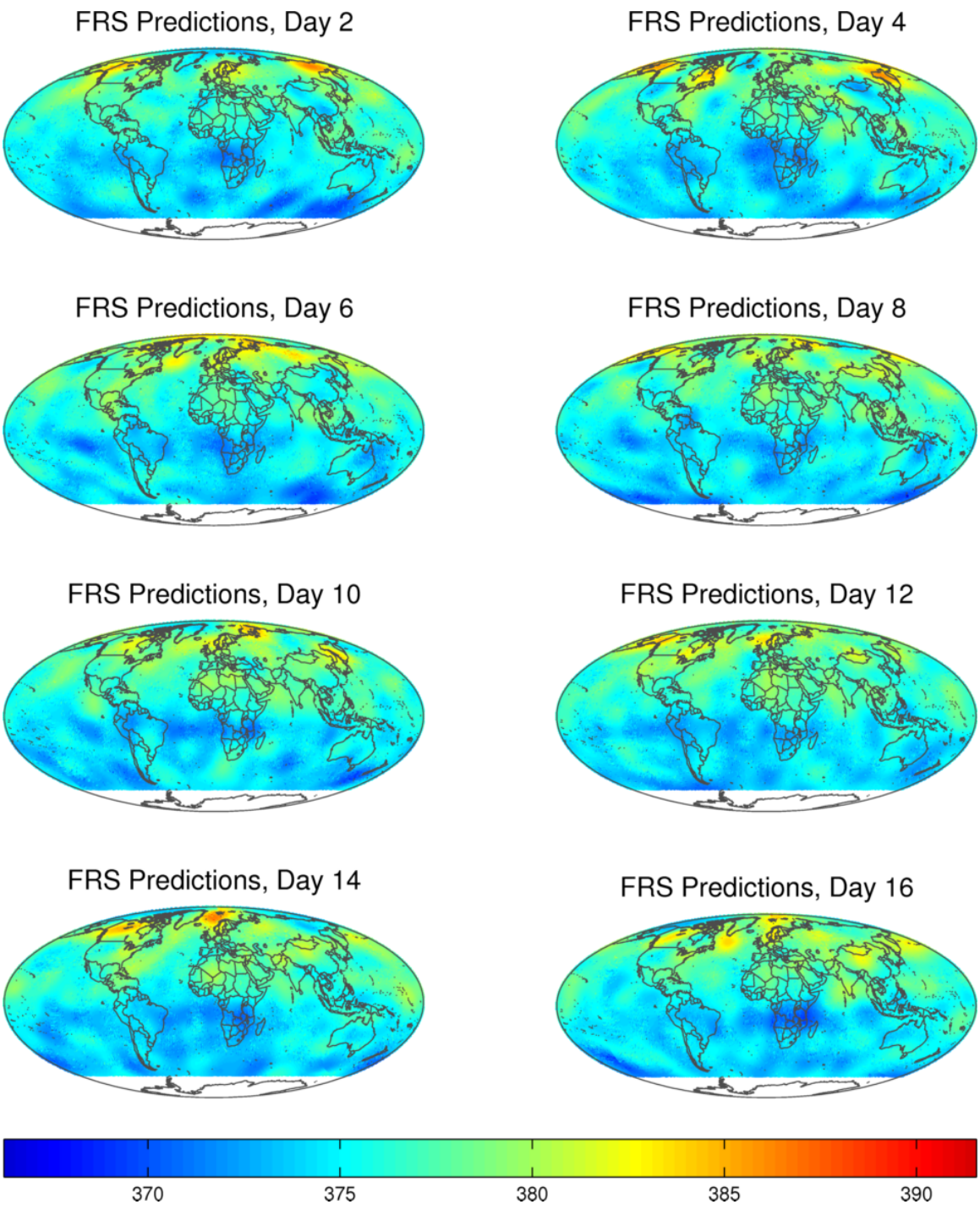

Figure 2.7: FRS predictions using EM parameter estimates of mid-tropospheric CO_2 (in ppm) from AIRS data, for eight days (here, the even days) in the study period. Units are ppm.

Currently, we make no assumptions on the structure of the matrix parameters K_0, H, and U (other than K_0 and U being positive-definite). This results in a large number of parameters that have to be estimated, which in turn can favor overfitting of the temporal evolution (as observed in Sections 2.4.2 and 2.5.3, where the assessment of uncertainty in the smoothing predictions was very liberal). This stems from both an underestimation of the variances in the covariance matrix U and from our empirical-Bayes approach, which does not account for uncertainty in the parameter estimation. The number of unknown parameters could be reduced in some cases if the temporal evolution of the process can be described through partial differential equations, in the spirit of Xu and Wikle (2007). The use of prior distributions in a fully Bayesian approach results in both regularization of the many parameters and a more correct assessment of the parameter-estimation uncertainty (see Chapter 3).

In conclusion, we have presented a coherent approach to spatio-temporal smoothing and parameter estimation for potentially massive datasets. The methodology is well suited to the analysis of remote-sensing data, crucially allowing for very fast computation.

Chapter 3: Bayesian Hierarchical Spatio-Temporal Smoothing for Very Large Datasets

3.1 Introduction

This chapter is concerned with spatio-temporal smoothing of very-large-to-massive datasets. The increase in the amount of data being collected has provided statisticians in all disciplines with the challenge of how to cope with the new wealth of data, but it is particularly challenging for spatial and spatio-temporal statistics. Spatial statistical analyses, such as kriging and maximum likelihood estimation, typically require solving systems of linear equations involving the covariance matrix of the data vector, which quickly becomes infeasible as the size of the dataset increases. For simplicity, we refer to solving these equations as "inversion" of the data vector's covariance matrix, since the computational complexity of the two operations is the same. Computational infeasibility becomes more of a problem when data are also collected over time, which makes for even larger datasets. Due to these special computational challenges for spatial and spatio-temporal statistical analyses, in this chapter we take the term "very large dataset" to mean a dataset size that is on the order of 10^4 to 10^7 observations. Massive datasets are orders of magnitude beyond this.

Here, we develop statistical inference that does not break down as the size of the (spatio-temporal) dataset increases, by working on a reduced-dimensional space. We model the

process as a random linear combination of spatial basis functions plus a spatially heterogeneous fine-scale-variation term. Instead of describing the dependence in the data using a spatio-temporal covariance function, we make use of a vector-autoregressive dynamical model for the coefficients of the basis functions. Thus, the temporal dependence in the data is explained by specifying the temporal evolution of a reduced-dimensional spatial process given its past state. This so-called spatio-temporal random effects (STRE) model contains unknown parameters.

We operate in a fully Bayesian (FB) framework, and thus we specify prior distributions for all parameters, where some are calibrated according to the variability in the data. Bayesian inference has a number of advantages: Not only does it allow for correct assessment of prediction variability, it also results in a phenomenon called shrinkage, which can be very helpful in situations with high-dimensional parameter spaces, where individual parameters are often unidentifiable without prior information. However, computational feasibility is crucial in FB inference, as inferences often rely on computationally intensive Markov chain Monte Carlo (MCMC) simulations from the posterior distribution. In this chapter, we propose a prior that induces sparsity and shrinkage on the autoregressive parameters describing the temporal evolution of the basis-function coefficients, and we describe how MCMC sampling can be achieved in a computationally feasible way.

Examples of large spatio-temporal datasets are readily available from measurements made by Earth-observing satellites. These datasets provide a distinct set of statistical challenges: Despite the large size of daily datasets, the observations can still be sparse relative to the spatial domain of interest (often the entire globe). This requires the statistical analyst to take full advantage of spatial and temporal correlations in the true process, in order to fill spatial gaps. However, no process on the globe will satisfy the stationarity assumptions

that are typically made in traditional spatial statistics. The STRE model proposed in this chapter accounts for these issues, and it can also easily handle the measurements at areal footprints and the change-of-support issues that are typical for remotely sensed data. We apply the STRE model to a dataset of global satellite CO_2 measurements; from a fixed window of daily measurements obtained from the Atmospheric InfraRed Sounder (AIRS) on the Aqua satellite, we give a sequence of complete daily maps of global CO_2 fields during this window, together with maps of their associated prediction standard errors.

An extensive review of the literature on dynamical spatio-temporal models in a hierarchical statistical framework can be found in Cressie and Wikle (2011, Chap. 7). We highlight and extend portions of the literature that are especially relevant to our work. First, the large literature on state-space modeling can be seen as a part of the hierarchical-statistical-modeling literature by noting that the measurement equation can be viewed as the data model, and the state equation can be viewed as the process model. Consequently, the STRE model referred to earlier is the state equation in a state-space model for time series (Hamilton, 1994, Chap. 13). Shumway and Stoffer (2006, Chap. 6) give an overview of various types of state-space implementations from a general time-series perspective.

An integral part of any state-space model is the observation matrix (Shumway and Stoffer, 2006, p. 325), which maps the state variables on the reduced dimension to the process or observations at the original or physical dimension. If a state-space model is applied in a spatio-temporal context, the observation matrix typically consists of known spatial basis functions (Smith et al., 1996; Kaplan et al., 1998). Here we propose methodology that allows for the use of any type of (orthogonal or non-orthogonal) spatial basis functions, and we choose bisquare functions for illustration in Sections 3.3 and 3.4. Other possible choices for the basis functions include empirical orthogonal functions (e.g., Aubry et al., 1993) and

wavelets (e.g., Nychka et al., 2002), but both of these basis-function types are most useful for gridded data. Overviews of possible sets of basis functions in spatial and spatio-temporal applications are given in Antoulas (2005) and Wikle (2010). Alternatively, the observation matrix can be obtained by discretizing a process-convolution model (Higdon, 1998). For the basis-function approach using positive integrable functions (e.g., bisquare functions found in Cressie and Johannesson, 2008), the two approaches are similar, since the basis functions can be interpreted as smoothing kernels. Instead of assuming a known observation matrix, Lopes et al. (2008) place a (strong) prior on it.

Another important component of a state-space model is the form of the temporal evolution of the state variables, for which many parameterizations are possible. We assume here that the evolution is linear and first-order Markov, and therefore the evolution is determined by a single propagator matrix. (For a more general science-based approach to the specification of the temporal evolution, see Wikle and Hooten, 2010.) This allows for Kalman-filter-type inference on the state variables (Kalman, 1960). Sparse parameterizations can be achieved by assuming that the propagator matrix is the identity (which corresponds to a random walk; see, e.g., Stroud et al., 2001) or diagonal (which corresponds to separable autoregressive models; see, e.g., Lopes et al., 2008). Less restrictive parameterizations that still depend on only a small number of parameters can be achieved by deriving the propagator matrix from a discretization of partial differential equations (e.g., Wikle, 2003; Cangelosi and Hooten, 2009; Stroud et al., 2010) or integro-difference equations (e.g., Kot et al., 1996; Wikle and Cressie, 1999; Xu et al., 2005; Dewar et al., 2009). If the dimension of the state space is sufficiently low, it is also possible to include more general lagged-nearest-neighbor models (Wikle et al., 1998), or it might even be possible to leave the propagator matrix completely general. In this chapter, we take the latter approach, albeit

on the reduced-dimensional space. Our formulation is based on Kalman smoothing, not filtering, and it is feasible even if the number of basis functions is moderately large. This is achieved by inducing strong shrinkage and sparsity through a multiresolutional prior (which results in a "soft" lagged-nearest-neighbor approach) inspired by the Minnesota prior in the time-series literature (Litterman, 1986; George et al., 2008). The Minnesota prior shrinks the autoregressive coefficients toward a random-walk model, a feature that is also present in our prior model. We also develop a fast posterior-sampling scheme based on conditional simulation, which is most commonly used in spatial statistics (e.g., Cressie, 1993, Sec. 3.6.2).

The methodology proposed in this chapter is specifically designed to scale up to very large or massive datasets. Early examples of dimension reduction using basis functions in state-space models applied to large spatio-temporal datasets can be found in Mardia et al. (1998), Wikle and Cressie (1999), and Wikle et al. (2001). Other approaches to statistical analysis of very large spatio-temporal datasets include multi-resolutional tree-structured models (Johannesson et al., 2007) and predictive-process models (discussed briefly in the spatio-temporal setting by Banerjee et al., 2008); in the latter case, the kriging equations are approximated by replacing the data locations with a smaller number of knots. In the process-convolution framework, the temporal evolution can either be modeled using a spatio-temporal smoothing kernel (e.g., Higdon, 2002) or a dynamical model for the state variables (e.g., Calder et al., 2002).

The specific dynamical spatio-temporal state-space model used in this chapter is a reduced-rank model called the STRE model (referred to earlier). This approach was proposed by Cressie et al. (2010), who were motivated by the spatial-only fixed-rank model of

Cressie and Johannesson (2006, 2008). Aside from a strong focus on computational scalability and no requirement for orthogonality of the basis functions, this framework has the added feature of incorporating a fine-scale-variation component (Wikle and Cressie, 1999; Cressie and Johannesson, 2008; Jun and Stein, 2008; Kang et al., 2009; Cressie and Kang, 2010). In this chapter, we generalize the distributional assumptions on this component to allow for spatially heterogeneous variances using a suggestion made in Section 2.3.4. In recent papers, estimation of the STRE-model parameters has relied on method-of-moments estimation (Kang et al., 2010) and expectation-maximization (EM) estimation (Chapter 2), which are not Bayesian. In the spatial-only setting, Kang and Cressie (2011) give FB inference for the spatial-random-effects model and its parameters. In the spatio-temporal setting of this chapter, we propose a multiresolutional sparsity- and shrinkage-inducing prior for the propagator matrix of the basis-function coefficients. Together with priors on the other model parameters, this allows us to carry out FB inference for the STRE model, and its parameters, in the context of spatio-temporal smoothing.

The rest of this chapter is organized as follows. Section 3.2 describes the methodology: We introduce the STRE model, explain the prior distributions assumed for the parameters, and give an overview on how to sample from the posterior distribution in a computationally efficient way. We then compare our methodology to an empirical-Bayesian, STRE-model approach that uses the EM algorithm for estimating parameters. We make the comparison in a simulation study (Section 3.3) and in an application to a dataset of global CO_2 measurements (Section 3.4). Discussion and conclusions are given in Section 3.5. Appendix A (p. 134ff) contains many details on the posterior inference and the MCMC algorithm upon which it is based.

3.2 Bayesian Spatio-Temporal Smoothing

3.2.1 The Spatio-Temporal Random-Effects Model

We are interested in a spatio-temporal process $\{Y_t(\mathbf{s}) : \mathbf{s} \in D_s, t \in 1, 2, \ldots\}$ on a continuous spatial domain D_s and at discrete time points $\{1, 2, \ldots\}$. As is often done in spatial statistics, we assume that the process $Y_t(\cdot)$ can be decomposed as follows,

$$Y_t(\mathbf{s}) = \mu_t(\mathbf{s}) + \nu_t(\mathbf{s}),\ \mathbf{s} \in D_s,\ t = 1, 2, \ldots, \tag{3.1}$$

where $\mu_t(\cdot)$ is large-scale trend, and $\nu_t(\cdot)$ accounts for spatial (and here, temporal) correlation. In what follows, we assume that $\mu_t(\cdot) := \mathbf{x}_t(\cdot)'\boldsymbol{\beta}_t$, which is a linear combination of p known covariates, $x_{t,1}(\cdot), \ldots, x_{t,p}(\cdot)$.

While our interest is in inference on $Y_t(\cdot)$, we cannot observe it perfectly. Our measurements are affected by additive measurement error and cannot be taken at every $(\mathbf{s}, t) \in D_s \times \{1, 2, \ldots\}$. Our focus in this chapter is on smoothing, namely after collecting the $n_1 + \ldots + n_T$ measurements,

$$Z_t(\mathbf{s}_{i,t}) = Y_t(\mathbf{s}_{i,t}) + \epsilon_t(\mathbf{s}_{i,t}),\ i = 1, \ldots, n_t,\ t = 1, \ldots, T,$$

we are interested in predicting the unknown quantity $Y_t(\mathbf{s})$ at every $\mathbf{s} \in D_s$ for all time points $t = 1, \ldots, T$.

We assume that the measurement-error process, $\epsilon_t(\cdot)$, is distributed as,

$$\epsilon_t(\cdot) \sim N(0, \sigma^2_{\epsilon,t} v_{\epsilon,t}(\cdot)),\ t = 1, 2, \ldots,$$

independent of $Y_t(\cdot)$, and independent in time and space. For identifiability reasons, both the measurement-error variance $\sigma^2_{\epsilon,t}$ and the function $v_{\epsilon,t}(\cdot) > 0$ will be assumed known for the remainder of this chapter. While it is common that $v_{\epsilon,t}(\cdot)$ is known, there may be

no information on $\sigma^2_{\epsilon,t}$; in this case, $\sigma^2_{\epsilon,t}$ can be estimated from the data via an estimation technique based on extrapolating the variogram (Kang et al., 2009). If prior instrument-calibration experiments have been done, $\sigma^2_{\epsilon,t}$ may in fact be known as well.

To exploit the spatio-temporal correlation in $Y_t(\cdot)$, we now specify a covariance function for the measurements and use this to form the covariance matrix, Σ, of the vector of all measurements, $\mathbf{Z}_{1:T} := [\mathbf{Z}'_1, \ldots, \mathbf{Z}'_T]'$, where $\mathbf{Z}_t := [Z_t(\mathbf{s}_{1,t}), \ldots, Z_t(\mathbf{s}_{n_t,t})]'$. Let $n_+ := \sum_{t=1}^{T} n_t$ denote the total number of observations taken at all time points combined. Now, statistical inference typically requires inversion of the $n_+ \times n_+$ matrix Σ, possibly repeatedly so at successive iterations of an estimation procedure. Since the inversion of a general $n_+ \times n_+$ matrix requires on the order of n_+^3 computations, this is infeasible for the very large spatio-temporal datasets of interest here, where $\{n_t\}$ (and possibly also T) are very large.

To achieve both computational feasibility and a flexible nonstationary model, we assume a *spatio-temporal random effects* (STRE) model (Cressie et al., 2010) for the component $\nu_t(\cdot)$ in (3.1):

$$\nu_t(\mathbf{s}) = \mathbf{b}_t(\mathbf{s})'\boldsymbol{\eta}_t + \delta_t(\mathbf{s}),\ \mathbf{s} \in D_s,\ t = 1, 2, \ldots, \tag{3.2}$$

where $\mathbf{b}_t(\cdot) := [b_{t,1}(\cdot), \ldots, b_{t,r_t}(\cdot)]'$ is an r_t-dimensional vector of known spatial basis functions; $\boldsymbol{\eta}_t$ is a random coefficient vector of length r_t; and the fine-scale variation,

$$\delta_t(\cdot) \sim N(0, \sigma^2_{\delta,t} v_{\delta,t}(\cdot)), \tag{3.3}$$

is *a priori* independent of $\{\boldsymbol{\eta}_t\}$ and independent in time and space. The basis functions in $\mathbf{b}_t(\cdot)$ do *not* have to be orthogonal. However, it is recommended that they be of different spatial resolutions $1, \ldots, C$ (Cressie and Johannesson, 2008), which can capture different scales of spatial variation. The fine-scale variation, $\{\delta_t(\cdot) :\ t = 1, 2, \ldots\}$, can be viewed

as an attempt to account for the error introduced by the dimension reduction. The temporal evolution of $\{Y_t(\cdot)\}$ is determined by a vector-autoregressive (VAR) model for $\{\boldsymbol{\eta}_t : t = 0, 1, \ldots, T\}$:

$$\boldsymbol{\eta}_t | \boldsymbol{\eta}_{t-1} \sim N_{r_t}(H_t \boldsymbol{\eta}_{t-1}, U_t),\ t = 1, \ldots, T, \tag{3.4}$$

with initial state $\boldsymbol{\eta}_0 \sim N_{r_0}(\mathbf{0}, K_0)$. The $r_t \times r_{t-1}$ matrix H_t and the $r_t \times r_t$ matrix U_t will be referred to as the propagator matrix and the innovation covariance matrix, respectively.

While models (3.2) and (3.4) have been introduced with a lot of generality, in this chapter we assume henceforth that $\mathbf{b}_t(\cdot) \equiv \mathbf{b}(\cdot)$, $r_t \equiv r$, $\sigma^2_{\delta,t} \equiv \sigma^2_\delta$, $v_{\delta,t}(\cdot) \equiv v_\delta(\cdot)$, $H_t \equiv H$, and $U_t \equiv U$, during the period $\{0, \ldots, T\}$. Strictly speaking, this is not needed in a FB framework; however, assumptions of this sort allow practical identifiability and result in well mixed MCMC samples from the posterior distribution.

As the number of basis functions, r, is much smaller than the sample sizes $\{n_t\}$, assumption (3.2) results in dimension reduction, since the computational complexity for processing the measurements taken at time point t is reduced to $\mathcal{O}(n_t r^3)$ from $\mathcal{O}(n_t^3)$; see Cressie et al. (2010). Additionally, the VAR model (3.4) is a state-space model that allows for sequential (in time) processing of data observed at subsequent time points via Kalman-filter- and Kalman-smoother-type algorithms. This ensures that the computational cost of inference on the process components $\{\boldsymbol{\eta}_t\}$ and $\{\delta_t(\mathbf{s})\}$ (given the unknown parameters) for all observed time points $t = 1, \ldots, T$ is linear in n_+; that is, inference for our model can scale up to very-large-to-massive datasets.

In summary, we have introduced the *data model*,

$$\mathbf{Z}_t = \mathbf{Y}_t + \boldsymbol{\epsilon}_t,\ t = 1, \ldots, T, \tag{3.5}$$

where, based on the STRE model, we can write,

$$\mathbf{Y}_t = X_t\boldsymbol{\beta}_t + B_t\boldsymbol{\eta}_t + \boldsymbol{\delta}_t,\ t = 1, \ldots, T. \tag{3.6}$$

Together, (3.4) and (3.6) specify a *process model* for $\{\mathbf{Y}_t\}$. In (3.5) and (3.6), we have stacked scalars into vectors and row vectors into matrices such that, for example, the i-th row of the $n_t \times r$ matrix B_t is given by $\mathbf{b}(\mathbf{s}_{i,t})'$. The corresponding covariance matrices are $K_t := var(\boldsymbol{\eta}_t)$, $t = 1, \ldots, T$, and

$$D_t := var(\boldsymbol{\delta}_t + \boldsymbol{\epsilon}_t) = \sigma_\delta^2 V_{\delta,t} + \sigma_{\epsilon,t}^2 V_{\epsilon,t},\ t = 1, \ldots, T,$$

which is a diagonal matrix with $V_{\delta,t} := diag(v_\delta(\mathbf{s}_{1,t}), \ldots, v_\delta(\mathbf{s}_{n_t,t}))$ and $V_{\epsilon,t} := diag(v_\epsilon(\mathbf{s}_{1,t}), \ldots, v_\epsilon(\mathbf{s}_{n_t,t}))$.

The process model is not fully specified yet. We often want to predict $Y_t(\mathbf{s})$ at a set of spatial locations that is different from the measurement locations. Here we assume (without loss of generality) that the set of prediction locations at time t is a *superset* of the measurement locations observed at time t, so that we can write $\mathbf{Y}_t = M_t\mathbf{Y}_t^P$, $t = 1, \ldots, T$, where $\mathbf{Y}_t^P$ is the process vector of all m_t prediction locations for time t. This is achieved by allowing the observation locations to be included in the set of all prediction locations. We assume that there are no duplicate measurements, and hence M_t is an $n_t \times m_t$ incidence matrix of mostly 0s and a 1 in each row. In practice, the set of prediction locations has often been a fine grid over the spatial domain of interest, resulting in $m_t \equiv m$, where the data locations are moved to their nearest respective grid cells.

We use a superscript P (as in "prediction") when a vector or matrix has been obtained by evaluating appropriate processes at all prediction locations, so that the process model is,

$$\mathbf{Y}_t^P = X_t^P\boldsymbol{\beta}_t + B_t^P\boldsymbol{\eta}_t + \boldsymbol{\delta}_t^P,\ t = 1, \ldots, T, \tag{3.7}$$

where $\{\boldsymbol{\eta}_t\}$ satisfies (3.4). This implies that $X_t = M_t X_t^P$, $B_t = M_t B_t^P$, and $\boldsymbol{\delta}_t = M_t \boldsymbol{\delta}_t^P$. The diagonal matrix $var(\boldsymbol{\delta}_t^P) =: \sigma_\delta^2 V_{\delta,t}^P$ is an $m_t \times m_t$ matrix, where $V_{\delta,t}^P$ will be modeled below.

3.2.2 Prior Distributions

Until now, we have implicitly assumed a *known* vector of process-model parameters $\boldsymbol{\theta}_P$, which contains the trend coefficients $\{\boldsymbol{\beta}_t : t = 1, \ldots, T\}$, the fine-scale-variation variance σ_δ^2, the function $v_\delta(\cdot)$, and the elements defining the matrices that describe the VAR process, K_0, H, and U. Of course, $\boldsymbol{\theta}_P$ will rarely be known in practice. We could take an empirical-Bayesian approach to inference, in which we estimate the parameters either via a method-of-moments technique (Wikle and Cressie, 1999; Kang et al., 2010) or via the EM algorithm (Xu and Wikle, 2007; Fassò and Cameletti, 2009b; Chapter 2). Instead, in this chapter, we take a Bayesian approach and specify prior distributions for all unknown parameters (e.g., Wikle et al., 1998). This results in a *parameter model* (usually called a prior), which, together with the data model (3.5) and the process model (3.7) given earlier, leads to posterior inference via Bayes' Theorem. Recall that our goal is smoothing; inference is implemented using Markov chain Monte Carlo (MCMC) simulations described in Section 3.2.4 and Appendix A.

All parameters in $\boldsymbol{\theta}_P$ are assumed to be *a priori* independent, unless specifically stated otherwise. For the parameters $\{\boldsymbol{\beta}_t\}$ and σ_δ^2, we assume (almost) noninformative priors (see Appendix A for details). The prior distributions for the covariance matrices K_0 and U are each taken to be a multiresolutional Givens-angle prior (Kang and Cressie, 2011). As this prior distribution has been considered in detail in previous work, we only give a brief review in Appendix A.

The function $v_\delta(\cdot)$ determines the form of the heterogeneity of the fine-scale-variation variance in (3.3), namely $var(\delta(\cdot)|\sigma^2_\delta, v_\delta(\cdot)) = \sigma^2_\delta v_\delta(\cdot)$. Following a suggestion made in Chapter 2, we assume a stochastic volatility model of the form,

$$v_\delta(\cdot) := \exp\{\mathbf{b}_\delta(\cdot)'\boldsymbol{\eta}_\delta\}, \tag{3.8}$$

where $\mathbf{b}_\delta(\cdot)$ is a known vector of r_δ basis functions and, for example, could be a sub-vector of $\mathbf{b}(\cdot)$. The prior distribution on $v_\delta(\cdot)$ is induced by $\boldsymbol{\eta}_\delta \sim N_{r_\delta}(\mathbf{0}, \sigma^2_{\eta_\delta} I_{r_\delta})$, where $\sigma^2_{\eta_\delta}$ is a *known* hyperparameter. This model allows for flexible estimation of the heterogeneity (in space), in that $v_\delta(\mathbf{s})$ can multiplicatively modify the overall fine-scale-variation variance, σ^2_δ, at any location $\mathbf{s} \in D_s$. The exponential function in (3.8) ensures that the resulting variance of $\delta(\cdot)$ is positive, and the prior mean, $E(\boldsymbol{\eta}_\delta) = \mathbf{0}$, approximately induces shrinkage of the resulting variance of $\delta(\cdot)$ toward the overall variance parameter, σ^2_δ, at any point in space. By modeling the function (on the log-scale) as a linear combination of basis functions, we ensure fast computation even when the function has to be evaluated at a large number of observed or prediction locations. The hyperparameter $\sigma^2_{\eta_\delta}$ can be chosen in accordance with prior beliefs on how different the fine-scale variation is expected to be in different parts of the spatial domain of interest. Consider the variance ratio $R := var(\delta(\mathbf{s}_1))/var(\delta(\mathbf{s}_2)) = v_\delta(\mathbf{s}_1)/v_\delta(\mathbf{s}_2)$, where $\mathbf{s}_1$ and $\mathbf{s}_2$ are chosen to be locations at the centers of two distant (normalized) basis functions, so that $(\mathbf{b}_\delta(\mathbf{s}_1) - \mathbf{b}_\delta(\mathbf{s}_2))'(\mathbf{b}_\delta(\mathbf{s}_1) - \mathbf{b}_\delta(\mathbf{s}_2)) \approx 2$. This implies that, approximately, $\exp\{R\} \sim N(0, 2\sigma^2_{\eta_\delta})$. When $1/2$ and 2 are chosen as the lower and upper endpoints, respectively, of a 95% credible interval for R, this results in a value of $\sigma^2_{\eta_\delta} \approx 0.25^2$ for the hyperparameter.

3.2.3 The Prior on the Propagator Matrix H

Let us now turn to the prior assumptions for the propagator matrix H. We first develop a two-stage prior that ensures that the full-conditional distribution of $\mathbf{h} := vec(H')$ is available in closed form (see Appendix A), and then we give some insight into the ramifications of this prior specification. To motivate our prior on H, note that the impact of H on the temporal evolution of the process is plainly obvious from,

$$E(\eta_{t,i}|\mathbf{h}_i, \boldsymbol{\eta}_{t-1}) = \mathbf{h}_i'\boldsymbol{\eta}_{t-1} = \textstyle\sum_{j=1}^{r} h_{ij}\eta_{t-1,j},\ t = 1, 2, \ldots,$$

where $\mathbf{h}_i'$ denotes the i-th row of H. Thus, h_{ij} describes the autoregressive effect of $\eta_{t-1,j}$ on $\eta_{t,i}$; intuitively, we want this effect to diminish as the (physical) distance between the j-th basis function and the i-th basis function increases. This intuition is complicated by the fact that we want the basis functions in $\mathbf{b}(\cdot)$ to be made up of C (say) resolutions. We can write H (after appropriate ordering) as a block matrix,

$$H =: \left[\begin{array}{c|c|c} H_{11} & \cdots & H_{1C} \\ \hline \vdots & \ddots & \vdots \\ \hline H_{C1} & \cdots & H_{CC} \end{array}\right], \tag{3.9}$$

where the block H_{kl} contains the elements of H that describe how the basis-function coefficients of resolution k at time point t are affected by the basis-function coefficients of resolution l at the previous time point $t-1$.

In light of this role that the elements of H play on the temporal evolution of the process, we assume that the (i, j)-th element of H has the (conditional) prior distribution,

$$h_{ij}|\boldsymbol{\theta}_H \overset{ind.}{\sim} N(\mu_{c_i} I(i = j), \tau^2_{c_i,c_j} g(d_{ij}; \alpha_{c_i,c_j}, \gamma_{c_i,c_j})^2),\ i = 1, \ldots, r,\ j = 1, \ldots, r, \tag{3.10}$$

where c_i denotes the resolution to which the i-th basis function belongs; the quantity $d_{ij} \in [0, 1]$, with $\max\{d_{ij}\} = 1$, is the normalized distance between the centers of the i-th and

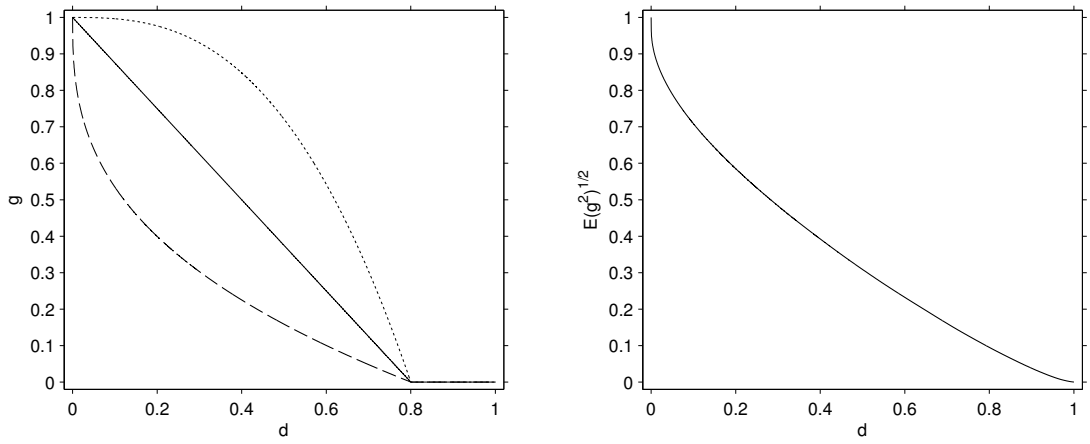

Figure 3.1: Left panel: The $g(\,\cdot\,;\alpha,\gamma)$ function for $\alpha = 0.8$ and $\gamma = 0$ (solid line), $\gamma = -1$ (dashed), and $\gamma = 1$ (dotted). Right Panel: The function $\sqrt{E(g(\,\cdot\,;\alpha_{kl},\gamma_{kl})^2)}$ describes the shrinkage (on the standard-deviation scale) induced by the prior on H as a function of the basis-function distance; see (3.14).

the j-th basis functions; $\boldsymbol{\theta}_H$ is a vector of parameters describing the prior distribution of H that consists of $\{\mu_k\colon\ k = 1,\ldots,C\}$, $\{\tau^2_{kl}\colon\ k,l = 1,\ldots,C\}$, $\{\alpha_{kl}\colon\ k,l = 1,\ldots,C\}$, and $\{\gamma_{kl}\colon\ k,l = 1,\ldots,C\}$; and

$$g(d;\alpha,\gamma) := \begin{cases} 1-(d/\alpha)^{\exp(\gamma)}, & d \le \alpha \\ 0, & d > \alpha \end{cases} \tag{3.11}$$

is a function of normalized distance with (random) range parameter $\alpha \in [0,1]$ and (random) shape parameter $\gamma \in \mathbb{R}$ (see the left panel of Figure 3.1; more details are given below in this subsection). Note that, to include the case $\alpha = 0$, we define $0/0 = 0$ in (3.11).

At the second level of the prior distribution on H, we assume that all parameters in $\boldsymbol{\theta}_H$ are independently distributed according to,

$$\begin{aligned}
\mu_k &\overset{ind}{\sim} N(1, \sigma^2_{\mu,k}),\ k = 1, \ldots, C, \\
\tau^2_{kl} &\overset{ind}{\sim} IG(a_{\tau,kl}, b_{\tau,kl}),\ k, l = 1, \ldots, C, \\
\alpha_{kl} &\overset{iid}{\sim} U(0, 1),\ k, l = 1, \ldots, C, \\
\gamma_{kl} &\overset{iid}{\sim} N(\mu_\gamma, \sigma^2_\gamma),\ k, l = 1, \ldots, C,
\end{aligned} \tag{3.12}$$

where are all parameters specifying these distributions are fixed, as follows: First, the choice of $E(\mu_k) = 1$ is based on our desire to center the noninformative prior of H at the identity matrix, which is a random-walk model. Second, the parameters $\{\gamma_{kl}\}$ determine the shape of (3.11) on the interval $(0, \alpha)$. A natural centering for these parameters is $\mu_\gamma = 0$, as $g(\,\cdot\,; \alpha, \gamma = 0)$ is a straight line from the point $(0, 1)$ to $(\alpha, 0)$. To find a good value for σ^2_γ, consider that square-root distances, absolute distances, and squared distances are often used in practice. To ensure that these values are contained in an *a priori* 95% credible interval for the exponent of the distance d in the function (3.11), we set $\sigma^2_\gamma = 0.5^2$, so that the endpoints of the credible interval are approximately given by $1/e = 0.37$ and $e = 2.72$ (see the dashed and dotted lines in the left panel of Figure 3.1). Finally, the remaining parameters $\{\sigma^2_{\mu,k}\}$, $\{a_{\tau,kl}\}$, and $\{b_{\tau,kl}\}$ are calibrated to the data through an initial point estimate of H, such as the EM estimate (see Appendix A for details).

We shall now interpret the prior assumptions on H made above, and discuss their ramifications. As noted earlier, the prior distribution for H implies that the temporal evolution of $\{\boldsymbol{\eta}_t\}$ is *a priori* centered on the random walk, $E(H) = E(E(H|\{\mu_k\})) = I_r$, due to the assumption $E(\mu_k) = 1$, $k = 1, \ldots, C$, in (3.12). In light of this, it should be noted that our prior for H is a noninformative prior. It only uses information on where the basis functions are located in the spatial domain, D_s, and to which resolutions the basis functions

belong. The prior is ideally suited for applications in which no prior information about the temporal evolution of the process is available, or in situations where it is the goal to validate or check existing scientific models about the temporal evolution of the process based on a complete map of predictions based on data. Our prior mean meets the constant-mean and mass-balance requirements for propagator matrices formulated in Gelpke and Künsch (2001).

Integrating over the parameter $\{\tau_{kl}\}$, we have,

$$h_{ij}|\mu_{c_i}, \alpha_{c_i,c_j}, \gamma_{c_i,c_j} \sim t_{2a_{\tau,c_ic_j}}\left(\mu_{c_i} I(i=j), \frac{b_{\tau,c_ic_j}}{a_{\tau,c_ic_j}-1} g(d_{ij};\alpha_{c_i,c_j},\gamma_{c_i,c_j})^2\right), \qquad (3.13)$$

where $t_\nu(\mu, \sigma^2)$ denotes a generalized t-distribution with ν degrees of freedom, location parameter μ, and scale parameter σ (Bishop, 2006, Sec. 2.3.7).

The off-diagonal elements of H are shrunk towards zero (or even set equal to zero), depending on the distance and the resolutions of the corresponding basis functions. The marginal variance of h_{ij} is,

$$var(h_{ij}) = varE(h_{ij}|\boldsymbol{\theta}_H) + E\, var(h_{ij}|\boldsymbol{\theta}_H) = \sigma^2_{\mu,c_i} I(i=j) + \frac{b_{\tau,c_ic_j}}{a_{\tau,c_ic_j}-1} E\big(g(d_{ij};\alpha_{c_i,c_j},\gamma_{c_i,c_j})^2\big), \qquad (3.14)$$

where the square root of the expectation on the right-hand side (i.e., on the standard-deviation scale) is shown in the right panel of Figure 3.1. Thus, the prior variance of h_{ij} is monotone decreasing as a function of d_{ij}.

The prior distributions on the parameters, $\{\alpha_{kl}\}$ and $\{\gamma_{kl}\}$, can be interpreted as controlling the sparsity and the shrinkage on the elements of H, respectively. The range parameters $\{\alpha_{kl}\}$ induce sparsity in H in that, assuming a uniform prior on α_{c_i,c_j},

$$P(h_{ij}=0) = P(g(d_{ij};\alpha_{c_i,c_j},\gamma_{c_i,c_j})=0) = P(\alpha_{c_i,c_j} \le d_{ij}) = d_{ij}.$$

Hence, it becomes more and more likely that h_{ij} is zero with increasing distance between the centers of basis functions i and j. That is, we are essentially specifying the (random) dimension of the parameter space by generating $\{\alpha_{kl}\}$. Other choices than a uniform distribution for the priors of $\{\alpha_{kl}\}$ are possible and can result in interesting dynamical structure for the model. For example, if the priors on $\{\alpha_{kl}\}$ are all point masses at zero, then $g(d; 0, \gamma) = I(d = 0)$, which results in a diagonal H with parameter μ_k down the diagonal of H_{kk}, $k = 1, \ldots, C$. Even this simple multiresolutional H induces complex spatio-temporal dependence, since $cov(\boldsymbol{\eta}_{t+1}, \boldsymbol{\eta}_t) = HK_t$ has non-trivial, nonstationary cross-dependence.

Given $\{\tau_{kl}\}$ and $\{\alpha_{kl}\}$, the parameters $\{\gamma_{kl}\}$ control the amount of shrinkage of the nonzero elements of H as a function of the basis-function distances. If the parameter γ_{kl} is nonnegative then, conditional on α_{kl}, the function $g_{kl}(\cdot)$ is concave on the interval $(0, \gamma_{kl})$; for nonpositive γ_{kl}, the function is convex on the interval (see the left panel of Figure 3.1). Therefore, very large values for $\{\gamma_{kl}\}$ make for little shrinkage (up to distances smaller than $\{\alpha_{kl}\}$).

Lastly, the marginal covariance between two elements, h_{i_1,j_1} and h_{i_2,j_2}, of H, after integrating out the prior distributions on $\boldsymbol{\theta}_H$ given by (3.12), is,

$$cov(h_{i_1,j_1}, h_{i_2,j_2}) = \{var(h_{i_1,j_1} + h_{i_2,j_2}) - var(h_{i_1,j_1}) - var(h_{i_2,j_2})\}/2$$

$$= \sigma^2_{\mu,k} I(i_1 = j_1) I(i_2 = j_2) I(c_{i_1} = c_{i_2} = k).$$

This means that the only *a priori* nonzero correlations between elements of H are those where both are a diagonal element within the same resolution. However, as mentioned above, all elements within each block H_{kl} are *a priori* statistically dependent, for $k, l = 1, \ldots, C$.

Note that our prior for H makes use of distances between basis functions. This distance is quite intuitive if the basis functions have a clear "center" (e.g., bisquare functions). For

other basis functions, one could use the "center of energy" (e.g., Wickerhauser, 1994, p. 164), defined as $\int \mathbf{s}\, b(\mathbf{s})^2 d\mathbf{s} / \int b(\mathbf{s})^2 d\mathbf{s}$ for a continuous basis function $b(\cdot)$. However, this center of energy might not be easily interpretable for basis functions with non-compact support (e.g., Fourier functions or empirical orthogonal functions), and so our prior for H might not be generally applicable for those functions (unless $\alpha_{kl} \equiv 0$ for all $k, l = 1, \ldots, C$, in which case recall that $0/0 = 0$ in (3.11)).

Since H refers to a reduced-dimensional space, it might seem unnecessary to look for sparsity in the $r \times r$ matrix H. However, the number of basis functions, r, can be moderately large, and it is usually larger than T (e.g., in Section 3.4, we have $r = 380$ and $T = 16$). This may result in practical nonidentifiability, for which regularization (here, sparsity and shrinkage) would be needed (see Appendix A for more details).

3.2.4 MCMC Inference

For a set of generic vectors $\{\mathbf{x}_t\}$, define $\mathbf{x}_{t_1:t_2} := [\mathbf{x}'_{t_1}, \ldots, \mathbf{x}'_{t_2}]'$. Recall that our goal in this spatio-temporal context is smoothing, not filtering. After having observed data $\mathbf{Z}_{1:T} = \mathbf{z}_{1:T}$, FB inference is based on the posterior distribution of $\mathbf{Y}^P_{1:T}$ (i.e., that of $\boldsymbol{\eta}_{1:T}$ and $\boldsymbol{\delta}^P_{1:T}$) and the unknown parameters, given the data. Unfortunately, this posterior distribution is not available in closed form. Instead, we sample from the posterior distribution via Markov chain Monte Carlo (MCMC) simulation. As the methodology developed in this chapter is intended to be used on very large (or even massive) datasets, computational feasibility and speed are of great concern. We employ a Gibbs sampler (Geman and Geman, 1984) with some Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) updates where necessary. In this section, we give an overview of the techniques used to sample the unknowns in the MCMC; details are given in Appendix A.

First, consider the basis-function coefficients $\boldsymbol{\eta}_{0:T}$. Due to their strong temporal dependence, it is not advised to update each $\boldsymbol{\eta}_t$ individually, which would result in slow convergence of the MCMC. Instead, we update the entire vector $\boldsymbol{\eta}_{0:T}$ at once, using a technique called the forward-filtering, backward-sampling algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994). The number of operations required for this is linear in each n_t for each iteration of the MCMC, which is essential to scalability of the algorithm as a whole.

Updating the propagator matrix H (or, equivalently, $\mathbf{h} := vec(H')$) and its random hyperparameters poses its own challenges. As $\mathbf{h}$ is an r^2-dimensional vector, direct sampling becomes impossible when r, the number of basis functions, is moderately large. To get around this, we employ a technique similar to conditional simulation used in geostatistics (for details on spatial conditional simulation, see, e.g., Cressie, 1993, Sec. 3.6.2). Considering the hyperparameters $\{\alpha_{kl}\}$, which control the sparsity of H, we notice that there is actually a change of dimension in the parameter space of $\mathbf{h}$. Depending on the value of $\{\alpha_{kl}\}$ sampled in the MCMC algorithm, a number of elements of $\mathbf{h}$ will have a variance of zero, and their full conditional distributions will be point masses at zero. If we marginalize over $\mathbf{h}$ when updating $\{\alpha_{kl}\}$ and use the conditional-simulation technique for sampling $\mathbf{h}$ mentioned earlier, we avoid having to use an explicit reversible-jump MCMC; see Appendix A.

We recommend updating all parameters with analytically intractable conditional distributions ($\boldsymbol{\eta}_\delta$, $\{\alpha_{kl}\}$, and the parameters in K_0 and U) using the adaptive Metropolis-Hastings algorithm of Haario et al. (2001); see Appendix A for details. The full conditional distributions of $\{\boldsymbol{\beta}_t\}$ and σ_δ^2 are also given in Appendix A.

Let $\boldsymbol{\theta}$ be a vector containing all unknowns, $\boldsymbol{\eta}_{0:T}$, $\boldsymbol{\delta}_{1:T}^{P}$, $\boldsymbol{\theta}_P$, and $\boldsymbol{\theta}_H$. Samples from the posterior distribution of $\boldsymbol{\theta}$ given the data are obtained as follows: We begin the MCMC sampler with some starting value $\boldsymbol{\theta}^{[0]}$, and then we obtain $\boldsymbol{\theta}^{[l]}$, $l = 1, 2, \ldots$, by updating each component of $\boldsymbol{\theta}$ given the most recent value of all other components as described in Appendix A. After L_b iterations, the algorithm should be sampling from the target (joint posterior) distribution. From a total number of L_a iterations, the first L_b are discarded, and we consider the set $\{\boldsymbol{\theta}^{[L_b+1]}, \ldots, \boldsymbol{\theta}^{[L_a]}\}$ to be a sample from the joint posterior distribution of all unknowns given the data.

To return to the issue of scalability of the algorithm for very large datasets, we note that the number of computations required at each iteration of the MCMC is linear in each n_t. However, each update of H requires inversion of a sparse $rT \times rT$ matrix with at most $rT(r + T - 1)$ nonzero elements (see Appendix A). This implies that if the number of basis functions, r, and the number of time points, T, are both very large, the algorithm can become fairly slow.

3.3 Simulation Study: FB-FRS vs. EM-FRS

Instead of specifying prior distributions for all parameters in the model described in Section 3.2.1, we could pursue empirical-Bayesian inference via Fixed Rank Smoothing (FRS), as described in Cressie et al. (2010). To do this, we must first estimate the parameters, and then we obtain the posterior distribution of $\mathbf{Y}_{1:T}^{P}$ given the data, by assuming that all parameters are known and fixed at their estimated values. We can estimate the parameters in the STRE model using maximum-likelihood estimation via the EM algorithm, which is shown in Chapter 2 to be preferable to the binned method of moments when the data are Gaussian. This EM-FRS procedure is therefore a natural candidate for comparison to the

fully Bayesian inference (FB-FRS) proposed in Section 3.2. In this section, we carry out a simulation study to assess parameter estimation, accuracy of predictions, and the accuracy of inferred prediction uncertainties.

3.3.1 Simulation Setup

The simulated data are meant to be a simplistic version of satellite data. The spatial domain is one-dimensional, $D_s := \{1, \ldots, 256\}$, and there are $T = 16$ time points. The "satellite" has a repeat cycle of two time units. The two tracks of the satellite have a width of 64: For t odd, the tracks are $\{1, \ldots, 64\}$ and $\{129, \ldots, 192\}$; for t even, the tracks are $\{65, \ldots, 128\}$ and $\{193, \ldots, 256\}$. To simulate non-retrievals due to cloud cover and other problems, 50% of the values within each track at each time point are declared missing at random. This results in $n_t = 64$ observations at each time point.

The basis functions we use are bisquare functions,

$$f_{bi}(\mathbf{s}) := \{1 - (\|\mathbf{s} - \mathbf{c}\|/w)^2\}^2 I(\|\mathbf{s} - \mathbf{c}\| < w), \tag{3.15}$$

where $\mathbf{c}$ is the center point, $w > 0$ is the specified range, and $I(\cdot)$ is an indicator function. In this simulation study, we have $r = 5$ bisquare basis functions from $C = 2$ resolutions, as depicted in Figure 3.2. The one basis function of the first resolution has a range of $w = 144$ and is centered at 128. The four basis functions of the second resolution have a range of $w = 38$ and are centered at 32.5, 96.5, 160.5, and 224.5, respectively.

With the exception of the ranges and center points of the basis functions, this setup with $C = 2$ resolutions is exactly the same as the one used in the simulation study given in Section 2.4. The parameters used in the simulation are also calibrated in the same way as in that chapter. The true matrix parameters are calibrated to match as closely as possible (as measured by the Frobenius norm) a stationary exponential spatial covariance of the form

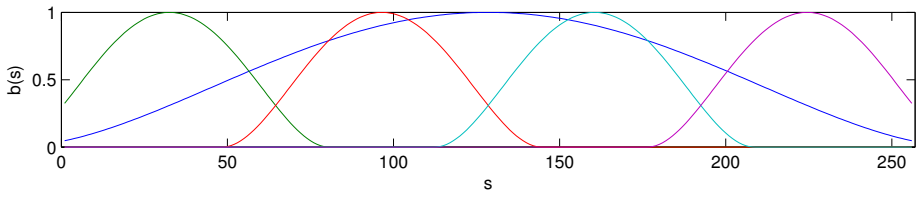

Figure 3.2: The five basis functions ($r = 5$) of two resolutions ($C = 2$) used in the simulation study.

$cov(\nu_t(i), \nu_t(j)) = \exp(-|i-j|/25)$ and a lag-1 temporal correlation of 0.8 (see Cressie et al., 2010, Sect. 3, for more details). The fine-scale-variation proportion is set to 0.05, which results in $\sigma^2_\delta = 0.0297$, held constant over time. The function $v_\delta(\cdot) = \exp\{\mathbf{b}_\delta(\cdot)'\boldsymbol{\eta}_\delta\}$ is determined by two bisquare functions of range $w = 96$, centered at 64 and 192, and the true $\boldsymbol{\eta}_\delta$ is simulated at each simulation iteration from a $N_2(\mathbf{0}, 0.25^2 I_2)$ distribution (as suggested in Section 3.2.2). For the EM algorithm, the parameter vector $\boldsymbol{\eta}_\delta$ was "estimated" by finding the mode of its posterior distribution. Finding a maximum a posteriori estimator for one set of parameters and maximum likelihood estimators (MLEs) for all others could be considered inappropriate; alternatively, one could consider $\boldsymbol{\eta}_\delta$ as a fixed parameter in the EM-FRS procedure and estimate it in the EM algorithm.

The measurement-error variance is also constant over time. It is determined by the signal-to-noise ratio (SNR; defined as in Section 3.1 of Cressie et al., 2010, but here we temporarily assume that $\boldsymbol{\eta}_\delta \equiv \mathbf{0}$), for which we have chosen two levels: SNR=2, resulting in $\sigma^2_\epsilon = 0.2968$, and SNR=5, resulting in $\sigma^2_\epsilon = 0.1187$. The variance σ^2_ϵ is assumed known for both the FB and the EM-FRS procedures. Finally, a constant mean of $\mu = 5$ is chosen (i.e., $x_t(\mathbf{s}) \equiv 1$ and $\beta_t \equiv \mu$). An example of the data simulated from the STRE model

(Section 3.2.1) is shown in Figure 3.3. (We only show the first four time points; the setup for the remaining time points is analogous.)

3.3.2 Simulation Results

Using this setup, we generate 1000 data sets for both levels of the SNR. For each dataset, we obtain posterior samples from our MCMC algorithm, and we calculate the posterior means and the posterior 2.5- and 97.5-percentiles (based on the prior distributions given in Sections 3.2.2, 3.2.3, and Appendix A). In addition, we obtain FRS predictions and standard errors based on *EM parameter estimation* and, as a reference, we also obtain FRS predictions and prediction standard errors using the *true parameters* $\boldsymbol{\theta}$. We use the true parameter values to initialize the EM algorithm and to calibrate the priors for the FB procedure. We save both Bayesian posterior samples of all parameters and EM parameter estimates for each dataset. Figure 3.3 shows the predictions and credible/prediction intervals for all three procedures for (the first four time points of) one simulated dataset, to illustrate inference on the process $\{Y_t(\cdot)\colon\ t = 1, \ldots, T\}$.

We summarize the results in Table 3.1. Generally, all summaries of the results are computed over all 1000 simulated datasets. However, the EM algorithm failed to converge for 12 of the datasets for SNR=2 (see "Success rate"), and so the results from these datasets were excluded from the analysis. The first mean squared prediction error (MSPE) is taken over all 256 spatial locations at all T=16 time points. The summaries denoted "on track" and "off track" are only taken over the spatial locations for each time point that were considered on or off track, respectively, as described in the previous subsection. The interval score (IS) is defined as (Gneiting and Raftery, 2007, Sect. 6.2),

$$\text{IS}_\alpha(l, u; y) = (u - l) + 2\{(l - y)_+ + (y - u)_+\}/\alpha,$$

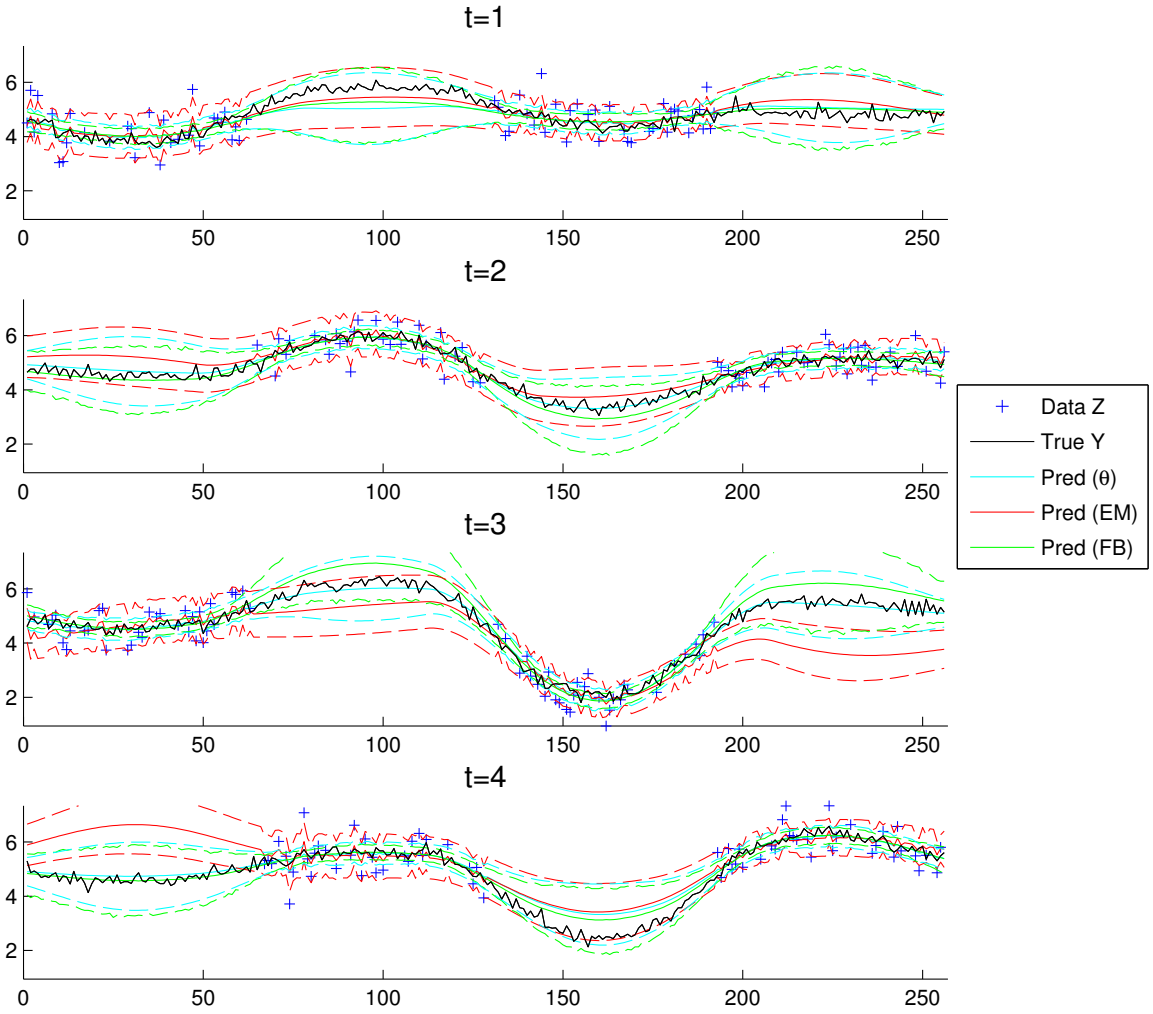

Figure 3.3: One realization of the data (blue crosses) observed at the first four time points in the simulation study for SNR=2. Also shown are FRS predictions using the true parameter values (light blue), FRS predictions using the EM parameter estimates (red), and Bayesian posterior means (green); dashed lines are the respective 95% credible/prediction intervals. The true process values are shown in black.

Table 3.1: Results for $\{\hat{Y}_t(\cdot)\}$ compared to $\{Y_t(\cdot)\}$, obtained from the simulation experiment, with the following acronyms: MSPE = (Empirical) Mean Squared Prediction Error, IS = Interval Score (see text), CIW = Credible/Prediction Interval Width (nominal 95% intervals), CIC = Credible/Prediction Interval Coverage (target is 95%)

	SNR = 5				SNR = 2			
	True θ	EM	FB	EM/FB	True θ	EM	FB	EM/FB
Success rate	—	1.000	1.000	—	—	0.988	1.000	—
MSPE	0.118	0.196	0.148	1.325	0.142	0.243	0.197	1.232
MSPE - on track	0.034	0.046	0.037	1.244	0.044	0.063	0.049	1.281
MSPE - off track	0.201	0.347	0.260	1.337	0.240	0.422	0.344	1.226
IS - on track	0.863	1.247	0.916	1.362	0.978	1.658	1.097	1.512
IS - off track	2.026	3.636	2.384	1.525	2.217	4.698	2.761	1.702
CIW - on track	0.718	1.086	0.736	1.476	0.819	1.367	0.858	1.592
CIW - off track	1.700	1.587	1.917	0.828	1.859	1.744	2.230	0.782
CIC - on track (t=8, s=96)	0.946	0.985	0.947	—	0.942	0.974	0.925	—
CIC - off track (t=2,s=32)	0.949	0.761	0.921	—	0.958	0.715	0.937	—

where l and u are, respectively, the lower and upper endpoints of a $(1-\alpha)$ confidence interval (we use $\alpha = 0.05$), y is the true value, and $(x)_+ := xI(x > 0)$. This scoring rule combines the width of the confidence interval with a penalty for not containing the true value.

We can see from Table 3.1 that the posterior mean from our FB-FRS procedure is a considerably better predictor than the FRS predictions based on EM estimates, both on and off track. At least for SNR=5, the MSPE of the posterior mean is fairly close to the MSPE of the FRS procedure using the true parameter values (i.e., for "perfect" parameter estimation). The difference between FB-FRS and EM-FRS is even greater when we consider the prediction-uncertainty assessment. Possibly due to an overestimation of σ_δ^2, the confidence intervals for EM-FRS are too wide on track, but too narrow off track (see also Figure 3.3). The IS for FB-FRS is much closer to the IS for FRS using the true parameters than it is to

Table 3.2: Mean squared estimation errors for the scalar parameters (Bayes estimates are posterior means).

	SNR = 5			SNR = 2		
	EM	FB	EM/FB	EM	FB	EM/FB
Success rate	1.000	1.000	—	0.988	1.000	—
μ_t	0.230	0.067	3.443	0.213	0.116	1.836
σ_δ^2 ($\times 100$)	0.486	0.009	55.166	1.577	0.031	51.466
$\eta_{\delta,1}$	0.059	0.058	1.002	0.074	0.065	1.140
$\eta_{\delta,2}$	0.057	0.055	1.036	0.077	0.071	1.081

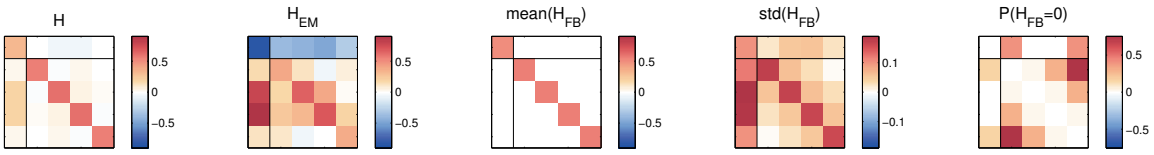

Figure 3.4: Propagator matrices. The left-hand panel shows the true H. All values in the other panels are elementwise medians over the 1000 simulations from the simulation study (SNR=5). Shown are the EM estimates, the posterior means, the posterior standard deviations, and the posterior probabilities of the elements being zero. The black lines divide H as in (3.9).

the IS for EM-FRS. From Table 3.2, the FB posterior means of the parameters also result in smaller mean squared estimation errors (MSEEs) than the EM estimates.

Finally, we show the inference on the propagator matrix H in Figure 3.4 by taking elementwise medians of the estimates and posterior summaries based on each of the 1000 simulated datasets for SNR=5. Clearly, the lack of regularization in the EM estimates leads to a complete mis-estimation of the first row of the matrix H. Since the parameter estimates are simply plugged into the FRS equations, this mis-estimation is not accounted for in the FRS-prediction uncertainties. While it seems from the FB posterior mean of H that the shrinkage induced by the prior might be too strong, we can see that the estimated posterior

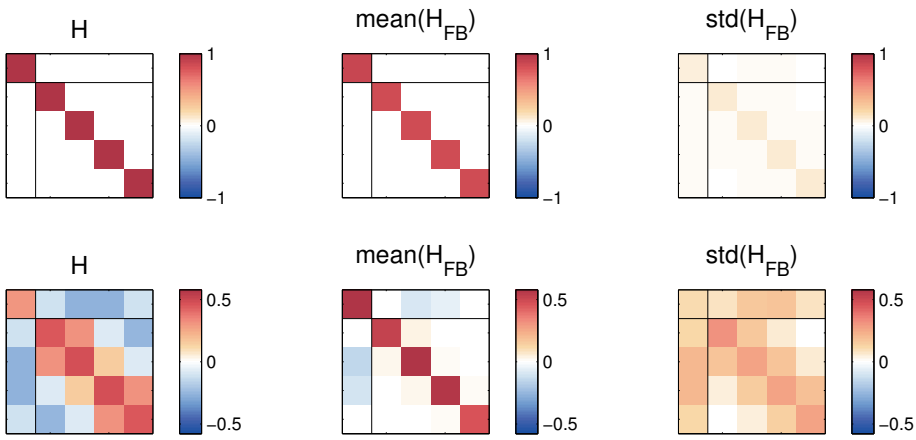

Figure 3.5: Top row: true $H = 0.8I_r$. Bottom row: true $H = P_K diag(.05, .08, .10, .94, .97)P_K$. Shown are the true H (first column), and the element-wise medians over the 1000 simulations from the two additional simulation studies (SNR=5): posterior means (middle column) and posterior standard deviations (right column). The black lines divide H as in (3.9).

standard deviations actually reflect the magnitude of the true off-diagonal elements (left panel) quite well. The estimated probability of each element being equal to zero is also in agreement with how close to zero the true values are.

To see how well a sparse H and a full H can be recovered by our model, we carried out two more simulations. The setup was kept the same as before (with SNR=5), except that now we specified the true H to be $H = 0.8I_r$ and $H = P_K diag(.05, .08, .10, .94, .97)P_K$, respectively, where P_K is the eigenvector matrix of the true K (which was again calibrated against an exponential covariance model). The results, shown in Figure 3.5, indicate that our model adapts well to very sparse or full propagator matrices.

3.4 Analysis of Global CO_2 Data

This section contains an application of our proposed fully Bayesian STRE methodology to a very large real-world dataset of global CO_2 measurements. We obtain the posterior

distribution of all parameters and the spatio-temporal process of interest, and we compare results to those from empirical-Bayesian STRE methodology based on the EM algorithm.

3.4.1 Spatio-Temporal Data: Mid-Tropospheric CO_2 Measurements from AIRS

The spatio-temporal dataset under consideration consists of 16 days of measurements of global mid-tropospheric CO_2, which were recorded by the Atmospheric InfraRed Sounder (AIRS) on board NASA's Aqua satellite (Chahine et al., 2006). The dataset is available from `http://airs.jpl.nasa.gov/AIRS_CO2_Data/`, and it is the same as that analyzed in Katzfuss and Cressie (2011b). Only CO_2 measurements between -60° and 90° latitude are available, since corresponding data at latitudes south of -60° have not been released by AIRS yet. The unit of measurement is parts per million (ppm). The measurements are taken at roughly 1:30pm local time, and we considered here those for May 1 through May 16, 2003, which from now on are referred to as days 1 through 16, respectively.

We have gridded the data onto a very fine grid, as in Section 2.5.1, which allows us to compare the results described in that paper. However, we would like to emphasize that neither methodology requires gridded data. The hexagonal grid (ISEA Aperture 3 Hexagons at resolution 8) of size $m_t \equiv 61,236$ was obtained using DGGRID software (Sahr, 2003). On each day, roughly 12,000, or 20%, of the grid cells contained data; orbit geometry, cloud cover, and retrieval convergence criteria caused the remaining grid cells to contain no data. If a particular grid cell contained more than one of the original measurements on a particular day, the data value at that grid cell was taken to be the average of those measurements and the measurement-error covariance matrix was modified correspondingly, so that $v_{\epsilon,t}(S_{i,t}) = 1/N_t(S_{i,t})$, where $N_t(S_{i,t})$ is the number of measurements contained in grid

cell $S_{i,t}$ at time t. For illustration, the gridded data of day 16 are shown in the top panel of Figure 3.6.

The measurement-error variances, $\{\sigma^2_{\epsilon,t}\}$, are assumed known in our model (Section 3.2.1); in reality, these variances were estimated prior to the actual analysis using the variogram-extrapolation technique described in Kang et al. (2009) and adapted here in Chapter 2. There we obtained a pooled estimate, $\hat{\sigma}^2_\epsilon = 5.6062\,\text{ppm}^2$, for all days $t = 1, \ldots, 16$.

The large-scale spatial trend was assumed to be determined by an intercept and a latitudinal gradient; that is, we set $\mathbf{x}_t(\cdot) = [1 \;\; lat(\cdot)]'$, independent of t.

Our model in Section 3.2.1 is described for measurements made at a point level, but our gridded data has areal (hexagonal) support. This change-of-support problem can be handled quite easily in the STRE model, by replacing the quantities in (3.5)–(3.7) by averages over the respective grid cells; more details can be found in Section 2.5.1.

For the basis functions, we used $r = 380$ bisquare functions defined by (3.15), from three resolutions. The set of functions was identical to that used in Section 2.5.2, where the reader can also find a brief discussion of how basis functions can be chosen.

We need to ensure that the distance measure d in (3.11) is normalized so that $\max\{d_{ij}\} = 1$. Here, for our basis functions located on the globe, we normalized the spherical distances between each pair of basis-function centers by dividing them by $\pi \cdot$ earth's radius $\approx \pi \cdot 6371km$, which is the maximum great-arc distance that two points on the globe can be apart.

The as-of-yet unspecified values of the hyperparameters in the priors on $\{\boldsymbol{\beta}_t\}$, σ^2_δ, $\boldsymbol{\theta}_H$, K_0, and U were calibrated using the EM estimates of the respective parameters as described in Appendix A.

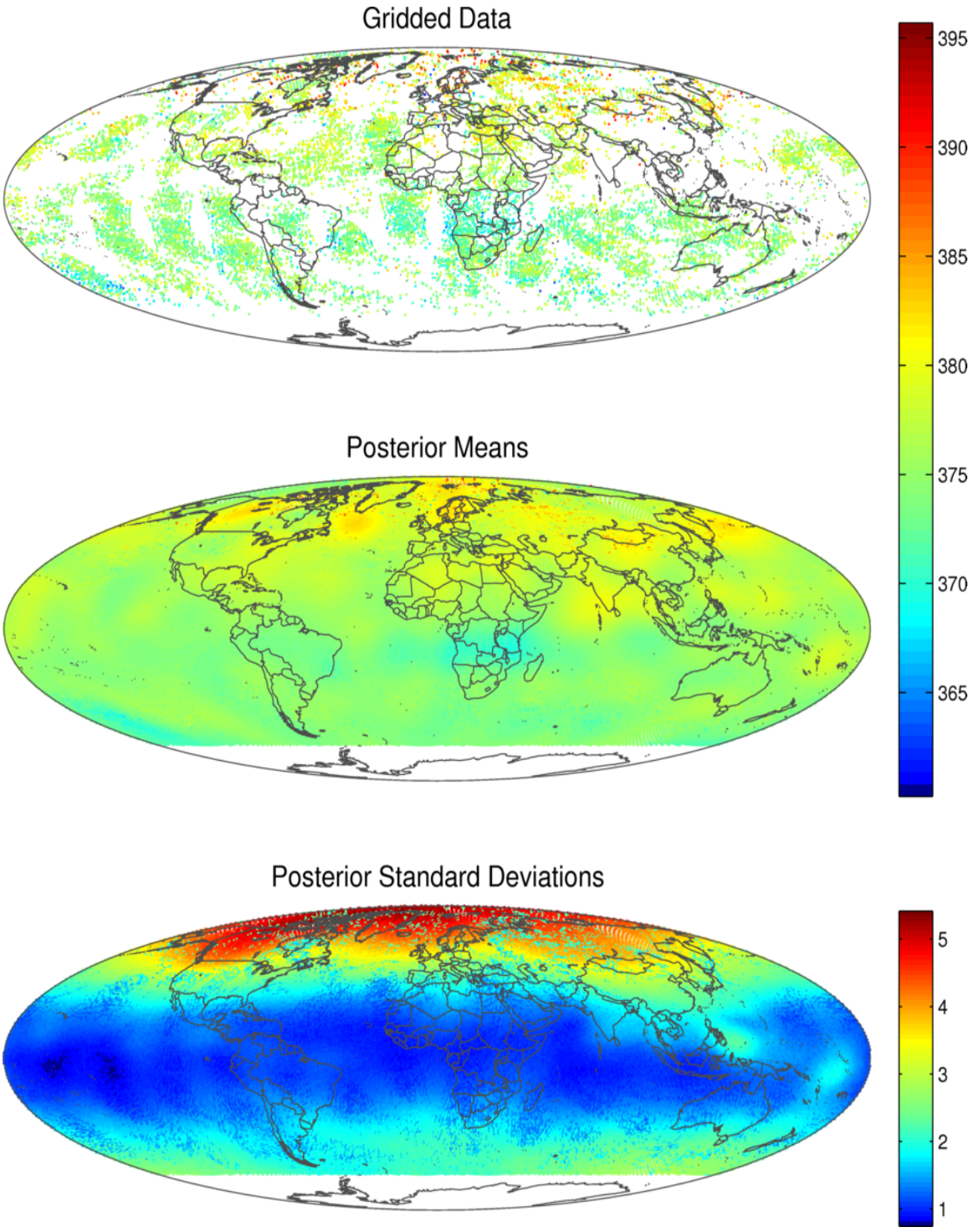

Figure 3.6: Gridded AIRS measurements of mid-tropospheric CO_2 on May 16, 2003 (top), and posterior means (middle) and posterior standard deviations (bottom) of $\{Y_{16}(\mathbf{s}) : \mathbf{s} \in D_s\}$. Units are ppm.

3.4.2 Posterior Results

We ran an MCMC for 8000 iterations using Matlab, where one iteration of the MCMC took about 30 seconds to compute on an eight-core machine (Intel Xeon X5560, with 94.5 GB RAM). Thus, the 16 days worth of data can be processed in less than three days. Trace plots show that convergence to stationarity had been reached after roughly 1500 iterations, so that we considered the first 2000 iterations as burn-in. We computed the posterior distribution of $\{Y_t(\cdot)\}$ at all $t = 1, \ldots, 16$ time points and all $m = 61,236$ hexagons. The posterior means and standard deviations for $t = 16$ are shown in Figure 3.6. The posterior standard deviations are lowest around the equator, where the process seems to be the smoothest. Notice the higher standard deviations over Southeast Asia, which reflect a lack of data in that region. In the northern part of the globe, many unusually high values in the data resulted in a large estimate of the fine-scale variation in that area. The uncertainty is also relatively large around -60° latitude, which is likely caused by the lack of data south of that latitude.

To evaluate the prediction performance of our FB-FRS procedure and of the EM-FRS approach, 500 grid cells containing observations at time point t=10 were randomly selected into a test set $\mathcal{S}_{\text{test}}$. These data were unavailable for the model fitting and were used to measure out-of-sample-prediction accuracy. As the true process $\{Y_t(\cdot)\}$ was not available in this example, we used the measurements $\{Z_t(S) : \ S \in \mathcal{S}_{\text{test}}\}$ directly as a reference, and we evaluated our predictions using an average-squared-distance criterion, ASD $:=$ $\sum_{S \in \mathcal{S}_{\text{test}}} (\hat{Y}_{10}(S) - Z_{10}(S))^2/500$. For the EM-FRS procedure, we obtained $\text{ASD}_{EM} = 9.1011$, and for the FB-FRS we obtained $\text{ASD}_{FB} = 8.7879$ when using posterior means as predictors. Thus, the FB-FRS approach had a small advantage. As a baseline predictor,

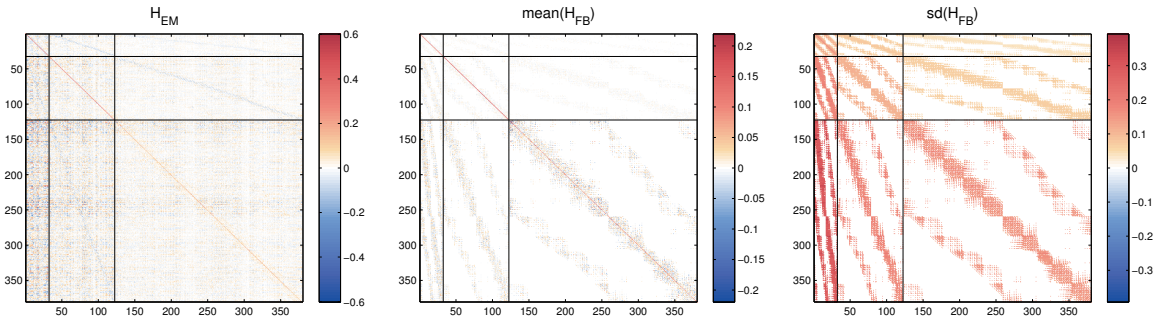

Figure 3.7: EM estimate, and mean and standard deviation of the posterior distribution of H from the AIRS data. The black lines divide H as in (3.9).

we simply calculated, for a $1^\circ \times 1^\circ$ grid, the binned means of the data from all 16 days, resulting in $\text{ASD}_{BM} = 12.0924$.

Figure 3.7 shows the EM estimate and summaries of the posterior distribution of the propagator matrix H. The EM estimate exhibits very little structure, other than a strong (positive) diagonal and two lines of negative elements, where each element corresponds to two basis functions of two different resolutions that are close in space. The FB posterior is much more structured, and we can see the sparsity induced by the prior distribution by examining the elements with zero standard deviation in the panel on the right. The eigenvalues of the posterior mean are all smaller than one, indicating a non-explosiveness of the process.

In Figure 3.8, we tie the inference on the parameters on the reduced-dimensional space back to the covariance structure of the data. For a reference point on the globe (0° longitude, 30° latitude, on day $t = 1$), we show the directional root-semivariograms for four spatio-temporal directions: for increasing longitude, increasing latitude, increasing time (days), and increasing longitude and time. For each of the four directions, the empirical root-semivariogram is compared to the theoretical quantities using plug-in EM estimates

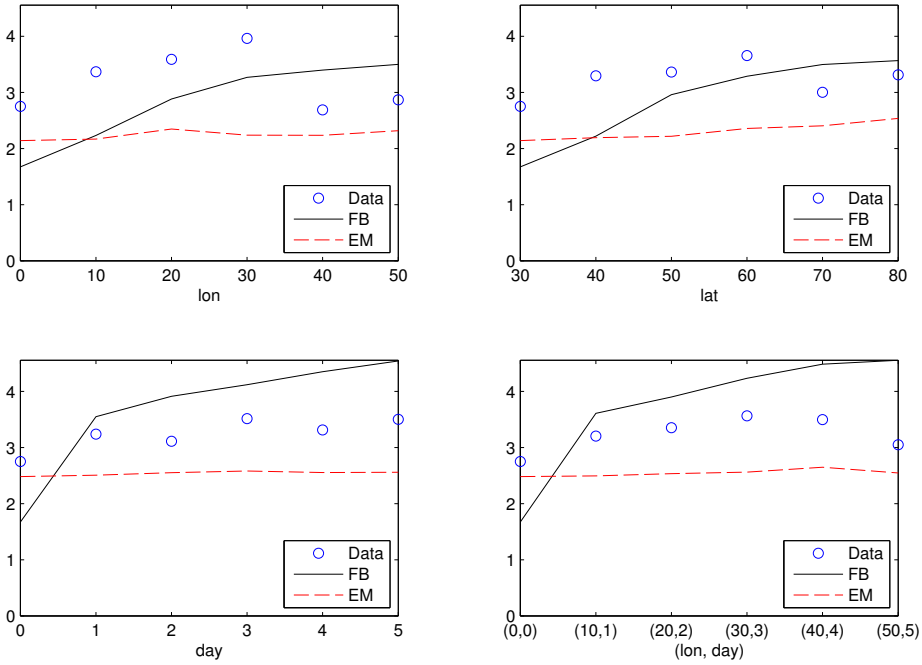

Figure 3.8: Directional root-semivariograms for the AIRS data at reference point 0° longitude, 30° latitude, on day $t = 1$. The spatio-temporal directions are longitude (top left), latitude (top right), time in days (bottom left), and longitude and time (bottom right). Shown are the empirical root-semivariograms (circles), together with the theoretical quantities using the EM estimates (dotted line) and FB inference (solid line), as estimated from the data.

and FB inference (posterior medians of root-semivariograms), respectively. The fit of the theoretical to the empirical quantities is not exact, likely due to the fact that K_0, H, and U were held constant over time and hence had to account for the spatial and temporal covariance structure on all days (i.e., not only for the reference time point $t = 1$ shown here).

3.5 Discussion and Conclusions

In this chapter, we have presented a fully Bayesian, hierarchical approach to spatio-temporal smoothing of very large datasets. By projecting the spatio-temporal process of interest onto a low-dimensional space spanned by basis functions, the STRE model makes Bayesian model fitting feasible, even when the number of observations is large. The temporal evolution of the process on the lower-dimensional space is governed by two covariance matrices and a propagator matrix. Apart from positive-definiteness of the covariance matrices, we do not require any restrictions (e.g., diagonal matrices) on these potentially large matrix parameters, but instead we specify prior distributions to achieve regularization and identifiability. In Appendix A, we give detailed instructions for posterior inference, and we provide computational speed-ups to achieve feasible computation times.

Another key ingredient of our approach is that we do not ignore the error introduced by the dimension reduction. Instead, we attempt to separate the discrepancy between observations and the reduced-dimension process into two types of error, one due to the measurement process, and one due to the dimension reduction. Both types of error are allowed to vary with space (and we could in principle let both vary with time also). Estimating the spatial heterogeneity of the variance of the fine-scale variation is important, as can be seen from the example using AIRS mid-tropospheric CO_2 in Section 3.4. From Figure 3.6,

there is an indication that there are different degrees of smoothness in different parts of the globe; see, in particular, the heterogeneity of variances at different latitudes in the bottom panel. As the basis-function space is limited in the amount of roughness it can exhibit, the variation in the excess roughness should be reflected in spatially heterogeneous fine-scale variation over the globe.

In the two comparisons of the FB framework to the empirical-Bayes EM-FRS procedure presented in Sections 3.3 and 3.4, the FB approach results in better predictions than the EM-FRS approach, particularly in the simulation experiment where the true process was available for comparison. The true worth of the FB inference is apparent from its more accurate assessment of prediction uncertainty in the simulation. Here, FB outperforms EM-FRS by up to 70% (in terms of the interval score; see Table 3.1). In Section 3.4, we expect that the relative prediction accuracy of the FB procedure would be even larger if the test set consisted of a contiguous region of the globe so that there would be no data nearby.

Our current prior on H (Section 3.2.3) allows for sparsity and shrinkage on each element h_{ij} as a function of the distance of the two basis functions i and j, but it assumes prior independence conditional on the parameters in $\boldsymbol{\theta}_H$. While inference as described in Appendix A should work even when the prior on H specifies conditional dependence (within rows of H), we encountered difficulties with our MCMC in that case.

It would be of interest to generalize further the joint-distributional assumptions on the fine-scale variation. In this chapter, we have assumed spatial independence and allowed the variance to vary spatially, but we could also allow for short-range spatial dependence. One would need to balance the requirement of rapid inversion of the data's covariance matrix (e.g., by using a strong taper) with a realistic covariance structure. Assuming spatial independence, the maps of predictions and standard errors look "spiky" in locations where data

were observed; short-range dependence would mitigate against this. Chapter 4 explores some of these suggestions.

Future work could also include further optimization of the computer code. The current code already exploits some computational speed-ups (as described in Appendix A), and parallelization is employed where possible to allow for efficient use of a multi-core computer. However, faster computation could be achieved by implementing the MCMC in a compiled language like C++, instead of Matlab.

Finally, an elegant solution to spatio-temporal *filtering* would be of interest; we have presented spatio-temporal smoothing here. If parameters are fixed across time (as they are in this chapter), their posterior distributions need to be updated when a new set of data becomes available. This can quickly become infeasible as one goes forward in time. Using sequential Monte Carlo methods (e.g., Doucet et al., 2001) and/or letting parameters such as H_t and U_t vary with t might provide a solution.

Chapter 4: Bayesian Nonstationary Spatial Modeling for Very Large Datasets

4.1 Introduction

With the proliferation of satellite measurements of environmental variables on a global scale, a need has arisen for statistical methods suitable for the analysis of *large* spatial datasets observed on *large* spatial domains. Statistical analyses of such datasets provide two main challenges: First, traditional spatial-statistical techniques are often unable to handle large numbers of observations (more than 10,000 or so) in a computationally feasible way. This is especially true for Bayesian models, for which posterior inference often requires computations to be carried out at each of many iterations in an MCMC sampler. The second challenge is that for large spatial domains (such as the entire globe), it is often not appropriate to assume that a process of interest is stationary over the entire domain. Processes often exhibit different scales of dependence, and while stationary correlation models (e.g., the widely used Matérn model; see Stein, 1999, p. 12) are often adequate approximations to the local behavior of a process, the long-range dependence often does not follow any simple parametric form. For example, the Matérn model does not allow negative dependence, and due to its exponential decay it is generally unable to describe long-range dependence (Stein, 2005). Appropriate characterization of long-range dependence is of concern, because even massive datasets are often sparse with respect to the large domains

of interest here, and so one often has to resort to observations that are relatively far away when predicting the process at certain data-poor areas of the domain.

The first problem of computational feasibility has been addressed (for nongridded data) mainly from two angles: An approach termed covariance tapering (Furrer et al., 2006; Kaufman et al., 2008; Shaby and Ruppert, 2011) relies on compactly supported correlation functions (e.g., Gneiting, 2002) to produce sparse covariance matrices, in which only a small number of elements are nonzero. While the number of computations required for finding the Cholesky decomposition of an $n \times n$ matrix and solving a system of linear equations involving that matrix is generally of order n^3, the number of computations required for these tasks may be as low as order n if the matrix is sparse (see Furrer et al., 2006, and Section 4.4.5). This may allow scalability of a covariance-tapering approach, even for very large datasets with 100,000 or more observations. However, by definition, covariance tapering does not allow long-range dependence to be modeled. In addition, Furrer et al. (2006), Kaufman et al. (2008), and Shaby and Ruppert (2011) do not use correlation models that are flexible enough to deal with the anticipated nonstationarity of processes viewed on a global scale. A recent overview of covariance tapering and other geostatistical models for large datasets can be found in Sun et al. (2011).

A second approach to achieving computational feasibility for large spatial datasets is through low-rank models. These models include a component that can be written as a linear combination of spatial basis functions (hereafter referred to as an SBF component),

$$\nu(\cdot) = \sum_{j=1}^{r} b_j(\cdot)\eta_j = \mathbf{b}(\cdot)'\boldsymbol{\eta}, \tag{4.1}$$

where $\boldsymbol{\eta}|\mathbf{K} \sim N_r(\mathbf{0}, \mathbf{K})$, and the number of basis functions, r, is much smaller than the number of observations, n. Many models that include an SBF component have been

proposed (for a recent overview, see Wikle, 2010). The models differ in how the functions in $\mathbf{b}(\cdot)$ and the covariance matrix $\mathbf{K}$ are parameterized and estimated; for Bayesian approaches, different prior distributions for $\mathbf{K}$ are possible. For discretized convolution models (i.e., convolution models whose integrals are discretized; see, e.g., Higdon, 1998; Calder, 2007; Lemos and Sansó, 2009), $\mathbf{b}(\cdot)$ contains the convolution kernels, and $\mathbf{K}$ is often assumed to be a multiple of the identity. Other authors view $\mathbf{b}(\cdot)$ as a vector of fixed basis functions. Examples of such functions include empirical orthogonal functions (e.g. Mardia et al., 1998; Wikle and Cressie, 1999), equatorial normal modes (e.g., Wikle et al., 2001), Fourier basis functions (e.g., Xu et al., 2005), W-wavelets (e.g., Shi and Cressie, 2007; Cressie et al., 2010; Kang and Cressie, 2011), and bisquare functions (e.g., Cressie and Johannesson, 2008; Chapters 2 and 3). If the basis functions used are not orthogonal, there is no obvious reason why the covariance matrix $\mathbf{K}$ should be assumed diagonal (Cressie and Johannesson, 2008). Kang and Cressie (2011) propose a prior distribution for non-orthogonal basis functions and nondiagonal $\mathbf{K}$. The prior takes into account that their wavelet basis functions are grouped into different spatial resolutions. Another way of specifying $\mathbf{b}(\cdot)$ and $\mathbf{K}$ is through the so-called predictive process (Banerjee et al., 2008; Finley et al., 2009; Banerjee et al., 2010). Here, both $\mathbf{b}(\cdot)$ and $\mathbf{K}$ are chosen to approximate a true "parent process," for which a parametric correlation model is chosen. To our knowledge, the effects of this approximation have not been fully investigated.

SBF models (4.1) allow for fast computation via the Sherman-Morrison-Woodbury formula (Sherman and Morrison, 1950; Woodbury, 1950; Henderson and Searle, 1981), as is made clear in Cressie and Johannesson (2006) and Shi and Cressie (2007). For general $\mathbf{K}$, they are also very flexible in that the covariance of an SBF component, namely $\mathbf{b}(\cdot)'\mathbf{K}\mathbf{b}(\cdot)$,

is not of traditional parametric form (such as, e.g., the Matérn covariance). The SBF covariance structure can, for example, easily adapt to processes with negative correlations. This flexibility, together with the fast computability, makes SBF components very well suited to modeling medium-range to long-range spatial dependence.

By itself, (4.1) is typically too smooth to capture fine spatial scales of variability. A fine-scale-variation (FSV) component, $\delta(\cdot)$, added to $\nu(\cdot)$ in (4.1), results in a model that Cressie and Johannesson (2008) call the spatial random effects (SRE) model, namely $\mathbf{b}(\cdot)'\boldsymbol{\eta} + \delta(\cdot)$.

In this chapter, our focus is on predicting the true process at observed and unobserved spatial locations from a large number of incomplete observations made with measurement error. A secondary interest is in a correct characterization of the covariance structure of the process of interest (but not necessarily in estimation of individual parameters in the model). We address the two challenges mentioned in the first paragraph (fast computation, together with flexible, nonstationary modeling of covariance structure) with a fully Bayesian model that combines the SBF component (4.1) (that allows for flexible modeling of medium-to-long-range dependence via a set of spatial basis functions) with an FSV component (that allows for modeling of short-range dependence). The resulting process that combines these two components is a more general SRE model than hitherto presented, and it can capture highly nonstationary spatial dependence at all scales.

One contribution of this chapter is the inclusion of a more general FSV component that is spatially dependent. Due to the dimension reduction inherent in (4.1), SBF components are typically not able to model "rough" (i.e., non-smooth) short-range dependence by themselves (see, e.g., Stein, 2008; Finley et al., 2009). Some efforts have been made to address this problem (going back to Wikle and Cressie, 1999), but many of them only attempt to remedy the underestimation of the variance (i.e., they do not correct for oversmoothing

of short-range covariance structure) by including a spatially uncorrelated component FSV component in the model (e.g., Cressie and Johannesson, 2008; Kang et al., 2009; Finley et al., 2009). The combination of an SBF component and a spatially dependent FSV component has been considered previously by Wikle and Cressie (1999), Berliner et al. (2000), Wikle et al. (2001), Stein (2008), and Kang et al. (2010), in specific circumstances.

In this chapter, we propose to use a tapered covariance for the FSV component because of its computational advantages, and we generalize existing tapering approaches (see above) by allowing the underlying covariance function to be nonstationary. To facilitate modeling on the entire globe, we extend a general class of nonstationary covariance functions for $\mathbb{R}^d$ (Stein, 2005; Paciorek and Schervish, 2006) to the sphere (see Section 4.3.2 for more details).

For the SBF component, we make inference on unknowns $\mathbf{b}(\cdot)$, $\boldsymbol{\eta}$, and $\mathbf{K}$ in (4.1). This Bayesian source separation task (see, e.g., Knuth, 2005), where both the "source signal" $\boldsymbol{\eta}$ and the mixing coefficients $\mathbf{b}(\cdot)$ have to be estimated from a set of observations, can be achieved by putting a strong prior on both components. This has been attempted in the context of discretized-convolution models by Lemos and Sansó (2009), who infer (spatially varying) parameters determining the shapes of their kernels. Lopes et al. (2008) also consider a model of the form (4.1) where both $\mathbf{b}(\cdot)$ and $\boldsymbol{\eta}$ are random, but as each basis function is itself a Gaussian process, their approach offers no computational advantage for large spatial datasets. Our prior model for the SBF component is motivated by the predictive-process approach given by Banerjee et al. (2008), where the true ("parent") process is approximated with what is essentially a model of the form (4.1). This provides a natural way of inferring the shapes of the basis functions in a Bayesian framework, where

the parameters determining the correlation structure of the parent process are assumed to be random.

Then, conditional on the correlation structure of the parent process, Banerjee et al. (2008) consider the matrix $\mathbf{K}$ to be fixed. Here, we don't consider the parent process to be the truth, and we assume $\mathbf{K}$ to be random and distributed *a priori* according to an inverse-Wishart distribution whose mean is determined by the correlation structure of the parent process. This results in more flexible modeling of long-range dependence that can include negative correlations for the SBF component (induced by the random elements of $\mathbf{K}$), even if the covariance function of the parent process is nonnegative (see Section 4.2.3 for more details). In addition to allowing the shapes of the basis functions to be estimated, our model also allows for their number and locations to be random.

Posterior inference for our model is extremely fast, even for very large datasets. We take advantage of sparse-matrix operations to ensure fast computation, and we employ the marginalization strategies described in van Dyk and Park (2008) to achieve satisfactory mixing of the Markov chain. To estimate the number of basis functions, we make use of a reversible-jump Markov chain Monte Carlo (RJMCMC) algorithm (Green, 1995).

This chapter is organized as follows: In Section 4.2, we introduce our spatial model that combines a low-dimensional SBF component and a tapered FSV component. In Section 4.3, we give details on the covariance functions used in Section 4.2. We describe how a process can be modeled on a sphere (such as the globe), and we give prior distributions for the spatially varying parameters that determine the covariance functions. Section 4.4 deals with posterior inference on the unknown quantities in the model via reversible jump Markov chain Monte Carlo (RJMCMC) sampling. We discuss prediction of the true process at observed and unobserved locations, and we also discuss computational issues

related to the analysis of very large datasets. We then conduct a simulation study in Section 4.5, where our model is compared to several other models conceived for very large spatial datasets, plus a stationary "baseline" model. In Section 4.6, the model is applied to a real-world dataset of global CO_2 measurements obtained by a satellite remote-sensing instrument. Conclusions and avenues of further research are given in Section 4.7.

4.2 The Model

4.2.1 Model Overview

Let $\{Y(\mathbf{s}) : \ \mathbf{s} \in \mathcal{D}\}$, or $Y(\cdot)$, denote the process of interest on a spatial domain $\mathcal{D}$. Suppose we have n observations on $Y(\cdot)$, namely $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$, where n is very large, and we assume additive measurement error:

$$Z(\mathbf{s}_i) := Y(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \ i = 1, \ldots, n, \tag{4.2}$$

where $\epsilon(\cdot)|v_\epsilon(\cdot) \sim N(0, v_\epsilon(\cdot))$ independent of $Y(\cdot)$ and independent across space (given $v_\epsilon(\cdot)$). The function $v_\epsilon(\cdot)$ is assumed to be a known function up to a (possibly random) parameter vector $\boldsymbol{\theta}_\epsilon$. For simplicity and to ensure identifiability, throughout this chapter we will assume $v_\epsilon(\cdot) \equiv \sigma^2_\epsilon$ (i.e., $\boldsymbol{\theta}_\epsilon$ only consists of one parameter, σ^2_ϵ). It should be noted that our approach is easily generalized to $v_\epsilon(\cdot) = \sigma^2_\epsilon v(\cdot)$, for $v(\cdot)$ a known function. The standard deviation, σ_ϵ, of the measurement error can either be assumed known (from instrument experiments, or estimated from the data by extrapolating the variogram to the origin as described in Kang et al., 2009), or it can be assumed to have a prior distribution, such as $\sigma_\epsilon \sim \log N(\mu_{\sigma_\epsilon}, \sigma^2_{\sigma_\epsilon})$, where μ_{σ_ϵ} and $\sigma^2_{\sigma_\epsilon}$ are fixed hyperparameters.

We model $Y(\cdot)$ as the sum of three components,

$$Y(\cdot) := \mu(\cdot) + \nu(\cdot) + \delta(\cdot), \tag{4.3}$$

where $\mu(\cdot)$ is the large-scale trend, $\nu(\cdot)$ describes the medium-range to long-range spatial dependence, and the fine-scale variation $\delta(\cdot)$ accounts for local (or short-range) dependence.

The trend component will be assumed to be a linear combination of p known spatial trend terms (including an intercept),

$$\mu(\cdot) := \mathbf{x}(\cdot)'\boldsymbol{\beta}. \tag{4.4}$$

In (4.4), $\mathbf{x}(\cdot)$ is a p-dimensional vector of covariates, and the prior on $\boldsymbol{\beta}$ is assumed uniform on $\mathbb{R}^p$ (i.e., distributed according to an improper multivariate normal distribution with infinite variances).

4.2.2 The Fine-Scale-Variation (FSV) Component

The fine-scale-variation component, $\delta(\cdot)$, in (4.3) will be assumed to be a Gaussian process with mean zero and a compactly supported covariance function (see, e.g., Gneiting, 2002), so that its covariance matrix (when evaluated at a large number of locations) is sparse and hence quickly invertible. Given two vectors of parameters, $\boldsymbol{\theta}_\sigma$ and $\boldsymbol{\theta}_\delta$, the fine-scale covariance function is assumed to be of the form,

$$C_\delta(\mathbf{s}_1, \mathbf{s}_2) := cov(\delta(\mathbf{s}_1), \delta(\mathbf{s}_2)|\boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_\delta) = \sigma_\delta(\mathbf{s}_1)\sigma_\delta(\mathbf{s}_2)\rho_\delta(\mathbf{s}_1, \mathbf{s}_2),\ \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}, \tag{4.5}$$

where the function $\sigma_\delta : \mathcal{D} \to \mathbb{R}_0^+$ is determined by $\boldsymbol{\theta}_\sigma$, and $\rho_\delta(\cdot, \cdot)$ is a compactly supported, nonstationary correlation function determined by $\boldsymbol{\theta}_\delta$. More details and specific choices for the correlation function and its parameters are given in Section 4.3.

As discussed in Section 4.1, an FSV term is an important component in a reduced-dimensional spatial model. It can "correct" for underestimation of the variance and the non-smooth short-range correlation structure of the true process $Y(\cdot)$. Computational issues

related to posterior inference on the FSV component for very large datasets are discussed in Sections 4.4.4 and 4.4.5.

4.2.3 The Spatial-Basis-Function (SBF) Component

The component describing the medium-range to long-range spatial dependence is assumed to be of the form (4.1), conditional on a set of parameters described later.

To motivate the prior distribution of $\mathbf{b}(\cdot)$ and $\mathbf{K}$, consider a non-dimension-reduced, mean-zero Gaussian process $\tilde{\nu}(\cdot)$, referred to as the "parent process" by Banerjee et al. (2008). In contrast to Banerjee et al. (2008), we do *not* assume $\tilde{\nu}(\cdot)$ to be the true process, nor do we assume our SBF component to be its approximation; we merely use the parent process to motivate a prior distribution for our (more flexible) SBF component. Assume that, given vectors of parameters $\boldsymbol{\theta}_\sigma$ and $\boldsymbol{\theta}_{\tilde{\nu}}$, the parent process $\tilde{\nu}(\cdot)$ has covariance,

$$C_{\tilde{\nu}}(\mathbf{s}_1, \mathbf{s}_2) := cov(\tilde{\nu}(\mathbf{s}_1), \tilde{\nu}(\mathbf{s}_2) | \boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}) = \sigma_{\tilde{\nu}}(\mathbf{s}_1)\sigma_{\tilde{\nu}}(\mathbf{s}_2)\rho_{\tilde{\nu}}(\mathbf{s}_1, \mathbf{s}_2),\ \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}, \tag{4.6}$$

where the function $\sigma_{\tilde{\nu}}(\cdot) : \mathcal{D} \to \mathbb{R}_0^+$ is determined by parameters $\boldsymbol{\theta}_\sigma$, and $\rho_\nu(\cdot, \cdot)$ is a compactly supported, nonstationary correlation function that is relatively smooth (i.e., making it suitable to describe mainly medium-range to long-range dependence) and determined by parameters $\boldsymbol{\theta}_{\tilde{\nu}}$. The choice of the covariance function and its parameters is discussed in Section 4.3.

Given a covariance function as in (4.6), the predictive-process (PP) approach of Banerjee et al. (2008) is a natural reduced-dimensional approximation to $\tilde{\nu}(\cdot)$. The PP is obtained by conditioning on the parent process at a number of reference locations or "centers,"

$$\mathcal{C} := \{\mathbf{c}_1, \ldots, \mathbf{c}_r\}, \tag{4.7}$$

such that $\nu_{\text{PP}}(\cdot) := E(\tilde{\nu}(\cdot) | \tilde{\boldsymbol{\nu}}(\mathcal{C}), \boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}})$, where $\tilde{\boldsymbol{\nu}}(\mathcal{C}) := [\tilde{\nu}(\mathbf{c}_1), \ldots, \tilde{\nu}(\mathbf{c}_r)]'$. Conditional on $\boldsymbol{\theta}_\sigma$, $\boldsymbol{\theta}_{\tilde{\nu}}$, and $\mathcal{C}$, the PP can be written as a linear combination of basis functions in the

form (4.1), as follows:

$$\begin{aligned}
\nu_{\text{PP}}(\mathbf{s}) &:= E(\tilde{\nu}(\mathbf{s})|\tilde{\boldsymbol{\nu}}(\mathcal{C}), \boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}) \\
&= cov(\tilde{\nu}(\mathbf{s}), \tilde{\boldsymbol{\nu}}(\mathcal{C})|\boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}) \big(var(\tilde{\boldsymbol{\nu}}(\mathcal{C})|\boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}})\big)^{-1} \tilde{\boldsymbol{\nu}}(\mathcal{C}) \\
&= \sigma_{\tilde{\nu}}(\mathbf{s}) \big(\rho_{\tilde{\nu}}(\mathbf{s}, \mathbf{c}_1), \ldots, \rho_{\tilde{\nu}}(\mathbf{s}, \mathbf{c}_r)\big) \mathbf{D}_{\tilde{\nu}} \big(\mathbf{D}_{\tilde{\nu}} \mathbf{R}_{\tilde{\nu}} \mathbf{D}_{\tilde{\nu}}\big)^{-1} \tilde{\boldsymbol{\nu}}(\mathcal{C}) \\
&=: \mathbf{b}_{\text{PP}}(\mathbf{s})' \boldsymbol{\eta}_{\text{PP}},
\end{aligned} \tag{4.8}$$

where

$$\begin{aligned}
\mathbf{D}_{\tilde{\nu}} &:= diag(\sigma_{\tilde{\nu}}(\mathbf{c}_1), \ldots, \sigma_{\tilde{\nu}}(\mathbf{c}_r)) \\
\mathbf{R}_{\tilde{\nu}} &:= (\rho_{\tilde{\nu}}(\mathbf{c}_i, \mathbf{c}_j))_{i,j=1,\ldots,r} \\
\mathbf{b}_{\text{PP}}(\cdot) &:= \sigma_{\tilde{\nu}}(\cdot) \left(\rho_{\tilde{\nu}}(\cdot, \mathbf{c}_1), \ldots, \rho_{\tilde{\nu}}(\cdot, \mathbf{c}_r)\right)' \\
\boldsymbol{\eta}_{\text{PP}} &:= \mathbf{R}_{\tilde{\nu}}^{-1} \mathbf{D}_{\tilde{\nu}}^{-1} \tilde{\boldsymbol{\nu}}(\mathcal{C}).
\end{aligned} \tag{4.9}$$

Since $var(\tilde{\nu}(\mathcal{C})|\boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}) = \mathbf{D}_{\tilde{\nu}} \mathbf{R}_{\tilde{\nu}} \mathbf{D}_{\tilde{\nu}}$, then $var(\boldsymbol{\eta}_{\text{PP}}|\boldsymbol{\theta}_{\tilde{\nu}}) = \mathbf{R}_{\tilde{\nu}}^{-1}$. That is, $\boldsymbol{\eta}_{\text{PP}}|\boldsymbol{\theta}_{\tilde{\nu}} \sim N_r(\mathbf{0}, \mathbf{K}_{\text{PP}})$, where $\mathbf{K}_{\text{PP}} := \mathbf{R}_{\tilde{\nu}}^{-1}$.

Thus, once we have specified a covariance structure of the form (4.6) for a parent process $\tilde{\nu}(\cdot)$, the PP approach provides us with a relatively simple and intuitive way of obtaining a low-dimensional (specifically, r-dimensional) SBF approximation. In what follows, we shall use this idea to motivate a prior distribution for the SBF component.

In this research, we are interested in processes on very large domains, such as the globe, where stationary parametric models (e.g., the Matérn model) for the correlation function in (4.6) are not flexible enough to describe properly the medium-range to long-range dependence of the process, even if the correlation model is allowed to exhibit nonstationarity through spatially varying parameters, as described later in Section 4.3.1. Therefore, we generalize the PP approach as follows.

Conditional on parameters $\boldsymbol{\theta}_\sigma$, $\boldsymbol{\theta}_{\tilde{\nu}}$, and the centers $\mathcal{C}$, we assume

$$\nu(\cdot) = \mathbf{b}(\cdot)' \boldsymbol{\eta}, \tag{4.10}$$

where $\boldsymbol{\eta}|\mathbf{K} \sim N_r(\mathbf{0}, \mathbf{K})$, and $\mathbf{K}$ and $\mathbf{b}(\cdot)$ depend on $\boldsymbol{\theta}_\sigma$, $\boldsymbol{\theta}_{\tilde{\nu}}$, and $\mathcal{C}$. In our model, we would like the basis functions to be normalized to have a maximum of one; we define them to be,

$$\mathbf{b}(\cdot) := \big(\rho_{\tilde{\nu}}(\cdot, \mathbf{c}_1), \ldots, \rho_{\tilde{\nu}}(\cdot, \mathbf{c}_r)\big)'. \tag{4.11}$$

Instead of fixing $\mathbf{K} = \mathbf{R}_{\tilde{\nu}}^{-1}$ given $\boldsymbol{\theta}_{\tilde{\nu}}$ (as in the PP approach), we assume that the covariance function of the parent process, $C_\nu(\cdot, \cdot)$, determines the prior mean of $\mathbf{K}$, namely $E(\mathbf{K}|\boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}, \mathcal{C}) = \mathbf{D}_{\tilde{\nu}}\mathbf{R}_{\tilde{\nu}}^{-1}\mathbf{D}_{\tilde{\nu}}$, where $\mathbf{D}_{\tilde{\nu}}$ is defined in (4.9).

When comparing our model for $\mathbf{b}(\cdot)$ and $\mathbf{K}$ to the PP assumptions in (4.9), we have replaced $\mathbf{b}_{\text{PP}}(\cdot)$ in (4.9) by,

$$\big(\sigma_{\tilde{\nu}}(\mathbf{c}_1)\rho_{\tilde{\nu}}(\cdot, \mathbf{c}_1), \ldots, \sigma_{\tilde{\nu}}(\mathbf{c}_r)\rho_{\tilde{\nu}}(\cdot, \mathbf{c}_r)\big)' = \mathbf{D}_{\tilde{\nu}}\left(\rho_{\tilde{\nu}}(\cdot, \mathbf{c}_1), \ldots, \rho_{\tilde{\nu}}(\cdot, \mathbf{c}_r)\right)'.$$

and then we "moved" the standard deviation $\sigma_{\tilde{\nu}}(\cdot)$, in form of the diagonal matrix $\mathbf{D}_{\tilde{\nu}}$, from the basis functions to (the conditional mean of) $\mathbf{K}$. The effect of this change should be small, as long as the basis functions have compact support and $\sigma_{\tilde{\nu}}(\cdot)$ is assumed to vary relatively smoothly over space. And since the form of $\mathbf{b}_{\text{PP}}(\cdot)$ in (4.9) is used only for motivation, there is no loss and much to be gained: First, in our model the spatially varying parameter $\sigma_{\tilde{\nu}}(\cdot)$ is not able to modify the basis functions arbitrarily at each point in space. Second, standardized basis functions allow for easier calibration of prior distributions, and they tie in well with previous work in this area (e.g., Cressie and Johannesson, 2008; Chapters 2 and 3) and with the literature on discretized kernel convolutions (e.g. Lemos and Sansó, 2009). Third, "pulling" $\mathbf{D}_{\tilde{\nu}}$ into $\mathbf{K}$ makes $\mathbf{K}$ a covariance matrix (instead of an inverse correlation matrix), which provides us with the following conjugate prior for $\mathbf{K}$:

We assume the complete (conditional) prior distribution of $\mathbf{K}$ to be,

$$\mathbf{K}|\boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}, \mathcal{C} \sim IW_r(\mathbf{M}, u), \tag{4.12}$$

where $IW_r(\mathbf{M}, u)$ denotes an inverse Wishart distribution; the first parameter, $\mathbf{M} := (u-1)\mathbf{D}_{\tilde{\nu}}\mathbf{R}_{\tilde{\nu}}^{-1}\mathbf{D}_{\tilde{\nu}}$, depends implicitly on $\boldsymbol{\theta}_\sigma$, $\boldsymbol{\theta}_{\tilde{\nu}}$ and $\mathcal{C}$; and u is a (fixed) hyperparameter greater than 1 that determines the variability of the prior distribution of $\mathbf{K}$ around its mean, $E(\mathbf{K}|\boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}, \mathcal{C}) = \mathbf{D}_{\tilde{\nu}}\mathbf{R}_{\tilde{\nu}}^{-1}\mathbf{D}_{\tilde{\nu}}$. Now, u can be interpreted as a gauge of our prior belief in how close $C_\nu(\cdot,\cdot)$, the true process' covariance function, is to $C_{\tilde{\nu}}(\cdot,\cdot)$ in (4.6). If we think that $C_{\tilde{\nu}}(\cdot,\cdot)$ is indeed correct, we can set u to a very large value, so that the distribution of $\mathbf{K}$ is very "tight" around $\mathbf{D}_\nu\mathbf{R}^{-1}\mathbf{D}_\nu$. If we have rather little faith in $C_{\tilde{\nu}}(\cdot,\cdot)$, we can set $u = 2$, which essentially weights the data (through $\boldsymbol{\eta}$) and the covariance model $C_{\tilde{\nu}}(\cdot,\cdot)$ equally in the full conditional distribution of $\mathbf{K}$ (see (4.23) below).

For any $r \times r$ positive-definite matrix $\mathbf{K}$, the probability density function (PDF) of $\mathbf{K}$ obtained from (4.12) is denoted as $IW_r(\mathbf{K}|\mathbf{M}, u)$ and given by,

$$IW_r(\mathbf{K}|\mathbf{M}, u) = \frac{2^{-r(r+u)/2}}{\Gamma_r((r+u)/2)} \frac{|\mathbf{M}|^{(r+u)/2}}{|\mathbf{K}|^{(2r+u+1)/2}} \exp\left\{-\frac{1}{2}trace(\mathbf{M}\mathbf{K}^{-1})\right\},$$

with mean $\mathbf{M}/(u-1)$, and $\Gamma_r(\cdot)$ is the multivariate gamma function (e.g., James, 1964). We will need this formula when constructing the full conditional distributions in Section 4.4.

The general SBF component (as described, e.g., by Cressie and Johannesson, 2008) can be used with any arbitrarily chosen basis functions. However, our specific SBF model here is motivated by a "parent process" that must have a valid covariance function. Thus, the choice of basis functions for our model here is obtained directly from a valid correlation function on $\mathcal{D}$ (see (4.9)), and so it should be noted that in (4.10) we cannot just choose any basis functions (e.g., bisquare functions are not positive-definite, and would be used in (4.10)). On the other hand, the motivation above provides us with a natural center for the prior distribution of $\mathbf{K}$, as in (4.12). An informative prior of this kind can be quite helpful

in spatial-only analyses where no repeated measurements are available and estimation of a general $r \times r$ covariance matrix $\mathbf{K}$ might be problematic.

4.2.4 The Prior Distribution of the Basis-Function Centers, $\mathcal{C}$

In what follows, we avoid requiring the basis-function centers $\mathcal{C}$ in (4.7) to be fixed and pre-specified (as in Banerjee et al., 2008), by putting a prior distribution on both the number, r, and the locations of the centers. This approach is inspired by Holmes and Mallick (2001), who propose a piecewise linear spline regression model, for which both the number and the locations of the splines are random.

As discussed later at the end of Section 4.4.3, it is not necessary to strongly penalize large r through the prior on $\mathcal{C}$, and so we assume that the prior for $\mathcal{C}$ has "density,"

$$[\mathcal{C}] \propto \zeta^{\psi_\xi(\mathcal{C})}.$$

which is similar to a Strauss process (e.g., Møller and Waagepetersen, 2004, p. 85). The locations of the centers are penalized for being too close to each other through the term,

$$\psi_\xi(\mathcal{C}) = \sum_{i \neq j} I(\|\mathbf{c}_i - \mathbf{c}_j\| \leq \xi), \tag{4.13}$$

where $\zeta \in [0, 1]$ determines the severity of the penalization, and ξ is the distance up to which penalization occurs. Because of the FSV component in the SRE model, two centers in the SBF component that are very close to each other are likely not of much more use than one center at that location. Moreover, two close centers can result in numerical instability when inverting $\mathbf{R}$, because the correlation between the two centers will be very close to one. Therefore, we set to zero the probability of two centers being very close in space, by setting the penalization parameter $\zeta = 0$ and defining $0^0 = 1$ and $0^x = 0$ for $x > 0$.

Our prior for the set of centers, $\mathcal{C}$, therefore follows an inhibitory point process that is more regular than a homogeneous Poisson process, because no pair of centers is allowed to

have a distance of less than or equal to ξ. In our experience, a good value for the distance ξ is 10% of the the support of the covariance function of $\tilde{\nu}(\cdot)$ in (4.6).

4.3 Covariance Functions for the SBF and FSV Components

In this section, we describe the covariance functions that will be used in our model for the functions $C_\delta(\cdot,\cdot)$ in (4.5) and $C_{\tilde{\nu}}(\cdot,\cdot)$ in (4.6). In subsection 4.3.1, we will recapitulate a class of nonstationary covariance functions valid in $\mathbb{R}^d$, $d = 1, 2, \ldots$, proposed by Paciorek and Schervish (2006) and extended in an unpublished technical report by Stein (2005). We then combine them with a tapering function (as suggested by Gneiting, 2002) to obtain a class of covariance functions with compact support . In subsection 4.3.2, we discuss how this class of models can be adapted for use in $C_\delta(\cdot,\cdot)$ and $C_{\tilde{\nu}}(\cdot,\cdot)$ when the domain, $\mathcal{D}$, is a sphere (the globe). Finally, in subsection 4.3.3, we describe our prior distributions for the parameters in the covariance models.

4.3.1 A Class of Compactly Supported, Nonstationary Covariance Functions

For domain, $\mathcal{D} = \mathbb{R}^d$, $d = 1, 2, \ldots$, many valid isotropic correlation functions are available. In this chapter, we need the Matérn correlation function (Stein, 1999, p. 12), given by,

$$\mathcal{M}_\upsilon(h) = (2h\sqrt{\upsilon})^\upsilon \mathcal{K}_\upsilon(2h\sqrt{\upsilon}) 2^{1-\upsilon}/\Gamma(\upsilon),\ h \geq 0, \tag{4.14}$$

where $\mathcal{K}_\upsilon(\cdot)$ is the modified Bessel function of order $\upsilon > 0$. The smoothness parameter υ determines the differentiability of (4.14) at the origin (and therefore the smoothness of the corresponding process).

If $h(\mathbf{s}_1, \mathbf{s}_2)$ is defined as the Euclidean distance between two locations $\mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^d$, then $\mathcal{M}_\upsilon(h(\mathbf{s}_1, \mathbf{s}_2))$ is a valid, stationary, and isotropic correlation function. A nonstationary

generalization of this function (Paciorek and Schervish, 2006) can be obtained by defining a new, spatially varying Mahalanobis-like distance,

$$q(\mathbf{s}_1, \mathbf{s}_2) = \{2(\mathbf{s}_1 - \mathbf{s}_2)'(\mathbf{\Sigma}(\mathbf{s}_1) + \mathbf{\Sigma}(\mathbf{s}_2))^{-1}(\mathbf{s}_1 - \mathbf{s}_2)\}^{1/2}, \tag{4.15}$$

where $\mathbf{\Sigma}(\mathbf{s})$ is a $d \times d$ positive-definite matrix describing the (local) geometric anisotropy at location $\mathbf{s}$. Specific parameterizations of this matrix are discussed in the next subsection. Using this spatially varying distance (4.15), a nonstationary Matérn covariance function is given by,

$$\widetilde{\mathcal{M}}(\mathbf{s}_1, \mathbf{s}_2) = c(\mathbf{s}_1, \mathbf{s}_2)\mathcal{M}_{(\upsilon(\mathbf{s}_1)+\upsilon(\mathbf{s}_2))/2}(q(\mathbf{s}_1, \mathbf{s}_2)). \tag{4.16}$$

If the normalization term is chosen as,

$$c(\mathbf{s}_1, \mathbf{s}_2) := |\mathbf{\Sigma}(\mathbf{s}_1)|^{1/4}|\mathbf{\Sigma}(\mathbf{s}_2)|^{1/4}|(\mathbf{\Sigma}(\mathbf{s}_1) + \mathbf{\Sigma}(\mathbf{s}_2))/2|^{-1/2}, \tag{4.17}$$

then $\widetilde{\mathcal{M}}(\mathbf{s}, \mathbf{s}) = 1$ and (4.16) is a valid correlation function (Paciorek and Schervish, 2006; Stein, 2005). This nonstationary Matérn class is very flexible, in that it allows for a spatially varying range and geometric anisotropy through the matrix $\mathbf{\Sigma}(\cdot)$, and the smoothness of the corresponding process at location $\mathbf{s}$ is determined by $\upsilon(\mathbf{s})$, where $\upsilon : \mathcal{D} \to \mathbb{R}^+$.

Unfortunately, (4.16) does not satisfy our requirement of compact support for $C_\delta(\cdot, \cdot)$ and $C_{\tilde{\nu}}(\cdot, \cdot)$, stated below (4.5) and (4.6), respectively. However, compact support can easily be achieved by multiplying (4.16) with a valid correlation model that does exhibit compact support (Gneiting, 2002). Specifically, we will use Kanter's function (Kanter, 1997) defined on $[0, \infty)$:

$$\mathcal{T}(x) = (1 - x)\frac{\sin(2\pi x)}{2\pi x} + \frac{1 - \cos(2\pi x)}{2\pi^2 x}, \text{ for } x \in (0, 1);$$

$\mathcal{T}(x) = 0$, for $x \geq 1$; and we set $\mathcal{T}(0) = 1$. The function $\mathcal{T}(\|\mathbf{h}\|)$ is positive-definite for $\mathbf{h} \in \mathbb{R}^3$, it is twice differentiable at the origin, and it minimizes the curvature at 0 within

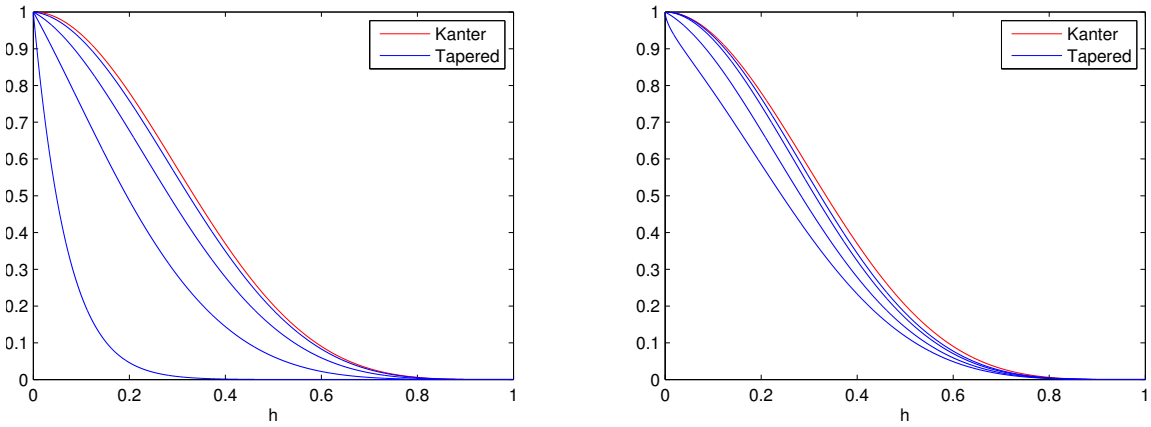

Figure 4.1: Kanter's function (in red) and the correlation model $\rho(0, h)$ in (4.18) for $L = 1$ (in blue). Here, both $\upsilon(\cdot)$ in (4.18) and $\mathbf{\Sigma}(\cdot)$ in (4.15) are held constant, and $\mathbf{\Sigma}$ is a 1×1 matrix (i.e., a scalar), denoted Σ. Left panel: $\upsilon = 0.5$ and $\Sigma = 0.1, 0.6, 2, 10$ (from left to right). Right panel: $\Sigma = 2$, and $\upsilon = 0.3, 0.5, 1, 2$ (from left to right).

the class of all compactly supported, valid (in $\mathbb{R}^3$) correlation functions (Gneiting, 2002). Kanter's function is shown in Figure 4.1 (in red).

In summary, conditional on a set of hyperparameters (discussed in Section 4.3.3 below), we have defined a nonstationary, compactly supported correlation function in $\mathbb{R}^3$ of the form,

$$\rho(\mathbf{s}_1, \mathbf{s}_2) = c(\mathbf{s}_1, \mathbf{s}_2)\mathcal{M}_{(\upsilon(\mathbf{s}_1)+\upsilon(\mathbf{s}_2))/2}(q(\mathbf{s}_1, \mathbf{s}_2))\, \mathcal{T}(\|\mathbf{s}_1 - \mathbf{s}_2\|/L),\ \mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^3, \qquad (4.18)$$

where $\mathcal{T}(x) = 0$ for $x \geq 1$, and so L is the tapering length (i.e., $\rho(\mathbf{s}_1, \mathbf{s}_2)$ is zero if the distance between $\mathbf{s}_1$ and $\mathbf{s}_2$ is greater than L). The correlation function (4.18) will be used in both (4.5) and (4.6)

4.3.2 Extending the Class of Nonstationary Covariance Functions to the Sphere

Let $\mathcal{S}^2$ denote the unit 2-sphere (in $\mathbb{R}^3$). Due to the curvature of $\mathcal{S}^2$, finding a valid correlation function for a process observed on the globe is not trivial (e.g., Jones, 1963). Das (2000) obtains several such functions in closed form, but these covariance functions produce realizations that are unrealistically smooth for most applications (Stein, 1999). Here, we follow instead the idea of Yaglom (1987), by restricting a valid covariance function in $\mathbb{R}^3$ to $\mathcal{S}^2$ and using chordal distance as a measure of distance (e.g., Banerjee, 2005). Of course, great-arc distance is the more relevant distance for processes that reside on the surface of the globe, but the approach nonetheless works well here, because the covariance functions $C_\delta(\cdot,\cdot)$ and $C_{\tilde{\nu}}(\cdot,\cdot)$ in (4.5) and (4.6), respectively, are constructed to be compactly supported, and so they only describe dependence for relatively short distances, for which the difference between chordal distance and great-arc distance is small. Specifically, the relationship between great-arc distance and chordal distance on $\mathcal{S}^2$ is given by,

$$ch = 2\sin(ga/2),$$

where ch is the chordal distance between two points on $\mathcal{S}^2$, and ga is the corresponding great-arc distance. Because $\sin(x) \approx x$ for $x \in [0, 0.5]$, two points on the sphere that are less than the sphere's radius (here, 1) apart have approximately equal chordal distance and great-arc distance. We will respect this when choosing the compact supports of $C_\delta(\cdot,\cdot)$ in (4.5) and $C_{\tilde{\nu}}(\cdot,\cdot)$ in (4.6).

Now, to apply the nonstationary correlation model in (4.18) to points on $\mathcal{S}^2$, we first need to convert the longitude-latitude coordinates to (x, y, z)-coordinates of a three-dimensional Cartesian coordinate system. Without loss of generality, assume that $\mathcal{S}^2$ is centered at the origin, $(0, 0, 0)$, and that the intersection of the prime meridian and the equator, $\mathbf{c} := (0, 0)$

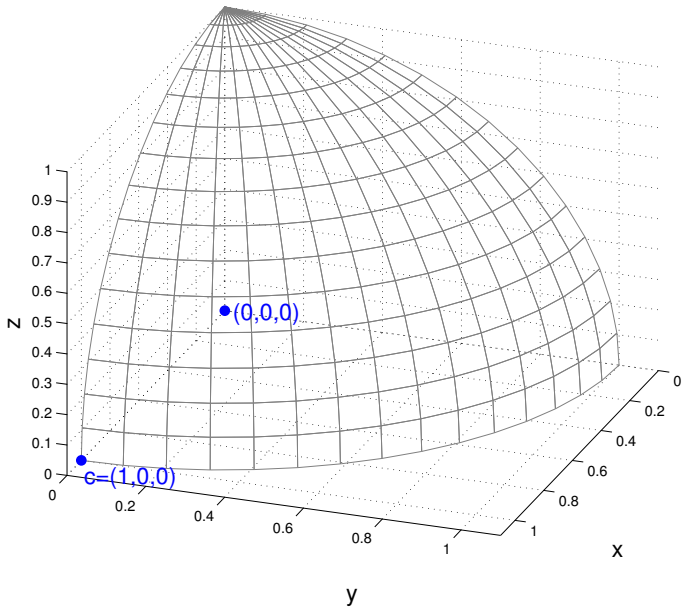

Figure 4.2: The part of a unit sphere centered at the origin that lies in the first octant of the Cartesian coordinate system, where all coefficients are positive. The origin and the point $\mathbf{c}$ referred to in the text are shown in blue.

($0°$ longitude, $0°$ latitude), is located on the x-axis (i.e., it has Euclidean coordinates $\tilde{\mathbf{c}} := (1, 0, 0)$). The part of such a sphere that lies in the first (positive) octant of a Cartesian coordinate system is shown for illustration in Figure 4.2. The (x, y, z)-coordinates of any point $\mathbf{s} = (s_1, s_2)$ with longitude s_1 and latitude s_2 on $\mathcal{S}^2$ are then given by,

$$x = \cos(s_2)\cos(s_1)$$

$$y = \cos(s_2)\sin(s_1)$$

$$z = \sin(s_2).$$

The more challenging task in generalizing (4.18) to the sphere is finding a sensible parameterization for the anisotropy matrix $\boldsymbol{\Sigma}(\cdot)$. For d-dimensional Euclidean space, we can parameterize this matrix using d scaling parameters and $d-1$ rotation parameters (see, e.g., Banerjee et al., 2008). But while $\mathcal{S}^2$ "lives" in $\mathbb{R}^3$, the surface of $\mathcal{S}^2$ is really (locally)

an approximately two-dimensional space at any point $\mathbf{s} \in \mathcal{S}^2$. This means that we would really only want to use two local scaling parameters and one local rotation parameter.

To parameterize $\boldsymbol{\Sigma}(\cdot)$ using these three parameters, consider again the (Euclidean) reference point $\tilde{\mathbf{c}} := (1, 0, 0)$ with spherical coordinates $\mathbf{c} = (0, 0)$. When we only consider a small area centered at $\mathbf{c}$, the sphere in this area is essentially flat and can be described by its y- and z-coordinates (see Figure 4.2). We therefore introduce two scaling parameters, $\gamma_1(\mathbf{c}) > 0$ and $\gamma_2(\mathbf{c}) > 0$, that describe the correlation length in the y- and z-directions, respectively. Defining a diagonal scaling matrix, $\mathbf{D}(\boldsymbol{\gamma}) := diag\{1, \gamma_1, \gamma_2\}$, the local scaling matrix at the reference point $\tilde{\mathbf{c}}$ (spherical coordinates $\mathbf{c}$) with parameters $\boldsymbol{\gamma}(\mathbf{c}) := (\gamma_1(\mathbf{c}), \gamma_2(\mathbf{c}))'$ is given by,

$$\mathbf{D}(\boldsymbol{\gamma}(\mathbf{c})) := diag\{1, \gamma_1(\mathbf{c}), \gamma_2(\mathbf{c})\}.$$

We introduce a rotation parameter, $\kappa(\mathbf{c}) \in [0, \pi/2)$, which rotates the (y, z)-coordinate system at $\tilde{\mathbf{c}}$ about the x-axis through the rotation matrix $\mathcal{R}_x(\kappa(\mathbf{c}))$. This rotation matrix is defined for a general rotation parameter κ, as

$$\mathcal{R}_x(\kappa) := \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\kappa & -\sin\kappa \\ 0 & \sin\kappa & \cos\kappa \end{pmatrix},$$

The local anisotropy matrix at the reference point $\tilde{\mathbf{c}}$ (spherical coordinates $\mathbf{c}$) is then given by,

$$\tilde{\boldsymbol{\Sigma}}(\mathbf{c}) := \mathcal{R}_x(\kappa(\mathbf{c}))\mathbf{D}(\boldsymbol{\gamma}(\mathbf{c}))\mathcal{R}_x(\kappa(\mathbf{c}))'. \tag{4.19}$$

Now let $\mathbf{s}$ be an arbitrary location on $\mathcal{S}^2$, with longitude s_1 and latitude s_2, and corresponding scaling parameters $\gamma_1(\mathbf{s})$ and $\gamma_2(\mathbf{s})$ and rotation parameter $\kappa(\mathbf{s})$. The functions $\gamma_j : \mathcal{S}^2 \to \mathbb{R}^+$, $j = 1, 2$, and $\kappa : \mathcal{S}^2 \to [0, \pi/2)$ are described further in Subsection 4.3.3 below. Let $\tilde{\mathbf{c}}$ and $\tilde{\mathbf{s}}$ be two vectors of the Euclidean coordinates corresponding to the spherical locations $\mathbf{c} = (0, 0)$ and $\mathbf{s} = (s_1, s_2)$, respectively. Let $\mathcal{R}_y(\theta)$ and $\mathcal{R}_z(\theta)$ be matrices

that rotate a vector in $\mathbb{R}^3$ by angle θ about the y-axis and the z-axis, respectively. The two rotation matrices are given by,

$$\mathcal{R}_y(\theta) := \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \quad \text{and} \quad \mathcal{R}_z(\theta) := \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

To extend the anisotropy matrix in (4.19) from our reference point, $\mathbf{c}$, to an arbitrary point, $\mathbf{s} \in \mathcal{S}^2$, we can simply rotate $\tilde{\mathbf{s}}$ about the y-axis and z-axis to the point $\tilde{\mathbf{c}}$ by writing $\tilde{\mathbf{c}} = \mathcal{R}_y(-s_2)\mathcal{R}_z(-s_1)\tilde{\mathbf{s}}$. We then combine these rotation matrices with the anisotropy matrix for $\mathbf{c}$ in (4.19) to obtain a quadratic form as in (4.15). Specifically,

$$\tilde{\mathbf{c}}'\tilde{\boldsymbol{\Sigma}}(\mathbf{s})^{-1}\tilde{\mathbf{c}} = \Big(\mathcal{R}_y(-s_2)\mathcal{R}_z(-s_1)\,\tilde{\mathbf{s}}\Big)'\tilde{\boldsymbol{\Sigma}}(\mathbf{s})^{-1}\Big(\mathcal{R}_y(-s_2)\mathcal{R}_z(-s_1)\,\tilde{\mathbf{s}}\Big) =: \tilde{\mathbf{s}}'\boldsymbol{\Sigma}(\mathbf{s})^{-1}\tilde{\mathbf{s}},$$

where $\tilde{\boldsymbol{\Sigma}}(\mathbf{s}) := \mathcal{R}_x(\kappa(\mathbf{s}))\mathbf{D}(\boldsymbol{\gamma}(\mathbf{s}))\mathcal{R}_x(\kappa(\mathbf{s}))'$, and

$$\boldsymbol{\Sigma}(\mathbf{s}) := \mathcal{R}_z(s_1)\mathcal{R}_y(s_2)\tilde{\boldsymbol{\Sigma}}(\mathbf{s})\mathcal{R}_y(s_2)'\mathcal{R}_z(s_1)'$$

is the anisotropy matrix that will be used in (4.15) and (4.17) for spherical domains (i.e., when $\mathcal{D} \subseteq \mathcal{S}^2$).

4.3.3 The Prior Distributions of the Parameters in the Covariance Models

In the previous two subsections, we have introduced a number of spatially varying parameters for the nonstationary correlation function (4.18) that will be used for the FSV component, $\delta(\cdot)$, in (4.5) and the parent process, $\tilde{\nu}(\cdot)$, of the SBF component in (4.6). Specifically, we have to specify a model for the smoothness parameter, $\upsilon(\cdot)$, the scaling parameter(s), $\gamma_j(\cdot)$ (hereafter referred to generically as $\gamma(\cdot)$), and the rotation parameter, $\kappa(\cdot)$, for each of the two components. In addition, we have to specify the standard deviation parameter, $\sigma(\cdot)$, used in (4.5) and (4.6), and a tapering length, L, used in (4.18), for each of the two components.

Both tapering lengths will be assumed fixed. The tapering length for the FSV component, L_δ, is mainly determined by computational feasibility (see Section 4.4.5 for more details). For the SBF component, we would still like to choose a finite tapering length, L_ν. For the sphere, this is due to our use of chordal distance as a measure of distance (see previous subsection). However, even in Euclidean space, where such considerations play no role, it is known from the wavelet literature (e.g., Daubechies, 1992) that compact support of the basis functions (which is determined by L_ν) is crucial for local adaptability to a highly variable function (or process).

For the remaining parameters, we first introduce a general prior model, and then give specifics for the covariance parameters of the two components, $\delta(\cdot)$ and $\nu(\cdot)$, in our model.

Let $\theta(\cdot)$ be a generic notation for one of the spatially varying parameters used in the covariance functions, as listed in the first two columns of Table 4.1. We assume that all covariance parameters are of the form,

$$\theta(\cdot) = g_\theta(\tilde{\theta} + \mathbf{b}_\theta(\cdot)'\boldsymbol{\eta}_\theta), \tag{4.20}$$

where $\tilde{\theta} \sim N(\mu_\theta, \sigma_\theta^2)$, $\boldsymbol{\eta}_\theta \sim N_{r_\theta}(\mathbf{0}, \tau_\theta^2 \mathbf{I}_{r_\theta})$, and $\mathbf{b}_\theta(\cdot)$ is an r_θ-dimensional vector of *fixed* basis functions, each normalized to $[0, 1]$. The functions $g_\theta(\cdot)$ are transformations from $\mathbb{R}$ to the range of the function $\theta(\cdot)$. Our specific choices are given in Table 4.1. Note that the smoothness parameter $\upsilon(\cdot)$ can theoretically take on any positive value, but we restrict it to the interval $[0, 2]$, as "the data can rarely inform about smoothness of higher orders" (Banerjee et al., 2008).

In the rest of this chapter, the basis functions, $\mathbf{b}_\theta(\cdot)$, that describe the parameters in (4.20), are taken to be the same for all parameters. Any choice of basis functions is possible but, assuming that the covariance parameters vary smoothly over space, we recommend choosing a relatively small number of bisquare functions with a relatively large (fixed)

Table 4.1: Spatially varying covariance parameters (generically denoted by $\theta(\cdot)$), together with their ranges, and the corresponding transformations, $g_\theta : \mathbb{R} \to \text{range}(\theta)$; see the text for details.

Parameter	Symbol $\theta(\cdot)$	Range of θ	Transformation $g_\theta(\cdot)$
Standard deviation	$\sigma(\cdot)$	$\mathbb{R}^+$	$\exp(\cdot)$
Smoothness	$\upsilon(\cdot)$	$[0, 2]$	$2\Phi(\cdot)$
Scale	$\gamma(\cdot)$	$\mathbb{R}^+$	$\exp(\cdot)$
Rotation angle	$\kappa(\cdot)$	$[0, \pi/2]$	$(\pi/2)\Phi(\cdot)$

radius. Specific choices depend on the domain $\mathcal{D}$ and are given in Sections 4.5 and 4.6. To achieve identifiability, we also set τ_θ^2 in the prior distribution of $\boldsymbol{\eta}_\theta$, given below (4.20), to be a fairly small value; in this chapter, we use the value $\tau_\theta^2 = (0.25)^2$ for all spatially varying covariance parameters. In Chapter 3, we show that this value roughly results in $1/2$ and 2 as the lower and upper endpoints, respectively, of a 95% credible interval for the ratio $\theta(\mathbf{s}_1)/\theta(\mathbf{s}_2)$, if $\mathbf{s}_1$ and $\mathbf{s}_2$ are two distance locations in the spatial domain $\mathcal{D}$.

This completes the general description of our models for spatially varying covariance parameters for the nonstationary Matérn covariance function given by (4.18). Recall that there are two covariance models in our model, $C_{\tilde{\nu}}(\cdot, \cdot)$ in (4.6), and $C_\delta(\cdot, \cdot)$ in (4.5), each with its own set of standard-deviation, smoothness, scale, and rotation parameters. The parameter models for the SBF component and the FSV component are both given by (4.20) except that, for the former all quantities will have a subscript $\tilde{\nu}$, and for the latter all quantities will have a subscript δ. All that remains is a specification of the hyperparameters determining the prior distributions below (4.20) for each set of parameters. Some parameters will be fixed at sensible values to improve mixing and identifiability.

The SBF component, $\nu(\cdot)$, only models smooth variation, and so we fix the smoothness parameter $\upsilon_{\tilde{\nu}}(\cdot) \equiv 2$ (i.e., we fix $\tilde{\upsilon}_{\tilde{\nu}} = \infty$ and set $\tau^2_{\upsilon_{\tilde{\nu}}} = 0$ in the prior distributions below

(4.20)) for the correlation function, $\rho_{\tilde{\nu}}(\cdot,\cdot)$, of its parent process. For the rotation angle, $\kappa_{\tilde{\nu}}(\cdot)$, we choose an approximate uniform distribution on $[0,\pi/2]$, which is attained by setting the hyperparameters, $\mu_{\kappa_{\tilde{\nu}}}=0$ and $\sigma^2_{\kappa_{\tilde{\nu}}}=1$ (if $\tau^2_{\kappa_{\tilde{\nu}}}$ were zero, the prior would be exactly uniform). The hyperparameters of $\gamma_{\tilde{\nu}}(\cdot)$ should reflect prior beliefs about the correlation lengths for the process of interest, which will depend on the application and the size of the domain $\mathcal{D}$; we discuss specific choices in Sections 4.5 and 4.6. Ignoring the parameters that have been fixed, the parameters determining the spatially varying correlation parameters of the parent process of the SBF component through the model (4.20), will be collected in a vector denoted by $\boldsymbol{\theta}_{\tilde{\nu}} := (\tilde{\gamma}_{\tilde{\nu}}, \boldsymbol{\eta}'_{\gamma_{\tilde{\nu}}}, \tilde{\kappa}_{\tilde{\nu}}, \boldsymbol{\eta}'_{\kappa_{\tilde{\nu}}})'$.

For the correlation function, $\rho_\delta(\cdot,\cdot)$, of the FSV component, we fix $\gamma_\delta(\cdot)\equiv 2L_\delta$ (to avoid having to make inference on too many parameters), but we let the smoothness parameter, $\nu_\delta(\cdot)$, follow an approximate uniform distribution on $[0,2]$, which is attained for our model of the form (4.20) by specifying the hyperparameters as $\mu_{\upsilon_\delta}=0$ and $\sigma^2_{\upsilon_\delta}=1$ (again, the prior would be exactly uniform if $\tau^2_{\upsilon_\delta}$ were zero). The resulting correlation function is illustrated for different smoothness parameters in the right panel of Figure 4.1. If the domain $\mathcal{D}$ is in two- or three-dimensional space, there will be multiple scale parameters, $\gamma_\delta(\cdot)$, and one or more rotation parameters, $\kappa_\delta(\cdot)$. However, since the scale parameters for $\delta(\cdot)$ are all fixed at 2, the rotation does not matter, and we simply set the rotation parameters equal to zero; that is, $\kappa_\delta(\cdot)\equiv 0$. The parameters determining the correlation function of $\delta(\cdot)$ through (4.20) are therefore $\boldsymbol{\theta}_\delta := (\tilde{\upsilon}_\delta, \boldsymbol{\eta}'_{\upsilon_\delta})'$.

Now for the standard deviation parameters, $\sigma_{\tilde{\nu}}(\cdot)$ and $\sigma_\delta(\cdot)$, we make use of an insight by Finley et al. (2009). The FSV component's role is in part to correct for the underestimation of the variability in the process that results from the dimension reduction in the SBF component. The FSV component is important (i.e., its variance should be large) in

areas where the SBF component only picks up a small part of the overall variance of the process (around its mean, $\mu(\cdot)$), and the FSV variance should be small at locations close to a basis-function center of the SBF component. We let $\sigma_{\tilde{\nu}}(\cdot)$ vary as described in (4.20) and in Table 4.1 (specific choices for its hyperparameters are discussed in Sections 4.5 and 4.6). Given $\sigma_{\tilde{\nu}}(\cdot)$, the variance of the predictive process (4.8) at location $\mathbf{s}$ (conditional on $\boldsymbol{\theta}_{\tilde{\nu}}$ and $\mathcal{C}$) is given by,

$$(\sigma_{\nu_{\text{PP}}}(\cdot))^2 := var(\nu_{\text{PP}}(\cdot)|\sigma_{\tilde{\nu}}(\cdot), \boldsymbol{\theta}_{\tilde{\nu}}, \mathcal{C}) = (\sigma_{\tilde{\nu}}(\cdot))^2 \mathbf{b}(\cdot)' \mathbf{R}_{\tilde{\nu}}^{-1} \mathbf{b}(\cdot).$$

It makes sense to let the standard deviation of the FSV component, $\sigma_\delta(\cdot)$, be determined by the difference between the variance of the parent process, $(\sigma_{\tilde{\nu}}(\cdot))^2$, and the variance of the predictive process, $(\sigma_{\nu_{\text{PP}}}(\cdot))^2$, such that,

$$\sigma_\delta(\cdot) = \sqrt{(\sigma_{\tilde{\nu}}(\cdot))^2 - (\sigma_{\nu_{\text{PP}}}(\cdot))^2} = \sigma_{\tilde{\nu}}(\cdot)\sqrt{1 - \mathbf{b}(\cdot)'\mathbf{R}^{-1}\mathbf{b}(\cdot)}. \tag{4.21}$$

From the law of total variance and the definition of the predictive process in Section 4.2.3, we have,

$$(\sigma_\nu(\cdot))^2 - (\sigma_{\nu_{\text{PP}}}(\cdot))^2 = var(\tilde{\nu}(\cdot)) - var(E(\tilde{\nu}(\cdot)|\tilde{\boldsymbol{\nu}}(\mathcal{C}))) = E(var(\tilde{\nu}(\cdot)|\tilde{\boldsymbol{\nu}}(\mathcal{C}))) \geq 0,$$

and so our model for $\sigma_\delta(\cdot)$ is valid. Of course, our model for ν is different from the predictive process, and so it seems natural to replace $(\sigma_{\nu_{\text{PP}}}(\cdot))^2 = var(\nu_{\text{PP}}(\cdot)|\sigma_{\tilde{\nu}}(\cdot), \boldsymbol{\theta}_{\tilde{\nu}}, \mathcal{C})$ in (4.21) by $var(\nu(\cdot)|\sigma_{\tilde{\nu}}(\cdot), \boldsymbol{\theta}_{\tilde{\nu}}, \mathcal{C})$, where $\nu(\cdot)$ is given by (4.10)–(4.12). However, the law of total variance cannot be applied as above, and this choice can result in an invalid $\sigma_\delta(\cdot)$.

In summary, the spatially varying parameters $\sigma_\delta(\cdot)$ and $\sigma_{\tilde{\nu}}(\cdot)$ are both determined by $\boldsymbol{\theta}_\sigma := (\tilde{\sigma}_{\tilde{\nu}}, \boldsymbol{\eta}'_{\sigma_{\tilde{\nu}}})'$ through (4.20) and (4.21).

4.4 Posterior Inference

4.4.1 Summary of the Hierarchical Model

We begin by writing the model developed in Sections 4.2 and 4.3 in vector notation.

The data model is given by,

$$\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}_\epsilon \sim N_n(\mathbf{Y}, \mathbf{V}_\epsilon),$$

where $\mathbf{Z} := (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$, $\mathbf{Y} := (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$, and $\mathbf{V}_\epsilon := diag(v_\epsilon(\mathbf{s}_1), \ldots, v_\epsilon(\mathbf{s}_n))'$.

The process model is given by,

$$\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\eta}, \mathcal{C}, \boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}, \boldsymbol{\theta}_\delta \sim N_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\eta}, \mathbf{V}_\delta),$$

and $\boldsymbol{\eta}|\mathbf{K} \sim N_r(\mathbf{0}, \mathbf{K})$, where the i-th row of matrix $\mathbf{X}$ is given by $\mathbf{x}(\mathbf{s}_i)$; the $n \times r$ matrix $\mathbf{B}$ has (i, j)-th element $(\rho_\nu(\mathbf{s}_i, \mathbf{c}_j))$ and is determined by $\boldsymbol{\theta}_{\tilde{\nu}}$ and $\mathcal{C}$; and $\mathbf{V}_\delta := (C_\delta(\mathbf{s}_i, \mathbf{s}_j))_{i,j=1,\ldots,n}$ is the sparse $n \times n$ covariance matrix of the FSV vector $\boldsymbol{\delta} := (\delta(\mathbf{s}_1), \ldots, \delta(\mathbf{s}_n))'$. We further define $\mathbf{V} := \mathbf{V}_\delta + \mathbf{V}_\epsilon$.

The parameter model consists of the prior distributions of the parameters $\boldsymbol{\theta}_\epsilon$ and $\boldsymbol{\beta}$ (see Section 4.2.1), $\mathbf{K}$ (see (4.12)), $\mathcal{C}$ (Section 4.2.4), and $\boldsymbol{\theta}_\sigma$, $\boldsymbol{\theta}_{\tilde{\nu}}$, and $\boldsymbol{\theta}_\delta$ (Section 4.3.3). All parameters are assumed to be *a priori* independent, unless explicitly stated otherwise.

4.4.2 Overview of the MCMC Sampler

For posterior inference, we will employ a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm (Green, 1995) based on a Gibbs sampler (Geman and Geman, 1984) with some Metropolis-Hastings (MH) steps (Metropolis et al., 1953; Hastings, 1970). In what is to follow, $[A]$ will denote the distribution of a random variable A, $[A|B]$ will denote the conditional distribution of A given B, and $[A|\cdot]$ will denote the full conditional distribution of A, which is defined as the conditional distribution of A given the data and given

all other unknowns in the model. Sometimes we will emphasize dependence of a matrix on a set of parameters by placing the parameters in parentheses; for example, the matrix of basis functions depends on $\boldsymbol{\theta}_{\tilde{\nu}}$, and so we sometimes write, $\mathbf{B}(\boldsymbol{\theta}_{\tilde{\nu}})$, for clarity.

Let $\Omega := \{\boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{K}, \mathcal{C}, \boldsymbol{\theta}_{\sigma}, \boldsymbol{\theta}_{\tilde{\nu}}, \boldsymbol{\theta}_{\delta}, \boldsymbol{\theta}_{\epsilon}\}$ be the set of the unknown, random quantities in the model (except for $\delta(\cdot)$; see below). The joint distribution of the data, $\mathbf{Z}$, and the unknowns, Ω, is given by,

$$[\mathbf{Z}, \Omega] = [\mathbf{Z}|\Omega][\Omega] = [\mathbf{Z}|\Omega][\boldsymbol{\beta}][\boldsymbol{\eta}|\mathbf{K}][\mathbf{K}|\boldsymbol{\theta}_{\sigma}, \boldsymbol{\theta}_{\tilde{\nu}}, \mathcal{C}][\mathcal{C}][\boldsymbol{\theta}_{\sigma}][\boldsymbol{\theta}_{\tilde{\nu}}][\boldsymbol{\theta}_{\delta}][\boldsymbol{\theta}_{\epsilon}], \tag{4.22}$$

and all distributions on the right-hand side are described or referenced in the previous subsection. The full conditional distributions needed for the updates in the Gibbs sampler are proportional to the joint distribution in (4.22).

Due to the large number of variables in Ω, it is imperative to ensure good mixing of the MCMC. We will "integrate out" (equivalently, "marginalize over") quantities where possible and where it is computationally feasible, according to the general recipe provided by van Dyk and Park (2008). Simply speaking, if we have two (generic) random variables A and B, and data $\mathbf{Z}$, mixing of the MCMC can often be improved if we replace the standard Gibbs update of A from $[A|B, \mathbf{Z}]$, by instead sampling from $[A|\mathbf{Z}]$ (i.e., we have integrated out B here). Convergence of the resulting MCMC to the joint posterior distribution, $[A, B|\mathbf{Z}]$, is still guaranteed as long as we do not condition on B in any subsequent update during the same MCMC iteration before we have updated B from $[B|A, \mathbf{Z}]$.

In our case, notice that $\delta(\cdot)$ is not included in Ω. Hence, we will integrate $\delta(\cdot)$ out of *all* updates in the main MCMC sampler, which improves mixing. Inference on $\delta(\cdot)$ is described separately in Section 4.4.4. For certain updates, we will also integrate out other variables (e.g., $\boldsymbol{\eta}$), as indicated below.

The MCMC sampler consists of the following steps:

1. Sample $\boldsymbol{\beta}$ from its full conditional distribution,

$$[\boldsymbol{\beta}|\cdot] = N_p((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}+\sigma_\beta^{-2}\mathbf{I}_p)^{-1}(\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Z}-\mathbf{B}\boldsymbol{\eta})+\sigma_\beta^{-2}\boldsymbol{\mu}_{\boldsymbol{\beta}}), (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}+\sigma_\beta^{-2}\mathbf{I}_p)^{-1})$$

2. Sample $\boldsymbol{\theta}_\sigma$ using an MH step from,

$$[\boldsymbol{\theta}_\sigma|\{\mathbf{Z},\Omega\}\backslash\{\boldsymbol{\theta}_\sigma,\boldsymbol{\eta}\}] \propto [\boldsymbol{\theta}_\sigma]\, IW_r(\mathbf{K}|\mathbf{M}(\boldsymbol{\theta}_\sigma),u)\, N_n(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta},\mathbf{BKB}'+\mathbf{V}(\boldsymbol{\theta}_\sigma))$$

3. Sample $\boldsymbol{\theta}_{\tilde{\nu}}$ using an MH step from,

$$[\boldsymbol{\theta}_{\tilde{\nu}}|\{\mathbf{Z},\Omega\}\backslash\{\boldsymbol{\theta}_{\tilde{\nu}},\boldsymbol{\eta}\}] \propto [\boldsymbol{\theta}_{\tilde{\nu}}]\, IW_r(\mathbf{K}|\mathbf{M}(\boldsymbol{\theta}_{\tilde{\nu}}),u)\, N_n(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta},\mathbf{B}(\boldsymbol{\theta}_{\tilde{\nu}})\mathbf{KB}(\boldsymbol{\theta}_{\tilde{\nu}})'+\mathbf{V}(\boldsymbol{\theta}_{\tilde{\nu}})).$$

4. Sample $\boldsymbol{\theta}_\delta$ using an MH step from,

$$[\boldsymbol{\theta}_\delta|\{\mathbf{Z},\Omega\}\backslash\{\boldsymbol{\theta}_\delta,\boldsymbol{\eta}\}] \propto [\boldsymbol{\theta}_\delta]\, N_n(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta},\mathbf{BKB}'+\mathbf{V}(\boldsymbol{\theta}_\delta)).$$

5. Sometimes $\boldsymbol{\theta}_\epsilon$ is known from measurement calibrations. If it is assumed random, sample $\boldsymbol{\theta}_\epsilon$ using an MH step from,

$$[\boldsymbol{\theta}_\epsilon|\{\mathbf{Z},\Omega\}\backslash\{\boldsymbol{\theta}_\epsilon,\boldsymbol{\eta}\}] \propto [\boldsymbol{\theta}_\epsilon]\, N_n(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta},\mathbf{BKB}'+\mathbf{V}(\boldsymbol{\theta}_\epsilon)).$$

6. Sample $\mathcal{C}$ and $\boldsymbol{\eta}$ jointly from $[\mathcal{C},\boldsymbol{\eta}|\{\mathbf{Z},\Omega\}\backslash\{\mathcal{C},\boldsymbol{\eta},\mathbf{K}\}]$, as described below in Section 4.4.3.

7. Sample $\mathbf{K}$ from its full conditional distribution,

$$[\mathbf{K}|\cdot] = IW_r(\mathbf{M}+\boldsymbol{\eta}\boldsymbol{\eta}', u+1), \tag{4.23}$$

with mean $(\mathbf{M}+\boldsymbol{\eta}\boldsymbol{\eta}')/u$. For example, for $u=2$, we have $E(\mathbf{K}|\cdot) = (\mathbf{D}_\nu\mathbf{R}^{-1}\mathbf{D}_\nu + \boldsymbol{\eta}\boldsymbol{\eta}')/2$.

8. Sample $\boldsymbol{\eta}$ from its full conditional distribution,

$$[\boldsymbol{\eta}|\cdot] = N_r((\mathbf{B}'\mathbf{V}^{-1}\mathbf{B} + \mathbf{K}^{-1})^{-1}\mathbf{B}'\mathbf{V}^{-1}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta}), (\mathbf{B}'\mathbf{V}^{-1}\mathbf{B} + \mathbf{K}^{-1})^{-1}). \quad (4.24)$$

If the elements of Ω are updated in this order, convergence of the MCMC to the joint posterior distribution, $[\Omega|\mathbf{Z}]$, is guaranteed (see van Dyk and Park, 2008, for more details).

4.4.3 Details on Sampling the Basis-Function Centers

The update for $\mathcal{C}$ is complicated due to the possible change of r, the reduced dimension; the dimensions of $\boldsymbol{\eta}$ and $\mathbf{K}$ depend on r. We can integrate out one of the two, but not both. Some experimenting has shown that higher acceptance rates can be achieved by integrating out the higher-dimensional parameter of the two (i.e., $\mathbf{K}$), and so we will produce a joint reversible jump proposal for $\mathcal{C}$ and $\boldsymbol{\eta}$.

We begin by proposing a change to the set $\mathcal{C}$, which indexes all basis functions currently in the model by their centers. We propose to do one of three possible actions, each with probability $1/3$: add a basis function, delete one of the basis functions, or move one of the basis functions.

- To add a basis function, we draw a new center, $\mathbf{c}_{r+1}$, from a uniform distribution on $\mathcal{D}$, and let $\mathcal{C}^* := \mathcal{C} \cup \{\mathbf{c}_{r+1}\}$ be the proposed set of basis-function centers, which now has size $r^* = r + 1$.

- If we want instead to delete a basis function, we will select one uniformly at random from the existing ones; that is, we draw $J \sim U(1, 2, \ldots, r)$, and then we set $\mathcal{C}^* := \mathcal{C}\backslash\{\mathbf{c}_J\}$ to be the proposed set of centers, with size $r^* = r - 1$.

- Moving a basis function is essentially a combination of a deletion and an addition: We first select a basis function uniformly at random to be deleted (moved), and then

we select a location uniformly on $\mathcal{D}$ at which to add a new one (i.e., move the old one). This also results in a new set of basis-function centers, $\mathcal{C}^*$, of size $r^* = r$.

If the domain $\mathcal{D}$ is a unit sphere, a new center location can be drawn uniformly on the sphere by setting longitude $= X_1$ and latitude $= \cos^{-1}(X_2) - \pi/2$, where $X_1 \sim U(-\pi, \pi)$ and $X_2 \sim U(-1, 1)$. Due to the prior distribution of $\mathcal{C}$ (see (4.13) in Section 4.2.4), the acceptance probability will always be zero if the proposal is to add or move a basis function, and the new location is within distance ξ of one of the locations of one of the current basis-function centers. In this case, we reject the proposal immediately and continue on with step 7 of the MCMC sampler of Section 4.4.2.

We now have to find a good proposal $\boldsymbol{\eta}^*$ conditional on the new set of centers $\mathcal{C}^*$. We denote the proposal distribution for a generic set of centers, $\mathcal{C}$, by $\mathcal{Q}_{\boldsymbol{\eta}}(\mathcal{C})$ and its PDF evaluated at $\boldsymbol{\eta}$ by $\mathcal{Q}_{\boldsymbol{\eta}}(\boldsymbol{\eta}|\mathcal{C})$. We would like to use $[\boldsymbol{\eta}|\{\mathbf{Z}, \Omega\}\backslash\{\boldsymbol{\eta}, \mathbf{K}\}]$ for $\mathcal{Q}_{\boldsymbol{\eta}}(\boldsymbol{\eta}|\mathcal{C})$ (see discussion at the end of this subsection), but we cannot sample from this distribution directly. Instead, we draw the proposal, $\boldsymbol{\eta}^*$, from,

$$\mathcal{Q}_{\boldsymbol{\eta}}(\mathcal{C}^*) := N_{r^*}(\mathbf{A}(\mathcal{C}^*)^{-1}\mathbf{B}(\mathcal{C}^*)'\mathbf{V}(\mathcal{C}^*)^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}), \mathbf{A}(\mathcal{C}^*)^{-1}), \tag{4.25}$$

where $\mathbf{A}(\mathcal{C}^*) = \mathbf{B}(\mathcal{C}^*)'\mathbf{V}(\mathcal{C}^*)^{-1}\mathbf{B}(\mathcal{C}^*) + \mathbf{K}_{\mathcal{C}^*}^{-1}$. The distribution $\mathcal{Q}_{\boldsymbol{\eta}}(\mathcal{C}^*)$ in (4.25) is similar to the full conditional distribution of $\boldsymbol{\eta}$ given by (4.24), but where $\mathbf{K}$ is replaced with a value motivated by the mean of (4.23),

$$\mathbf{K}_{\mathcal{C}^*} := E(\mathbf{K}|\mathbf{M}(\mathcal{C}^*), \boldsymbol{\eta} = \boldsymbol{\eta}_{\mathcal{C}^*}) = (\mathbf{M}(\mathcal{C}^*) + \boldsymbol{\eta}_{\mathcal{C}^*}\boldsymbol{\eta}'_{\mathcal{C}^*})/u,$$

and

$$\begin{aligned}\boldsymbol{\eta}_{\mathcal{C}^*} &:= E(\boldsymbol{\eta}|\mathbf{Z}, \mathbf{B}(\mathcal{C}^*), \mathbf{V}(\mathcal{C}^*), \mathbf{K} = \mathbf{M}(\mathcal{C}^*)) \\ &= (\mathbf{B}(\mathcal{C}^*)'\mathbf{V}(\mathcal{C}^*)^{-1}\mathbf{B}(\mathcal{C}^*) + \mathbf{M}(\mathcal{C}^*)^{-1})^{-1}\mathbf{B}(\mathcal{C}^*)'\mathbf{V}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}).\end{aligned}$$

The acceptance probability for the proposed pair $(\mathcal{C}^*, \boldsymbol{\eta}^*)$ is determined by the product of the likelihood ratio, the prior ratio, the proposal ratio, and a Jacobian that is equal to one here and can hence be ignored (for details, see Green, 1995). Specifically, the acceptance probability is given by, $\min\{1, \alpha\}$, where

$$\alpha := \frac{N_n(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta}, \mathbf{B}(\mathcal{C}^*)\boldsymbol{\eta}^*, \mathbf{V}(\mathcal{C}^*))}{N_n(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta}, \mathbf{B}(\mathcal{C})\boldsymbol{\eta}, \mathbf{V}(\mathcal{C}))} \frac{[\boldsymbol{\eta}^*|\mathcal{C}^*, \boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}]}{[\boldsymbol{\eta}|\mathcal{C}, \boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}]} \frac{[\mathcal{C}^*]}{[\mathcal{C}]} \frac{\mathcal{Q}(\{\mathcal{C}^*, \boldsymbol{\eta}^*\}, \{\mathcal{C}, \boldsymbol{\eta}\})}{\mathcal{Q}(\{\mathcal{C}, \boldsymbol{\eta}\}, \{\mathcal{C}^*, \boldsymbol{\eta}^*\})}. \tag{4.26}$$

Since we have already rejected a proposed set of centers, $\mathcal{C}^*$ if any pair of centers in the set has a distance of less than or equal to ξ (due to (4.13)), the ratio of the PDFs of the prior distributions of $\mathcal{C}^*$ and $\mathcal{C}$ is equal to one, and so $[\mathcal{C}^*]/[\mathcal{C}]$ can be ignored in (4.26). The prior distribution of $\boldsymbol{\eta}$ given $\mathcal{C}$ (and $\boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}$) in (4.26) is given generically by,

$$\begin{aligned}[\boldsymbol{\eta}|\mathcal{C}, \boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}] &= \int N_r(\boldsymbol{\eta}|\mathbf{0}, \mathbf{K}) IW_r(\mathbf{K}|\mathbf{M}(\mathcal{C}), u) d\mathbf{K} \\ &= \frac{1}{\pi^{r/2}} \frac{|\mathbf{M}(\mathcal{C})|^{(r+u)/2}}{|\mathbf{M}(\mathcal{C}) + \boldsymbol{\eta}\boldsymbol{\eta}'|^{(r+u+1)/2}} \frac{\Gamma_r((r+u+1)/2)}{\Gamma_r((r+u)/2)},\end{aligned}$$

where $\mathbf{M}(\mathcal{C})$ is introduced below (4.12). The function $\mathcal{Q}(\{\mathcal{C}, \boldsymbol{\eta}\}, \{\mathcal{C}^*, \boldsymbol{\eta}^*\})$ describes the probability of proposing a move from $\{\mathcal{C}, \boldsymbol{\eta}\}$ to $\{\mathcal{C}^*, \boldsymbol{\eta}^*\}$. It can be decomposed as,

$$\frac{\mathcal{Q}(\{\mathcal{C}^*, \boldsymbol{\eta}^*\}, \{\mathcal{C}, \boldsymbol{\eta}\})}{\mathcal{Q}(\{\mathcal{C}, \boldsymbol{\eta}\}, \{\mathcal{C}^*, \boldsymbol{\eta}^*\})} = \frac{\mathcal{Q}_\mathcal{C}(\mathcal{C}^*, \mathcal{C})}{\mathcal{Q}_\mathcal{C}(\mathcal{C}, \mathcal{C}^*)} \frac{\mathcal{Q}_{\boldsymbol{\eta}}(\boldsymbol{\eta}|\mathcal{C})}{\mathcal{Q}_{\boldsymbol{\eta}}(\boldsymbol{\eta}^*|\mathcal{C}^*)},$$

where $\mathcal{Q}_{\boldsymbol{\eta}}(\boldsymbol{\eta}^*|\mathcal{C}^*)$ is the PDF corresponding to the distribution in (4.25) evaluated at $\boldsymbol{\eta}^*$, and $\mathcal{Q}_{\boldsymbol{\eta}}(\boldsymbol{\eta}|\mathcal{C})$ is defined analogously. A proposed addition of a basis function results in, $\mathcal{Q}_\mathcal{C}(\mathcal{C}^*, \mathcal{C})/\mathcal{Q}_\mathcal{C}(\mathcal{C}, \mathcal{C}^*) = 1/(r+1)$, a proposed deletion results in $\mathcal{Q}_\mathcal{C}(\mathcal{C}^*, \mathcal{C})/\mathcal{Q}_\mathcal{C}(\mathcal{C}, \mathcal{C}^*) = r$, and a proposed move results in, $\mathcal{Q}_\mathcal{C}(\mathcal{C}^*, \mathcal{C})/\mathcal{Q}_\mathcal{C}(\mathcal{C}, \mathcal{C}^*) = 1$.

Note that for $r = 0$, deleting or moving a basis function is impossible, and so in this case we always propose to add a basis function (i.e., $r^* = 1$ with probability one). As a result, we have to adjust α in (4.26) slightly by dividing it by 3 when $r = 0$.

Finally, there might be a concern that for very large datasets, the data might always favor a very large set of centers if there is no strong penalization for large r through the prior distribution on $\mathcal{C}$. If the values of r in the MCMC sampler were frequently close to n, we would, of course, lose all computational advantages over traditional spatial models. However, this concern is unnecessary. To see this, we will assume, temporarily for the rest of this subsection, that the spatial trend, $\mu(\cdot)$, and the covariance functions of the FSV component ($C_\delta(\cdot,\cdot)$ in (4.5)) and the parent process ($C_{\tilde{\nu}}(\cdot,\cdot)$ in (4.6)) are fixed, and so all derivations are implicitely assumed to be conditional on $\boldsymbol{\beta}$, $\boldsymbol{\theta}_\sigma$, $\boldsymbol{\theta}_{\tilde{\nu}}$ and $\boldsymbol{\theta}_\delta$. Following the derivations in Holmes and Mallick (2000, App. I), we note that Bayes' Theorem gives,

$$[\boldsymbol{\eta}|\mathbf{Z},\mathcal{C}] = [\mathbf{Z}|\boldsymbol{\eta},\mathcal{C}][\boldsymbol{\eta}|\mathcal{C}]/[\mathbf{Z}|\mathcal{C}],$$

and so we can write α in (4.26) as,

$$\begin{aligned}\alpha &= \frac{[\mathbf{Z}|\mathcal{C}^*,\boldsymbol{\eta}^*]}{[\mathbf{Z}|\mathcal{C},\boldsymbol{\eta}]}\frac{[\boldsymbol{\eta}^*|\mathcal{C}^*]}{[\boldsymbol{\eta}|\mathcal{C}]}\frac{[\mathcal{C}^*]}{[\mathcal{C}]}\frac{\mathcal{Q}_\mathcal{C}(\mathcal{C}^*,\mathcal{C})}{\mathcal{Q}_\mathcal{C}(\mathcal{C},\mathcal{C}^*)}\frac{\mathcal{Q}_{\boldsymbol{\eta}}(\boldsymbol{\eta}|\mathcal{C})}{\mathcal{Q}_{\boldsymbol{\eta}}(\boldsymbol{\eta}^*|\mathcal{C}^*)}\frac{[\boldsymbol{\eta}|\mathbf{Z},\mathcal{C}]}{[\boldsymbol{\eta}^*|\mathbf{Z},\mathcal{C}^*]}\frac{[\boldsymbol{\eta}^*|\mathbf{Z},\mathcal{C}^*]}{[\boldsymbol{\eta}|\mathbf{Z},\mathcal{C}]} \\ &= \frac{[\mathbf{Z}|\mathcal{C}^*]}{[\mathbf{Z}|\mathcal{C}]}\frac{[\mathcal{C}^*]}{[\mathcal{C}]}\frac{\mathcal{Q}_\mathcal{C}(\mathcal{C}^*,\mathcal{C})}{\mathcal{Q}_\mathcal{C}(\mathcal{C},\mathcal{C}^*)}\left(\frac{[\boldsymbol{\eta}^*|\mathbf{Z},\mathcal{C}^*]}{[\boldsymbol{\eta}|\mathbf{Z},\mathcal{C}]}\frac{\mathcal{Q}_{\boldsymbol{\eta}}(\boldsymbol{\eta}|\mathcal{C})}{\mathcal{Q}_{\boldsymbol{\eta}}(\boldsymbol{\eta}^*|\mathcal{C}^*)}\right),\end{aligned}$$

where the last term in the parentheses can be viewed as the acceptance probability of a Metropolis-Hastings proposal $\boldsymbol{\eta}^*$ when the proposal distribution has PDF $\mathcal{Q}_{\boldsymbol{\eta}}(\boldsymbol{\eta}^*|\mathcal{C}^*)$ and the target PDF is $[\boldsymbol{\eta}^*|\mathbf{Z},\mathcal{C}^*]$. This term would therefore be equal to one if $\mathcal{Q}_{\boldsymbol{\eta}}(\boldsymbol{\eta}^*|\mathcal{C}^*) = [\boldsymbol{\eta}^*|\mathbf{Z},\mathcal{C}^*]$, but this distribution is not available in closed form (as discussed above (4.25)). Regardless, we see that the acceptance probability for a proposed set of centers, $\mathcal{C}$, can be written as the product of the Bayes factor of $\mathcal{C}^*$ versus $\mathcal{C}$, the ratio of prior probabilities of $\mathcal{C}^*$ and $\mathcal{C}$, and terms depending on the proposal distributions chosen for $\mathcal{C}^*$ and $\boldsymbol{\eta}^*$. This is reassuring, as "the Bayes factor functions as a fully automatic Occam's razor" (Kass and Raftery, 1995, p. 790), and so there is intuition that strong penalization through the prior on $\mathcal{C}$ is not necessary to keep r from becoming too large.

4.4.4 Spatial Prediction

In spatial statistics, the main interest is often in predicting the process $Y(\cdot)$ at a set of prediction locations, $\{\mathbf{s}_1^P, \ldots, \mathbf{s}_{n_P}^P\}$, which might or might not include the set of observed locations. Often the set of prediction locations is a fine grid over the domain of interest, $\mathcal{D}$. Let $\mathbf{Y}^P := [Y(\mathbf{s}_1^P), \ldots, Y(\mathbf{s}_{n_P}^P)]'$ denote a vector containing the true process evaluated at all locations of interest. This vector can be written as,

$$\mathbf{Y}^P = \mathbf{X}^P\boldsymbol{\beta} + \mathbf{B}^P\boldsymbol{\eta} + \boldsymbol{\delta}^P, \tag{4.27}$$

where the superscript P generically denotes evaluation of a process at the set of prediction locations. In the previous section, we have described how to obtain MCMC samples from the posterior distributions of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\mathcal{C}$, $\boldsymbol{\theta}_\sigma$, and $\boldsymbol{\theta}_{\tilde{\nu}}$, from which it is straightforward to obtain samples of $\mathbf{X}^P\boldsymbol{\beta} + \mathbf{B}^P(\mathcal{C}, \boldsymbol{\theta}_{\tilde{\nu}})\,\boldsymbol{\eta}$, the first two terms of (4.27). Therefore, posterior samples of $\boldsymbol{\delta}^P$ are all we need to make inference on $\mathbf{Y}^P$. Sampling the potentially very large vector, $\boldsymbol{\delta}^P$, is computationally expensive, and so we do so only for thinned samples after convergence of the main MCMC of Section 4.4.2. It is possible to do this without jeopardizing the convergence of the MCMC to the correct joint posterior distribution, because we have integrated out $\delta(\cdot)$ from all MCMC updates in Section 4.4.2, and so none of the updates there depend on $\delta(\cdot)$. Specifically, we are interested in,

$$[\Omega, \boldsymbol{\delta}^P|\mathbf{Z}] = [\Omega|\mathbf{Z}]\,[\boldsymbol{\delta}^P|\Omega, \mathbf{Z}],$$

where samples of the first term on the right-hand side were obtained in Section 4.4.2, and samples of the second term are obtained here, but only for the thinned Markov chain.

Assume now that, after appropriate reordering, we can write $\boldsymbol{\delta}^P = [\boldsymbol{\delta}', \boldsymbol{\delta}^{U\prime}]'$, where $\boldsymbol{\delta}$ is obtained by evaluating $\delta(\cdot)$ at all observed locations, and $\boldsymbol{\delta}^U$ represents $\delta(\cdot)$ at all

unobserved locations at which predictions are of interest. If we define $\mathbf{V}_\delta^P := var(\boldsymbol{\delta}^P)$ and $\mathbf{V}_\delta^{P,O} := cov(\boldsymbol{\delta}^P, \boldsymbol{\delta})$, the full conditional distribution of $\boldsymbol{\delta}^P$ is given by,

$$[\boldsymbol{\delta}^P | \cdot] = N_{n_P}(\mathbf{V}_\delta^{P,O}\mathbf{V}^{-1}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\boldsymbol{\eta}), \mathbf{V}_\delta^P - \mathbf{V}_\delta^{P,O}\mathbf{V}^{-1}\mathbf{V}_\delta^{P,O\prime}), \tag{4.28}$$

where recall that $\mathbf{V} = \mathbf{V}_\delta + \mathbf{V}_\epsilon$ (as defined in Section 4.4.1). To avoid having to obtain $\mathbf{V}_\delta^{P,O}\mathbf{V}^{-1}\mathbf{V}_\delta^{P,O\prime}$ explicitly, we calculate instead the quantity,

$$\tilde{\boldsymbol{\delta}}^P + \mathbf{V}_\delta^{P,O}\mathbf{V}^{-1}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\boldsymbol{\eta} - \tilde{\boldsymbol{\delta}} - \tilde{\boldsymbol{\epsilon}}),$$

which has the distribution given by (4.28) if we set $\tilde{\boldsymbol{\delta}}^P := (\tilde{\boldsymbol{\delta}}', \tilde{\boldsymbol{\delta}}^{U\prime})' = (\mathbf{V}_\delta^P)^{1/2}\mathbf{W}_1$ and $\tilde{\boldsymbol{\epsilon}} := \mathbf{V}_\epsilon^{1/2}\mathbf{W}_2$, where $\mathbf{W}_1 \sim N_{n_P}(\mathbf{0}, \mathbf{I}_{n_P})$ and $\mathbf{W}_2 \sim N_n(\mathbf{0}, \mathbf{I}_n)$, independently. This sampling technique is essentially what is known as conditional simulation in spatial statistics (e.g., Cressie, 1993, Sect. 3.6.2).

4.4.5 Computational Issues

Let $\boldsymbol{\Sigma}_Z := var(\mathbf{Z}|\boldsymbol{\beta}, \mathbf{K}, \mathcal{C}, \boldsymbol{\theta}_\sigma, \boldsymbol{\theta}_{\tilde{\nu}}, \boldsymbol{\theta}_\delta, \boldsymbol{\theta}_\epsilon) = \mathbf{B}\mathbf{K}\mathbf{B}' + \mathbf{V}$, which is a dense (i.e., non-sparse) $n \times n$ matrix. Many of the MCMC updates described above require $\boldsymbol{\Sigma}_Z^{-1}$, but naive calculation of this inverse is impossible or at least computationally infeasible for large n. Due to the dimension reduction inherent in our model, we can write the inverse as (Sherman and Morrison, 1950; Woodbury, 1950; Henderson and Searle, 1981),

$$\boldsymbol{\Sigma}_Z^{-1} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{B}(\mathbf{K}^{-1} + \mathbf{B}'\mathbf{V}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{V}^{-1},$$

and a similar formula (e.g., Cressie and Johannesson, 2008) gives,

$$|\boldsymbol{\Sigma}_Z| = |\mathbf{V}||\mathbf{I}_r + \mathbf{K}\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}|.$$

Calculating the inverse and determinant of the dense $n \times n$ matrix $\boldsymbol{\Sigma}_Z$ can therefore be reduced to calculating the inverse and determinants of $r \times r$ matrices and of the sparse

$n \times n$ matrix $\mathbf{V}$. This idea was proposed in an empirical-Bayes setting by Cressie and Johannesson (2006). Close examination of the MCMC updating steps in Section 4.4.2 reveals that the only calculations involving the inverse of $\mathbf{V}$ are $\mathbf{V}^{-1/2}\mathbf{X}$, $\mathbf{V}^{-1/2}\mathbf{B}$, and $\mathbf{V}^{-1/2}\mathbf{z}$, where $\mathbf{V}^{1/2}$ is the (lower-triangular) Cholesky factor of $\mathbf{V}$ (see Stein, 2008). For this reason, when $\mathbf{V}$ is sparse (which it is when defined via a tapered covariance function), the MCMC updates can be carried out quite rapidly, as follows.

Finding the Cholesky decomposition of this sparse matrix and solving the systems of linear equations must be done at each iteration of the MCMC sampler. Fortunately, these tasks can be carried out significantly faster if $\mathbf{V}$ is ordered in a way that results in a sparse Cholesky factor (e.g., Furrer et al., 2006). Also, since our tapering range is fixed, the sparsity structure (i.e., the position of the nonzero elements) of $\mathbf{V}$ is the same for all MCMC iterations. Hence, we can find an efficient ordering (e.g., the minimum-degree ordering) once, at the beginning of the algorithm, and then we can use that ordering when computing the Cholesky decompositions at each MCMC iteration. To sample the vector $\boldsymbol{\delta}^P$ for thinned iterations of the main MCMC, we can again find an efficient ordering, this time for the set of all locations (observed and unobserved).

Cressie and Johannesson (2008) considered a model that is similar to ours, but in their model the matrix $\mathbf{V}$ is diagonal. The required number of computations for inference using their model is linear in the number of observations, n. Here, because we assume $\mathbf{V}$ to be sparse, not diagonal, we cannot achieve this same theoretical computational complexity. In general, the number of computations required for operations involving a sparse matrix is proportional to the number of nonzero elements of that matrix (Gilbert et al., 1992). It is difficult to make general statements about the computational complexity of the Cholesky decomposition of a sparse matrix, because it depends on the bandwidth of the matrix (i.e.,

the largest distance of a nonzero element to the diagonal) and the locations of the nonzero elements. However, Furrer et al. (2006) have provided some numerical results on a fixed domain, with fixed tapering length, and using a regular sampling grid; the results indicate that the time required to compute the Cholesky decomposition of a tapered $n \times n$ covariance matrix increases roughly linearly with n, which in turn indicates that the computational complexity of our algorithm is approximately of order n. In fact, we have considerable control over the speed of the MCMC algorithm through selection of the tapering range, L_δ, of the FSV component. For extremely massive datasets, we can set L_δ to a very small value (maybe even zero), to achieve computational feasibility.

Theoretical-computational-complexity issues aside, in our experience the majority of computation time at each of our MCMC iterations was actually not spent on Cholesky decompositions, but on simply creating the matrices V_δ and $\mathbf{B}$. For the nonzero elements of these two matrices (more precisely, for V_δ we only need the nonzero upper-triangular elements), we have to evaluate a Matérn function (4.14). This involves evaluation of the modified Bessel function, which is rather slow, because no closed-form solution is available. But again, in situations in which rapid computation is crucial, we can achieve faster computation by setting L_δ and L_ν to small values.

4.5 Simulation Study in One Spatial Dimension

In this Section, we compare our model to a "baseline" model and to two other models conceived for very large spatial datasets, in three simulation studies. The models are:

SRE: Our SRE model described in the previous subsections, a combination of an SBF component with random basis functions and an FSV component with tapered spatial dependence.

SMC: A stationary Matérn covariance model (4.14), which is infeasible for large datasets due to its computational complexity of order n^3.

CTO: A covariance-tapering-only model, which is essentially the same model as the SRE model, except that the SBF term is not part of the model (i.e., $\nu(\cdot) \equiv 0$).

KCG: The Kang and Cressie (2011) SRE model with a multi-resolutional Givens-angle prior for $\mathbf{K}$. The FSV component is modeled to be spatially independent conditional on a constant variance.

The results presented here are preliminary, since a full study would involve identifying the important factors, their levels, and performing a (partial) factorial experiment.

The true process is is taken to be one-dimensional with domain $\mathcal{D} = \{1, 2, \ldots, 256\}$. In Simulation Study 1, it has a nonstandard and nonstationary correlation function. The covariance function is the product of a wave correlation function and a nonstationary Matérn covariance of the form introduced in Section 4.3. Defining the wave correlation function as,

$$\mathcal{W}(h) = \sin(h)/h, \ h > 0,$$

and $\mathcal{W}(0) = 1$, the true covariance function for Simulation Study 1 is given by,

$$C(s_1, s_2) = \sigma(s_1)\sigma(s_2)\frac{\gamma(s_1)^{1/4}\gamma(s_2)^{1/4}}{\gamma(s_1, s_2)^{1/2}}\mathcal{M}_{\upsilon(s_1,s_2)}\left(\frac{|s_1 - s_2|}{\gamma(s_1, s_2)^{1/2}}\right)\mathcal{W}\left(\frac{|s_1 - s_2|}{10}\right), \tag{4.29}$$

for $s_1, s_2 \in \{1, 2, \ldots, 256\}$, and where $\gamma(s_1, s_2) := (\gamma(s_1) + \gamma(s_2))/2$ and $\upsilon(s_1, s_2) := (\upsilon(s_1) + \upsilon(s_2))/2$. The three parameters of the Matérn covariance in the first simulation

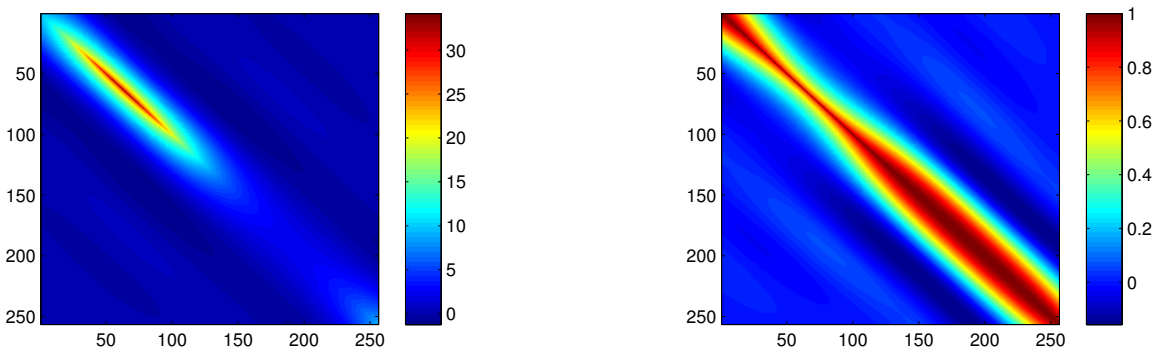

Figure 4.3: The true covariance structure (left panel) and correlation structure (right panel) given by (4.29) and assumed in Simulation Study 1.

vary spatially as follows:

$$\begin{aligned} \sigma(s) &= 3\exp\big(\sin(s\,2\pi/256)/1.5\big) \\ \gamma(s) &= 9000\exp\big(-2\sin(s\,2\pi/256)\big) \\ \upsilon(s) &= 0.5\Phi\big(-0.8\sin(s\,2\pi/256)\big). \end{aligned} \tag{4.30}$$

The resulting covariance and correlation structure are shown in the left and the right panel, respectively, of Figure 4.5. The corresponding true process has high variance and quickly decaying dependence around $s = 70$, and so it will be very "rough" in that area. On the other hand, around $s = 180$, the variance is much smaller, and the process should be rather smooth. One sample of $Y(\cdot)$ with covariance function (4.29) is shown in Figure 4.4 (in blue). Part of our more complete simulation experiment will be to explore the combinations of low variance and quickly decaying dependence, and large variance and long correlation ranges, as part of the factor space.

For the SRE and TCO models, the tapering lengths were chosen such that their size relative to the size of the domain $\mathcal{D}$ was similar to the relative size of these parameters in the analysis of global data in Section 4.6 below. In Section 4.6, we chose $L_\delta = 0.08$,

which corresponded to $0.08/\pi \approx 2.55\%$ as a ratio of the maximum distance on the globe (assuming unit radius). Here, the maximum distance on the domain is given by 255, and so we set $L_\delta = 6.5 \approx (2.55\%)(255)$. To find a good value for L_ν, note that on a unit sphere, the ratio of the radius (determined to be the maximum value for L_ν in Section 4.3.2) to the maximum distance two points can have, is $1/\pi$, and so we set $L_\nu = 80 \approx (1/\pi)255$.

For the SRE and CTO models, we took $\mathbf{b}_\theta(\cdot)$ to be made up of two bisquare functions with radius 64, centered at locations 64 and 196, respectively. We chose $u = 4$ for the SRE model. To determine a comparable number of basis functions for the KCG model, we did a pilot study using the SRE model that showed that the posterior probability of $r > 11$ was negligible, and so we used eleven bisquare basis functions of two resolutions for the KCG model: two functions had radius 192 and were centered at locations 64 and 192, and nine bisquare functions had radius 48 and were centered at 0, 32, 64, 96, 128, 160, 192, 224, and 256. Part of a more complete simulation experiment will be to assess the effect of adding more basis functions to the KCG model, to see how flexible it *can* be, at the price of higher model complexity and longer computation times.

In general, all prior distributions for the remaining parameters in all four models were chosen to be fairly vague but centered at values that resulted in model covariances that were as close as possible to the true covariance. Notation for the hyperparameters specified here was introduced in Section 4.3.3. The hyperparameters for $\sigma(\cdot)$ were chosen as $\mu_\sigma = \log(3)$ and $\sigma^2_\sigma = \log(6^2)$ for both the SRE model and the CTO model. The hyperparameters for the scaling parameter of the SBF component of the SRE model were $\mu_{\gamma_\delta} = \log(10L_\nu)$ and $\sigma^2_{\gamma_\delta} = \log((L_\nu - L_\delta)^2)$. The scaling parameter of the FSV component in the CTO model was allowed to vary. We centered it at the same value at which the parameter was fixed for the CBF model, $\mu_{\gamma_\delta} = \log(2L_\delta)$, and the variance was taken to be, $\sigma^2_{\gamma_\delta} = \log((2L_\delta)^2)$. The

SMC model had the same hyperparameters for σ, and for the scaling parameter we chose $\mu_\gamma = \log(640)$ and $\sigma_\gamma^2 = \log(640^2)$. The hyperparameters for the eigenvalues and Givens angles of $\mathbf{K}$ for the KCG model were calibrated as described in Kang and Cressie (2011) from $\mathbf{K}_{\text{F}}$. They used binned-method-of-moments parameter estimates for calibration, but here we chose $\mathbf{K}_{\text{F}}$ to be the value of $\mathbf{K}$ that miminized $\|\mathbf{B}_{\text{KCG}}^P \mathbf{K} \mathbf{B}_{\text{KCG}}^{P\prime} - \mathbf{\Sigma}_Y\|_F$, where $\mathbf{B}_{\text{KCG}}^P$ was a matrix obtained by evaluating the eleven bisquare functions chosen for the KCG model at all 256 locations in the domain, $\mathbf{\Sigma}_Y$ was the true covariance matrix shown in the left panel of Figure 4.5, and $\|\cdot\|_F$ is the Frobenius norm. The prior distribution of the FSV variance of the KCG model was assumed to be of the form of the FSV variance of our SRE model as described in Section 4.3.3 (i.e., it was different than in Kang and Cressie, 2011), with $\mu_\sigma = \log\sqrt{\sigma_{\text{F}}^2}$ and $\sigma_\sigma^2 = \mu_\sigma^2$, where $\sigma_{\text{F}}^2 = avg\{|diag(\mathbf{B}_{\text{KCG}}^P \mathbf{K}_{\text{F}} \mathbf{B}_{\text{KCG}}^{P\prime} - \mathbf{\Sigma}_Y)|\}$.

The simulation study consisted of 100 iterations. This number will be increased in a future simulation experiment, to ensure that the simulation error in the results is negligible. For each iteration, we simulated the true process from the covariance model described above. We then simulated data by adding a measurement-error term with variance 0.9, which corresponded to a signal-to-noise ratio of 10 (when taking $(\sigma(0))^2$ from (4.30) to be the "signal"). To ensure comparability of the results, we took the measurement-error variance to be known for all four models. To mimic the nonretrieval encountered with satellite data, we assumed that two intervals of length 25 each were not observed (i.e., the imaginary satellite did not cover these regions): The first non-randomly missing region, $\text{MNR}_1 = \{61, 62, \ldots, 85\}$, was in the "rough" portion of the domain, and the second missing region, $\text{MNR}_2 = \{201, 202, \ldots, 225\}$, was in the "smooth" portion of the domain. In addition, one third of the remaining locations were selected at random at each iteration of

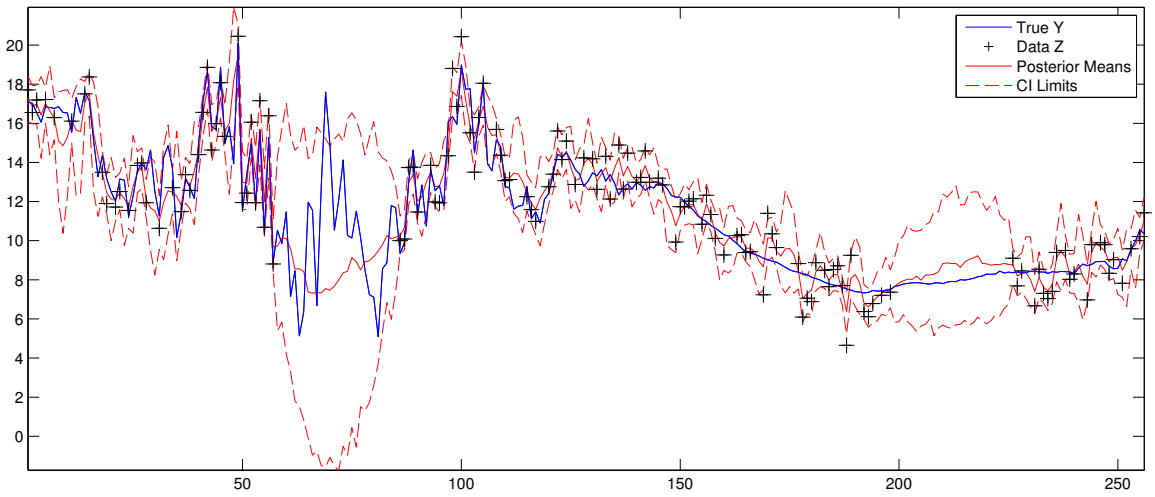

Figure 4.4: For Simulation Study 1, one example of a simulated true process $Y(\cdot)$ (in blue) and a set of simulated data (black crosses), together with the posterior mean of $Y(\cdot)$ and the posterior 2.5- and 97.5-percentiles (i.e., the endpoints of a 95% credible interval) as estimated from the data using our SRE model.

the simulation study as unobserved (e.g., due to heavy cloud cover or other retrieval problems for the imaginary satellite). These missing locations that were scattered among the observed locations will be denoted MAR (missing at random). The remaining, observed locations will be denoted OBS. The true process and the data for one iteration of the simulation study are shown in Figure 4.4. Each of the four models was then run for 2000 MCMC iterations (thinned by a factor of 5), the first 1000 of which were taken as burn-in. In a more complete study, this number will be increased.

The comparison of the four models was based on the mean squared prediction error (MSPE), the squared difference between the posterior means for each of the four models and the true process itself, averaged over locations (where the locations were stratified into the four groups: OBS, MAR, MNR_1, and MNR_2) and the 100 iterations. To quantify the accuracy of the uncertainty estimation, we used the interval score (IS). The IS combines

the width of a credible interval with a penalty for not containing the true value, and it is defined as (Gneiting and Raftery, 2007, Sect. 6.2),

$$\text{IS}_\alpha(l, u; y) = (u - l) + 2\{(l - y)_+ + (y - u)_+\}/\alpha, \tag{4.31}$$

where l and u are, respectively, the lower and upper endpoints of a $(1-\alpha)$ credible interval (we use $\alpha = 0.05$), y is the true value, and $(x)_+ := xI(x > 0)$. The goal is for small IS.

Table 4.2: Summary of the results of Simulation Study 1.

	SRE	SMC	SMC/SRE	CTO	CTO/SRE	KCG	KCG/SRE
Time (sec)	23.88	72.68	3.04	4.26	0.18	86.66	3.63
MSPE (OBS)	0.56	0.62	1.10	0.62	1.11	0.74	1.31
MSPE (MAR)	2.55	2.52	0.99	3.25	1.28	4.01	1.57
MSPE (MNR_1)	29.36	22.97	0.78	31.26	1.06	29.25	1.00
MSPE (MNR_2)	1.23	0.92	0.75	2.47	2.00	2.33	1.89
IS (OBS)	3.61	3.83	1.06	3.70	1.03	3.89	1.08
IS (MAR)	8.19	8.78	1.07	8.56	1.05	12.35	1.51
IS (MNR_1)	31.19	32.43	1.04	32.55	1.04	56.16	1.80
IS (MNR_2)	6.77	7.97	1.18	6.99	1.03	7.02	1.04

A summary of the results of the simulation study is shown in Table 4.2. We see that the CTO model produced the fastest computation time (on average, only slightly more than 4 seconds for 2,000 MCMC iterations). The SMC actually outperformed all other models in terms of the MSPE, except at locations where data were available. This is likely due to the the fact that the true covariance model in (4.29) was actually rather close to a Matérn model for small spatial lags, except that its variance varied significantly over space (which the SMC model was unable to capture, of course). The SRE model did the best in terms of uncertainty quantification as measured by IS. As expected, the CTO model performed much worse than the SRE model in terms of the MSPE in regions where the process is smooth but

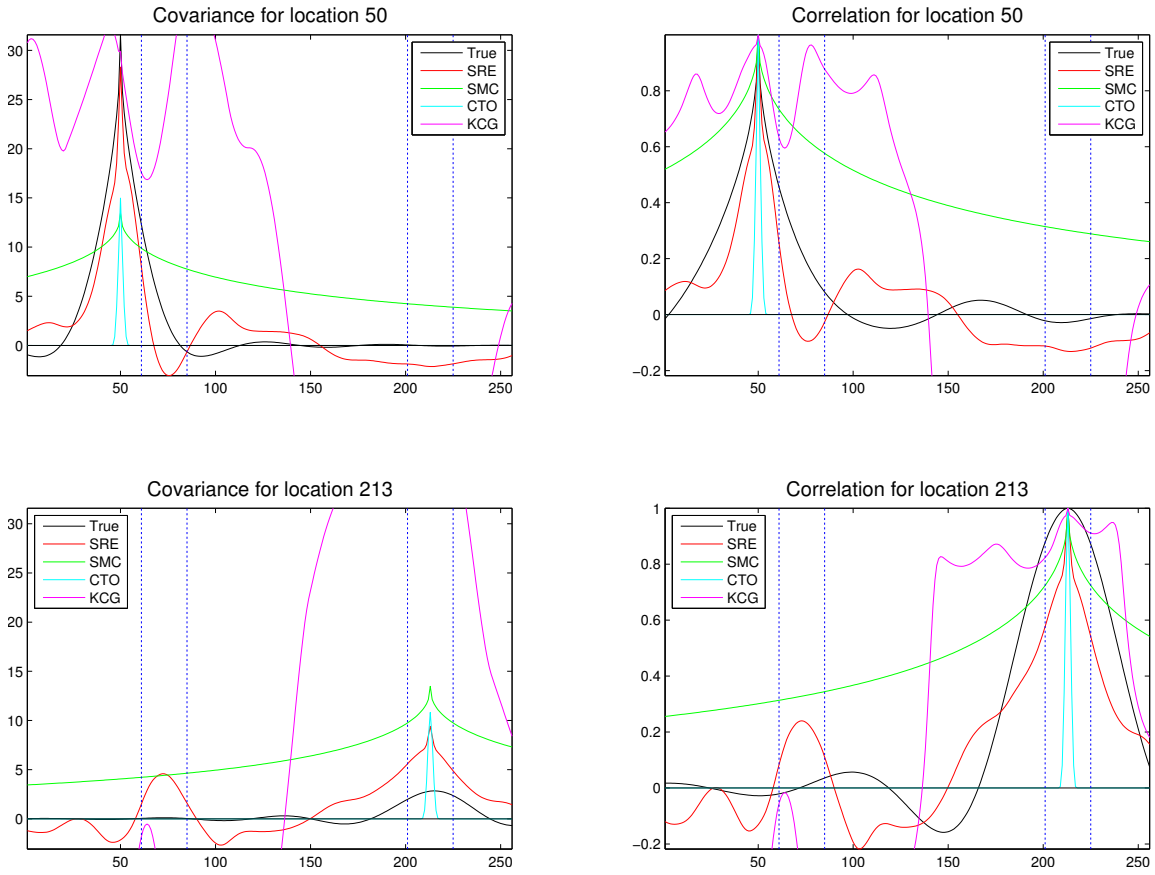

Figure 4.5: The true covariance (left column) and correlation (right column) over space for reference location 50 (first row) and reference location 213 (second row), together with the point-wise posterior means of the same quantities estimated using our SRE model, the SMC model, the CTO model, and the KCG model, for one sample from Simulation Study 1. The vertical dotted blue lines indicate the regions of missing data, MNR_1 and MNR_2.

no data were available (MNR_2). The KCG model's predictions were worse at locations that were close to observed locations (MAR), likely due to its spatially uncorrelated FSV term. In a more complete simulation study, we shall examine the effect of spatial dependence in the FSV component and whether a larger number of basis functions would result in better predictions.

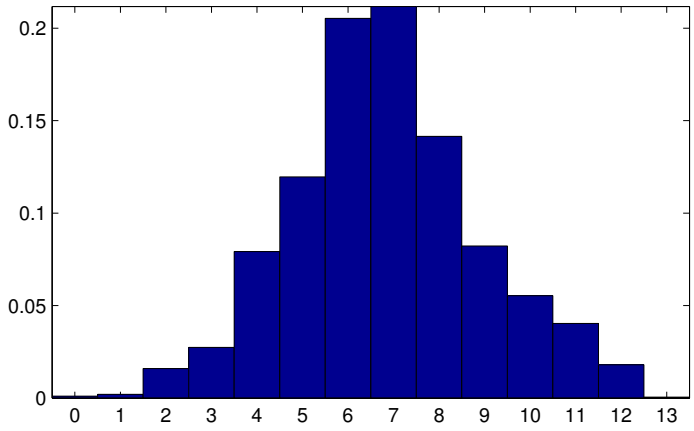

Figure 4.6: The posterior distribution of r, the number of basis functions for our SRE model for one sample from Simulation Study 1.

For one simulated dataset, we show the posterior means and credible intervals for the SRE model in Figure 4.4, the pointwise posterior means of the covariances over space for two reference locations for all four models in Figure 4.5, and the posterior distribution of r, the number of basis functions in Figure 4.6. For this realization, the KCG model has trouble estimating the covariance structure and the correlation structure; with only 11 basis functions and the considerable heterogeneity in the true model, its performance is likely to improve when a larger number of basis functions is chosen. The CTO model produced estimates of the correlation that vanish much too quickly for increasing distance (it is limited by $L_\delta = 6.5$, of course), and the SMC model overestimated the scaling parameter for this specific simulated dataset.

In Simulation Study 2, the true covariance function was assumed to be stationary. We chose an exponential covariance (i.e., a Matérn covariance with smoothness $\upsilon(\cdot) \equiv 0.5$) with standard deviation $\sigma(\cdot) \equiv 3$ and scaling parameter $\gamma(\cdot) \equiv 640$. The scaling parameter, γ, is much smaller than in (4.30), because the Matérn covariance function here is not

multiplied by the wave function as in (4.29). All hyperparameters for the four models were exactly the same as before. The results are shown in Table 4.3. The overall pattern of relative results is somewhat similar to the pattern of relative results of Simulation Study 1 in Table 4.2. But now, the SMC model, which is the correct model here, is doing comparatively better than the SRE model, whereas the CTO model and KCG model are doing comparatively worse than the SRE model overall. This indicates that our SRE model can adjust to simple covariance structures rather easily.

Table 4.3: Summary of the results of Simulation Study 2.

	SRE	SMC	SMC/SRE	CTO	CTO/SRE	KCG	KCG/SRE
Time (sec)	25.44	72.24	2.84	4.28	0.17	85.61	3.36
MSPE (OBS)	0.50	0.46	0.92	0.56	1.12	0.72	1.43
MSPE (MAR)	1.33	0.97	0.73	1.82	1.36	2.72	2.04
MSPE (MNR_1)	5.73	4.74	0.83	7.80	1.36	7.03	1.23
MSPE (MNR_2)	5.52	4.38	0.79	7.61	1.38	6.53	1.18
IS (OBS)	3.49	3.41	0.98	3.55	1.02	3.88	1.11
IS (MAR)	6.37	5.61	0.88	6.39	1.00	8.73	1.37
IS (MNR_1)	11.69	11.07	0.95	13.58	1.16	22.38	1.92
IS (MNR_2)	11.99	11.10	0.93	13.69	1.14	21.46	1.79

In Simulation Study 3, the true covariance function was assumed to be of SRE form, with the same basis functions used in the KCG model and spatially independent and homogeneous FSV; these are the assumptions underlying the KCG model. The true $\mathbf{K}$, say $\mathbf{K}_0$, was obtained as, $\mathbf{K}_0 = \underset{\mathbf{K}}{\operatorname{argmin}} \|\mathbf{B}^P_{\text{KCG}}\mathbf{K}\mathbf{B}^P_{\text{KCG}}{}' - \mathbf{\Sigma}_Y\|_F$, where now $\mathbf{\Sigma}_Y$ denotes the true covariance matrix at all locations from Simulation Study 2 (i.e., stationary Matérn). The true FSV variance was taken to be $\sigma^2_{\delta,0} = avg\{|diag(\mathbf{B}^P_{\text{KCG}}\mathbf{K}_0\mathbf{B}^P_{\text{KCG}}{}' - \mathbf{\Sigma}_Y)|\}$. The hyperparameters for the four models were again the same as before, except that the KCG-model

hyperparameters were now calibrated against the EM estimates of $\mathbf{K}$ and σ_δ^2 (see Katzfuss and Cressie, 2009, on how to obtain these estimates). The results are shown in Table 4.4. Clearly, the KCG model now makes the best predictions, especially in the areas MNR_1 and MNR_2. This model can find good values for the coefficients of the basis functions in areas where data are available, and once the basis-function coefficients are estimated, good predictions in the missing regions are given automatically. Due to the fact that there are only eleven basis functions, the true variance, $\mathbf{b}_{\text{KCG}}(\cdot)'\mathbf{K}_0\mathbf{b}_{\text{KCG}}(\cdot) + \sigma^2_{\delta,0}$, now varies considerably over space, in that it is high close to the basis function centers and low in areas far from any basis function centers. This variance heterogeneity cannot be captured correctly by the SRE, SMC, or CTO models. The SMC model assumes constant variance, and in the SRE and CTO models, there are only two basis functions available to estimate the spatial variation in the variance. However, the SRE model is very close to the CTO model in terms of the IS.

Table 4.4: Summary of the results of Simulation Study 3.

	SRE	SMC	SMC/SRE	CTO	CTO/SRE	KCG	KCG/SRE
Time (sec)	22.75	69.56	3.06	4.41	0.19	87.32	3.84
MSPE (OBS)	0.75	0.70	0.92	0.75	1.00	0.72	0.96
MSPE (MAR)	4.54	3.20	0.70	4.72	1.04	2.88	0.63
MSPE (MNR_1)	8.24	4.54	0.55	10.02	1.22	3.82	0.46
MSPE (MNR_2)	7.78	4.84	0.62	8.93	1.15	3.65	0.47
IS (OBS)	4.19	4.09	0.98	4.18	1.00	4.18	1.00
IS (MAR)	10.56	9.47	0.90	10.58	1.00	11.13	1.05
IS (MNR_1)	14.75	12.86	0.87	14.98	1.02	14.12	0.96
IS (MNR_2)	13.26	11.80	0.89	14.54	1.10	12.41	0.94

4.6 Analysis of Global CO_2 Data from the AIRS Instrument

In this section, we illustrate the use of our SRE model on a large real-world spatial dataset. The dataset (available from `http://airs.jpl.nasa.gov/AIRS_CO2_Data/`) consists of 13,911 measurements of global mid-tropospheric CO_2, which were recorded at roughly 1:30pm local time on May 1, 2003 by the Atmospheric InfraRed Sounder (AIRS) on board NASA's Aqua satellite (e.g., Chahine et al., 2006). The unit of measurement is parts per million (ppm). Data at latitudes south of -60° latitude have not been released yet by the AIRS team, and so all available measurements are north of -60° latitude. The dataset is shown in the top panel of Figure 4.7. While the measurements are really averages over the "footprint" of the AIRS instrument, we assume for simplicity that they are made at point locations at the centers of the footprints.

We again compared our SRE model to the covariance-tapering-only (TCO) model and the Kang and Cressie (2011) Givens-angle SRE model (referred to in this chapter as the KCG model), both described in Section 4.5. To estimate prediction accuracy, we left out both a large area (to assess long-range prediction) and a random sample (to assess short-range prediction) of observations. We created two test sets, one consisting of the 77 non-randomly selected observations (hereafter referred to as MNR) in the region 30° to 47° longitude and 34° to 46° latitude, and the other one consisting of a random sample of 200 of the remaining measurements (hereafter MAR). The test data were only used for model evaluation, and they were not available for model fitting. Therefore, the number of observations was given by, $n = 13,911 - 77 - 200 = 13,634$. We also wanted to assess the accuracy of the uncertainty estimation via the interval score (IS) given by (4.31). From the models, we obtained samples from, $[\{Y(\mathbf{s}_j) : \ \mathbf{s}_j \in \text{MNR} \cup \text{MAR}\}|\mathbf{Z}]$, the posterior distribution of the process at the test set locations, given the observations $\mathbf{Z}$ that were not

in either of the test sets. By adding an independent measurement-error component, $\epsilon(\mathbf{s}_i) \sim N(0, \sigma^2_\epsilon)$, to these samples, we obtained samples from, $[\{Z(\mathbf{s}_j) : \mathbf{s}_j \in \text{MNR} \cup \text{MAR}\}|\mathbf{Z}]$, which in turn could be used for assessment of the uncertainty-estimation accuracy when compared to the test data, $\{Z(\mathbf{s}_i)\}$, using the IS (for more details on this idea, see, e.g., Cressie and Wikle, 2011, Sect. 2.2.2).

Because we could only compare our posterior distributions to the measurements (which include measurement error) and not the corresponding true-process values $\{Y(\mathbf{s}_i)\}$, we obtained samples from the posterior distribution of $Z(\mathbf{s}_i)$ by adding an independent measurement-error component, $\epsilon(\mathbf{s}_i) \sim N(0, \sigma^2_\epsilon)$, to the samples from the posterior distributions of $Y(\mathbf{s}_i)$, where $\mathbf{s}_i \in$ MNR and $\mathbf{s}_i \in$ MAR We used the samples from the posterior distribution of the $Z(\mathbf{s}_i)$.

To ensure comparability between the three models we were comparing here, we assumed the measurement-error variance to be known for all three models; in reality, we estimated the variance using a variogram-extrapolation technique (Kang et al., 2009) to be $\sigma^2_\epsilon = 5.4221\text{ppm}^2$. For the mean term, $\mathbf{x}(\cdot)'\boldsymbol{\beta}$, we used only an intercept, resulting in the one-dimensional covariate vector $x(\cdot) \equiv 1$.

For our SRE model, the tapering lengths were set to $L_\nu = 1$ (as recommended in Section 4.3.2) and $L_\delta = 0.08$. Further, we chose $u = 4$, $\mu_\sigma = \log(\sqrt{\hat{\sigma}^2_Z})$, and $\sigma^2_\sigma = 2\mu_\sigma$, where $\hat{\sigma}^2_Z$ is the empirical variance of the data. For both of the two scale parameters of the SBF component, we chose the hyperparameters, $\mu_{\gamma_\nu} = \log(2L_\nu)$ and $\sigma^2_{\gamma_\nu} = 2\mu_{\gamma_\nu}$. The basis functions, $\mathbf{b}_\theta(\cdot)$, for the spatially-varying covariance parameters were taken to be 32 bisquare functions with radius 6241.11 km of great-arc distance, centered at ISEA Aperture 3 Hexagon centers at resolution 1 obtained using DGGRID software (Sahr, 2003).

For the CTO model, we chose the same L_δ, $\mathbf{b}_\theta(\cdot)$, and the same prior distribution for $\boldsymbol{\theta}_\sigma$ as for the SRE model. However, we allowed for random rotation and two random scale parameters of the FSV component, with $\mu_{\gamma_\delta} = \log(2L_\delta)$ and $\sigma^2_{\gamma_\delta} = 2\mu_{\gamma_\delta}$ for both scale parameters.

For the KCG model, we chose 124 bisquare basis functions of two resolutions. This number was determined to be comparable to the SRE model, because the estimated posterior probability of $r > 95$ was zero for our SRE model. The set of basis functions for the KCG model was identical to the first two resolutions chosen in Section 2.5.2. The hyperparameters for $\mathbf{K}$ and σ^2_δ were calibrated as described in Kang and Cressie (2011), but using the parameters' EM estimates (see Katzfuss and Cressie, 2009, on how to obtain these estimates).

We ran an MCMC for each of the three models for 12,000 iterations, of which 2,000 were considered burn-in, and we only used every 20th of the remaining iterations for inference. We also obtained the posterior distribution of $Y(\cdot)$ at a grid of 20,422 locations on the globe, given by ISEA Aperture 3 Hexagon centers at resolution 7 north of -60° latitude; see the DGGRID software (Sahr, 2003). Summaries of the posterior distribution of $Y(\cdot)$ obtained from our SRE model at the 20,422 grid centers are shown in the middle and bottom panels of Figure 4.7. In Figure 4.8, we also show posterior means of the spatially varying parameters determining $C_{\tilde{\nu}}(\cdot,\cdot)$ in (4.6) and $C_\delta(\cdot,\cdot)$ in (4.5). The images of the posterior mean of $\sigma_{\tilde{\nu}}(\cdot)$ and $\upsilon_\delta(\cdot)$, together with the posterior standard deviation of $Y(\cdot)$ shown in Figure 4.7, indicate that midtropospheric CO_2 is rather smooth around the equator, in that $\sigma_{\tilde{\nu}}(\cdot)$ is relatively small there, and the smoothness parameter of the FSV component is relatively large. In higher latitudes the variance $(\sigma_{\tilde{\nu}}(\cdot))^2$ increases and the smoothness parameter $\upsilon_\delta(\cdot)$ decreases.

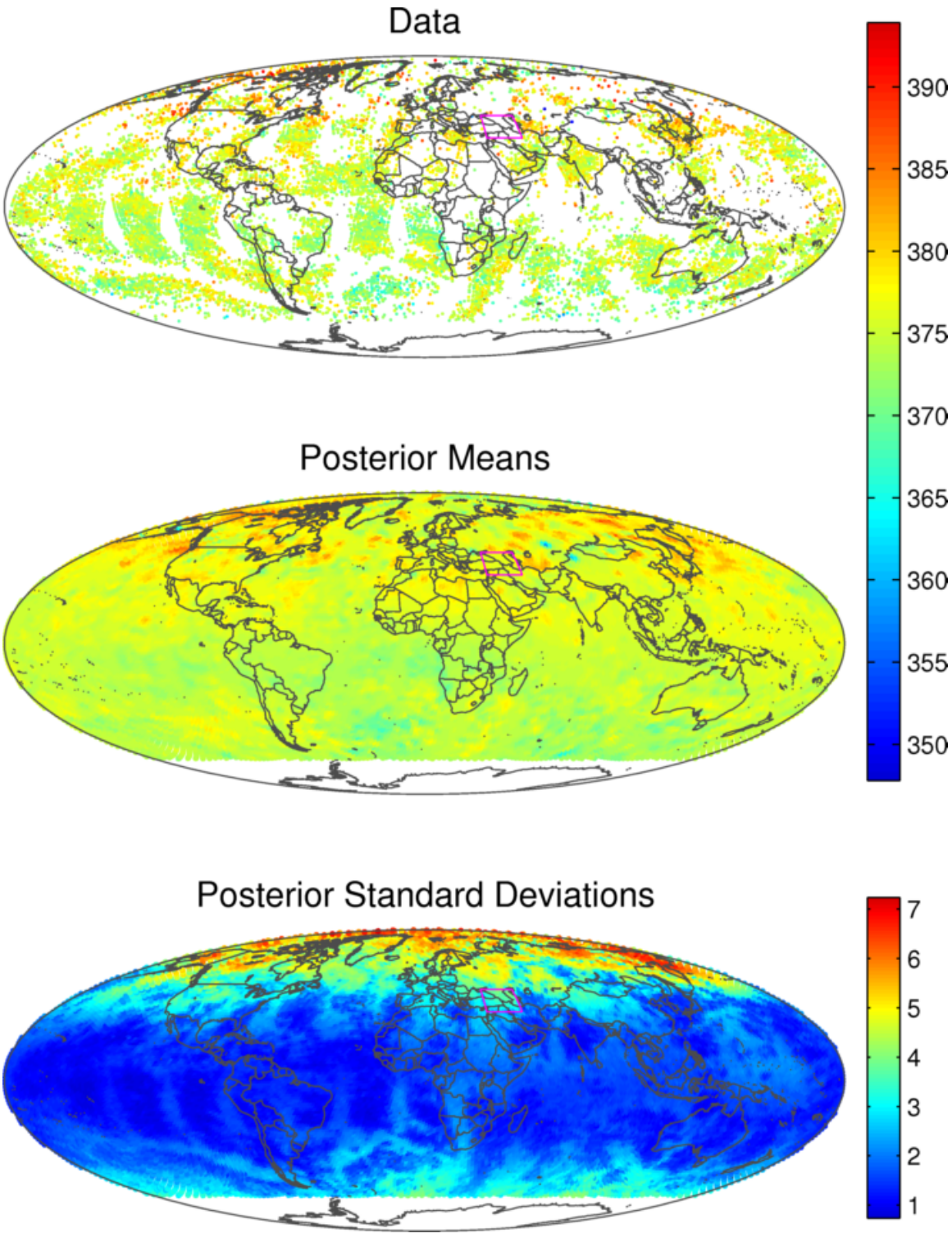

Figure 4.7: AIRS data and posterior summaries of $Y(\cdot)$ obtained from our SRE model. The pink box indicates the MNR region. Units are ppm.

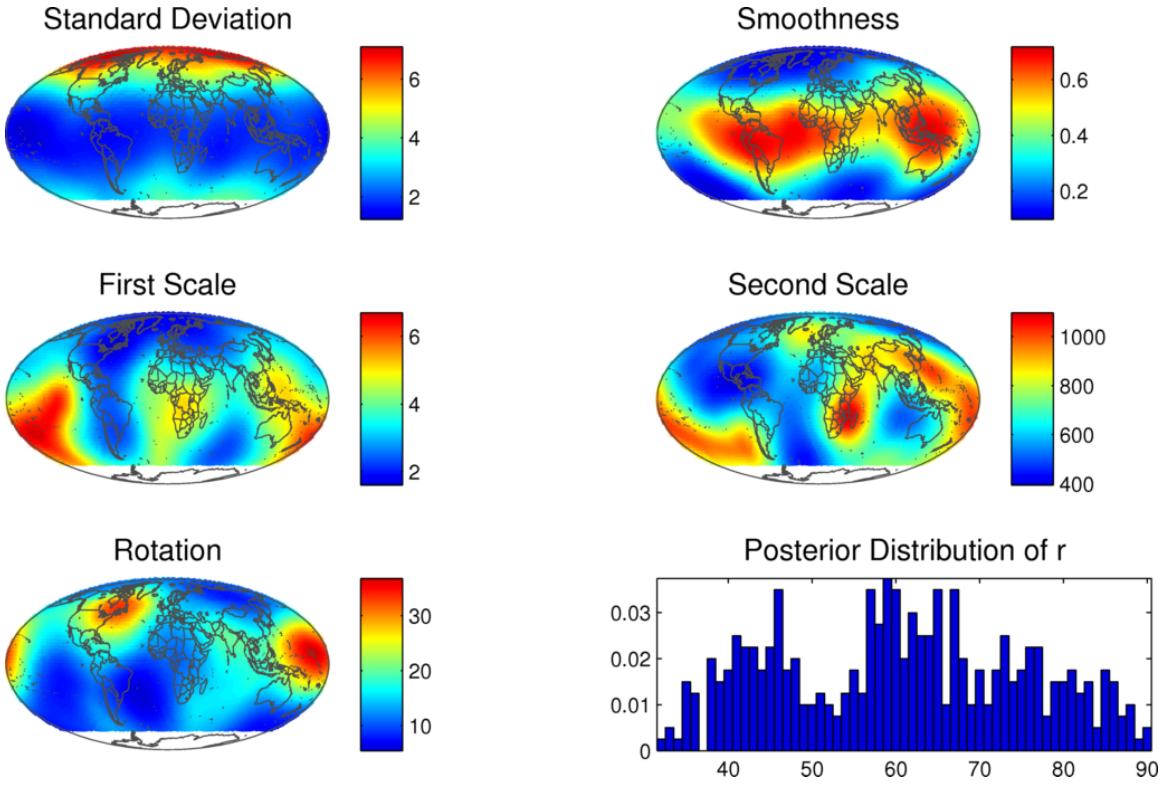

Figure 4.8: Posterior means of the spatially varying covariance parameters of the SRE model in $C_{\tilde{\nu}}(\cdot,\cdot)$ in (4.6) and $C_{\delta}(\cdot,\cdot)$ in (4.5): standard deviation $\sigma_{\tilde{\nu}}(\cdot)$ (top left), smoothness $\upsilon_{\delta}(\cdot)$ (top right), the scale parameters $\gamma_{\tilde{\nu},1}(\cdot)$ and $\gamma_{\tilde{\nu},1}(\cdot)$ (middle row), the rotation parameter $\kappa_{\tilde{\nu}}(\cdot)$ (bottom left), and the posterior distribution of the number of basis functions r (bottom right), as estimated from the AIRS data in Section 4.6.

The average squared distance (ASD) of the posterior means of $Y(\mathbf{s}_i)$ at the test locations and the corresponding measurements, $\{Z(\mathbf{s}_i)\}$, in the two test sets are shown in Table 4.5 for all three models. We also show the IS for 95% credible intervals for the measurements $\{Z(\mathbf{s}_i)\}$ in the test sets. The SRE model performs best according to all but two criteria: It takes the longest to fit, and its ASD for MAR is worse than the ASD for the CTO. This provides some indication that fixing the scale and rotation parameters for the FSV results in worse short-range prediction performance in this real-world dataset. These parameters can, of course, be allowed to be random, but we do need to make sure that the acceptance rate for the MH step for $\boldsymbol{\theta}_\delta$ (Step 4 in Section 4.4.2) stays at a reasonable level. The KCG model performs relatively better here (especially in terms of the ASD in the MNR set) than in Simulation Study 1, likely due to the fact that the true covariance structure of mid-tropospheric CO_2 is very different from parametric covariance functions (even if the parameters vary spatially as in Simulation Study 1).

Table 4.5: Summary of the results of the AIRS data analysis.

	SRE	CTO	CTO/SRE	KCG	KCG/SRE
Time (hrs)	18.57	11.38	0.61	13.82	0.74
ASD (MAR)	18.34	17.81	0.97	19.53	1.06
ASD (MNR)	19.12	20.81	1.09	19.20	1.00
IS (MAR)	24.47	25.02	1.02	30.07	1.23
IS (MNR)	31.92	33.93	1.06	34.08	1.07

4.7 Conclusions

In this chapter, we have presented an SRE model that combines an SBF component with a spatially dependent FSV component. For the SBF component, we make inference

on the number, locations, and shapes of the basis functions. The FSV component is allowed to exhibit nonstationarity, and compact support of its covariance function ensures fast computation, even for very large datasets.

The results of a preliminary simulation study (Section 4.5) and a validation exercise on global CO_2 (Section 4.6) indicate that our SRE model may provide considerable improvements when compared with two other spatial statistical models used for the analysis of large spatial datasets. Compared with a model containing only a tapered covariance component, we improve long-range prediction. Compared with the Givens-angle-based model of Kang and Cressie (2011) that has a spatially independent FSV component, we obtain better predictions at locations that are close to observed locations. There is also some qualitative (from visual inspection of estimated covariance structure) and quantitative (from IS) indication that we improve the estimation of the covariance structure and the prediction uncertainty. Of interest would also be a comparison of the performance of our model to that of the predictive-process model of Banerjee et al. (2008) and Finley et al. (2009).

Up to Section 4.4, we allowed for the measurement-error variance to be random; however, we have fixed the parameter in both the Simulation Study and the AIRS data analysis. This ensured comparability between the different models considered there. But while estimating the measurement-error variance is easy in theory, there might be considerable identifiability problems between the variance of the FSV component and the variance of the measurement error, if the fine-scale correlation structure of the true process is not sufficiently smooth. Further investigation of this issue is warranted.

We have also claimed that our model allows for feasible computation times. In principle, the model should scale well for increasing sample size, and its computational speed

can be controlled via the choice of the tapering lengths L_δ and L_ν. Since most of the computation time is spent on evaluating the Matérn covariance functions in $\mathbf{B}$ and $\mathbf{V}_\delta$, further significant speed-ups could be achieved in two ways: Assuming a double exponential correlation function for the SBF component (i.e., setting $\upsilon_\nu(\cdot) \equiv \infty$) would let us avoid having to evaluate the modified Bessel function in the general expression of the Matérn covariance functions when calculating the non-zero elements of $\mathbf{B}$. Second, if we assumed a discrete distribution for $\upsilon_\delta(\cdot) \equiv \upsilon_\delta$, $\mathbf{V}_\delta$ and its Cholesky decomposition could be precomputed for every possible value of υ_δ, and so we would not have to do so at each iteration of the MCMC.

Finally, while this chapter includes an attempt to unify some of the many related models for large spatial datasets, we do not give much discussion of multiresolutional structure of the basis functions. Multiresolutional structure is a common theme in much of the wavelet literature, and the use of basis functions of different resolutions has also been advocated by Cressie and Johannesson (2008), for example. In our model, we do not allow several resolutions of basis functions, and instead we allow for spatially varying (random) shape of the basis functions. One issue that should be further investigated is whether it is possible to induce a multiresolutional structure for our basis functions through the parent process.

Appendix A: Details of Posterior Inference for the Model of Chapter 3

For generic (sets of) random variables X and Y, let $[X]$ denote the (marginal) distribution of X, $[X|Y]$ denote the conditional distribution of X given Y, and $[X|\cdot]$ denote the full conditional distribution of X, which is defined as the conditional distribution of X given all other variables (including the data). We sample from the posterior distribution, the distribution of the unknowns given the data, using a Markov chain Monte Carlo (MCMC) algorithm in form of a Gibbs sampler (Geman and Geman, 1984) with some Metropolis-Hastings (MH) updates (Metropolis et al., 1953; Hastings, 1970) where necessary. In a Gibbs sampler, each unknown variable is updated from its full conditional distribution. These full conditional distributions are proportional to the joint distribution of all variables. Due to (conditional) independencies, the joint distribution can be written as the product of the data model, the process model, and the parameter (prior) model:

$$[\mathbf{Z}_{1:T}, \boldsymbol{\eta}_{0:T}, \boldsymbol{\delta}^P_{1:T}, \boldsymbol{\theta}_P, \boldsymbol{\theta}_H] = [\mathbf{Z}_{1:T}|\boldsymbol{\eta}_{0:T}, \boldsymbol{\delta}^P_{1:T}, \boldsymbol{\beta}_{1:T}][\boldsymbol{\eta}_{0:T}, \boldsymbol{\delta}^P_{1:T}|\boldsymbol{\theta}_P][\boldsymbol{\theta}_P, \boldsymbol{\theta}_H], \tag{A.1}$$

where most of the terms on the right-hand side are given in Section 3.2. We recommend using the adaptive MH algorithm of Haario et al. (2001) for all blocks of parameters for which the full conditional distribution is not available in closed form, as this allows the complicated covariance structure of the parameters in each block to be more fully exploited by the algorithm. Chapter 3 does not give a complete specification of the prior distributions;

we shall do so at the beginning of the relevant subsections below. At each step of the MCMC, each unknown is updated as described below, given the most recently sampled values of the other unknowns.

Posterior Inference on the Process Variables and the Trend

We begin by making the prior assumptions for the trend terms more explicit:

$$\boldsymbol{\beta}_t \overset{iid}{\sim} N_p(\boldsymbol{\mu}_\beta, \sigma_\beta^2 I_p),\ t = 1, \ldots, T,$$

where $\boldsymbol{\mu}_\beta$ and σ_β^2 are known hyperparameters. The prior mean, $\boldsymbol{\mu}_\beta$, can be set equal to a point estimate of the trend-coefficient vector. The variance σ_β^2 is set to some very large value (e.g., 10^{15}), to make the prior distribution essentially noninformative.

The prior distribution on $\{\boldsymbol{\delta}_t^P\}$ is also normal and is described in Section 3.2.1. Making use of standard normal-normal conjugacy, we see from (A.1) that the full conditional distributions of $\{\boldsymbol{\beta}_t\}$ and $\{\boldsymbol{\delta}_t^P\}$ are multivariate normal distributions of the form $N(A^{-1}\mathbf{k}, A^{-1})$, where (using obvious notation),

$$A_{\boldsymbol{\beta}_t} := X_t'(\sigma_{\epsilon,t}^2 V_{\epsilon,t})^{-1} X_t + \sigma_\beta^{-2} I_p$$

$$\mathbf{k}_{\boldsymbol{\beta}_t} := X_t'(\sigma_{\epsilon,t}^2 V_{\epsilon,t})^{-1}(\mathbf{z}_t - M_t \boldsymbol{\delta}_t^P - B_t \boldsymbol{\eta}_t) + \sigma_\beta^{-2} \boldsymbol{\mu}_\beta,$$

and

$$A_{\boldsymbol{\delta}_t^P} := M_t'(\sigma_{\epsilon,t}^2 V_{\epsilon,t})^{-1} M_t + (\sigma_\delta^2 V_{\delta,t}^P)^{-1}$$

$$\mathbf{k}_{\boldsymbol{\delta}_t^P} := M_t'(\sigma_{\epsilon,t}^2 V_{\epsilon,t})^{-1}(\mathbf{z}_t - X_t \boldsymbol{\beta}_t - B_t \boldsymbol{\eta}_t),$$

respectively.

As mentioned in Section 3.2.4, the basis-function-coefficient vectors $\{\boldsymbol{\eta}_t\}$ can be sampled using a forward-filtering, backward-sampling algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994). Telescoping the full conditional distribution and exploiting the Markov

structure of the random-effects vectors, we have,

$$[\boldsymbol{\eta}_{0:T}|\cdot] = [\boldsymbol{\eta}_T|\mathbf{z}_{1:T}, \boldsymbol{\theta}] \prod_{t=0}^{T-1} [\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t+1}, \mathbf{z}_{1:t}, \boldsymbol{\theta}].$$

At iteration $l+1$ of the Gibbs sampler, we wish to update $\boldsymbol{\eta}_{0:T}$ given all other variables, namely $\mathbf{z}_{1:T}$, $\boldsymbol{\delta}_t^{P^{[l]}}$, and $\boldsymbol{\theta}^{[l]}$. To do this, we first run the Kalman filter as described in Shumway and Stoffer (2006, Chap. 6), with the transformed measurements $\{\tilde{\mathbf{z}}_t := \mathbf{z}_t - X_t\boldsymbol{\beta}_t^{[l]} - M_t\boldsymbol{\delta}_t^{P^{[l]}} : t = 1, \ldots, T\}$ as the data, to obtain the filtering quantities, $\boldsymbol{\eta}_{t|t}^{[l]} := E(\boldsymbol{\eta}_t|\mathbf{z}_{1:t}, \boldsymbol{\theta}^{[l]})$, $\boldsymbol{\eta}_{t|t-1}^{[l]} := E(\boldsymbol{\eta}_t|\mathbf{z}_{1:(t-1)}, \boldsymbol{\theta}^{[l]})$, $P_{t|t}^{[l]} := var(\boldsymbol{\eta}_t|\mathbf{z}_{1:t}, \boldsymbol{\theta}^{[l]})$, and $P_{t|t-1}^{[l]} := var(\boldsymbol{\eta}_t|\mathbf{z}_{1:(t-1)}, \boldsymbol{\theta}^{[l]})$, $t = 1, \ldots, T$, where $P_{0|0}^{[l]} = K_0^{[l]}$. Then, we sample $\boldsymbol{\eta}_T^{[l+1]} \sim N_r(\boldsymbol{\eta}_{T|T}^{[l]}, P_{T|T}^{[l]})$, and for $t = T-1, T-2, \ldots, 0$, we sample,

$$\boldsymbol{\eta}_t^{[l+1]} \sim N_r(\boldsymbol{\eta}_{t|t}^{[l]} + J_t^{[l]}[\boldsymbol{\eta}_{t+1}^{[l+1]} - \boldsymbol{\eta}_{t+1|t}^{[l]}], P_{t|t}^{[l]} - J_t^{[l]} P_{t+1|t}^{[l]} J_t^{[l]\prime}),$$

where $J_t^{[l]} := P_{t|t}^{[l]} H^{[l]\prime} (P_{t+1|t}^{[l]})^{-1}$.

Posterior Inference on the Fine-Scale-Variation Variance

For the standard deviation, σ_δ, of the fine-scale variation $\{\delta_t(\cdot)\}$ given in Section 2.1 of the main document, we assume a flat prior of the form,

$$\sigma_\delta \sim U(0, \kappa_\delta),$$

where $\kappa_\delta^2 := 5\hat{\sigma}_\delta^2$ (e.g., Kang and Cressie, 2011). Here, $\hat{\sigma}_\delta^2$ is a point estimate of σ_δ^2 (e.g., the EM estimate). This prior results in a full conditional distribution of closed form,

$$\sigma_\delta^2|\cdot \sim InvGamma\left(n_+/2 - 1/2, \textstyle\sum_{t=1}^T \boldsymbol{\delta}_t^{P\prime} M_t' V_{\delta,t}^{-1} M_t \boldsymbol{\delta}_t^P/2; \kappa_\delta^2\right),$$

which we define to be an inverse-gamma distribution truncated above at κ_δ^2 (and renormalized).

For the basis-function coefficients $\boldsymbol{\eta}_\delta$ in the function $v_\delta(\cdot) = \exp\{\mathbf{b}_\delta(\cdot)'\boldsymbol{\eta}_\delta)\}$, the full conditional distribution is not available in closed form. Hence, the vector $\boldsymbol{\eta}_\delta$ is updated as a block using a MH step from,

$$[\boldsymbol{\eta}_\delta|\cdot] \propto [\boldsymbol{\eta}_\delta]\,[\boldsymbol{\delta}_{1:T}|\sigma^2_\delta, \boldsymbol{\eta}_\delta] = N_{r_\delta}(\boldsymbol{\eta}_\delta|\mathbf{0}, \sigma^2_{\eta_\delta} I_{r_\delta}) \prod_{t=1}^{T} N(\boldsymbol{\delta}_t|\sigma^2_\delta V_{\delta,t}),$$

where $N(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ generically denotes the density function of a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ evaluated at $\mathbf{x}$.

Posterior Inference on the Covariance Matrices K_0 and U

The prior distributions for the covariance matrices K_0 and U are taken to be multiresolutional Givens-angle priors. We follow Kang and Cressie (2011) and decompose each of the two matrices into the $r \times r$ diagonal matrix of eigenvalues and the Givens angle matrices with $r(r+1)/2$ parameters. The eigenvalues and Givens angles of K_0 and U corresponding to different resolutions will be treated differently. Let the r_c eigenvalues corresponding to resolution c for the two matrices be denoted by $\{\lambda^{K_0}_{c,1}, \ldots, \lambda^{K_0}_{c,r_c}\}$ and $\{\lambda^{U}_{c,1}, \ldots, \lambda^{U}_{c,r_c}\}$, $c = 1, \ldots, C$, respectively. The $r(r+1)/2$ Givens angles for the two matrices will be denoted as $\{\rho^{K_0}_{ij} : i = 1, \ldots, r-1;\ j = i+1, \ldots, r\}$ and $\{\rho^{U}_{ij} : i = 1, \ldots, r-1;\ j = i+1, \ldots, r\}$, respectively. We transform both sets of parameters to avoid having a restricted domain. The log-eigenvalues, $\tilde{\lambda}_{c,j} := \log \lambda_{c,j}$, and the transformed Givens angles,

$$\tilde{\rho}_{ij} := \log \frac{\pi/2 + \rho_{ij}}{\pi/2 - \rho_{ij}},\ i = 1, \ldots, r-1,\ j = i+1, \ldots, r,$$

both have support on the entire real line. Then, for $c = 1, \ldots, C$, the log-eigenvalues corresponding to resolution c for K_0 and U are *a priori* distributed as the order statistics of an iid sample from the normal distribution,

$$\begin{aligned} \{\tilde{\lambda}^{K_0}_{c,1}, \ldots, \tilde{\lambda}^{K_0}_{c,q_c}\} &\sim OSN(\mu^{K_0}_c, w^{K_0}_c) \\ \{\tilde{\lambda}^{U}_{c,1}, \ldots, \tilde{\lambda}^{U}_{c,q_c}\} &\sim OSN(\mu^{U}_c, w^{U}_c), \end{aligned} \tag{A.2}$$

for $c = 1, \ldots, C$, where $\{\mu_c^{K_0}\}$, $\{w_c^{K_0}\}$, $\{\mu_c^U\}$, and $\{w_c^U\}$ are fixed hyperparameters (see below). We also impose the constraint that no eigenvalue of the $(c+1)$-th resolution can be larger than any eigenvalue of the c-th resolution, so that all eigenvalues are ordered even across resolutions. Letting $R_k := \{(i,j): \ c_i = c_j = k\}$, for $k = 1, \ldots, C$, and $R_0 := \{(i,j): \ c_i \neq c_j\}$ (c_i is defined below (10) in the main document), we have the prior distributions,

$$\begin{aligned} \tilde{\rho}_{i,j}^{K_0} &\overset{iid}{\sim} N(m_c^{K_0}, (\tau_c^{K_0})^2), \ (i,j) \in R_c, \\ \tilde{\rho}_{i,j}^{U} &\overset{iid}{\sim} N(m_c^{U}, (\tau_c^{U})^2), \ (i,j) \in R_c, \end{aligned} \tag{A.3}$$

for $c = 0, 1, \ldots, C$.

To determine the (fixed) hyperparameters for the prior distributions of the parameters in K_0 and U, we consider point estimates $\hat{K}_0$ for K_0 and $\hat{U}$ for U (e.g., the EM estimates). Using these estimates, we follow the hyperparameter-estimation approach of Kang and Cressie (2011). We calculate the empirical eigenvalues and empirical Givens angles of $\hat{K}_0$ and $\hat{U}$, and we calibrate the means and the (inflated) variances of the hyperparameters in (A.2) and (A.3) from the log empirical eigenvalues and transformed empirical Givens angles.

The full conditional distributions of the parameters in K_0 and U cannot be obtained analytically. Instead, we update them using random-walk MH steps. We update the log-eigenvalues blocked by resolutions with normal proposals. For the transformed Givens angles, all angles within a particular resolution, and the between-resolution angles, are updated as blocks, independently for K_0 and U, and again with normal proposals.

Posterior Inference on the Propagator Matrix H

From (3.10), the joint (conditional) prior for the vector $\mathbf{h} := vec(H')$ is given by,

$$\mathbf{h}|\boldsymbol{\theta}_H \sim N_{r^2}(\boldsymbol{\mu}_H, \Sigma_H),$$

where the mean vector $\boldsymbol{\mu}_H$ is stacked in the same manner as $\mathbf{h}$ and its elements are given in (3.10). The covariance matrix is given by, $\Sigma_H := diag(\sigma_{11}^2, \ldots, \sigma_{1r}^2, \ldots, \sigma_{r1}^2, \ldots, \sigma_{rr}^2)$, where $\sigma_{ij} := \tau_{c_i,c_j} g(d_{ij}; \alpha_{c_i,c_j}, \gamma_{c_i,c_j})$ is the standard deviation of h_{ij} (conditional on $\boldsymbol{\theta}_H$) given by (3.10).

We must complete the prior specification on H by describing the calibration of the hyperparameters $\{\sigma_{\mu,k}^2\}$, $\{a_{\tau,kl}\}$, and $\{b_{\tau,kl}\}$. We begin by calibrating the latter two (sets of) parameters: We choose ν to be the desired degrees of freedom in the (conditional) t-distribution for h_{ij} (see Section 2.3 of the main document; we use $\nu = 10$ in Section 3; in Section 4, we set $\nu = 1000$ to give more weight to the prior means of $\{\tau_{kl}\}$ in light of the large number of basis functions used, namely $r = 380$), which implies $a_{\tau,kl} = \nu/2$. Then, $\{b_{\tau,kl}\}$ are chosen so that the estimated variance of the off-diagonal elements of a point estimate $\hat{H}_{kl}$ (e.g., the EM estimate) matches the theoretical value derived in (3.14). This leads to $b_{\tau,kl} = (a_{\tau,kl} - 1)w_{kl}$, where

$$w_{kl} := avg\{\hat{h}_{ij}^2 \colon\ i \neq j, c_i = k, c_j = l\} / avg\{E(g(d_{ij}; \alpha_{kl}, \gamma_{kl})^2) \colon\ i \neq j, c_i = k, c_j = l\}.$$

Finally, values for $\{\sigma_{\mu,k}^2\}$ are chosen so that $\sigma_{\mu,k}^2 = avg\{(\hat{h}_{ii} - 1)^2 \colon\ c_i = k\} - b_{\tau,kk}/(a_{\tau,kk} - 1)$, which is derived by setting $i = j$ in the expression for the marginal variance of h_{ij} given in (3.14).

To update the elements of $\mathbf{h} := vec(H')$ and the parameters $\boldsymbol{\theta}_H$, we begin by updating $\boldsymbol{\theta}_H$. The full conditional distributions of $\{\mu_k \colon\ k = 1, \ldots, C\}$ are given by,

$$\mu_k | \cdot \sim N([1/\sigma_\mu^2 + \sum_{i \colon c_i = k} 1/\tau_{kk}^2]^{-1} [1/\sigma_\mu^2 + \sum_{i \colon c_i = k} h_{ii}/\tau_{kk}^2], [1/\sigma_\mu^2 + \sum_{i \colon c_i = k} 1/\tau_{kk}^2]^{-1}),\ k = 1, \ldots, C.$$

The elements of $\{\tau_{kl}^2\}$ are sampled from their full conditional distributions,

$$\tau_{kl}^2 | \cdot \sim InvGamma(a_{\tau,kl} + \sum_{(i,j) \in \mathcal{I}_{kl}} 1/2, b_{\tau,kl} + (\mathbf{h}_{kl} - \boldsymbol{\mu}_{H,kl})' \Sigma_{h,kl}^{-1} (\mathbf{h}_{kl} - \boldsymbol{\mu}_{H,kl})/2),$$

where $\mathcal{I}_{kl} := \{(i,j) : \ c_i = k, c_j = l, \alpha_{kl} > d_{ij}\}$, $\mathbf{h}_{kl} := vec(H'_{kl})$, $\boldsymbol{\mu}_{H,kl} := E(\mathbf{h}_{kl}|\boldsymbol{\theta}_H)$, and $\Sigma_{h,kl} := var(\mathbf{h}_{kl}|\boldsymbol{\theta}_H)/\tau_{kl}^2$. For the shape parameters $\{\gamma_{kl}\}$, the full conditional distributions,

$$[\gamma_{kl}|\cdot] \propto [\mathbf{h}_{kl}|\boldsymbol{\mu}_{H,kl}, \tau_{kl}^2, \alpha_{kl}]\,[\gamma_{kl}], \ k,l = 1,\ldots,C,$$

are not available in closed form. Therefore, we update $\{\gamma_{kl}\}$ using MH steps with normal proposals: For $k,l = 1,\ldots,C$, we draw a proposal γ_{kl}^* from $N(\gamma_{kl}, \sigma^2_{\gamma,\text{prop}})$, where γ_{kl} is the value from the previous MCMC iteration and $\sigma^2_{\gamma,\text{prop}}$ is the proposal variance. The proposal is then accepted with probability,

$$\min\left\{1, \frac{N(\mathbf{h}_{kl}|\boldsymbol{\mu}_{H,kl}, \tau_{kl}^2\Sigma^*_{h,kl})}{N(\mathbf{h}_{kl}|\boldsymbol{\mu}_{H,kl}, \tau_{kl}^2\Sigma_{h,kl})}\frac{N(\gamma^*_{kl}|\mu_\gamma, \sigma^2_\gamma)}{N(\gamma_{kl}|\mu_\gamma, \sigma^2_\gamma)}\right\},$$

where $\Sigma^*_{h,kl}$ is the same as $\Sigma_{h,kl}$ above, but with γ^*_{kl} instead of γ_{kl}, and recall that $N(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ denotes the density function of a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ evaluated at $\mathbf{x}$.

The only parameters that remain are $\{\alpha_{kl} : \ k,l = 1,\ldots,C\}$ and $\mathbf{h}$. We will update them as a block by decomposing their joint full conditional distribution in the form,

$$[\{\alpha_{kl}\}, \mathbf{h}|\tilde{\boldsymbol{\theta}}] = [\{\alpha_{kl}\}|\tilde{\boldsymbol{\theta}}]\,[\mathbf{h}|\{\alpha_{kl}\}, \tilde{\boldsymbol{\theta}}], \tag{A.4}$$

where $\tilde{\boldsymbol{\theta}}$ is a vector containing all unknowns except for $\mathbf{h}$ and $\{\alpha_{kl} : \ k,l = 1,\ldots,C\}$, as well as the data. To sample $\{\alpha_{kl}\}$ efficiently, we sample their transformations, $\tilde{\alpha}_{kl} := \Phi^{-1}(\alpha_{kl})$, $k,l = 1,\ldots,C$, as a block from,

$$[\{\tilde{\alpha}_{kl}\}|\tilde{\boldsymbol{\theta}}] \propto [\{\tilde{\alpha}_{kl}\}] \int [\boldsymbol{\eta}_{1:T}, \mathbf{h}|\boldsymbol{\eta}_0, \boldsymbol{\mu}_H, \Sigma_H, U]d\mathbf{h} = [\{\tilde{\alpha}_{kl}\}]\,[\boldsymbol{\eta}_{1:T}, \mathbf{h}|\boldsymbol{\eta}_0, \boldsymbol{\mu}_H, \Sigma_H, U],$$

where $[\{\tilde{\alpha}_{kl}\}] = \prod_{k,l} N(\tilde{\alpha}_{kl}|0,1)$, and then we transform $\{\tilde{\alpha}_{kl}\}$ back to $\{\alpha_{kl}\}$. To calculate the acceptance probability, we need to find $[\boldsymbol{\eta}_{1:T}, \mathbf{h}|\boldsymbol{\eta}_0, \boldsymbol{\mu}_H, \Sigma_H, U]$; that is, we need to

marginalize over $\mathbf{h}$:

$$[\boldsymbol{\eta}_{1:T}, \mathbf{h}|\boldsymbol{\eta}_0, \boldsymbol{\mu}_H, \Sigma_H, U] = \int [\boldsymbol{\eta}_{1:T}|\boldsymbol{\eta}_0, \mathbf{h}, U]\,[\mathbf{h}|\boldsymbol{\mu}_H, \Sigma_H]\,d\mathbf{h} = N_{rT}(vec((\eta^{1:T})')|\Upsilon'\boldsymbol{\mu}_H, \Upsilon'\Sigma_H\Upsilon+\tilde{U}), \tag{A.5}$$

where $\tilde{U} := U \otimes I_T$, $\Upsilon := I_r \otimes \eta^{0:(T-1)}$, and $\eta^{t_1:t_2}$ is a matrix with the columns $\boldsymbol{\eta}_{t_1}, \ldots, \boldsymbol{\eta}_{t_2}$. We sample $\{\tilde{\alpha}_{kl}\}$ as a block using adaptive normal MH proposals; then we accept the proposed set of $\{\tilde{\alpha}_{kl}\}$ with a probability that is the ratio of the distribution (A.5) evaluated at the proposed set, divided by (A.5) evaluated at the current set of $\{\tilde{\alpha}_{kl}\}$.

Finally, we would like to sample $\mathbf{h}$ as implied by the second term on the right-hand side of (A.4), which is its full conditional distribution. This distribution is of the form, $N(A_h^{-1}\mathbf{k}_h, A_h^{-1})$, where

$$\begin{aligned} A_h &:= \textstyle\sum_{t=1}^{T}(I_r \otimes \boldsymbol{\eta}_{t-1})U^{-1}(I_r \otimes \boldsymbol{\eta}_{t-1}') + \Sigma_H^{-1} = \Upsilon\tilde{U}^{-1}\Upsilon' + \Sigma_H^{-1} \\ \mathbf{k}_h &:= \textstyle\sum_{t=1}^{T}(I_r \otimes \boldsymbol{\eta}_{t-1})U^{-1}\boldsymbol{\eta}_t + \Sigma_H^{-1}\boldsymbol{\mu}_h = \Upsilon\tilde{U}^{-1}vec(\eta^{1:T\prime}) + \Sigma_H^{-1}\boldsymbol{\mu}_h. \end{aligned} \tag{A.6}$$

Note that we need to invert the $r^2 \times r^2$ matrix A_h, which, depending on the number of basis functions used in an application, can be a very large matrix. We make use of a Sherman-Morrison-Woodbury formula (Sherman and Morrison, 1950; Woodbury, 1950; Henderson and Searle, 1981) to replace the (direct) inversion of A_h with that of an $rT \times rT$ matrix:

$$A_h^{-1} = \Sigma_H - \Sigma_H\Upsilon(\tilde{U} + \Upsilon'\Sigma_H\Upsilon)^{-1}\Upsilon'\Sigma_H. \tag{A.7}$$

However, this matrix can often be too large to hold in memory and to sample from directly. To avoid this, we employ a technique similar to *conditional simulation* used in spatial statistics (for details on spatial conditional simulation, see Cressie, 1993, Sec. 3.6.2). We first sample $\tilde{\mathbf{h}} \sim N_{r^2}(\boldsymbol{\mu}_H, \Sigma_H)$ and $\tilde{\boldsymbol{\xi}}_t \sim N_r(\mathbf{0}, U)$, $t = 1, \ldots, T$. If we set the new vector, $\mathbf{h}^*$ say, to be,

$$\mathbf{h}^* = \tilde{\mathbf{h}} + \Sigma_H\Upsilon(\tilde{U} + \Upsilon'\Sigma_H\Upsilon)^{-1}(vec((\eta^{1:T})') - \Upsilon'\tilde{\mathbf{h}} - \tilde{\boldsymbol{\xi}}_{1:T}), \tag{A.8}$$

then $\mathbf{h}^*|\cdot \sim N_{r^2}(A_h^{-1}\mathbf{k}_h, A_h^{-1})$, which is the correct full conditional distribution (A.6). As can be easily seen from the definitions below equation (A.6), the matrix $(\tilde{U} + \Upsilon'\Sigma_H\Upsilon)$ is a sparse matrix of dimension $rT \times rT$ with at most $r^2T + rT^2 - rT = rT(r + T - 1)$ nonzero elements. This allows for fast sampling of $\mathbf{h}$.

Note that there is actually a change of dimension in the parameter space, induced by sampling $\{\alpha_{kl}\}$. We have not made this explicit in the formulas relating to the updating of $\mathbf{h}$ and $\{\alpha_{kl}\}$, and no explicit reversible-jump MCMC is needed. Additionally, neither calculating (A.5) nor (A.8) requires taking the inverse of Σ_H (which will have some rows and columns that are exactly zero) directly; all terms containing Σ_H^{-1} in (A.5) cancel out after simplifying.

We conclude this section with a short discussion of why regularization of H is important. Assuming that all quantities other than H in the model are fixed, we are essentially trying to make inference on the r^2 elements of H from $T+1$ replications of an r-dimensional vector, $\{\boldsymbol{\eta}_t : \ t = 0, \ldots, T\}$. If we put $\Sigma_H = \omega I_{r^2}$, where $\omega \to \infty$ (i.e., no regularization), we can see from (A.6) that $E(\mathbf{h}|\cdot) = (\Upsilon\tilde{U}^{-1}\Upsilon')^{-1}\Upsilon\tilde{U}^{-1}vec((\eta^{1:T})')$, which is equivalent to the generalized-least-squares estimator in a regression model with data $vec((\eta^{1:T})')$, matrix of covariates Υ, and covariance matrix $\tilde{U}$. That is, we are trying to infer the r^2 elements of $\mathbf{h}$ from the rT-dimensional vector, $vec((\eta^{1:T})')$, a problem that almost demands regularization, especially if $T < r$. By assuming finite prior variances in Σ_H (and even some variances equal to zero, for sparsity), we essentially obtain a ridge-regression estimator for the mean of the full conditional distribution of $\mathbf{h}$.

Bibliography

Antoulas, A. (2005), *Approximation of Large-Scale Dynamical Systems*, Philadelphia, PA: SIAM.

Aubry, N., Lian, W., and Titi, E. (1993), "Preserving symmetries in the proper orthogonal decomposition," *SIAM Journal on Scientific Computing*, 14, 483–505.

Banerjee, S. (2005), "On geodetic distance computations in spatial modeling," *Biometrics*, 61, 617–625.

Banerjee, S., Finley, A. O., Waldmann, P., and Ericsson, T. (2010), "Hierarchical spatial process models for multiple traits in large genetic trials," *Journal of the American Statistical Association*, 105, 506–521.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society, Series B*, 70, 825–848.

Berliner, L. M. (1996), "Hierarchical Bayesian time series models," in *Maximum Entropy and Bayesian Methods*, eds. Hanson, K. and Silver, R., Dordrecht: Kluwer Academic Publishers, pp. 15–22.

Berliner, L. M., Wikle, C. K., and Cressie, N. (2000), "Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling," *Journal of Climate*, 13, 3953–3968.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, New York, NY: Springer.

Calder, C. A. (2007), "Dynamic factor process convolution models for multivariate space-time data with application to air quality assessment," *Environmental and Ecological Statistics*, 14, 229–247.

Calder, C. A., Holloman, C., and Higdon, D. (2002), "Exploring space-time structure in ozone concentration using a dynamic process convolution model," in *Case Studies in Bayesian Statistics*, eds. Gatsonis, C., Kass, R., Carriquiry, A., Gelman, A., Higdon, D., Pauler, D., and Verdinelli, I., New York, NY: Springer, vol. VI, pp. 165–176.

Cangelosi, A. and Hooten, M. B. (2009), "Models for bounded systems with continuous dynamics," *Biometrics*, 65, 850–856.

Carter, C. and Kohn, R. (1994), "On Gibbs sampling for state space models," *Biometrika*, 81, 541–553.

Chahine, M., Pagano, T. S., Aumann, H., Atlas, R., Barnet, C., Blaisdell, J., Chen, L., Divakarla, M., Fetzer, E., Goldberg, M., Gautier, C., Granger, S., Hannon, S., Irion, F., Kakar, R., Kalnay, E., Lambrigtsen, B., Lee, S.-Y., Marshall, J. L., Mcmillian, W. W., Mcmillin, L., Olsen, E. T., Revercomb, H., Rosenkranz, P., Smith, W., Staelin, D., Strow, L., Susskind, J., Tobin, D., Wolf, W., and Zhou, L. (2006), "AIRS - Improving weather forecasting and providing new data on greenhouse gases," *Bulletin of the American Meteorological Society*, 87, 911–926.

Christensen, O. and Waagepetersen, R. (2002), "Bayesian prediction of spatial count data using generalized linear mixed models," *Biometrics*, 58, 280–286.

Cressie, N. (1993), *Statistics for Spatial Data, revised edition*, New York, NY: John Wiley & Sons.

Cressie, N. and Johannesson, G. (2006), "Spatial prediction of massive datasets," in *Mastering the Data Explosion in the Earth and Environmental Sciences: Proceedings of the Australian Academy of Science Elizabeth and Frederick White Conference*, Canberra, Australia: Australian Academy of Science.

— (2008), "Fixed rank kriging for very large spatial data sets," *Journal of the Royal Statistical Society, Series B*, 70, 209–226.

Cressie, N. and Kang, E. L. (2010), *High-resolution digital soil mapping: Kriging for very large datasets*, Dordrecht, NL: Springer, chap. 4, pp. 49–63.

Cressie, N. and Kornak, J. (2003), "Spatial statistics in the presence of location error with an application to remote sensing of the environment," *Statistical Science*, 18, 436–456.

Cressie, N., Shi, T., and Kang, E. L. (2010), "Fixed rank filtering for spatio-temporal data," *Journal of Computational and Graphical Statistics*, 19, 724–745.

Cressie, N. and Wikle, C. K. (2002), *Space-time Kalman filter*, Chichester: John Wiley & Sons, vol. 4, pp. 2045–2049.

— (2011), *Statistics for Spatio-Temporal Data*, Hoboken, NJ: Wiley.

Das, B. (2000), "Global covariance modeling: A deformation approach to anisotropy," Unpublished doctoral dissertation, University of Washington.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Philadelphia, PA: SIAM.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1 – 38.

Dewar, M., Scerri, K., and Kadirkamanathan, V. (2009), "Data-driven spatio-temporal modeling using the integro-difference equation," *IEEE Transactions on Signal Processing*, 57, 83–91.

Doucet, A., De Freitas, N., and Gordon, N. (2001), *Sequential Monte Carlo Methods in Practice*, New York, NY: Springer.

Farrell, B. and Ioannou, P. (2001), "State estimation using a reduced-order Kalman filter," *Journal of the Atmospheric Sciences*, 58, 3666–3680.

Fassò, A. and Cameletti, M. (2009a), "A unified statistical approach for simulation, modeling, analysis and mapping of environmental data," *Simulation*, 86, 139–153.

— (2009b), "The EM algorithm in a distributed computing environment for modelling environmental space-time data," *Environmental Modelling & Software*, 24, 1027–1035.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009), "Improving the performance of predictive process modeling for large datasets." *Computational Statistics & Data Analysis*, 53, 2873–2884.

Frühwirth-Schnatter, S. (1994), "Applied state space modelling of non-Gaussian time series using integration-based Kalman filtering," *Statistics and Computing*, 4, 259–269.

Furrer, R., Genton, M. G., and Nychka, D. W. (2006), "Covariance tapering for interpolation of large spatial datasets," *Journal of Computational and Graphical Statistics*, 15, 502–523.

Furrer, R., Sain, S. R., Nychka, D. W., and Meehl, G. (2007), "Multivariate Bayesian analysis of atmosphere-ocean general circulation models," *Environmental and Ecological Statistics*, 14, 249–266.

Gelpke, V. and Künsch, H. R. (2001), "Estimation of motion from sequences of images: Daily variability of Total Ozone Mapping Spectrometer ozone data," *Journal of Geophysical Research*, 106, 11825–11834.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

George, E. I., Sun, D., and Ni, S. (2008), "Bayesian stochastic search for VAR model restrictions," *Journal of Econometrics*, 142, 553–580.

Gilbert, J. R., Moler, C., and Schreiber, R. (1992), "Sparse Matrices in MATLAB: Design and Implementation," *SIAM Journal on Matrix Analysis and Applications*, 13, 333–356.

Gneiting, T. (2002), "Compactly Supported Correlation Functions," *Journal of Multivariate Analysis*, 83, 493–508.

Gneiting, T. and Raftery, A. E. (2007), "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, 102, 359–378.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation Bayesian model determination," *Biometrika*, 82, 711.

Haario, H., Saksman, E., and Tamminen, J. (2001), "An adaptive Metropolis algorithm," *Bernoulli*, 7, 223–242.

Hamilton, J. (1994), *Time Series Analysis*, Princeton, NJ: Princeton University Press.

Hannan, E. and Deistler, M. (1988), *The Statistical Theory of Linear Systems*, New York, NY: Wiley.

Hastings, W. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Heaton, M. J., Katzfuss, M., Ramachandar, S., Pedings, K., Gilleland, E., Mannshardt-Shamseldin, E., and Smith, R. L. (2011), "Spatio-temporal models for large-scale indicators of extreme weather," *Environmetrics*, 22, 294–303.

Henderson, H. and Searle, S. (1981), "On deriving the inverse of a sum of matrices," *SIAM Review*, 23, 53–60.

Higdon, D. (1998), "A process-convolution approach to modelling temperatures in the North Atlantic Ocean," *Environmental and Ecological Statistics*, 5, 173–190.

— (2002), "Space and space-time modeling using process convolutions," in *Quantitative Methods for Current Environmental Issues*, eds. Anderson, C., Barnett, V., Chatwin, P., and El-Shaarawi, A., London: Springer, pp. 37–56.

Holmes, C. and Mallick, B. (2001), "Bayesian regression with multivariate linear splines," *Journal of the Royal Statistical Society: Series B*, 63, 3–17.

Holmes, C. and Mallick, B. K. (2000), "Bayesian wavelet networks for nonparametric regression," *IEEE Transactions on Neural Networks*, 11, 27–35.

James, A. (1964), "Distributions of matrix variates and latent roots derived from normal samples," *Annals of Mathematical Statistics*, 35, 475–501.

Johannesson, G. and Cressie, N. (2004), "Variance-covariance modeling and estimation for multi-resolution spatial models," in *GeoENV IV - Geostatistics for Environmental Applications*, eds. Sanchez-Vila, X., Carrera, J., and Gomez-Hernandez, J., Kluwer, Dordrecht, pp. 319–330.

Johannesson, G., Cressie, N., and Huang, H.-C. (2007), "Dynamic multi-resolution spatial models," *Environmental and Ecological Statistics*, 14, 5–25.

Jones, R. (1963), "Stochastic processes on a sphere," *Annals of Mathematical Statistics*, 34, 213–218.

Jun, M. and Stein, M. L. (2008), "Nonstationary covariance models for global data," *Annals of Applied Statistics*, 2, 1271–1289.

Kalman, R. (1960), "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, 82, 35–45.

Kang, E. L. and Cressie, N. (2011), "Bayesian inference for the spatial random effects model," *Journal of the American Statistical Association*, 106, in press.

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, 271–289.

Kang, E. L., Liu, D., and Cressie, N. (2009), "Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models," *Computational Statistics & Data Analysis*, 53, 3016–3032.

Kanter, M. (1997), "Unimodal spectral windows," *Statistics & Probability Letters*, 34, 403–411.

Kaplan, A., Cane, M., Kushnir, Y., Clement, A., Blumenthal, M., and Rajagopalan, B. (1998), "Analyses of global sea surface temperature 1856-1991," *Journal of Geophysical Research*, 103, 18567–18589.

Kass, R. and Raftery, A. (1995), "Bayes factors," *Journal of the American Statistical Association*, 90, 773–795.

Katzfuss, M. and Cressie, N. (2009), "Maximum likelihood estimation of covariance parameters in the spatial-random-effects model," in *Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association, pp. 3378–3390.

— (2011a), "Bayesian hierarchical spatio-temporal smoothing for massive datasets," Technical Report No. 853, Department of Statistics, The Ohio State University, Columbus, OH.

— (2011b), "Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets," *Journal of Time Series Analysis*, 32, 430–446.

Kaufman, C., Schervish, M., and Nychka, D. W. (2008), "Covariance tapering for likelihood-based estimation in large spatial data sets," *Journal of the American Statistical Association*, 103, 1545–1555.

Knuth, K. (2005), "Informed source separation: A Bayesian tutorial," in *European Signal Processing Conference*, eds. Sanjur, B., Cetin, E., Tekalp, E., and Kuruoglu, E., Antalya, Turkey.

Kot, M., Lewis, M., and van Den Driessche, P. (1996), "Dispersal data and the spread of invading organisms," *Ecology*, 77, 2027–2042.

Landgrebe, D. A. (2003), *Signal Theory Methods in Multispectral Remote Sensing*, Hoboken, NJ: Wiley.

Lemos, R. T. and Sansó, B. (2009), "A spatio-temporal model for mean, anomaly, and trend fields of North Atlantic sea surface temperature," *Journal of the American Statistical Association*, 104, 5–18.

Litterman, R. (1986), "Forecasting with Bayesian vector autoregressions: Five years of experience," *Journal of Business & Economic Statistics*, 4, 25–38.

Lopes, H. F., Salazar, E., and Gamerman, D. (2008), "Spatial dynamic factor analysis," *Bayesian Analysis*, 3, 759–792.

Mardia, K., Goodall, C., Redfern, E., and Alonso, F. (1998), "The kriged Kalman filter," *Test*, 7, 217–282.

McLachlan, G. and Krishnan, T. (2008), *The EM Algorithm and Extensions*, New York, NY: Wiley.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, 21, 1087–1092.

Møller, J. and Waagepetersen, R. (2004), *Statistical Inference and Simulation for Spatial Point Processes*, Boca Raton, FL: Chapman & Hall/CRC.

Nychka, D. W., Wikle, C. K., and Royle, J. A. (2002), "Multiresolution models for nonstationary spatial covariance functions," *Statistical Modelling*, 2, 315–331.

Paciorek, C. and Schervish, M. (2006), "Spatial modelling using a new class of nonstationary covariance functions," *Environmetrics*, 17, 483–506.

Sahr, K. (2003), "DGGRID Software," `http://webpages.sou.edu/~sahrk/dgg/dggrid/dggrid.html`, version 4.3b.

Shaby, B. and Ruppert, D. (2011), "Tapered Covariance: Bayesian Estimation and Asymptotics," *Journal of Computational and Graphical Statistics*, accepted.

Sherman, J. and Morrison, W. (1950), "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix," *Annals of Mathematical Statistics*, 21, 124–127.

Shi, T. and Cressie, N. (2007), "Global statistical analysis of MISR aerosol data: A massive data product from NASA's Terra satellite," *Environmetrics*, 18, 665–680.

Shumway, R. (2006), "Dynamic mixed models for irregularly observed time series," *Resenhas-Reviews of the Institute of Mathematics and Statistics*, 4, 433–456.

Shumway, R. and Stoffer, D. (1982), "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, 3, 253–264.

— (2006), *Time Series Analysis and Its Applications: With R Examples*, New York, NY: Springer.

Smith, T., Reynolds, R., Livezey, R., and Stokes, D. (1996), "Reconstruction of historical sea surface temperatures using empirical orthogonal functions," *Journal of Climate*, 9, 1403–1420.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York, NY: Springer.

— (2005), "Nonstationary spatial covariance functions," Technical Report No. 21, Center for Integrating Statistical and Environmental Science, The University of Chicago.

— (2008), "A modeling approach for large spatial datasets," *Journal of the Korean Statistical Society*, 37, 3–10.

Stroud, J., Müller, P., and Sansó, B. (2001), "Dynamic models for spatiotemporal data," *Journal of the Royal Statistical Society, Series B*, 63, 673–689.

Stroud, J., Stein, M. L., Lesht, B., and Schwab, D. (2010), "An ensemble Kalman filter and smoother for satellite data assimilation," *Journal of the American Statistical Association*, 105, 978–990.

Sun, Y., Li, B., and Genton, M. G. (2011), "Geostatistics for large datasets," in *Space-Time Processes and Challenges Related to Environmental Problems: Proceedings of the Spring School "Advances And Challenges In Space-Time Modelling Of Natural Events"*, eds. Montero, J., Porcu, E., and Schlather, M., Springer, to appear.

Tzeng, S., Huang, H.-C., and Cressie, N. (2005), "A fast, optimal spatial-prediction method for massive datasets," *Journal of the American Statistical Association*, 100, 1343–1357.

van Dyk, D. A. and Park, T. (2008), "Partially collapsed Gibbs samplers: Theory and methods," *Journal of the American Statistical Association*, 103, 790–796.

Voutilainen, A., Pyhälahti, T., Kallio, K., Pulliainen, J., Haario, H., and Kaipio, J. (2007), "A filtering approach for estimating lake water quality from remote sensing data," *International Journal of Applied Earth Observation and Geoinformation*, 9, 50–64.

Wickerhauser, M. (1994), *Adapted Wavelet Analysis from Theory to Software*, Wellesley, MA: A K Peters.

Wikle, C. K. (2003), "Hierarchical Bayesian models for predicting the spread of ecological processes," *Ecology*, 84, 1382–1394.

— (2010), "Low-rank representations for spatial processes," in *Handbook of Spatial Statistics*, eds. Gelfand, A. E., Fuentes, M., Guttorp, P., and Diggle, P., Boca Raton, FL: Chapman and Hall/CRC, pp. 107 – 118.

Wikle, C. K., Berliner, L. M., and Cressie, N. (1998), "Hierarchical Bayesian space-time models," *Environmental and Ecological Statistics*, 5, 117–154.

Wikle, C. K. and Cressie, N. (1999), "A dimension-reduced approach to space-time Kalman filtering," *Biometrika*, 86, 815–829.

Wikle, C. K. and Hooten, M. B. (2006), "Hierarchical Bayesian spatio-temporal models for population spread," in *Applications of Computational Statistics in the Environmental Sciences: Hierarchical Bayes and MCMC Methods*, eds. Clark, J. and Gelfand, A. E., Oxford, UK: Oxford University Press, pp. 145–169.

— (2010), "A general science-based framework for dynamical spatio-temporal models," *Test*, 19, 417–451.

Wikle, C. K., Milliff, R., Nychka, D. W., and Berliner, L. M. (2001), "Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds," *Journal of the American Statistical Association*, 96, 382–397.

Woodbury, M. (1950), "Inverting modified matrices," Memorandum Report 42, Statistical Research Group, Princeton University.

Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *Annals of Statistics*, 11, 95 – 103.

Xu, B., Wikle, C. K., and Fox, N. (2005), "A kernel-based spatio-temporal dynamical model for nowcasting radar precipitation," *Journal of the American Statistical Association*, 100, 1133–1144.

Xu, K. and Wikle, C. K. (2007), "Estimation of parameterized spatio-temporal dynamic models," *Journal of Statistical Planning and Inference*, 137, 567–588.

Yaglom, A. (1987), *Correlation Theory of Stationary and Related Random Functions, Vol. 1*, New York, NY: Springer.

Zhang, H. (2002), "On estimation and prediction for spatial generalized linear mixed models," *Biometrics*, 58, 129–136.

Zhao, Y., Staudenmayer, J., Coull, B., and Wand, M. P. (2006), "General design Bayesian generalized linear mixed models," *Statistical Science*, 21, 35–51.

Zhu, H., Gu, M., and Peterson, B. (2007), "Maximum likelihood from spatial random effects models via the stochastic approximation expectation maximization algorithm," *Statistics and Computing*, 17, 163–177.