Centre for Statistical & Survey Methodology Working Paper Series

Faculty of Engineering and Information Sciences

2013

Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions

Aritra Sengupta
Ohio State University

Noel Cressie
University of Wollongong, ncressie@uow.edu.au

Recommended Citation

Sengupta, Aritra and Cressie, Noel, Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 7-13, 2013, 56.
https://ro.uow.edu.au/cssmwp/121

NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

07-13

Hierarchical Statistical Modeling of Big Spatial Datasets Using the Exponential Family of Distributions

Aritra Sengupta and Noel Cressie

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email:
anica@uow.edu.au

Hierarchical Statistical Modeling of Big Spatial Datasets Using the Exponential Family of Distributions

Aritra Sengupta[*] Noel Cressie[*][†]

Abstract

Big spatial datasets are very common in scientific problems, such as those involving remote sensing of the earth by satellites, climate-model output, small-area samples from national surveys, and so forth. In this article, our interest lies primarily in very large, non-Gaussian datasets. We consider a hierarchical statistical model consisting of a conditional exponential-family model for the data and an underlying (hidden) geostatistical process for some transformation of the (conditional) mean of the data model. Within this hierarchical model, dimension reduction is achieved by modeling the geostatistical process as a linear combination of a fixed number of spatial basis functions, which results in substantial computational speed-ups. These models do not rely on specifying a spatial-weights matrix, and no assumptions of homogeneity, stationarity, or isotropy are made. Our approach to inference using these models is empirical-Bayesian in nature. We develop maximum likelihood (ML) estimates of the unknown parameters using Laplace approximations in an expectation-maximization (EM) algorithm. We illustrate the performance of the resulting empirical hierarchical model using a simulation study. We also apply our methodology to analyze a remote sensing dataset of aerosol optical depth.

Keywords: Aerosol optical depth; EM algorithm; empirical Bayes; geostatistical process; Laplace approximation; maximum likelihood estimation; MCMC; SRE model

[*] Department of Statistics, The Ohio State University

[†] National Institute for Applied Statistics Research Australia, University of Wollongong, Australia (ncressie@uow.edu.au)

1 Introduction

Big spatial datasets are very common in scientific problems, such as those involving remote sensing of the earth by satellites, climate-model output, small-area samples from national surveys, and so forth. In this article, our interest lies primarily in datasets that are very large and non-Gaussian in form. We consider a hierarchical statistical model consisting of two levels. At the first level, we have an exponential-family model for the data given a spatial process and parameters (which we call the data model). At the second level, we assume a geostatistical process given parameters (which we call the process model), for some transformation of the mean of the data model.

The exponential family of distributions include commonly used continuous and discrete distributions; for a detailed review, see McCullagh and Nelder (1989, Section 2.2.2). All members of the exponential family have a density or probability mass function that can be written as:

$$p(z|\gamma) = \exp\left\{(z\gamma - b(\gamma))/\tau^2 - c(z,\tau)\right\}, \tag{1}$$

where γ is called the canonical parameter or the natural parameter, $b(\gamma)$ is a function that depends only on γ, $c(z,\tau)$ is a function independent of γ, and τ is a scaling constant. The representation above is called the canonical form, or the natural form, of the exponential family.

Here, and in what follows, we use the notation $[A|B]$ to denote the conditional probability distribution of A given B. Suppose we have data, $Z_1,\ldots,Z_n$, coming from a member of the exponential family such that $\{[Z_i|\gamma_1,\ldots,\gamma_n] : i = 1,\ldots,n\}$ are mutually independent, and $[Z_i|\gamma_1,\ldots,\gamma_n] \equiv [Z_i|\gamma_i]$, where $[Z_i|\gamma_i]$ has density given by (1). Then one may proceed by modeling a transformation of the expectation of $[Z_i|\gamma_i]$, namely $E(Z_i|\gamma_i) = b'(\gamma_i)$, as

$$g(E(Z_i|\gamma_i)) = \mathbf{X}_i^\top \boldsymbol{\beta}, \tag{2}$$

where $g(\cdot)$ is the link function, $\mathbf{X}_i$ denotes a p-dimensional vector of known covariates, and $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients. There are a lot of possible choices for $g(\cdot)$. The maximum likelihood (ML) estimator of $\boldsymbol{\beta}$ can be obtained via iteratively reweighted least squares.

For a detailed review of the literature on GLMs, see McCullagh and Nelder (1989) or McCulloch et al. (2001).

When $Z_1, \ldots, Z_n$ are associated with locations in space, the assumption of independence is doubtful. A way to extend the framework above, that takes into account spatial variability, is to replace γ in (1) with a spatial process, $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$, where D is the spatial domain of interest. The covariance between $Y(\mathbf{s})$ and $Y(\mathbf{u})$, for $\mathbf{s}, \mathbf{u} \in D$, is defined as:

$$C_Y(\mathbf{s}, \mathbf{u}) \equiv \mathrm{cov}(Y(\mathbf{s}), Y(\mathbf{u})).$$

Now consider spatial data $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$ from a GLM such that $\{[Z(\mathbf{s}_i)|Y(\cdot)] : i = 1, \ldots, n\}$ are mutually independent, and

$$g(E(Z(\mathbf{s}_i)|Y(\cdot))) = Y(\mathbf{s}_i); \; i = 1, \ldots, n, \tag{3}$$

where $g(\cdot)$ is the link function. The hierarchical modeling framework defined above yields a spatial version of the GLM framework; it was proposed by Diggle et al. (1998), who assumed a Gaussian model for $Y(\cdot)$ and a prior distribution on its parameters. See also Omre and Tjelmeland (1997) for an exposition of the same framework for solving complex problems in petroleum geostatistics.

Lindley and Smith (1972) introduced a Bayesian-linear-model framework, where conditional and prior distributions come from a multivariate Gaussian distribution. In the spatial context, Omre (1987) defined Bayesian kriging for the linear model; for further extensions, see Cressie (1993, Sec. 3.4.4). Besag et al. (1991) showed how a spatial model for counts in small areas could be decomposed hierarchically, where the hidden process $Y(\cdot)$ was used to model the spatial dependence. They assumed that the counts were (conditionally) Poisson distributed, and that the log means were a Gaussian spatial process, specifically a Gaussian Markov Random Field (MRF) known as the conditional autoregressive (CAR) model. However, a simultaneous autoregressive (SAR) model, or a geostatistical model could also be used. Indeed Diggle et al. (1998) employed spatial generalized linear mixed models (GLMMs) for spatially dependent non-Gaussian variables observed potentially anywhere in D, and they assumed a hidden geostatistical processes $Y(\cdot)$ with both fixed effects and random effects. Their hierarchical model was fully Bayesian and required a Markov chain Monte Carlo (MCMC) algorithm to obtain the posterior distribution. In a spatio-temporal context, Wikle et al. (1998) developed a fully Bayesian hierarchical-model formulation for modeling a dataset of monthly maximum temperatures.

In contrast, Heagerty and Lele (1998) developed a method for binary data where they used a composite-likelihood (e.g., Lindsay, 1988) approach to estimate the spatial hierarchical model parameters. Zhang (2002) gave a Monte Carlo version of the EM Gradient Algorithm to analyze non-Gaussian data, and Monestiez et al. (2006) developed a method called Poisson kriging for mapping the relative abundance of species.

Despite the popularity of the spatial models discussed above, these models might suffer from two major drawbacks: (1) there might be spatial confounding, and (2) there is often a computational bottleneck when the size of the dataset is large. Spatial confounding between the fixed and the random effects was pointed out in articles by Reich et al. (2006), Hodges and Reich (2010), and Paciorek (2010). Reich et al. (2006) and Hodges and Reich (2010) proposed a modeling approach that gets around the problem of spatial confounding by introducing random effects that are orthogonal to the column space of the matrix of covariates. We shall discuss this in more detail in Section 2.2.

The computational bottleneck arises due to the general computational cost of $O(n^3)$ to obtain the inverse of an $n \times n$ covariance matrix. It is often referred to as a “big n” problem. Many geophysical and environmental datasets are high-dimensional. When the data are Gaussian, reduced-rank-modeling approaches for the hidden Gaussian process $Y(\cdot)$ have been developed to deal with this computational challenge (e.g., Wikle et al., 2001; Cressie and Johannesson, 2006, 2008; Banerjee et al., 2008; Stein, 2008; Lopes et al., 2008). When the data are non-Gaussian, Lopes et al. (2011) take the GLMM approach in Diggle et al. (1998), but with reduced-rank factor analysis models for $Y(\cdot)$ in place of the intrinsically stationary models used by Diggle et al. (1998). A number of spatial and spatio-temporal applications for very-large-to-massive datasets center around this reduced-rank representation of the hidden continuous Gaussian process (e.g., see the review in Wikle, 2010).

The reduced-rank methods discussed above are based on geostatistical models, where a continuously indexed Gaussian process $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ is used to specify the hidden process. In the case where $D \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ is a spatial lattice of sites, a geostatistical model for $\mathbf{Y} \equiv (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_N))^\top$ can still be used; such a model captures the spatial dependence through the covariance matrix, $\mathbf{\Sigma}_Y \equiv \mathrm{cov}(\mathbf{Y})$.

A Gaussian MRF that is used to capture the spatial dependence in $\mathbf{Y}$, does so through the (typically sparse) precision matrix $\mathbf{\Sigma}_Y^{-1}$. A detailed discussion of this can be found in Rue and Held (2005, Chapter 5) and Cressie and Wikle (2011, Pages 185-186). Rue and Held (2005, Chapter 5) discuss a way to approximate a geostatistical model with a sparse CAR model, and this relationship has been used by Lindgren et al. (2011) and Simpson et al. (2012) to build hierarchical spatial models with Gaussian-MRF process models that allow fast computations. However, by necessity, they use only a small number of parameters, which could be problematic when modeling spatial dependence over large, continental-scale, heterogeneous regions. In a recent article, Hughes and Haran (2013) consider a Bayesian hierarchical model with a hidden Gaussian MRF and use a dimension-reduction approach to deal with spatial confounding and computational complexity that arise when analyzing a large spatial dataset. They parameterize the precision matrix using an underlying graph, $G = (V, E)$, where edges represent spatial dependence, and they assume only a small number of parameters.

In this article, we assume that there are small areas $\{A_i : i = 1, \ldots, N\}$ at locations $D \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$, respectively. The order of the small areas is immaterial, so we choose to order them such that $A_1, \ldots, A_n$ have observations $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$, respectively, associated with them, where $n \leq N$. Define the observation vector (i.e., data) to be

$$\mathbf{Z}_O = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^\top;\ 1 \leq n \leq N.$$

We propose a flexible class of spatial models for analyzing these (potentially) non-Gaussian lattice data. The models are hierarchical, where the data model comes from the exponential family of distributions, and the process model is geostatistical and nonstationary (Section 2). These models are computationally efficient to implement, and we take an empirical hierarchical modeling (EHM) approach where any unknown parameters are estimated by ML estimation. Hence, the model is not fully Bayesian, but Bayes' Theorem is used to obtain the all-important predictive distribution; for the special case where data are spatial counts, we have demonstrated its feasibility (Sengupta and Cressie, 2013). For a more complete discussion of the EHM approach, see Cressie and Wikle (2011, Chapter 2).

Our spatial statistical analysis of the lattice data $\mathbf{Z}_O$ is a combination of the GLMM framework of Diggle et al. (1998), the use of the Spatial Random Effects (SRE) model of Cressie and Johannesson (2006, 2008), developed for Gaussian data with a continuous spatial index, and a fast EM algorithm for estimating any unknown parameters. The SRE model is a geostatistical model that achieves dimension reduction by modeling the underlying spatial process as a linear combination of specified spatial basis functions on a spatially continuous domain; in what is to follow, we use it on a discrete spatial lattice. The dimension reduction is important for spatial best linear unbiased prediction (i.e., kriging), since it involves inverting the $n \times n$ covariance matrix of $\mathbf{Z}_O$. Using the SRE model, the matrix inversion is a relatively simple task, the model is well suited to change-of-support, and it avoids any stationarity assumptions for the covariance matrix. Unlike the model used in Lopes et al. (2011), the SRE model does not assume a diagonal covariance matrix for the spatial random effects. Instead, it captures spatial-statistical dependence using both the modeler-specified spatial basis functions *and* correlated random effects. Assuming the data are Gaussian, Katzfuss and Cressie (2009) gave an EM algorithm to obtain ML estimates for SRE-model parameters; and there is also a Bayesian-hierarchical-model (BHM) version that puts prior distributions on the parameters rather than estimating them (Kang and Cressie, 2011).

When the data are non-Gaussian, estimation of the parameters in a hierarchical statistical model is not as straightforward. In the EHM proposed in Section 2, we use the EM algorithm (Dempster et al., 1977) to obtain ML estimates of the parameters in the model. Since the expectations in the E-step of the algorithm are not available in closed form, we use a Laplace approximation to approximate the intractable integrals. Having obtained the estimates for the unknown parameters, we substitute them into the predictive distribution and use an MCMC algorithm to generate samples from it. Thus, our use of EHM for non-Gaussian data, with parameter estimates substituted into optimal predictors, is the direct analogue of kriging (used ubiquitously in geostatistical and environmental applications). We handle big spatial datasets by embedding the SRE model into our hierarchical statistical model.

The plan of this article is as follows. In Section 2, we describe a hierarchical model for non-Gaussian spatial data, whose data model comes from the exponential family and whose process model is based on a hidden SRE model. We also address the issue of spatial confounding in Section 2. In Section 3, we outline statistical inference based on generating MCMC samples from the predictive distribution. Then, in Section 4, we describe the EM algorithm for obtaining ML estimates of the model parameters described in Section 2. In Section 5, we carry out a simulation experiment to assess the performance of our EHM approach. In Section 6, we use our EHM approach to analyze a large, spatial, remote sensing dataset of aerosol optical depth (AOD) from the MISR instrument on the Terra satellite. Discussion and conclusions follow in Section 7, and technical derivations are given in the Appendix.

2 Hierarchical Statistical Model

In this section, we give details of the hierarchical statistical model that we use to model non-Gaussian data. Specifically, the *data model* comes from the exponential family of distributions, and the *process model* is a (transformed) Gaussian spatial process. We consider lattice data obtained from among small areas $\{A_i : i = 1, \ldots, N\}$, located at $\{\mathbf{s}_i : i = 1, \ldots, N\}$, respectively, although some locations have missing data. Thus, the spatial domain is the discrete spatial lattice $D \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$. Without loss of generality, the locations where there are observations are denoted as $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset D$, where $1 \leq n \leq N$. Hence, the set of unobserved locations are $\{\mathbf{s}_i : i = n+1, \ldots, N\}$, if $n < N$.

2.1 Components of the Hierarchical Statistical Model

1. Conditional distribution of the data given the process (data model)

Recall $\mathbf{Z}_O = (Z(\mathbf{s}_1),\ldots,Z(\mathbf{s}_n))^\top$ denotes the vector of observations, and $Y(\mathbf{s})$ denotes the hidden process at location $\mathbf{s} \in D$. Further, define the random process $Y(\cdot) \equiv \{Y(\mathbf{s}) : \mathbf{s} \in D\}$. Then assume that $[Z(\mathbf{s}_i)|Y(\cdot)] = [Z(\mathbf{s}_i)|Y(\mathbf{s}_i)]$, and furthermore that it is a member of the exponential family (e.g., McCullagh and Nelder, 1989, Chapter 2). Conditional independence of the data given the process yields,

$$[\mathbf{Z}_O|Y(\cdot)] = \prod_{i=1}^{n} [Z(\mathbf{s}_i)|Y(\mathbf{s}_i)],$$

where

$$Z(\mathbf{s}_i)|Y(\mathbf{s}_i) \sim \text{ ind. exponential family}\left(\mu_{Z|Y}(\mathbf{s}_i), V(\mu_{Z|Y}(\mathbf{s}_i))\right),\ i = 1,\ldots n; \tag{4}$$

the conditional mean, $\mu_{Z|Y}(\mathbf{s}_i) \equiv E(Z(\mathbf{s}_i)|Y(\mathbf{s}_i))$, depends on $Y(\mathbf{s}_i)$; and the variance of the conditional distribution, $[Z(\mathbf{s}_i)|Y(\mathbf{s}_i)]$, is expressed as a function of the conditional mean through $V(\mu_{Z|Y}(\mathbf{s}_i))$. The function $V(\cdot)$ denotes the mean-variance relationship for the exponential family. The distribution in (4) can be written as:

$$f_{Z|Y}(z(\mathbf{s}_i)|Y(\mathbf{s}_i)) = \exp\left\{(z(\mathbf{s}_i)\gamma(\mathbf{s}_i) - b(\gamma(\mathbf{s}_i)))/\tau^2 - c(z(\mathbf{s}_i),\tau)\right\}, \tag{5}$$

where for convenience we have written the distribution in its *canonical form*. The quantities $\gamma(\mathbf{s}_i)$ and $b(\gamma(\mathbf{s}_i))$ depend on $Y(\mathbf{s}_i)$ in a way determined by which member of the exponential family in (4) is chosen.

2. Link function

We proceed by modeling a transformation, $g(\cdot)$, of the mean $\mu_{Z|Y}(\cdot)$ as a sum of the two components:

$$g(\mu_{Z|Y}(\mathbf{s})) = t(\mathbf{s}) + \nu(\mathbf{s});\ \mathbf{s} \in D, \tag{6}$$

where $g(\mu_{Z|Y}(\mathbf{s}))$ is the *link function* evaluated at the (conditional) mean, $t(\mathbf{s})$ is deterministic large-scale spatial variation (or the trend term), and $\nu(\mathbf{s})$ denotes random, mean-zero, small-scale spatial variation, which is assumed to be a Gaussian process. If $g(\mu_{Z|Y}(\cdot)) \equiv \gamma(\cdot)$ in (5), then $g(\cdot)$ is the canonical link function, which plays an important role in the GLM (McCullagh and Nelder, 1989, Section 2.2.3). Examples of canonical links include the logit link for the Binomial distribution, the log link for the Poisson distribution, and the inverse link for the Gamma distribution. However, the canonical link is not the only choice. Some popular non-canonical links include the probit link for the Binomial distribution and the log link for the Gamma distribution (Section 6.2).

3. Process model

   The process $Y(\cdot)$ is defined as:

   $$Y(\cdot) \equiv g(\mu_{Z|Y}(\cdot)). \tag{7}$$

   Thus, $Y(\cdot)$ is related to the mean of the observed process through the link function. If we work with the canonical link, we have the special case $Y(\cdot) \equiv \gamma(\cdot)$.

   From (6),

   $$Y(\cdot) = t(\cdot) + \nu(\cdot), \tag{8}$$

   where recall that $t(\cdot)$ is the *deterministic* spatial trend and $\nu(\cdot)$ is a *random* mean-zero spatial Gaussian process.

4. Spatial trend

   The trend, or large-scale spatial variation, is modeled as a linear combination of known covariates, $\mathbf{X}(\mathbf{s}) \equiv (X_1(\mathbf{s}), \ldots, X_p(\mathbf{s}))^\top$:

   $$t(\mathbf{s}) = C(\mathbf{s}) + \mathbf{X}(\mathbf{s})^\top \boldsymbol{\beta}, \tag{9}$$

   where $C(\mathbf{s})$ is a known offset term, and $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression coefficients that need to be estimated. Recall that $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_N))^\top$, and hence (8) becomes,

   $$\mathbf{Y} = \mathbf{C} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\nu}, \tag{10}$$

where $\mathbf{X} \equiv \left(\mathbf{X}_O^\top, \mathbf{X}_U^\top\right)^\top$, $\mathbf{X}_O \equiv (\mathbf{X}(\mathbf{s}_1), \ldots, \mathbf{X}(\mathbf{s}_n))^\top$, $\mathbf{X}_U \equiv (\mathbf{X}(\mathbf{s}_{n+1}), \ldots \mathbf{X}(\mathbf{s}_N))^\top$, $\boldsymbol{\nu} \equiv \left(\boldsymbol{\nu}_O^\top, \boldsymbol{\nu}_U^\top\right)^\top$, $\boldsymbol{\nu}_O \equiv (\nu(\mathbf{s}_1), \ldots, \nu(\mathbf{s}_n))^\top$, $\boldsymbol{\nu}_U \equiv (\nu(\mathbf{s}_{n+1}), \ldots, \nu(\mathbf{s}_N))^\top$, and $\mathbf{C} \equiv (C(\mathbf{s}_1), \ldots, C(\mathbf{s}_N))^\top$.

5. Spatial Random Effects (SRE) model for $\nu(\cdot)$

We use a geostatistical model for $\nu(\cdot)$, in contrast to the MRF used by Besag et al. (1991) and Lindgren et al. (2011). In what follows, $\text{Gau}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is an abbreviation for a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The possibility of big data, $\mathbf{Z}_O$, motivates us to propose the Spatial Random Effects (SRE) model:

$$\nu(\cdot) = \mathbf{S}(\cdot)^\top \boldsymbol{\eta} + \xi(\cdot), \tag{11}$$

where $\mathbf{S}(\cdot)$ is an r-dimensional vector of known spatial basis functions; $\boldsymbol{\eta}$ is a vector of random effects that is assumed to have a $\text{Gau}(\mathbf{0}, \mathbf{K})$ distribution; and $\xi(\cdot)$ is a fine-scale-variation component that is assumed to be spatially independent with a $\text{Gau}(0, \text{v}_\xi(\cdot)\sigma_\xi^2)$ distribution and $\text{v}_\xi(\cdot)$ known. Other possible approaches to spatial prediction where datasets are very-large-to-massive are discussed in Section 1.

Recall that $|D| = N \geq n$, where n may be very large; however, the random-effects vector $\boldsymbol{\eta}$ is only of dimension r ($r \ll n$). We do not assume any particular structure for the $r \times r$ covariance matrix $\mathbf{K}$, nor do we necessarily try to parameterize it using just a few parameters. The spatial dependence in $\mathbf{Y}$ is captured using both $\mathbf{K}$ and the spatial basis functions $\mathbf{S}(\cdot)$. Dimension reduction is achieved by modeling the underlying N-dimensional spatial process as a linear combination of r fixed spatial basis functions over the entire spatial domain of interest. In Section 5, we show that this leads to substantial computational gain, which is especially significant when dealing with very large datasets. As well as computational speed-ups, the hierarchical model given by (5), (10), and (11) avoids making second-order stationarity assumptions, and it is well suited to change-of-support.

2.2 Spatial Confounding of Fixed and Random Effects

Our interest in this article lies primarily in inference on the hidden spatial process $Y(\cdot)$ or, equivalently, in inference on $\mu_{Z|Y}(\cdot) = g^{-1}(Y(\cdot))$. That is, we wish to predict $Y(\cdot)$ over the entire spatial domain D, based on the data $\mathbf{Z}_O = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^\top$. We first discuss confounding for the case where there is no dimension reduction, namely for a full-rank spatial generalized linear mixed model (SGLMM). The process model for a full-rank SGLMM is given by:

$$g(\mu_{Z|Y}(\cdot)) = \mathbf{X}(\cdot)^\top \boldsymbol{\beta} + \nu(\cdot), \tag{12}$$

where recall that $\mathbf{X}(\cdot)$ is a p-dimensional vector of known covariates, $\boldsymbol{\beta}$ is a p-dimensional vector of fixed but unknown regression coefficients, and $\nu(\cdot)$ is the random effect. Define $\mathbf{g}_O \equiv (g(\mu_{Z|Y}(\mathbf{s}_1)), \ldots, g(\mu_{Z|Y}(\mathbf{s}_n)))^\top$, and rewrite (12) in vector notation as,

$$\mathbf{g}_O = \mathbf{X}_O \boldsymbol{\beta} + \boldsymbol{\nu}_O = \mathbf{X}_O \boldsymbol{\beta} + \mathbf{I}_n \boldsymbol{\nu}_O, \tag{13}$$

where $\mathbf{X}_O \equiv (\mathbf{X}(\mathbf{s}_1), \ldots, \mathbf{X}(\mathbf{s}_n))^\top$, and $\boldsymbol{\nu}_O \equiv (\nu(\mathbf{s}_1), \ldots, \nu(\mathbf{s}_n))^\top$. The last equality emphasizes the matrix coefficients of the fixed and random effects. Reich et al. (2006) and Hodges and Reich (2010) used a reparameterization of (13) to show that such a SGLMM exhibits spatial confounding for fully Bayesian inference. Specifically, posterior inference for $\boldsymbol{\beta}$ tends to be biased, and its posterior variance is inflated. This happens because a subspace of the column space of $\mathbf{I}_n$ coincides with the column space of $\mathbf{X}_O$ (see Paciorek, 2010). They also proposed a way to mitigate this spatial confounding by setting some random effects equal to zero, but Hughes and Haran (2013) pointed out that for a Gaussian MRF, this can result in negative spatial dependence. Hughes and Haran (2013) proposed a model that alleviates spatial confounding, reduces the dimension of the random effects, and only allows for positive spatial dependence among the random effects.

Our approach to modeling is also based on reducing the dimension of the random effects. We use spatial basis functions to achieve dimension reduction but allow general dependence between the random effects. Recall the SRE model (11), which gives

$$\boldsymbol{\nu}_O = \mathbf{S}_O\boldsymbol{\eta} + \boldsymbol{\xi}_O, \tag{14}$$

where $\mathbf{S}_O \equiv (\mathbf{S}(\mathbf{s}_1), \ldots, \mathbf{S}(\mathbf{s}_n))^\top$ is typically sparse, and $\boldsymbol{\xi}_O \equiv (\xi(\mathbf{s}_1), \ldots, \xi(\mathbf{s}_n))^\top$. The basis functions are introduced to capture the small-scale spatial variation in the model, and their optimal choice is an area of ongoing research (e.g., Bradley et al., 2011). As long as $\mathbf{X}_O$ is not perfectly collinear with $\mathbf{S}_O$, the large-scale variability that is captured by the fixed-effects component will not be fully explained by the random effects. In this article, we take an empirical-Bayesian approach, where we use the EM algorithm to estimate the unknown parameters (Section 4), and then we substitute in the estimates to obtain MCMC samples from the empirical predictive distribution (Section 3). That is, the EM estimate of $\boldsymbol{\beta}$ (and $\mathbf{K}$ and σ^2_ξ) is held fixed in the MCMC, which is consistent with the treatment of large-scale variation in kriging when, in practice, the spatial trend (and the variogram) is unknown and has to be estimated (e.g., Cressie, 1993, Section 3.5). When $\boldsymbol{\beta}$ is held fixed in the MCMC, (empirical) Bayesian inference on the random-effects term is no longer confounded. Consequently, an EHM approach mitigates spatial confounding in the SGLMM (12) used in the process model.

3 Empirical-Bayesian Inference

Our main focus in this paper is on prediction of $Y(\cdot)$ or of $\mu_{Z|Y}(\cdot)$. That is, after having observed $\mathbf{Z}_O$ at locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, we wish to make inference on $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_N))^\top$ or some function of $\mathbf{Y}$. The parameters $\boldsymbol{\theta} \equiv \left\{\boldsymbol{\beta}, \mathbf{K}, \sigma^2_\xi\right\}$ are also of interest, but instead of putting a prior distribution on them, we *estimate* them using an EM algorithm (Section 4). Our hierarchical model becomes an empirical hierarchical model when we substitute the estimated parameters $\hat{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$, into the predictive distribution, $[\mathbf{Y}|\mathbf{Z}_O, \boldsymbol{\theta}]$. With a slight abuse of notation, we write this as $[\mathbf{Y}|\mathbf{Z}_O, \hat{\boldsymbol{\theta}}]$ and refer to it as the *empirical* predictive distribution.

Recall that $\mathbf{Z}_O = (Z(\mathbf{s}_1),\ldots,Z(\mathbf{s}_n))^\top$, and write $\mathbf{Y} \equiv \left(\mathbf{Y}_O^\top, \mathbf{Y}_U^\top\right)^\top$, where

$$\mathbf{Y}_O \equiv (Y(\mathbf{s}_1),\ldots,Y(\mathbf{s}_n))^\top, \text{ and } \mathbf{Y}_U \equiv (Y(\mathbf{s}_{n+1}),\ldots,Y(\mathbf{s}_N))^\top.$$

Similarly, $\mathbf{X} \equiv \left(\mathbf{X}_O^\top, \mathbf{X}_U^\top\right)^\top$, $\mathbf{S} \equiv \left(\mathbf{S}_O^\top, \mathbf{S}_U^\top\right)^\top$, and $\boldsymbol{\xi} \equiv \left(\boldsymbol{\xi}_O^\top, \boldsymbol{\xi}_U^\top\right)^\top$. Now,

$$\begin{aligned}
[\boldsymbol{\xi}_U|\mathbf{Z}_O,\boldsymbol{\eta},\boldsymbol{\xi}_O,\boldsymbol{\theta}] &= \frac{[\boldsymbol{\xi}_O,\boldsymbol{\xi}_U,\mathbf{Z}_O,\boldsymbol{\eta},|\boldsymbol{\theta}]}{[\boldsymbol{\xi}_O,\mathbf{Z}_O,\boldsymbol{\eta},|\boldsymbol{\theta}]} \\
&= \frac{[\mathbf{Z}_O|\boldsymbol{\eta},\boldsymbol{\xi}_O,\boldsymbol{\theta}][\boldsymbol{\eta}|\mathbf{K}][\boldsymbol{\xi}_O|\sigma_\xi^2][\boldsymbol{\xi}_U|\sigma_\xi^2]}{\int[\mathbf{Z}_O|\boldsymbol{\eta},\boldsymbol{\xi}_O][\boldsymbol{\eta}|\mathbf{K}][\boldsymbol{\xi}_O|\sigma_\xi^2][\boldsymbol{\xi}_U|\sigma_\xi^2]d\boldsymbol{\xi}_U} \\
&= [\boldsymbol{\xi}_U|\sigma_\xi^2].
\end{aligned} \tag{15}$$

Thus, given $\boldsymbol{\theta}$, $\boldsymbol{\xi}_U$ is conditionally independent of $(\mathbf{Z}_O,\boldsymbol{\eta},\boldsymbol{\xi}_O)$, and hence for an unobserved site in $\{\mathbf{s}_i : i = n+1,\ldots,N\}$, we have:

$$\begin{aligned}
E\left(Y(\mathbf{s}_i)|\mathbf{Z}_O,\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2\right) &= C(\mathbf{s}_i) + \mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta} + \mathbf{S}(\mathbf{s}_i)^\top E\left(\boldsymbol{\eta}|\mathbf{Z}_O,\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2\right) \\
\text{var}\left(Y(\mathbf{s}_i)|\mathbf{Z}_O,\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2\right) &= \mathbf{S}(\mathbf{s}_i)^\top\text{var}\left(\boldsymbol{\eta}|\mathbf{Z}_O,\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2\right)\mathbf{S}(\mathbf{s}_i) + \sigma_\xi^2 v_\xi(\mathbf{s}_i).
\end{aligned} \tag{16}$$

For a site $\mathbf{s}_i \in \{\mathbf{s}_1,\ldots,\mathbf{s}_n\}$, where an observation is available, we have

$$\begin{aligned}
E\left(Y(\mathbf{s}_i)|\mathbf{Z}_O,\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2\right) =& C(\mathbf{s}_i) + \mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta} + \mathbf{S}(\mathbf{s}_i)^\top E\left(\boldsymbol{\eta}|\mathbf{Z}_O,\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2\right) + E\left(\xi(\mathbf{s}_i)|\mathbf{Z}_O,\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2\right) \\
\text{var}\left(Y(\mathbf{s}_i)|\mathbf{Z}_O,\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2\right) =& \mathbf{S}(\mathbf{s}_i)^\top\text{var}\left(\boldsymbol{\eta}|\mathbf{Z}_O,\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2\right)\mathbf{S}(\mathbf{s}_i) + \text{var}\left(\xi(\mathbf{s}_i)|\mathbf{Z}_O,\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2\right) \\
&+ 2S(\mathbf{s}_i)^\top\text{cov}\left(\boldsymbol{\eta},\xi(\mathbf{s}_i)|\mathbf{Z}_O,\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2\right).
\end{aligned} \tag{17}$$

The goal here is to predict $\mathbf{Y}$ (or some function of $\mathbf{Y}$), given the data. However, the predictive distribution, $[\mathbf{Y}|\mathbf{Z}_O,\boldsymbol{\theta}]$, is not available in closed form, nor is $\boldsymbol{\theta}$ known. We shall use a combination of EM estimation of $\boldsymbol{\theta}$ to yield $\hat{\boldsymbol{\theta}}_{EM}$, and we shall use an MCMC algorithm (see, e.g., Robert and Casella, 2004) to yield samples from the predictive distribution, $[\mathbf{Y}|\mathbf{Z}_O,\boldsymbol{\theta}]$, where $\hat{\boldsymbol{\theta}}_{EM}$ is substituted in for $\boldsymbol{\theta}$. In actuality, this is achieved by obtaining samples from the predictive distribution, $[\boldsymbol{\eta},\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}]$, and the distribution $[\boldsymbol{\xi}_U|\sigma_\xi^2]$, where $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{EM}$ and $\sigma_\xi^2 = \hat{\sigma}_{\xi;EM}^2$ are respectively substituted in. The EM algorithm to obtain $\hat{\boldsymbol{\theta}}_{EM}$ is presented in the next section, where it is seen that the E-step cannot be evaluated exactly; we propose a Laplace approximation. The MCMC algorithm to obtain the predictive distribution is described in the Appendix.

4 EM Estimation of Parameters

In this section, we obtain the ML estimates of the parameters using the EM algorithm. The EM algorithm (Dempster et al., 1977) has been employed for estimation of parameters in the presence of missing data; for more details, see McLachlan and Krishnan (2008). For the hierarchical model described in Section 2, the random effects, $\boldsymbol{\eta}$, and the fine-scale variation, $\boldsymbol{\xi}_O$, are not known and can be treated as "data" that complete the likelihood. The EM algorithm involves iterating between an E (expectation)-step and an M (maximization)-step, and in our case the E-step is the most problematic. We resolve this problem by using Laplace approximations to evaluate the expectations required in the E-step.

Recall that

$$g(\mu_{Z|Y}(\cdot)) = Y(\cdot),$$

where $g(\cdot)$ is the link function. We now rewrite $\gamma(\cdot)$ and $b(\gamma(\cdot))$ in (5) as functions of $Y(\cdot)$. Define:

$$\begin{aligned} \gamma(\cdot) &\equiv h_1(Y(\cdot)) \\ b(\gamma(\cdot)) &\equiv h_2(Y(\cdot)). \end{aligned} \tag{18}$$

Then, under this re-parameterization, the conditional density of $[Z(\mathbf{s})|Y(\mathbf{s})]$, for $\mathbf{s} \in \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, is given by:

$$f_{Z|Y}(z(\mathbf{s})) = \exp\left\{(z(\mathbf{s})h_1(Y(\mathbf{s})) - h_2(Y(\mathbf{s})))/\tau^2 - c(z(\mathbf{s}), \tau)\right\}. \tag{19}$$

Note that if the canonical link is considered, we have $\gamma(\cdot) = Y(\cdot)$, and hence

$$\begin{aligned} h_1(Y(\cdot)) &= Y(\cdot) \\ h_2(Y(\cdot)) &= b(Y(\cdot)). \end{aligned} \tag{20}$$

The "complete data" log likelihood, L_c, for the unknown parameters is made up of the observations $\mathbf{Z}_O$ and the unobserved $\boldsymbol{\eta}$ and $\boldsymbol{\xi}_O$. Then L_c is simply the logarithm of the joint distribution of $\mathbf{Z}_O$, $\boldsymbol{\eta}$, and $\boldsymbol{\xi}_O$, given the parameters $\boldsymbol{\theta} = \left\{\boldsymbol{\beta}, \mathbf{K}, \sigma_\xi^2\right\}$. That is,

$$\begin{aligned} L_c(\boldsymbol{\theta}|\mathbf{Z}_O, \boldsymbol{\eta}, \boldsymbol{\xi}_O) =& \log\left[\mathbf{Z}_O|\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\xi}_O\right] + \log\left[\boldsymbol{\eta}|\mathbf{K}\right] + \log\left[\boldsymbol{\xi}_O|\sigma_\xi^2\right] \\ =& \text{const.} + \left\{\sum_{i=1}^n Z(\mathbf{s}_i) h_1(C(\mathbf{s}_i) + \mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta} + \mathbf{S}(\mathbf{s}_i)^\top \boldsymbol{\eta} + \xi(\mathbf{s}_i)) \right. \\ & \left. - \sum_{i=1}^n h_2(C(\mathbf{s}_i) + \mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta} + \mathbf{S}(\mathbf{s}_i)^\top \boldsymbol{\eta} + \xi(\mathbf{s}_i))\right\} / \tau^2 \\ & - \frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\text{trace}\left(\boldsymbol{\eta}\boldsymbol{\eta}^\top \mathbf{K}^{-1}\right) - \frac{n}{2}\log\sigma_\xi^2 - \frac{1}{2\sigma_\xi^2}\text{trace}\left(\boldsymbol{\xi}_O\boldsymbol{\xi}_O^\top \mathbf{V}_{\xi;O}^{-1}\right), \end{aligned} \tag{21}$$

where recall that $[\mathbf{A}|\mathbf{B}]$ denotes the density function of $\mathbf{A}$ given $\mathbf{B}$, $\mathbf{V}_{\xi;O} \equiv \text{diag}(\text{v}_\xi(\mathbf{s}_1), \ldots, \text{v}_\xi(\mathbf{s}_n))$, and "const." denotes a generic constant that does not depend on $\boldsymbol{\theta}$. The EM algorithm is based on L_c and an iteration procedure that we now describe. Assume we have completed the l-th iteration of the EM algorithm; that is, we have an estimate $\boldsymbol{\theta}^{[l]}$ of $\boldsymbol{\theta}$.

4.1 The E-step

At the $(l+1)$-th iteration, the E-step is:

$$\begin{aligned}
Q(\boldsymbol{\theta},\boldsymbol{\theta}^{[l]}) \equiv\ & E\left(L_c(\boldsymbol{\theta}|\mathbf{Z}_O,\boldsymbol{\eta},\boldsymbol{\xi}_O)|\boldsymbol{\theta}^{[l]}\right) \\
=\ & \text{const.} + \left\{\sum_{i=1}^{n} Z(\mathbf{s}_i)E\left(h_1(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\boldsymbol{\eta}+\xi(\mathbf{s}_i))|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right)\right. \\
& \left.-\sum_{i=1}^{n} E\left(h_2(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\boldsymbol{\eta}+\xi(\mathbf{s}_i))|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right)\right\}/\tau^2 \\
& -\frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\text{trace}\left(E\left(\boldsymbol{\eta}\boldsymbol{\eta}^\top|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right)\mathbf{K}^{-1}\right) \\
& -\frac{n}{2}\log\sigma_\xi^2 - \frac{1}{2\sigma_\xi^2}\text{trace}\left(E\left(\boldsymbol{\xi}_O\boldsymbol{\xi}_O^\top|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right)\mathbf{V}_{\xi;O}^{-1}\right). \qquad (22)
\end{aligned}$$

The expectations involved in the E-step of the EM algorithm are with respect to the unobserved variables $\boldsymbol{\eta}$ and $\boldsymbol{\xi}_O$, and they are not available in closed form.

When the integrals in the E-step are problematic, one approach may be to implement a stochastic EM (SEM) algorithm (e.g., see Robert and Casella, 2004; McLachlan and Krishnan, 2008), where the expectations are evaluated using Monte Carlo integration. When datasets are large, this computation can be very slow, and hence the EM algorithm can be very slow to converge. In our approach, we derive Laplace approximations (LA) to approximate the expectations involved in (22), which are based on second-order Taylor-series expansions of the logarithm of the integrands around their respective modes.

To apply the LA, we need to obtain the mode, $(\hat{\boldsymbol{\eta}}^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})$, of L_c considered as a function of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}_O$. Sengupta and Cressie (2013) use a coordinate-wise ascent method for the Poisson GLM and canonical log link, which maximizes alternately with respect to $\boldsymbol{\eta}$, and then with respect to $\boldsymbol{\xi}_O$, until convergence. We do the same here for the general hierarchical model described in Section 2.

We use a second-order Taylor-series approximation to approximate the posterior distribution of $[\boldsymbol{\eta},\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}]$ with a Gaussian distribution with mean and variance given by the posterior mode and the inverse of the negative Hessian of the posterior evaluated at the mode; see the justification given in Kass and Steffey (1989). Details of our approximations can be found in the Appendix, where it is seen that the posterior distribution, $[\boldsymbol{\eta}, \boldsymbol{\xi}_O | \mathbf{Z}_O, \boldsymbol{\theta}^{[l]}]$, is approximately a multivariate Gaussian density, with approximate mean and approximate variance given by

$$E\left(\begin{pmatrix}\boldsymbol{\eta}\\ \boldsymbol{\xi}_O\end{pmatrix}\Big|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]}\right) = \begin{pmatrix}\hat{\boldsymbol{\eta}}^{[l]}\\ \hat{\boldsymbol{\xi}}_O^{[l]}\end{pmatrix}, \tag{23}$$

and

$$\mathrm{var}\left(\begin{pmatrix}\boldsymbol{\eta}\\ \boldsymbol{\xi}_O\end{pmatrix}\Big|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]}\right) = \left\{\begin{pmatrix}-\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}\left(L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \boldsymbol{\eta}, \boldsymbol{\xi}_O)\right) & -\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\xi}_O^\top}\left(L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \boldsymbol{\eta}, \boldsymbol{\xi}_O)\right)\\ -\frac{\partial^2}{\partial\boldsymbol{\xi}_O\partial\boldsymbol{\eta}^\top}\left(L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \boldsymbol{\eta}, \boldsymbol{\xi}_O)\right) & -\frac{\partial^2}{\partial\boldsymbol{\xi}_O\partial\boldsymbol{\xi}_O^\top}\left(L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \boldsymbol{\eta}, \boldsymbol{\xi}_O)\right)\end{pmatrix}\Big|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^{[l]}, \boldsymbol{\xi}_O=\hat{\boldsymbol{\xi}}_O^{[l]}}\right\}^{-1}, \tag{24}$$

respectively. To obtain $\mathrm{var}(\boldsymbol{\eta}|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]})$ and $\mathrm{var}(\boldsymbol{\xi}_O|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]})$, we need to invert the matrix of partial derivatives shown just above. Let $\mathbf{A}$ denote an $r \times r$ matrix and $\mathbf{B}$ denote an $n \times n$ matrix. Further, let $\mathbf{U}$ be any $r \times n$ matrix and $\mathbf{V}$ be any $n \times r$ matrix. Then, a block-matrix-inversion formula (e.g., Duncan, 1944) is given by:

$$\begin{pmatrix}\mathbf{A} & \mathbf{U}\\ \mathbf{V} & \mathbf{B}\end{pmatrix}^{-1} = \begin{pmatrix}(\mathbf{A}-\mathbf{U}\mathbf{B}^{-1}\mathbf{V})^{-1} & -(\mathbf{A}-\mathbf{U}\mathbf{B}^{-1}\mathbf{V})^{-1}\mathbf{U}\mathbf{B}^{-1}\\ -(\mathbf{B}-\mathbf{V}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}\mathbf{A}^{-1} & (\mathbf{B}-\mathbf{V}\mathbf{A}^{-1}\mathbf{U})^{-1}\end{pmatrix}. \tag{25}$$

Now recall the Sherman-Morrison-Woodbury formula (e.g., Henderson and Searle, 1981):

$$(\mathbf{B}-\mathbf{V}\mathbf{A}^{-1}\mathbf{U})^{-1} = \mathbf{B}^{-1} + \mathbf{B}^{-1}\mathbf{V}(\mathbf{A}-\mathbf{U}\mathbf{B}^{-1}\mathbf{V})^{-1}\mathbf{U}\mathbf{B}^{-1}.$$

We use this formula in the block-matrix-inversion formula (25) to obtain the following equivalent block-matrix-inversion formula, which we use to obtain the inverse in (24):

$$\begin{pmatrix}\mathbf{A} & \mathbf{U}\\ \mathbf{V} & \mathbf{B}\end{pmatrix}^{-1} = \begin{pmatrix}(\mathbf{A}-\mathbf{U}\mathbf{B}^{-1}\mathbf{V})^{-1} & -(\mathbf{A}-\mathbf{U}\mathbf{B}^{-1}\mathbf{V})^{-1}\mathbf{U}\mathbf{B}^{-1}\\ -\mathbf{B}^{-1}\mathbf{V}(\mathbf{A}-\mathbf{U}\mathbf{B}^{-1}\mathbf{V})^{-1} & \mathbf{B}^{-1}+\mathbf{B}^{-1}\mathbf{V}(\mathbf{A}-\mathbf{U}\mathbf{B}^{-1}\mathbf{V})^{-1}\mathbf{U}\mathbf{B}^{-1}\end{pmatrix}, \tag{26}$$

where the lower off-diagonal block is obtained using the Sherman-Morrison-Woodbury formula as follows:

$$
\begin{aligned}
(\mathbf{B}-\mathbf{V}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}\mathbf{A}^{-1} &= \left\{\mathbf{B}^{-1}\mathbf{V}(\mathbf{A}-\mathbf{U}\mathbf{B}^{-1}\mathbf{V})^{-1}\mathbf{U}\mathbf{B}^{-1}+\mathbf{B}^{-1}\right\}\mathbf{V}\mathbf{A}^{-1} \\
&= \mathbf{B}^{-1}\mathbf{V}(\mathbf{A}-\mathbf{U}\mathbf{B}^{-1}\mathbf{V})^{-1}\left\{\mathbf{U}\mathbf{B}^{-1}\mathbf{V}\mathbf{A}^{-1}+(\mathbf{A}-\mathbf{U}\mathbf{B}^{-1}\mathbf{V})\mathbf{A}^{-1}\right\} \\
&= \mathbf{B}^{-1}\mathbf{V}(\mathbf{A}-\mathbf{U}\mathbf{B}^{-1}\mathbf{V})^{-1}. \qquad (27)
\end{aligned}
$$

Now, for generic variables $\mathbf{u}$ and $\mathbf{v}$, define

$$
J(\mathbf{u}_0,\mathbf{v}_0) = -\frac{\partial^2}{\partial\mathbf{u}\partial\mathbf{v}^\top}\left(L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O,\mathbf{u},\mathbf{v})\right)\Big|_{\mathbf{u}=\mathbf{u}_0,\mathbf{v}=\mathbf{v}_0}.
$$

We consider the different component matrices in the $(r+n)\times(r+n)$ matrix of partial derivatives given in (24). The matrix $J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})$ is an $n\times n$ diagonal matrix; its inversion is easy. The matrix $J(\hat{\boldsymbol{\eta}}^{[l]},\hat{\boldsymbol{\eta}}^{[l]})$ is of dimension $r\times r$, where $r\ll n$. The other two matrices, $J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\eta}}^{[l]})$ and $J(\hat{\boldsymbol{\eta}}^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})$, have dimension $n\times r$ and $r\times n$, respectively. We can then use formula (26) to invert the matrix in (24), which gives, approximately,

$$
\begin{aligned}
\mathrm{var}(\boldsymbol{\eta}|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}) &= \left(J(\hat{\boldsymbol{\eta}}^{[l]},\hat{\boldsymbol{\eta}}^{[l]})-J(\hat{\boldsymbol{\eta}}^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})^{-1}J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\eta}}^{[l]})\right)^{-1} \\
\mathrm{var}(\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}) &= J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})^{-1}+J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})^{-1}J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\eta}}^{[l]}) \\
&\quad\times\left(J(\hat{\boldsymbol{\eta}}^{[l]},\hat{\boldsymbol{\eta}}^{[l]})-J(\hat{\boldsymbol{\eta}}^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})^{-1}J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\eta}}^{[l]})\right)^{-1}J(\hat{\boldsymbol{\eta}}^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})^{-1} \\
\mathrm{cov}(\boldsymbol{\eta},\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}) &= -\left(J(\hat{\boldsymbol{\eta}}^{[l]},\hat{\boldsymbol{\eta}}^{[l]})-J(\hat{\boldsymbol{\eta}}^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})^{-1}J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\eta}}^{[l]})\right)^{-1} \\
&\quad\times J(\hat{\boldsymbol{\eta}}^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})^{-1}. \qquad (28)
\end{aligned}
$$

In the formulas given just above, all we need to invert is the $n\times n$ diagonal matrix, $J(\hat{\boldsymbol{\xi}}_O^{[l]},\hat{\boldsymbol{\xi}}_O^{[l]})$, and some fixed-rank $r\times r$ matrices. This makes the computations extremely efficient and allows us to obtain the expressions for $E(\boldsymbol{\eta}\boldsymbol{\eta}^\top|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})$ and $E(\boldsymbol{\xi}_O\boldsymbol{\xi}_O^\top|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})$ in (22) as follows:

$$\begin{aligned}
E(\boldsymbol{\eta}\boldsymbol{\eta}^\top|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}) &= \mathrm{var}(\boldsymbol{\eta}|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}) + E(\boldsymbol{\eta}|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})E(\boldsymbol{\eta}|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})^\top \\
E(\boldsymbol{\xi}_O\boldsymbol{\xi}_O^\top|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}) &= \mathrm{var}(\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}) + E(\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})E(\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})^\top, \qquad (29)
\end{aligned}$$

where the terms on the right-hand side of (29) are evaluated approximately using (23) and (28).

The remaining terms in (22), for which we need an approximation, are

$$E\left(h_k(C(\mathbf{s}) + \mathbf{X}(\mathbf{s})^\top\boldsymbol{\beta} + \mathbf{S}(\mathbf{s})^\top\boldsymbol{\eta} + \xi(\mathbf{s}))|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right);\ \mathbf{s}\in\{\mathbf{s}_1,\ldots\mathbf{s}_n\},\ k=1,2.$$

For the particular case of count data and the canonical link considered in Sengupta and Cressie (2013), analytical expressions were obtained based on the Gaussian approximation for $[\boldsymbol{\eta},\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}]$ discussed above. In the general case considered here, a second-order Taylor-series expansion is needed to evaluate the required expectations. From the Appendix, we see that, approximately,

$$\begin{aligned}
E\left(h_k(C(\mathbf{s}_i) + \mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta} + S(\mathbf{s}_i)^\top\boldsymbol{\eta} + \xi(\mathbf{s}_i))|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right) &= h_k(C(\mathbf{s}_i) + \mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta} + S(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]} + \hat{\xi}^{[l]}(\mathbf{s}_i)) \\
&+ \frac{1}{2}h_k''(C(\mathbf{s}_i) + \mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta} + S(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]} + \hat{\xi}^{[l]}(\mathbf{s}_i)) \times \Big(\mathbf{S}(\mathbf{s}_i)^\top\mathrm{var}(\boldsymbol{\eta}|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})\mathbf{S}(\mathbf{s}_i) \\
&+ 2\mathbf{S}(\mathbf{s}_i)^\top\mathrm{cov}(\boldsymbol{\eta},\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})\mathbf{e}(\mathbf{s}_i) + \mathbf{e}(\mathbf{s}_i)^\top\mathrm{var}(\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})\mathbf{e}(\mathbf{s}_i)\Big), \qquad (30)
\end{aligned}$$

where $k=1,2$, and $\mathbf{e}(\mathbf{s}_i)$ is a vector of length n whose i-th element is 1 and all other entries are 0, for $i=1,\ldots,n$.

4.2 The M-step

Following the E-step, we perform the M-step, which involves maximizing (22) with respect to each of the parameters in $\boldsymbol{\theta}$. The maximization with respect to $\mathbf{K}$ and σ_ξ^2 is obtained by differentiating (22) with respect to $\mathbf{K}$ and σ_ξ^2, equating to zero, and solving the resulting equations. The solutions at the $(l+1)$-th iteration are:

$$\sigma_{\xi}^{2[l+1]} = \frac{1}{n}\text{trace}\left(\left(E(\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})E(\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})^\top + \text{var}\left(\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right)\right)\mathbf{V}_{\xi;O}^{-1}\right)$$

$$\mathbf{K}^{[l+1]} = E(\boldsymbol{\eta}|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})E(\boldsymbol{\eta}|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]})^\top + \text{var}\left(\boldsymbol{\eta}|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right). \tag{31}$$

However, the maximization of (22) with respect to $\boldsymbol{\beta}$ is not available in closed form; we use a Newton-Raphson update at each M-step as follows:

$$\boldsymbol{\beta}^{[l+1]} = \boldsymbol{\beta}^{[l]} - \left[\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{R}(\boldsymbol{\theta})\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[l]}}\mathbf{R}(\boldsymbol{\theta}^{[l]}). \tag{32}$$

In (32), $\mathbf{R}(\boldsymbol{\theta})$ denotes the score function obtained by taking the partial derivative of $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{[l]})$, given by (22), with respect to $\boldsymbol{\beta}$, and $\mathbf{R}(\boldsymbol{\theta}^{[l]})$ is obtained by evaluating $\mathbf{R}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^{[l]}$. The score function and the derivative required in (32) are evaluated in the Appendix.

4.3 Starting Values for the EM Algorithm

In order to implement the EM algorithm, we need to specify some starting values for the parameters. Although in the simulation study described in Section 5, we use the true parameter values as our starting values, for real data applications we do not have that luxury. In this section, we give a recommendation for initializing the EM algorithm. We shall use this method to obtain the starting values for the EM algorithm when analyzing the large remote sensing dataset in Section 6.

One may proceed by using the classical fixed-effects GLM estimate, $\hat{\boldsymbol{\beta}}_{GLM}$, as the starting value for $\boldsymbol{\beta}$; here, $\hat{\boldsymbol{\beta}}_{GLM}$ is obtained using the iterated reweighted least squares algorithm (see McCulloch et al., 2001, Chapter 5).

Recall that the spatial trend is

$$t(\mathbf{s}_i) = C(\mathbf{s}_i) + \mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta};$$

consider the detrended process,

$$U(\mathbf{s}_i) \equiv Y(\mathbf{s}_i) - t(\mathbf{s}_i), \tag{33}$$

which has mean zero and

$$\text{var}(U(\mathbf{s}_i)) = \mathbf{S}(\mathbf{s}_i)^\top \mathbf{K}\mathbf{S}(\mathbf{s}_i) + \sigma^2_\xi \text{v}_\xi(\mathbf{s}_i). \tag{34}$$

Writing $\mathbf{U}_O \equiv (U(\mathbf{s}_1), \ldots, U(\mathbf{s}_n))^\top$, we obtain:

$$\text{cov}(\mathbf{U}_O) \equiv \boldsymbol{\Sigma}_{U;O} = \mathbf{S}_O\mathbf{K}\mathbf{S}_O^\top + \sigma^2_\xi \mathbf{V}_{\xi;O}, \tag{35}$$

where recall that $\mathbf{V}_{\xi;O}$ is a known diagonal matrix.

To obtain method-of-moments estimates of $\mathbf{K}$ and σ^2_ξ that can be used as starting values, we replace $Y(\mathbf{s}_i)$ with $g(Z(\mathbf{s}_i) + c)$, where c is some user-specified constant that is added to the data to ensure that the transformation is defined everywhere within the range of the data and recall that $g(\cdot)$ is the link function. For example, for Poisson data and the canonical log link, $\log(Z(\mathbf{s}_i) + 0.5)$ avoids a singularity when $Z(\mathbf{s}_i) = 0$.

Consequently, an approximation for $U(\cdot)$ is obtained as:

$$\hat{U}(\mathbf{s}_i) \equiv g(Z(\mathbf{s}_i) + c) - C(\mathbf{s}_i) - \mathbf{X}(\mathbf{s}_i)^\top \hat{\boldsymbol{\beta}}_{GLM}, \; i = 1, \ldots, n. \tag{36}$$

Define $s^2_U \equiv \frac{1}{n}\sum_{i=1}^n \hat{U}(\mathbf{s}_i)^2$, and choose

$$\hat{\boldsymbol{\Sigma}}_{U;O} = s^2_U \mathbf{I}_n, \tag{37}$$

simply to capture the total variation through the trace operator. We apportion approximately 90% of this to the smooth small-scale variation and 10% to the fine-scale variation (e.g., Katzfuss and

Cressie, 2011). That is, we select our starting values for $\mathbf{K}$ and σ^2_ξ to satisfy

$$\begin{aligned} &\mathbf{S}_O\mathbf{K}^{[0]}\mathbf{S}_O^\top \approx 0.9 \times \hat{\mathbf{\Sigma}}_{U;O} \\ &\sigma_\xi^{2[0]} = 0.1 \times \text{trace}(\hat{\mathbf{\Sigma}}_{U;O})/\text{trace}(\mathbf{V}_{\xi;O}), \end{aligned} \tag{38}$$

as follows. Using (38), and the *Q*-*R* decomposition, $\mathbf{S}_O = \mathbf{Q}_S\mathbf{R}_S$, we obtain the starting value for $\mathbf{K}$ as

$$\mathbf{K}^{[0]} = \mathbf{R}_S^{-1}\mathbf{Q}_S^\top \left(0.9 \times \hat{\mathbf{\Sigma}}_{U;O}\right)\mathbf{Q}_S(\mathbf{R}_S^\top)^{-1}. \tag{39}$$

Note that this approximate 90-10 apportionment of the total variability could be done differently, depending on the data's smooth-scale variation relative to their fine-scale variation.

4.4 Properties of the Resulting EM Algorithm

Suppose that the algorithm is initialized with parameter values $\boldsymbol{\theta}^{[0]} \in \Theta$, where Θ is the parameter space. Then it can be seen from (31) that $\boldsymbol{\theta}^{[l]} \in \Theta$, $l = 1, 2, \ldots$, which is a desirable property. For example, this means that if the starting value for $\mathbf{K}$ is a covariance matrix, then all future EM updates will also be symmetric and at least non-negative definite. Likewise, if we choose $\sigma_\xi^{2[0]} > 0$, then it is guaranteed that the EM estimate satisfies $\hat{\sigma}^2_{\xi;EM} \geq 0$.

The most appealing feature of the resulting EM algorithm is computational. The E-step requires one optimization to obtain the posterior mode. Then the SRE-model assumption and the Sherman-Morrison-Woodbury formula make the LA computations extremely efficient. The computational complexity of the EM algorithm is linear in the sample size *n* (see Section 5.4). This is a highly desirable property when dealing with big data. In Section 5, the computational performance of this algorithm and the variability of the estimates are assessed through simulation.

5 A Simulation Study

In this section, we investigate statistical properties of our EHM approach using a simulation experiment, where we simulate *Poisson* data over a regular spatial domain using the hierarchical model set-up as described in Section 2. Further, we demonstrate the computational gain that is achieved by using an EHM approach as opposed to a BHM approach. The R-functions for the EM algorithm and the MCMC algorithm relevant to our EHM are available on request.

5.1 Simulation Set-Up

We generated count data from a Poisson distribution whose mean was obtained by exponentiating an underlying spatial Gaussian process $Y(\cdot)$. We considered a regular spatial domain, $D = \{\mathbf{s}_1, \dots \mathbf{s}_N\}$, consisting of $N = 300 \times 300 = 90,000$ points on $\{-149.5, \dots, -0.5, 0.5, \dots, 149.5\}^2$. In this simulation, the hidden process $Y(\cdot)$ given by (8), (9), and (11) was made up of three additive components:

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})^\top \boldsymbol{\beta} + \mathbf{S}(\mathbf{s})^\top \boldsymbol{\eta} + \xi(\mathbf{s});\ \mathbf{s} \in D, \tag{40}$$

where the fine-scale heterogeneity term $\mathrm{v}_\xi(\cdot) = 1$, and the offset term $C(\cdot) = 0$. The large-scale variation, or trend, was assumed to be,

$$\mathbf{X}(\mathbf{s})^\top \boldsymbol{\beta} = \beta_0 + \beta_1 \times s_2, \tag{41}$$

where $\mathbf{s} = (s_1, s_2)^\top$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$.

Recall that the random-effects vector $\boldsymbol{\eta} \sim \mathrm{Gau}(\mathbf{0}, \mathbf{K})$, and here $\xi(\cdot)$ is a process of independent and identically distributed (i.i.d.) $\mathrm{Gau}(0, \sigma_\xi^2)$ random variables, independent of $\boldsymbol{\eta}$. For the vector of basis functions, $\mathbf{S}(\cdot)$, we used the bisquare functions. The centers of the bisquare functions were selected using two scales of resolution and were regularly spaced within a resolution. The number of basis functions used at the two resolutions were, respectively, 4 and 25. Consequently, $r = 4 + 25 = 29$.

To specify the SRE model's covariance matrix $\mathbf{K}$, we started with an exponential covariance function given by

$$C(\mathbf{u}, \mathbf{v}) = c_0 \exp\left(-\frac{||\mathbf{u} - \mathbf{v}||}{a_0}\right), \tag{42}$$

where c_0 is the sill and a_0 is the scale parameter. Here we specified $c_0 = 1$ (without loss of generality) and $a_0 = 100$ (to capture moderate-to-strong spatial dependence). Let $\boldsymbol{\nu} \equiv (\nu(\mathbf{s}_1), \ldots \nu(\mathbf{s}_N))^\top$ be a mean-zero spatial Gaussian process defined over D, whose covariance matrix is obtained from the exponential covariance model (42); that is, $\boldsymbol{\nu} \sim \text{Gau}(\mathbf{0}, \boldsymbol{\Sigma}_\nu)$. We calibrated $\mathbf{K}$ and σ_ξ^2 using the procedure given in Kang and Cressie (2011). For just the calibration, we considered only 9,000 regularly spaced locations (sampling every tenth location from the list of all 90,000 locations) that covered the entire spatial domain, rather than using all 90,000 locations.

First we calculated $\mathbf{K}^0$ such that $||\mathbf{S}\mathbf{K}^0\mathbf{S}^\top - \boldsymbol{\Sigma}_\nu||$ was minimized, where $||\cdot||$ is the Frobenius norm (e.g., Cressie and Johannesson, 2008). Finally, to control the variability of $\mathbf{Y}$, we chose $\mathbf{K} = k\mathbf{K}^0$, where k was chosen to preserve the total variation. That is,

$$\text{trace}(\boldsymbol{\Sigma}_\nu)/N = 1 = \text{trace}(k\mathbf{S}\mathbf{K}^0\mathbf{S}^\top + \sigma_\xi^2\mathbf{I}_N)/N. \tag{43}$$

For selecting the large-scale-variation parameter $\boldsymbol{\beta}$, we defined the variation of the "signal," V_s, as:

$$V_s \equiv \frac{1}{N}\text{trace}\left(\mathbf{S}\mathbf{K}\mathbf{S}^\top + \sigma_\xi^2\mathbf{I}_N\right) + \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta} - \underset{\mathbf{s}_i \in D}{\text{ave}}(\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta})\right)^2.$$

The parameter $\boldsymbol{\beta}$ was selected such that V_s was approximately 2 (see Aldworth and Cressie, 1999, Section 3.2.4). Note that β_0 is a free parameter that does not impact V_s. We fixed $\beta_0 = 2$. Specifying $\beta_1 = 0.0125$ gives $V_s = 2.17$. Consequently, in our simulation study, $\boldsymbol{\beta} = (2, 0.0125)^\top$. Additionally, we specified the *fine-scale-variation proportion (FVP)*,

$$FVP \equiv \frac{\text{trace}\left(\sigma_\xi^2\mathbf{I}_N\right)}{\text{trace}\left(\mathbf{S}\mathbf{K}\mathbf{S}^\top + \sigma_\xi^2\mathbf{I}_N\right)}, \tag{44}$$

which from (43) is equal to σ_ξ^2. In our simulation, FVP was held at 5%; hence, $\sigma_\xi^2 = 0.05$. Using (43), we obtained $k = 1.22$.

We simulated $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ from the Gaussian process defined above and then, using (40), we obtained $\mathbf{Y}$ over the entire domain D. Next, we used the inverse of the log link function,

$$\mu_{Z|Y}(\cdot) = \exp\left(Y(\cdot)\right), \tag{45}$$

to simulate a realization of the conditionally (conditional on $Y(\cdot)$) independent Poisson random variables, $\mathbf{Z}_O$, for only n locations ($n \leq N$); the n locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ were randomly sampled without replacement from the $N = 90,000$ possible locations.

We will use this set-up to investigate the performance of the EM-based parameter estimates (Section 5.2), to compare the predictive performance of our EHM approach to that of an independent hierarchical GLM (Section 5.3), to compare the computational efficiency of our EHM approach to that of a competing Bayesian hierarchical modeling (BHM) approach (Section 5.4), and finally to do a sensitivity study of the EHM and the BHM approaches (Section 5.5). In Sections 5.2, 5.3, and 5.5, we hold n fixed at 20,000. In Section 5.4, we vary n and tabulate the computational efficiency as a function of n. We use the true parameter values as starting values for the EM algorithm and for specifying hyperparameters for the BHM approach.

5.2 Assessment of the EM Estimates

In this section, we assess the performance of the EM estimates. Holding n fixed at 20,000, we simulated 1600 vectors $\mathbf{Z}_O^{[1]}, \ldots, \mathbf{Z}_O^{[1600]}$ as specified in Section 5.1. For each of the simulated datasets, $\mathbf{Z}_O^{[l]}$, where $l = 1, \ldots, 1600$, we used the EM algorithm described in Section 4 to estimate the unknown parameters.

We calculated the average and the empirical root mean squared error (RMSE) for the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ and σ_ξ^2; the results are summarized in Table 1, and they show very good agreement with the true values.

—— Table 1 approximately here ——

Now we consider the EM estimate of $\mathbf{K}$. The elementwise mean of the EM estimates, $\left\{\hat{\mathbf{K}}_{EM}^{[l]} : l = 1, \ldots, 1600\right\}$, was computed as:

$$\mathrm{ave}(\hat{\mathbf{K}}_{EM}) \equiv \frac{1}{1600} \sum_{l=1}^{1600} \hat{\mathbf{K}}_{EM}^{[l]}. \tag{46}$$

Figure 1 shows an image plot of the matrix $\mathbf{H} \equiv \left\{\mathrm{ave}(\hat{\mathbf{K}}_{EM})\mathbf{K}_T^{-1}\right\}$, where $\mathbf{K}_T$ is the true covariance matrix for $\boldsymbol{\eta}$. We compare the matrix $\mathbf{H}$ to the identity matrix, which gives a visual representation of how close the mean of the EM estimate of $\mathbf{K}$ is to the true value $\mathbf{K}_T$.

—— Figure 1 approximately here ——

We also computed trace$(\hat{\mathbf{K}}_{EM}^{[l]}\mathbf{K}_T^{-1})$, for $l = 1,\ldots,1600$. Now, had we observed $\boldsymbol{\eta}^{[l]}$, the ML estimate of $\mathbf{K}$ would be given by:

$$\hat{\mathbf{K}}_{ML;\eta}^{[l]} = \boldsymbol{\eta}^{[l]}\boldsymbol{\eta}^{[l]\top}, \tag{47}$$

for which

$$\text{trace}(\hat{\mathbf{K}}_{ML;\eta}\mathbf{K}_T^{-1}) = \text{trace}(\boldsymbol{\eta}^{[l]}\boldsymbol{\eta}^{[l]\top}\mathbf{K}_T^{-1}) = \boldsymbol{\eta}^{[l]\top}\mathbf{K}_T^{-1}\boldsymbol{\eta}^{[l]} \sim \chi_r^2. \tag{48}$$

Consequently, we might expect the distribution of trace$(\hat{\mathbf{K}}_{EM}^{[l]}\mathbf{K}_T^{-1})$ to look similar to a χ_r^2 distribution. Recall that $r = 29$ in our case. Figure 2 shows a histogram of $\left\{\text{trace}(\hat{\mathbf{K}}_{EM}^{[l]}\mathbf{K}_T^{-1}) : l = 1,\ldots,1600\right\}$, upon which a χ_{29}^2 density is superimposed. The sample mean and the sample variance of $\left\{\text{trace}(\hat{\mathbf{K}}_{EM}^{[l]}\mathbf{K}_T^{-1})\right\}$ are 29.4194 and 59.821, respectively, which we compare to $\text{E}(\chi_{29}^2) = 29$ and $\text{var}(\chi_{29}^2) = 58$.

—— Figure 2 approximately here ——

Overall, the EM algorithm seems to perform well, despite the approximations involved in the E-step of the EM algorithm. Next, we shall investigate the predictive properties of our EHM approach.

5.3 Predictive Properties

In this section, we assess the predictive properties for the EHM approach described in Sections 2–4. Here, we again held n fixed at $20,000$, and we generated 100 datasets $\mathbf{Z}_O^{[1]},\ldots,\mathbf{Z}_O^{[100]}$. For each of the simulated datasets $\left\{\mathbf{Z}_O^{[l]} : l = 1,\ldots,100\right\}$, we implemented the EM algorithm to obtain $\hat{\boldsymbol{\theta}}_{EM}^{[l]} \equiv (\hat{\boldsymbol{\beta}}_{EM}^{[l]}, \hat{\mathbf{K}}_{EM}^{[l]}, \hat{\sigma}_{\xi;EM}^{2[l]})$. Then, using the MCMC algorithm described in Section 3, we obtained samples from the empirical predictive distribution, $[\boldsymbol{\eta}, \boldsymbol{\xi}_O | \mathbf{Z}_O^{[l]}, \hat{\boldsymbol{\theta}}_{EM}^{[l]}]$: For each of the 100 simulated datasets, we generated 25,000 MCMC samples, after discarding a burn-in sample of size 2,000. Recall that our EHM approach yields the predictor of $Y(\cdot)$ based on $\mathbf{Z}_O^{[l]}$, as the mean of the resulting

MCMC samples from the empirical predictive distribution $[Y(\cdot)|\mathbf{Z}_O^{[l]}, \hat{\boldsymbol{\theta}}_{EM}^{[l]}]$. Here we compare this to one derived from a spatially independent GLM, namely

$$Y(\cdot) = \mathbf{X}(\cdot)^\top \boldsymbol{\beta} + \xi(\cdot), \tag{49}$$

where $\xi(\cdot) \sim$ i.i.d. Gau$(0, \sigma_\xi^2)$. To estimate the parameters of the resulting EHM, we used the EM algorithm described in Section 4 with $\boldsymbol{\eta} = \mathbf{0}$, that is, with no spatial random-effects component. The MCMC algorithm from which the empirical predictive distribution is obtained is, likewise, a special case of that given in Section 3, with $\boldsymbol{\eta} = \mathbf{0}$.

In what follows, we denote the 20,000 locations with data as D_O and the complementary set of 70,000 locations without data as D_U. Recall that D_O was obtained by random sampling from D without replacement; for the 100 datasets, the set of locations D_O (and hence D_U) are held fixed.

Using obvious notation where "S" denotes "spatial" and "I" denotes "independent," define $\hat{Y}_{SEHM}^{[l]}(\cdot)$ and $\hat{Y}_{IEHM}^{[l]}(\cdot)$ to be the means of their respective predictive distributions, $[Y(\cdot)|\mathbf{Z}_O^{[l]}, \hat{\boldsymbol{\theta}}_{SEM}^{[l]}]$ and $[Y(\cdot)|\mathbf{Z}_O^{[l]}, \hat{\boldsymbol{\theta}}_{IEM}^{[l]}]$. Importantly, $\mathbf{Z}_O^{[1]}, \ldots, \mathbf{Z}_O^{[100]}$ were simulated according to the set-up given in Section 5.1.

Consider the ratio of the mean squared prediction errors,

$$e(\mathbf{s}) \equiv \frac{\frac{1}{100}\sum_{l=1}^{100}(\hat{Y}_{SEHM}^{[l]}(\mathbf{s}) - Y^{[l]}(\mathbf{s}))^2}{\frac{1}{100}\sum_{l=1}^{100}(\hat{Y}_{IEHM}^{[l]}(\mathbf{s}) - Y^{[l]}(\mathbf{s}))^2};\ \mathbf{s} \in D, \tag{50}$$

where $Y^{[l]}(\cdot)$ is the true process (Section 5.1). From (50), we made kernel-density plots showing the distribution of $e(\cdot)$ for locations in D_O and for those in D_U, separately. These plots are shown in the left panel of Figure 3, from which we see that SEHM has higher relative efficiency for locations in D_U than for those in D_O. Clearly, for locations without data (i.e., D_U), SEHM borrows strength efficiently from nearby observations, and hence it performs much better than IEHM in terms of smaller mean squared prediction error.

—— Figure 3 approximately here ——

Now we shall investigate the performance of our EHM approach for the locations with and without data. We made kernel-density plots that compare the distribution of mean squared prediction errors,

$$\frac{1}{100}\sum_{l=1}^{100}(\hat{Y}^{[l]}_{SEHM}(\mathbf{s})-Y^{[l]}(\mathbf{s}))^2,$$

for locations **s** in D_O to those in D_U (see Figure 3, right panel). Generally, the right panel of Figure 3 shows that mean squared prediction errors are smaller in D_O than in D_U. Since a datum $Z(\mathbf{s})$ at location **s** is very informative about the hidden value $Y(\mathbf{s})$ at **s**, this is to be expected.

5.4 Computational Time: EHM versus BHM

In this section, we illustrate the computational gain achieved by using an EHM approach as opposed to using a comparable BHM approach. In what follows, whenever we say EHM (BHM), we mean a spatial EHM (spatial BHM).

Recall that part of our EHM approach involves estimating the unknown parameters using an EM algorithm, followed by an MCMC algorithm that generates samples from the empirical predictive distribution, $[\boldsymbol{\eta},\boldsymbol{\xi}_O|\mathbf{Z}_O,\hat{\boldsymbol{\theta}}_{EM}]$, where $\hat{\boldsymbol{\theta}}_{EM}\equiv(\hat{\boldsymbol{\beta}}_{EM},\hat{\mathbf{K}}_{EM},\hat{\sigma}^2_{\xi;EM})$. In a BHM approach, priors are put on $\boldsymbol{\beta}$, **K**, and σ^2_{ξ}, and an MCMC algorithm is used to generate samples from the posterior distribution, $[\boldsymbol{\eta},\boldsymbol{\xi},\boldsymbol{\theta}|\mathbf{Z}_O]$. Priors are assigned following Kang and Cressie (2011), the details of which are given in the Appendix.

Generally, the MCMC algorithm mixes more slowly for the BHM than for the EHM. Hence, we need to calibrate the MCMC sample sizes properly before we can compare the computational times. Suppose the number of MCMC samples from the empirical predictive distribution, $[\boldsymbol{\eta},\boldsymbol{\xi}_O|\mathbf{Z}_O,\hat{\boldsymbol{\theta}}_{EM}]$, is L_{EHM}, and suppose that L_{BHM} is the number of MCMC samples obtained from the posterior distribution, $[\boldsymbol{\eta},\boldsymbol{\xi},\boldsymbol{\theta}|\mathbf{Z}_O]$.

To calibrate the MCMC sample sizes, there are different diagnostic measures that could be used (e.g., Robert and Casella, 2004, Chapter 12). In this article, we shall use the diagnostics proposed by Gelman and Rubin (1992) and Brooks and Gelman (1998). The Gelman-Rubin statistic, or potential scale reduction factor (PSRF), is based on the idea of generating several MCMC chains, each of length L, and then comparing the variability based on these individual chains to that based on the combined chain. If PSRF is close to 1, we can conclude that each set of L simulated values is close to the target distribution; if PSRF is large, L may be too small. Brooks and Gelman (1998) proposed the multivariate potential scale reduction factor (MPSRF), which is a multivariate extension of the PSRF, that can be used for assessing convergence of several parameters simultaneously.

For fixed data size n, we generated five MCMC chains, each of length L. Then we found the values of L_{EHM} and L_{BHM} that had comparable MSPRFs close to 1. We started with $n = 5,000$ and found that for the elements of $\boldsymbol{\xi}$, mixing was achieved quickly for both EHM and BHM. However, mixing for $\boldsymbol{\eta}$ is comparatively slow for EHM and even slower for BHM, so we calibrated the MCMC sample sizes based on the convergence diagnostics for $\boldsymbol{\eta}$. Figure 4 shows plots of the MPSRF and the maximum of elementwise PSRFs as functions of L. From Figure 4, we selected $L_{EHM} = 15,000$, and $L_{BHM} = 40,000$, which resulted in MPSRFs of 1.08 for EHM and 1.07 for BHM.

—— Figure 4 approximately here ——

Next we investigated how the MPSRF and the PSRFs changed as n changed. By holding $L_{EHM} = 15,000$ and $L_{BHM} = 40,000$, and varying n, Table 2 shows that the Gelman-Rubin and Gelman-Brooks statistics are robust to change in the sample size, n. Consequently, we compare the computational times for EHM and BHM, for all n, using $L_{EHM} = 15,000$ and $L_{BHM} = 40,000$.

—— Table 2 approximately here ——

The simulation experiment was performed on a dual quad core 2.8 GHz 2x Xeon X5560 processor, with 96 Gbytes of memory. The computational times for the EHM and BHM are given in Table 3. From Table 3 we see that EHM is on the order of 6-10 times faster than BHM. Nevertheless, in both cases, the computational time increases approximately linearly in n, which is due to the dimension reduction afforded by the SRE model given by (11).

—— Table 3 approximately here ——

5.5 Sensitivity Study Comparing EHM to BHM

In this section, we describe a sensitivity study to demonstrate the precision and accuracy of the EHM predictions, when compared to BHM predictions (e.g., Kang et al., 2009).

Using the methods described in Section 5.1, we simulated $\mathbf{Z}_O$, with $n = 20,000$. From those simulated data, we obtained samples from the empirical predictive distribution $[Y(\cdot)|\mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM}]$, which is our EHM approach, and from the posterior distribution $[Y(\cdot)|\mathbf{Z}_O]$, which is the BHM approach. First, we did a visual assessment of the predictions, $\hat{Y}_{SEHM}(\cdot) \equiv E(Y(\cdot)|\mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM})$ and $\hat{Y}_{SBHM}(\cdot) \equiv E(Y(\cdot)|\mathbf{Z}_O)$, which are shown in Figure 5, along with the data, $\{Z(\mathbf{s}_i), i = 1, \ldots, n = 20,000\}$, and the true underlying process, $Y(\cdot)$. Figure 5 gives the visual impression that there is no difference in the predictions obtained using EHM and BHM, which is confirmed with a kernel-density plot showing the distribution of the difference, $\hat{Y}_{SEHM}(\cdot) - \hat{Y}_{SBHM}(\cdot)$; see Figure 6 (left panel).

—— Figure 5 approximately here ——

—— Figure 6 approximately here ——

Next we computed the ratio,

$$r(\cdot) = \frac{(\mathrm{var}(Y(\cdot)|\mathbf{Z}_O))^{1/2}}{(\mathrm{var}(Y(\cdot)|\mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM})^{1/2}}. \tag{51}$$

The distribution of the ratio of the standard deviations is shown on the right panel of Figure 6, separately for locations in D_O (where data are observed) and D_U (where data are not observed). From the right panel of Figure 6, we see that the ratio is mostly larger than 1; it is always larger than 1 in D_U, and it is larger than 1 for 87.5% of locations in D_O. Thus, our EHM approach tends to yield credible intervals for $Y(\cdot)$ that are narrower than those obtained from a BHM approach. From this experiment, we see that for $\mathbf{s} \in D_O$, EHM-based credible intervals tend to be narrower by a factor of 0.8, while for $\mathbf{s} \in D_U$, the factor is 0.75. These results are consistent with other spatial studies (e.g., Kang et al., 2009).

6 Analysis of Aerosol Optical Depth from the MISR Instrument

In this section, we use the methodology presented in the previous sections to analyze a large, spatial, remotely sensed dataset on aerosol optical depth (AOD) retrieved by the Multi-angle Imaging SpectroRadiometer (MISR) instrument on NASA's Terra satellite. An analysis of this dataset was done by Shi and Cressie (2007); they used a log transformation of the data and then analyzed log(AOD) using a Gaussian model, however they did not obtain spatial predictions back on the original AOD scale. The key feature of our current analysis is to model AOD directly, using a hierarchical spatial statistical model with a Gamma data model. The methodology we have developed in the previous sections allows us to obtain optimal spatial predictions, posterior standard errors, and 95% prediction intervals on the original AOD scale.

6.1 Background to the Dataset

The Terra satellite was launched by NASA on December 18, 1999, as part of the Earth Observing System (EOS). The MISR instrument is one of the key instruments on board that collects global aerosol information, and it covers the entire globe in 16 days. Level-2 AOD data are collected at a 17.6 km $\times$ 17.6 km spatial resolution; they can then be converted to level-3 AOD data at a lower spatial resolution (of $0.5^\circ \times 0.5^\circ$) by averaging all the level-2 observations that fall within the level-3 pixels. (Here, and in what follows, when we say level-3 pixel, we mean a pixel at the spatial resolution of $0.5^\circ \times 0.5^\circ$.) Due to orbit geometry, clouds, or non-retrievals, data can be missing in many regions. We use our model to predict the true AOD at level-3 pixels, both where there are data and where there are no data.

We analyze here a spatial dataset of lattice data consisting of level-3 AOD values observed between August 2-9, 2001, within a study region D bounded by longitudes -125° and $+3^\circ$ and latitudes -20° and $+44^\circ$. This is the same dataset that was analyzed in Shi and Cressie (2007), and was part of a spatio-temporal dataset in Kang et al. (2010), although exclusively on the log(AOD) scale. The region covers North and South America, the western part of the Sahara desert in Africa, the Iberian Peninsula in Europe, and parts of the Atlantic and Pacific Oceans (see Kang et al., 2010, for a map of the study region). There are $N \equiv 128 \times 256 = 32,768$ level-3 pixels in D. The $n = 21,759$ data in D_O are shown in the top-left panel of Figure 9, where white pixels define the no-data locations (i.e., D_U); a histogram for the data is shown on the top-right panel of Figure 9.

6.2 Hierarchical Spatial Statistical Modeling of AOD

In this section, we do some initial data analysis of the AOD dataset by fitting a weighted generalized linear model that does not contain spatial dependence (McCullagh and Nelder, 1989), followed by a full spatial analysis of the dataset. Recall from Section 6.1 that $Z(\mathbf{s}_i)$ is the average AOD obtained by averaging all the level-2 observations that fall within the level-3 pixel located at $\mathbf{s}_i$. Let $m(\mathbf{s}_i)$ denote the number of level-2 observations that are averaged to obtain $Z(\mathbf{s}_i)$, for $i = 1, \ldots, n$. We denote the level-2 observations within the level-3 pixel located at $\mathbf{s}_i$ as $Z_j(\mathbf{s}_i)$, $j = 1, \ldots, m(\mathbf{s}_i)$, so that $Z(\mathbf{s}_i) \equiv \sum_{j=1}^{m(\mathbf{s}_i)} Z_j(\mathbf{s}_i)/m(\mathbf{s}_i)$.

Conditional on an underlying spatial process $Y(\cdot)$, we assume independent Gamma distributions for the level-2 observations. That is, conditional on $Y(\cdot)$, $Z_j(\mathbf{s})$ and $Z_k(\mathbf{u})$ are independent, except when $\mathbf{s} = \mathbf{u}$ and $j = k$. We further assume local homogeneity within a level-3 pixel; that is,

$$Z_j(\mathbf{s}_i)|Y(\mathbf{s}_i) \sim \text{ i.i.d Gamma}(\nu, \mu_{Z|Y}(\mathbf{s}_i)/\nu);\ j = 1, \ldots m(\mathbf{s}_i), \tag{52}$$

where $\mu_{Z|Y}(\mathbf{s}_i) \equiv E(Z(\mathbf{s}_i|Y(\cdot)) = E(Z(\mathbf{s}_i)|Y(\mathbf{s}_i))$ is the mean of the conditional distribution $[Z_j(\mathbf{s}_i)|Y(\mathbf{s}_i)]$; $\nu > 0$ is the shape parameter of the Gamma distribution; and, consequently, $\mu_{Z|Y}(\mathbf{s}_i)/\nu$ (> 0) is its scale parameter for the level-3 pixel at $\mathbf{s}_i$. That is, the density function for $Z_j(\mathbf{s}_i)|Y(\mathbf{s}_i)$, under this parameterization, is

$$f_{Z|Y}(z_j(\mathbf{s}_i)|Y(\mathbf{s}_i)) = \frac{(z_j(\mathbf{s}_i)\nu)^{\nu} \exp(-z_j(\mathbf{s}_i)\nu/\mu_{Z|Y}(\mathbf{s}_i))}{z_j(\mathbf{s}_i)\Gamma(\nu)\mu_{Z|Y}(\mathbf{s}_i)^{\nu}};\ z_j(\mathbf{s}_i) \geq 0. \tag{53}$$

From (52), and (53), we obtain the conditional distribution of the level-3 datum at $\mathbf{s}_i$ as,

$$Z(\mathbf{s}_i)|Y(\mathbf{s}_i) \sim \text{Gamma}(m(\mathbf{s}_i)\nu, \mu_{Z|Y}(\mathbf{s}_i)/(m(\mathbf{s}_i)\nu));\ i = 1, \ldots, n, \tag{54}$$

where the distributions are assumed independent. Thus, we see that the between-pixel heterogeneity shows up in the scale and the shape parameters, although $E(Z(\mathbf{s}_i)|Y(\mathbf{s}_i))$ is $\mu_{Z|Y}(\mathbf{s}_i)$ and does not depend on $m(\mathbf{s}_i)$. This yields the loglikelihood,

$$\begin{aligned} L(\boldsymbol{\beta}, \nu) = \sum_{i=1}^{n} \Bigg\{ &(m(\mathbf{s}_i)\nu - 1)\log(Z(\mathbf{s}_i)) + m(\mathbf{s}_i)\nu \log(m(\mathbf{s}_i)\nu) - \frac{Z(\mathbf{s}_i)m(\mathbf{s}_i)\nu}{\exp(X(\mathbf{s}_i)^\top \boldsymbol{\beta})} \\ &- \log\Gamma(m(\mathbf{s}_i)\nu) - m(\mathbf{s}_i)\nu(\mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta}) \Bigg\}. \end{aligned} \tag{55}$$

The canonical link for the Gamma distribution is the reciprocal link, namely, $\gamma(\mathbf{s}) = (\mu_{Z|Y}(\mathbf{s}))^{-1}$, which leads to constraints on the conditional mean that are not easy to model. Guided by previous analyses of AOD where log data were analyzed, we use a log link. That is,

$$\log(\mu_{Z|Y}(\mathbf{s}_i)) = \mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta};\ i = 1, \ldots, N, \tag{56}$$

where $\mathbf{X}(\mathbf{s}_i)$ is a p-dimensional vector of known covariates, and there is no offset term $C(\cdot)$ in this model. After some initial exploratory data analysis considering the covariates used in Kang et al. (2010), we selected the covariates in (56) to be the indicator functions for each of the Americas, Africa (the Sahara desert), the south-western tip of Europe (Iberian Peninsular), and oceans; and we also included latitude as a covariate.

From the weighted GLM (WGLM) given by (53) and (56), we obtained the ML estimate, $\hat{\boldsymbol{\beta}}_{WGLM}$, of $\boldsymbol{\beta}$, which does not depend on ν. Note that the estimate $\hat{\boldsymbol{\beta}}_{WGLM}$ is different than what one would obtain using a standard R or Matlab package, since they do not consider the different $\{m(\mathbf{s}_i) : i = 1, \ldots, n\}$ that appear in the loglikelihood given by (55). The maximum likelihood estimate of ν is obtained by maximizing $L(\hat{\boldsymbol{\beta}}_{WGLM}, \nu)$ with respect to ν and results in $\hat{\nu} = 0.3637$. These ML estimates are used in the hierarchical statistical analysis that follows.

As an aside, if we transform the data as, $\tilde{Z}(\mathbf{s}_i) \equiv m(\mathbf{s}_i)Z(\mathbf{s}_i)$; $i = 1, \ldots, n$, then the distribution of $\tilde{Z}(\mathbf{s}_i)$ is Gamma$(m(\mathbf{s}_i)\nu, \mu_{\tilde{Z}|Y}(\mathbf{s}_i))$, where $\mu_{\tilde{Z}|Y}(\mathbf{s}_i) \equiv m(\mathbf{s}_i)\mu_{Z|Y}(\mathbf{s}_i)$. Hence, the log link is:

$$\log(\mu_{\tilde{Z}|Y}(\mathbf{s}_i)) = \log(m(\mathbf{s}_i)\mu_{Z|Y}(\mathbf{s}_i)) = \log(m(\mathbf{s}_i)) + \mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta}, \tag{57}$$

where there is now an offset term $C(\mathbf{s}_i) = \log(m(\mathbf{s}_i))$. Since the information content of $\left\{\tilde{Z}(\mathbf{s}_i)\right\}$ and $\{Z(\mathbf{s}_i)\}$ are the same, the ML estimates of $\boldsymbol{\beta}$ and ν are unchanged.

Our spatial hierarchical statistical model consists of a data model and a process model; recall that unknown parameters are estimated. The data model is given by (54), where $\nu = 0.3637$, obtained above. We assume the log link,

$$Y(\cdot) = \log(\mu_{Z|Y}(\cdot)), \tag{58}$$

and the process model is:

$$Y(\mathbf{s}_i) = \mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta} + \mathbf{S}(\mathbf{s}_i)^\top \boldsymbol{\eta} + \xi(\mathbf{s}_i);\ i = 1, \ldots N, \tag{59}$$

where recall that $N = 128 \times 256 = 32,768$ level-3 pixels, and $\mathbf{X}(\cdot)$ is a 5-dimensional vector made up of the same covariates used in the initial data analysis. In (59), the r-dimensional vector of random effects, $\boldsymbol{\eta}$, is assumed to have a Gau$(\mathbf{0}, \mathbf{K})$ distribution, where the covariance matrix $\mathbf{K}$ is fixed but unknown and will be estimated. We use mutiresolutional W-wavelet basis functions for $\mathbf{S}(\cdot)$; see Kang et al. (2010) and Kang and Cressie (2011). That is, we choose all 32 W-wavelets from the first resolution, and 62 W-wavelets from the second resolution, resulting in $r = 32 + 62 = 94$. The $N \times r$ matrix $\mathbf{S}$ of basis functions is further rescaled by dividing each column of $\mathbf{S}$ by the standard deviation of the elements of the corresponding column. Finally, the component $\xi(\cdot)$ denotes the fine-scale-variation parameter, and we model it using a Gau$(0, \sigma_\xi^2)$ distribution.

6.3 Parameter Estimation and Optimal Spatial Mapping of AOD

We use the EM algorithm (Section 4) to estimate the parameters $\boldsymbol{\theta} = \left\{\boldsymbol{\beta}, \mathbf{K}, \sigma_\xi^2\right\}$. To implement the EM algorithm, we obtain the starting values using the methods discussed in Section 4.3, with $\hat{\boldsymbol{\beta}}_{WGLM}$ used as the starting value for $\boldsymbol{\beta}$. The EM estimates, $\hat{\boldsymbol{\theta}}_{EM} \equiv \left\{\hat{\boldsymbol{\beta}}_{EM}, \hat{\mathbf{K}}_{EM}, \hat{\sigma}^2_{\xi;EM}\right\}$, are then substituted into an MCMC algorithm (Appendix C) to obtain samples from the empirical predictive distribution, $[\boldsymbol{\eta}, \boldsymbol{\xi}_O | \mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM}]$. We generated 20,000 MCMC samples, after discarding 2,000 samples as burn-in. These MCMC samples, together with MCMC samples from $[\boldsymbol{\xi}_U | \hat{\sigma}^2_{\xi;EM}]$, give us the entire empirical predictive distribution, $[\mathbf{Y} | \mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM}]$, or any desired transformation or summary of it. For example, we can obtain $[\boldsymbol{\mu}_{Z|Y} | \mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM}]$, where $\boldsymbol{\mu}_{Z|Y} \equiv (\mu_{Z|Y}(\mathbf{s}_1), \ldots, \mu_{Z|Y}(\mathbf{s}_N))^\top$ and $\mu_{Z|Y}(\cdot) = \exp(Y(\cdot))$, whose moments and quantiles are immediately computable.

Using the MCMC samples, we first computed the predictive mean and the predictive standard deviation of the process $Y(\cdot)$; see the left panels of Figure 7. These panels are comparable to the optimal predictions in Shi and Cressie (2007), Kang et al. (2010), and Kang and Cressie (2011), which are on the log scale. The predictive mean of $Y(\cdot)$ shows that high aerosol particles are emitted from the Sahara desert and make their way across the Atlantic Ocean to North America via mid-latitude trade winds. The map of predictive standard deviations reflects the satellite tracks and regions of missing data, as it should. The additive nature of the model for $Y(\cdot)$ allows us to map and interpret different sources of variability separately. Specifically, the right panels of Figure 7 show image plots for the trend component $\mathbf{X}(\cdot)^\top \hat{\boldsymbol{\beta}}_{EM}$, for the predictive mean of the small-scale variation component $\mathbf{S}(\cdot)^\top \boldsymbol{\eta}$, and for the predictive mean of the fine-scale-variation component $\xi(\cdot)$. Adding them together, we obtain the predictive mean of $Y(\cdot)$ shown in the middle-left panel of Figure 7.

—— Figure 7 approximately here ——

Recall that the datum $Z(\mathbf{s}_i)$ was obtained by averaging $m(\mathbf{s}_i)$ level-2 observations observed in the level-3 pixel located at $\mathbf{s}_i$; $i = 1, \ldots, n$. We incorporated that heterogeneity in our hierarchical model through (54), and to assess its impact we made side-by-side boxplots showing how the predictive standard deviation of $Y(\cdot)$ varies for different values of $m(\mathbf{s}_i)$; see Figure 8. As expected, the predictive standard deviation of $Y(\mathbf{s}_i)$ decreases as $m(\mathbf{s}_i)$ increases, reflecting the importance of the data model in this spatial statistical analysis.

—— Figure 8 approximately here ——

Our goal in this analysis is to make inference on the original AOD scale. Here we obtained maps of the mean, the standard deviation, the 2.5 percentile, and the 97.5 percentile of each of the N elements of $\boldsymbol{\mu}_{Z|Y}$ in the (empirical) predictive distribution $[\boldsymbol{\mu}_{Z|Y}|\mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM}]$; see Figure 9. Notice that the map of the predictive standard deviation shows a mean-variance relationship, which is the consequence of the Lognormal process model for $\mu_{Z|Y}(\cdot)$. The maps showing the 2.5 percentile and the 97.5 percentile give the upper bound and lower bound, respectively, of pixelwise 95% credible intervals. All panels in Figure 9 show maps on the original AOD scale, where they are most interpretable scientifically.

—— Figure 9 approximately here ——

7 Discussion and Conclusions

In this article, we have developed a hierarchical spatial statistical model where the data model belongs to the exponential family of distributions. The process model is spatially dependent and is based on a hidden SRE model for the underlying latent random process. This allows for nonstationarity and dimension reduction, which is advantageous when analyzing big, spatially heterogeneous datasets. The spatially independent fine-scale variation term is an important component of the SRE model and is an attempt to account for the variability that the fixed-rank random-effects do not capture. The fixed-rank random-effects term, coupled with the spatially independent fine-scale variability term, enables efficient computation via repeated use of the Sherman-Morrison-Woodbury formula. The model parameters are assumed fixed but unknown and are estimated.

The spatial independence of the fine-scale variation term, $\xi(\cdot)$, assumed in this article can be generalized to allow for some spatial dependence, for which sparse-matrix-inversion techniques can be used to invert its covariance matrix. This situation has been explored in Nguyen et al. (2012), where the orbit geometry of the satellite leads to spatial dependence in the fine-scale variation term.

The model proposed in this article is spatial-only. However, it could be extended to a hierarchical spatio-temporal model in an obvious way. We could use the same data model and a process model where the reduced-dimensional basis function coefficients evolve over time (e.g., Wikle et al., 2001; Cressie et al., 2010). There remain the problems of estimation of spatio-temporal-model parameters and optimal filtering, smoothing, and forecasting from the empirical predictive distribution.

Because of our *empirical* hierarchical modeling (EHM) approach, we are able to avoid spatial confounding between fixed-effects and random-effects terms in the process model. We have developed an EM algorithm to estimate the unknown parameters; since the expectations required in the E-step of the EM algorithm are not available in closed form, we developed a Laplace approximation for them.

Based on a simulation experiment, we assessed the performance of EM estimation of the parameters, and then we investigated the predictive properties of our EHM approach. We further used the simulation set-up to compare the performance of our EHM approach to that of a comparable BHM approach, both in terms of computational efficiency (EHM is 6-10 times faster) and in terms of width of credible intervals (EHM is 75-80% more liberal).

Finally, we used our methodology to analyze a big, spatially heterogeneous dataset on AOD. Based on a Gamma data model and a Lognormal process model, and after properly accounting for sources of heterogeneity, we obtained a map of optimal spatial predictions of AOD on the original scale, along with maps quantifying the uncertainty of that prediction.

In conclusion, we have presented an empirical hierarchical modeling (EHM) approach that captures non-linear, non-Gaussian, spatial variability, has a geostatistical process model, and is well suited to the analysis of big data.

Acknowledgments

This research was partially supported by NASA's Earth Science Technology Office through its Advanced Information Systems Technology Program (Sengupta) and by the NSF and Census Bureau under NSF Grant SES-1132031, funded through the NCRN program (Cressie). We would like to thank Christopher K. Wikle for his valuable comments on the issue of spatial confounding and Amy Braverman for her feedback on various aspects of this research. We also express our thanks to the editor and the anonymous referees for their valuable suggestions.

References

Aldworth, J. and Cressie, N. (1999). "Sampling designs and prediction methods for Gaussian spatial processes." In *Multivariate Analysis, Designs of Experiments, and Survey Sampling*, ed. S. Ghosh, 1–54. New York, NY: Markel Dekker, Inc.

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.

Besag, J., York, J., and Mollié, A. (1991). "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics*, 43, 1–20.

Bradley, J. R., Cressie, N., and Shi, T. (2011). "Selection of rank and basis functions in the Spatial Random Effects model." In *Proceedings of the 2011 Joint Statistical Meetings*, 3393–3406. Alexandria, VA: American Statistical Association.

Brooks, S. P. and Gelman, A. (1998). "General methods for monitoring convergence of iterative simulations." *Journal of Computational and Graphical Statistics*, 7, 434–455.

Cressie, N. (1993). *Statistics for Spatial Data*. rev. ed. New York, NY: Wiley.

Cressie, N. and Johannesson, G. (2006). "Spatial prediction for massive data sets." In *Australian*

Academy of Science Elizabeth and Frederick White Conference, 1–11. 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., 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. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley.

Dempster, A. P., Laird, N., and Rubin, D. (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). "Model-based geostatistics." *Journal of the Royal Statistical Society, Series C*, 47, 299–350.

Duncan, W. J. (1944). "Some devices for the solution of large sets of simultaneous linear equations (with an appendix on the reciprocation of partitioned matrices)." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Seventh Series*, 35, 660–670.

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7, 457–472.

Heagerty, P. J. and Lele, S. R. (1998). "A composite likelihood approach to binary spatial data." *Journal of the American Statistical Association*, 93, 1099–1111.

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

Hodges, J. S. and Reich, B. J. (2010). "Adding spatially-correlated errors can mess up the fixed effect you love." *The American Statistician*, 64, 325–334.

Hughes, J. and Haran, M. (2013). “Dimension reduction and alleviation of confounding for spatial generalized linear mixed models.” *Journal of the Royal Statistical Society, Series B*, 75, 139–159.

Kang, E. L. and Cressie, N. (2011). “Bayesian inference for the Spatial Random Effects model.” *Journal of the American Statistical Association*, 106, 972–983.

Kang, E. L., Cressie, N., and Shi, T. (2010). “Using temporal variability to improve spatial mapping with application to satellite data.” *Canadian Journal of Statistics*, 38, 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.

Kass, R. E. and Steffey, D. (1989). “Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models).” *Journal of the American Statistical Association*, 84, 717–726.

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

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

Lindgren, F., Rue, H., and Lindström, J. (2011). “An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach.” *Journal of the Royal Statistical Society, Series B*, 73, 423–498.

Lindley, D. V. and Smith, A. F. M. (1972). “Bayes estimates for the linear model.” *Journal of the Royal Statistical Society, Series B*, 34, 1–41.

Lindsay, B. G. (1988). “Composite likelihood methods.” *Contemporary Mathematics*, 80, 221–239.

Lopes, H. F., Gamerman, D., and Salazar, E. (2011). “Generalized spatial dynamic factor models.” *Computational Statistics and Data Analysis*, 55, 1319 – 1330.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. London, UK: Chapman and Hall.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2001). *Generalized, Linear, and Mixed Models*. New York, NY: Wiley.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. 2nd ed. New York, NY: Wiley-Interscience.

Monestiez, P., Dubroca, L., Bonnin, E., Durbec, J.-P., and Guinet, C. (2006). “Geostatistical modelling of spatial distribution of *Balaenoptera physalus* in the Northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts.” *Ecological Modelling*, 193, 615 – 628.

Nguyen, H., Cressie, N., and Braverman, A. (2012). “Spatial statistical data fusion for remote sensing applications.” *Journal of the American Statistical Association*, 107, 1004–1018.

Omre, H. (1987). “Bayesian Kriging – merging observations and qualified guesses in kriging.” *Mathematical Geology*, 19, 25–39.

Omre, H. and Tjelmeland, H. (1997). “Petroleum geostatistics.” In *Geostatistics Wollongong ’96 (Vol. 1)*, eds. E. Y. Baafi and N. A. Schofield, 41–52. Dordrecht, NL: Kluwer Academic Publishers.

Paciorek, C. J. (2010). “The importance of scale for spatial-confounding bias and precision of spatial regression estimators.” *Statistical Science*, 25, 107–125.

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006). “Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models.” *Biometrics*, 62, 1197–1206.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York, NY: Springer.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. London, UK: Chapman & Hall/CRC.

Sengupta, A. and Cressie, N. (2013). “Empirical hierarchical modeling for count data using the Spatial Random Effects model.” *Spatial Economic Analysis*, forthcoming.

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.

Simpson, D., Lindgren, F., and Rue, H. (2012). “Think continuous: Markovian Gaussian models in spatial statistics.” *Spatial Statistics*, 1, 16 – 29.

Stein, M. L. (2008). “A modeling approach for large spatial datasets.” *Journal of the Korean Statistical Society*, 37, 3 – 10.

Wikle, C. K. (2010). “Low-rank representations for spatial processes.” In *Handbook of Spatial Statistics*, eds. A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, 107–118. Boca Raton, FL: Chapman and Hall/CRC.

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

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

Appendix

A Approximations Involved in the EM Algorithm

Let $\boldsymbol{\delta} \equiv (\boldsymbol{\eta}^\top, \boldsymbol{\xi}^\top)^\top$ be an m $(m = r+n)$-dimensional vector. Here we derive the Laplace approximation to the density $[\boldsymbol{\delta}|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]}]$. Let $\hat{\boldsymbol{\delta}}^{[l]}$ maximize the complete data log likelihood, $L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \boldsymbol{\delta})$. Now, the density for the distribution of $[\boldsymbol{\delta}|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]}]$ is given by:

$$p(\boldsymbol{\delta}|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]}) \propto \exp\left(L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \boldsymbol{\delta})\right). \tag{A.1}$$

A second-order Taylor-series approximation of $L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \boldsymbol{\delta})$ around $\hat{\boldsymbol{\delta}}^{[l]}$ yields:

$$\begin{aligned} L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \boldsymbol{\delta}) &= L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \hat{\boldsymbol{\delta}}^{[l]}) + \frac{1}{2}(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}^{[l]})^\top \left[\frac{\partial^2}{\partial\boldsymbol{\delta}^\top\partial\boldsymbol{\delta}} L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \boldsymbol{\delta})\right]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}^{[l]}} (\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}^{[l]}) \\ &\quad + \text{higher-order terms} \\ &\approx L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \hat{\boldsymbol{\delta}}^{[l]}) - \frac{1}{2}(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}^{[l]})^\top \mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]}, \boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}^{[l]}), \end{aligned} \tag{A.2}$$

where $\mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]}, \boldsymbol{\theta}^{[l]}|\mathbf{Z}_O) \equiv -\left[\frac{\partial^2}{\partial\boldsymbol{\delta}^\top\partial\boldsymbol{\delta}} L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \boldsymbol{\delta})\right]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}^{[l]}}$. In (A.2) above, notice that the first-order linear term is zero since the first-order derivative of $L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$, evaluated at $\boldsymbol{\delta} = \hat{\boldsymbol{\delta}}^{[l]}$, is zero (recall that $\hat{\boldsymbol{\delta}}^{[l]}$ maximizes $L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \boldsymbol{\delta})$). Therefore, for the density of $[\boldsymbol{\delta}|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]}]$, we have approximately,

$$p(\boldsymbol{\delta}|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]}) \propto \exp\left(L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \hat{\boldsymbol{\delta}}^{[l]})\right) \times \exp\left(-\frac{1}{2}(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}^{[l]})^\top \mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]}, \boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}^{[l]})\right). \tag{A.3}$$

Thus, $p(\boldsymbol{\delta}|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]})$ is approximately proportional to a Gaussian density. Evaluating the proportionality constant on the right-hand side of (A.3) yields the approximation:

$$\int p(\boldsymbol{\delta}|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]}) d\boldsymbol{\delta} = \exp\left(L_c(\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O, \hat{\boldsymbol{\delta}}^{[l]})\right)(2\pi)^{m/2}|\mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]}, \boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)|^{-1/2}, \tag{A.4}$$

and hence the first two moments are approximately,

$$
\begin{aligned}
&E(\boldsymbol{\delta}|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}) = \hat{\boldsymbol{\delta}}^{[l]} \\
&\mathrm{var}(\boldsymbol{\delta}|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}) = \mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]},\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)^{-1}.
\end{aligned}
\tag{A.5}
$$

Next, for $k = 1,2$, we derive the expectation:

$$
E\left(h_k\left(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\boldsymbol{\eta}+\boldsymbol{\xi}(\mathbf{s}_i)\right)|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right) \equiv E\left(h_k\left(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{q}(\mathbf{s}_i)^\top\boldsymbol{\delta}\right)|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right).
$$

Using a second-order Taylor-series expansion of $h_k(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{q}(\mathbf{s}_i)^\top\boldsymbol{\delta})$ around $\hat{\boldsymbol{\delta}}^{[l]}$, we obtain:

$$
\begin{aligned}
&h_k(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{q}(\mathbf{s}_i)^\top\boldsymbol{\delta}) \\
&= h_k\left(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{q}(\mathbf{s}_i)^\top\hat{\boldsymbol{\delta}}^{[l]}\right) \\
&\quad + (\boldsymbol{\delta}-\hat{\boldsymbol{\delta}}^{[l]})^\top\left(h_k'\left(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{q}(\mathbf{s}_i)^\top\hat{\boldsymbol{\delta}}^{[l]}\right)\times\mathbf{q}(\mathbf{s}_i)\right) \\
&\quad + \frac{1}{2}(\boldsymbol{\delta}-\hat{\boldsymbol{\delta}}^{[l]})^\top\left(h_k''\left(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{q}(\mathbf{s}_i)^\top\hat{\boldsymbol{\delta}}^{[l]}\right)\times\mathbf{q}(\mathbf{s}_i)\mathbf{q}(\mathbf{s}_i)^\top\right)(\boldsymbol{\delta}-\hat{\boldsymbol{\delta}}^{[l]}) \\
&\quad + \text{higher-order terms},
\end{aligned}
\tag{A.6}
$$

where the vector $h_k'(\mathbf{x}_0) \equiv \frac{d}{d\mathbf{x}}h_k(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}_0}$, and the matrix $h_k''(\mathbf{x}_0) \equiv \frac{d^2}{d\mathbf{x}^\top d\mathbf{x}}h_k(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}_0}$.

Taking expectations, we obtain:

$$
\begin{aligned}
&E\left(h_k(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{q}(\mathbf{s}_i)^\top\boldsymbol{\delta})|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right) \\
&\approx h_k\left(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{q}(\mathbf{s}_i)^\top\hat{\boldsymbol{\delta}}^{[l]}\right) \\
&\quad + E\left((\boldsymbol{\delta}-\hat{\boldsymbol{\delta}}^{[l]})|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right)^\top\left(h_k'\left(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{q}(\mathbf{s}_i)^\top\hat{\boldsymbol{\delta}}^{[l]}\right)\times\mathbf{q}(\mathbf{s}_i)\right) \\
&\quad + \frac{1}{2}\mathrm{tr}\Bigg\{E\left((\boldsymbol{\delta}-\hat{\boldsymbol{\delta}}^{[l]})(\boldsymbol{\delta}-\hat{\boldsymbol{\delta}}^{[l]})^\top|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right) \\
&\quad \times\left(h_k''\left(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{q}(\mathbf{s}_i)^\top\hat{\boldsymbol{\delta}}^{[l]}\right)\times\mathbf{q}(\mathbf{s}_i)\mathbf{q}(\mathbf{s}_i)^\top\right)\Bigg\}.
\end{aligned}
\tag{A.7}
$$

The second term in (A.7) is zero, since $\hat{\boldsymbol{\delta}}^{[l]}$ is the expectation of the Gaussian density that approximates the posterior density, $[\boldsymbol{\delta}|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]}]$; see (A.5). Consequently, we obtain:

$$
\begin{aligned}
&E\left(h_k(C(\mathbf{s}_i) + \mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta} + \mathbf{q}(\mathbf{s}_i)^\top \boldsymbol{\delta})|\mathbf{Z}_O, \boldsymbol{\theta}^{[l]}\right) \\
&\approx h_k\left(\mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta} + \mathbf{q}(\mathbf{s}_i)^\top \hat{\boldsymbol{\delta}}^{[l]}\right) \\
&\quad + \frac{1}{2}\mathrm{tr}\left\{\mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]}, \boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)^{-1}\left(h_k''\left(C(\mathbf{s}_i) + \mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta} + \mathbf{q}(\mathbf{s}_i)^\top \hat{\boldsymbol{\delta}}^{[l]}\right) \times \mathbf{q}(\mathbf{s}_i)\mathbf{q}(\mathbf{s}_i)^\top\right)\right\} \\
&= h_k\left(C(\mathbf{s}_i) + \mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta} + \mathbf{q}(\mathbf{s}_i)^\top \hat{\boldsymbol{\delta}}^{[l]}\right) \\
&\quad + \frac{1}{2}h_k''\left(C(\mathbf{s}_i) + \mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta} + \mathbf{q}(\mathbf{s}_i)^\top \hat{\boldsymbol{\delta}}^{[l]}\right) \times \mathbf{q}(\mathbf{s}_i)^\top \mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]}, \boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)^{-1}\mathbf{q}(\mathbf{s}_i).
\end{aligned}
\tag{A.8}
$$

Recall that $\boldsymbol{\delta} \equiv (\boldsymbol{\eta}^\top, \boldsymbol{\xi}^\top)^\top$. Therefore, from (A.5) and (A.8), we obtain the approximations to the expectations involved in the E-step of the EM algorithm, that are used in (23), (24), and (30).

B Evaluations for the One-Step Newton-Raphson Update for $\boldsymbol{\beta}$

In this part of the Appendix, we evaluate the expressions involved in the one-step Newton-Raphson update for $\boldsymbol{\beta}$, which was discussed at the end of Section 4.2. Specifically, we will evaluate the score function $\mathbf{R}(\boldsymbol{\theta})$ and its derivative with respect to $\boldsymbol{\beta}$, assuming as many derivatives for $h_1(\cdot)$ and $h_2(\cdot)$ as necessary.

The expression for $Q(\cdot, \cdot)$ given by (22), after substituting in the approximations to the required expectations, becomes

$$\begin{aligned}Q(\boldsymbol{\theta},\boldsymbol{\theta}^{[l]}) =\ & \text{const.} + \Bigg\{\sum_{i=1}^{n} Z(\mathbf{s}_i)\Big\{h_1(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]}+\hat{\xi}^{[l]}(\mathbf{s}_i)) \\ & +\frac{1}{2}h_1''\left(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]}+\hat{\xi}^{[l]}(\mathbf{s}_i)\right)\times\mathbf{q}(\mathbf{s}_i)^\top\mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]},\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)^{-1}\mathbf{q}(\mathbf{s}_i)\Big\} \\ & -\sum_{i=1}^{n}\Big\{h_2(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]}+\hat{\xi}^{[l]}(\mathbf{s}_i)) \\ & +\frac{1}{2}h_2''\left(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]}+\hat{\xi}^{[l]}(\mathbf{s}_i)\right)\times\mathbf{q}(\mathbf{s}_i)^\top\mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]},\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)^{-1}\mathbf{q}(\mathbf{s}_i)\Big\}\Bigg\}/\tau^2 \\ & -\frac{1}{2}\log|\mathbf{K}|-\frac{1}{2}\text{trace}\left(\hat{E}\left(\boldsymbol{\eta}\boldsymbol{\eta}^\top|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right)\mathbf{K}^{-1}\right) \\ & -\frac{n}{2}\log\sigma_\xi^2-\frac{1}{2\sigma_\xi^2}\text{trace}\left(\hat{E}\left(\boldsymbol{\xi}_O\boldsymbol{\xi}_O^\top|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right)\mathbf{V}_{\xi;O}^{-1}\right),\end{aligned} \tag{B.1}$$

where $\mathbf{q}(\mathbf{s})$ and $\mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]},\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)$ are defined in Appendix A; the approximations, $\hat{E}\left(\boldsymbol{\eta}\boldsymbol{\eta}^\top|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right)$ and $\hat{E}\left(\boldsymbol{\xi}_O\boldsymbol{\xi}_O^\top|\mathbf{Z}_O,\boldsymbol{\theta}^{[l]}\right)$, to the respective expectations, are given by (29) (which follows from Appendix A).

Now, to obtain the score function, $\mathbf{R}(\boldsymbol{\theta})$, we differentiate (B.1) with respect to $\boldsymbol{\beta}$, resulting in:

$$\begin{aligned}\mathbf{R}(\boldsymbol{\theta}) = & \Bigg\{\sum_{i=1}^{n} Z(\mathbf{s}_i)\Big\{h_1'(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]}+\hat{\xi}^{[l]}(\mathbf{s}_i)) \\ & +\frac{1}{2}h_1'''\left(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]}+\hat{\xi}^{[l]}(\mathbf{s}_i)\right)\times\mathbf{q}(\mathbf{s}_i)^\top\mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]},\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)^{-1}\mathbf{q}(\mathbf{s}_i)\Big\}\mathbf{X}(\mathbf{s}_i) \\ & -\sum_{i=1}^{n}\Big\{h_2'(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]}+\hat{\xi}^{[l]}(\mathbf{s}_i)) \\ & +\frac{1}{2}h_2'''\left(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]}+\hat{\xi}^{[l]}(\mathbf{s}_i)\right)\times\mathbf{q}(\mathbf{s}_i)^\top\mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]},\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)^{-1}\mathbf{q}(\mathbf{s}_i)\Big\}\mathbf{X}(\mathbf{s}_i)\Bigg\}/\tau^2\end{aligned} \tag{B.2}$$

The Newton-Raphson update (32) also requires the partial derivative of $\mathbf{R}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}$, which is given by:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}}\mathbf{R}(\boldsymbol{\theta}) = \Bigg\{ & \sum_{i=1}^{n} Z(\mathbf{s}_i)\Big\{h_1''(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]}+\hat{\xi}^{[l]}(\mathbf{s}_i)) \\ & +\frac{1}{2}h_1^{iv}\Big(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]}+\hat{\xi}^{[l]}(\mathbf{s}_i)\Big)\times\mathbf{q}(\mathbf{s}_i)^\top\mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]},\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)^{-1}\mathbf{q}(\mathbf{s}_i)\Big\}\mathbf{X}(\mathbf{s}_i)\mathbf{X}(\mathbf{s}_i)^\top \\ & -\sum_{i=1}^{n}\Big\{h_2''(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]}+\hat{\xi}^{[l]}(\mathbf{s}_i)) \\ & +\frac{1}{2}h_2^{iv}\Big(C(\mathbf{s}_i)+\mathbf{X}(\mathbf{s}_i)^\top\boldsymbol{\beta}+\mathbf{S}(\mathbf{s}_i)^\top\hat{\boldsymbol{\eta}}^{[l]}+\hat{\xi}^{[l]}(\mathbf{s}_i)\Big)\times\mathbf{q}(\mathbf{s}_i)^\top\mathbf{Q}_{LA}(\boldsymbol{\delta}^{[l]},\boldsymbol{\theta}^{[l]}|\mathbf{Z}_O)^{-1}\mathbf{q}(\mathbf{s}_i)\Big\}\mathbf{X}(\mathbf{s}_i)\mathbf{X}(\mathbf{s}_i)^\top\Bigg\}/\tau^2.\end{aligned} \tag{B.3}$$

Then (B.3) is evaluated at $\boldsymbol{\theta}=\boldsymbol{\theta}^{[l]}$, and its matrix inverse is taken; it is then substituted into (32).

C MCMC Algorithm

Here we describe the MCMC procedure that is used to obtain samples from the predictive distribution, $[\boldsymbol{\eta},\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}]$. We implement the MCMC procedure with a Gibbs sampler, incorporating Metropolis-Hastings steps where necessary. The full conditional distributions, as well as details of the Metropolis Hastings steps, are described in the following paragraph.

The joint distribution, $[\mathbf{Z}_O,\boldsymbol{\eta},\boldsymbol{\xi}_O|\boldsymbol{\theta}]$, can be written as:

$$[\mathbf{Z}_O,\boldsymbol{\eta},\boldsymbol{\xi}_O|\boldsymbol{\theta}]\equiv[\mathbf{Z}_O|\boldsymbol{\eta},\boldsymbol{\xi}_O,\boldsymbol{\beta}]\times[\boldsymbol{\eta}|\mathbf{K}]\times[\boldsymbol{\xi}_O|\sigma_\xi^2]. \tag{C.1}$$

Let "$[\mathbf{A}|\mathbf{B},\cdot]$" denote the full conditional distribution of the unknown **A** given **B** and all other unknowns (and the data). The Gibbs sampler uses the following steps to generate samples from the predictive distribution, $[\boldsymbol{\eta},\boldsymbol{\xi}_O|\mathbf{Z}_O,\boldsymbol{\theta}]$.

1. At $t=0$, we select starting values $\boldsymbol{\eta}^{[0]}$ and $\boldsymbol{\xi}_O^{[0]}$.

2. t=t+1; simulate successively from the full conditionals, $[\boldsymbol{\eta}^{[t+1]}|\boldsymbol{\xi}_O^{[t]},\cdot]$ and $[\boldsymbol{\xi}_O^{[t+1]}|\boldsymbol{\eta}^{[t+1]},\cdot]$.

3. Repeat step 2 to generate as many samples as needed.

4. Discard an initial number of samples as "burn-in."

The full conditionals are not available in closed form, so we use a Metropolis-Hastings step within the Gibbs sampler. A generic version of the algorithm that we have used to draw samples from the full conditionals, $[\boldsymbol{\eta}^{[t+1]}|\boldsymbol{\xi}_O^{[t]},\cdot]$ and $[\boldsymbol{\xi}_O^{[t+1]}|\boldsymbol{\eta}^{[t+1]},\cdot]$ (at the $(t+1)$-th stage), is discussed below. Suppose $\mathbf{a}$ is the random variable (or a block of random variables) that we are updating, and $\mathbf{a}_0$ is the most recently sampled value. We follow the steps below to obtain a new sample of $\mathbf{a}$:

1. Draw a trial value $\mathbf{a}_1$ from a proposal density, $\text{Gau}(\mathbf{a}_0,\boldsymbol{\Sigma}_a)$.
2. Generate U_1 uniformly on $(0,1)$.
3. Compute the joint density of $\mathbf{a}$ and all other unknowns, $l(\mathbf{a}_0,\text{rest})$ and $l(\mathbf{a}_1,\text{rest})$ where "rest" denotes all the other unknowns fixed at their most recently sampled value.
4. If $U_1 < \min\left\{\frac{l(\mathbf{a}_1,\text{rest})}{l(\mathbf{a}_0,\text{rest})},1\right\}$, accept the trial value $\mathbf{a}_1$ and keep it for the most current iteration; otherwise, the value $\mathbf{a}_0$ is retained.

When sampling from $[\boldsymbol{\eta}^{[t+1]}|\boldsymbol{\xi}_O^{[t]},\cdot]$, we update $\boldsymbol{\eta}$ as a block. To sample from $[\boldsymbol{\xi}_O^{[t+1]}|\boldsymbol{\eta}^{[t+1]},\cdot]$, we update $\boldsymbol{\xi}_O$ elementwise.

D BHM: Prior Specifications and the MCMC Algorithm

In this part of the Appendix, we present the prior distributions (or the parameter model) of BHM and fully Bayesian inference using the MCMC algorithm.

Following Kang and Cressie (2011), the prior distribution of $\boldsymbol{\theta} = (\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2)$ is assumed to be made up of mutually independent components:

$$[\boldsymbol{\beta},\mathbf{K},\sigma_\xi^2] = [\boldsymbol{\beta}]\cdot[\mathbf{K}]\cdot[\sigma_\xi^2]. \tag{D.1}$$

Next we assume that the p-dimensional fixed-effects parameters, $\boldsymbol{\beta}$, have a Gaussian prior distribution,

$$\boldsymbol{\beta} \sim \text{Gau}(\boldsymbol{\mu}_\beta,\boldsymbol{\Sigma}_\beta), \tag{D.2}$$

where $\boldsymbol{\mu}_\beta$ and $\boldsymbol{\Sigma}_\beta \equiv \text{diag}(\sigma^2_{\beta;1}, \ldots, \sigma^2_{\beta;p})$ are known hyperparameters. For fine-scale-variance parameter σ^2_ξ, we assume that $\sigma_\xi \sim$Uniform$(0, \kappa_\xi)$, where κ_ξ is a known hyperparameter. Finally, the prior distribution on **K** is based on the spectral decomposition,

$$\mathbf{K} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^\top, \tag{D.3}$$

where $\boldsymbol{\Lambda} \equiv \text{diag}(\lambda_1, \ldots, \lambda_r)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$, and **P** is an orthogonal matrix that can be parametrized in terms of the $r(r-1)/2$ Givens angles,

$$\boldsymbol{\theta}_G \equiv \left\{\theta_{ij} : i = 1, \ldots, r-1, j = i+1, \ldots, r\right\}.$$

In terms of these Givens angles, we can write **P** as (e.g., Kang and Cressie, 2011):

$$\mathbf{P} = (\mathbf{G}_{12}\mathbf{G}_{13}\ldots\mathbf{G}_{1r})(\mathbf{G}_{23}\ldots\mathbf{G}_{2r})\ldots\mathbf{G}_{(r-1)r},$$

where $\mathbf{G}_{ij}$ is the Givens rotation matrix corresponding to the Givens angle θ_{ij}, which is obtained by modifying the $r \times r$ identity matrix as follows: The i^{th} and the j^{th} diagonal elements of 1 are both replaced by cos(θ_{ij}), and the $(i,j)^{\text{th}}$ and $(j,i)^{\text{th}}$ elements of 0 are replaced by $-$sin(θ_{ij}) and sin(θ_{ij}), respectively.

We assign priors to the eigenvalues $\{\lambda_i : i = 1, \ldots, r\}$ and the Givens angles $\boldsymbol{\theta}_G$, using models discussed in Kang and Cressie (2011). That is,

$$[\lambda_1, \ldots, \lambda_r] = [\lambda_{1,1}, \ldots, \lambda_{1,q_1}] \cdots [\lambda_{K,1}, \ldots, \lambda_{K,q_K} | \lambda_{K-1,q_{K-1}}], \tag{D.4}$$

where $\lambda_{k,1}, \ldots, \lambda_{k,q_k}$ are the eigenvalues corresponding to the q_k basis functions from the k-th resolution, $k = 1, \ldots, K$, and $\sum_{k=1}^K q_k = r$. Finally, $\lambda_{k,1}, \ldots, \lambda_{k,q_k}$ are assumed to be distributed as order statistics corresponding to i.i.d. truncated Lognormal random variables with known hyperparameters, mean μ_k and variance σ^2_k, for $k = 1, \ldots, K$, where the Lognormal distribution is restricted to $(0, \lambda_{k-1,q_{k-1}})$.

We define the prior on θ_{ij} through a prior on the logit transformation of θ_{ij}, namely

$$h(\theta_{ij}) \equiv \log\left(\frac{\pi/2+\theta_{ij}}{\pi/2-\theta_{ij}}\right). \tag{D.5}$$

Then we assign independent priors on $h(\theta_{ij})$ as

$$h(\theta_{ij}) \sim \mathrm{Gau}(c_k, \tau_k^2), \tag{D.6}$$

if i, j both belong to the same resolution k, where $k = 1, \ldots, K$; otherwise,

$$h(\theta_{ij}) \sim \mathrm{Gau}(0, \tau_0^2), \tag{D.7}$$

if i, j belong to different resolutions. The hyperparameters $\{c_k\}$, $\{\tau_k^2\}$, and τ_0^2 are assumed known.

We also specify the hyperparameters following the recommendations in Kang and Cressie (2011). In the simulation study described in this article, the true parameter values, $\boldsymbol{\theta}_T$, were used to specify the hyperparameters. We selected $\mu_\beta = \boldsymbol{\beta}_T$, and the elements of the covariance matrix $\boldsymbol{\Sigma}_\beta$ were specified as three times the square of the standard-errors obtained by fitting a classical fixed-effects Poisson GLM (e.g., McCullagh and Nelder, 1989, Chapter 6) to the data, with the same covariates that were used for the simulation. Next we chose $\kappa_\xi = 10\sigma_{\xi;T}$.

Finally, to specify the hyperparameters in the prior on $\mathbf{K}$, we first obtained:

$$\mathbf{K}_T = \mathbf{P}_T \boldsymbol{\Lambda}_T \mathbf{P}_T^\top,$$

where $\boldsymbol{\Lambda}_T \equiv (\lambda_{1;T}, \ldots, \lambda_{r;T})$. We also computed the Givens angles for $\mathbf{K}_T$, namely,

$$\left\{\theta_{ij;T} : i = 1, \ldots, r-1, ; j = i+1, \ldots, r\right\}.$$

For $k = 1, \ldots, K$, we specified:

$$\begin{aligned} \mu_k &= \sum_{i=1}^{q_k} \log(\lambda_{k,i;T})/q_k \\ \sigma_k^2 &= \sum_{i=1}^{q_k} (\log(\lambda_{k,i;T} - \mu_k)^2/(q_k - 1). \end{aligned} \tag{D.8}$$

Similarly, we specified $\{c_k\}$, $\left\{\tau_k^2\right\}$, and τ_0^2 as:

$$\begin{aligned} c_k &= \sum_{(i,j)\in N_k} h(\theta_{ij;T})/|N_k|, \\ \tau_k^2 &= \sum_{(i,j)\in N_k} (h(\theta_{ij;T}) - c_k)^2/(|N_k| - 1), \\ \tau_0^2 &= \sum_{(i,j)\in N_0} h(\theta_{ij;T})^2/|N_0|, \end{aligned} \tag{D.9}$$

where $h(\cdot)$ is given by (D.5), $N_k \equiv \{(i,j) :$ the i-th and the j-th basis functions are both of the k-th resolution$\}$, $k = 1, \ldots, K$, and $N_0 \equiv \{(i,j) :$ the i-th and the j-th basis functions are of different resolutions$\}$.

Finally, we implemented the MCMC procedure with a Gibbs sampler to generate samples from the posterior distribution, $[\boldsymbol{\eta}, \boldsymbol{\xi}_O, \boldsymbol{\xi}_U, \boldsymbol{\theta}|\mathbf{Z}_O]$. The full conditionals of σ_ξ^2 and $\boldsymbol{\xi}_U$ can be derived in closed form. The full conditional of $\boldsymbol{\xi}_U$ is:

$$[\boldsymbol{\xi}_U|\mathbf{Z}_O, \boldsymbol{\eta}, \boldsymbol{\xi}_O, \boldsymbol{\theta}] = [\boldsymbol{\xi}_U|\boldsymbol{\theta}].$$

The full conditional of σ_ξ^2 is a truncated Inverse-Gamma distribution, namely, $\mathrm{IG}((N-1)/2, \boldsymbol{\xi}^\top\boldsymbol{\xi}/2) \cdot I(0 < \sigma_\xi < k)$ (see Kang and Cressie, 2011), where recall that $\boldsymbol{\xi} = (\boldsymbol{\xi}_O^\top, \boldsymbol{\xi}_U^\top)^\top$. The other full conditionals are not available in closed form, so we incorporated a Metropolis-Hastings step, with random walk proposals, to simulate from them. Details of the Metropolis-Hastings algorithm is given in Appendix B. We updated $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ in blocks, and $\boldsymbol{\xi}_O$ elementwise. When sampling the eigenvalues, we updated in blocks according to resolution. If the total ordering of the eigenvalues was broken, we rejected the sample and a new sample was drawn until the ordering of the eigenvalues was preserved (Kang and Cressie, 2011). When sampling the Givens angles, we updated the Givens angles corresponding to the same resolution, $\left\{\theta_{ij} : (i, j) \in N_k\right\}$, as a block, for $k = 1, \ldots, K$, and the Givens angles $\left\{\theta_{ij} : (i, j) \in N_0\right\}$ were updated as a block.

Figure Captions

Figure 1: The left panel shows the identity matrix, and the right panel shows the matrix, $\mathrm{ave}(\hat{\mathbf{K}}_{EM})\mathbf{K}_T^{-1}$, where $\mathrm{ave}(\hat{\mathbf{K}}_{EM})$ is the elementwise average of the EM estimates $\left\{\hat{\mathbf{K}}_{EM}^{[l]} : l = 1, \ldots, 1600\right\}$. The common color bar is shown on the right.

Figure 2: Plot showing a histogram of $\left\{\mathrm{trace}(\hat{\mathbf{K}}_{EM}^{[l]}\mathbf{K}_T^{-1}) : l = 1, \ldots, 1600\right\}$. The chi-squared density with degrees of freedom equal to $r = 29$ is overlayed on the histogram.

Figure 3: The left panel corresponds to kernel-density plots showing the distribution of the SEHM mean squared prediction error divided by the IEHM mean squared prediction error, for locations with data (solid line) and for locations without data (dashed line). The right panel corresponds to kernel-density plots comparing the SEHM mean squared prediction errors obtained for locations with data (solid line) and for locations without data (dashed line)

Figure 4: Plots showing the Gelman-Rubin-Brooks statistics, for EHM (left panel) and for BHM (right panel), as a function of the number of MCMC samples. The solid line corresponds to the MPSRF for $\boldsymbol{\eta}$; the dashed line corresponds to the maximum of the elementwise PSRFs for $\boldsymbol{\eta}$. Here, the number of observations is $n = 5,000$.

Figure 5: Plots show the observed data (top-left panel), the true simulated process, $Y(\cdot)$ (top-right panel), the mean of the empirical predictive distribution, $\hat{Y}_{SEHM}(\cdot) \equiv E(Y(\cdot)|\mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM})$ (bottom-left panel), and the mean of the posterior distribution, $\hat{Y}_{SBHM}(\cdot) \equiv E(Y(\cdot)|\mathbf{Z}_O)$ (bottom-right panel).

Figure 6: The left panel corresponds to the kernel-density plot showing the distribution of the difference, $\hat{Y}_{SEHM}(\cdot) - \hat{Y}_{SBHM}(\cdot)$. The right panel corresponds to kernel-density plots showing the distribution of the ratio, $(\mathrm{var}(Y(\mathbf{s})|\mathbf{Z}_O))^{1/2}/(\mathrm{var}(Y(\mathbf{s})|\mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM})^{1/2}$, separately for locations with data and for locations without data.

Figure 7: Maps to the left show the log(AOD) (top-left panel), the mean (middle-left panel) and standard deviation (bottom-left panel) of the predictive distribution of $Y(\cdot)$, namely $[Y(\cdot)|\mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM}]$. Maps to the right show the predictive mean of the different components of variability in $Y(\cdot)$, namely, the components due to trend,$X(\cdot)^\top \hat{\boldsymbol{\beta}}_{EM}$ (top-right panel), the random-effects component, $E[\mathbf{S}(\cdot)^\top \boldsymbol{\eta}|\mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM}]$ (middle-right panel), and the fine-scale-variation component, $E[\boldsymbol{\xi}(\cdot)|\mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM}]$ (bottom-right panel). The middle-left panel which is a map of the mean of the predictive distribution of $Y(\cdot)$, namely $E[Y(\cdot)|\mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM}]$, is the sum of the three panels shown on the right

Figure 8: Boxplots showing the variability of the predictive standard deviation of $Y(\mathbf{s}_i)$ for values of $m(\mathbf{s}_i) = 1, 2, \ldots, 21$.

Figure 9: AOD data in D (top-left panel) and histogram showing their distribution (top-right panel). Maps show the predictive mean (middle-left panel), the pixelwise predictive standard deviation (middle-right panel), the pixelwise predictive 2.5 percentile (bottom-left panel), and the pixelwise predictive 97.5 percentile (bottom-right panel) obtained from the empirical predictive-distribution, $[\mu_{Z|Y}(\cdot)|\mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM}]$. The plots of the predictive mean and the predictive percentiles have the same color scale, where any value greater than 1 has been assigned the highest color-value.

Tables

Table 1: True parameter values and the sample mean of the EM parameter estimates based on 1600 simulated datasets. Each dataset is of size $n = 20,000$. The empirical root mean squared errors (RMSEs) of the parameter estimates are also reported.

Parameter	True value	Sample mean based on the 1600 simulated datasets	RMSE
β_1	2.0	1.922	0.0954
β_2	0.0125	0.01262	0.0002
σ^2_ξ	0.05	0.0507	0.002

Table 2: Gelman-Rubin-Brooks statistics for varying sample sizes (*n*). The number of MCMC samples generated are L=15,000 for EHM, and L=40,000 for BHM. MPSRF is the multivariate potential scale reduction factor, and max(PSRF) is the maximum of the elementwise potential scale reduction factors (PSRFs).

	EHM (L=15,000)			BHM (L=40,000)		
	$\boldsymbol{\eta}$		$\boldsymbol{\xi}_O$	$\boldsymbol{\eta}$		$\boldsymbol{\xi}$
Sample size (n)	MPSRF	max(PSRF)	max(PSRF)	MPSRF	max(PSRF)	max(PSRF)
5,000	1.08	1.028	1.0025	1.07	1.021	1.0011
10,000	1.07	1.028	1.0028	1.09	1.016	1.0011
15,000	1.09	1.027	1.0027	1.06	1.018	1.0014
20,000	1.07	1.027	1.0028	1.09	1.014	1.0012

Table 3: Computational time for varying sample sizes (n). For EHM, the EM algorithm was used to estimate the parameters, and then an MCMC algorithm was used to generate $L_{EHM} = 15,000$ samples from the empirical predictive distribution, $[\boldsymbol{\eta}, \boldsymbol{\xi}_O | \mathbf{Z}_O, \hat{\boldsymbol{\theta}}_{EM}]$. For BHM, an MCMC algorithm was used to generate $L_{BHM} = 40,000$ samples from the posterior distribution, $[\boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\theta} | \mathbf{Z}_O]$.

Sample size (n)	Computational Time (in hours)				Ratio of computational time (BHM/EHM)
	EHM (L=15,000)			BHM (L=40,000)	
	EM Estimation	MCMC Implementation	Total	MCMC Implementation	
5,000	0.02	0.16	0.18	3.95	21.94
20,000	0.02	0.62	0.64	5.79	9.04
35,000	0.02	1.01	1.03	7.61	7.38
50,000	0.04	1.45	1.49	8.70	5.83