Bayesian Inference for Big Spatial Data Using Non-stationary Spectral Simulation

Hou-Cheng Yang[*] and Jonathan R. Bradley[†]

Abstract

It is increasingly understood that the assumption of stationarity is unrealistic for many spatial processes. In this article, we combine dimension expansion with a spectral method to model big non-stationary spatial fields in a computationally efficient manner. Specifically, we use Mejía and Rodríguez-Iturbe (1974)'s spectral simulation approach to simulate a spatial process with a covariogram at locations that have an expanded dimension. We introduce Bayesian hierarchical modelling to dimension expansion, which originally has only been modeled using a method of moments approach. We consider a novel scheme to re-weight levels in a Bayesian spatial hierarchical model that allows one to use non-stationary spectral simulation within a collapsed Gibbs sampler. Our method is both full rank and non-stationary, and can be applied to big spatial data because it does not involve storing and inverting large covariance matrices. We demonstrate the wide applicability of our approach through simulation studies, and an application using ozone data obtained from the National Aeronautics and Space Administration (NASA).

[*] (to whom correspondence should be addressed) Department of Statistics, Florida State University, 117 N. Woodward Ave, Tallahassee, FL 32306, hy15e@my.fsu.edu

[†] Department of Statistics, Florida State University, 117 N. Woodward Ave, Tallahassee, FL 32306, bradley@stat.fsu.edu

1 Introduction

There is increasing interest in using spatial statistical methods to model environmental processes. This is partially due to the emergence of remote sensing instruments and the popularity of Geographic Information Systems (GIS) software (e.g. see, Stein et al., 2006; Kalkhan, 2011, for standard references). The main goal of these analyses is to make predictions at observed and unobserved locations and provide uncertainty quantification. Early works make the assumption that the process is weakly stationary (e.g., see Cressie, 1993, for a review); that is, the covariance between the response at two different locations is a function of the spatial lag. However, non-stationary processes are much more common in environmental systems observed over large heterogeneous spatial domains (see Bradley et al., 2016, for a discussion). There are many models for non-stationary spatial data, and reduced ranks basis function expansions have become a popular choice (Banerjee et al., 2008; Cressie and Johannesson, 2008). However, there are inferential issues with reduced rank methods in the spatial setting (Stein, 2014), and consequently, there is renewed interest in proposing computationally efficient approximately full-rank models (Nychka et al., 2015; Datta et al., 2016a; Katzfuss, 2017; Bradley et al., 2020; Katzfuss et al., 2020). Thus, in this article our primary goal is to develop an efficient full rank non-stationary spatial statistical model.

There are numerous methods available to model non-stationary spatial data. For example, process convolution (Higdon, 1998; Paciorek and Schervish, 2006; Neto et al., 2014) convolves a known spatially referenced function with a spatial process typically assumed to be Gaussian. There are several related, but different approaches available. For example, using a finite integral representation of a process convolution results in a basis function expansion (Cressie and Wikle, 2011, page 157). Several parameterizations of basis function expansions are available, including: fixed rank kriging (Cressie and Johannesson, 2008), lattice kriging (Nychka et al., 2015), the predictive process (Banerjee et al., 2008), and a stochastic partial differential equation approach (Lindgren et al., 2011), among others.

An alternative to modeling non-stationarity with spatial basis functions is to assume a deformation (Sampson and Guttorp, 1992). Here, Euclidean space is “deformed,” or warped, so that far away locations can be more correlated, and vice versa. The parameter space for this method is considerably smaller than many parameterizations using spatial basis function expansions (e.g., see Cressie and Johannesson, 2008; Kang and Cressie, 2011, for examples), and is full rank. A similar but different approach to deformation is referred to as “dimension expansion” (Bornn et al., 2012). This method involves extending the dimension of the locations to a higher dimensional space. This methodology is based on the surprising result that every non-stationary covariance function in $\mathbb{R}^d$ can be written as a stationary covariogram defined on locations in $\mathbb{R}^{2d}$ (Perrin and Meiring, 2003). Recently, Bornn et al. (2012) proposed a method of moments approach to analyzing spatio-temporal data using dimension expansion. To our knowledge, the dimension expansion approach has not been implemented using a Bayesian framework.

Thus, our first contribution is to introduce dimension expansion to the Bayesian setting to analyze big spatial data. To achieve a computationally efficient approach to dimension expansion in the Bayesian setting we offer three technical results. In our first technical result, we provide a “non-stationary version” of Bochner’s Theorem (Bochner, 1959). That is, we show that a non-stationary covariance function can be written as a convolution of the cosine function with a spectral density. The proof of this result simply involves combining Perrin and Meiring (2003)’s dimension expansion result with Bochner’s Theorem. This result opens up new opportunities to use spectral methods to model non-stationary spatial process. Other methods exist (e.g. see, Priestley, 1965; Martin, 1982) to model non-stationary data using spectral densities. However, these methods involve difficult to interpret types of “quasi-stationarity” assumptions (see, Sayeed and Jones, 1995, for a discussion), while our approach can be easily interpreted through dimension expansion. Castruccio and Guinness (2017) have also proposed an approach that uses evolutionary spectrum and incorporates an axial symmetric structure into their model.

The second technical result developed in this manuscript follows from our non-stationary version of Bochner's Theorem. Specifically, we extend Mejía and Rodríguez-Iturbe (1974)'s method for spectral simulation of a stationary spatial processes to non-stationary spatial processes. This makes it straightforward to simulate in the high-dimensional non-stationary setting because spectral simulation does not require the inverse and storage of a high-dimensional covariance matrix (i.e., is matrix free). In practice, Gaussian spatial datasets correspond to a likelihood that is difficult to compute in high dimensions (i.e., when the dimension of the data n is large) because this requires $O(n^3)$ computation and $O(n^2)$ dynamic memory. While non-stationary spectral simulation is "matrix free" the implementation of our model will require additional steps that include operations on low dimensional matrices. Consequently, we describe our method as "large-matrix-free."

To further aid in computation we develop a new technique to re-weight the data model in a spatial hierarchical model. Our re-weighting method imposes conditional independence between the spatially co-varying random effect at observed locations and the data given the covariance parameters, and conditional independence between variance/covariance parameters and a large subset of the data. As a result, our re-weighting method allows us to use non-stationary spectral simulation to update the spatially co-varying random vector within a collapsed Gibbs sampler (Liu, 1994), and update other parameters based on a low-dimensional sub-sample. Furthermore, these conditional independence assumptions do not change the marginal distribution for the data, and consequently, the marginal statistical properties of the data are invariant to our added assumptions of conditional independence. This re-weighting technique is similar to what is done in a recent paper by Bradley (2019). Overall, our method is computationally feasible, full-rank, does not require storage of large matrices, and can be implemented on irregularly spaced locations. This last feature is particularly important as spectral methods based on the discrete Fourier transform often require regularly spaced locations (Fuentes, 2002; Fuentes et al., 2008).

The remaining sections of this article are organized as follows. Section 2 introduces our proposed statistical model, our first two theoretical results, and our re-weighting technique. In Section 3, we describe our implementation using a collapsed Gibbs sampler. In Section 4, we present a simulation study and compare our approach to the Nearest Neighbor Gaussian Process (NNGP; Datta et al., 2016b) and the general Vecchia approximation (Katzfuss and Guinness, 2019) when data are generated from an additive model with nonlinear fixed effects and stationary spatial random effects. Additionally, we present a simulation study where we compare to a spatial process convolution (SPC; Paciorek and Schervish, 2006) approach, and a stochastic partial differential equation (SPDE; Lindgren et al., 2011) approach in the purely non-stationary setting. In Section 5, we implement our model using the benchmark ozone dataset analyzed in (Cressie and Johannesson, 2008) and (Zhang et al., 2019). Finally, Section 6 contains a discussion. For ease of exposition, all proofs are given in the appendices.

2 Methodology

Let $Z(\cdot)$ be a spatial process defined for all $\mathbf{s}{\in}D{\subset}\mathbb{R}^d$, where D is the spatial domain of interest in d-dimensional Euclidean space, $\mathbb{R}^d$. We observe the value of $Z(\cdot)$ at a finite set of locations $\mathbf{s}_1$, …, $\mathbf{s}_n{\in}D$. The data is decomposed additively with

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s}),$$

where $\mathbf{s}{\in}D$, $Y(\cdot)$ is the Gaussian process of principal interest, and the Gaussian process $\epsilon(\cdot)$ represents measurement error. The measurement error $\epsilon(\cdot)$ is assumed to be uncorrelated with mean-zero and variance σ_ϵ^2.

The process $Y(\cdot)$ is further decomposed as

$$Y(\mathbf{s}) = \mathbf{x}'(\mathbf{s})\boldsymbol{\beta} + \nu(\mathbf{s});\ \mathbf{s} \in D,$$

where $\mathbf{x}(s)$ is a known p-dimensional vector of covariates and $\boldsymbol{\beta} \in \mathbb{R}^p$ is unknown. For any collection of locations $\mathbf{u}_1, \ldots, \mathbf{u}_m$, the random vector $\boldsymbol{\nu} = (\nu(\mathbf{u}_1), \ldots, \nu(\mathbf{u}_m))'$ is assumed to have the probability density function (pdf),

$$p(\boldsymbol{\nu} \mid \boldsymbol{\theta}, \delta^2) = \int_{\mathbb{R}^m} p(\boldsymbol{\nu} \mid \boldsymbol{\theta}, \widetilde{\boldsymbol{\nu}}, \delta^2) p(\widetilde{\boldsymbol{\nu}} \mid \boldsymbol{\theta}) d\widetilde{\boldsymbol{\nu}}, \tag{1}$$

where $p(\boldsymbol{\nu} \mid \boldsymbol{\theta}, \widetilde{\boldsymbol{\nu}}, \delta^2)$ is the multivariate normal distribution density with mean $\widetilde{\boldsymbol{\nu}} \in \mathbb{R}^m$, covariance matrix $\delta^2\mathbf{I}_m$, and $\mathbf{I}_m$ is an $m \times m$ identity matrix. The pdf $p(\widetilde{\boldsymbol{\nu}} \mid \boldsymbol{\theta})$ will be specified in Section 2.2, but is approximately normal with mean zero, and $m \times m$ covariance matrix $\mathbf{C}(\boldsymbol{\theta})$, where the (i, j)-th element of $\mathbf{C}(\boldsymbol{\theta})$ is

$$C(\mathbf{s}_i, \mathbf{s}_j) = \sigma_\nu^2 \exp\left(-\frac{E(\mathbf{s}_i, \mathbf{s}_j)}{\phi}\right),$$

where

$$E(\mathbf{s}_i, \mathbf{s}_j) = \|\binom{\mathbf{s}_i}{\boldsymbol{\psi}'(\mathbf{s}_i)\boldsymbol{\eta}} - \binom{\mathbf{s}_j}{\boldsymbol{\psi}'(\mathbf{s}_j)\boldsymbol{\eta}}\|,$$

$\boldsymbol{\theta} = (\phi, \sigma_\nu^2, \boldsymbol{\eta}')'$, and $\| \cdot \|$ is a Euclidean distance. This covariance function uses the aforementioned dimension expansion approach from Bornn et al. (2012). Here, $\boldsymbol{\psi}(\mathbf{s}_i)$ is an $r \times d$ matrix consisting of known basis functions. This use of spatial basis functions is similar to the model in Shand and Li (2017). It will be useful to organize the n-dimensional vectors $\mathbf{Z} = \{Z(\mathbf{s}_1) \dots Z(\mathbf{s}_n)\}'$ and $\mathbf{Y} = \{Y(\mathbf{s}_1) \dots Y(\mathbf{s}_n)\}'$, and the $n \times p$ matrix $\mathbf{X} = (\mathbf{x}(\mathbf{s}_1), \dots, \mathbf{x}(\mathbf{s}_n))'$, where $\{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subset \{\mathbf{u}_1, \dots, \mathbf{u}_m\}$.

There is terminology in the spatial statistical models literature that describes separating the mixed effects model into four components as we do above (for a standard reference, see Cressie and Wikle, 2011, Section 9.2.1): large-scale variability (i.e., $\mathbf{X}\boldsymbol{\beta}$), small-scale variability (i.e., $\widetilde{\boldsymbol{\nu}}$), fine-scale variability (i.e., a term we have marginalized across with variance δ^2), and a measurement error term (i.e., a term we have marginalized across with variance σ_ϵ^2). The variance δ^2 is similar to a "nugget" used in classical spatial statistics (Banerjee et al., 2015). The introduction of δ^2 may be problematic, since the parameters δ^2 and σ_ϵ^2 may be unidentifiable in some settings. However, in the Bayesian setting, the parameters δ^2 and σ_ϵ^2 can be made identifiable through the specification of prior distributions. Additionally, inclusion of an additional random effect has been shown to lead to increased performance in prediction in several settings and removing δ^2 as a source of variability in the latent process $Y(\cdot)$ can lead to over-smoothing (Finley et al., 2009; Bradley et al., 2020).

The introduction of fine-scale variability is often interpreted as a "model correction term." For example, Finley et al. (2009) introduced a "modified" predictive process approach, which modifies the traditional predictive process (Banerjee et al., 2008) by introducing an additional random effect (similar to our inclusion of δ^2) to account for possible over-smoothing of their reduced rank small-scale- variability term. Several spatial statistical models include a fine-scale-variability term while others trust that the small-scale variability term captures all the variability in the latent process and choose to remove the fine-scale variability term. See Bradley et al. (2016) for a review of several recent spatial prediction methods based on this large-scale, small-scale, (possibly) fine-scale, and measurement error variability decomposition.

2.1 The Bayesian Hierarchical Model

In this section, we summarize the statistical model used for inference. The model is organized using the "data model," "process model," and "parameter model" notation used in Cressie and Wikle (2011), as follows:

$$\begin{aligned}
&\textbf{Data Model}: \mathbf{Z} \mid \boldsymbol{\beta}, \boldsymbol{\nu}_O, \widetilde{\boldsymbol{\nu}}_O, \boldsymbol{\theta}, \boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\nu}_O, \sigma_\epsilon^2 \mathbf{I}_n) W(\boldsymbol{\theta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{\nu}}_O, \mathbf{Z})\\
&\textbf{Process Model 1}: \boldsymbol{\nu} \mid \widetilde{\boldsymbol{\nu}}, \delta^2 \sim \mathrm{N}(\widetilde{\boldsymbol{\nu}}, \delta^2 \mathbf{I}_n)\\
&\textbf{Process Model 2}: \widetilde{\boldsymbol{\nu}} \mid \boldsymbol{\theta} \sim p(\widetilde{\boldsymbol{\nu}} \mid \boldsymbol{\theta})\\
&\textbf{Parameter Model 1}: \boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_p)\\
&\textbf{Parameter Model 2}: \boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \sigma_\eta^2 \mathbf{I}_r)\\
&\textbf{Parameter Model 3}: \sigma_\nu^2 \sim \mathrm{IG}(\alpha_1, \beta_1)\\
&\textbf{Parameter Model 4}: \sigma_\beta^2 \sim \mathrm{IG}(\alpha_2, \beta_2)\\
&\textbf{Parameter Model 5}: \sigma_\eta^2 \sim \mathrm{IG}(\alpha_3, \beta_3)\\
&\textbf{Parameter Model 6}: \phi \sim \mathrm{U}(0, \mathrm{U})\\
&\textbf{Parameter Model 7}: \delta^2 \sim \mathrm{IG}(\alpha_4, \beta_4).
\end{aligned} \tag{2}$$

In Equation (2), $\boldsymbol{\nu}_O = \mathbf{O}\boldsymbol{\nu}$, $\widetilde{\boldsymbol{\nu}}_O = \mathbf{O}\widetilde{\boldsymbol{\nu}}$, and $\mathbf{O}$ is an $n \times m$ incidence matrix; $\mathbf{0}_p$ is a p-dimensional vector of zeros; "$\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$" is a shorthand for a multivariate normal distribution with mean $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$; "$\widetilde{\boldsymbol{\nu}} \mid \boldsymbol{\theta} \sim p(\widetilde{\boldsymbol{\nu}} \mid \boldsymbol{\theta})$" should be read as $\widetilde{\boldsymbol{\nu}}$ given $\boldsymbol{\theta}$ is distributed according to the density $p(\widetilde{\boldsymbol{\nu}} \mid \boldsymbol{\theta})$; "IG$(\alpha,\kappa)$" is a shorthand for the inverse gamma distribution with shape $\alpha > 0$ and scale $\kappa > 0$; and "U(L,U)" is a shorthand for a uniform distribution with lower bound L and upper bound U. All hyperparameters are chosen so that the corresponding prior distribution is "flat," and example specifications are provided in Section 4 and Section 5. The term $W(\boldsymbol{\theta},\boldsymbol{\gamma},\widetilde{\boldsymbol{\nu}}_O,\mathbf{Z})$ is a positive-valued function of the $(r+2)$-dimensional parameter vector $\boldsymbol{\theta} = (\sigma^2_\nu, \phi, \boldsymbol{\eta}')'$, the 3-dimensional parameter vector $\boldsymbol{\gamma} = (\sigma^2_\beta, \delta^2, \sigma^2_\epsilon)$, $\widetilde{\boldsymbol{\nu}}_O$, and $\mathbf{Z}$. The explicit definition of W is given in Section 2.3. We refer to our model as an expanded spectral density (ESD) approach. We will assume σ^2_ϵ is known, which is a common assumption in spatial additive models (e.g.,see Bradley et al., 2016, among others). However, one can place a prior on σ^2_ϵ. We suggest placing an informative prior on σ^2_ϵ to avoid potential issues with weak identifiability.

The $n \times m$ incidence matrix $\mathbf{O}$ allows one to identify which of the m locations are observed. That is, to form $\mathbf{O}$, first set $\mathbf{O}$ equal to a $m \times m$ identity matrix and then remove the i-th row if $Z(\mathbf{u}_i)$ is unobserved. This incidence matrix does not require $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ to be defined on a regular grid, which as discussed in the Introduction, is an important contribution, since spectral methods based on the discrete Fourier transform often require regularly spaced locations.

The joint distribution used for statistical inference is found by multiplying all the conditional distributions and marginal distributions implied by (2). To write the joint distribution explicitly, let the densities of the processes and parameters in Equation (2) be denoted with p, and let $p(\boldsymbol{\theta}) = p(\sigma^2_\nu)p(\phi)p(\boldsymbol{\eta})$ and $p(\boldsymbol{\gamma}) = p(\sigma^2_\beta)p(\delta^2)p(\sigma^2_\epsilon)$. Also, let $h(\mathbf{Z}|\boldsymbol{\beta},\boldsymbol{\nu},\sigma^2_\epsilon)$ be the density of a normal distribution with mean $\mathbf{X}\boldsymbol{\beta} + \mathbf{O}\boldsymbol{\nu}$ and covariance matrix $\sigma^2_\epsilon \mathbf{I}_n$. Then the joint distribution of $\mathbf{Z}$, $\boldsymbol{\nu}_O$, $\widetilde{\boldsymbol{\nu}}_O$, $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, and σ^2_η is given by

$$p(\mathbf{Z},\boldsymbol{\nu}_O,\widetilde{\boldsymbol{\nu}}_O,\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},\sigma^2_\eta) = h(\mathbf{Z}|\boldsymbol{\beta},\boldsymbol{\nu}_O,\sigma^2_\epsilon)p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O,\delta^2)p(\boldsymbol{\beta}|\sigma^2_\beta)p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta})p(\boldsymbol{\theta}|\sigma^2_\eta)p(\boldsymbol{\gamma})p(\sigma^2_\eta)W(\boldsymbol{\theta},\boldsymbol{\gamma},\widetilde{\boldsymbol{\nu}}_O,\mathbf{Z}). \tag{3}$$

The standard way to define the Data Model (e.g., see Cressie and Wikle, 2011) sets $W(\boldsymbol{\theta},\boldsymbol{\gamma},\widetilde{\boldsymbol{\nu}}_O,\mathbf{Z}) \equiv 1$. We write the special case of ESD when $W(\boldsymbol{\theta},\boldsymbol{\gamma},\widetilde{\boldsymbol{\nu}}_O,\mathbf{Z}) \equiv 1$ as

$$h(\mathbf{Z},\boldsymbol{\nu}_O,\widetilde{\boldsymbol{\nu}}_O,\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},\sigma^2_\eta) = h(\mathbf{Z}|\boldsymbol{\beta},\boldsymbol{\nu}_O,\sigma^2_\epsilon)p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O,\delta^2)p(\boldsymbol{\beta}|\sigma^2_\beta)p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta})p(\boldsymbol{\theta}|\sigma^2_\eta)p(\boldsymbol{\gamma})p(\sigma^2_\eta). \tag{4}$$

We call the joint distribution in (4) the "un-weighted model." The definition of W and the consequences of setting $W \neq 1$ requires more detail, which we will provide in Section 2.3.

Process Model 1, Parameter Model 1, and Parameter Models 3−5 are fairly standard assumptions for Gaussian data, as they lead to easy to sample full-conditional distributions within a Gibbs sampler (Cressie and Wikle, 2011). Parameter model 6 and 7 are used to avoid identifiability issues and leads to a conjugate full-conditional distribution (see Banerjee et al., 2015, page 124). It is common to assume that $p(\widetilde{\boldsymbol{\nu}} \mid \boldsymbol{\theta})$ in Process Model 2 is the multivariate normal distribution with mean zero and covariance matrix $\mathbf{C}(\boldsymbol{\theta})$ (Banerjee et al., 2015; Cressie and Wikle, 2011). However, in our model $p(\widetilde{\boldsymbol{\nu}} \mid \boldsymbol{\theta})$ is only approximately normal with mean-zero and covariance matrix $\mathbf{C}(\boldsymbol{\theta})$ (see Section 2.2 for details).

There are several different ways in which one could decompose a Bayesian hierarchical model with our parameters and random effects. For example, the Data Model could replace the normal density with $p(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_\epsilon^2)$. However, it will be important for us to isolate $\widetilde{\boldsymbol{\nu}}$ and $\boldsymbol{\nu}$ to lead to easy to sample full-conditional distributions in a collapsed Gibbs sampler (see Section 3).

2.2 Non-Stationary Spectral Simulation using Dimension Expansion

A non-stationary extension of Bochner's Theorem is stated in Theorem 1.

Theorem 1:Let $C(\boldsymbol{s}_i, \boldsymbol{s}_j)$ be a positive definite function on D, which is assumed to be compact. Then there exists a function $\boldsymbol{f}: D \rightarrow \mathbb{R}^d$ and a measure $G_\theta(\boldsymbol{\omega})$ such that for any pair of locations $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$,

$$C(\mathbf{s}_i, \mathbf{s}_j) = \int_{-\infty}^{\infty} \cos[\{\boldsymbol{f}(\mathbf{s}_i) - \boldsymbol{f}(\mathbf{s}_j)\}'\boldsymbol{\omega}_1 + (\mathbf{s}_i - \mathbf{s}_j)'\boldsymbol{\omega}_2] G_\theta(d\boldsymbol{\omega}), \tag{5}$$

where the 2d-dimensional vector $\boldsymbol{\omega} = (\boldsymbol{\omega}_1', \boldsymbol{\omega}_2')'$.

Proof: See Appendix A.

The proof of Theorem 1 involves a simple combination of the result in Perrin and Meiring (2003) and Bochner's Theorem. We use Theorem 1 to define a non-stationary covariance function $C(\cdot,\cdot)$. That is, in our model we choose a specific form for $G_\theta(d\boldsymbol{\omega})$ and $\mathbf{f}(\cdot)$, and we use Equation (5) to define our non-stationary covariance function.

Additionally, in practice we approximate $\mathbf{f}(\mathbf{s})$ by assuming $\boldsymbol{f}(\mathbf{s}) = \boldsymbol{\psi}'(\mathbf{s})\boldsymbol{\eta}$, which is similar to the strategy used in Bornn et al. (2012) and Shand and Li (2017). This leads naturally to questions on how to specify spatial basis functions. In general, we use sparse radial basis functions with equally spaced knot locations as suggested in Nychka (2001) and Cressie and Johannesson (2008). One might also consider the use of information criteria to adaptively select knot locations (Bradley et al., 2011; Tzeng and Huang, 2018).

There are several things we can learn from Theorem 1. First, every non-stationary covariance function can be written as a convolution in $2d$-dimensional space according to (5). Second, $\{\boldsymbol{f}(\mathbf{s}_i) - \boldsymbol{f}(\mathbf{s}_j)\}'\boldsymbol{\omega}_1$ is a deformation, which shows an explicit connection between dimension expansion and deformation. Furthermore, this deformation induces non-stationarity, since $\{\boldsymbol{f}(\mathbf{s}_i) - \boldsymbol{f}(\mathbf{s}_j)\}'\boldsymbol{\omega}_1 = 0$ leads to the classical version of Bochner's Theorem, and hence, this model simultaneously allows for stationarity (i.e., when $\mathbf{f}(\mathbf{s}) \equiv \mathbf{0}_d$ or $\boldsymbol{\eta} = \mathbf{0}_r$) and non-stationarity (i.e., when $\mathbf{f}(\mathbf{s}) \neq \mathbf{0}_d$ or $\boldsymbol{\eta} \neq \mathbf{0}_r$). Third, if we assume a specific form of $G(d\boldsymbol{\omega})$, we can use Equation (5) to approximate the covariance function. For example, when $C(\cdot,\cdot)$ is the exponential covariance function (as is the case in (2)), then $G_\theta(d\boldsymbol{\omega})$ has a corresponding Cauchy density (Stein, 2012). Denote the density corresponding to $G_\theta(d\boldsymbol{\omega})$ with $\frac{g_\theta(\boldsymbol{\omega})}{C(0,0)}$. Moreover, the ability to simulate from the spectral density without mathematical operations of covariance matries, allows us to completely circumvent computing and storing a large covariance matrix (Mejía and Rodríguez-Iturbe, 1974).

Theorem 2: Let $\boldsymbol{\omega}_i = (\boldsymbol{\omega}_{1,i}', \boldsymbol{\omega}_{2,i}')'$, $\boldsymbol{\omega}_i \overset{ind}{\sim} G_\theta(d\boldsymbol{\omega})$, *and* $\kappa_i \overset{ind}{\sim} U(-\pi, \pi)$. *Then for a given* $\boldsymbol{f}\colon \mathbb{R}^d \to \mathbb{R}^d$ *the random process,*

$$\widetilde{\nu}(\mathbf{s}) \equiv \sigma_\nu \left(\frac{2}{K}\right)^{\frac{1}{2}} \sum_{i=1}^{\mathrm{K}} \cos(\boldsymbol{f}(\mathbf{s})'\boldsymbol{\omega}_{1,i} + \mathbf{s}'\boldsymbol{\omega}_{2,i} + \kappa_i), \tag{6}$$

has $E\{\widetilde{\nu}(\boldsymbol{s})\} = 0$, $E\{\widetilde{\nu}(\boldsymbol{s}_i)\widetilde{\nu}(\boldsymbol{s}_j)\} = C(\boldsymbol{s}_i, \boldsymbol{s}_j)$, *and converges in distribution (as* $K \to \infty$*) to a mean-zero Gaussian process with covariance function* $C(\cdot, \cdot)$ *in Equation* (5) *with spectral density* $\prod \frac{g_\theta(\omega_{jk})}{C(0,0)}$, *for* $\boldsymbol{s}$ *contained within a d-dimensional ball in* $D \subset \mathbb{R}^d$.

Proof: See Appendix A.

The proof of Theorem 2 involves a simple combination of the result in Perrin and Meiring (2003) and Mejía and Rodríguez-Iturbe (1974) (also see, Cressie, 1993, pg. 204 for more discussion). In practice, to use Theorem 2, we need to specify the spectral density. In our implementation, we assume that $\omega_{j,i} \overset{ind}{\sim} \frac{g_\theta(\omega)}{C(0,0)}$, where for each i, $\boldsymbol{\omega}_i = (\omega_{1,i}, \ldots, \omega_{2d,i})'$ and $g_\theta(\cdot)$ is the $Cauchy(0, 1/\phi)$ density. This choice of the Cauchy density leads to the exponential covariogram (Cressie, 1993).

It is arguably more common to simulate $\widetilde{\boldsymbol{\nu}}$ using a Cholesky decomposition. However, this requires order n^3 computation and order n^2 memory. Theorem 2 allows us to simulate $\boldsymbol{\nu}$ without these memory and computational problems. It follows from the transformation theorem (Resnick, 2013) that the pdf of $\widetilde{\boldsymbol{\nu}}$ is given, under our specification, by

$$\begin{aligned} &p(\widetilde{\boldsymbol{\nu}} \mid \boldsymbol{\theta}) \\ &= \int_{\widetilde{\boldsymbol{\nu}}:\widetilde{\nu}(\mathbf{s})=\sigma_\nu\left(\frac{2}{K}\right)^{\frac{1}{2}}\sum_{i=1}^{\mathrm{K}}\cos(\boldsymbol{f}(\mathbf{s})'\boldsymbol{\omega}_{1,i}+\mathbf{s}'\boldsymbol{\omega}_{2,i}+\kappa_i)\}} \prod_{jk} \frac{g_\theta(\omega_{j,k})}{C(0,0)} \prod_{i=1}^{K} \frac{1}{2\pi} I(-\pi < \kappa < \pi) d\boldsymbol{\omega}_{1,i} d\boldsymbol{\omega}_{2,i} d\kappa_1 \ldots d\kappa_n, \end{aligned} \tag{7}$$

where $I(\cdot)$ is the indicator function. Again, from Theorem 2 the pdf in (7) is roughly Gaussian with mean zero and covariance $\mathbf{C}(\boldsymbol{\theta})$.

2.3 Re-Weighting Levels in a Spatial Hierarchical Model

We will impose the following conditional independence assumption: $p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta},\mathbf{Z}) = p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta})$. We argue that conditional dependence is not necessarily needed for prediction at observed locations because spatial extrapolation is not needed at observed locations. Furthermore, this added assumption will have additional computational benefits. That is, Gibbs sampling from the posterior distribution would traditionally require one to simulate from the full conditional distribution $p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta},\mathbf{Z})$ and not from $p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta})$, and conditional independence will allow us to use Theorem 2 to jointly update $\widetilde{\boldsymbol{\nu}}_O$. However, this conditional independence assumption changes the likelihood, which for transparency, we now describe explicitly. Define the weight in the Data Model in Equation (2) to be the following

$$W(\boldsymbol{\theta},\boldsymbol{\gamma},\widetilde{\boldsymbol{\nu}}_O,\mathbf{Z}) = \frac{h(\mathbf{Z}_{-w}|\mathbf{Z}_w)h(\mathbf{Z}_w|\boldsymbol{\theta},\boldsymbol{\gamma})}{\int\int h(\mathbf{Z}|\boldsymbol{\beta},\boldsymbol{\nu}_O,\sigma_\epsilon^2)p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O,\delta^2)p(\boldsymbol{\beta}|\sigma_\beta^2)d\boldsymbol{\nu}_O d\boldsymbol{\beta}}, \tag{8}$$

where we subset the data to define $\mathbf{Z}_w$ and $\mathbf{Z}_{-w}$. Specifically, let $\mathbf{O}_w$ be the $w \times n$ subset incidence matrix that is formed by removing the rows of $\mathbf{I}_n$ that do not belong to a pre-defined subset of $\{1,\ldots,n\}$ of size $w \ll n$. Likewise, let $\mathbf{O}_{-w}$ be the $(n-w)\times n$ matrix that is formed by removing the rows of $\mathbf{I}_n$ that belong to the pre-defined subset. Then, the w-dimensional vector $\mathbf{Z}_w = \mathbf{O}_w\mathbf{Z}$ and the $(n-w)$-dimensional vector $\mathbf{Z}_{-w} = \mathbf{O}_{-w}\mathbf{Z}$. Additionally, let $h(\mathbf{Z}_w|\boldsymbol{\theta},\boldsymbol{\gamma})$ be the normal distribution density with mean $\mathbf{0}_w$ and $w \times w$ covariance matrix $\sigma_\beta^2\mathbf{O}_w\mathbf{X}\mathbf{X}'\mathbf{O}_w' + \mathbf{O}_w\mathbf{C}(\boldsymbol{\theta})\mathbf{O}_w' + (\delta^2+\sigma_\epsilon^2)\mathbf{I}_w$, so that $h(\mathbf{Z}_w|\boldsymbol{\theta},\boldsymbol{\gamma})$ is the density for $\mathbf{Z}_w|\boldsymbol{\theta},\boldsymbol{\gamma}$ derived from the un-weighted model. Then define,

$$h(\mathbf{Z}_{-w}|\mathbf{Z}_w) = \frac{\int\int\int\int\int\int h(\mathbf{Z},\boldsymbol{\nu}_O,\widetilde{\boldsymbol{\nu}}_O,\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},\sigma_\eta^2)\, d\boldsymbol{\nu}_O\, d\boldsymbol{\beta}\, d\widetilde{\boldsymbol{\nu}}_O\, d\sigma_\eta^2\, d\boldsymbol{\theta}\, d\boldsymbol{\gamma}}{\int\int\int h(\mathbf{Z}_w|\boldsymbol{\theta},\boldsymbol{\gamma})p(\boldsymbol{\theta}|\sigma_\eta^2)p(\boldsymbol{\gamma})p(\sigma_\eta^2)\, d\boldsymbol{\theta}\, d\boldsymbol{\gamma}\, d\sigma_\eta^2},$$

where recall $h(\mathbf{Z},\boldsymbol{\nu}_O,\widetilde{\boldsymbol{\nu}}_O,\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},\sigma_\eta^2)$ is defined in (4). We call the joint distribution in (3) with W specified with (8) the "weighted ESD model."

There are several important properties of the weighted ESD model. First, the marginal distribution of the data $\mathbf{Z}$ is the *same* for both the weighted ESD model and the un-weighted model.

That is, we have

$$\begin{aligned} p(\mathbf{Z}) &= \int \ldots \int p(\mathbf{Z}, \boldsymbol{\nu}_O, \widetilde{\boldsymbol{\nu}}_O, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_\eta^2)\, d\boldsymbol{\nu}_O\, d\boldsymbol{\beta}\, d\widetilde{\boldsymbol{\nu}}_O\, d\sigma_\eta^2\, d\boldsymbol{\theta}\, d\boldsymbol{\gamma} \\ &= \int \ldots \int h(\mathbf{Z}, \boldsymbol{\nu}_O, \widetilde{\boldsymbol{\nu}}_O, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_\eta^2)\, d\boldsymbol{\nu}_O\, d\boldsymbol{\beta}\, d\widetilde{\boldsymbol{\nu}}_O\, d\sigma_\eta^2\, d\boldsymbol{\theta}\, d\boldsymbol{\gamma} = h(\mathbf{Z}), \end{aligned} \tag{9}$$

where the algebraic details are provided in Appendix A. Since the weighted ESD model and un-weighted model have the same marginal distribution of the data, both models imply the same assumptions on the data (marginally). Furthermore, Equation (9) implies that the weighted ESD model is proper (i.e., integrates to one), since the un-weighted model is proper.

Another important property of the weighted ESD model is that the following relationships hold,

$$p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta}, \mathbf{Z}) = p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta}) \tag{10}$$

$$p(\boldsymbol{\theta}|\widetilde{\boldsymbol{\nu}}_w, \boldsymbol{\gamma}, \sigma_\eta^2, \mathbf{Z}) \propto h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma}) p(\widetilde{\boldsymbol{\nu}}_w|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\sigma_\eta^2) \tag{11}$$

$$p(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\nu}}_O, \boldsymbol{\theta}, \sigma_\eta^2, \mathbf{Z}) \propto h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma}), \tag{12}$$

where $\widetilde{\boldsymbol{\nu}}_w = \mathbf{O}_w\mathbf{O}\widetilde{\boldsymbol{\nu}}$ so that the conditional distribution in (11) collapses across $\boldsymbol{\beta}$, $\boldsymbol{\nu}$, and all $\widetilde{\nu}(\mathbf{u})$ that are not elements of $\widetilde{\boldsymbol{\nu}}_w$, and where $p(\widetilde{\boldsymbol{\nu}}_w|\boldsymbol{\theta})$ is approximately a multivariate normal distribution with mean zero and $w \times w$ covariance matrix $\mathbf{O}_w\mathbf{O}\mathbf{C}(\boldsymbol{\theta})\mathbf{O}'\mathbf{O}_w'$. Similarly, the conditional distribution in (12) collapses across $\boldsymbol{\nu}_O$ and $\boldsymbol{\beta}$.

Equation (10) $-$ (12) show that several types of conditional independence arises in the weighted ESD model. Specifically, $\widetilde{\boldsymbol{\nu}}_O$ is conditionally independent of $\mathbf{Z}$ given $\boldsymbol{\theta}$. This is particularly useful for our purposes, since this allows one to use non-stationary spectral simulation, via Theorem 2, to update $\widetilde{\boldsymbol{\nu}}_O$ in a collapsed Gibbs sampler. Additionally, one can use (11) and (12) to update $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ efficiently in a collapsed Gibbs sampler as well, since the distributions in (11) and (12) can be computed efficiently using a subset of the data. This is because $\mathbf{Z}_w$ is a low-dimensional subset of the high-dimensional dataset. Equation (11) shows that the weighted ESD model assumes that $\boldsymbol{\theta}$ is conditionally independent of $\mathbf{Z}_{-w}$ given $\widetilde{\boldsymbol{\nu}}_w$, $\boldsymbol{\gamma}$, σ_η^2, and $\mathbf{Z}_w$. Likewise, Equation (12) shows that the weighted ESD model assumes that $\boldsymbol{\gamma}$ is conditionally independent of $\mathbf{Z}_{-w}$ given $\widetilde{\boldsymbol{\nu}}_O$, $\boldsymbol{\gamma}$, σ_η^2,

Algorithm 1 Implementation: Collapsed Gibbs sampler

```
1: Initialize $\boldsymbol{\beta}^{[1]}$, $\boldsymbol{\nu}_O^{[1]}$,$\widetilde{\boldsymbol{\nu}}_O^{[1]}$ $\boldsymbol{\eta}^{[1]}$,$\sigma_\nu^{2[1]}$,$\sigma_\beta^{2[1]}$, $\sigma_\eta^{2[1]}$, $\phi^{[1]}$ $\delta^{2[1]}$, $\boldsymbol{\theta}^{[1]} = (\sigma_\nu^{2[1]}, \phi^{[1]}, \boldsymbol{\eta}^{[1]})'$, and $\boldsymbol{\gamma}^{[1]} = (\sigma_\beta^{2[1]}, \delta^{2[1]}, \sigma_\epsilon^{2[1]})$.
2: Set $b = 2$.
3: Simulate $\boldsymbol{\beta}^{[b]}$ from $p(\boldsymbol{\beta}|\boldsymbol{\nu}_O^{[b-1]}, \boldsymbol{\eta}^{[b-1]}, \sigma_\nu^{2[b-1]}, \sigma_\beta^{2[b-1]}, \sigma_\eta^{2[b-1]}, \phi^{[b-1]}, \delta^{2[b-1]}, \mathbf{Z})$.
4: Simulate $\widetilde{\boldsymbol{\nu}}_O^{[b]}$ from $p(\widetilde{\boldsymbol{\nu}}_O \mid \boldsymbol{\theta}^{[b-1]})$ using Theorem 2 with $K$ "large," where recall from (10), $p(\widetilde{\boldsymbol{\nu}}_O \mid \boldsymbol{\theta}) = p(\widetilde{\boldsymbol{\nu}}_O \mid \boldsymbol{\theta}, \mathbf{Z})$.
5: Simulate $\boldsymbol{\nu}_O^{[b]}$ from $p(\boldsymbol{\nu}_O|\boldsymbol{\beta}^{[b]}, \boldsymbol{\eta}^{[b-1]}, \sigma_\nu^{2[b-1]}, \sigma_\beta^{2[b-1]}, \sigma_\eta^{2[b-1]}, \phi^{[b-1]}, \delta^{2[b-1]}, \widetilde{\boldsymbol{\nu}}_O^{[b]}, \mathbf{Z})$.
6: Simulate $\boldsymbol{\theta}^{[b]}$ from $p(\boldsymbol{\theta}|\mathbf{O}_w\widetilde{\boldsymbol{\nu}}_O^{[b]}, \boldsymbol{\gamma}^{[b-1]}, \sigma_\eta^{2[b-1]}, \mathbf{Z})$ and set $(\sigma_\nu^{2[b]}, \phi^{[b]}, \boldsymbol{\eta}^{[b]}) = \boldsymbol{\theta}^{[b]}$, where recall from (11), $p(\boldsymbol{\theta}|\mathbf{O}_w\widetilde{\boldsymbol{\nu}}_O^{[b]}, \boldsymbol{\gamma}^{[b-1]}, \sigma_\eta^{2[b-1]}, \mathbf{Z}) \propto h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma}^{[b-1]})p(\mathbf{O}_w\widetilde{\boldsymbol{\nu}}_O^{[b]}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\sigma_\eta^2)$.
7: Simulate $\boldsymbol{\gamma}^{[b]}$ from $p(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\nu}}_O^{[b]}, \boldsymbol{\theta}^{[b]}, \sigma_\eta^{2[b]}, \mathbf{Z})$ and set $(\sigma_\beta^{2[b]}, \delta^{2[b]}, \sigma_\epsilon^{2[b]}) = \boldsymbol{\gamma}^{[b]}$, where recall from (12), $p(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\nu}}_O^{[b]}, \boldsymbol{\theta}^{[b]}, \sigma_\eta^{2[b]}, \mathbf{Z}) \propto h(\mathbf{Z}_w|\boldsymbol{\theta}^{[b]}, \boldsymbol{\gamma})p(\boldsymbol{\gamma}^{[b]})$.
8: Let $b = b + 1$.
9: If $b < B$ (a prespecified value) repeat Steps $3 - 12$, otherwise stop.
```

and $\mathbf{Z}_w$. In practice, we specify the subset of our dataset to be w roughly equally spaced over the observed locations. The value of w can be selected using cross-validation.

These properties aid in the interpretation of W in the weighted ESD model. That is, W in the weighted ESD model is the quantity that enforces the conditional independence assumptions implied by (10) − (12), while simultaneously ensuring that our (marginal) assumptions on the data do not change (i.e., see (9)).

3 Computation and Prediction

3.1 Computation: Collapsed Gibbs Sampling

In this section, we outline the steps needed for collapsed Gibbs sampling. Gibbs sampling requires simulating from full-conditional distributions (Gelfand and Smith, 1990). In a collapsed Gibbs sampler, some of the events conditioned on in the full-conditional distribution are integrated out (Liu, 1994). In Algorithm 1, we present the steps needed for our proposed collapsed Gibbs sampler. The expressions for the full-conditional distributions listed in Algorithm 1 are derived in Appendix B. This collapsed Gibbs sampler can easily be modified to allow for heterogeneous variances, component-wise updating, and allow for other choices of proper prior distributions on $\boldsymbol{\theta}$

and $\boldsymbol{\gamma}$.

An important motivation for collapsed Gibbs sampling is that Step 4 of Algorithm 1 is computationally straightforward using non-stationary spectral simulation. Additionally, in Step 5, the full-conditional distribution has a known, and easy to sample from expression. This is significant, as this full-conditional distribution traditionally involves inverses and determinants of high-dimensional matrices. Specifically, the following relationship holds,

$$f(\boldsymbol{\nu}_O|\cdot) \propto \exp\left\{-\frac{(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\nu}_O)'\mathbf{V}_\epsilon^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\nu}_O)}{2}\right\} f(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O,\delta,\boldsymbol{\theta}),$$

where $\mathbf{V}_\epsilon = \sigma_\epsilon^2$ and $f(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O,\delta,\boldsymbol{\theta}) \propto \exp\left\{-\frac{(\boldsymbol{\nu}_O-\widetilde{\boldsymbol{\nu}}_O)'(\boldsymbol{\nu}_O-\widetilde{\boldsymbol{\nu}}_O)}{2\delta^2}\right\}$. Then,

$$f(\boldsymbol{\nu}_O|\cdot) \propto \exp\left\{-\frac{\boldsymbol{\nu}_O'(\delta^{-2}\mathbf{I}+\mathbf{V}_\epsilon^{-1})\boldsymbol{\nu}_O}{2}+\boldsymbol{\nu}_O'\left(\mathbf{V}_\epsilon^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})+\frac{1}{\delta^2}\widetilde{\boldsymbol{\nu}}_O\right)\right\}.$$

This gives

$$f(\boldsymbol{\nu}|\cdot) = \mathrm{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*), \tag{13}$$

where $\boldsymbol{\mu}^* = \boldsymbol{\Sigma}^*\{\mathbf{V}_\epsilon^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})+\frac{1}{\delta^2}\widetilde{\boldsymbol{\nu}}_O\}$, and $(\boldsymbol{\Sigma}^*)^{-1} = \delta^{-2}\mathbf{I}_n+\mathbf{V}_\epsilon^{-1}$, where we emphasize that $\boldsymbol{\Sigma}^*$ is a computationally advantageous diagonal matrix. We have found that Algorithm 1 leads to fast mixing, as several of the covariance structures are diagonal (i.e., $\boldsymbol{\nu}_O$ and $\boldsymbol{\eta}$ have diagonal covariances), and there is no Metropolis-Hasting steps required to update the random effect $\widetilde{\boldsymbol{\nu}}_O$, which has a more complex covariance structure.

Algorithm 1 provides a perspective on the relationship between $\boldsymbol{\nu}_O$, $\widetilde{\boldsymbol{\nu}}_O$, and δ^2 under our conditional independence assumption $p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta},\mathbf{Z}) = p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta})$. That is, the mean of $\boldsymbol{\nu}_O$ is updated while the mean of $\tilde{\boldsymbol{\nu}}_O$ is zero, which suggests that the role of $\tilde{\boldsymbol{\nu}}_O$ is purely to induce (possibly) non-stationary spatial correlations into $\boldsymbol{\nu}_O$. This also emphasizes our need for $\delta^2 > 0$, since its presence allows for $\boldsymbol{\nu}_O \neq \widetilde{\boldsymbol{\nu}}_O$ (almost surely), and hence, have non-zero mean. Algorithm 1 also provides a perspective on updating $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. In particular, the updates in Steps 6 and 7 are the same as the updates for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ in a Gibbs sampler for the un-weighted model that (1) does not use non-stationary spectral simulation and (2) marginalizes $\boldsymbol{\beta}$, $\boldsymbol{\nu}_O$, $\widetilde{\boldsymbol{\nu}}_O$, and $\mathbf{Z}_{-w}$.

There are parameters in our model that are not explicitly modeled with a prior distribution that require specification. In particular, the number of basis functions r, the size of the equally spaced subset w, the number of components to average in spectral simulation K, and the spectral density. In general, for both simulations and the application, we have found that values of r that are too small lead to larger prediction errors, and vice versa. Additionally, we have found that the method is robust to small values of K, but we generally found $K = 100$ leads to small prediction errors and is computationally feasible. In practice, we use cross-validation to choose the value of r, w, and K. One could also consider a more general model not considered in this paper, where prior distributions are placed on r, w, and K. This choice will possibly lead to computationally prohibitive reversible jumps in the MCMC. When the latent spatial process is non-stationary, we have found that the prediction errors are robust to the specification of the spectral density. However, in practice one might consider other spectral densities besides the Cauchy distribution.

3.2 Spatial Prediction

Spatial prediction at observed locations follows from Algorithm 1, where the b-th posterior replicate of $Y(\mathbf{s})$ is computed as

$$Y(\mathbf{s})^{[b]} = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}^{[b]} + \mathbf{e}(\mathbf{s})'\boldsymbol{\nu}_O^{[b]},$$

where $\mathbf{s} \in \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is observed, $\mathbf{e}(\mathbf{s}) = (I(\mathbf{s} = \mathbf{s}_1), \ldots, I(\mathbf{s} = \mathbf{s}_n))'$, and $I(\cdot)$ is an indicator function. Then posterior means and variances of $Y(\mathbf{s})^{[b]}$ across b can be used to perform inference on $Y(\mathbf{s})$.

To predict at missing location $\mathbf{u} \in \{\mathbf{u}_1, \ldots, \mathbf{u}_m\} \cap \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}^c$ we first define an $w^* \times n$ incidence matrix $\mathbf{O}_w^*$, where each row indicates the observed location that is a w^* nearest neighbor of $\mathbf{u}$, where w^* is not necessarily the same as w. Set the w^*-dimensional vector $\boldsymbol{\nu}_w^* = \mathbf{O}_w^* \boldsymbol{\nu}_O$, and denote the $(n - w^*)$-dimensional vector $\boldsymbol{\nu}_{-w}^*$ to consist of all $\nu(\mathbf{s})$ such that $\mathbf{s}$ is *not* a w^* nearest neighbor of $\mathbf{u}$. Then it follows from standard result for multivariate normal distributions (Ravishanker and Dey, 2020),

$$\nu(\mathbf{u})|\boldsymbol{\nu}_w^*, \boldsymbol{\theta} \sim N\left\{\mathbf{C}_O(\mathbf{u}|\boldsymbol{\theta})\mathbf{C}_O(\boldsymbol{\theta})^{-1}\boldsymbol{\nu}_w^*, var(\nu(\mathbf{u})|\boldsymbol{\theta}) - \mathbf{C}_O(\mathbf{u}|\boldsymbol{\theta})\mathbf{C}_O(\boldsymbol{\theta})^{-1}\mathbf{C}_O(\mathbf{u}|\boldsymbol{\theta})'\right\}, \tag{14}$$

where the w-dimensional vector $\mathbf{C}_O(\mathbf{u}|\boldsymbol{\theta}) = cov(\nu(\mathbf{u}), \boldsymbol{\nu}_w^*|\boldsymbol{\theta})$, the $w^* \times w^*$ matrix $\mathbf{C}_O(\boldsymbol{\theta}) = cov(\boldsymbol{\nu}_w^*|\boldsymbol{\theta})$, and we have implicitly assumed K is large enough so that $\widetilde{\boldsymbol{\nu}}$ is normally distributed. To predict at missing locations we use the posterior predictive distribution of a "new" replicate of $\nu(\mathbf{u})$, which we denote with $\nu(\mathbf{u})^{new}$. Here, $p(\nu(\mathbf{u})^{new}|\boldsymbol{\nu}_w^*, \boldsymbol{\theta})$ is assumed to be the same as (14). Furthermore, we assume the following,

$$\begin{aligned} p(\nu(\mathbf{u})^{new}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{Z}) &= \int\int p(\nu(\mathbf{u})^{new}, \boldsymbol{\nu}_O|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{Z})d\boldsymbol{\nu}_{-w}^* d\boldsymbol{\nu}_w^* \\ &= \int\int p(\boldsymbol{\nu}_{-w}^*, \nu(\mathbf{u})^{new}|\boldsymbol{\nu}_w^*, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{Z})p(\boldsymbol{\nu}_w^*|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{Z})d\boldsymbol{\nu}_{-w}^* d\boldsymbol{\nu}_w^* \\ &= \int p(\nu(\mathbf{u})^{new}|\boldsymbol{\nu}_w^*, \boldsymbol{\theta})p(\boldsymbol{\nu}_w^*|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{Z})d\boldsymbol{\nu}_w^*. \end{aligned} \tag{15}$$

The posterior predictive replicate $\nu(\mathbf{u})^{new}$ is generated under the assumption of conditional independence of $\mathbf{Z}$ (and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) given $\boldsymbol{\nu}_w$. This distinction is important because $\nu(\mathbf{u})$ itself is conditionally independent of $\mathbf{Z}$ (and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) given $\boldsymbol{\nu}_O$ *and not* $\boldsymbol{\nu}_w^*$. When $w^* = n$ we have $\mathbf{O}_w^*$ is equal to the identity matrix and $p(\nu(\mathbf{u})^{new}|\boldsymbol{\nu}_w^*, \boldsymbol{\theta}) = p(\nu(\mathbf{u})^{new}|\boldsymbol{\nu}_O, \boldsymbol{\theta}) = p(\nu(\mathbf{u})^{new}|\boldsymbol{\nu}_O, \boldsymbol{\theta}, \mathbf{Z})$. When setting $w^* < n$ we are implicitly assuming that the dependence between $\nu(\mathbf{u})$ and $\nu(\mathbf{s})$ is negligible for $\mathbf{s}$ that are not w^* nearest neighbors of $\mathbf{u}$ so that $p(\nu(\mathbf{u})^{new}|\boldsymbol{\nu}_w^*, \boldsymbol{\theta}) \approx p(\nu(\mathbf{u})^{new}|\boldsymbol{\nu}_O, \boldsymbol{\theta})$. Consequently, $\nu(\cdot)^{new}$ should be interpreted as a different process that approximates $\nu(\cdot)$. This added assumption/simplification is incorporated into our predictions mainly for computational purposes (see discussion at the end of this section). Similar simplifications have been used before in different contexts (e.g., local kriging, Pronzato and Rendas, 2017).

Notice that w^* is possibly different from the w neighbors used in the likelihood re-weighting technique used in Section 2.3. Ultimately, Equation (15) can be seen as a type of collapsed Gibbs sampler, where we collapse across $\boldsymbol{\nu}_{-w}^*$. From Equation (13),

$$\boldsymbol{\nu}_w^*|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{Z} \sim N\left\{\mathbf{O}_w^*\boldsymbol{\Sigma}^*\mathbf{V}_\epsilon^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}), \left(\frac{1}{\delta^2}\right)^2 \mathbf{O}_w^*\boldsymbol{\Sigma}^*\mathbf{C}(\boldsymbol{\theta})\boldsymbol{\Sigma}^{*\prime}\mathbf{O}_w^{*\prime} + \mathbf{O}_w^*\boldsymbol{\Sigma}^*\mathbf{O}_w^{*\prime}\right\}. \tag{16}$$

From (15), (16), and (14) we have,

$$\nu(\mathbf{u})^{new}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\theta},\mathbf{Z} \sim N\left\{\mathbf{C}_O(\mathbf{u}|\boldsymbol{\theta})\mathbf{C}_O(\boldsymbol{\theta})^{-1}\mathbf{O}_w^{*}\boldsymbol{\Sigma}^{*}\mathbf{V}_{\epsilon}^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}),\sigma^{*2}\right\}, \tag{17}$$

where $\sigma^{*2} = \mathbf{C}_O(\mathbf{u}|\boldsymbol{\theta})\mathbf{C}_O(\boldsymbol{\theta})^{-1}\mathbf{M}(\boldsymbol{\gamma},\boldsymbol{\theta})\mathbf{C}_O(\boldsymbol{\theta})^{-1}\mathbf{C}_O(\mathbf{u}|\boldsymbol{\theta})'+var(\nu(\mathbf{u})|\boldsymbol{\theta})-\mathbf{C}_O(\mathbf{u}|\boldsymbol{\theta})\mathbf{C}_O(\boldsymbol{\theta})^{-1}\mathbf{C}_O(\mathbf{u}|\boldsymbol{\theta})'$ and $\mathbf{M}(\boldsymbol{\gamma},\boldsymbol{\theta}) = \left(\frac{1}{\delta^2}\right)^2\mathbf{O}_w^{*}\boldsymbol{\Sigma}^{*}\mathbf{C}(\boldsymbol{\theta})\boldsymbol{\Sigma}^{*\prime}\mathbf{O}_w^{*\prime}+\mathbf{O}_w^{*}\boldsymbol{\Sigma}^{*}\mathbf{O}_w^{*\prime} = \left(\frac{1}{\delta^2}\right)^2\left(\frac{1}{\frac{1}{\delta^2}+\frac{1}{\sigma_\epsilon^2}}\right)^2\mathbf{C}_O(\boldsymbol{\theta})+\left(\frac{1}{\frac{1}{\delta^2}+\frac{1}{\sigma_\epsilon^2}}\right)\mathbf{I}_{w^*}$. Finally predictions at missing locations are computed as averages across b (after a burn-in) of,

$$Y(\mathbf{u})^{[b]} = \mathbf{x}(\mathbf{u})'\boldsymbol{\beta}^{[b]}+\nu(\mathbf{u})^{new[b]}, \tag{18}$$

where $\nu(\mathbf{u})^{new[b]}$ is the b-th replicate from (17). To sample from (17) posterior replicates of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are required, and these are obtained using the collapsed Gibbs sampler.

One could choose $w^* = n$ so that $\mathbf{O}_w^{*}$ is equal to the identity matrix and $p(\nu(\mathbf{u})^{new}|\boldsymbol{\nu}_w^{*},\boldsymbol{\theta}) = p(\nu(\mathbf{u})^{new}|\boldsymbol{\nu}_O,\boldsymbol{\theta}) = p(\nu(\mathbf{u})^{new}|\boldsymbol{\nu}_O,\boldsymbol{\theta},\mathbf{Z})$. However, choosing $w^* \ll n$ is needed to produce a $w^* \times w^*$ matrix $\mathbf{C}_O(\boldsymbol{\theta})^{-1}$ in (17) that is straightforward to compute in practice. We also update $\nu(\mathbf{u})^{new}$ pointwise, which limits our inference to prediction and estimation of prediction variances. Of course joint updates of $\nu(\cdot)$ at missing locations can be done in a similar manner, but this leads to storage issues of the cross-covariance matrix of $\{\nu(\mathbf{u})\}$ and $\boldsymbol{\nu}_w^{*}$.

4 Simulation Studies

We simulate data in a variety of settings, and compare to several state-of-the-art methods in spatial statistics including the nearest neighbor Gaussian process (NNGP) model (Datta et al., 2016b), the general Vecchia approximation (Katzfuss and Guinness, 2019; Katzfuss et al., 2020), spatial process convolution (SPC; Paciorek and Schervish, 2006), and a stochastic partial differential equation (SPDE; Lindgren et al., 2011) approach. We refer to the general Vecchia approximation to a Gaussian model with stationary covariances as an "approximate stationary model" instead of a "stationary model." The expression of the covariance function from the general Vecchia approximation can be found in Proposition 1 in Katzfuss and Guinness (2019). It is immediate from this expression that the covariance function is *not* a function of the spatial lag only, but also depends on an indexing set, and consequently, is non-stationary by definition. Their use of this indexing set is extremely important as it leads to sparsity that produces massive computational gains. See Katzfuss and Guinness (2019) for more details.

We will consider two different simulation setups, where stationarity and non-stationarity arises in different ways. In Simulation Model 1, we consider spatial data, where the large-scale variability is nonstationary and the small-scale spatial variability is stationary. In Simulation Model 2, we consider generating non-stationary data directly from existing non-stationary spatial models (i.e., SPC and SPDE). In Sections (4.1) and (4.2), we give the details surrounding Simulation Model 1 and the results of this analysis, respectively. Likewise, in Sections (4.3) and (4.4), we give the details surrounding Simulation Model 2 and the results of this analysis, respectively. In all simulation studies we treat the measurement error variance as known.

4.1 Simulation Model 1 Set-Up

We assume $\mathbf{Z} = \mathbf{Y} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is a fixed and known n-dimensional vector and $\boldsymbol{\epsilon} \sim N(\mathbf{0}_n, \sigma_\epsilon^2 \mathbf{I}_n)$. We choose σ_ϵ^2 based on the signal-noise-ratio (SNR); in particular, we choose SNR equal to 3. We also present results with 5% of the data missing at random.

We simulate $\mathbf{Y}$ to be a mixture of a one-dimensional SPC Paciorek and Schervish (2006) with zero mean. Specifically, let $f_0(i)$ be a SPC generated using functions made available by the R-package `convoSPAT` (e.g., see Risser and Calder, 2015). We use five mixture components with the default specifications of the functions "`f_mc_kernels`" and "`NSconvo_sim`" to generate $f_0(i)$. More details on SPC are provided in Appendix C. We add a subscript "0" on f to distinguish f_0 from $\mathbf{f}$.

Then we propose five simulation cases,

$$
\begin{aligned}
&\text{Case 1:} \quad Y(i) = f_0(i)\\
&\text{Case 2:} \quad Y(i) = 0.85f_0(i) + 0.15\zeta(i)\\
&\text{Case 3:} \quad Y(i) = 0.5f_0(i) + 0.5\zeta(i)\\
&\text{Case 4:} \quad Y(i) = 0.15f_0(i) + 0.85\zeta(i)\\
&\text{Case 5:} \quad Y(i) = \zeta(i),
\end{aligned}
$$

and $\boldsymbol{\zeta} = (\zeta(1), \ldots, \zeta(m))' \sim \mathrm{N}(\mathbf{0}_m, \mathbf{R})$, where the $m \times m$ matrix $\mathbf{R} = \{\sigma_\zeta^2 \exp(-\phi_\zeta \|i - j\|)\}$. Thus, $\boldsymbol{\zeta}$ is a stationary term. In Case 1, the data is generated from a highly non-linear process and $\mathbf{x}_i$ is a five dimensional vector consisting of independent draws from a uniform distribution. In Case 3, we weight the process half with a nonlinear term and half with a stationary term. In Case 5, we only have a stationary term. So the data is rough in Case 1 and gradually becomes smoother as we consider other cases. We show examples of the data in Figure (1).

In Figure 1, Case 1 shows no spatial structure, yet the SPC model implies spatial cross-correlations. The smallest spatial cross-correlation is roughly 0.12 (computed empirically from 100 independently generated spatial fields from the SPC model in Case 1). Furthermore, even when spatial structure is present in a plot of the data, the latent process may or may not be stationary (e.g., possible functional patterns in the large-scale term, patterns of local stationarity may be present, etc.). The difficulty in using visualizations of the data to assess non-stationarity motivates the need to test for non-stationarity. For the ESD model this amounts to testing whether or not the elements of $\boldsymbol{\eta}$ are equal to zero.

We generate 1,000 observations over this one-dimensional domain $[0, 1]$ for each case. For Case 2 to Case 4, σ_ζ^2 is set to be equal to the sample variance of the elements in the vector $(f_0(\mathbf{x}_1), \ldots, f_0(\mathbf{x}_n))'$. We fixed $\phi_\zeta = 0.3$. SNR is defined to be

$$
\mathrm{SNR} = \frac{\sum_{i=1}^{n}(Y(i) - \frac{1}{n}\sum_{j=1}^{n} Y(j))^2}{(n-1)\sigma_\epsilon^2}
$$

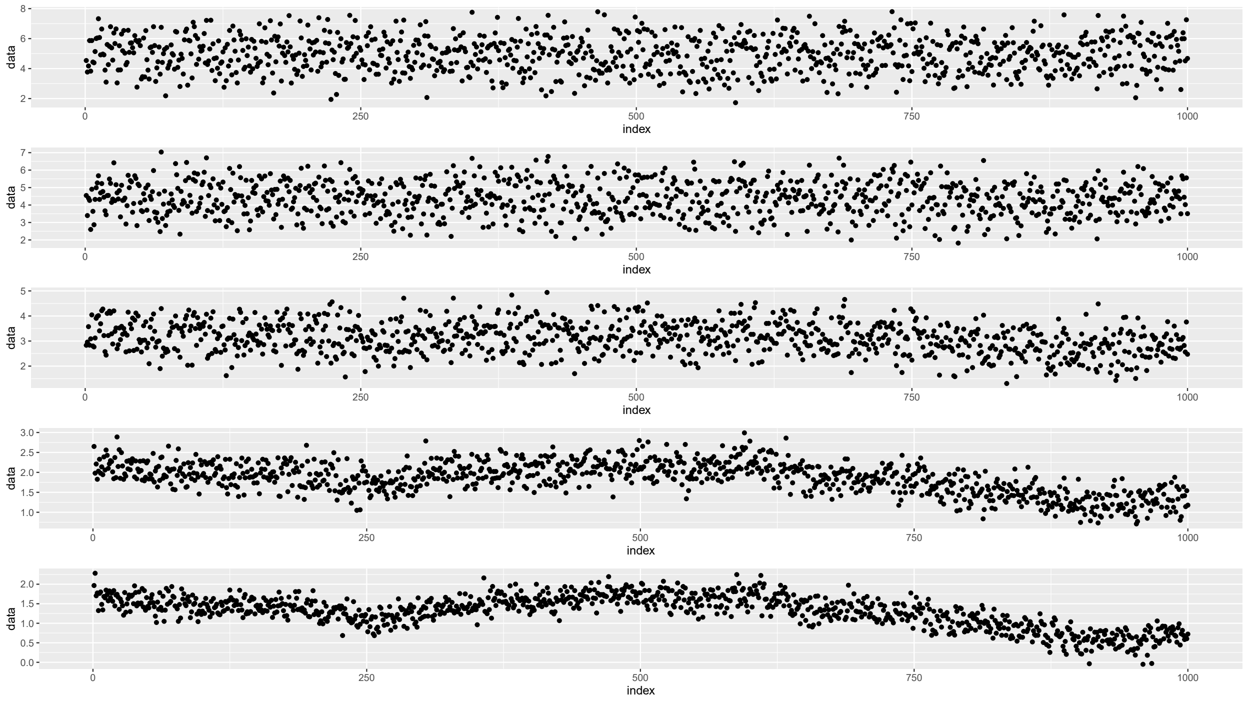

Figure 1: Simulation Model 1 data with SNR=3; First row to last row are examples of Case 1 to Case 5.

so that

$$\sigma_\epsilon^2 = \frac{\sum_{i=1}^{n}(Y(i) - \frac{1}{n}\sum_{j=1}^{n} Y(j))^2}{(n-1)\mathrm{SNR}}.$$

To implement ESD, we specify a 20-dimensional vector $\boldsymbol{\psi}(i)$ to consist of bisquare radial basis functions over equally spaced knots. That is, $\boldsymbol{\psi}(i) = (\psi_1(i) \ldots \psi_{20}(i))'$, where

$$\phi_k(i) = \left\{1 - \frac{|i - c_k|}{\tau}\right\}^2 I(|i - c_k| < \tau);\ k = 1, \ldots, 20,$$

$I(\cdot)$ is an indicator function, $\{c_1, \ldots, c_{20}\}$ are the equally-spaced knots points over $\{1, \ldots, 1,000\}$, and τ is equal to 1.5 times the median of non-zero distances between the points in $\{1, \ldots, 1,000\}$. These basis functions are used to model $\mathbf{f}$ in ESD, and are also used as covariates in all three models under comparison.

When implementing ESD, NNGP, and the Vecchia method we use the same covariates and 20 bisquare basis functions. Thus, ESD, NNGP, and Vecchia model the nonstationary function ρf_0 in the same way (for $\rho = 1, 0.85, 0.5, 0.15$, and 0 for Cases $1 - 5$), and hence, all methods are, in a sense, using basis functions/important covariates to model non-stationarity. The process $(1-\rho)\zeta(\cdot)$, used to generate data, is simulated according to a Gaussian process with mean zero and an exponential covariogram. Thus, to fairly compare each method we specify the covariance functions of ESD, Vecchia, and NNGP to best model the stationary process $(1-\rho)\zeta(\cdot)$. That is, NNGP and the Vecchia approximation are specified with an exponential covariogram, and the ESD is based on $Cauchy(0, 1/\phi)$ (corresponding to the exponential covariogram), where recall ESD allows for the stationary case when $\boldsymbol{\eta} = \mathbf{0}_r$.

Each of the models that are compared in this simulation study are mis-specified. As such, there is no guarantee that any method will perform better than others. This is arguably more realistic, as the possibility of a mis-specified model is always present in real applications. In Simulation Study 2, we consider the case where there is a gold standard.

To implement NNGP, we use 15 nearest neighbors which is consistent with what is suggested in Datta et al. (2016a). For the general Vecchia approximation, we use 15 nearest neighbors which is the same as NNGP. We also use the R-package `spNNGP` (Finley et al., 2020) and `GpGp` (Guinness and Katzfuss, 2018). The weighted ESD is implemented with $1,000$ replicates, with a burn-in of 500. ESD uses the aforementioned covariates to model $\boldsymbol{\beta}$ and the 20 bisquare functions to model $\mathbf{f}$. We also set $w = w^* = 50$ and $K = 100$. Convergence was assessed visually with trace plots and no lack of convergence was detected.

In general, we have found that ESD is robust to the choice of the spectral density sensitivity. The tuning parameters w $(=w^*)$ and K are also fairly robust to the choice of spectral density provided that the rank r is chosen to be "large enough." We compared the posterior mean of $\mathbf{C}(\boldsymbol{\theta})$ when fitting different ESD models each based on different specifications of the spectral density, and when simulating from an ESD with a Cauchy spherical density. In particular, we considered fitting ESD with the Cauchy density, triangular density, and the Gaussian density. On average when $r = 20$ the elements of these matrices differed on the order of 10^{-7} and saw no appreciable difference in the predictions, but note for small values of r (in this case less than 20) we saw very

Table 1: RMSPE (over both observed and missing locations) for the data in the first row of Figure (1). We call our method the Expanded Spectral Density (ESD) method.

Method	**RMSPE**
ESD	0.607
NNGP	1.015
Vecchia Approximation	1.013

poor estimates of the covariance matrix.

4.2 Results for Simulation Model 1

In Table (1), we provide the root mean squared prediction error (RMSPE) of each model for the data in the first row in Figure (1). The RMSPE is defined as

$$\sqrt{\frac{1}{A}\left\{\sum_{i\in A}\{Y(i)-\hat{Y}(i)\}^2\right\}}, \tag{19}$$

where $\hat{Y}(i)$ is the posterior mean from fitting the each model and $A = \{1, \ldots, 1000\}$. Here we see that ESD outperforms NNGP and the general Vecchia approximation. There are $m = 1,000$ prediction locations, and thus, for 5% missing at random we observe $n = 950$, and the results in Table 1 are aggregated across both observed and missing locations.

To assess each method over multiple replicates we record the performance (in terms of RMSPE) over the SNR and each of the five cases. The results are shown in Figure (2) for Case 1 to Case 5, and for A defined to be either the set of randomly selected observed locations or the set of missing locations, respectively. The results were fairly consistent across SNR, and we only show the results for SNR = 3. The first row contains the RMSPE computed over missing locations and in the second row the RMSPE is computed over observed locations. The first column to the last column in Figure (2) displays results for Case 1 to Case 5, respectively. We assume 5% missing data in Figure (2). We also consider 10% and 20% missing data, however, the conclusions are similar to the case of 5% and are consequently not shown. The boxplot is computed over 100 independent replicates of the $m(= 1,000)$-dimensional spatial field.

For Cases $1 - 3$, we find that our method outperforms the NNGP and the general Vecchia approximation at observed locations. In Case 4, ESD performs slightly worse than NNGP and general Vecchia approximation at observed locations. In Case 5, at observed locations the general Vecchia approximation slightly outperforms NNGP, which outperforms ESD. At missing locations the ESD performs similar to NNGP and general Vecchia approximation for Cases $1 - 2$, and slightly worse in Case 3. For missing locations in Cases $2 - 5$, ESD performs worse in terms of RMSPE than NNGP and the general Vecchia approximation at both missing and observed. Based on the results, we believe our model performs well (comparable to NNGP and the general Vecchia approximation) at observed locations (missing locations) in the highly non-stationary large-scale variability and stationary small-scale variability setting.

Case 1 in Figure 1 is meant to represent a difficult to estimate non-stationary spatial data setting. As such, it is reasonable to ask whether or not it is even possible to predict in this case. Thus, as a baseline comparison, we also compute predictions based on a model that assumes the data are independent and identically distributed. That is, we fit a model that assumes $Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s})$, where $Y(\cdot)$ is independently and identically distributed $N(0, \delta^2)$ and $\epsilon(\cdot)$ is independently and identically distributed $N(0, \sigma_\epsilon^2)$, and conjugate priors are assumed. For Case 1 the RMSPE is roughly 1.49 at missing locations and 1.48 at observed locations, over 100 independent replicates. Upon comparison to Figure 2, we see that other methods compared in this simulation study do consistently better. This is to be expected as the data are generated with spatial correlations present, and the result shows that models that allow for spatial correlations perform better.

All methods use the same covariates to model the non-stationary large-scale term f_0, however, these covariates may not perfectly model f_0. As a result, a model that allows for non-stationarity (i.e., ESD) may be able to partially model the residual between f_0 and the imperfect model $\mathbf{X}\boldsymbol{\beta}$. This is one reason why we might see that ESD outperforms the other methods in Case $1 - 3$, and is more competitive to less competitive in Cases 4 and 5. Of course one could specify the NNGP and the general Vecchia approximation based on a non-stationary covariance function, which may

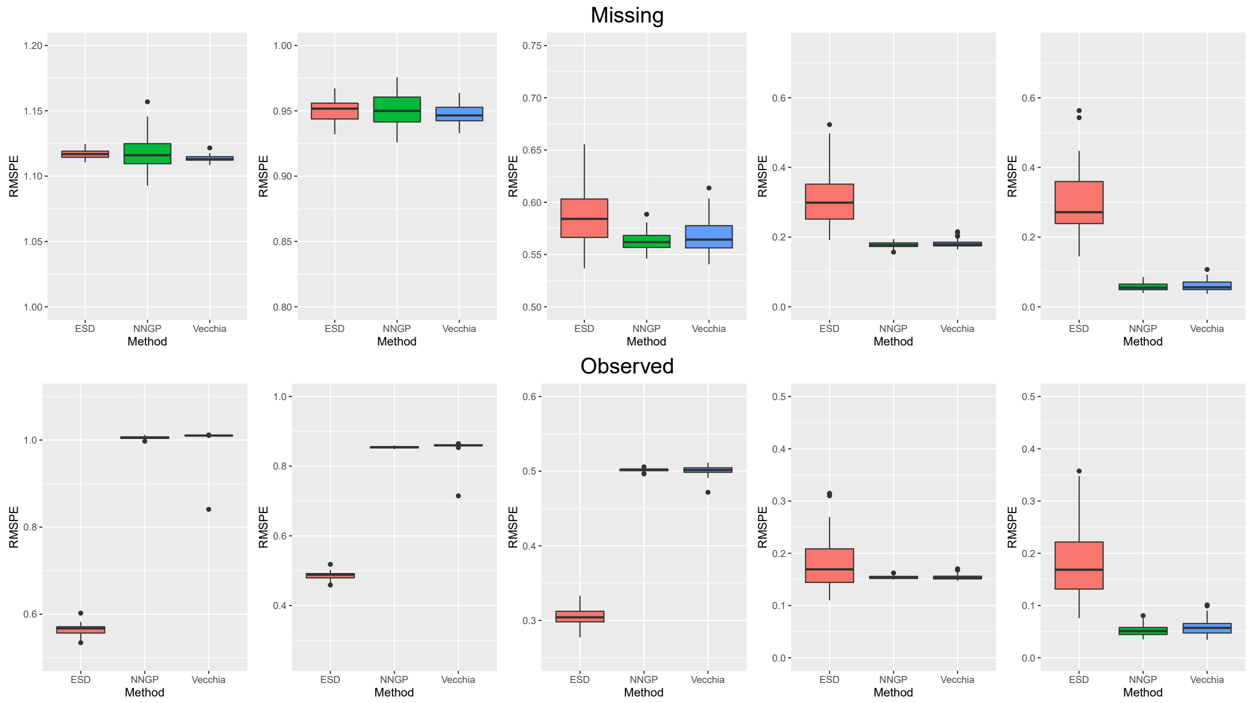

Figure 2: From first column to last column is Case 1 to Case 5. The SNR is set equal to 3. RMSPE on the first row is averaged over missing replicates, and the RMPSE on the second row is averaged over observed replicates. RMSPE is on the y-axis, and the method of prediction is labeled on the x-axis. Boxplots are made over 100 independent replicates.

perform better for this simulation setup. However, this requires an extra step of model selection for the non-stationary covariance function in NNGP and the general Vecchia approximation. Additionally, one could add basis functions to better model f_0 in the NNGP and the general Vecchia approximation. Ultimately, this highlights ESD's flexibility, since ESD simultaneously allow for a weakly stationary processes (i.e., $\boldsymbol{\eta} = \mathbf{0}_r$) and non-stationary processes (i.e., $\boldsymbol{\eta} \neq \mathbf{0}_r$) without the need for additional model selection when the large-scale-variability term is miss-specified. One could also envision a general Vecchia/NNGP approximation of ESD, however, this is a topic of future interest as there many ways one could do this.

4.3 Simulation Model 2 Set-Up

In our second simulation, we specify a 12 × 12 grid $D = \{(i,j)' : i,j = 0, \frac{1}{12}, \ldots, 12\}$, so that $m = 144$. Our main goal is to simulate non-stationary spatial processes, and compare the predictive performance of our method to the gold standard. This is done for two reasons, the first goal is to to assess the consequences of our model assumptions and approximations on prediction and computation. The second goal, is to illustrate the performance of our method in a two-dimensional non-stationary small-scale-variability setting.

We assume $\mathbf{Z} = \mathbf{Y} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is generated from a known non-stationary spatial model. The posterior mean (or approximated/estimated posterior mean) from this known non-stationary spatial model is considered the "gold standard" that should outperform ESD, since the data is generated from the gold standard. Thus, if the RMSPE based on our model is similar to that of the gold standard, this suggests that the effect of our assumptions/approximations have a small impact on prediction, and vice versa. We choose σ_ϵ^2 to achieve a moderately large SNR $= 5$. We allow for 5% of the locations in D to be missing at random.

We propose two simulation cases,

$$\begin{aligned}
&\text{Case 6:} \quad Y(\mathbf{u}) \text{ is generated from a zero mean SPC model}\\
&\text{Case 7:} \quad Y(\mathbf{u}) \text{ is generated from a zero mean SPDE model,}
\end{aligned}$$

where note that Cases $1 - 5$ fall under Simulation Model 1. In Case 6 we generate the data and fit the SPC model using functions made available by the R-package `convoSPAT` (e.g., see Risser and Calder, 2015). Specifically, for Case 6, we use five mixture components with the default specifications of the functions "`f_mc_kernels`" and "`NSconvo_sim`" to generate $Y(\mathbf{u})$. In Case 7, we generate the data and fit the SPDE model using functions made available by the R-package `inla` (e.g., see Lindgren and Rue, 2015). Specifically, three B-splines are specified to model the variance and spatial range parameter on the log-scale with coefficients 1, 3, and 1 (e.g., see Krainski et al. (2018) to see the specific functions to call to generate the SPDE model in this manner). For Case 7, observations are re-scaled to be roughly between -0.2 to 0.2. We implement a weighted ESD using 69 equally spaced bisquare basis functions to model both $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ using the R package `FRK`

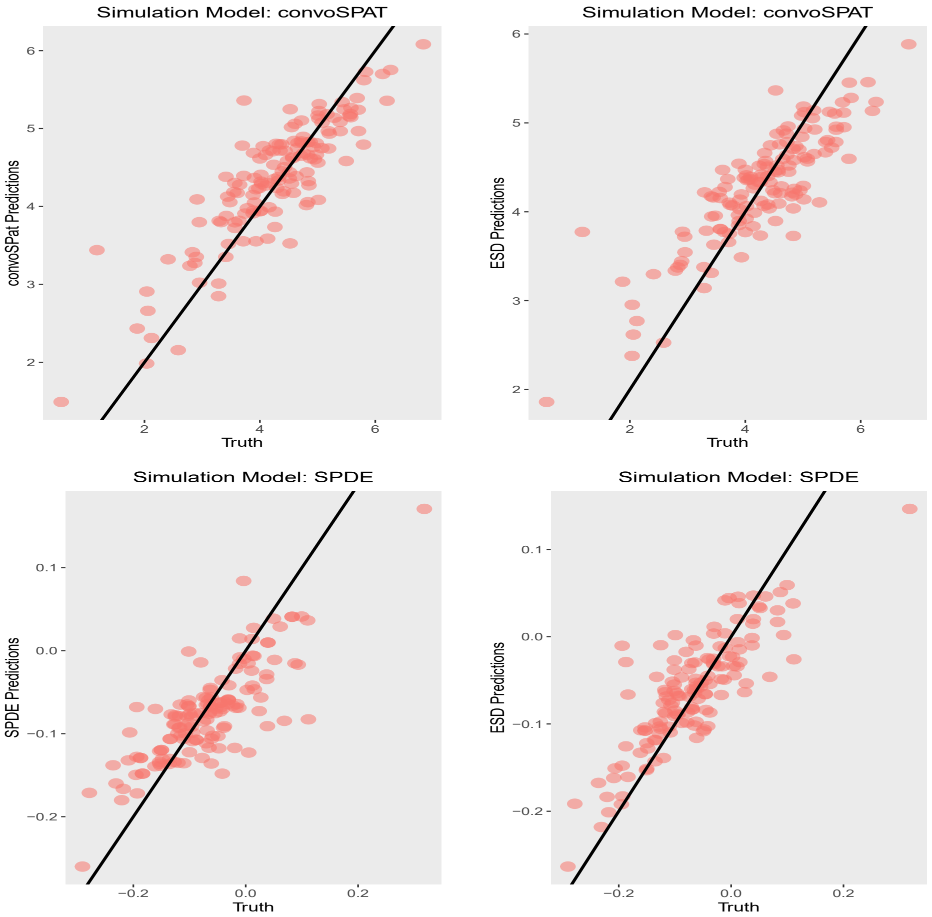

Figure 3: In the top left panel we plot the posterior mean from ESD versus the true $Y(\cdot)$ for Case 6. In the top right panel we plot the gold standard versus the true $Y(\cdot)$ Case 6. In the bottom left panel we plot the posterior mean from ESD versus the true $Y(\cdot)$ for Case 7. In the bottom right panel we plot the gold standard versus the true $Y(\cdot)$ Case 7. The gold standard in Case 6 are the predictions using SPC (SPDE) computed using the R package convoSPAT (inla).

(Zammit-Mangion and Cressie, 2017). We also set $w = w^* = 50$ and $K = 100$. We again use the exponential covariogram by setting the spectral density equal to a $Cauchy(0, 1/\phi)$. We informally considered a similar proof-of-concept baseline comparison to an i.i.d. model as done in Simulation Study 1, and found that SPC, SPDE, and ESD outperform an i.i.d. model in terms of RMSPE under this simulation design.

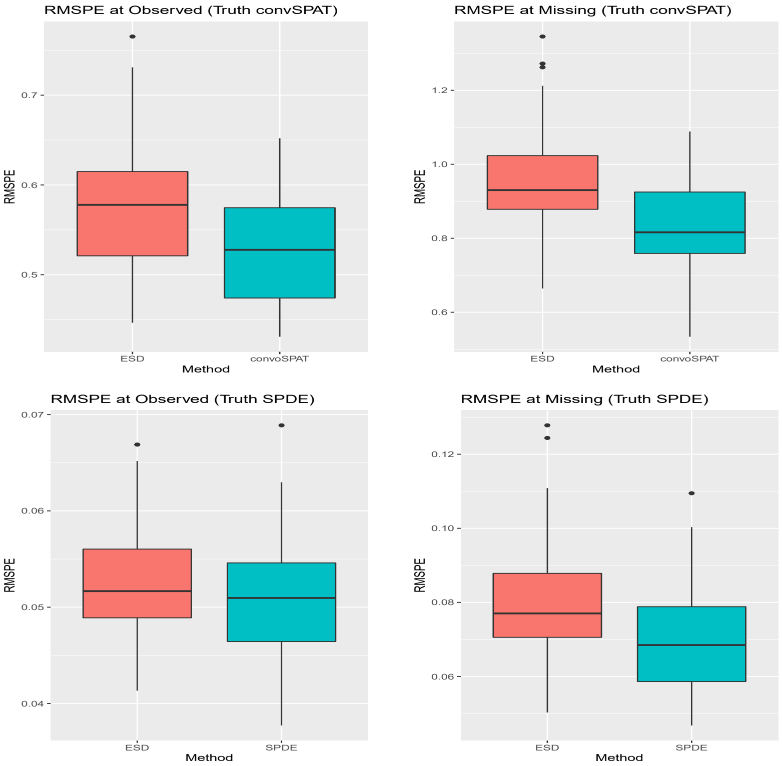

Figure 4: Top Row: Data generated from convoSPAT. Bottom Row: Data generated from SPDE. Left Column: Observed locations. Right Column: Missing Locations. Boxplots computed over 50 independent replicates. Scales are different across panels for visualization purposes.

4.4 Results for Simulation Model 2

For illustration, in Figure (3) we plot the predicted ESD versus the truth and the gold standard versus the truth for two different replicates generated from Cases 6 and 7, respectively. These plots suggest that under both cases, we obtain very similar predictions, however, as expected the gold standard in each case performs better in terms of RMSPE. In Case 7, it appears that ESD predictions slightly oversmooth the truth compared to the gold standard. We repeat this simulation study 50 times and compute boxplots of the respective RMSPEs, which are displayed in Figure (4). The results again suggest that the prediction errors of ESD are similar, albeit larger, than the respective gold standards, and this difference is larger at missing locations. Consequently, the ESD appears to perform well when data is generated from current models for non-stationary spatial

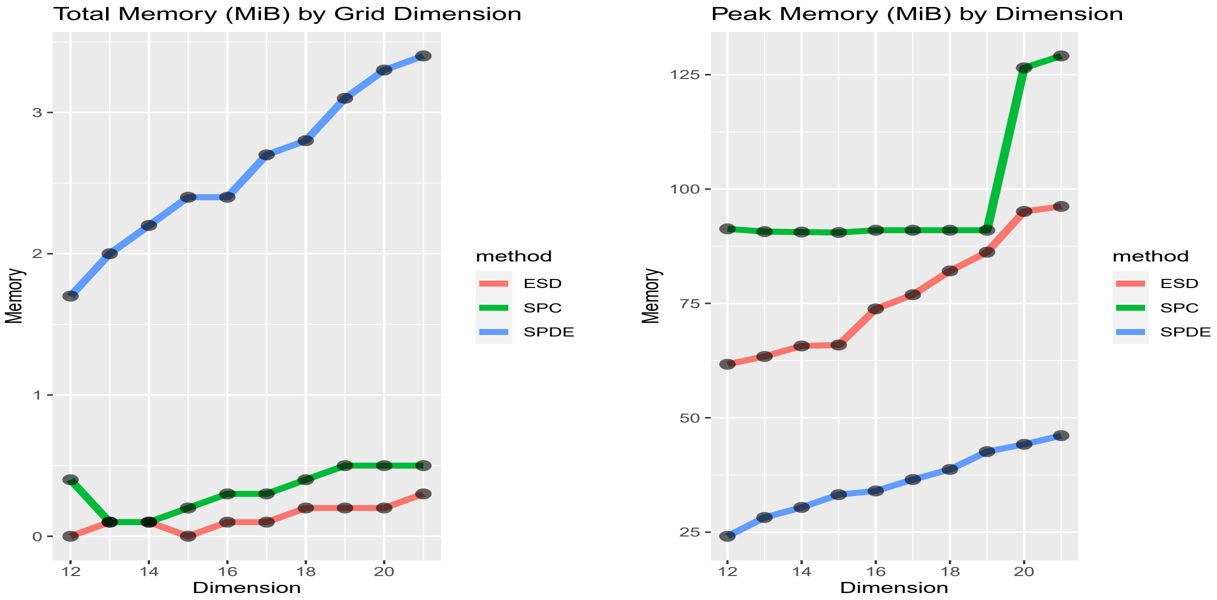

Figure 5: Total memory and average peak memory in MiB for simulations from Case 7 by method (i.e., SPC, ESD, SPDE). The dimension t is found by changing the dimensions of D from 12×12 to $t \times t$ so that $m = t^2$. We again assume 5% missing for each m.

data.

We also compare the computational performances of these methods in terms of storage, since our avoidance of large covariance matrices aids with decreasing memory complexity. If the basis function matrices (used to model $\mathbf{f}$) are dense, and the subset size w ($=w^*$) is small, then ESD has memory complexity on the order of mdr. However, bisquare basis functions are sparse. Furthermore, when using the R function `auto_basis` in the R package `FRK` to compute the bisquare basis function, we find that the sparsity of the basis functions often increases with the number of prediction locations m. For example, when $m = 144$ roughly 77% of the elements in the basis function matrix are zero, and when $m = 441$ the percentage of zero elements in the basis function matrix increases to roughly 81%. This is exciting, as we can actively increase the sparsity of the bisquare basis functions in a manner that decreases the memory complexity of ESD as m increases. A negative consequence is that it is not clear how to analytically derive the memory complexity of ESD. Thus, to investigate this further, we empirically investigate the memory complexity of ESD relative to SPC and SPDE.

In Figure (5), we plot the total random access memory (RAM) and peak RAM in mebibytes (MiB) by method (ESD, SPC, and SPDE) and dimension of D (i.e., t, where $m = t^2$), where peak RAM represents the maximum amount of memory used and total RAM measures the overall memory used. In general, peak RAM is the more crucial quantity. These quantities are computed using the R package `peakRam` (Quinn, 2017). The total RAM for each method generally increases as t increases. We see that ESD requires slightly less total RAM than SPC, and both ESD and SPC require less total RAM than SPDE. However, ESD produces higher peak RAM than SPDE, but generally less peak RAM than SPC. In general, the peak RAM of ESD is roughly constant across these dimensions. Thus, empirically our methods appears to require less total RAM than SPDE and SPC, and we suggest using sparse basis functions to control the peak RAM. However, these results are based on our coding of ESD, the specifications of one computer (i.e., Figure (5) is based on Windows 10 and Intel(R) CORE(TM) i5-8250U CPU with 1.60Ghz), and the current versions of `convoSPAT` and `inla`. Thus, these results should be seen as subjective aids at understanding the memory complexity of ESD relative to standard non-stationary prediction models.

5 Real Data Application

5.1 Ozone Data Application: Data Description

As an illustration, we analyze the ozone dataset used in Cressie and Johannesson (2008), which has become a benchmark dataset in the spatial statistics literature (e.g., see Zhang et al., 2019). This dataset consists of $n = 173,405$ values of total column ozone (TCO) in Dobson units (see Figure (6) for a plot of the data). The dataset was obtained through a Dobson spectrophotometer on board the Nimbus-7 polar orbiting satellite on October 1st, 1988. For details on how these data were collected see Cressie and Johannesson (2008). This dataset is made publically available by the Centre for Environmental Informatics at the University of Wollongong's National Institute for Applied Statistics Research Australia (`https://hpc.niasra.uow.edu.au/ckan/`).

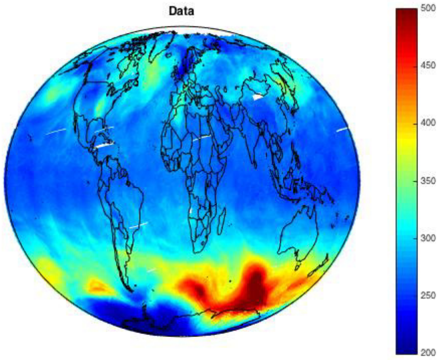

Figure 6: Level 2 total column ozone data (in Dobson units) collected on October 1st, 1988, and analyzed by Cressie and Johannesson (2008).

5.2 Analysis

We present an analysis of the ozone dataset using ESD. We partition the data into a training set and a prediction set. We randomly generated 5% of the observed values to be included in the validation dataset for evaluating the prediction performance of all methods. A total of 2,000 MCMC iterations of the Gibbs sampler in Algorithm 1 were used. The first 1,000 iterations were treated as a burn-in. We informally check trace plots for convergence, and no lack of convergence was detected. Since d=2, $\boldsymbol{\psi}(\mathbf{s})$ is an $r \times 2$ dimensional matrix, which we denote with $\boldsymbol{\psi}(\mathbf{s}) = \{\boldsymbol{\phi}_1(\mathbf{s}), \boldsymbol{\phi}_2(\mathbf{s})\}$, where $\boldsymbol{\phi}_i(\mathbf{s})$ is an r-dimensional vector, i=1,2. Using the R-package `FRK`, we consider two possible values for r (i.e., $r = 92$ and 364) equally-spaced bisquare basis functions on a spherical domain, which defines either 92 or 364-dimensional vectors $\boldsymbol{\zeta}(\mathbf{s})$ (Zammit-Mangion and Cressie, 2017). Then, we set $\boldsymbol{\phi}_1(\mathbf{s}) = \{\mathbf{0}_3', \boldsymbol{\zeta}(\mathbf{s})'\}'$, and we take $\boldsymbol{\phi}_2(\mathbf{s}) = (1, \mathbf{s}', \mathbf{0}_r')'$. This choice of $\boldsymbol{\phi}_2(\mathbf{s})$ isolates the effect of the latitude and longitude on the non-stationarity of the process. The covariates are defined to be $\mathbf{x}(\mathbf{s}) = (1, \boldsymbol{\zeta}(\mathbf{s})')'$. We use the estimate of σ_ϵ^2 from Cressie and Johannesson (2008).

Figure (7) displays the prediction and prediction variances using non-stationary spectral simulation. Upon comparison of Figure (6) to Figure (7a), we see that we obtain small in-sample error. Additionally, (7b) shows that our prediction error is relatively constant over the globe. We com-

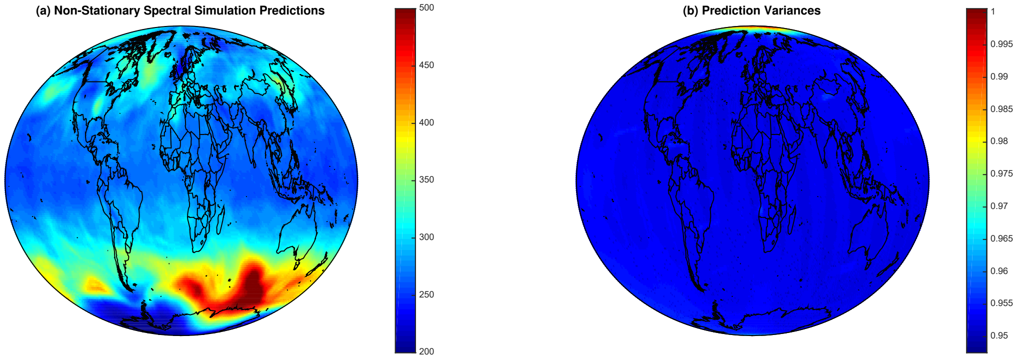

Figure 7: Weighted ESD results for 364 basis functions and $w = w^* = 20$. In (a), we plot the posterior means (in Dobson units) from the model in (2), which was implemented using the Gibbs sampler outlined in Algorithm 1. The corresponding posterior variances (in Dobson units squared) are presented in (b).

pute RMSPE over the validation locations. Specifically, we compute the average square distance between the validation data and its corresponding prediction, and then we take the square root. RMSPE for our method decreases as r increases (see Figure 8). We also computed the RMSPE for the fixed rank kriging method as implemented through the R-package `FRK` (Zammit-Mangion and Cressie, 2017). The FRK predictor is based on $\mathbf{x}(\mathbf{s}) = 1$ and uses $\boldsymbol{\zeta}(\mathbf{s})$ as its basis set. The RMSPE for FRK is approximately 7.12, and hence, we outperform the FRK predictor in terms of RMSPE when $r = 364$ and $w = w^* = 20$. Of course, one could increase the number of basis functions in FRK, but for comparison purposes we wish to include the same number of basis functions in both ESD and FRK. In Zhang et al. (2019), they compare the Smoothed Full-Scale Approximation (SFSA), Full-Scale Approximation using a block modulating function (FSAB) (Sang and Huang, 2012), NNGP, and a local Gaussian process method with adaptive local designs (LaGP) (Gramacy and Apley, 2015). Their results show that the RMSPE for SFSA, NNGP, and FSAB are all around 5.2, and the RMSPE for LaGP is around 6.2. Thus, our method also outperforms these methods in terms of RMSPE. The general Vecchia approximation has a similar result with RMSPE equal to 5.15, where the ESD had RMSPE of 5.05. This small difference may be accounted for by Monte

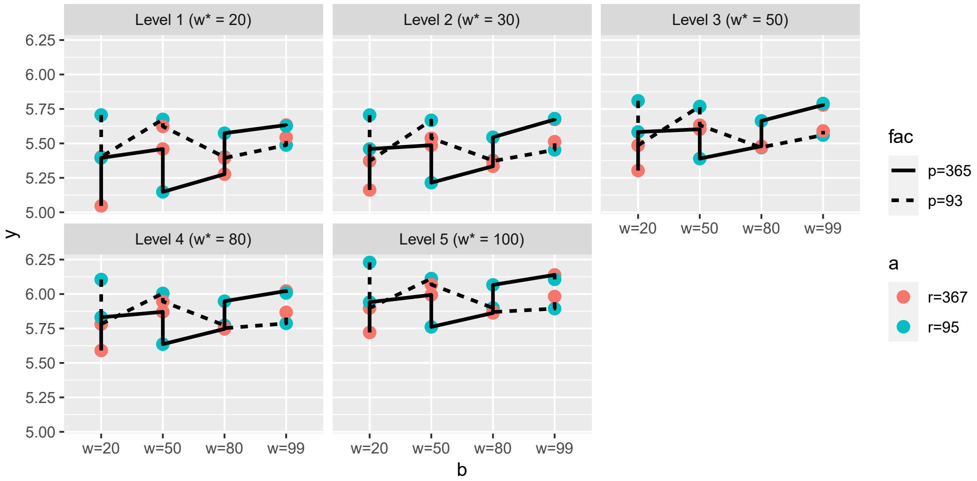

Figure 8: Sensitivity Study: RMSPE (over holdout locations) by p, r, w, and w^*.

Carlo error and thus we conclude that similar predictions are produced by both methods. Ultimately, we perform (marginally) the best in this hold-out study, and this a well-known benchmark dataset used in the nonstationary data setting (Cressie and Johannesson, 2008). Our results for the ESD, the general Vecchia approximation, and FRK are based on the validation dataset that we generated. However, SFSA, NNGP, FSAB, and LaGP are based on the validation dataset generated in Zhang et al. (2019).

We also investigated the sensitivity on the choice of w, w^*, r, and p. The choice of w effectively subsets the data when updating $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ and is also used in the posterior predictive step to produce our predictions. In general, w is restricted to be small to obtain computational advantages, and there are no theoretical guarantees that small values of w will lead to precise predictions. However, several recent papers have suggested that subsetting in the spatial setting often leads to reasonable predictions (Katzfuss and Guinness, 2019; Bradley, 2019). In general, we have found worse (in terms of prediction) results as w^* becomes larger than w, better predictive performance as r and p increase, and smaller choices for w^* is preferable. In Figure 8 we give the RMSPEs for $p = 93, 365$, $r = 95, 367$, $w = 20, 50, 80, 99$, and $w^* = 20, 50, 80, 100$, where the different levels of p and r are

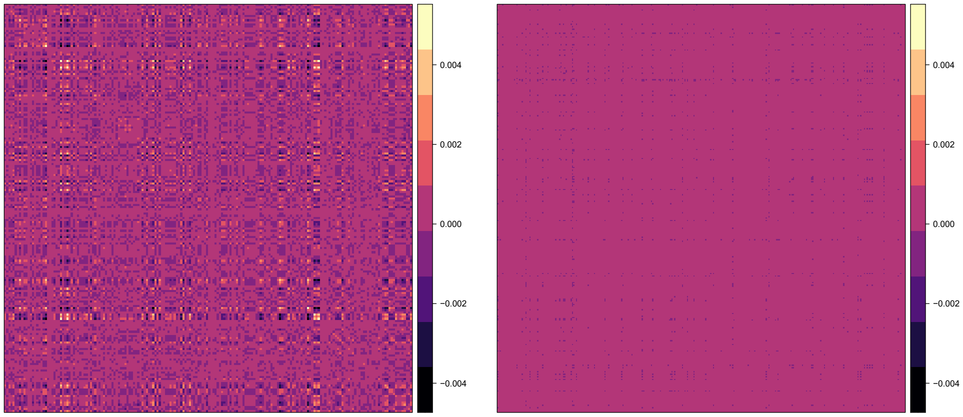

Figure 9: Posterior Covariance Matrix Image Plot. From left to right is 92, 364 basis function.

based on the two basis function specifications of $\boldsymbol{\zeta}(\mathbf{s})$. The RMSPE tends to be roughly between 5 and 6, and appears to be fairly robust to different specifications. Thus, choosing small values of w for computational purposes appears to lead to reasonable predictions for this dataset.

We also include the computation times in Table (2). Our method is less competitive. Although we avoid storing and inverting a high dimensional covariance matrix, we require nested loops, which can be computationally intensive (i.e., a loop in the Gibbs sampler and a loop over $i = 1 \ldots K$ in Theorem 2).

In terms of inference on parameters, we are particularly interested in $\boldsymbol{\eta}$ and $\mathbf{f}$. This is because when $\boldsymbol{\eta}$ is zero (and $\mathbf{f}$ is a constant function at zero) we obtain a stationary process (see Theorem 1). In Figure (9) we plot the posterior covariance matrix of $\boldsymbol{\eta}$. Our main goal with this plot is to show empirically what happens to the covariance matrix of $\boldsymbol{\eta}$ under different model specifications. As r increases, we see the variances and covariances appear to be close to zero. In both cases several credible intervals for elements of $\boldsymbol{\eta}$ do not contain zero, which suggests that non-stationarity is present in this dataset. Furthermore, we emphasize the ability of ESD to do inference on $\boldsymbol{\eta}$, which acts as a proxy for the presence/absence of non-stationarity (see Theorem 1). The competing methods can not immediately perform inference on $\mathbf{f}$.

The best predictive performance of ESD occurs when $w = w^* = 20$, which is a small subset size. As the subset size decreases to zero point estimation on $\boldsymbol{\eta}$ favors the prior mean zero (i.e., stationarity). This specification is reasonable considering that the most comparable method in terms of RMSPE, the general Vecchia approximation, approximates a model with stationarity covariances. This suggests that if non-stationarity is present then locally stationary (Donoho et al., 1996) covariances over large regions may be expected, since the global stationary covariance assumption leads to high performing predictions. When $\mathbf{f}(\mathbf{s}) = \mathbf{f}(\mathbf{u})$ for $\mathbf{s} \neq \mathbf{u}$ and $\mathbf{s}, \mathbf{u} \in A \subset D$, from Theorem 1, the expression of the covariance is a stationary covariogram defined over the subregion A. As such, maps with several such subregions suggest non-stationarity, and specifically, local stationary covariances. In Figure 10, we plot the components of $\mathbf{f}(\cdot) = (f_1(\cdot), f_2(\cdot))'$. Here, we see large regions of similar values of f_1 and f_2 suggesting this type of locally stationary behavior is present. Thus, for this illustration small w (and w^*) is reasonable, since $\mathbf{f}$ does not appear to be (functionally) highly variable and does not appear to differ wildly from zero, which enforces a type of local stationarity. When $\mathbf{f}$ has large changes functionally then a larger value of w is likely needed and computations will be extensive in this setting.

We reiterate that we do not recommend generating a just single subsample without some sort of cross-validation procedure, as this could possibly produce poor predictions and estimates. That is, we generate multiple subsamples under multiple model specifications and choose w and the placement of w the locations by minimizing a hold-out error. In this manner the entire dataset is used to select the subsample w (i.e., the training data is used to predict, and the holdout data ia used to validate). If the hold-out error suggests that $w = 0$ is preferable, our algorithm will produce covariances that are close to the prior mean of $\mathbf{f}$, which is zero and the stationary case (see Theorem 1 where $\mathbf{f}$ identically equal to zero produces Bochner's theorem). In this example, the estimated w is greater zero, and the function $\mathbf{f}$ is significantly different from zero using pointwise

Table 2: Computation Time by r and method.

Method	r	**Time (seconds)**
ESD	92	2345
	364	13394
Vecchia Approximation	-	200

credible intervals. This suggests non-stationarity. Furthermore, the maps in Figure 10 suggest patterns consistent with local stationarity.

6 Discussion

Bayesian analysis of big Gaussian spatial data is a challenging and important problem. We propose a Bayesian approach using non-stationary spectral simulation. To develop non-stationary spectral simulation we combine Bochner's theorem with dimension expansion (Perrin and Meiring, 2003), and apply Mejía and Rodríguez-Iturbe (1974)'s spectral simulation method. The advantage is that no large matrix inversion or storage is needed to approximately simulate a non-stationary full-rank Gaussian process. Additionally, the proposed method is extremely broad, since every positive definite non-stationary covariance function can be written according to (5). A novel strategy that involves weighting the data model is introduced so that (1) the data is conditionally independent of the spatially covarying random effects at observed locations given the covariance parameters to allow one to use non-stationary spectral simulation within a collapsed Gibbs sampler, and (2) the covariance parameters are conditionally independent of a large subset of the data. Furthermore, these computational compromises (i.e., adding conditional independence assumptions) do not impose any additional assumptions on our (marginal) model for the data (i.e., see (9)).

In Section 4, Simulation Study 1 is used to illustrate the high-performance of our method when the large-scale-variability term is nonlinear and the small-scale-variability is stationary. We compare to current methods to analyze non-stationary large-scale-variability and stationary small-scale-variability; namely, the nearest neighbor Gaussian process (NNGP; Datta et al., 2016a) model

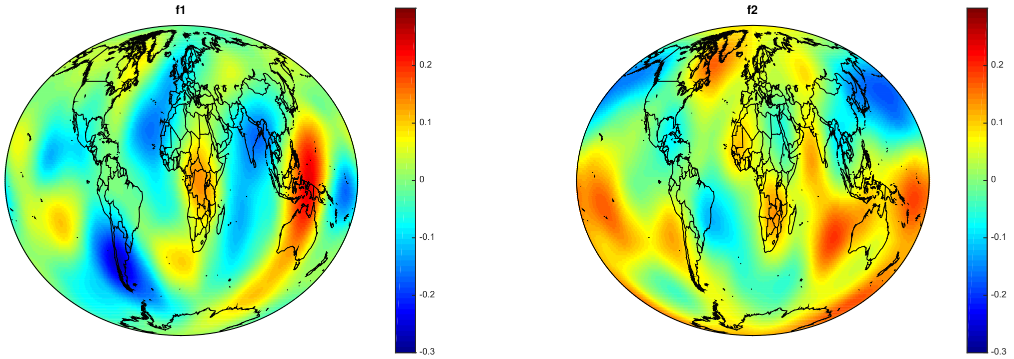

Figure 10: Let $\mathbf{f}(\cdot) = (f_1(\cdot), f_2(\cdot))'$. In the left and right panels we plot the posterior means of f_1 and f_2, respectively.

and Vecchia approximation (Katzfuss and Guinness, 2019). In this setup, the ESD is consistently preferable in terms of root mean squared prediction error (RMPSE) at observed locations, and has comparable RMSPE to NNGP at missing locations in several cases. We generate data that is different from our model, and we find our method has better results in different scenarios based on how nonlinear the process is. In Simulation Study 2, we generate data from standard non-stationary spatial models, whose predictions are treated as a gold standard; namely SPC (Paciorek and Schervish, 2006) and SPDE (Lindgren et al., 2011). Here, we see that the RMSPE of ESD is comparable to the gold standard method in terms of RMSPE. Moreover, our approach appears to require less total RAM than SPC and SPDE, and peak RAM larger than SPDE. In Section 5, we analyze the total column ozone dataset from Cressie and Johannesson (2008). We obtain predictions that have small in-sample error, and that appears to outperform fixed rank kriging (FRK), SFSA, FSAB, NNGP, and LaGP in terms of out-of-sample error. The hold-out error was similar in value to the general Vecchia approximation. Additionally, our framework allows one to perform inference on the presence of generic non-stationarity, which is not immediately possible using the competing methods.

Environmental studies are often based on high-dimensional spatial Gaussian datasets with complex patterns of non-stationarity. Several studies focus on simplifying matrix valued operations and storage (Higdon et al., 1999; Paciorek and Schervish, 2006; Cressie and Johannesson, 2008; Banerjee et al., 2008; Lindgren et al., 2011; Nychka et al., 2015). Thus, our "large-matrix-free" approach offers a unique solution to this important problem. Despite these advantageous there are limitations that offer opportunities for future research. The ESD is fairly competitive when it comes to memory complexity through its use of spectral simulation. However, computation time is noticeably less competitive, since we have nested loops within our algorithm; namely the iterative non-stationary spectral simulation is nested within a collapsed Gibbs sampler. Additionally, there are several parameters that are fixed and chosen using cross-validation in practice, including the number of iterations in the non-stationary spectral simulator, the selection of the subset of the dataset used to update covariance parameters, and the rank of the basis function expansion.

The assumption of conditional independence between the spatially varying random effects at observed locations and the data is a limitation of ESD, since this type of conditional dependence may be present at the observed locations. However, in general, we do not need to spatially interpolate the process at observed locations, and as a result, this particular conditional independence assumption at observed locations still allows one to incorporate large-scale and small-scale variability into the predictions. We have found that this assumption leads to competitive RMSPEs and computational speed-ups through the use of nonstationary spectral simulation.

Acknowledgments

Jonathan Bradley and Hou-Cheng Yang's research was partially supported by the U.S. National Science Foundation (NSF) grant SES-1853099. We would like to thank the reviewers for their helpful comments, and we are especially indebted to one reviewers comments, which led to the development of Section 3.2.

Appendix A: Proofs

Proof of Theorem 1:

It follows from Perrin and Meiring (2003) that for every non-stationary positive definite function C and every pair of locations $\mathbf{s}_1$ and $\mathbf{s}_2$ there exists a $\mathbf{w}_1$ and $\mathbf{w}_2$ such that $C(\mathbf{s}_1, \mathbf{s}_2) = \rho\left\{\left(\begin{smallmatrix}\mathbf{s}_1\\ \mathbf{w}_1\end{smallmatrix}\right), \left(\begin{smallmatrix}\mathbf{s}_2\\ \mathbf{w}_2\end{smallmatrix}\right)\right\}$, where ρ is a stationary covariogram. Let $\mathbf{f}$ be the function that maps generic locations $\mathbf{s}_1, \mathbf{s}_2 \in D$ to its corresponding expanded dimension $\mathbf{w} \in \mathbb{R}^d$ so that $C(\mathbf{s}_1, \mathbf{s}_2) = \rho\left\{\left(\begin{smallmatrix}\mathbf{s}_1\\ \mathbf{w}_1\end{smallmatrix}\right), \left(\begin{smallmatrix}\mathbf{s}_2\\ \mathbf{w}_2\end{smallmatrix}\right)\right\}$.

It follows from Bochner's theorem (Bochner, 1959) that $\rho\left\{\left(\begin{smallmatrix}\mathbf{s}_i\\ \boldsymbol{f}(\mathbf{s}_i)\end{smallmatrix}\right), \left(\begin{smallmatrix}\mathbf{s}_j\\ \boldsymbol{f}(\mathbf{s}_j)\end{smallmatrix}\right)\right\}$ is positive definite (and equivalently so is $C(s_i, s_j)$) if and only if

$$C(s_i, s_j) = \rho\left\{\left(\begin{smallmatrix}\mathbf{s}_i\\ \boldsymbol{f}(\mathbf{s}_i)\end{smallmatrix}\right), \left(\begin{smallmatrix}\mathbf{s}_j\\ \boldsymbol{f}(\mathbf{s}_j)\end{smallmatrix}\right)\right\} = \int_{-\infty}^{\infty} \cos\left\{\{\boldsymbol{f}(\mathbf{s}_i) - \boldsymbol{f}(\mathbf{s}_j)\}'\boldsymbol{\omega}_1 + (\mathbf{s}_i - \mathbf{s}_j)'\boldsymbol{\omega}_2\right\} G_\theta(d\boldsymbol{\omega}).$$

This completes the result.

Proof of Theorem 2:

We have that,

$$E\{\widetilde{\nu}(\mathbf{s})\} = \frac{(2K)^{\frac{1}{2}}}{2\pi\sigma_\nu} \int_{-\infty}^{\infty} \int_{-\pi}^{\pi} \cos(\boldsymbol{f}(\mathbf{s})'\boldsymbol{\omega}_{1,i} + \mathbf{s}'\boldsymbol{\omega}_{2,i} + \kappa_i) g_\theta(\boldsymbol{\omega}) d\kappa d\boldsymbol{\omega} = 0,$$

since $\int_{-\pi}^{\pi} \cos(\kappa) d\kappa = \int_{-\pi}^{\pi} \sin(\kappa) d\kappa = 0$. Also,

$$\begin{aligned}
&E\{\widetilde{\nu}(s_i)\widetilde{\nu}(s_j)\} \\
&= \frac{2\sigma_\nu^2}{K} \sum_{i=1}^{K} \sum_{i=1}^{K} E[\cos(\boldsymbol{f}(\mathbf{s}_i)'\boldsymbol{\omega}_{1,i} + \mathbf{s}_i'\boldsymbol{\omega}_{2,i} + \kappa_i) \cos(\boldsymbol{f}(\mathbf{s}_j)'\boldsymbol{\omega}_{1,i} + \mathbf{s}_j'\boldsymbol{\omega}_{2,i} + \kappa_i)] \\
&= \int \cos\{(\boldsymbol{f}(\mathbf{s}_i)' - \boldsymbol{f}(\mathbf{s}_j)')\boldsymbol{\omega}_{1,i} + (s_i - s_j)\boldsymbol{\omega}_{2,i}\} g_\theta(\boldsymbol{\omega}_1) g_\theta(\boldsymbol{\omega}_2) d\boldsymbol{\omega}_1 d\boldsymbol{\omega}_2 \\
&= \rho\left\{\left(\begin{smallmatrix}\mathbf{s}_i\\ \boldsymbol{f}(\mathbf{s}_i)\end{smallmatrix}\right) - \left(\begin{smallmatrix}\mathbf{s}_j\\ \boldsymbol{f}(\mathbf{s}_j)\end{smallmatrix}\right)\right\} = C(s_i, s_j).
\end{aligned}$$

As $K \to \infty$, $\{\widetilde{\nu}(\mathbf{u}) : \mathbf{u} \in \mathcal{B}\}$ converges to a Gaussian process, where $\mathcal{B}$ is a ball in $D \subset \mathbb{R}^d$ (e.g., see Cressie, 1993, pg. 204).

Proof of Equation (9):

From (3) we have

$$p(\mathbf{Z}) = \int \dots \int p(\mathbf{Z}, \boldsymbol{\nu}_O, \widetilde{\boldsymbol{\nu}}_O, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2_\eta)\, d\boldsymbol{\nu}_O\, d\boldsymbol{\beta}\, d\widetilde{\boldsymbol{\nu}}_O\, d\sigma^2_\eta\, d\boldsymbol{\theta}\, d\boldsymbol{\gamma}$$

$$= \int \dots \int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma^2_\epsilon) p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2) p(\boldsymbol{\beta}|\sigma^2_\beta) p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\sigma^2_\eta) p(\boldsymbol{\gamma}) p(\sigma^2_\eta) W(\boldsymbol{\theta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{\nu}}_O, \mathbf{Z})\, d\boldsymbol{\nu}_O\, d\boldsymbol{\beta}\, d\widetilde{\boldsymbol{\nu}}_O\, d\sigma^2_\eta\, d\boldsymbol{\theta}\, d\boldsymbol{\gamma}$$

$$= \int \dots \int \frac{h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma^2_\epsilon) p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2) p(\boldsymbol{\beta}|\sigma^2_\beta) p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\sigma^2_\eta) p(\boldsymbol{\gamma}) p(\sigma^2_\eta) h(\mathbf{Z}_{-w}|\mathbf{Z}_w) h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})}{\int \int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma^2_\epsilon) p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2) p(\boldsymbol{\beta}|\sigma^2_\beta) d\boldsymbol{\nu}_O d\boldsymbol{\beta}}\, d\boldsymbol{\nu}_O\, d\boldsymbol{\beta}\, d\widetilde{\boldsymbol{\nu}}_O\, d\sigma^2_\eta\, d\boldsymbol{\theta}\, d$$

$$=$$

$$\int \dots \int \left[\int \int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma^2_\epsilon) p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2) p(\boldsymbol{\beta}|\sigma^2_\beta)\, d\boldsymbol{\nu}_O\, d\boldsymbol{\beta} \right] \frac{p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\sigma^2_\eta) p(\boldsymbol{\gamma}) p(\sigma^2_\eta) h(\mathbf{Z}_{-w}|\mathbf{Z}_w) h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})}{\int \int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma^2_\epsilon) p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2) p(\boldsymbol{\beta}|\sigma^2_\beta) d\boldsymbol{\nu}_O d\boldsymbol{\beta}}\, d\widetilde{\boldsymbol{\nu}}_O\, d\sigma$$

$$= \int \int \int \left[\int p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta})\, d\widetilde{\boldsymbol{\nu}}_O \right] p(\sigma^2_\eta)\, d\sigma^2_\eta p(\boldsymbol{\theta}|\sigma^2_\eta) p(\boldsymbol{\gamma}) p(\sigma^2_\eta) h(\mathbf{Z}_{-w}|\mathbf{Z}_w) h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})\, d\sigma^2_\eta\, d\boldsymbol{\theta}\, d\boldsymbol{\gamma}$$

$$= h(\mathbf{Z}_{-w}|\mathbf{Z}_w) \int \int \int p(\boldsymbol{\theta}|\sigma^2_\eta) p(\boldsymbol{\gamma}) p(\sigma^2_\eta) h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})\, d\sigma^2_\eta\, d\boldsymbol{\theta}\, d\boldsymbol{\gamma}$$

$$= \frac{\int \dots \int h(\mathbf{Z}, \boldsymbol{\nu}_O, \widetilde{\boldsymbol{\nu}}_O, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2_\eta)\, d\boldsymbol{\nu}_O\, d\widetilde{\boldsymbol{\nu}}_O\, d\boldsymbol{\beta}\, d\boldsymbol{\theta}\, d\boldsymbol{\gamma}\, d\sigma^2_\eta}{\int \int \int p(\boldsymbol{\theta}|\sigma^2_\eta) p(\boldsymbol{\gamma}) p(\sigma^2_\eta) h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})\, d\sigma^2_\eta\, d\boldsymbol{\theta}\, d\boldsymbol{\gamma}} \int \int \int p(\boldsymbol{\theta}|\sigma^2_\eta) p(\boldsymbol{\gamma}) p(\sigma^2_\eta) h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})\, d\sigma^2_\eta\, d\boldsymbol{\theta}\, d\boldsymbol{\gamma}$$

$$= \int \dots \int h(\mathbf{Z}, \boldsymbol{\nu}_O, \widetilde{\boldsymbol{\nu}}_O, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2_\eta)\, d\boldsymbol{\nu}_O\, d\boldsymbol{\beta}\, d\widetilde{\boldsymbol{\nu}}_O\, d\sigma^2_\eta\, d\boldsymbol{\theta}\, d\boldsymbol{\gamma}$$

$$= h(\mathbf{Z}),$$

which completes the result.

Proof of Equation (10):

We have that,

$$p(\widetilde{\boldsymbol{\nu}}_O, \boldsymbol{\theta}, \mathbf{Z}) = \int \int \int \frac{h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma^2_\epsilon) p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2) p(\boldsymbol{\beta}|\sigma^2_\beta) p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\sigma^2_\eta) p(\boldsymbol{\gamma}) p(\sigma^2_\eta) h(\mathbf{Z}_{-w}|\mathbf{Z}_w) h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})}{\int \int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma^2_\epsilon) p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2) p(\boldsymbol{\beta}|\sigma^2_\beta) d\boldsymbol{\nu}_O d\boldsymbol{\beta}}\, d\boldsymbol{\nu}_O\, d\boldsymbol{\beta}\, d\sigma$$

$$= \int \int \left[\int \int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma^2_\epsilon) p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2) p(\boldsymbol{\beta}|\sigma^2_\beta) d\boldsymbol{\nu}_O d\boldsymbol{\beta} \right] \frac{p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\sigma^2_\eta) p(\boldsymbol{\gamma}) p(\sigma^2_\eta) h(\mathbf{Z}_{-w}|\mathbf{Z}_w) h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})}{\int \int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma^2_\epsilon) p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2) p(\boldsymbol{\beta}|\sigma^2_\beta) d\boldsymbol{\nu}_O d\boldsymbol{\beta}}\, d\sigma^2_\eta\, d\boldsymbol{\gamma}$$

$$= p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta}) h(\mathbf{Z}_{-w}|\mathbf{Z}_w) \int \int p(\boldsymbol{\theta}|\sigma^2_\eta) p(\sigma^2_\eta) h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma})\, d\sigma^2_\eta\, d\boldsymbol{\gamma}.$$

It follows that

$$p(\boldsymbol{\theta}, \mathbf{Z}) = \int p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta}) h(\mathbf{Z}_{-w}|\mathbf{Z}_w) \int \int p(\boldsymbol{\theta}|\sigma^2_\eta) p(\sigma^2_\eta) h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma})\, d\sigma^2_\eta\, d\boldsymbol{\gamma}\, d\widetilde{\boldsymbol{\nu}}_O$$
$$= h(\mathbf{Z}_{-w}|\mathbf{Z}_w) \int \int p(\boldsymbol{\theta}|\sigma^2_\eta) p(\sigma^2_\eta) h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma})\, d\sigma^2_\eta\, d\boldsymbol{\gamma},$$

so that $p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta}, \mathbf{Z}) = p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta})$, which completes the result.

Proof of Equation (11):

Let the $(m-w)$-dimensional vector $\widetilde{\boldsymbol{\nu}}_{-w} = (\widetilde{\nu}(\mathbf{u}_i) : i \in \{1, \ldots, m\}$ and $\widetilde{\nu}(\mathbf{u}_i)$ is not an element of $\widetilde{\boldsymbol{\nu}}_w)'$. Then,

$$
\begin{aligned}
&p(\boldsymbol{\theta}, \widetilde{\boldsymbol{\nu}}_w, \boldsymbol{\gamma}, \sigma_\eta^2, \mathbf{Z})\\
&= \int\int\int \frac{h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma_\epsilon^2)p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2)p(\boldsymbol{\beta}|\sigma_\beta^2)p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta})p(\boldsymbol{\theta}|\sigma_\eta^2)p(\boldsymbol{\gamma})p(\sigma_\eta^2)h(\mathbf{Z}_{-w}|\mathbf{Z}_w)h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})}{\int\int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma_\epsilon^2)p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2)p(\boldsymbol{\beta}|\sigma_\beta^2)d\boldsymbol{\nu}_O d\boldsymbol{\beta}}\, d\boldsymbol{\nu}_O\, d\boldsymbol{\beta}\, d\widetilde{\boldsymbol{\nu}}_{-w}\\
&= \int\left[\int\int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma_\epsilon^2)p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2)p(\boldsymbol{\beta}|\sigma_\beta^2)d\boldsymbol{\nu}_O d\boldsymbol{\beta}\right] \frac{p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta})p(\boldsymbol{\theta}|\sigma_\eta^2)p(\boldsymbol{\gamma})p(\sigma_\eta^2)h(\mathbf{Z}_{-w}|\mathbf{Z}_w)h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})}{\int\int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma_\epsilon^2)p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2)p(\boldsymbol{\beta}|\sigma_\beta^2)d\boldsymbol{\nu}_O d\boldsymbol{\beta}}\, d\widetilde{\boldsymbol{\nu}}_{-w}\\
&= p(\boldsymbol{\theta}|\sigma_\eta^2)h(\mathbf{Z}_{-w}|\mathbf{Z}_w)h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})p(\boldsymbol{\gamma})p(\widetilde{\boldsymbol{\nu}}_w|\boldsymbol{\theta})p(\sigma_\eta^2),
\end{aligned}
$$

so that $p(\boldsymbol{\theta}|\widetilde{\boldsymbol{\nu}}_w, \boldsymbol{\gamma}, \sigma_\eta^2, \mathbf{Z}) \propto p(\boldsymbol{\theta}, \widetilde{\boldsymbol{\nu}}_w, \boldsymbol{\gamma}, \sigma_\eta^2, \mathbf{Z}) \propto p(\boldsymbol{\theta}|\sigma_\eta^2)h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})p(\widetilde{\boldsymbol{\nu}}_w|\boldsymbol{\theta})$, which completes the result.

Proof of Equation (12):

We have that

$$
\begin{aligned}
&p(\boldsymbol{\gamma}, \widetilde{\boldsymbol{\nu}}_O, \boldsymbol{\theta}, \sigma_\eta^2, \mathbf{Z})\\
&= \int\int \frac{h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma_\epsilon^2)p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2)p(\boldsymbol{\beta}|\sigma_\beta^2)p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta})p(\boldsymbol{\theta}|\sigma_\eta^2)p(\boldsymbol{\gamma})p(\sigma_\eta^2)h(\mathbf{Z}_{-w}|\mathbf{Z}_w)h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})}{\int\int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma_\epsilon^2)p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2)p(\boldsymbol{\beta}|\sigma_\beta^2)d\boldsymbol{\nu}_O d\boldsymbol{\beta}}\, d\boldsymbol{\nu}_O\, d\boldsymbol{\beta}\\
&= \left[\int\int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma_\epsilon^2)p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2)p(\boldsymbol{\beta}|\sigma_\beta^2)d\boldsymbol{\nu}_O d\boldsymbol{\beta}\right] \frac{p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta})p(\boldsymbol{\theta}|\sigma_\eta^2)p(\boldsymbol{\gamma})p(\sigma_\eta^2)h(\mathbf{Z}_{-w}|\mathbf{Z}_w)h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})}{\int\int h(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\nu}_O, \sigma_\epsilon^2)p(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta^2)p(\boldsymbol{\beta}|\sigma_\beta^2)d\boldsymbol{\nu}_O d\boldsymbol{\beta}}\\
&= p(\boldsymbol{\theta}|\sigma_\eta^2)h(\mathbf{Z}_{-w}|\mathbf{Z}_w)h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})p(\boldsymbol{\gamma})p(\widetilde{\boldsymbol{\nu}}_O|\boldsymbol{\theta})p(\sigma_\eta^2),
\end{aligned}
$$

so that $p(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\nu}}_O, \boldsymbol{\theta}, \sigma_\eta^2, \mathbf{Z}) \propto p(\boldsymbol{\gamma}, \widetilde{\boldsymbol{\nu}}_O, \boldsymbol{\theta}, \sigma_\eta^2, \mathbf{Z}) \propto p(\boldsymbol{\gamma})h(\mathbf{Z}_w|\boldsymbol{\theta}, \boldsymbol{\gamma})$, which completes the result.

Appendix B: Full Conditional Distributions

In this section, we outline the collapsed Gibbs sampler.

1. The full conditional distribution for $\boldsymbol{\beta}$ is

$$f(\boldsymbol{\beta}|\cdot) \propto \exp\left\{-\frac{(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\nu}_O)'\mathbf{V}_\epsilon^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\nu}_O)}{2}\right\}\exp\left\{-\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2\sigma_\beta^2}\right\}$$

$$\propto \exp\left\{-\frac{\boldsymbol{\beta}'(\mathbf{X}'\mathbf{V}_\epsilon^{-1}\mathbf{X}+\sigma_\beta^{-2}\mathbf{I}_p)\boldsymbol{\beta}}{2}+\boldsymbol{\beta}'\mathbf{X}'\mathbf{V}_\epsilon^{-1}(\mathbf{Z}-\boldsymbol{\nu}_O)\right\}$$

$$\propto \exp\left\{-\frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\beta}}{2}+\boldsymbol{\beta}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\mu}_*\right\}$$

$$\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_*)'\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_*)\right\}.$$

Thus, $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$, where $\boldsymbol{\mu}_* = \boldsymbol{\Sigma}_*^{-1}\boldsymbol{X}'\mathbf{V}_\epsilon^{-1}(\mathbf{Z}-\boldsymbol{\nu}_O)$ and $\boldsymbol{\Sigma}_*^{-1} = \boldsymbol{X}'\mathbf{V}_\epsilon^{-1}\mathbf{X}+\sigma_\beta^{-2}\mathbf{I}_p$, $\mathbf{X} = (\mathbf{x}(\mathbf{s}_1)\ldots\mathbf{x}(\mathbf{s}_n))'$.

2. From (10), we sample $\widetilde{\boldsymbol{\nu}}$ from $f(\widetilde{\boldsymbol{\nu}}|\boldsymbol{\theta})$ using the non-stationary spectral simulation method from Theorem 2. For example, when using the exponential covariogram, we simulate the elements of $\boldsymbol{\omega}_i$ from a $Cauchy(0, 1/\phi)$ and κ_i from a $U(-\pi, \pi)$ for $i = 1, \ldots, K$. Then for $j = 1, \ldots, N$ we compute $\boldsymbol{f}(\mathbf{u}_j) = \boldsymbol{\psi}'(\mathbf{u}_j)\boldsymbol{\eta}$, substitute into the expression of $\widetilde{\nu}(\mathbf{u}_j)$ according in (6), and stack over j to create $\widetilde{\boldsymbol{\nu}}$.

3. The full-conditional distribution is

$$f(\boldsymbol{\nu}_O|\cdot) \propto \exp\left\{-\frac{(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\nu}_O)'\mathbf{V}_\epsilon^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\nu}_O)}{2}\right\} f(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta, \boldsymbol{\theta}),$$

where $\mathbf{V}_\epsilon = \sigma_\epsilon^2$ and $f(\boldsymbol{\nu}_O|\widetilde{\boldsymbol{\nu}}_O, \delta, \boldsymbol{\theta}) \propto \exp\left\{-\frac{(\boldsymbol{\nu}_O-\widetilde{\boldsymbol{\nu}}_O)'(\boldsymbol{\nu}_O-\widetilde{\boldsymbol{\nu}}_O)}{2\delta^2}\right\}$. Then,

$$f(\boldsymbol{\nu}_O|\cdot) \propto \exp\left\{-\frac{\boldsymbol{\nu}_O'(\delta^{-2}\mathbf{I}+\mathbf{V}_\epsilon^{-1})\boldsymbol{\nu}_O}{2}+\boldsymbol{\nu}_O'\left(\mathbf{V}_\epsilon^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})+\frac{1}{\delta^2}\widetilde{\boldsymbol{\nu}}_O\right)\right\}.$$

This gives

$$f(\boldsymbol{\nu}|\cdot) = \mathrm{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*),$$

where $\boldsymbol{\mu}^* = \boldsymbol{\Sigma}^*\{\mathbf{V}_\epsilon^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})+\frac{1}{\delta^2}\widetilde{\boldsymbol{\nu}}_O\}$, and $(\boldsymbol{\Sigma}^*)^{-1} = \delta^{-2}\mathbf{I}+\mathbf{V}_\epsilon^{-1}$.

4. It follows from (11) that the full conditional distribution for $\boldsymbol{\eta}$, which collapses across $\widetilde{\boldsymbol{\nu}}_{-w}$, $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$, is given by

$$p(\boldsymbol{\eta}|\sigma_\nu^2, \phi, \widetilde{\boldsymbol{\nu}}_w, \boldsymbol{\gamma}, \sigma_\eta^2, \mathbf{Z}) \propto h(\mathbf{Z}_w|\boldsymbol{\theta}, \gamma)p(\widetilde{\boldsymbol{\nu}}_w|\boldsymbol{\theta})p(\boldsymbol{\eta}|\sigma_\eta^2).$$

We use Metropolis-Hasting to sample $\boldsymbol{\eta}$ and we use a multivariate normal distribution for the proposal distribution.

5. It follows from (11) that the full conditional distribution for σ_ν^2, which collapses across $\widetilde{\boldsymbol{\nu}}_{-w}$, $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$, is given by

$$p(\sigma_\nu^2|\boldsymbol{\eta}, \phi, \widetilde{\boldsymbol{\nu}}_w, \boldsymbol{\gamma}, \sigma_\eta^2, \mathbf{Z}) \propto p(\sigma_\nu^2)h(\mathbf{Z}_w|\boldsymbol{\theta}, \gamma)p(\widetilde{\boldsymbol{\nu}}_w|\boldsymbol{\theta}),$$

where $\alpha_1 = \beta_1 = 1$. We use Metropolis-Hasting to sample σ_ν^2 and we use an inverse gamma distribution for the proposal distribution. There are three contributions to the full-conditional distribution for σ_ν^2. The first multiplicative term is the prior distribution, the second term represents the likelihood after one collapses across $\boldsymbol{\nu}$ and $\boldsymbol{\beta}$, and the third term arises from collapsing across $\widetilde{\boldsymbol{\nu}}_{-w}$. Notice in our expression that the vectors $\mathbf{Z}_w$ and $\widetilde{\boldsymbol{\nu}}_w$ are w-dimensional, where w is small in value.

6. It follows from (12) that the full conditional distribution for σ_β^2, which collapses across $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$, is given by

$$p(\sigma_\beta^2|\boldsymbol{\theta}, \delta^2, \sigma_\epsilon^2, \sigma_\eta^2, \mathbf{Z}) \propto p(\sigma_\beta^2)h(\mathbf{Z}_w|\boldsymbol{\theta}, \gamma).$$

We use Metropolis-Hasting to sample σ_β^2 and we use an inverse gamma distribution for the proposal distribution.

7. The full conditional distribution of σ_η^2 is easily obtained and given by,

$$f(\sigma_\eta^2|\cdot) \propto (\sigma_\eta^2)^{-(\alpha_3+\frac{r}{2})-1}\exp\left\{-\frac{\boldsymbol{\eta}'\boldsymbol{\eta}}{2\sigma_\eta^2} + \sigma_\eta^{-2}\beta_3\right\},$$

which is an inverse gamma distribution with shape parameter $\alpha_3 + \frac{r}{2}$ and scale parameter $\frac{\boldsymbol{\eta}'\boldsymbol{\eta}}{2} + \beta_3$.

8. The prior for ϕ is Uniform Distribution$(0, U)$. Then, it follows from (11) that the full conditional distribution for σ_{ν}^2, which collapses across $\widetilde{\boldsymbol{\nu}}_{-w}$, $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$, is given by

$$p(\phi|\sigma_{\nu}^2, \boldsymbol{\eta}, \widetilde{\boldsymbol{\nu}}_w, \boldsymbol{\gamma}, \sigma_{\eta}^2, \mathbf{Z}) \propto I(0 \leq \phi \leq U) h(\mathbf{Z}_w|\boldsymbol{\theta}, \gamma) p(\widetilde{\boldsymbol{\nu}}_w|\boldsymbol{\theta}),$$

where $I(\cdot)$ is the indicator function. We use Metropolis-Hasting to sample ϕ and we use a uniform distribution for the proposal distribution.

9. It follows from (12) that the full conditional distribution for δ^2, which collapses across $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$, is given by

$$p(\delta^2|\boldsymbol{\theta}, \sigma_{\beta}^2, \sigma_{\epsilon}^2, \sigma_{\eta}^2, \mathbf{Z}) \propto h(\mathbf{Z}_w|\boldsymbol{\theta}, \gamma) p(\delta^2).$$

We use Metropolis-Hasting to sample δ^2 and we use an inverse gamma distribution for the proposal distribution.

10. We assume σ_{ϵ}^2 is known, but provide details on updating this parameter if an informative IG prior is preferred. In particular, it follows from (12) that the full conditional distribution for σ_{ϵ}^2, which collapses across $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$, is given by

$$p(\sigma_{\epsilon}^2|\boldsymbol{\theta}, \sigma_{\beta}^2, \delta^2, \sigma_{\eta}^2, \mathbf{Z}) \propto h(\mathbf{Z}_w|\boldsymbol{\theta}, \gamma) p(\sigma_{\epsilon}^2).$$

We can use a Metropolis-Hasting step to sample σ_{ϵ}^2 and we use an inverse gamma distribution for the proposal distribution.

Appendix C: Review of Spatial Process Convolution

The SPC model for a spatial random process $Y(\cdot)$ is defined via a kernel convolution:

$$Y(\mathbf{s}) = \int_D g(\mathbf{s}, \mathbf{s} - \mathbf{u}) dw(\mathbf{u}) \tag{20}$$

where $w(\cdot)$ is a spatial random process defined on $D \subset \mathbb{R}^d$ and g is referred to as a "kernel function." These kernel functions require the first and second moments to be finite, or $\int_{\mathbb{R}^d} g(\mathbf{s}, \mathbf{u}) d\mathbf{u} < \infty$ and $\int_{\mathbb{R}^d} g(\mathbf{s}, \mathbf{u}) d\mathbf{u} < \infty$ for every $\mathbf{s} \in D$.

Higdon (1998) develops a nonstationary process specification of $Y(\cdot)$ in this framework by specifying $w(\cdot)$ to be a white-noise process. Paciorek and Schervish (2006) specify the SPC model with $\mathbf{g}(\mathbf{s},\cdot)$ set equal to to be a multivariate normal density with mean $\mathbf{s}$. They show that the nonstationary covariance function has the form

$$cov(Y(\mathbf{s}_i), Y(\mathbf{s}_j)) =$$
$$\sigma^2 \mathrm{det}\left(\frac{\Sigma(\mathbf{s}_i)+\Sigma(\mathbf{s}_j)}{2}\right)^{-1/2} \det\left(\Sigma(\mathbf{s}_i)\right)^{1/4} \det\left(\Sigma(\mathbf{s}_j)\right)^{1/4} \exp\left\{-\frac{1}{2}\left(\mathbf{s}_i-\mathbf{s}_j\right)'\left(\frac{\Sigma(\mathbf{s}_i)+\Sigma(\mathbf{s}_j)}{2}\right)^{-1}\left(\mathbf{s}_i-\mathbf{s}_j\right)\right\}.$$

where $\sigma^2 > 0$ and $\mathbf{s}_i, \mathbf{s}_j \in D$. Risser and Calder (2015) extends this approach by introducing "mixture components." Specifically, let $\{\mathbf{b}_k : k = 1,\ldots,L\} \in D$ with mixture covariance matrices Σ_k. Then, Risser and Calder (2015) let

$$\Sigma(\mathbf{s}) = \sum_{k=1}^{L} h_k(\mathbf{s})\Sigma_k \tag{21}$$

where

$$h_k(\mathbf{s}) = \exp\left\{\frac{-\left\|\mathbf{s}-\mathbf{b}_k\right\|^2}{2\lambda_w}\right\}. \tag{22}$$

where $\sum_{k=1}^{L} w_k(\mathbf{s}) = 1$, and λ_w is a tuning parameter and it is considered fixed. We used an R-package **convoSPAT** to simulate data from SPC model. The covariance matrices Σ_k are defined through a spectral decomposition

$$\mathbf{\Sigma}_k = \begin{bmatrix}\cos(\eta_k) & -\sin(\eta_k)\\ \sin(\eta_k) & \cos(\eta_k)\end{bmatrix}\begin{bmatrix}\lambda_{1k} & 0\\ 0 & \lambda_{2k}\end{bmatrix}\begin{bmatrix}\cos(\eta_k) & \sin(\eta_k)\\ -\sin(\eta_k) & \cos(\eta_k)\end{bmatrix} \tag{23}$$

where $\lambda_{1k} > 0$ and $\lambda_{2k} > 0$ are log-linear regression coefficents, and η_k is real-valued. We use the R-code function **f_mc_kernels** to calculate these mixture component matrices. Next, the setting for mixture component kernel matrices are the following. We define the number of mixture components points equal $L = 5$. We define $(\lambda_{11},\ldots,\lambda_{15})' = (-1.3, 0.5, -0.6, 0.2, -0.1)$, $(\lambda_{21},\ldots,\lambda_{25})' = (-1.4, -0.1, 0.2, 1, -1)$, and $(\eta_1,\ldots,\eta_5) = (0, -0.15, 0.15, 0.1, 0.12)$. The true value for a constant mean, nugget effect and process variance σ^2 are 4, 0.1, and 1, respectively. The function **NSconvo_sim** is used to generate data from this model. This specification is used when $d = 1$ in Simulation Study 1 and $d = 2$ in Simulation Study 2.

References

Anderes, E. B. and Stein, M. L. (2011). "Local likelihood estimation for nonstationary random fields." *Journal of Multivariate Analysis*, 102, 506–520.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical modeling and analysis for spatial data*. Boca Raton, FL, CRC Press.

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.

Bochner, S. (1959). *Lectures on Fourier Integrals*. Princeton, NY: Princeton University Press.

Bornn, L., Pillai, N. S., Smith, A., and Woodard, D. (2017). "The use of a single pseudo-sample in approximate Bayesian computation." *Statistics and Computing*, 27, 3, 583–590.

Bornn, L., Shaddick, G., and Zidek, J. V. (2012). "Modeling nonstationary processes through dimension expansion." *Journal of the American Statistical Association*, 107, 281–289.

Bradley, J. R. (2019). "What is the best predictor that you can compute in five minutes using a given Bayesian hierarchical model?" *arXiv preprint arXiv:1912.04542*.

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.

Bradley, J. R., Cressie, N., Shi, T., et al. (2016). "A comparison of spatial predictors when datasets could be very large." *Statistics Surveys*, 10, 100–131.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2020). "Hierarchical models for spatial data with errors that are correlated with the latent process." *Statistica Sinica*, 30, 81–109.

Castruccio, S. and Guinness, J. (2017). "An evolutionary spectrum approach to incorporate large-scale geographical descriptors on global processes." *Journal of the Royal Statistical Society: Series C*, 66, 329–344.

Cressie, N. (1993). "Statistics for spatial data."

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

Cressie, N. and Wikle, C. K. (2011). "Statistics for Spatio-Temporal Data."

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016a). "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets." *Journal of the American Statistical Association*, 111, 800–812.

— (2016b). "On nearest-neighbor Gaussian process models for massive spatial data." *Wiley Interdisciplinary Reviews: Computational Statistics*, 8, 162–171.

Donoho, D. L., Mallat, S., and Sachs, R. (1996). "Estimating covariances of locally stationary processes: Consistency of best basis methods." In *Proceedings of Third International Symposium on Time-Frequency and Time-Scale Analysis (TFTS-96)*, 337–340. IEEE.

Finley, A., Datta, A., and Banerjee, S. (2020). "Package 'spNNGP'." *R package Version 0.1.4*.

Finley, A. O., Banerjee, S., and Gelfand, A. E. (2013). "spBayes for large univariate and multivariate point-referenced spatio-temporal data models." *Journal of Statistical Software*, 63, 1–28.

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

Friedman, J. H. et al. (1991). "Multivariate adaptive regression splines." *The Annals of Statistics*, 19, 1–67.

Fuentes, M. (2002). "Spectral methods for nonstationary spatial processes." *Biometrika*, 89, 197–210.

Fuentes, M., Chen, L., and Davis, J. M. (2008). "A class of nonseparable and nonstationary spatial temporal covariance functions." *Environmetrics*, 19, 487–507.

Fuentes, M. and Smith, R. L. (2001). "A new class of nonstationary spatial models." Tech. rep., unpublished manuscript, available at www.stat.ncsu.edu/information/library/papers/nonstat.ps.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2015). "Interpretable priors for hyperparameters for Gaussian random fields." *arXiv preprint arXiv:1503.00256*.

Gelfand, A. E. (2000). "Gibbs sampling." *Journal of the American Statistical Association*, 95, 1300–1304.

Gelfand, A. E. and Smith, A. F. (1990). "Sampling-based approaches to calculating marginal densities." *Journal of the American statistical association*, 85, 410, 398–409.

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1, 515–534.

Gramacy, R. B. and Apley, D. W. (2015). "Local Gaussian process approximation for large computer experiments." *Journal of Computational and Graphical Statistics*, 24, 2, 561–578.

Guinness, J. and Katzfuss, M. (2018). "GpGp: fast Gaussian process computation using Vecchia's Approximation." *R package version 0.1. 0*.

Guinness, J. and Stein, M. L. (2013). "Interpolation of nonstationary high frequency spatial–temporal temperature data." *The Annals of Applied Statistics*, 7, 1684–1708.

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

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., and Katzfuss, M. (2019). "A case study competition among methods for analyzing large spatial data." *Journal of Agricultural, Biological and Environmental Statistics*, 24, 398–425.

Higdon, D. (1998). "A process-convolution approach to modelling temperatures in the North Atlantic Ocean." *Environmental and Ecological Statistics*, 5, 2, 173–190.

Higdon, D., Swall, J., and Kern, J. (1999). "Non-stationary spatial modeling." *Bayesian statistics*, 6, 1, 761–768.

Horrell, M. T. and Stein, M. L. (2017). "Half-spectral space–time covariance models." *Spatial Statistics*, 19, 90–100.

Im, H. K., Stein, M. L., and Zhu, Z. (2007). "Semiparametric estimation of spectral density with irregular observations." *Journal of the American Statistical Association*, 102, 478, 726–735.

Kalkhan, M. A. (2011). *Spatial statistics: geospatial information modeling and thematic mapping*. Boca Raton, FL, CRC Press.

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

Katzfuss, M. (2013). "Bayesian nonstationary spatial modeling for very large datasets." *Environmetrics*, 24, 3, 189–200.

— (2017). "A multi-resolution approximation for massive spatial datasets." *Journal of the American Statistical Association*, 112, 517, 201–214.

Katzfuss, M. and Guinness, J. (2019). "A general framework for Vecchia approximations of Gaussian processes." *arXiv preprint arXiv:1708.06302*.

Katzfuss, M., Guinness, J., Gong, W., and Zilber, D. (2020). "Vecchia approximations of Gaussian-process predictions." *Journal of Agricultural, Biological and Environmental Statistics*, 1–32.

Kleiber, W. and Nychka, D. (2012). "Nonstationary modeling for multivariate spatial processes." *Journal of Multivariate Analysis*, 112, 76–91.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. CRC Press.

Lindgren, F. and Rue, H. (2015). "Bayesian spatial modelling with R-INLA." *Journal of Statistical Software*, 63, 1–25.

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, 4, 423–498.

Liu, J. S. (1994). "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem." *Journal of the American Statistical Association*, 89, 958–966.

Martin, W. (1982). "Time-frequency analysis of random signals." In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, vol. 7, 1325–1328. IEEE.

Mejía, J. M. and Rodríguez-Iturbe, I. (1974). "On the synthesis of random field sampling from the spectrum: An application to the generation of hydrologic spatial processes." *Water Resources Research*, 10, 705–711.

Neal, R. M. (2003). "Slice sampling." *Annals of statistics*, 31, 705–741.

Neto, J. H. V., Schmidt, A. M., and Guttorp, P. (2014). "Accounting for spatially varying directional effects in spatial covariance structures." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63, 103–122.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). "A multiresolution Gaussian process model for the analysis of large spatial datasets." *Journal of Computational and Graphical Statistics*, 24, 2, 579–599.

Nychka, D. W. (2001). "Spatial process estimates as smoothers." In *Smoothing and Regression: Approaches, Computation and Applications, rev. ed*, ed. M. G. Schmiek, 393–424. New York, NY: Wiley.

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

Perrin, O. and Meiring, W. (2003). "Nonstationarity in Rn is second-order stationarity in R2n." *Journal of Applied Probability*, 40, 815–820.

Priestley, M. B. (1965). "Evolutionary spectra and non-stationary processes." *Journal of the Royal Statistical Society. Series B*, 27, 204–237.

Prokhorov, Y. V. (1956). "Convergence of random processes and limit theorems in probability theory." *Theory of Probability and Its Applications*, 1, 157–214.

Pronzato, L. and Rendas, M. (2017). "Bayesian local kriging." *Technometrics*, 59, 3, 293–304.

Quinn, T. (2017). "peakRAM: Monitor the total and peak RAM used by an expression or function." *R package Version 1.0.2*.

Ravishanker, N. and Dey, D. K. (2020). *A first course in linear model theory*. CRC Press.

Resnick, S. I. (2013). *A probability path*. Berlin Germany: Springer Science & Business Media.

Risser, M. D. (2016). "Nonstationary spatial modeling, with emphasis on process convolution and covariate-driven approaches." *arXiv preprint arXiv:1610.02447*.

Risser, M. D. and Calder, C. A. (2015). "Local likelihood estimation for covariance functions with spatially-varying parameters: the convoSPAT package for R." *arXiv preprint arXiv:1507.08613*.

Robert, C. P. (2004). *Monte carlo methods*. Hoboken, NJ: Wiley.

Sampson, P. D. and Guttorp, P. (1992). "Nonparametric estimation of nonstationary spatial covariance structure." *Journal of the American Statistical Association*, 87, 108–119.

Sang, H. and Huang, J. Z. (2012). "A full scale approximation of covariance functions for large spatial data sets." *Journal of the Royal Statistical Society: Series B*, 74, 111–132.

Sayeed, A. M. and Jones, D. L. (1995). "Optimal kernels for nonstationary spectral estimation." *IEEE Transactions on Signal Processing*, 43, 478–491.

Schmidt, A. M. and O'Hagan, A. (2003). "Bayesian inference for non-stationary spatial covariance structure via spatial deformations." *Journal of the Royal Statistical Society: Series B*, 65, 743–758.

Shand, L. and Li, B. (2017). "Modeling nonstationarity in space and time." *Biometrics*, 73, 759–768.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors." *Statistical Science*, 32, 1–28.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B*, 64, 583–639.

Stein, A., van der Meer, F. D., and Gorte, B. (2006). *Spatial statistics for remote sensing*. Berlin, Germany: Springer Science & Business Media.

Stein, M. L. (1999). *Statistical Interpolation of spatial Data: Some Theory for Kriging*. Place Springer.

— (2012). *Interpolation of spatial data: some theory for kriging*. Berlin Germany: Springer Science & Business Media.

— (2014). "Limitations on low rank approximations for covariance matrices of spatial data." *Spatial Statistics*, 8, 1–19.

— (2015). "When does the screening effect not hold?" *Spatial Statistics*, 11, 65–80.

Stein, M. L., Chi, Z., and Welty, L. J. (2004). "Approximating likelihoods for large spatial data sets." *Journal of the Royal Statistical Society: Series B*, 66, 275–296.

Tierney, L. (1994). "Markov chains for exploring posterior distributions." *the Annals of Statistics*, 1701–1728.

Tzeng, S. and Huang, H.-C. (2018). "Resolution adaptive fixed rank kriging." *Technometrics*, 60, 198–208.

Zammit-Mangion, A. and Cressie, N. (2017). "FRK: An R package for spatial and spatio-temporal prediction with large datasets." *arXiv preprint arXiv:1705.08105*.

Zhang, B., Sang, H., and Huang, J. Z. (2019). "Smoothed full-scale approximation of Gaussian process models for computation of large spatial datasets." *Statistica Sinica*, 29, 1711–1737.

Ziemke, J., Chandra, S., and Bhartia, P. (2005). "A 25-year data record of atmospheric ozone in the Pacific from Total Ozone Mapping Spectrometer (TOMS) cloud slicing: Implications for ozone trends in the stratosphere and troposphere." *Journal of Geophysical Research: Atmospheres*, 110, 1–15.