# FULL-SCALE APPROXIMATIONS OF SPATIO-TEMPORAL COVARIANCE MODELS FOR LARGE DATASETS

Bohai Zhang, Huiyan Sang and Jianhua Z. Huang

*Texas A&M University*

*Abstract:* Various continuously-indexed spatio-temporal process models have been constructed to characterize spatio-temporal dependence structures, but the computational complexity for model fitting and predictions grows in a cubic order with the size of dataset and application of such models is not feasible for large datasets. This article extends the full-scale approximation (FSA) approach by Sang and Huang (2012) to the spatio-temporal context to reduce computational complexity. A reversible jump Markov chain Monte Carlo (RJMCMC) algorithm is proposed to select knots automatically from a discrete set of spatio-temporal points. Our approach is applicable to nonseparable and nonstationary spatio-temporal covariance models. We illustrate the effectiveness of our method through simulation experiments and application to an ozone measurement dataset.

*Key words and phrases:* Covariance approximation, Gaussian process, knot selection, reversible jump Markov chain Monte Carlo, sparse matrix.

## 1. Introduction

Spatio-temporal datasets arising from climatology, geology, and other scientific fields have generated considerable interests in statistical modeling. The primary interests in analyzing such data are to detect meaningful spatio-temporal dependence patterns, and to subsequently smooth and predict in space-time domain.

Recent developments are mainly in spatio-temporal process models. We focus on a paradigm in which both space and time are continuously indexed. A key ingredient is a valid spatio-temporal covariance model that characterizes spatio-temporal dependence structure. A simple but commonly used class of spatio-temporal covariance model assumes a separable form that factors into a purely spatial and a purely temporal component. However, separable models do not allow for space-time interaction and often fail to model a physical process adequately. There is a growing literature on methods for constructing more flexible spatio-temporal covariance functions. Cressie and Huang (1999) introduced several classes of nonseparable spatio-temporal covariance functions based on the spectral density of nonnegative finite measures. Gneiting (2002) extended their

work and introduced a broader class of nonseparable spatio-temporal covariance functions, which does not depend on closed forms of inverse Fourier transforms. Stein (2005) developed a class of asymmetric and nonstationary space-time covariance functions on the sphere.

Parameter estimation and spatio-temporal prediction with these models typically require $\mathcal{O}(n^3)$ operations for a spatio-temporal dataset of size $n$, imposing computational challenges. One approach to the computation seeks to simplify the model fitting method mainly through likelihood approximations. Composite likelihood (CL) methods (Varin and Vidoni (2005)) have been applied to deal with spatial and spatio-temporal datasets due to their simplicity and sound asymptotic properties. The idea is to use a pseudo-likelihood by combining likelihood objects constructed from conditional or marginal models as a surrogate for the ordinary likelihood. Recently, Bevilacqua et al. (2012) introduced a weighted composite likelihood (WCL) method based on pairwise differences of spatio-temporal observations. They showed that the estimators of their methods are consistent and asymptotic normal under increasing domain asymptotics (Cressie (1993)). Bai, Song, and Raghunathan (2012) also developed a CL method based on pairwise differences, forming a joint estimation function based on spatial, temporal and spatio-temporal group-based estimation functions. A second approach seeks to simplify model specifications of covariance structures to achieve computational efficiency. Many literature following this path have emerged but primarily focusing on spatial or spatial discrete-time contexts (Furrer, Genton, and Nychka (2006); Rue and Held (2005); Cressie and Johannesson (2008); Katzfuss and Cressie (2011)). Banerjee et al. (2008) proposed a class of spatial predictive processes models that is applicable to fitting spatio-temporal process models with large data sets. The idea of this reduced rank type of approach is to approximate a spatio-temporal process with a predictive process, the prediction of a given spatio-temporal process conditional on the random spatio-temporal vector at a selected location set of reduced size.

Recently, Sang and Huang (2012) developed a covariance approximation method, referred to as *full-scale approximation* (FSA), for the implementation of univariate spatial process models with large datasets. Combining merits of reduced rank techniques (Cressie and Johannesson (2008); Banerjee et al. (2008)) and sparse matrix algorithms (Furrer, Genton, and Nychka (2006)), the FSA approach provides a high quality approximation to the covariance function at both large and small spatial scales and achieves computational efficiency.

We extend the FSA approach to the spatio-temporal process context. We propose a general-purpose full-scale approximation that can approximate well the original covariance function at both large and small spatio-temporal scales. Here the first step produces a reduced rank spatio-temporal covariance that is

effective in capturing large-scale spatio-temporal dependence and the second step a sparse covariance matrix that can approximate well the small-scale spatio-temporal dependence unexplained by the first part. Our method yields a new full-scale spatio-temporal covariance function for any given covariance function that maintains the flexibility and the richness of spatio-temporal structure while substantially reducing computational cost. Spatio-temporal predictions of the full-scale covariance approximation models can be carried out efficiently following the conventional prediction procedure in Gaussian processes.

The method requires careful selection of knots in the reduced rank step, an issue not addressed by Sang and Huang (2012), to achieve a good balance between model fitting and computational time. We propose an adaptive and automatic way to select both the knot number and knot locations by treating them as random variables. We consider selecting knots either from a set containing all spatio-temporally observed locations or a fine grid covering entire space-time domain following a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm (Green (1995)).

The rest of the article is organized as follows. In Section 2, we present the FSA approach for spatio-temporal processes and detail the Bayesian implementation of the FSA approach, including knot selection. Section 3 presents results of some simulation studies to investigate performance of the method and to compare it with existing methods. In Section 4, we apply our method to a dataset of the maximum 8-hour ozone measurements across the eastern US. Section 5 gives some concluding remarks.

## 2. The FSA Approach

### 2.1. Model

Let $Y(\mathbf{s}, t)$ be a response variable observed at location $\mathbf{s}$ and time $t$, where $\mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^2$, $t \in [0,\ T] \subseteq \mathbb{R}^+$. A general formulation of spatio-temporal process model is

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + w(\mathbf{s}, t) + \epsilon(\mathbf{s}, t), \tag{2.1}$$

where $\mu(\mathbf{s}, t)$ is a deterministic mean function, $w(\mathbf{s}, t)$ is a zero-mean Gaussian process characterizing spatio-temporal variations, and $\epsilon(\mathbf{s}, t)$ is a Gaussian white noise process independent of $w(\mathbf{s}, t)$. The variance of $\epsilon(\mathbf{s}, t)$ is usually assumed to be a constant $\tau^2$, called the "nugget effect", to account for measurement errors. The spatio-temporal dependence structure of $w(\mathbf{s}, t)$ is specified by a spatio-temporal covariance function $\Gamma_w(\mathbf{s}, t; \mathbf{s}', t') \equiv \text{Cov}(w(\mathbf{s}, t), w(\mathbf{s}', t'))$. In the spatio-temporal regression framework, we assume $\mu(\mathbf{s}, t) = Z^T(\mathbf{s}, t)\boldsymbol{\beta}$, where $Z(\mathbf{s}, t)$ is a $p \times 1$ column vector of space-time covariates and $\boldsymbol{\beta}$ is the associated coefficient vector.

To simplify notation, denote a spatio-temporal point $\mathbf{x}$ by $(\mathbf{s}, t)$. Let $\mathscr{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ be a collection of all observed spatio-temporal points, $\mathbf{w} = (w(\mathbf{x}_1), w(\mathbf{x}_2), \cdots, w(\mathbf{x}_n))^T$ and $\boldsymbol{\epsilon} = (\epsilon(\mathbf{x}_1), \epsilon(\mathbf{x}_2), \cdots, \epsilon(\mathbf{x}_n))^T$. It follows that $\mathbf{w} \sim \text{MVN}(\mathbf{0}, \Sigma_\mathbf{w})$ and $\boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, \Sigma_{\boldsymbol{\epsilon}})$, where MVN stands for the multivariate normal distribution, $\Sigma_\mathbf{w} = [\Gamma_\mathbf{w}(\mathbf{x}_i, \mathbf{x}_j)]_{i=1:n, j=1:n}$, and $\Sigma_{\boldsymbol{\epsilon}} = \tau^2 \mathbf{I}_n$.

The marginal distribution of $\mathbb{Y} = (Y(\mathbf{x}_1), \cdots, Y(\mathbf{x}_n))^T \sim \text{MVN}(\mathbb{Z}\boldsymbol{\beta}, \Sigma_\mathbf{w} + \tau^2 \mathbf{I}_n)$, where $\mathbb{Z} = (Z(\mathbf{x}_1), Z(\mathbf{x}_2), \cdots, Z(\mathbf{x}_n))^T$. To make likelihood-based or Bayesian inferences, we need to evaluate the likelihood of $\mathbb{Y}$; this requires $\mathcal{O}(n^3)$ computational time due to the inversion of $\Sigma_\mathbf{w} + \tau^2 \mathbf{I}_n$. A similar computational bottleneck is encountered when performing spatio-temporal prediction.

## 2.2. Covariance approximation for large computation of spatio-temporal process

We propose a full-scale covariance approximation for the latent spatio-temporal process $w$ by a sum of two independent processes,

$$w^\dagger(\mathbf{x}) = w_l(\mathbf{x}) + w_s(\mathbf{x}), \tag{2.2}$$

where $w_l(\mathbf{x})$ is a low-rank process that captures the large-scale spatio-temporal dependence structure and $w_s(\mathbf{x})$ is a residual process that models the small-scale spatio-temporal dependence not captured by $w_l(\mathbf{x})$. We model the low rank process using a predictive process on spatio-temporal domain. The predictive process, proposed by Banerjee et al. (2008), has been shown to be flexible enough to model the large-scale dependence structure of a spatial process. Given a set of points $\mathscr{X}^* = \{\mathbf{x}_1^*, \mathbf{x}_2^*, \cdots, \mathbf{x}_m^*\}$, called spatio-temporal knots, the spatio-temporal predictive process is

$$w_l(\mathbf{x}) = \mathcal{C}(\mathbf{x}, \mathscr{X}^*)\mathcal{C}^{*-1}\mathbf{w}^*,$$

where $\mathbf{w}^* = (w(\mathbf{x}_1^*), w(\mathbf{x}_2^*), \cdots, w(\mathbf{x}_m^*))^T$, $\mathcal{C}(\mathbf{x}, \mathscr{X}^*) = [\Gamma_w(\mathbf{x}, \mathbf{x}_j^*)]_{j=1:m}$, and $\mathcal{C}^* = \mathcal{C}(\mathscr{X}^*, \mathscr{X}^*)$ is the covariance matrix of $\mathbf{w}^*$. It follows that the covariance function of $w_l$ is given by

$$\Gamma_{w_l}(\mathbf{x}, \mathbf{x}') = \mathcal{C}(\mathbf{x}, \mathscr{X}^*)\mathcal{C}^{*-1}\mathcal{C}^T(\mathbf{x}', \mathscr{X}^*). \tag{2.3}$$

The covariance matrix of $w_l$ evaluated at $\mathscr{X}$ is $\Sigma_{w_l} = \mathcal{C}(\mathscr{X}, \mathscr{X}^*)\mathcal{C}^{*-1}\mathcal{C}^T(\mathscr{X}, \mathscr{X}^*)$ where $\mathcal{C}(\mathscr{X}, \mathscr{X}^*) = [\Gamma_w(\mathbf{x}_i, \mathbf{x}_j^*)]_{i=1:n, j=1:m}$. One often chooses $m \ll n$, which results in a low-rank matrix $\Sigma_{w_l}$ and hence leads to efficient computations. If the knot set is chosen to be $\mathscr{X}$, the original spatio-temporal covariance is fully recovered.

The residual process $w_s(\mathbf{x})$ is an important supplement to $w_l(\mathbf{x})$ for better approximating the original process $w(\mathbf{x})$, while maintaining computational efficiency. The idea is to use a sparse covariance matrix to approximate the covariance of the exact residual process $w(\mathbf{x}) - w_l(\mathbf{x})$. The covariance function of

$w(\mathbf{x}) - w_l(\mathbf{x})$ is $\Gamma_w(\mathbf{x}, \mathbf{x}') - \Gamma_{w_l}(\mathbf{x}, \mathbf{x}')$, and we take the covariance function of $w_s(\mathbf{x})$ to be

$$\Gamma_{w_s}(\mathbf{x}, \mathbf{x}') = \{\Gamma_w(\mathbf{x}, \mathbf{x}') - \Gamma_{w_l}(\mathbf{x}, \mathbf{x}')\}\mathcal{K}(\mathbf{x}, \mathbf{x}'), \tag{2.4}$$

where $\mathcal{K}(\mathbf{x}, \mathbf{x}')$, referred to as a modulating function, is chosen to ensure $\Gamma_{w_s}$ is a valid positive semidefinite function and that is zero for a large proportion of possible spatio-temporal pairs $(\mathbf{x}, \mathbf{x}')$ evaluated at $\mathscr{X}$. The choice of $\mathcal{K}(\mathbf{x}, \mathbf{x}')$ ensures that the resulting residual covariance matrix $\Sigma_{w_s}$ can be handled efficiently.

We describe several strategies for specifying the modulating function $\mathcal{K}$. The first is to use a tapering function, the result is referred to as *FSA-Taper*, which sets the correlation of distant spatio-temporal pairs to zero. In the univariate spatial case, a number of compactly supported covariance functions have been used for covariance tapering, for example the spherical covariance function, the family of Wendland covariance functions, and the bisquare function, to name a few (Wendland (1995, 1998); Gneiting (2002); Cressie and Johannesson (2008)). In the spatio-temporal context, we consider tapering functions as Schur products of a purely spatial and a purely temporal tapering function. Let $\mathcal{K}_u(\mathbf{s}, \mathbf{s}'; \gamma_u)$ be a tapering function on the spatial domain satisfying $\mathcal{K}_u = 0$ when $\|\mathbf{s} - \mathbf{s}'\| > \gamma_u$, and $\mathcal{K}_v(t, t'; \gamma_v)$ be a tapering function on the temporal domain such that $\mathcal{K}_v = 0$ when $\|t - t'\| > \gamma_v$. Here, $\gamma_u$ and $\gamma_v$ are referred to as the spatial taper range and the temporal taper range, respectively. Then $\mathcal{K}(\mathbf{s}, t; \mathbf{s}', t') = \mathcal{K}_u(\mathbf{s}, \mathbf{s}'; \gamma_u)\mathcal{K}_v(t, t'; \gamma_v)$.

A second specification of $\mathcal{K}$ uses local partitioning: residuals are assumed to be independent across partitioned space-time subregions, while the original residual covariance is preserved within each subregion. Let $B_1, B_2, \cdots, B_K$ be a partition of the space-time domain $\mathcal{S} \times [0, \ T]$, referred to as blocks. Then the modulating function is

$$\mathcal{K}_{block}(\mathbf{x}, \mathbf{x}') = \begin{cases} 1 \text{ if } \mathbf{x}, \mathbf{x}' \in B_i, \ i = 1, \ldots, K; \\ 0 \text{ otherwise.} \end{cases}$$

By rearranging observation indices such that the observations within a block are grouped together, we obtain a block-diagonal modulating matrix $\mathcal{K}_{block}(\mathscr{X}, \mathscr{X})$ with $\mathbf{1}_{n_i}\mathbf{1}_{n_i}^T$ the $i$th block, where $\mathbf{1}_{n_i}$ is a column vector of 1s and $n_i$ is the number of observations within the $i$th block for $i = 1, 2, \ldots, K$. Thus the covariance matrix of the approximated residual process $w_s$ on $\mathscr{X}$ is also block-diagonal, whose inverse can be efficiently computed if the block size is not large. We refer the FSA approach with $\mathcal{K} = \mathcal{K}_{block}$ as the *FSA-Block* method.

Let $\Sigma_{w^\dagger}$ denote the covariance matrix of observations $\mathscr{X}$ given by the FSA-Block method. It is positive definite (PD) when the knot set does not overlap with the observation set, otherwise it is positive semidefinite (PSD). To see why, note that $\Sigma_{w^\dagger} = \Sigma_{w_l} + \Sigma_{w_s}$, where $\Sigma_{w_l}$ is the covariance matrix of the predictive process and $\Sigma_{w_s}$ is a residual covariance. Here $\Sigma_{w_s} = (\Sigma_w - \Sigma_{w_l}) \circ \mathcal{K}_{block}(\mathscr{X}, \mathscr{X})$,

where $\circ$ is the Schur product (entry-wise product) of matrices. Denote the observational locations in the block $B_k$ by $\mathscr{X}_k$. Then we obtain a block diagonal matrix $\Sigma_{w_s}$ with $\Sigma_{w_s}^k = \Gamma_w(\mathscr{X}_k, \mathscr{X}_k) - \Gamma_{w_l}(\mathscr{X}_k, \mathscr{X}_k)$ as its $k$th block. Since $\Sigma_{w_s}^k$ is the conditional covariance of $w(\mathscr{X}_k)$ given $w(\mathscr{X}^*)$, it is PD when $\mathscr{X}^* \cap \mathscr{X} = \emptyset$ and PSD otherwise. It follows that $\Sigma_{w_s}$ and $\Sigma_{w^\dagger}$ are PD when $\mathscr{X}^* \cap \mathscr{X} = \emptyset$ and PSD otherwise.

The reduced-rank part plus the residual part using local partitioning provides an exact recovery of the true covariance within each subregion. Specifically, the covariance function of $w^\dagger$ is

$$\Gamma_{w^\dagger}(\mathbf{x}, \mathbf{x}') = \begin{cases} \Gamma_w(\mathbf{x}, \mathbf{x}') & \text{if } \mathbf{x}, \mathbf{x}' \in B_i, \ i = 1, \ldots, K; \\ \Gamma_{w_l}(\mathbf{x}, \mathbf{x}') & \text{otherwise.} \end{cases} \tag{2.5}$$

As the covariance approximation errors induced by the FSA-Block only occur for pairs belonging to different subregions and most of these pairs some distance apart, the errors $\Gamma_w(\mathbf{x}, \mathbf{x}') - \Gamma_{w^\dagger}(\mathbf{x}, \mathbf{x}')$ are expected to be small for most pairs.

## 2.3. Fast computation of parameter estimation and spatio-temporal prediction using FSA

In this section, we show the implementation of a spatio-temporal regression model using the FSA method. Replacing the latent spatio-temporal process $w$ as (2.1) with its induced spatio-temporal FSA $w^\dagger$ as (2.2), we obtain the data model at $n$ observed locations,

$$\mathbb{Y} = \mathbb{Z}\boldsymbol{\beta} + \mathbf{w}^\dagger + \epsilon, \qquad \epsilon \sim \text{MVN}(\mathbf{0}, \Sigma_\epsilon), \tag{2.6}$$

where $\mathbf{w}^\dagger$ is an $n \times 1$ vector of $w^\dagger$ evaluated on $\mathscr{X}$. The data likelihood is then given by $\mathbb{Y} \sim \text{MVN}(\mathbb{Z}\boldsymbol{\beta}, \Sigma_{w_l} + \Sigma_{w_s} + \Sigma_\epsilon)$.

Here $\Sigma_{w_s} + \Sigma_\epsilon$ is a sparse matrix for FSA-Taper or a block diagonal matrix for FSA-Block, whose inversion can be handled efficiently. We apply the Sherman-Woodbury-Morrison formula to calculate the inverse of $\Sigma_\mathbb{Y}$

$$\begin{aligned} \Sigma_\mathbb{Y}^{-1} = {} & (\Sigma_{w_s} + \Sigma_\epsilon)^{-1} - (\Sigma_{w_s} + \Sigma_\epsilon)^{-1}\mathcal{C}(\mathscr{X}, \mathscr{X}^*) \\ & \times \{\mathcal{C}^* + \mathcal{C}^T(\mathscr{X}, \mathscr{X}^*)(\Sigma_{w_s} + \Sigma_\epsilon)^{-1}\mathcal{C}(\mathscr{X}, \mathscr{X}^*)\}^{-1} \\ & \times \mathcal{C}^T(\mathscr{X}, \mathscr{X}^*)(\Sigma_{w_s} + \Sigma_\epsilon)^{-1}. \end{aligned} \tag{2.7}$$

The determinant of $\Sigma_\mathbb{Y}$ can also be efficiently computed by applying Sylvester's determinant theorem,

$$\begin{aligned} |\Sigma_\mathbb{Y}| = {} & |\Sigma_{w_s} + \Sigma_\epsilon| \times |\mathcal{C}^*|^{-1} \\ & \times |\mathcal{C}^* + \mathcal{C}^T(\mathscr{X}, \mathscr{X}^*)(\Sigma_{w_s} + \Sigma_\epsilon)^{-1}\mathcal{C}(\mathscr{X}, \mathscr{X}^*)|. \end{aligned} \tag{2.8}$$

Likelihood-based inference uses maximum likelihood or restricted maximum likelihood. For Bayesian inference, we need to specify priors for model parameters. For the regression coefficient vector $\boldsymbol{\beta}$, we assign a vague multivariate normal prior $\boldsymbol{\beta} \sim \text{MVN}(\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$. For the variance of measurement errors $\tau^2$, we assign an inverse-gamma prior $\text{IG}(a, b)$ where the hyper-parameters $a, b$ are chosen with reasonable guesses of mean and variance. Denote the set of parameters in the spatio-temporal covariance function $\Gamma_w$ by $\boldsymbol{\theta}$, whose prior specification depends on the choice of the covariance function. Customarily, the inverse-gamma prior can be assigned on the variance parameter $\sigma^2$; the spatial/temporal range parameter can be assigned with a reasonably informative prior, e.g. a uniform prior with its support specified according to the belief on the practical spatial/temporal dependence range of the spatio-temporal dataset.

Let $\Omega = (\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2)$ be the collection of model parameters. The MCMC method is used to draw samples of parameters from the posterior

$$p(\Omega|\mathbb{Y}) \propto p(\boldsymbol{\beta})p(\boldsymbol{\theta})p(\tau^2)p(\mathbb{Y}|\Omega). \tag{2.9}$$

The Gibbs sampler is used to update $\boldsymbol{\beta}$ from $\text{MVN}(\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot}, \Sigma_{\boldsymbol{\beta}|\cdot})$, where

$$\Sigma_{\boldsymbol{\beta}|\cdot} = (\mathbb{Z}^T(\Sigma_{w_l} + \Sigma_{w_s} + \Sigma_{\boldsymbol{\epsilon}})^{-1}\mathbb{Z} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1},$$

$$\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot} = \Sigma_{\boldsymbol{\beta}|\cdot}(\mathbb{Z}^T(\Sigma_{w_l} + \Sigma_{w_s} + \Sigma_{\boldsymbol{\epsilon}})^{-1}\mathbb{Y} + \Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}).$$

For parameters without a closed-form of the full conditional distribution, we draw samples using the Metropolis-Hasting algorithm. For example, for spatial/temporal dependence range parameters, we can use truncated normal distribution centered at the current value as the proposal distribution. The log-normal proposal centered at the current value can also be used for dependence range parameters.

The spatio-temporal process regression model combined with the FSA provides a straightforward and efficient prediction using large spatio-temporal datasets. In classical geostatistics, assuming the model parameters are known, for a given new spatio-temporal point $\mathbf{x}_0$ the approximated best linear unbiased predictor (BLUP) of $Y(\mathbf{x}_0)$ is

$$\hat{Y}(\mathbf{x}_0) = Z^T(\mathbf{x}_0)\boldsymbol{\beta} + \mathcal{C}_{w^\dagger}(\mathbf{x}_0, \mathscr{X})\{\Sigma_{w_l} + \Sigma_{w_s} + \Sigma_{\boldsymbol{\epsilon}}\}^{-1}(\mathbb{Y} - \mathbb{Z}\boldsymbol{\beta}) \tag{2.10}$$

and the approximated mean square error (MSE) is

$$\text{MSE}(\hat{Y}(\mathbf{x}_0)) = \sigma^2 + \tau^2 - \mathcal{C}_{w^\dagger}(\mathbf{x}_0, \mathscr{X})\{\Sigma_{w_l} + \Sigma_{w_s} + \Sigma_{\boldsymbol{\epsilon}}\}^{-1}\mathcal{C}_{w^\dagger}^T(\mathbf{x}_0, \mathscr{X}), \tag{2.11}$$

where $\mathcal{C}_{w^\dagger}(\mathbf{x}_0, \mathscr{X}) = [\Gamma_{w_l}(\mathbf{x}_0, \mathbf{x}_i) + \Gamma_{w_s}(\mathbf{x}_0, \mathbf{x}_i)]_{i=1:n, \mathbf{x}_i \in \mathscr{X}}$ is a $1 \times n$ cross-covariance matrix between $w^\dagger(\mathbf{x}_0)$ and $\mathbf{w}$. In practice, data-based estimates of the parameters are plugged in the above expression.

The Bayesian approach generalizes to the case of prediction when the covariance parameters are unknown. The predictive distribution for $Y(\mathbf{x}_0)$ is a Gaussian distribution with predictive mean given by (2.10) and variance given by (2.11). Therefore, a random sample of $Y(\mathbf{x}_0)$ from the (posterior) predictive distribution can be obtained by a draw of $\Omega$ from the posterior followed by a draw from the conditional predictive distribution of $Y(\mathbf{x}_0)$ given $\Omega$.

Again, the calculation of BLUP and draws from the posterior predictive distribution involve the inversion of $\Sigma_{w_l} + \Sigma_{w_s} + \Sigma_{\boldsymbol{\epsilon}}$, which can be handled efficiently using the computational technique described in (2.7).

### 2.4. Selection of tuning parameters

Both the FSA-Taper and the FSA-Block involve tuning parameters: taper ranges and a knot set are required for the FSA-Taper; block partition in space-time domain and a knot set are required for the FSA-Block. The choices of these tuning parameters determine the approximation performance and the computational complexity of the FSA model.

Choice of knots is a key ingredient in the low rank component of the FSA. Typically, a denser knot design can lead to a better approximation of the parent process but at a cost of heavier computational burden. A heuristic way for selecting knots is to predetermine a knot number $m$ based on available computational resources, then to place knots with good space-time coverage. Possible options include random sampling, Latin hypercube sampling (McKay, Conover, and Beckman (1979); Stein (1987)) and using a regular grid. Alternatively, one may consider a random knot selection in which knot number $m$ and their locations are allowed to be chosen automatically.

For random knot selection, Guhaniyogi et al. (2011) introduced an adaptive predictive process model for spatial data. They fixed the knot number and modeled knot locations with a point pattern model. Katzfuss (2013) applied the FSA-Taper approach to a nonstationary Matérn covariance function for spatial process, where the knot number was assigned with an improper flat prior on the set of all positive integers and knot locations were assigned with a uniform prior over the whole spatial domain.

Motivated by this work, we propose a Bayesian approach to adaptively select knot number and knot locations for the spatio-temporal FSA method. A RJMCMC algorithm (Green (1995)) is offered to update the knot set from a discrete set of spatio-temporal points. Choices of candidate set include the set of all observed points or a regular grid covering the entire space-time domain, denoted by $\bar{\mathcal{L}}$. Let $m$ be the knot number and $\mathcal{L}$ be the set of selected knot locations. We propose to assign the knot number $m$ with a Poisson($\lambda$) prior truncated at $\lambda_0$, where $\lambda$ is chosen to balance the trade-off between computational capacity

and model fitting, and $\lambda_0 > 0$ is set to reflect the maximum tolerance of computational time. Conditional on the knot number, we assume knots are randomly chosen from the candidate knot set, $p(\mathcal{L}|m) = \binom{M}{m}^{-1}$, where $M$ is the size of $\bar{\mathcal{L}}$.

At each MCMC step, we consider three types of possible moves of selected knot set, changing from $(\mathcal{L}, m) \to (\mathcal{L}^*, m^*)$: (a) *birth*: add a knot by randomly selecting a point in $\bar{\mathcal{L}} \backslash \mathcal{L}$, so $m^* = m + 1$; (b) *death*: randomly delete a knot in $\mathcal{L}$, so $m^* = m - 1$; and (c) *change*: randomly choose a knot from $\mathcal{L}$ and then replace it with a randomly chosen point from $\bar{\mathcal{L}} \backslash \mathcal{L}$, so $m^* = m$. The acceptance ratio $\alpha$ of proposing a move is given by

$$\alpha = \min\left(1, \frac{p(\mathbb{Y}|\Omega, m^*, \mathcal{L}^*)p(\mathcal{L}^*|m^*)p(m^*)J((\mathcal{L}^*, m^*) \to (\mathcal{L}, m))}{p(\mathbb{Y}|\Omega, m, \mathcal{L})p(\mathcal{L}|m)p(m)J((\mathcal{L}, m) \to (\mathcal{L}^*, m^*))}\right).$$

Denote the probability of birth, death, and change moves with knot number $m$ by $b_m$, $d_m$ and $c_m$ respectively, then $b_m + d_m + c_m = 1$. If $m = 1$, we set $d_m = c_m = 0$; if $m = \lambda_0$, we set $b_m = 0$ and $d_m = c_m = 1/2$; and if $1 < m < \lambda_0$, we set $b_m = d_m = c_m = 1/3$. Then $J$ is calculated as follows,

$$J((\mathcal{L}, m) \to (\mathcal{L}^*, m^*)) = \begin{cases} \frac{b_m}{M-m} & \text{if } m^* = m+1, \\ \frac{d_m}{m} & \text{if } m^* = m-1, \\ \frac{c_m}{m(M-m)} & \text{if } m^* = m. \end{cases} \quad (2.12)$$

Following this RJMCMC algorithm, the knot number and locations are automatically selected at each iteration. We illustrate this algorithm in Section 3.2 through simulation experiments.

For the choice of block partition for the FSA-Block, one principle is to maximize residual correlations within blocks and minimize residual correlations across blocks so that most of the spatio-temporal correlations are preserved. If the spatio-temporal residual covariance is fairly isotropic, one simple strategy is to apply the K-means clustering algorithm on observed spatio-temporal points to find $K$ cluster centers (Kaufman and Rousseeuw (1990)) and then create partitions in space-time domain. For the choice of tapering range for the FSA-Taper, some pilot studies can be conducted to give a rough estimate of the practical spatial/temporal dependence range. For example, we can select several time points and consider purely spatial datasets to estimate the spatial dependence range; similarly, we can consider time series at properly selected locations to estimate the time dependence range. These pilot estimates of dependence range are then subsequently used to set proper/conservative taper range to balance the trade-off between covariance approximation accuracy and computation efficiency.

## 3. Simulation Study

In this section, we report on simulation studies to evaluate the performance of the spatio-temporal FSA approach. We show the effectiveness of FSA-Block in approximating stationary spatio-temporal covariance models, and compare it with the independent blocks model (denoted as "Block"), predictive process model (denoted as "PP") and modified predictive process model (denoted as "Modified PP") (Finley et al. (2009)). The implementations of all methods were written in Matlab and run on a processor with 2.9 GHz Xeon CPUs and 16GB memory. For likelihood function optimization, we used the matlab function fminunc which implements a Broyden-Fletcher-Goldfarb-Shanno (BFGS) based Quasi-Newton method. In the supplementary document, we also illustrate the FSA with random knot selection for a nonstationary spatio-temporal covariance model in a simulation study. In both simulation studies, the full covariance model (denoted as "FM") is also implemented to serve as the benchmark.

We randomly selected 4,000 spatio-temporal locations on a space-time domain $\mathcal{S} \times \mathcal{T}$, where $\mathcal{S} = [0, 20] \times [0, 20]$ and $\mathcal{T} = [0, 20]$. The selected locations were then divided into a training set of size 3,500 and a test set of size 500, where the test set included 243 points in a space-time hole $[5, 10] \times [5, 10] \times [0, 20]$ and 257 randomly selected points from the remaining space-time locations. We obtained realizations of the spatio-temporal process $Y(\mathbf{s}, t)$ at the selected points following the model in (2.1).

We first experimented with a nonseparable space-time covariance function proposed by Gneiting (2002),

$$C(\mathbf{h}, u) = \frac{\sigma^2}{((20|u|^{2\alpha}/a) + 1)} \exp\left(-\frac{3\|\mathbf{h}\|}{c\left((20|u|^{2\alpha}/a) + 1\right)^{\eta/2}}\right), (\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R},$$
(3.1)

where $a, c > 0$ are temporal and spatial dependence range parameters respectively; $\alpha \in (0, 1]$ is the smoothness parameter; and $\eta \in [0, 1]$ is the space-time interaction parameter. The mean of the regression model $\mu(\boldsymbol{s}, t)$ was set to 0 for the entire region. We used equal variance $\tau^2 = 0.01$ for the variance of $\epsilon(\mathbf{s}, t)$. The true values of the covariance parameters and other model parameters are shown in Table 1. Two parameter settings were considered: the first had $a = 10$ and $c = 20$, for a large-scale spatio-temporal dependence structure; the second had $a = 5$ and $c = 10$, for a small-scale spatio-temporal dependence structure. The maximum likelihood estimators (MLEs) were obtained based on the training set and the mean squared prediction errors (MSPE) were calculated based on the predictions of the test set for evaluation.

We implemented the FSA-Block approach using 500 spatio-temporal knots and 35 blocks. The knots were chosen randomly from $\mathcal{S} \times \mathcal{T}$ and the 35 block

centers were created by the K-means clustering algorithm based on Euclidean distances of space-time points. For comparisons, the independent blocks method with the same 35 blocks and the predictive process method with the same 500 knots were considered. The smoothness parameter $\alpha$ in the covariance model was fixed to be 0.5. The parameter estimations and the prediction results of different approaches are shown in the Table 1.

Under the first parameter setting, where the spatio-temporal dependence range was large, the FSA-Block approach clearly outperformed the other methods in terms of prediction. The independent blocks method gave less accurate predictions. The predictive process model and the modified predictive process model did not work well either in terms of prediction, possibly requiring a denser knot set for a satisfactory approximation.

The FSA-Block obtained reasonable estimates for the range parameters $a$ and $c$, but higher MSEs than the independent blocks method. The estimate of the nugget effect $\tau^2$ obtained by the FSA-Block was slightly higher than the truth. The biases may be attributed to its predictive process part, which underestimates the correlations between blocks due to the limited number of knots. The naive independent blocks method worked well in terms of parameter estimation, which is not surprising since local information may be enough for estimating a stationary model with dense observations. The parameter estimation results of the predictive process model had noticeable biases, again perhaps due to the use of limited knots. Besides, the predictive process model gives a much larger estimate of the nugget effect due to its underestimation of the variance at each location (Finley et al. (2009)). The modified predictive process provided a bias correction for the variance at each location, so its estimates of $\tau^2$ and $\sigma^2$ were better than those obtained from the predictive process model, but it still underestimated correlations, leading to biased estimation of range parameters. The FSA-Block provides bias-correction for the predictive process model within each block, thus the estimates of range parameters, the nugget, and the variance had much smaller biases than those obtained from the predictive process model and the modified predictive process model.

When the spatio-temporal dependence range was relatively small, the FSA-Block still gave comparable prediction performance with the full covariance model, while the predictive process model and the modified predictive process model gave worse prediction performance than under large-scale spatio-temporal dependence. The predictive process model fails to capture small-scale dependence and thus its performance is often sensitive to the strength of dependence, and its parameter estimation has fairly large biases. The FSA approach seems to be more robust and capable of adjusting the biases in the estimation of the range parameters at different scales of dependence range.

Table 1. The means and MSEs (in parenthesis) of each parameter and MSPE results for covariance model with a nugget. The results are based on 100 runs of simulations.

| Settings | Method | Mean and MSEs | | | | | MSPE |
|---|---|---|---|---|---|---|---|
| | | $a$ | $c$ | $\eta$ | $\sigma^2$ | $\tau^2$ | |
| Set-up 1 | | 10 | 20 | 0.5 | 1 | 0.01 | |
| | FM | 9.68 ( 1.75) | 19.81 ( 4.31) | 0.48 (0.0395) | 0.97 (0.0062) | 0.01 (0.0001) | 0.34 |
| | FSA-Block | 11.73 ( 5.84) | 25.09 ( 34.71) | 0.48 (0.0687) | 1.04 (0.0104) | 0.04 (0.0009) | 0.37 |
| | Block | 9.48 ( 2.38) | 19.89 ( 4.97) | 0.39 (0.0923) | 0.96 (0.0090) | 0.02 (0.0001) | 0.43 |
| | PP | 23.47 (206.36) | 37.54 (349.67) | 0.87 (0.2175) | 2.25 (1.6948) | 0.40 (0.1499) | 0.45 |
| | Modified PP | 26.16 (293.08) | 42.26 (539.68) | 0.86 (0.2224) | 1.51 (0.2989) | 0.03 (0.0011) | 0.46 |
| Set-up 2 | | 5 | 10 | 0.5 | 1 | 0.01 | |
| | FM | 5.10 ( 0.27) | 10.19 ( 0.40) | 0.47 (0.0368) | 0.99 (0.0020) | 0.02 (0.0004) | 0.60 |
| | FSA-Block | 5.82 ( 1.15) | 11.68 ( 3.85) | 0.46 (0.0809) | 0.97 (0.0035) | 0.06 (0.0031) | 0.63 |
| | Block | 5.03 ( 0.29) | 10.18 ( 0.60) | 0.43 (0.0560) | 0.97 (0.0028) | 0.03 (0.0006) | 0.66 |
| | PP | 17.75 (181.76) | 21.24 (135.58) | 0.62 (0.1947) | 1.52 (0.3322) | 0.64 (0.4006) | 0.73 |
| | Modified PP | 19.04 (221.35) | 21.87 (150.45) | 0.80 (0.2076) | 1.18 (0.0554) | 0.15 (0.0223) | 0.73 |

The Matérn class (Matérn et al. (1960); Stein (1999)) is another widely used stationary covariance family due to its flexibility in accommodating different smoothness. We simulated data from the Matérn covariance model with

$$C(\mathbf{h}, u) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} \left(3\sqrt{\frac{\|\mathbf{h}\|^2}{\phi_s^2} + \frac{|u|^2}{\phi_t^2}}\right)^{\nu} K_\nu \left(3\sqrt{\frac{\|\mathbf{h}\|^2}{\phi_s^2} + \frac{|u|^2}{\phi_t^2}}\right), \qquad (3.2)$$

where $\phi_s, \phi_t > 0$ are spatial and temporal range parameters, respectively, $\nu > 0$ is the smoothness parameter, and $K_\nu$ denotes the modified Bessel function of the second kind of order $\nu$. The results are included in the supplementary document.

## 4. Analysis of the Eastern US Ozone Data

We applied the spatio-temporal FSA to the daily surface ozone data collected at 513 monitoring stations in the eastern US from May 1, 1998 to October 31, 1999. The observations are the maxima of hourly means over 8 consecutive hours of ozone. The raw data can be downloaded from `www.image.ucar.edu/Data/Ozmax/`.

We followed the procedure described in Gilleland and Nychka (2005) and Bevilacqua et al. (2012) to pre-process the daily observations. The daily maximum 8-hour ozone measurement at station $\mathbf{s}$ and day $t$ is assumed to have the decomposition,

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + \sigma(\mathbf{s})w(\mathbf{s}, t),$$

where $\mu(\mathbf{s}, t) = a(\mathbf{s}) + \sum_{j=1}^{3} \{b_j(\mathbf{s}) \cos(2\pi jt/184) + c_j(\mathbf{s}) \sin(2\pi jt/184)\}$, modeling the seasonal effect. The coefficients in the seasonal effect $\mu(\mathbf{s}, t)$ were estimated

by ordinary least square and $\sigma(\mathbf{s})$ was estimated using the residuals after removing the seasonal effect. Following Gilleland and Nychka (2005), the estimated coefficients matrix of the seasonal effect were further smoothed over space.

We modeled the spatio-temporal component $w(\mathbf{s}, t)$ by a Gaussian process with mean 0 and a nonseparable spatio-temporal covariance function as in (3.1), with $\mathbf{s}$ defined on the sphere. Since the station locations are on the sphere, the transformed great circle distance (Gneiting (1999); Gilleland and Nychka (2005)) was used to ensure positive-definiteness of the covariance function: $d(\mathbf{s}, \mathbf{s}') = 2r \sin(\Delta\phi/2)$, where $r$ is the radius of the earth and $\Delta\phi \in [0, \ \pi]$ is the central angle between $\mathbf{s}$ and $\mathbf{s}'$. We used kilometer as the unit of spatial lags and day as the unit of temporal lags.

Using only the monthly data in June and July, 1998 and 1999, allows us to implement the full covariance model whose results can be used as a benchmark. For each monthly dataset containing roughly $15,000$ observations, we randomly selected 1,500 space-time data points as a hold-out set for prediction and used the rest as training data. We obtained the maximum likelihood estimates of model parameters of the full model, the FSA-Block method, and the weighted composite likelihood (WCL) method (Bevilacqua et al. (2012)) for the training data. For the FSA-Block approach, we applied Latin hypercube sampling to obtain 400 space-time knots and the K-means clustering algorithm to divide the monthly data into 14 blocks. The WCL method needs to specify a pair of spatio-temporal lags $(d_s, d_t)$ such that the weights $w_{ij} = 1$ when $d(\mathbf{s}, \mathbf{s}') \leq d_s$ and $|t_i - t_j| \leq d_t$, $w_{ij} = 0$ otherwise. Following Bevilacqua et al. (2012), we set $d_s = 400, d_t = 3$, obtained by minimizing the asymptotic variances of WCL estimators. The results are included in the supplementary document.

We considered larger datasets of around 45,000 daily ozone observations from June to August in 1998 and 1999. We randomly held out 4,500 space-time data points for prediction. We considered three covariance models to fit the data: model A is the separable covariance model in (3.1) with $\eta = 0$; model B is the nonseparable covariance model in (3.1); and model C is the Matérn covariance model in (3.2). Here MLEs of model parameters were only obtained for the FSA-Block and WCL methods only since the full covariance model is not computationally feasible. For the WCL method, the weights were chosen in the same way as in the monthly data analysis; for the FSA-Block method, we applied Latin hypercube sampling to obtain 400 space-time knots and the K-means clustering algorithm to divide each summer dataset into 54 blocks. Prediction was performed for both methods using partial training data.

Parameter estimations and prediction results are shown in Table 2. It appears that Gneiting's covariance models (Model A and B) outperform the Matérn covariance model (Model C) in terms of prediction. The separable and the nonseparable Gneiting models provide comparable prediction results, indicating that

Table 2. Parameter estimation and prediction results of summer ozone. Model A is the separable covariance model in (3.1) with space-time interaction parameter $\eta = 0$. Model B is the nonseparable covariance model in (3.1). And model C is the Matérn covariance model in (3.2).

| Year | Method | Model | $a$ | $c$ | $\eta$ | $\alpha$ | RMSPE |
|------|--------|-------|-----|-----|--------|----------|-------|
| | | | Gneiting's model | | | | |
| | FSA | A | 20.17 | 376.88 | − | 0.268 | 0.372 |
| 1998 | | B | 20.78 | 378.62 | 0.062 | 0.267 | 0.372 |
| | WCL | A | 29.22 | 1106.17 | − | 0.716 | 0.385 |
| | | B | 23.98 | 1060.31 | 1 | 0.775 | 0.382 |
| | | | Matérn model | | | | |
| | | Model | | $\phi_t$ | $\phi_s$ | $\nu$ | RMSPE |
| | FSA | C | | 2.55 | 1958.93 | 0.274 | 0.436 |
| | WCL | C | | 12.30 | 285.89 | 1.830 | 0.635 |
| | | | Gneiting's model | | | | |
| Year | Method | Model | $a$ | $c$ | $\eta$ | $\alpha$ | RMSPE |
| | FSA | A | 19.15 | 418.08 | − | 0.270 | 0.380 |
| 1999 | | B | 19.22 | 418.32 | 0.008 | 0.270 | 0.380 |
| | WCL | A | 20.01 | 1022.59 | − | 1 | 0.402 |
| | | B | 14.63 | 1001.88 | 1 | 1 | 0.401 |
| | | | Matérn model | | | | |
| | | Model | | $\phi_t$ | $\phi_s$ | $\nu$ | RMSPE |
| | FSA | C | | 1.49 | 2863.92 | 0.251 | 0.447 |
| | WCL | C | | 9.91 | 132.79 | 2.750 | 0.821 |

a simple separable covariance model may be capable of modeling the spatio-temporal dependence of the summer ozone datasets. For all three covariance models, the FSA-Block clearly outperforms the WCL method in terms of prediction performance. The parameter estimations using WCL seem problematic in some cases, for example, estimates of $\eta$ and $\alpha$ for Gneiting models are on the boundary of the parameter space for the dataset in the summer of 1999. We applied the RJMCMC algorithm described in Section 2.4 to automatically select knots when applying the FSA-Block on the summer ozone datasets in 1998 and 1999, and the results are included in the supplementary document.

## 5. Discussion

We have proposed a method FSA to approximate a spatio-temporal covariance function. Our construction provides a flexible framework for statistically and computationally efficient parameter estimation and prediction for modeling of large spatio-temporal datasets. We have focused on the FSA-Block variation that provides exact bias-corrections for spatio-temporal pairs of the covariance matrix within blocks.

We have used the K-means clustering algorithm to choose block centers and subsequently create the space-time partition. An interesting direction for the future work is to treat the partition as unknown and select it adaptively using a Bayesian method, such as the tree-generating process (Chipman, George, and McCulloch (1998); Gramacy and Lee (2008); Konomi, Sang, and Mallick (2013)).

## Acknowledgement

## References

Bai, Y.,Song, P. and Raghunathan, T. E. (2012). Joint composite estimating functions in spatiotemporal models. *J. Roy. Statist. Soc. Ser. B* **74**, 799-824.

Banerjee, S., Gelfand, A., Finley, A. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *J. Roy. Statist. Soc. Ser. B* **70**, 825-848.

Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2012). Estimating space and space-time covariance functions for large data sets: A weighted composite likelihood approach. *J. Amer. Statist. Assoc.* **107**, 268-280.

Chipman, H., George, E. and McCulloch, R. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93**, 935-960.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York.

Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.* **94**, 1330-1340.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *J. Roy. Statist. Soc. Ser. B* **70**, 209–226.

Finley, A., Sang, H., Banerjee, S. and Gelfand, A. (2009). Improving the performance of predictive process modeling for large datasets. *Comput. Statist. Data Anal.* **53**, 2873-2884.

Furrer, R., Genton, M. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15**, 502-523.

Gilleland, E. and Nychka, D. (2005). Statistical models for monitoring and regulating ground-level ozone. *Environmetrics* **16**, 535-546.

Gneiting, T. (1999). Correlation functions for atmospheric data analysis. *Quarterly J. Roy. Meteorolog. Soc.* **125**, 2449-2464.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *J. Amer. Statist. Assoc.* **97**, 590-600.

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Ann. Statist.* **103**, 1119-1130.

Green, P. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika* **82**, 711-732.

Guhaniyogi, R., Finley, A., Banerjee, S. and Gelfand, A. (2011). Adaptive gaussian predictive process models for large spatial datasets. *Environmetrics* **22**, 997–1007.

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

Katzfuss, M. and Cressie, N. (2011). Spatio-temporal smoothing and em estimation for massive remote-sensing datasets. *J. Time Ser. Anal.* **32**, 430-446.

Kaufman, L. and Rousseeuw, P. (1990). *Finding groups in data*. Volume 16. Wiley, New York.

Konomi, B. A., Sang, H. and Mallick, B. K. (2013). Adaptive bayesian nonstationary modeling for large spatial datasets using covariance approximations. *J. Comput. Graph. Statist.*, accepted.

Matérn, B. et al. (1960). Spatial variation. stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden fran statens Skogsforskningsinstitut* **49**(5).

McKay, M. D., Conover, W. J. and Beckman, R. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21**, 239-245.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Volume 104, Chapman & Hall.

Sang, H. and Huang, J. (2012). A full scale approximation of covariance functions for large spatial data sets. *J. Roy. Statist. Soc. Ser. B* **74**, 111-132.

Stein, M. (1987). Large sample properties of simulations using latin hypercube sampling. *Technometrics* **29**, 143-151.

Stein, M. (2005). Space-time covariance functions. *J. Amer. Statist. Assoc.* **100**, 310-321.

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

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92**, 519-528.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.* **4**, 389-396.

Wendland, H. (1998). Error estimates for interpolation by compactly supported radial basis functions of minimal degree 1. *J. Approx. Theory* **93**, 258-272.

Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, USA.

E-mail: zhang@stat.tamu.edu

Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, USA.

E-mail: huiyan@stat.tamu.edu

Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, USA.

E-mail: jianhua@stat.tamu.edu