





Adaptive Bayesian Nonstationary Modeling for Large Spatial Datasets Using Covariance Approximations

Bledar A. Konomi, Huiyan Sang & Bani K. Mallick


To cite this article: Bledar A. Konomi, Huiyan Sang & Bani K. Mallick (2014) Adaptive Bayesian Nonstationary Modeling for Large Spatial Datasets Using Covariance Approximations, Journal of Computational and Graphical Statistics, 23:3, 802-829, DOI: [10.1080/10618600.2013.812872](https://doi.org/10.1080/10618600.2013.812872)

To link to this article: <https://doi.org/10.1080/10618600.2013.812872>

 View supplementary material 

 Accepted author version posted online: 18 Jul 2013.
Published online: 23 Jun 2014.

 Submit your article to this journal 

 Article views: 577

 View related articles 

 View Crossmark data 

 Citing articles: 15 View citing articles 



Adaptive Bayesian Nonstationary Modeling for Large Spatial Datasets Using Covariance Approximations

Bledar A. KONOMI, Huiyan SANG, and Bani K. MALLICK

Gaussian process models have been widely used in spatial statistics but face tremendous modeling and computational challenges for very large nonstationary spatial datasets. To address these challenges, we develop a Bayesian modeling approach using a nonstationary covariance function constructed based on adaptively selected partitions. The partitioned nonstationary class allows one to knit together local covariance parameters into a valid global nonstationary covariance for prediction, where the local covariance parameters are allowed to be estimated within each partition to reduce computational cost. To further facilitate the computations in local covariance estimation and global prediction, we use the full-scale covariance approximation (FSA) approach for the Bayesian inference of our model. One of our contributions is to model the partitions stochastically by embedding a modified treed partitioning process into the hierarchical models that leads to automated partitioning and substantial computational benefits. We illustrate the utility of our method with simulation studies and the global Total Ozone Matrix Spectrometer (TOMS) data. Supplementary materials for this article are available online.

Key Words: Bayesian treed Gaussian process; Full-scale approximation; Kernel Convolution; Markov chain Monte Carlo; Reversible-jump Markov chain Monte Carlo.

1. INTRODUCTION

The motivating problem of our article is the analysis and spatial prediction of atmospheric ozone data. Ozone plays two significant roles in the atmosphere. Tropospheric ozone, which appears in the lower atmosphere, is an air pollutant that is harmful to breathe and damages crops, trees, and other vegetation. It is a main ingredient of urban smog. The troposphere generally extends to a level about 6 miles up, where it meets the second layer, the stratosphere. Stratospheric ozone is important for life on the Earth because it absorbs incoming ultraviolet (UV) radiation and constitutes a negative radioactive forcing of climate (World Meteorological Organization 2007, Chapter V). In an effort to better understand

Bledar A. Konomi, Pacific Northwest National Laboratory (E-mail: bledarkonomi@pnnl.gov). Huiyan Sang (E-mail: huiyan@stat.tamu.edu), and Bani K. Mallick (E-mail: bmallick@stat.tamu.edu), Department of Statistics, Texas A&M University, College Station, TX 77843.

© 2014 *American Statistical Association, Institute of Mathematical Statistics,*
and *Interface Foundation of North America*

Journal of Computational and Graphical Statistics, Volume 23, Number 3, Pages 802–829

DOI: [10.1080/10618600.2013.812872](https://doi.org/10.1080/10618600.2013.812872)

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jcgs.

atmospheric ozone, scientists from NOAA and NASA have developed various instruments for measuring ozone values. One important instrument is the satellite-based Total Ozone Mapping Spectrometer (TOMS) that provides global measurements of total column ozone (TCO) on a daily basis. The raw satellite data, called “Level 1” data, are calibrated and preprocessed by NASA to yield Level 2 TCO measurements, which are spatially and temporally irregular following satellite scanning tracks (measurements are eight seconds apart). Level 2 data usually have a significant number of missing observations due to cloud and tropospheric aerosol contaminations. For wider scientific use, Level 2 data are further processed to produce daily data on regular grids (1 degree latitude by 1.25 degrees longitude for pixels with latitude from 50 S to 50 N, see Krueger et al. (1998) for more details), called Level 3 data.

Our goal is to build a spatial model for the TOMS Level 2 data, investigating its global scale spatial dependence structure and predicting Level 3 data from Level 2 data efficiently. The huge dimensionality and the spatial heterogeneity feature, which usually accompany the satellite ozone data, make it challenging to study them on the global scale. Daily, we observe approximately 170,000 to 250,000 Level 2 TOMS data over the globe. The massive number of observations make it computationally infeasible for analyzing Level 2 data and predicting Level 3 TOMS data using conventional spatial statistics methods. Moreover, the heterogeneity feature of this data (Jun and Stein 2008) poses further methodological and computational challenges for the analysis.

The ability to model the heterogeneity of the data with accuracy is of particular importance in environmental and geophysical science. In recent years, a number of nonstationary covariance functions of Gaussian processes (GP) have been proposed to model this heterogeneity of spatial data. The spatial deformations have been used to model nonstationary covariance functions in GPs by Sampson and Guttorp (1992); Schmidt and O’Hagan (2003); Anderes and Stein (2008). But the implementation typically requires replications of the process to learn the deformation functions. The only exception is the work by Anderes and Stein (2008), where they implement the deformation model of an isotropic GP with approximate likelihood computations derived from partitioning the observations into subregions and assuming independence of the process across partitions. Jun and Stein (2008) used a parametric nonstationary covariance function on a global scale to model the TOMS ozone data. They considered the gridded Level 3 TOMS data, where computations can be facilitated via a discrete Fourier transformation. Other approaches use different sets of basis functions that can be modeled in space (e.g., wavelet, kernel). Nychka and Royle (2002) applied a wavelet approach to produce a nonstationary covariance function for gridded data.

Another popular nonstationary covariance class, which is also used for computation reduction, is based on kernel and process convolution techniques to create nonstationary covariance functions (Higdon 1998; Fuentes 2001; Smith 2001; Calder 2008; Higdon, Swall, and Kern 2011 among others). Despite the fact that the above methods can be adapted to accommodate spatial varying covariance matrices, the computational complexity increases. They typically assume finite approximates to the integrals in kernel and process convolutions to facilitate the computation with very large and irregularly spaced data. However, the computational complexity increases when we introduce nonstationarity in the above models. Paciorek and Schervish (2006) extended the kernel convolution approach (Higdon 1998) and created a general class of nonstationary covariance functions with

closed forms built upon familiar stationary covariance functions. One of the appealing advantages of this nonstationary class is that it can link piecewise GPs to create a global GP model for fitting and prediction of heterogeneous spatial data. In this way, it naturally allows local covariance parameters to be knitted together into a valid global nonstationary covariance. In the meantime, the computational cost of model fitting can be alleviated if we predetermine subregions with nearly stationary local behavior and estimate local covariance structures separately. However, it is still computationally expensive if the number of observations within each subregion is large. Prediction is also difficult since it involves the global nonstationary covariance. Another challenging question is how to determine spatial partitioning that can capture nonstationarity well.

Bayesian treed Gaussian processes (BTGP) models have also been used to model nonstationarity by stochastically partitioning the global region into stationary subregions using a binary tree-generating process (Gramacy and Lee 2008). The BTGP model partitions the spatial space into subregions and fits separately stationary GP models within each subregion. Despite their advantages, the BTGP models are not suitable for very large spatial datasets such as the ozone data considered in our study. This is because each subregion may still involve a large number of observations, making local GP model fitting computationally very challenging. Conditional on a treed partition, the prediction of a BTGP model is done independently within each subregion following the conventional Gaussian Process prediction technique (Hjort and Omre 1994), which often results in undesirable discontinuity in the predictive surface across the partition boundaries. Therefore, Bayesian model averaging over the posterior of tree space need to be used to produce smoother transitions between regions.

In this article we propose a nonstationary covariance model that allows us to model large spatial datasets flexibly and efficiently. To deal with large nonstationary spatial datasets, we use the closed-form nonstationary covariance function proposed by Paciorek and Schervish (2006) based on space partitions. The nonstationary covariance function links different reduced stationary and anisotropic covariance functions from separate subregions into a unique covariance function, which allows fitting separate local GPs to each subregion and making predictions by combining them. Space partitioning is achieved by employing a binary tree-generating process similar to that of BTGP.

We use a newly developed computational method proposed by Sang and Huang (2012), referred to as the full-scale approximation (FSA) approach, to facilitate the computations involved in both local fitting and global prediction. Many existing computational methods for large spatial GP models often assume the covariance structure is a reduced-rank form (Banerjee et al. 2008; Cressie and Johannesson 2008), or a sparse covariance form (Gneiting 2002; Furrer, Genton, and Nychka 2006; Kaufman, Schervish, and Nychka 2008). Unfortunately, both models have their failure modes: Reduced-rank models cannot capture small-scale covariance accurately (see, e.g., Stein 2008; Finley et al. 2009), while sparse covariance models cannot approximate accurately if the true covariances are not sparse (Kaufman, Schervish, and Nychka 2008). The FSA approach combines a reduced rank covariance approximation for the large-scale variation with a tapering or block covariance approximation to model the small-scale covariance unexplained by the reduced rank part. It can effectively capture both the large-scale and small-scale spatial variations while maintaining computational efficiency. The FSA approach can be easily integrated into the

conventional Bayesian framework to both fit, and do predictions for, GP models. It has been proven to work well in the case where a stationary covariance function is assumed (Sang and Huang 2012). In this article, we investigate its use with nonstationary covariance function.

The rest of the article is organized as follows: In Section 2, we describe the Gaussian process. In Section 3 we describe how we model the covariance function, including details of the treed partitioning process and the FSA approach. Section 4 describes the Bayesian inference for treed space exploration, parameter estimations, and predictions with the facilitation of the FSA techniques. In Section 5, we describe a simulation study where we compare our method with other existing methods and, in Section 6, we apply the method to the Level 2 TOMS data. Conclusions are presented in Section 7.

2. REVIEW OF GAUSSIAN PROCESS MODELS FOR SPATIAL DATASETS

In this section, we present a summary of Gaussian process models for spatial datasets. Our presentation of Gaussian process models is based on the standard treatment in Banerjee, Carlin, and Gelfand (2004) and Schabenberger and Gotway (2005). The basic geostatistical Gaussian regression model is of the form

$$Z(s) = \mu(s) + w(s) + \epsilon(s), \quad (1)$$

where the process is decomposed into a mean part $\mu(s)$, often modeled as a linear regression, that is, $\mu(s) = X(s)\beta$, and two independent error processes, $w(s)$ and $\epsilon(s)$: $w(s)$, is introduced to capture the spatial association and it is assumed to be a mean zero Gaussian spatial process; $\epsilon(s)$ models the measurement error, also known as the nugget effect, which is usually modeled with an independent Gaussian process.

Given n observations, the model in (1) is usually written as the vector form $\mathbf{Z} = \boldsymbol{\mu} + \mathbf{w} + \boldsymbol{\epsilon}$, where $\mathbf{Z} = (Z(s_1), \dots, Z(s_n))$, $\boldsymbol{\mu} = (\mu(s_1), \dots, \mu(s_n))$, $\mathbf{w} = (w(s_1), \dots, w(s_n))$, and $\boldsymbol{\epsilon} = (\epsilon(s_1), \dots, \epsilon(s_n))$. Let $\mathbf{C}_{n,n}$ denote the covariance matrix of \mathbf{w} and τ^2 denote the variance of $\epsilon(s)$. The log-likelihood function can be expressed as

$$l_n = -\frac{1}{2} \log(\det\{\mathbf{C}_{n,n} + \tau^2 \mathbf{I}_n\}) - \frac{1}{2} (\mathbf{Z} - \boldsymbol{\mu})' (\mathbf{C}_{n,n} + \tau^2 \mathbf{I}_n)^{-1} (\mathbf{Z} - \boldsymbol{\mu}).$$

Evaluation of the log-likelihood involves calculation of the inverse and determinant of the $n \times n$ matrix $(\mathbf{C}_{n,n} + \tau^2 \mathbf{I})$ which typically requires $\mathcal{O}(n^3)$ operations. This makes the use of GP impractical for large volumes of spatial data such as global satellite ozone datasets in which n is at the scale of 170,000 to 250,000.

3. MODEL

Most often, $w(s)$ is specified to have a stationary isotropic correlation function along with a constant scale, as in the case of the Matérn family (Stein 1999). However, a number of previous studies suggest that the ozone data often show geometric anisotropies and a nonstationary association structure on a global scale (see, e.g., Trenberth and Shea 2005; Stein 2007; Jun and Stein 2008; Stein 2008). In particular, Stein (2007) showed that

the covariance structure of the TOMS ozone data shows strong dependence on latitude, whereas within a narrow latitude band dependencies exhibit approximate isotropy and do not change significantly along the direction of longitude. Below, we first describe the covariance function that is used to characterize nonstationary dependence structures. This particular covariance function requires space partitions, which is done by assigning with a treed partitioning process prior that we describe in Section 3.2. Finally in Section 3.3, we introduce the FSA approach to approximate the (local/global) covariance function for computational purposes.

3.1 MODELING THE NONSTATIONARY COVARIANCE FUNCTION

To relax the stationarity assumption on $w(s)$, we choose to model its covariance based upon a parametric nonstationary correlation function constructed following the spirit of kernel convolution (see Paciorek and Schervish 2006). Specifically, if an isotropic correlation function, $\rho_0(\cdot)$, is positive definite on \mathbb{R}^d then a valid nonstationary correlation function on \mathbb{R}^d is defined by

$$\rho_{\text{NS}}(s_i, s_j) = |\mathbf{B}_i|^{\frac{1}{4}} |\mathbf{B}_j|^{\frac{1}{4}} \left| \frac{\mathbf{B}_i + \mathbf{B}_j}{2} \right|^{\frac{1}{2}} \rho_0(Q_{ij}), \quad (2)$$

where \mathbf{B}_i is referred to as the $d \times d$ kernel covariance matrix at location s_i and $Q_{ij} = \sqrt{(s_i - s_j)'((\mathbf{B}_i + \mathbf{B}_j)/2)^{-1}(s_i - s_j)}$, the Mahalanobis distance between s_i and s_j . The use of Mahalanobis distance allows to model the anisotropy of the correlation functions where the spatial correlation between two observations will not depend only on the absolute distance but also upon the separation vector between their locations (see Banerjee, Carlin, and Gelfand 2004). The eigenvalue decomposition of \mathbf{B}_i has clear geometric interpretations: The square roots of the eigenvalues of \mathbf{B}_i control the range of the spatial dependence and the eigenvector matrix corresponds to a rotation matrix (Banerjee, Carlin, and Gelfand 2004). Therefore, we usually reparameterize \mathbf{B}_i as $\mathbf{B}_i = \Psi_i \Lambda_i \Psi_i^T$, where Λ_i is a diagonal eigenvalue matrix and Ψ_i is the eigenvector matrix. Nonstationarity is introduced into the covariance function by letting \mathbf{B}_i depend on location s_i .

The proof of the validity of the correlation function defined in (2) is a simple application of the Schoenberg theorem as it is shown in Paciorek and Schervish (2006). This suggests that we can construct a unique and valid nonstationary covariance function from a given stationary correlation function such as power exponential, rational quadratic, and Matérn family.

It is straightforward to construct a covariance function from the above correlation function by including a standard deviation function. In doing so, the covariance can be defined as

$$C(s_i, s_j) = \sigma_i \sigma_j \rho_{\text{NS}}(s_i, s_j; \nu, \phi), \quad (3)$$

where σ_i is the standard deviation at s_i and ρ_{NS} is the correlation function defined in (2).

The use of the nonstationary covariance class defined in (3) also requires specifications of \mathbf{B}_i and σ_i . In general \mathbf{B}_i and σ_i can be modeled as smooth functions or as discrete functions of the space domain. Paciorek and Schervish (2006) developed an MCMC algorithm for the case where \mathbf{B}_i is a smooth function of space. However, the computational cost of this approach is too expensive, thus making it infeasible to be applied for the analysis of the

massive global ozone dataset in our study. Therefore, we opt for a relatively simpler model specification where the kernel covariance matrices \mathbf{B}_i and the standard deviations σ_i are modeled using piecewise discrete functions. Specifically, suppose we partition the input space of a zero mean GP into D nonoverlapping regions. Let $\xi(s_i) \in \{1, \dots, D\}$, denote the region that s_i belongs to. Each region has its corresponding kernel matrix and standard deviation, that is, $\mathbf{B}_i = \mathbf{B}_{\xi(s_i)} \in \{\mathbf{B}_\xi, \xi = 1, \dots, D\}$ and $\sigma_i = \sigma_{\xi(s_i)} \in \{\sigma_\xi, \xi = 1, \dots, D\}$.

The covariance defined in (3) with discrete space partition has a unique feature; the global covariance is constructed by knitting local covariances together. For example, in the case of two subregions each with local parameters (σ_1, \mathbf{B}_1) and (σ_2, \mathbf{B}_2) respectively, the global covariance matrix becomes

$$\begin{pmatrix} \mathbf{C}_1(\sigma_1, \mathbf{B}_1) & \mathbf{C}_{12}(\sigma_1, \mathbf{B}_1, \sigma_2, \mathbf{B}_2) \\ \mathbf{C}_{21}(\sigma_1, \mathbf{B}_1, \sigma_2, \mathbf{B}_2) & \mathbf{C}_2(\sigma_2, \mathbf{B}_2) \end{pmatrix}, \quad (4)$$

where $\mathbf{C}_1(\sigma_1, \mathbf{B}_1)$ and $\mathbf{C}_2(\sigma_2, \mathbf{B}_2)$ are the local covariance matrices within subregion 1 and 2, and $\mathbf{C}_{12}(\sigma_1, \mathbf{B}_1, \sigma_2, \mathbf{B}_2)$ is the cross-covariance matrix between subregion 1 and 2. This unique structure will help to reduce the computations in parameters estimations as will be described in Section 4.1.

3.2 BINARY TREE-GENERATING PROCESS FOR SPACE PARTITIONING

One remaining issue of the discrete specification for \mathbf{B}_i and σ_i is how to split the study region. In some applications, one may have clear guidance to predetermine subregions based on information such as topographical boundaries or previous studies. But in most applications, one may not have the necessary prior knowledge to make the choice of subregions. Many methods have been proposed in recent statistical literature to partition the study region. For instance, the Voronoi tessellation is used by Kim, Mallick, and Holmes (2005) to create and fit piecewise independent stationary GPs. In Section 3.2, we introduce a space partitioning method that is motivated by the treed partitioning process used in the Bayesian CART (classification and regression trees) models (Chipman, George, and McCulloch 1998; Denison, Mallick, and Smith 1998) and the treed GP process models (Gramacy and Lee 2008).

Tree-generating models typically partition the input space into nonoverlapping regions by making binary splits recursively, that is, each new partition is a subpartition of a previous one. A binary tree consists of internal nodes and terminal nodes (also called leaves). A tree partitions input spaces by splitting each internal node according to a splitting rule on the value of a single input variable so that partition boundaries are parallel to coordinate axes.

In the Bayesian framework, a binary tree is treated as random and assigned with a prior distribution through a tree-generating process (Chipman, George, and McCulloch 1998; Gramacy and Lee 2008). Starting with a null tree (all data in a single region), a leaf node $\eta \in \mathcal{T}$, representing a subregion of the input space, splits with probability $P_{\text{split}}(\eta, \mathcal{T}) = a(1 + d_\eta)^{-b}$, where d_η is the depth of $\eta \in \mathcal{T}$, a controls the shape (balance) of the tree, and b controls the size of the tree. The treed model prior is

$$P(\mathcal{T}) = P_{\text{rule}}(\rho|\eta, \mathcal{T}) \prod_{\eta_i \in \mathcal{I}} P_{\text{split}}(\eta_i, \mathcal{T}) \prod_{\eta_j \in \mathcal{D}} (1 - P_{\text{split}}(\eta_j, \mathcal{T})),$$

where \mathcal{I} and \mathcal{D} denote the internal and terminal nodes, respectively. While $P_{\text{rule}}(\rho|\eta, \mathcal{T})$ denote splitting prior of: first, the splitting dimension u which is chosen to be discrete uniform distribution; and the split location ζ which is chosen to be uniformly distributed in a continuous subset of the locations S in the u th dimension. We add that in practice one may further specify tree priors such that each new region have at least a minimal number of data points to ensure parameters of the model within each subregion can be efficiently estimated (Gramacy and Lee 2008).

3.3 FULL SCALE APPROXIMATION MODELS FOR LARGE SPATIAL DATASETS

Conditional on a treed space partition, the nonstationary global covariance model in (2) is knitted together by local covariances in each subregion. This unique structure allows one to estimate the region-specific parameters ϕ_ξ by fitting a local GP covariance model using only the data in the ξ th subregion (see more details in Section 4.1). In doing so, computations are reduced from fitting a global GP model to a number of local GP models. However, in the case with massive global ozone data, each partitioned subregion may consist of relatively dense set of observations (of size 10,000 to 40,000). Although dense observations help to yield accurate estimations of local parameters without borrowing information across subregions, the high-dimensional nature of the local covariance model requires applying a computationally feasible approach.

Various approximation techniques have been introduced in recent literature to reduce computational cost. Banerjee et al. (2008) proposed the predictive process (PP) which is a low-rank approximation of a given parent GP. Cressie and Johannesson (2008) developed the fixed rank kriging method constructed from a reduced number of basis functions. Both of these methods assume reduced rank covariances that capture middle- to large-scale variation well. However, these reduced rank models usually fail to accurately capture the small-scale dependence structure that exists (see, e.g., Stein 2008; Finley et al. 2009). Another popular approach is to use tapering (Furrer, Genton, and Nychka 2006; Kaufman, Schervish, and Nychka 2008) as a way to construct sparse covariance matrix approximations. This method works well in handling short-range dependence but it may not be effective in accounting for spatial dependence with long range, because the tapered covariance function with a relatively small taper range fails to provide a good approximation to the original covariance function.

A new approach proposed by Sang and Huang (2012) offers a covariance approximation method, referred to as the full-scale approximation (FSA), to facilitate the computations for GPs. For any given covariance function, the function is decomposed into two parts: a reduced-rank part using predictive process (PP) to characterize the large-scale dependence and a residual part to capture the small-scale spatial dependence that is unexplained by the reduced-rank part. The residual covariance is then approximated by a sparse covariance structure using tapering or blocking. By combining the reduced rank and the sparse technique, the FSA overcomes the deficiencies of each and provides a high-quality approximation to the original covariance function at both the small and large spatial scales. Indeed, Sang and Huang (2012) demonstrated that the FSA offers substantial computational savings and achieves accuracy in both model inference and prediction comparable to the full covariance approach.

Consider a generic zero mean GP $w(s)$ with covariance function $C(s, s')$. Conditional on a given tree, each local GP is such an example. Let $\mathcal{S}^* = \{s_1^*, \dots, s_m^*\}$ a fixed set of “knots” that are chosen from the study region, the FSA approximates the covariance function $C(s, s')$ with the covariance function $C^\dagger(s, s')$ consisting of two parts,

$$C^\dagger(s, s') = C_l(s, s') + C_s(s, s'). \quad (5)$$

$C_l(s, s')$ in (6) is defined in the same fashion as the reduced-rank covariance function of the predictive process (Banerjee et al. 2008)

$$C_l(s, s') = \mathbf{C}^T(s, \mathcal{S}^*) \mathbf{C}_{m,m}^{-1}(\mathcal{S}^*, \mathcal{S}^*) \mathbf{C}(s', \mathcal{S}^*), \quad (6)$$

where $\mathbf{C}_{m,m}(\mathcal{S}^*, \mathcal{S}^*) = [C(s_i^*, s_j^*)]_{i,j=1}^m$, an $m \times m$ covariance matrix at knots set \mathcal{S}^* and $\mathbf{C}(s, \mathcal{S}^*) = [C(s, s_j^*)]_{j=1}^m$.

The reduced-rank part $C_l(s, s')$ can capture reasonably well the long-range but not the short-range dependence because it discards entirely the residual covariance $C(s, s') - C_l(s, s')$. The novelty of the FSA is a more careful treatment of the covariance function that can both preserve most information present in the residual process and also achieve computational efficiency. Specifically, the FSA adds a small-scale part $C_s(s, s')$ to create a sparse approximate to the residual covariance $C(s, s') - C_l(s, s')$ by multiplying with a modulating function $K(s, s')$.

$$C_s(s, s') = \{C(s, s') - \mathbf{C}^T(s, \mathcal{S}^*) \mathbf{C}_{m,m}^{-1}(\mathcal{S}^*, \mathcal{S}^*) \mathbf{C}(s, \mathcal{S}^*)\} K(s, s'). \quad (7)$$

There are two possible ways to specify $K(s, s')$ so that the latter part can capture the local residual spatial dependence well while still allows efficient computations. The first specification is to assume that $K(s, s')$ is a compactly supported correlation function, also called the tapering function, that is identically zero whenever $|s - s'| \geq \gamma$ for a positive taper range γ (Gneiting 2002). Choices of valid compact support correlation functions include spherical function and a family of Wendland functions among others (see, e.g., Wendland 1998; Gneiting 2002). The tapering function preserves the diagonal entries of the true residual covariance and shrinks the off-diagonal entries to ensure positive definiteness. Since the residual covariance mainly characterizes fine-scale dependence, the approximation errors by tapering are expected to be small for a conservatively chosen taper range. Using covariance tapering, we obtain a sparse covariance matrix and can use efficient sparse matrix algorithms for likelihood inference and prediction.

The second specification of $K(s, s')$ is local blocking, where residuals are assumed to be independent across partitioned input blocks, but to keep the original residual dependence structure within each block. Specifically, let A_1, \dots, A_k be k disjoint blocks which divide the input space. The function $K(s, s')$ is taken to be $K(s, s') = 1$ if s and s' belong to the same block, and $K(s, s') = 0$ otherwise. Voronoi tessellation is one option to construct the disjoint blocks (Green and Sibson 1978). The reduced-rank part plus the residual part using local blocking provides an exact recovery of the true covariance within each block. The approximation errors are $C(s, s') - C_l(s, s')$ if (s, s') belong to different blocks. Since most of these pairs are farther away, the errors are expected to be small for most entries.

Applying the FSA approach to approximate the covariance function of $w(s)$, the resultant covariance matrix at n given locations is the summation of the following two matrices:

$$\mathbf{C}_l = \mathbf{C}_{n,m} \mathbf{C}_{m,m}^{-1} (\mathbf{S}^*, \mathbf{S}^*)^T \mathbf{C}_{n,m}^T, \quad (8)$$

where $\mathbf{C}_{n,m} = [C(s_i, s_j^*)]_{i=1:n, j=1:m}$, and

$$\mathbf{C}_s = (\mathbf{C}_{n,n} - \mathbf{C}_{n,m} \mathbf{C}_{m,m}^{-1} (\mathbf{S}^*, \mathbf{S}^*)^T \mathbf{C}_{n,m}^T) \circ \mathbf{T}, \quad (9)$$

where $\mathbf{C}_{n,n} = [C(s_i, s_j)]_{i=1:n, j=1:n}$ and $\mathbf{T} = [K(s_i, s_j)]_{i=1:n, j=1:n}$. Here, the “ \circ ” notation refers to the element-wise matrix product, also called Schur or Hadamard product.

The approximation of the covariance matrix as described above will facilitate the computations of the likelihood or Bayesian posterior sampling by applying the well-known Sherman-Woodbury-Morrison (SWM) formula for inverse matrices and using a sparse or a block matrix as shown below.

$$\begin{aligned} & (\mathbf{C}_{n,m} \mathbf{C}_{m,m}^{-1} \mathbf{C}_{n,m}^T + \mathbf{C}_s + \tau^2 \mathbf{I}_n)^{-1} \\ &= (\mathbf{C}_s + \tau^2 \mathbf{I}_n)^{-1} - (\mathbf{C}_s + \tau^2 \mathbf{I}_n)^{-1} \mathbf{C}_{n,m} \\ & \quad \times \{ \mathbf{C}_{m,m} + \mathbf{C}_{n,m}^T (\mathbf{C}_s + \tau^2 \mathbf{I}_n)^{-1} \mathbf{C}_{n,m} \}^{-1} \mathbf{C}_{n,m}^T (\mathbf{C}_s + \tau^2 \mathbf{I}_n)^{-1}. \end{aligned} \quad (10)$$

According to (9), the $n \times n$ matrix $\mathbf{C}_s + \tau^2 \mathbf{I}_n$ is a sparse or a block diagonal matrix. The right hand sides of (10) only involve the inversion $n \times n$ sparse or block diagonal matrices and $m \times m$ low-rank matrices. Therefore, the computational cost in fitting the spatial model can be greatly reduced relative to the expensive computational cost of using the original covariance function.

The FSA model requires selection of knots and taper range that involves trade-off analysis between inference accuracy and computational cost. To our knowledge, how to select knots and taper range is still an open research topic. In this article, we follow Banerjee et al. (2010) to select knots on a uniform grid overlaid on the domain if data locations are fairly evenly distributed, and on k -means clustering centers if data locations are irregularly distributed. The number of knots and the taper range or the block size are determined by doing a crude pilot trade-off analysis between prediction accuracy and corresponding run time using a subset of training data and holdout data, as is done in Sang and Huang (2012).

4. BAYESIAN IMPLEMENTATION

Let $\boldsymbol{\Omega} = (\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\tau}, \boldsymbol{\sigma}, \mathcal{T})$ denote collectively the model parameters and $\boldsymbol{\phi} = (\mathbf{B}, \boldsymbol{\tau}, \boldsymbol{\sigma})$ denote all the local correlation parameters. The likelihood of $p(\mathbf{Z}|\boldsymbol{\Omega})$ is well defined by the model and is assumed to be normally distributed with a generalized global mean and a nonstationary covariance function as described in Section 3.1.

The Bayesian inference for parameter estimations begins with assigning prior distribution for $\boldsymbol{\Omega}$. The prior distribution of $\boldsymbol{\Omega}$ is specified, independently for the regression coefficients and tree, and in a conditionally independent manner for the covariance parameters $\boldsymbol{\phi}$, as

$$\pi(\boldsymbol{\Omega}) = \pi(\boldsymbol{\beta})\pi(\mathcal{T})\pi(\boldsymbol{\phi}|\mathcal{T}) = \pi(\boldsymbol{\beta})\pi(\mathcal{T}) \prod_{\xi=1:D} \pi(\mathbf{B}_\xi)p(\boldsymbol{\tau}_\xi)p(\boldsymbol{\sigma}_\xi),$$

where D is the total number of disjoint subregions and corresponds to the total number of external nodes in a tree setting.

For the regression coefficients β , we choose a vague multivariate normal prior. To specify prior distributions for tree model \mathcal{T} we follow Chipman, George, and McCulloch (1998), as described in Section 3.2. Moreover, for ϕ we choose the same local covariance functions for each of the terminal nodes. Specifically, we assign an inverse gamma prior for the variances of spatial errors σ_i^2 and the nugget error variances τ_i^2 . The hyperparameters in the priors for the variance components are chosen such that the Inverse Gamma distributions have large variances and reasonable guess of means from the exploratory analysis for a subset of samples following an empirical Bayes paradigm. Assuming the anisotropic correlation function described in (2), one can use a Wishart prior for the matrix \mathbf{B} as given in Banerjee, Carlin, and Gelfand (2004). Alternatively, one may opt to eigen-decompose the matrix \mathbf{B} and specify priors for the eigenvalues λ 's and the rotation parameters θ . In the latter case, a noninformative uniform prior on $(0, \pi]$ is chosen for each θ while prior for each λ is set to be an inverse gamma distribution with large variance and reasonable guess of means. Altogether, our priors are very weak and our data size is very large, suggesting little reason for concern with regard to sensitivity analysis. However, we did implement some sensitivity study, primarily on the various prior uncertainties, which confirmed this. Details are available with the authors.

The posterior distribution of the model parameters has the form

$$p(\boldsymbol{\Omega}|\mathbf{Z}) = p(\mathbf{Z}|\boldsymbol{\Omega}, \mathcal{T})\pi(\boldsymbol{\beta})\pi(\mathcal{T}) \prod_{\xi=1:D} \pi(\mathbf{B}_\xi)\pi(\tau_\xi)\pi(\sigma_\xi), \quad (11)$$

and since this distribution does not have a closed form, Markov chain Monte Carlo (MCMC) techniques should be used to draw samples of the model parameters. Gibbs sampler (Smith 1990) is used to simulate each component of $(\mathcal{T}|\phi, \beta)$, $(\beta|\mathcal{T}, \phi)$, and $(\phi|\beta, \mathcal{T})$. Furthermore, we use Metropolis-Hasting and reversible jump MCMC to sample from $(\phi|\beta, \mathcal{T})$ and $(\mathcal{T}|\phi, \beta)$. In Section 4.1, we first describe a computational efficient method using the FSA to update the GP parameters conditional on a treed partition, that is, updating $(\beta|\mathcal{T}, \phi)$ and $(\phi|\beta, \mathcal{T})$. We then proceed to describe the MCMC algorithm to sample from the posterior of treed space partition in Section 4.2.

4.1 MCMC UPDATING OF THE GP CONDITIONAL ON A TREED PARTITION

As shown in Section 3, full likelihood evaluation $p(\mathbf{Z}|\boldsymbol{\Omega})$ involves matrix operations of a very large global covariance matrix given by (3). The computational issue will be further complicated within a full MCMC implementation where large matrix operations have to be repeated for every MCMC iteration. Motivated by the unique structure of the nonstationary correlation function in (2) in which the global covariance model is knitted together by local covariance (see (4) for example), we estimate the regional-specific parameters ϕ_ξ by fitting a local GP covariance model using only the data in the ξ th subregion (see Paciorek and Schervish 2006). In the Bayesian framework, conditional on β , independent MCMCs are applied to draw samples for the parameters $\phi_\xi = (\mathbf{B}_\xi, \sigma_\xi, \tau_\xi)$ for every subregion using Metropolis-Hastings within Gibbs steps (Mueller 1993).

As described in Section 3.3, by choosing a set of knots and a modulating function by tapering or blocking for each subregion, we adopt the FSA approach to approximate each local covariance matrix and hence facilitate the posterior computations involved in each Bayesian local model fitting by applying the SWM formula in (10) for large covariance matrix operations. Our simulation results in Section 5 illustrate the utility of the local fitting method using the FSA for estimating the local parameters compared with the full likelihood method.

Conditional on all the other parameters and the treed partition \mathcal{T} , $\boldsymbol{\beta}$ has a multivariate normal posterior distribution. Following a similar strategy by assuming independence among different subregions, we draw $\boldsymbol{\beta}$ from the $\text{MVN}(\boldsymbol{\mu}_{\boldsymbol{\beta}|}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}|})$, where

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}|} \approx \left[\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0} + \sum_{\xi=1}^D \mathbf{X}'_{\xi} \boldsymbol{\Sigma}_{\xi}^{-1} \mathbf{X}_{\xi} \right] \approx \left[\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0} + \sum_{\xi=1}^D \mathbf{X}'_{\xi} \{ \mathbf{C}_{l,\xi} + \mathbf{C}_{s,\xi} + \tau_{\xi}^2 \mathbf{I} \}^{-1} \mathbf{X}_{\xi} \right], \quad (12)$$

and the mean

$$\boldsymbol{\mu}_{\boldsymbol{\beta}|} \approx \left[\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0} + \sum_{\xi=1}^D \mathbf{X}'_{\xi} \boldsymbol{\Sigma}_{\xi}^{-1} \mathbf{Z}_{\xi} \right] \approx \left[\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0} + \sum_{\xi=1}^D \mathbf{X}'_{\xi} \{ \mathbf{C}_{l,\xi} + \mathbf{C}_{s,\xi} + \tau_{\xi}^2 \mathbf{I} \}^{-1} \mathbf{Z}_{\xi} \right], \quad (13)$$

where \mathbf{X}_{ξ} is the regression design matrix which corresponds to the ξ th subregion, $\boldsymbol{\mu}_{\boldsymbol{\beta}_0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0}$ are the mean and covariance matrix of the prior distribution of $\boldsymbol{\beta}$, and $\mathbf{C}_{l,\xi} + \mathbf{C}_{s,\xi}$ are the covariance matrix by applying the FSA at the observed spatial points in subregion ξ .

4.2 MCMC UPDATING OF THE TREED PARTITION

The implementation of the treed model is based on the *grow*, *prune*, *change*, and *swap-rotate* operations as discussed by Chipman, George, and McCulloch (1998) and Gramacy and Lee (2008). These four tree operations are randomly selected during each MCMC iteration. The key tool of the tree-generating algorithm is the reversible jump MCMC (RJ-MCMC), Green (1995), which is used in the *prune* and *grow* operations.

When a *grow* operation occurs, one terminal (external) node creates two smaller and disjoint children nodes. These children nodes become terminal nodes of the new tree. New parameters must be proposed for one of the created terminal nodes as well as for the new splitting point, ζ_{r+1} . The other created terminal node absorbs its parameters from the parent subregion. In the *prune* operation, we randomly select interior nodes with both children external nodes and merge them. Parameters from one of the children subregions will be absorbed from the new node. In the *change* operation, an internal node is picked uniformly from the current tree internal nodes and changes the splitting rule. The tree structure and the parameters remain the same; however, the likelihood changes, since terminal nodes will not have the same limit boundaries. In the *swap* operation, we randomly select two interior parent-child nodes and swap the parent and child nodes' splitting rules. Because of possible problems with having the same splitting rule in the same variables for parent and child, a *rotation* operation is also included in this case. Below we illustrate the treed updating algorithm with the *grow* operation. For more details regarding the treed GP we refer the reader to Gramacy and Lee (2008).

Grow (split) operation: The first step for this operation is to uniformly select a subregion and then split it into two. To avoid very small subregions, the split-point proposal distribution is chosen to be uniformly distributed from a reasonable interval of the selected subregion. Let ϕ_j denote all the parameters in the j th selected subregion (parent node) and ϕ_{jk} denote the parameters of the k th child of the j th subregion (children subregions). In this step, we increase the dimension of the parameters to include one of the splitting parts (ϕ_{jk}). Next, we generate random variables ϕ_{jk} from the prior distribution.

If we have an existing \mathcal{T} tree and propose a \mathcal{T}' with a grow (split) step, the M-H ratio is

$$\frac{P(\mathcal{T})}{P(\mathcal{T}')} \frac{q(\mathcal{T}', \mathcal{T})}{q(\mathcal{T}, \mathcal{T}')} \frac{f(\mathbf{Z}_{j1}, \mathbf{Z}_{j2} | \beta, \phi_{j1}, \phi_{j2}) \pi(\phi_{j1}) \pi(\phi_{j2})}{f(\mathbf{Z}_j | \beta, \phi_j) \pi(\phi_j) q(\phi_{j2})} \| \mathbf{J} \|,$$

where $P(\mathcal{T})$ is the prior distribution for \mathcal{T} tree, and $q(\mathcal{T}', \mathcal{T})$ is the proposal distribution of changing from \mathcal{T}' to \mathcal{T} , $q(\phi_{j2})$ is the proposal distribution of generating parameters for one of the two formed siblings. This transformation will give a unity Jacobian term which can be ignored in the above equation. The M-H ratio also involves calculations of $f(\mathbf{Z}_j | \cdot)$ and $f(\mathbf{Z}_{j1}, \mathbf{Z}_{j2} | \cdot)$. $f(\mathbf{Z}_j | \cdot)$ is the existing likelihood of the parent subregion, which is the density of $\text{MVN}(\mathbf{X}_j \beta, \mathbf{C}_j^{(\text{old})}(\mathbf{B}_j, \sigma_j) + \tau_j^{(\text{old})2})$, where $\mathbf{C}_j^{(\text{old})}$ is the covariance matrix of all the observations in subregion j with ij th entry defined in (3). With the use of the FSA, $\mathbf{C}_j^{(\text{old})}$ is approximated by the summation of a reduced rank matrix and a sparse or block diagonal matrix $\mathbf{C}_{l,j}^{(\text{old})} + \mathbf{C}_{s,j}^{(\text{old})}$. The likelihood function becomes $\text{MVN}(\mathbf{X}_j \beta, \mathbf{C}_{l,j}^{(\text{old})} + \mathbf{C}_{s,j}^{(\text{old})} + \tau_j^{(\text{old})2})$, which can be calculated efficiently using the SWM formula in (9). $f(\mathbf{Z}_{j1}, \mathbf{Z}_{j2} | \cdot)$ is the new likelihood of the split parent subregion, which is the density of $\text{MVN}(\mathbf{X}_j \beta, \mathbf{C}_j^{(\text{new})}(\sigma_{j1}, \mathbf{B}_{j1}, \sigma_{j2}, \mathbf{B}_{j2}) + \tau_j^{(\text{new})2})$, where $\mathbf{C}_j^{(\text{new})}$ is the covariance matrix of all the observations in the old subregion j after the split, which has a similar expression as in (4). Again, the FSA can be applied to $\mathbf{C}_j^{(\text{new})}$ to speed up the computation. We follow the similar approach to build the merge and change step algorithms. When we deal with high-dimensional data, we may use the assumption of independent subregions without loss of generality. As previously mentioned, in the global ozone data, each partitioned subregion consists of a relatively dense set of observations (e.g., of size 10,000 to 40,000).

In practice, one may also take into account available prior knowledge to build the tree model. For instance, it is believed that the heterogeneity of the TOMS ozone data mostly comes through latitude, whereas no significant change of the spatial pattern exists along the direction of longitude Jun and Stein (2008). Using this feature, we can simplify the tree-generating process by growing the tree only in the needed dimension. The simplification eliminates the need to use *Swap* operation and helps the mixing of the MCMC.

The treed model proposed here is largely motivated from the original BTGP proposed by Gramacy and Lee (2008) but our method differs from the BTGP model in several aspects. First, our model is designed for a larger scale study than the BTGP model in terms of sample size. By applying the FSA, our method is able to handle the case where each subregion contains about 10^4 number of observations, whereas the BTGP model is perhaps limited to the case with a moderate number of observations ($< 10^4$) within each subregion. Second, our model assumes a global regression mean part whereas the treed GP models partition the regression part along with the covariance. Third, because of the availability of the closed form of the global covariance, the M-H ratios in our case differ from the

ones used in the BTGP algorithm. For example, in the grow step, we use the joint likelihood $f(\mathbf{Z}_{j1}, \mathbf{Z}_{j2}|\cdot)$, whereas the BTGP assumes the observations between two split child subregions are independent, that is, $f(\mathbf{Z}_{j1}, \mathbf{Z}_{j2}|\cdot) = f(\mathbf{Z}_{j1}|\beta, \boldsymbol{\phi}_{j1})f(\mathbf{Z}_{j2}|\beta, \boldsymbol{\phi}_{j2})$. Finally, conditional on a treed partition, the prediction of a BTGP model is done independently within each subregion following the conventional Gaussian process prediction technique (Hjort and Omre 1994). In contrast, the availability of the global covariance for a given tree allows us to produce a smoother predictive surface. The details will be discussed in the next section.

4.3 PREDICTION

Unlike the parameter estimations, procedure using local fitting for the model in (3), spatial prediction is typically sensitive to the assumption of independence among subregions. One major problem is that the independence assumption leads to unstable prediction around boundary areas for a given tree. It is, therefore, desirable to use the global covariance in some fashion to produce a smoother prediction surface. Making prediction with the full global covariance matrix is certainly not trivial because it requires to handle computations of matrices of size at the order of 10^6 . Although the FSA method described in Section 3.3 is, conceptually, also applicable to the global covariance model for prediction purpose, it is perhaps limited to the order of 10^4 on modest single-processor machines (Sang and Huang 2012). Nevertheless, it may not be necessary to do prediction using entire global data since contributions from extremely distant regions are expected to be very limited. A compromised approach is to taper the global covariance model with a moderate taper range first and then apply the FSA approach to the tapered global covariance. Below we discuss this method for spatial prediction in more details.

Weighing the trade-off between computational accuracy and complexity, we propose to apply the FSA to a sparse approximation of the global covariance for prediction. To be specific, conditional on the parameter estimations based on local fitting, we can calculate the global covariance, denoted as \mathbf{C}_{NS} , by plugging in the parameter values to the covariance model in (3). We create a sparse approximation to the global covariance \mathbf{C}_{NS} by tapering with a carefully and conservatively selected taper range to exclude distant observations, that is, $\mathbf{C}_{\text{NS},T} = \mathbf{C}_{\text{NS}} \circ \mathbf{T}$, where \mathbf{T} is the tapering matrix. Alternatively, to predict at a given location s , one may include several neighboring subregions besides the subregion that covers the new locations. In this case, we specify the tapering matrix $\mathbf{T} = [K_{i,j}]_{i=1:n, j=1:n}$ as $K_{i,j} = 1$ if s_i and s_j belong to the subregion that covers the new locations as well as the neighboring subregions, and $K_{i,j} = 0$ otherwise. We then apply the FSA to approximating $\mathbf{C}_{\text{NS},T}$ to further reduce computational cost.

Let $\mathbf{C}_{l,T} + \mathbf{C}_{s,T}$ be the FSA covariance matrix to the tapered global covariance matrix $\mathbf{C}_{\text{NS},T}$, where $\mathbf{C}_{l,T}$ is a reduced-rank matrix as defined in (8) and $\mathbf{C}_{s,T}$ is a sparse matrix as defined in (9). For spatial prediction, we use the Bayesian method that samples from the predictive distribution $p(Z(s_0)|\mathbf{Z})$ at a new site s_0 . For a given sample of Ω value, $p(Z(s_0)|\mathbf{x}(s_0), \Omega, \mathbf{Z})$ is a Gaussian distribution with the mean and the variance given by

$$E[Z(s_0)|\Omega, \mathbf{Z}] = \mathbf{x}^T(s_0)\boldsymbol{\beta} + \mathbf{h}^T(s_0)(\mathbf{C}_{l,T} + \mathbf{C}_{s,T} + \tau^2\mathbf{I}_n)^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}) \quad (14)$$

and

$$\text{var}[Z(s_0)|\Omega, \mathbf{Y}] = \sigma^2 - \mathbf{h}^T(s_0)(\mathbf{C}_{l,T} + \mathbf{C}_{s,T} + \tau^2 \mathbf{I}_n)^{-1} \mathbf{h}(s_0) + \tau^2, \quad (15)$$

where $\mathbf{h}^T(s_0) = [C_{l,T}(s_0, s_i) + C_{s,T}(s_0, s_i)]_{i=1}^N$, in which $C_{l,T}(s_0, s_i)$ and $C_{s,T}(s_0, s_i)$ are derived from the FSA covariance functions using $C_{\text{NS},T}$ as the parent process as described in (5) and (6).

Inversion of the matrix $\mathbf{C}_{l,T} + \mathbf{C}_{s,T} + \tau^2 \mathbf{I}$ appeared in (14)–(15) can be reduced to computations involving low-rank matrices and sparse matrices by applying the Sherman–Woodbury–Morrison formula presented in (9). Our simulation study shows that this is a fast and effective way to predict with high accuracy even within the boundaries of the subregions because the approximated global covariance is used for prediction.

We remark here that although the above prediction procedure is described conditional on a single tree, the prediction surface produced by our method is fairly smooth because we use the global covariance in some fashion. Following the Bayesian model average (BMA) method suggested in the classical BTGP, it is straightforward to further smooth the prediction surface by averaging over the predictive posterior samples conditional on a set tree samples. We expect this method will lead to better prediction performance. But it also comes with more expensive computational cost unless one takes advantage of multiple processors.

5. SIMULATION

In this section, we conduct a number of simulation studies to evaluate the performance of the covariance model described in Section 3. We randomly select 2100 locations from the spatial region $[0, 200] \times [0, 200]$ and 100 of them are held out to assess prediction performance. As shown in Figure 1(a), the study region is partitioned into two subregions with a line parallel to the x axis. We then simulate the spatial response $Y(s)$ at these 2100 locations from a spatial Gaussian regression model with the anisotropic nonstationary

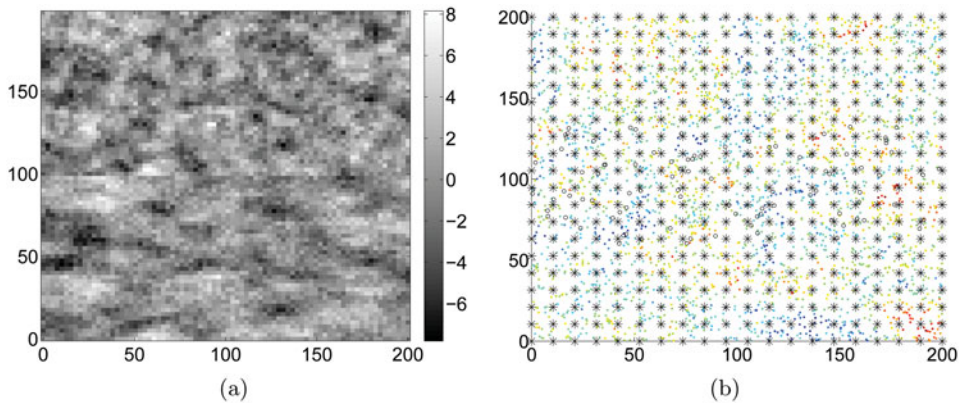


Figure 1. (a) Simulated random field, (b) simulated data (colored dots), hold out set (blue circles), and 400 knots (black stars).

Table 1. Posterior estimations of the model parameters and the MSPE

Param.	True	Full model	Separate FSA $m = 200, k = 20$	Joint FSA $m = 200, k = 20$	Separate PP $m = 200$
ψ_1	0.3	0.32(0.17)	0.25(0.18)	0.25(0.17)	0.56(0.22)
ψ_2	0.1	0.14(0.10)	0.16(0.11)	0.14(0.11)	0.14(0.11)
λ_{11}	70	79.88(16.02)	81.63(18.81)	81.02(18.71)	89.85(19.42)
λ_{12}	40	44.38(11.51)	49.69(12.07)	48.90(11.93)	57.75(15.83)
λ_{21}	10	9.07(3.41)	9.57(2.90)	9.47(3.03)	23.99(12.02)
λ_{22}	30	29.19(6.48)	35.44(8.04)	34.62(7.88)	37.82(13.68)
σ_1^2	5	4.44(0.41)	4.20(0.45)	4.37(0.44)	6.49(0.95)
σ_2^2	5	5.11(0.56)	5.02(0.48)	5.08(0.50)	8.21(1.43)
τ_1^2	1	1.04(0.21)	1.06(0.20)	1.04(0.20)	2.8(0.26)
τ_2^2	1	0.91(0.15)	0.85(0.17)	0.89(0.17)	3.18(0.21)
MSPE	—	3.2429	3.7890	3.7801	6.5644

covariance function as described in (3). The mean is modeled as $\mu(s) = 0$ for the entire region. The true parameter values are given in Table 1.

For the Bayesian posterior inference, priors are assigned as described in Section 4.

5.1 A SIMULATION WITH FIXED PARTITIONS

To examine the utility of using local fitting with the FSA for model inference and prediction, in the first simulation study we assume given partitions so that parameter estimations can be conveniently compared among various model inference methods. Specifically, in the *first approach* (denoted as “Full model”) we fit the global covariance model and do predictions using the full global covariance given in (3). In the *second approach* (denoted as “Separate FSA”) we use the local fitting for parameter estimation and global covariance for prediction using the FSA covariance. In the *third approach* (denoted as “Joint FSA”) we fit the global covariance model and do global prediction using the FSA method. Finally, for comparison purposes, we also include the fourth model (denoted as “Separate PP”) which is the local fitting for parameter estimation and global covariance for prediction using the predictive process (PP) approximation. Since the data points are uniformly distributed, for the FSA, we experiment with $m = 200$ regularly spaced knots and $k = 20$ blocks taken as equal-sized squares within each subregion.

And for the predictive process, we use the same 200 knots. For each method, we run 5000 iterations to collect posterior samples after a burn-in period of 1000 iterations. Good convergence of the respective marginal distributions is indicated by the trace plots of parameters.

Table 1 shows the Bayesian posterior sample means and standard deviations of the model parameters for each approach. In general, the posterior parameter estimations using the FSA are closer to the posterior parameter estimations using the full model than those using the PP approximation. Moreover the FSA yields slight loss in terms of mean square prediction errors (MSPE) whereas the MSPE using PP approximation is significantly larger than the MSPE under the full model.

Furthermore, we observe that the posterior parameter estimations from the local fitting method do not differ significantly from those obtained using the global covariance model fitting. As a result, the prediction performance of the *third approach* is only slightly better than that of the *second approach*. But the huge computational savings using the *second approach*, since we invert two smaller matrices, make it more appealing and practical to estimate covariance parameters in the presence of large datasets. On the other hand, prediction performance is very sensitive to the choice between doing prediction separately for each subregion and doing prediction jointly including all subregions. When we do both parameter estimations and predictions separately for each subregion, the MSPE is equal to 4.2622, which is much higher than the MSPE obtained from local fitting and global prediction. Because of the above findings, we will only concentrate on local fitting and global prediction in the case of known and unknown subregions and for real data analysis. Finally, we also consider the case in which isotropic Matérn correlation is used for local covariances. The MSPE value of this model using local fitting and global prediction with the FSA is 5.397, which is significantly larger than the MSPE of its anisotropic counterpart (MSPE=3.789). This finding clearly suggests that anisotropic nonstationary covariance function built by using the Mahalanobis distance leads to better performance.

5.2 A SIMULATION WITH TREED PARTITIONS

We fit the simulated data using local fitting with treed partitions based on full local covariance models, the FSA of local covariances models, and the PP approximation of local covariances models, respectively. For the FSA and the PP, we experiment with $m = 400$ knots. We fix the number of knots and place them in an equal-distance grid over the entire region, as shown in Figure 1. The green dots represent the locations where we simulate training data, the blue dots are the locations of the held-out data, and the red stars indicate the location of the knots kept unchanged over the MCMC iterations. To see the effect of the number of knots on the performance of model fitting, we also include the results for the FSA and the PP with a smaller number of knots ($m = 121$). For the FSA, we experiment with 25 blocks that are taken as equal-sized squares and remain unchanged.

We use the treed prior $0.6(1 + d_\eta)^{-2}$ as in Chipman, George, and McCulloch (1998). We also investigated the case when the treed prior is constant without using penalty, $P(\mathcal{T}) \propto 1$. Figure 2 shows the posterior distributions of the number of subregions for the full local covariance models under the aforementioned two treed priors, respectively. It seems that the use of the treed penalty helps the algorithm to converge to the right number of partitions. We next investigate how the use of the different covariance approximation methods (FSA and PP) affects the distribution of the number of subregions, that is, the number of the external nodes of a tree.

The distribution of the number of subregions using uniform prior, $P(\mathcal{T}) \propto 1$, and PP approximation with 121 and 400 knots are shown in Figure 3(a) and (c), respectively. The distribution of the number using the same prior and the FSA with 121 and 400 knots and 25 blocks are shown in Figure 3(b) and (d). The posterior distribution for the number of subregions using the treed prior is given in Figure 3(e) for the PP approximation with 400 knots and in Figure 3(f) for the FSA with 400 knots and 25 blocks.

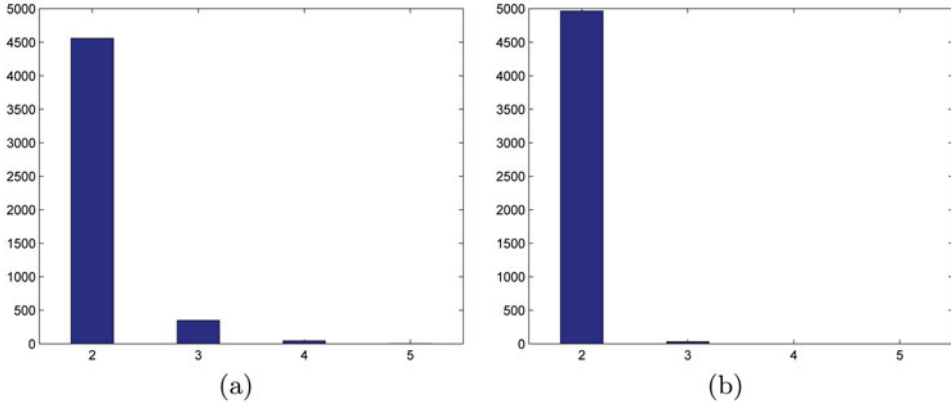


Figure 2. MCMC posterior distribution of the number of subregions when we use the full model: (a) without *treed* penalty (b) with a *treed* penalty.

Compared with the full model, both the FSA and the PP approaches entail a certain amount of loss in capturing the real number of subregions. The loss can be reduced by increasing the number of knots or decreasing the number of blocks. From the plots of the posterior distribution of the number of the subregions we can see that the number of knots does affect the estimations for space partitions. The PP with a small number of knots will usually overestimate the number of the subregions. For example, when we use 121 knots, the MCMC distribution for the number of subregions will capture the true number of partitions in only 48.8% of the MCMC draws. This is a weak performance compared to the posterior of the subregions using the full model. In contrast, the FSA with the same number of knots produces much closer results to that of the full model. When we increase the number of the knots to 400, we observe the PP has an increase of almost 20% in capturing the real situation but is still very ineffective compared to the FSA with the same number of knots. Moreover, we observe improved posterior estimations for the space partitions by adopting the tree prior with the penalty for the tree structure. For the PP with 400 knots, the MCMC distribution for the number of subregions captures the true situation in 70.8% of the MCMC draws. The use of the FSA with 400 knots and 25 blocks is very similar to the full model. In general, the FSA is preferred over PP approximation since the latter is more sensitive to the choice of the number of knots.

Finally, to demonstrate the prediction performance of using the *treed* model for space partition, we calculated the MSPE value based on predictions at the held-out test dataset. We include the full covariance model and the classical BTGP model as benchmarks to examine the performance of the FSA model. In summary, we consider five different cases as shown in Table 2: (a) global fitting and global prediction using full covariance model and assuming known correct partitions (denoted as “Full Model-Correct Partitions”); (b) classical BTGP for model fitting and prediction by averaging over predictions (denoted as “BTGP-BMA”); (c) local model fitting using BTGP and global prediction by averaging over predictions based on full covariance model (denoted as “BTGP-GBMA”); (d) local fitting and global prediction using the FSA and assuming a wrong partition (we take three equal-area subregions along the *y*-axis) (denoted as “FSA-Wrong Fixed Partition”); (e) local

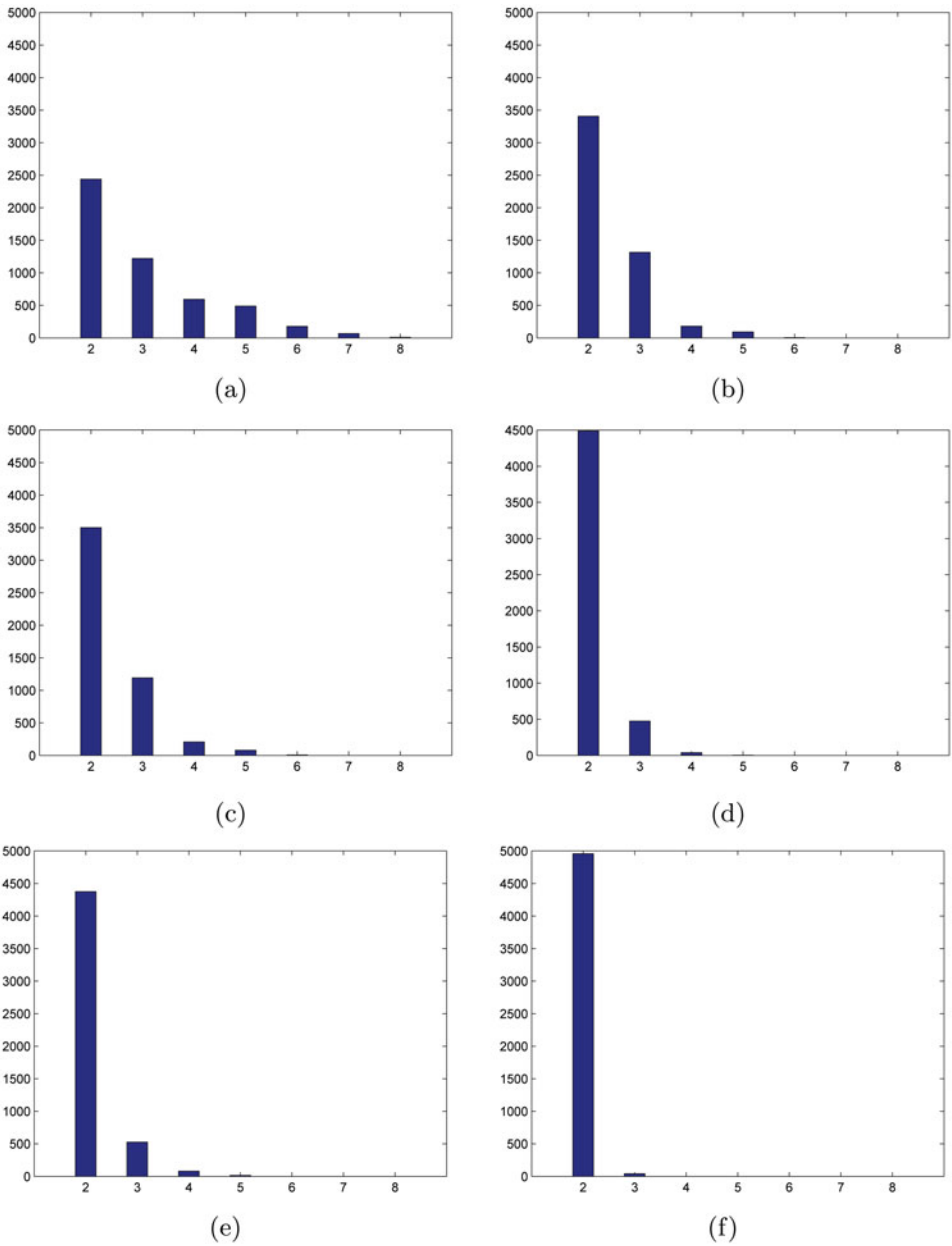


Figure 3. MCMC posterior distribution of the number of subregions: (a) PP with 121 knots and uniform prior; (b) FSA with 121 knots and blocks 25 and uniform prior; (c) predictive process with 400 knots and uniform prior; (d) FSA with 400 knots and 25 blocks and uniform prior; (e) PP with 400 knots and treed prior; (f) FSA with 400 knots and blocks 25 and treed prior.

fitting with stochastic treed partitions and global prediction conditional on the maximum a posteriori value of tree using the FSA (denoted as “FSA-MAP”).

The MSPE of these different approaches are given in Table 2. Not surprisingly, the case of known correct partitions with the full nonstationary covariance function gives us the

Table 2. MSPE for four different methods using FSA to approximate the covariance matrix

Method	MSPE
Full Model - Correct Partitions	3.243
BTGP - BMA	3.926
BTGP - GBMA	3.665
FSA - Misspecified Partitions	8.296
FSA - MAP	4.350

smallest MSPE = 3.243. The BTGP model using the full covariance for prediction yields the second best prediction performance, and the classical BTGP model gives the third best prediction performance. The MSPE difference between the classical BTGP and the BTGP, which uses dependent prediction, leads us to prefer the dependent form. However, these three approaches are not practical when the number of locations is at the order of 10^6 such as the satellite ozone data. The FSA model using stochastic treed partitions gives a slightly larger MSPE value than the BTGP. But this method has the advantage of scaling up to much higher dimensions as demonstrated in Section 6. The FSA model with stochastic treed partitions conditional on the MAP tree leads to slightly worse MSPE than in the case where the space partition is known (MSPE = 3.787 from Table 1), but much better prediction performance than the case of assuming wrong partition. In conclusion, in the case of unknown subregions, which is common in practice, the treed generating process model offers an automatic and reliable model-based way to partition space based on information provided by the dataset itself; whereas ad hoc-predetermined partitions could lead to unrobust and unsatisfying prediction performance.

6. DATA

TOMS Level 2 data are spatially and temporally irregular measurements of ozone following satellite scanning tracks (measurements are 8 sec apart) and there is a significant number of missing observations. TOMS Level 3 data are post processed from Level 2 data to regular grids (1 degree latitude by 1.25 degrees longitude for pixels with latitude from 50 S to 50 N, see Krueger et al. (1998) for more details) as daily averages. Although there is loss of information in Level 3 data, especially fine scale spatial and temporal variations, data on grids with global coverage and few missing observations reserve rich information on stratospheric ozone for scientific use. TOMS Level 3 data are obtained usually from ad-hoc methods to average Level 2 data pixel by pixel. The main difficulty with using a sophisticated statistical method lies in the huge computational cost associated with massive quantities of data.

To illustrate our method, we use the data collected by NIMBUS-7/TOMS satellite to measure the total column of ozone over the globe on October 1, 1988. To avoid huge variances and deal with not-very-good-quality data, we restrict our attention to pixels with latitude from 70 S to 70 N. Within the chosen latitude band we have 187,000 spatial data points from which we leave out 5000 observations as the test dataset uniformly

randomly chosen over the study region. These test data will be used to assess the prediction performance under different models.

We fit a spatial Gaussian regression model with the nonstationary covariance function as in (1). We represent the regression mean as linear combinations of spherical harmonics, which have been shown to be effective basis functions for capturing the large-scale patterns in the globe (Jun and Stein 2008). Specifically, we regress the ozone levels with $\mathbf{X}_m^n(\sin \vartheta, \varphi) | n = 0, 1, 2, \dots, m = -n, \dots, n$ for $n = 12$. This is similar to what has been done in Stein (2007); Jun and Stein (2008). The spherical harmonics terms capture most of the patterns in the mean and the results are not sensitive to the choice of n if it is larger than 12.

We model the spatial random effects \mathbf{w} in (1) as a zero mean Gaussian process with the nonstationary covariance function defined in (3). We model the local covariance using anisotropic Matérn defined on a sphere. The use of the covariance class defined in (2) for global data requires a valid distance metric to ensure the positive definiteness of the covariance function. The great circle distance is a natural distance metric on a sphere. But its use for spatial statistics on a sphere is troublesome since it cannot ensure the positive definiteness for many widely used correlation functions that are defined on Euclidean spaces (e.g., spherical and Matérn class). As an excellent approximation to the great circle distance, chordal distance (also called tunnel distance), measuring the great circle chord length between the points, becomes a popular choice (Banerjee 2005; Jun and Stein 2008) because it is indeed defined on the Euclidean space in \mathbb{R}^3 and thus ensures the validity of many correlation functions on a sphere. Specifically, for location i on the globe ($\text{lat}_i, \text{lon}_i$), where lat_i and lon_i denote latitude and longitude, we can convert the spherical coordinates to the three-dimensional Cartesian coordinates $\mathbf{s}_i = (r \cos(\text{lat}_i) \cos(\text{lon}_i), r \cos(\text{lat}_i) \sin(\text{lon}_i), r \sin(\text{lat}_i))$. Consider two locations on the globe with Cartesian coordinates \mathbf{s}_i and \mathbf{s}_j , respectively. The chordal distance is defined as the Euclidean metric $Q_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$. In addition, one may extend the Euclidean metric to define a Mahalanobis chordal distance

$$Q_{ij}^{NI} = \sqrt{(\mathbf{s}_i - \mathbf{s}_j)' \mathbf{B}^{-1} (\mathbf{s}_i - \mathbf{s}_j)}$$

as the distance used in (2) to model the anisotropy of the correlation functions.

Below, we investigate both the model with predetermined subregions and the model with adaptive subregions, and we compare their performance.

6.1 FIXED SUBREGIONS

For the model with fixed subregions, we first experimented with 50 disjoint equal subregions in a latitude range of $[-70, 70]$. This results in a moderate number of locations within each subregion, which allows us to fit each local model using full covariance and to compare with the local fitting results using the FSA and the PP. For simplicity, we used an isotropic covariance function within each subregion. For the FSA and the PP we chose 250 knots located on a uniform grid over each subregion domain. In addition for the FSA, we chose 60 blocks in each subregion. MCMC algorithms were run for each model for a total of 5000 iterations and posterior inference was based on the 1000 draws saved from every 4th iteration. We predict the values of the test data as described in Section 4.3 and

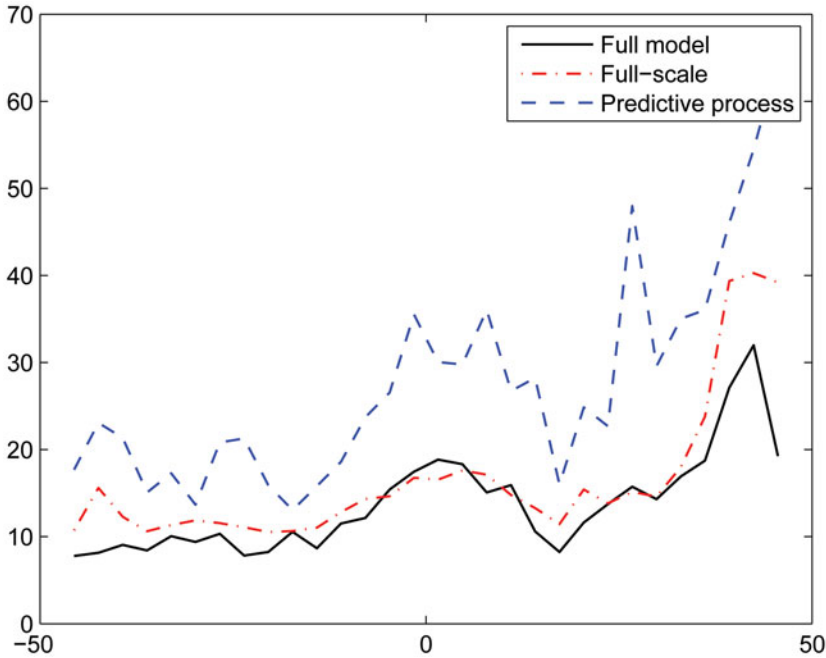


Figure 4. Comparing the MSPE for the three different methods as a function of latitude.

compute the MSPE from each subregion using the PP approximation, the FSA, and the full-model. The prediction of the training data have been computed using data from neighboring subregions. This reduces the prediction error as explained in Section 5 and maintains the low computational cost. The MSPE values of 30 subregions in the middle area are given in Figure 4. We exclude the MSPE for latitude bigger than 50 and smaller than -50 because they appear to have huge MSPE due to fewer data and big variance. (The data from latitude levels beyond this range have fewer than 288 observations per latitude.) The MSPE using the PP approximations is always larger than the MSPE using the FSA. In some cases, the MSPE using the PP is twice as big as the MSPE using the FSA. In general, the FSA gives comparable prediction performance with the full model (see Figure 4). Similar results are observed in latitude bigger than 50 and smaller than -50 .

Next, we consider taking 20 disjoint subregions obtained by equally partitioning the latitude range of $[-70^\circ, 70^\circ]$, aiming to compare the MSPE using anisotropic and isotropic covariance functions. In the simulation study, we have shown that the posterior estimations of the model parameters are not sensitive to the choice of global fitting or local fitting as long as the amount of data is relatively large within each subregion. For the ozone dataset, 20 partitions result in roughly 10,000 observations within each subregion, a size that contains rich information to learn local covariance structure and yet can be handled computationally using the FSA approach. We estimate the covariance parameters following the methods as described in Section 4.1. On the other hand, recall from the simulation study that the information from neighboring fields makes a significant contribution to the prediction accuracy, especially for data prediction near the boundary areas. We decide to include two

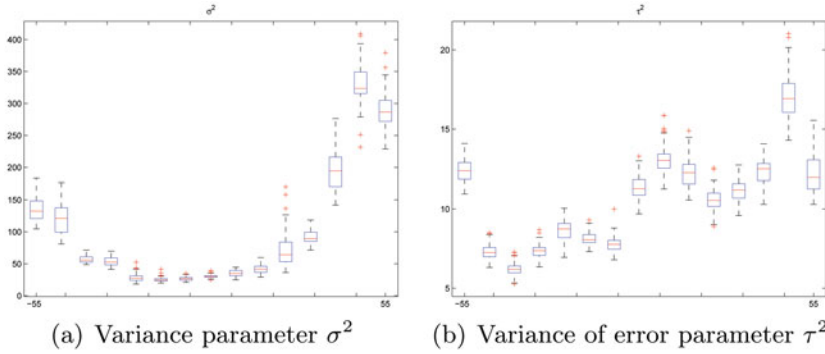


Figure 5. The posterior distribution of the variance parameter in the three-dimensional model as a function of latitude.

neighboring subregions and use the global covariance at all the selected points as defined in (3) for prediction as described in Section 4.3.

We use the FSA approach to fit the local covariance model. Since the observed locations are roughly evenly spaced over the globe, we chose 300 uniformly distributed regularly spaced knots and 40 equal blocks. Our pilot study based on a subset of data suggests that this choice of knots and blocks offers satisfying approximation to the full model. We then follow MCMC algorithms to run a total of 5000 iterations with a burn-in period of 1000. The boxplots of the posterior samples of the variance parameter are plotted in Figure 5. This figure clearly suggests that variation of ozone levels within a latitude is much lower near the equator than elsewhere. MSPE are calculated for a set of hold out test data as a criterion for model comparisons. Our result suggests that the anisotropic model offers superior prediction performance to the isotropic model based on 5000 held-out data. As shown in Figure 6, the MSPE using the anisotropic covariance matrix is in general smaller than the one using the isotropic covariance matrix, which indicates that the TOMS ozone data show strong geometric anisotropy. Indeed, it is known that stratospheric ozone, a major contributor to the TOMS ozone, is mainly produced at the tropics by solar UV radiation and then transported to middle and high latitudes by winds and broad circulation patterns (Solomon 1999). It is expected that local anisotropy can occur during this transport process.

6.2 ANISOTROPIC CASE WITH DYNAMICALLY CHOSEN SUBREGIONS

In this section, we fit an anisotropic model with treed generating process for space partitions under Bayesian framework. We choose a binary tree of depth five with 20 external nodes as an initial value in the MCMC run that gives 20 disjoint equal subregions and roughly 10,000 observations within each subregion. We make this choice mainly to avoid dealing with very large covariance model if we start with a very small tree (e.g., a single node tree). MCMC algorithms were run for each model for a total of 6000 iterations and with a burn-in of 1000 initial iterations. The movements and their associated proposal probabilities are taken to be similar to Chipman, George, and McCulloch (1998): growing a terminal node (0.25), pruning a pair of terminal nodes (0.25), changing a nonterminal rule (0.40), and swapping/rotating a rule between parent and child (0.10). To avoid computational

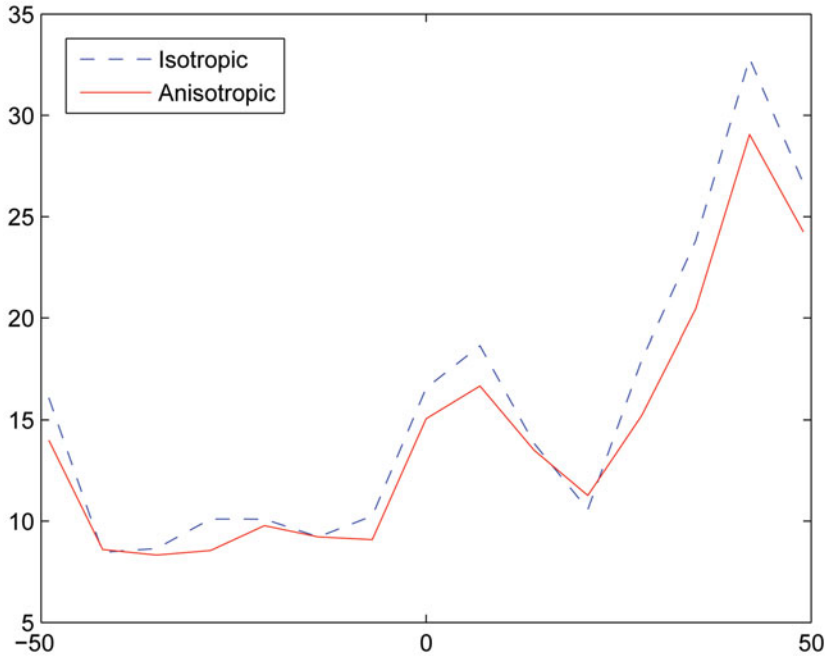


Figure 6. Comparing the MSPE of anisotropic with isotropic covariance.

complexity, the Bayesian tree is assumed to have independent subregions as it is described in Section 4. The prediction of the Level 3 values is done as described in Section 4.3 using the FSA reduction technique.

The FSA is used in both model fitting and parameter estimations as described in Section 4. For the FSA, we select 7000 knots uniformly distributed over the latitude band ranging from 70° S to 70° N and keep them fixed at every MCMC iteration. We also fix the number of blocks within each subregion as described in the previous section and keep them fixed in every MCMC iteration. The subregions may change when we update the treed partitions in the MCMC runs, resulting in a varying number of knots residing in each subregion. Generally, subregions with a smaller area contain fewer observations and fewer knots than larger subregions, which results in poorer local parameter estimations due to lack of observation and less accurate covariance approximations based on the FSA and the PP. To avoid this unbalanced performance across different subregions, we apply some restrictions when generating the partitions in the MCMC algorithm. Following an exploratory study, which was done without any restriction, we decide to take only subregions with at least 300 observations and 100 knots. These two restrictions are shown to be effective to avoid numerical instability arising from overly unbalanced space partitions. These are similar to the restriction in Bayesian treed Gaussian process (Gramacy and Lee 2008), with the difference that now we also introduce them to the knots. Usually restrictions on the knots implies restriction on the number of observations in each external node of the Bayesian tree. To ensure proper posterior distributions, we need a much smaller number of observations

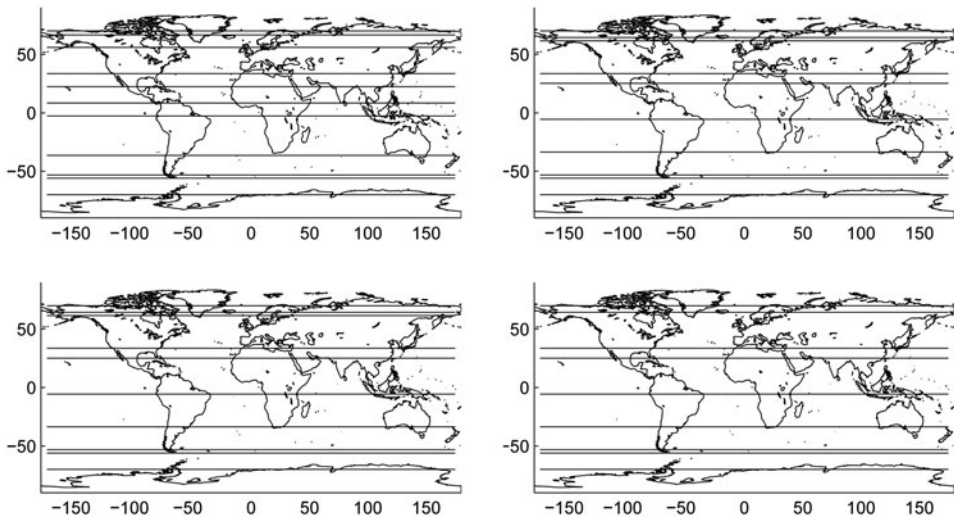


Figure 7. Four different partitions in the tree process.

and knots in each partition than the one we have chosen. However, the above restrictions are chosen to avoid numerical instabilities observed in practice.

By applying the method described in Section 4.2 we conclude that the number of subregions is between 8 and 11. More explicitly, after 5000 iterations and a burning period of 1000 iterations, we have approximately: 2% of the MCMC sample with 8 subregions, 68% of the MCMC sample with 9 subregions, 23% of the MCMC sample with 10 subregions, and 6% of the MCMC sample with 11 subregions. To better understand the treed partition of the space, we plot four posterior samples of the partitions in Figure 7. In addition, from the MCMC 5000 iterations we compute the MAP of the parameters and plot the MAP partition in Figure 8. As we can see from this figure, the area closer to the pole is partitioned into a relatively larger number of subregions, which may indicate stronger nonstationarity than

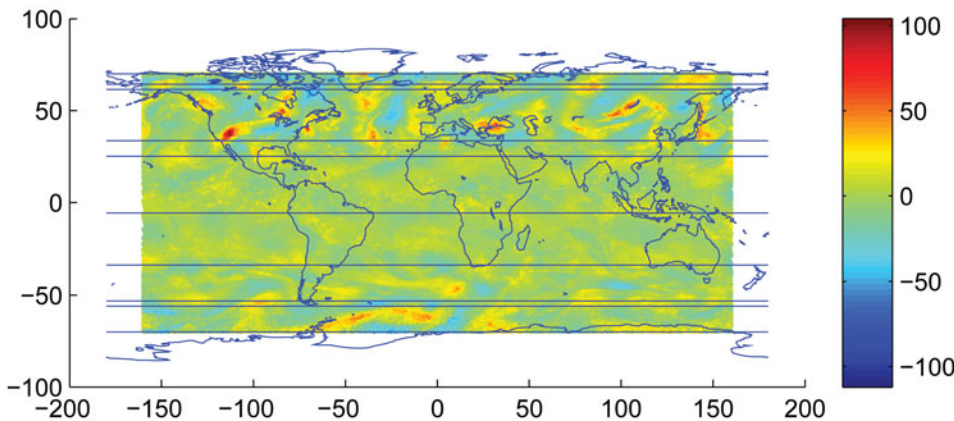


Figure 8. MAP estimation of the Bayesian treed GP.

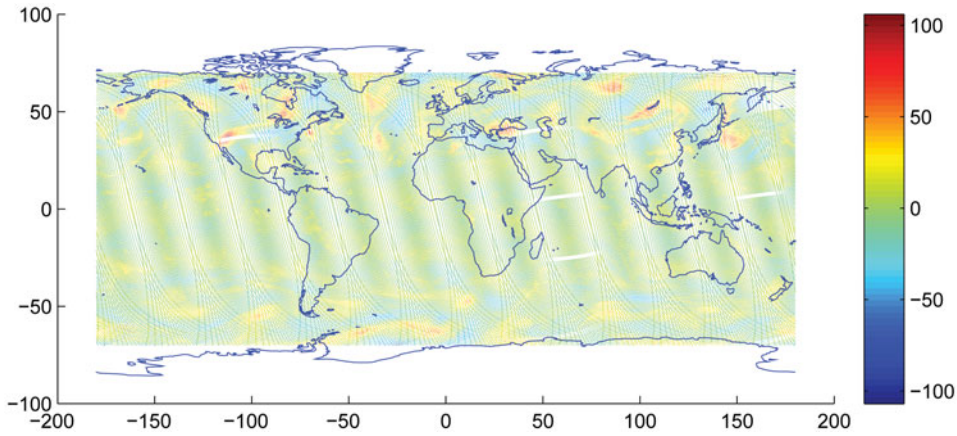


Figure 9. Level 2 “Real” residuals.

other areas of the globe. Especially between the latitude bound 60 to 70, we have three distinct subregions and in -70 to -55 we have two. Indeed, as shown in Figure 9, ozone is clearly more homogeneous near the tropical areas than near the poles.

The predicted Level 3 residuals of resolution 1×1.25 (1 degree latitude by 1.25 degrees longitude), using the MAP parameter estimation, are shown in Figure 10. The prediction of the training data have been computed using data from neighboring subregions using a tapering band of ± 10 for latitude and ± 5 longitude (this is chosen because the latitude correlation strength seems to be bigger). The predicted Level 3 residuals are very close to the real Level 2 residuals shown in Figure 9 and Figure 10 respectively. The red and blue patterns of the Level 3 and Level 2 data are similar in both graphs. As discussed in Section 4.3, we may further improve prediction performance by averaging the predictive surfaces over the sampled treed space. However, this is beyond the scope of the present article because of the computational constraint.

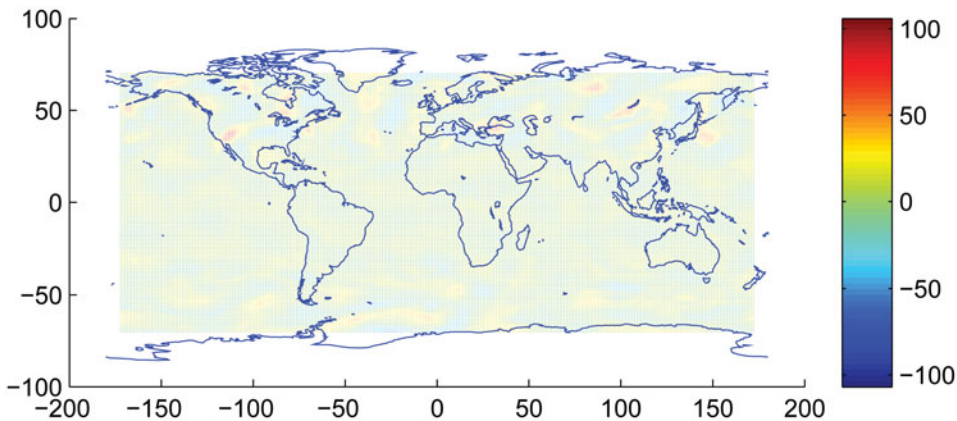


Figure 10. Level 3 Predicted residuals.

7. CONCLUDING REMARKS

In this article, we focus on modeling realistically massive global data with nonstationary GP models. The proposed method can be seen as a computationally efficient way of modeling nonstationarity and dealing with high-dimensional spatial datasets. We stochastically partition the space into smaller subregions with similar covariance structure. The local covariances are then linked together to produce a valid global nonstationary covariance function. To address the high dimensionality, we use the FSA approach to facilitate computations of massive covariance matrices. Our simulation study shows its utility in accurately estimating model parameters and producing reliable predictions compared with competing methods. For unknown partitions, we propose and develop a Bayesian treed generating model similar to the one used in the BTGP (Gramacy and Lee 2008). To make it computationally feasible, we apply the approximation techniques to simplify the MCMC operations. In addition, we improve the *grow*, *prune*, and *change* operation by considering dependence between subregions. Our results show that the use of treed generating model leads to substantially improved performance compared with the model with misspecified partitions.

The FSA approximation used in our study requires specifying knot and taper range or block size. In our studies, we chose knots and block size based on some exploratory analysis of a subset of data. It is desirable to use some model-based way to automatically choose them. A more comprehensive study of the selection of knots and taper range or block size is left for future research. There is research that chooses knots based upon minimization of a spatially averaged predictive variance criterion (e.g., Finley et al. 2009), or assigning priors to knots locations to adaptively select them (Guhaniyogi et al. 2011). We envision that these techniques would make the FSA approach more efficient. We will explore some strategies for knots and sparsity selections in future work.

A possible extension in model inference is to estimate the local covariance parameters at the locations of interest in a weighted or moving window fashion as proposed by Paciorek and Schervish (2006). We could use a moving window to include pairs of locations for which either one or both of the locations are near s and estimate or find the posterior of the parameters for the location s .

We apply the proposed algorithm to the TOMS data. Our results confirm that the TOMS data show strong patterns of anisotropy and nonstationarity. Specifically, the MCMC results indicate that the TOMS global data should be partitioned into roughly 8 to 11 subregions, where in each region similar anisotropic covariance structure can be assumed. For other types of global data, the nonstationarity may not only exist in the direction of latitude as in the TOMS data. Our method can be extended to such data by considering partitioning space beyond one dimension. In addition, another perhaps more interesting scientific extension is to investigate the vertical distribution of ozone and its connection with other atmospheric variables related to ozone chemistry and climate. NASA's newly launched instrument, the Ozone Mapper Profiler Suite (OMPS), is designed to measure the global distribution of the total atmospheric ozone column on a daily basis as well as the vertical distribution of ozone from about 15 km to 60 km. We envision the computational efficient method developed in this article will have some great potential to analyze such data.

SUPPLEMENTARY MATERIALS

Matlab-code for TGP_FSA: The “JCSG_code_TGP_FSA” file contains the main code (written in Matlab) to perform the adaptive Bayesian nonstationary with covariance approximations methods described in the article. The file also contains the datasets used as examples in the article. (JCSG_code_TGP_FSA.zip, zipped file)

ACKNOWLEDGMENTS

The research of Huiyan Sang was partially sponsored by National Science Foundation grant DMS-1007618 and the research of Bani Mallick was partially supported by NSF DMS 0914951. Bani Mallick and Huiyan Sang were also partially supported by award KUS-CI-016-04, made by King Abdullah University of Science and Technology. The authors thank the referees and the editors for valuable comments.

[Received May 2012. Revised May 2013.]

REFERENCES

- Anders, E. B., and Stein, M. L. (2008), “Estimating Deformations of Isotropic Gaussian Random Fields on the Plane,” *The Annals of Statistics*, 36, 719–741. [803]
- Banerjee, S. (2005), “On Geodetic Distance Computations in Spatial Modeling,” *Biometrics*, 61, 617–625. [821]
- Banerjee, S., Carlin, B., and Gelfand, A. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: Chapman & Hall-CRC. [805,806,811]
- Banerjee, S., Finley, A., Waldmann, P., and Ericsson, T. (2010), “Hierarchical Spatial Process Models for Multiple Traits in Large Genetic Trials,” *Journal of the American Statistical Association*, 105, 506–521. [810]
- Banerjee, S., Gelfand, A., Finley, A., and Sang, H. (2008), “Gaussian Predictive Process Models for Large Spatial Data Sets,” *Journal of the Royal Statistical Society, Series B*, 70, 825–848. [804,808,809]
- Calder, C. (2008), “A Dynamic Process Convolution Approach to Modeling Ambient Particulate Matter Concentrations,” *Environmetrics*, 19, 39–48. [803]
- Chipman, H., George, E., and McCulloch, R. (1998), “Bayesian CART Model Search,” *Journal of the American Statistical Association*, 93, 935–960. [807,811,812,817,823]
- Cressie, N., and Johannesson, G. (2008), “Fixed Rank Kriging for Very Large Spatial Data Sets,” *Journal of the Royal Statistical Society, Series B*, 70, 209–226. [804,808]
- Denison, D., Mallick, B., and Smith, A. (1998), “A Bayesian CART Algorithm,” *Biometrika*, 85, 363–377. [807]
- Finley, A., Sang, H., Banerjee, S., and Gelfand, A. (2009), “Improving the Performance of Predictive Process Modeling for Large Datasets,” *Computational Statistics and Data Analysis*, 53, 2873–2884. [804,808,827]
- Fuentes, M. (2001), “A High Frequency Kriging Approach for Non-Stationary Environmental Processes,” *Environmetrics*, 12, 469–483. [803]
- Furrer, R., Genton, M., and Nychka, D. (2006), “Covariance Tapering for Interpolation of Large Spatial Datasets,” *Journal of Computational and Graphical Statistics*, 15, 502–523. [804,808]
- Gneiting, T. (2002), “Compactly Supported Correlation Functions,” *Journal of Multivariate Analysis*, 83, 493–508. [804,809]
- Gramacy, R. B., and Lee, H. K. H. (2008), “Bayesian Treed Gaussian Process Models With an Application to Computer Modeling,” *Journal of the American Statistical Association*, 103, 1119–1130. [804,807,812,813,824,827]
- Green, P. (1995), “Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination,” *Biometrika*, 82, 711–732. [812]
- Green, P., and Sibson, R. (1978), “Computing Dirichlet Tessellations in Plane,” *Computer Journal*, 21, 168–173. [809]

- Guhaniyogi, R., Finley, A., Banerjee, S., and Gelfand, A. (2011), “Adaptive Gaussian Predictive Process Models for Large Spatial Datasets,” *Environmetrics*, 22, 997–1007. [827]
- Higdon, D. (1998), “A Process-Convolution Approach to Modeling Temperatures in the North Atlantic Ocean,” *Journal of Environmental and Ecological Statistics*, 5, 173–190. [803]
- Higdon, D. M., Swall, J., and Kern, J. C. (2011), “Non-Stationary Spatial Modeling,” in *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, Oxford University Press, UK. [803]
- Hjort, N., and Omre, H. (1994), “Topics in Spatial Statistics” (with discussion, comments and rejoinder), *Scandinavian Journal of Statistics*, 289–357. [804,814]
- Jun, M., and Stein, M. L. (2008), “Nonstationary Covariance Models for Global Data,” *Annals of Applied Statistics*, 2, 1271–1289. [803,805,813,821]
- Kaufman, C., Schervish, M., and Nychka, D. (2008), “Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets,” *Journal of the American Statistical Association*, 103, 1545–1555. [804,808]
- Kim, H.-M., Mallick, B., and Holmes, C. (2005), “Analyzing Nonstationary Spatial Data Using a Piecewise Gaussian Process,” *Journal of the American Statistical Association*, 470, 653–658. [807]
- Krueger, A. J., Bhartia, P. K., McPeters, R. D., Herman, J. R., Wellemeyer, C. G., Jaross, G., Seftor, C. J., Torres, O., Labow, G., Byerly, W., Taylor, S. L., Swisler, T., and Cebula, R. P. (1998), “ADEOS Total Ozone Mapping Spectrometer (TOMS) Data Products Users Guide,” available at http://toms.gsfc.nasa.gov/datainfo/adeos_userguide.pdf. [803,820]
- Mueller, P. (1993), “Alternatives to the Gibbs Sampling Scheme,” Technical Report, Institute of Statistics and Decision Sciences, Duke University. [811]
- Nychka, D., and Royle, C. W. K. A. (2002), “Multiresolution Models for Nonstationary Spatial Covariance Functions,” *Statistical Modelling*, 2, 315–332. [803]
- Paciorek, C., and Schervish, M. (2006), “Spatial Modelling Using a New Class of Nonstationary Covariance Functions,” *Environmetrics*, 17, 483–506. [803,804,806,811,827]
- Sampson, P. D., and Guttorp, P. (1992), “Nonparametric Estimation of Nonstationary Spatial Covariance Structure,” *Journal of the American Statistical Association*, 87, 108–119. [803]
- 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, 19–741. [804,808,810,814]
- Schabenberger, O., and Gotway, C. (2005), *Statistical Methods for Spatial Data Analysis* (2nd ed.), Boca Raton, FL: Chapman & Hall. [805]
- 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. [803]
- Smith, A. E. G. A. F. M. (1990), “Sampling-Based Approaches to Calculating Marginal Densities,” *Journal of the American Statistical Association*, 85, 398–409. [811]
- Smith, R. (2001), “Environmental Statistics,” Technical Report, Department of Statistics, University of North Carolina, Chapel Hill. [803]
- Solomon, S. (1999), “Stratospheric Ozone Depletion: A Review of Concepts and History,” *Reviews of Geophysics—Richmond Virginia Then Washington*, 37, 275–316. [823]
- Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging* (2nd ed.), New York: Springer. [805]
- (2007), “Spatial Variation of Total Column Ozone on a Global Scale,” *Annals of Applied Statistics*, 1, 191–210. [805,821]
- (2008), “A Modeling Approach for Large Spatial Datasets,” *Journal of the Korean Statistical Society*, 37, 3–10. [804,805,808]
- Trenberth, K. E., and Shea, D. J. (2005), “Atlantic Hurricanes and Natural Variability in 2005,” *Geophysical Research Letters*, 33, L12704. [805]
- Wendland, H. (1998), “Error Estimates for Interpolation by Compactly Supported Radial Basis Functions of Minimal Degree,” *Journal of Approximation Theory*, 93, 258–272. [809]