Received: 26 March 2012, Revised: 20 December 2012, Accepted: 21 December 2012, Published online in Wiley Online Library: 11 February 2013

(wileyonlinelibrary.com) DOI: 10.1002/env.2200

Bayesian nonstationary spatial modeling for very large datasets

Matthias Katzfuss[a]*

With the proliferation of modern high-resolution measuring instruments mounted on satellites, planes, ground-based vehicles, and monitoring stations, a need has arisen for statistical methods suitable for the analysis of large spatial datasets observed on large spatial domains. Statistical analyses of such datasets provide two main challenges: first, traditional spatial-statistical techniques are often unable to handle large numbers of observations in a computationally feasible way; second, for large and heterogeneous spatial domains, it is often not appropriate to assume that a process of interest is stationary over the entire domain. We address the first challenge by using a model combining a low-rank component, which allows for flexible modeling of medium-to-long-range dependence via a set of spatial basis functions, with a tapered remainder component, which allows for modeling of local dependence using a compactly supported covariance function. Addressing the second challenge, we propose two extensions to this model that result in increased flexibility: first, the model is parameterized on the basis of a nonstationary Matérn covariance, where the parameters vary smoothly across space; second, in our fully Bayesian model, all components and parameters are considered random, including the number, locations, and shapes of the basis functions used in the low-rank component. Using simulated data and a real-world dataset of high-resolution soil measurements, we show that both extensions can result in substantial improvements over the current state-of-the-art.

Keywords: covariance tapering; full-scale approximation; low-rank models; massive datasets; model selection; reversible-jump MCMC

1. INTRODUCTION

From remote sensing of environmental variables using satellite instruments to proximal sensing of soil properties using a ground-based gamma-radiometer, a vast number of spatial measurements are now being obtained every day. On the basis of such very large, noisy, non-gridded, and incomplete datasets, the goal is spatial prediction of a process of interest, together with rigorous quantification of prediction uncertainty. Computational feasibility for such datasets have been addressed from several angles: approximations by Gaussian Markov random fields (e.g., Lindgren *et al*., 2011), composite likelihoods (e.g., Lindsay, 1988; Curriero and Lele, 1999; Bevilacqua *et al*., 2012; Eidsvik *et al*., 2012), covariance tapering, and low-rank models. We focus here on the latter two.

Covariance tapering (Furrer *et al*., 2006; Kaufman *et al*., 2008; Shaby and Ruppert, 2012) relies on compactly supported correlation functions (e.g., Gneiting, 2002) to produce sparse covariance matrices containing only a moderate number of non-zero elements. Use of efficient sparse-matrix algorithms then may result in computational feasibility for large datasets. However, by definition, covariance tapering is most appropriate for modeling processes with weak long-range dependence.

A second approach to achieving computational feasibility for large spatial datasets is through low-rank models, which include a component that can be written as a linear combination of spatial basis functions,

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

where $\boldsymbol{\eta}|\mathbf{W} \sim N_r(\mathbf{0}, \mathbf{W})$, and the number of basis functions, r, is much smaller than the number of observations. Many models that include such a component have been proposed (for a recent overview, see Wikle, 2010). The models differ in the parameterizations and priors for the covariance matrix $\mathbf{W}$ and the functions in $\mathbf{b}(\cdot)$. For discretized convolution models (i.e., convolution models whose integrals are discretized; see, e.g., Higdon, 1998; Calder, 2007; Lemos and Sansó, 2009), $\mathbf{b}(\cdot)$ contains the convolution kernels, and $\mathbf{W}$ is often assumed to be a multiple of the identity. Other authors view $\mathbf{b}(\cdot)$ as a vector of fixed basis functions, such as empirical orthogonal functions (e.g. Mardia *et al*., 1998; Wikle and Cressie, 1999), equatorial normal modes (e.g., Wikle *et al*., 2001), Fourier basis functions (e.g., Xu *et al*., 2005), W-wavelets (e.g., Shi and Cressie, 2007; Cressie *et al*., 2010; Kang and Cressie, 2011), or bisquare functions (e.g., Cressie and Johannesson, 2008; Katzfuss and Cressie, 2011, 2012). Here, we use the predictive-process approach (Banerjee *et al*., 2008), where both $\mathbf{b}(\cdot)$ and $\mathbf{W}$ are parameterized according to a "parent process," for which a parametric covariance model is chosen.

* *Correspondence to: Matthias Katzfuss, Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany. E-mail: katzfuss@gmail.com*

[a] *Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany*

Models with low-rank components (1) allow for fast computation via the Sherman–Morrison–Woodbury formula, as is made clear in Cressie and Johannesson (2006) and Shi and Cressie (2007). For general $\mathbf{W}$, they are also flexible, in that the covariance of (1), namely $\mathbf{b}(\mathbf{s}_1)'\mathbf{W}\,\mathbf{b}(\mathbf{s}_2)$ for locations $\mathbf{s}_1$ and $\mathbf{s}_2$, is not of traditional parametric form. The fast computation and the flexibility make components of the form (1) very well suited to modeling medium-range to long-range spatial dependence. However, because of the dimension reduction inherent in (1), a low-rank component alone is typically not able to model "rough" (i.e., non-smooth) short-range dependence (see, e.g., Stein, 2008; Finley *et al*., 2009). Some efforts have been made to address this problem (e.g., Wikle and Cressie, 1999; Berliner *et al*., 2000; Wikle *et al*., 2001; Stein, 2008), including in the context of the predictive process (Katzfuss, 2011, ch. 4; Sang *et al*., 2011; Sang and Huang, 2012). Here, we follow the approach of Sang and Huang (2012), who divide a parent process into a predictive-process component and a remainder component. The covariance matrix of the remainder component is then made sparse by multiplication of its covariance function with a compactly supported tapering function. This approach allows for computationally feasible inference, even for large datasets.

The contributions of this article are two extensions of the approach by Sang and Huang (2012), which allow for more flexibility and nonstationarity. First, we specify a nonstationary Matérn model (Paciorek and Schervish, 2006; Stein, 2005) for the parent covariance, in which the parameters vary smoothly across space as linear combinations of spatial basis functions.

The second extension is that we allow the set of basis-function locations (henceforth referred to as "knots") in our low-rank component to be a random point process. This allows us to avoid choosing an arbitrary and fixed set of knots a priori. Here, $\mathbf{b}(\cdot)$, $\boldsymbol{\eta}$, and $\mathbf{W}$ in (1) are all treated as unknown and random. This Bayesian source separation task (see, e.g., Knuth, 2005), where both the "source signal" $\boldsymbol{\eta}$ and the "mixing coefficients" $\mathbf{b}(\cdot)$ have to be estimated, can be achieved by putting a prior on both components. This has been done in the context of discretized-convolution models by Lemos and Sansó (2009), who infer (spatially varying) parameters determining the shapes of their kernels. Lopes *et al.* (2008) also consider a model of the form (1) where both $\mathbf{b}(\cdot)$ and $\boldsymbol{\eta}$ are random, but as each basis function is itself a Gaussian process, their approach is infeasible for large spatial datasets. Recently, Guhaniyogi *et al.* (2011) also proposed a predictive-process model where the locations (but not the number) of the basis functions are assumed random. In this article, we implicitly make inference on the number, locations, and shapes of the basis functions. Our approach is a special case of that in Katzfuss (2011, ch. 4) and is inspired by Holmes and Mallick (2001), who propose a piecewise linear spline regression model for which both the number and the locations of the splines are random.

A third contribution of this article is partially philosophical in nature: we do not consider the parent process to be the truth that is to be approximated, but rather as a way of obtaining a prior for the two spatially dependent components in our model. The resulting process is more flexible than the parent process, and hence it is often preferable for modeling nonstationary real-world processes.

Posterior inference for our model is described in detail. It is fairly involved but computationally feasible, even for very large datasets. A reversible-jump Markov chain Monte Carlo algorithm (Green, 1995) allows us to infer the number of basis functions. We take advantage of sparse-matrix operations to ensure fast computation, and we employ marginalization strategies (e.g., van Dyk and Park, 2008) to achieve satisfactory mixing of the Markov chain.

This article is organized as follows: In Section 2, we introduce our nonstationary spatial model based on the model of Sang and Huang (2012). Section 3 deals with posterior inference on the unknown quantities in the model. In Section 4, we assess the effect of our extensions to the approach of Sang and Huang (2012), using simulated data and a real-world dataset of soil measurements. Conclusions are given in Section 5.

2. METHODOLOGY

2.1. A standard spatial statistical model

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

$$Z(\mathbf{s}_i) := Y(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \quad i = 1, \ldots, n, \tag{2}$$

where $\epsilon(\cdot)|\sigma_\epsilon^2 \sim \mathrm{GWN}\left(0, \sigma_\epsilon^2\right)$ is Gaussian white noise and independent of $Y(\cdot)$. For simplicity and to ensure identifiability throughout this article, we will assume that σ_ϵ^2 is fixed and known. In practice, if σ_ϵ^2 is not known (e.g., from instrument experiments), it can be estimated from the data by extrapolating the variogram to the origin as described in Kang *et al.* (2009).

In spatial statistics, the process model is often given by,

$$Y(\cdot) := \mu(\cdot) + \omega(\cdot), \tag{3}$$

where $\mu(\cdot) := \mathbf{x}(\cdot)'\boldsymbol{\beta}$ is the large-scale trend, $\boldsymbol{\beta}$ has an (improper) flat prior on $\mathbb{R}^p$, and $\omega(\cdot)$ is a spatially correlated component, which is typically modeled as a Gaussian process,

$$\omega(\cdot)|\boldsymbol{\theta} \sim \mathrm{GP}(0, C_P), \tag{4}$$

with mean zero and covariance function

$$C_P(\mathbf{s}_1, \mathbf{s}_2) = \sigma(\mathbf{s}_1)\sigma(\mathbf{s}_2)\,\rho_P(\mathbf{s}_1, \mathbf{s}_2), \quad \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}, \tag{5}$$

where $\sigma: \mathcal{D} \to \mathbb{R}_0^+$ and the correlation function $\rho_P: (\mathcal{D} \times \mathcal{D}) \to [-1, 1]$ are parameterized by $\boldsymbol{\theta}$.

2.2. A low-rank component with random basis functions

Although the standard spatial model described in Section 2.1 has been used extensively and successfully (e.g., Banerjee *et al*., 2004), it is computationally infeasible if n is very large (more than 10,000 or so) and C_P is a standard covariance function (e.g., the exponential covariance function). This is because it takes on the order of n^3 computations to evaluate the likelihood.

Many approximations or modeling approaches have been proposed to solve this problem (Section 1). We will focus here on the predictive process (Banerjee *et al.*, 2008). Given a so-called "parent process" $\omega(\cdot)$ as in (4), the predictive process is defined as, $\nu(\cdot) := E(\omega(\cdot)|\omega(\mathbf{k}_1), \ldots, \omega(\mathbf{k}_r))$, where

$$\mathcal{K} := \{\mathbf{k}_1, \ldots, \mathbf{k}_r\}, \quad \text{with } \mathbf{k}_j \in \mathcal{D},\ j = 1, \ldots, r, \tag{6}$$

is a set of knots. Conditional on $\boldsymbol{\theta}$ and $\mathcal{K}$, the predictive process can be written as a linear combination of basis functions, namely as $\nu(\cdot) = \mathbf{b}(\cdot)'\boldsymbol{\eta}$ with $\boldsymbol{\eta} \sim N_r(\mathbf{0}, \mathbf{W})$, where now

$$\mathbf{b}(\mathbf{s}) := \sigma(\mathbf{s})\ (\rho_P(\mathbf{s}, \mathbf{k}_1), \ldots, \rho_P(\mathbf{s}, \mathbf{k}_r))', \quad \mathbf{s} \in \mathcal{D}, \tag{7}$$

and $\mathbf{W} := \left((\rho_P(\mathbf{k}_i, \mathbf{k}_j))_{i,j=1,\ldots,r}\right)^{-1}$. Thus, we have $\nu(\cdot)|\boldsymbol{\theta}, \mathcal{K} \sim \text{GP}(0, C_\nu)$, where $C_\nu(\mathbf{s}_1, \mathbf{s}_2) := \mathbf{b}(\mathbf{s}_1)'\mathbf{W}\mathbf{b}(\mathbf{s}_2)$, $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}$.

In what follows, we do not choose a fixed set of knots $\mathcal{K}$ in (6). Instead, we model $\mathcal{K}$ as a random point process. As discussed later at the end of Section 3.2, it is not necessary to strongly penalize large numbers of basis functions, r, through the prior on $\mathcal{K}$. Thus, we assume a flat, noninformative, improper prior for $\mathcal{K}$ with density proportional to 1.

2.3. Adding a tapered remainder component

It was pointed out by Finley *et al.* (2009) that the predictive process can only account for smooth dependence. Hence, as in Sang and Huang (2012), we write

$$\omega(\cdot) = \nu(\cdot) + (\omega(\cdot) - \nu(\cdot)) =: \nu(\cdot) + \tilde{\delta}(\cdot). \tag{8}$$

Then $\tilde{\delta}(\cdot) = \omega(\cdot) - \nu(\cdot)$ is independent of $\nu(\cdot)$, and $\tilde{\delta}(\cdot) \sim \text{GP}(0, C_{\tilde{\delta}})$, where $C_{\tilde{\delta}}(\cdot,\cdot) = C_P(\cdot,\cdot) - C_\nu(\cdot,\cdot)$ is a valid covariance function. To achieve computational feasibility for large n, Sang and Huang (2012) proposed to replace $\tilde{\delta}(\cdot)$ in (8) by $\delta(\cdot) \sim \text{GP}(0, C_\delta)$, where

$$C_\delta(\mathbf{s}_1, \mathbf{s}_2) = \mathcal{T}(\|\mathbf{s}_1 - \mathbf{s}_2\|/L)\, C_{\tilde{\delta}}(\mathbf{s}_1, \mathbf{s}_2), \quad \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D} \tag{9}$$

is a tapered version of $C_{\tilde{\delta}}$. In (9), $\mathcal{T}(\cdot)$ is a compactly supported correlation function (e.g., Gneiting, 2002) that is equal to zero when its argument is greater than one. Multiplication of $C_{\tilde{\delta}}$ with $\mathcal{T}$ achieves that $C_\delta(\mathbf{s}_1, \mathbf{s}_2) = 0$ if $\|\mathbf{s}_1 - \mathbf{s}_2\| \geqslant L$, resulting in a covariance matrix that is sparse and quickly invertible (see Section 3.4 below). We will assume the tapering length L to be fixed and chosen to ensure computational feasibility.

In summary, our data model is given by (2), and our process model is given by

$$Y(\cdot) = \mathbf{x}(\cdot)'\boldsymbol{\beta} + \nu(\cdot) + \delta(\cdot), \tag{10}$$

where $\nu(\cdot)$ describes the medium-range to long-range spatial dependence, and $\delta(\cdot)$ accounts for local (or short-range) dependence. Both $\nu(\cdot)$ and $\delta(\cdot)$ are zero-mean Gaussian processes, whose covariance functions depend on a random set of knots, $\mathcal{K}$, with a flat prior distribution, and on a parent covariance function, C_P, parameterized by $\boldsymbol{\theta}$ and specified further in Section 2.4 as follows.

2.4. The parent covariance function

Let $\mathcal{M}_\upsilon$ denote the Matérn correlation function (Stein, 1999, p. 50),

$$\mathcal{M}_\upsilon(h) = \left(2h\sqrt{\upsilon}\right)^\upsilon \mathcal{K}_\upsilon\left(2h\sqrt{\upsilon}\right) 2^{1-\upsilon}/\Gamma(\upsilon), \quad h > 0, \tag{11}$$

and $\mathcal{M}_\upsilon(0) = 1$, where $\mathcal{K}_\upsilon$ is the modified Bessel function of the second kind of order $\upsilon > 0$. Also, let

$$q(\mathbf{s}_1, \mathbf{s}_2) = \left\{2(\mathbf{s}_1 - \mathbf{s}_2)'(\boldsymbol{\Sigma}_A(\mathbf{s}_1) + \boldsymbol{\Sigma}_A(\mathbf{s}_2))^{-1}(\mathbf{s}_1 - \mathbf{s}_2)\right\}^{1/2}, \quad \mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^d,\ d \in \mathbb{N}, \tag{12}$$

be a spatially varying (SV) Mahalanobis-like distance, where $\boldsymbol{\Sigma}_A(\mathbf{s})$ is a $d \times d$ positive-definite matrix describing (local) geometric anisotropy at location $\mathbf{s}$. We write, $\boldsymbol{\Sigma}_A(\mathbf{s}) := \boldsymbol{\mathcal{R}}(\mathbf{s})\,\boldsymbol{\Gamma}(\mathbf{s})\,\boldsymbol{\mathcal{R}}(\mathbf{s})'$, where $\boldsymbol{\Gamma}(\mathbf{s}) := \text{diag}\{\gamma_1(\mathbf{s}), \ldots, \gamma_d(\mathbf{s})\}$, $\{\gamma_j : \mathcal{D} \to \mathbb{R}^+,\ j = 1, \ldots, d\}$ are SV scale parameters, and $\boldsymbol{\mathcal{R}}$ is a rotation matrix parameterized by SV rotation angles $\{\kappa_j : \mathcal{D} \to [0, \pi/2],\ j = 1, \ldots, d-1\}$. A valid nonstationary Matérn correlation function (Paciorek and Schervish, 2006; Stein, 2005) is given by,

$$\widetilde{\mathcal{M}}(\mathbf{s}_1, \mathbf{s}_2) = c(\mathbf{s}_1, \mathbf{s}_2)\mathcal{M}_{(\upsilon(\mathbf{s}_1)+\upsilon(\mathbf{s}_2))/2}(q(\mathbf{s}_1, \mathbf{s}_2)), \quad \mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^d,\ d \in \mathbb{N}, \tag{13}$$

where $c(\mathbf{s}_1, \mathbf{s}_2) := |\boldsymbol{\Sigma}_A(\mathbf{s}_1)|^{1/4}|\boldsymbol{\Sigma}_A(\mathbf{s}_2)|^{1/4}|(\boldsymbol{\Sigma}_A(\mathbf{s}_1) + \boldsymbol{\Sigma}_A(\mathbf{s}_2))/2|^{-1/2}$.

Table 1. Details for the spatially varying covariance parameters of the form (15)

Parameter	Symbol $\theta(\cdot)$	Range of $\theta(\cdot)$	Transformation $g_\theta(\cdot)$	μ_θ	σ_θ^2
Standard deviation	$\sigma(\cdot)$	$\mathbb{R}^+$	$\exp(\cdot)$	(*)	$\sigma_\sigma^2 = 0.25$
Smoothness	$\upsilon(\cdot)$	$[0, 2]$	$2\Phi(\cdot)$	$\mu_\upsilon = 0$	$\sigma_\upsilon^2 = 1$
Scale	$\gamma_j(\cdot)$	$\mathbb{R}^+$	$\exp(\cdot)$	(*)	$\sigma_\gamma^2 = 0.25$
Rotation angle	$\kappa_j(\cdot)$	$[0, \pi/2]$	$(\pi/2)\Phi(\cdot)$	$\mu_\kappa = 0$	$\sigma_\kappa^2 = 1$

$\Phi(\cdot)$: Cumulative distribution function of the standard normal distribution.
(*): The prior means μ_σ and μ_γ depend on the application; see Section 4 for some specific choices.

Choosing $\rho_P := \widetilde{\mathcal{M}}$ in (5) results in the parent covariance

$$C_P(\mathbf{s}_1, \mathbf{s}_2) = \sigma(\mathbf{s}_1)\sigma(\mathbf{s}_2)\widetilde{\mathcal{M}}(\mathbf{s}_1, \mathbf{s}_2), \quad \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D} \subset \mathbb{R}^d, \ d \in \mathbb{N}. \tag{14}$$

This nonstationary Matérn class is very flexible, in that it allows for SV standard deviation $\sigma(\cdot)$, SV smoothness parameter $\upsilon(\cdot)$, and SV geometric anisotropy through SV scale parameters $\{\gamma_j(\cdot)\colon\ j = 1, \ldots, d\}$ and SV rotation angles $\{\kappa_j(\cdot)\colon\ j = 1, \ldots, d-1\}$.

To ensure computational feasibility, we let the parameters vary spatially according to linear combinations of spatial basis functions. We assume that all SV parameters are determined by the (random) parameter vector, $\boldsymbol{\theta} := \left(\tilde{\sigma}, \boldsymbol{\eta}_\sigma', \tilde{\upsilon}, \boldsymbol{\eta}_\upsilon', \tilde{\boldsymbol{\gamma}}', \boldsymbol{\eta}_{\boldsymbol{\gamma}}', \tilde{\boldsymbol{\kappa}}', \boldsymbol{\eta}_{\boldsymbol{\kappa}}'\right)'$, through models of the form,

$$\theta(\mathbf{s}) = g_\theta\left(\tilde{\theta} + \mathbf{b}_\theta(\mathbf{s})'\boldsymbol{\eta}_\theta\right), \quad \mathbf{s} \in \mathcal{D}, \tag{15}$$

where $\theta(\cdot)$ is a generic notation for one of the SV parameters, $\tilde{\theta} \sim N\left(\mu_\theta, \sigma_\theta^2\right)$, $\boldsymbol{\eta}_\theta \sim N_{r_\theta}\left(\mathbf{0}, \tau_\theta^2 \mathbf{I}_{r_\theta}\right)$, and $\mathbf{b}_\theta(\cdot)$ is an r_θ-dimensional vector of *fixed* basis functions (same for all parameters), each normalized to [0, 1]. The functions $g_\theta(\cdot)$ are transformations from $\mathbb{R}$ to the range of $\theta(\cdot)$.

Specific choices for $g_\theta(\cdot)$, μ_θ, and σ_θ^2 are given in Table 1. For example, we have $\sigma(\mathbf{s}) = \exp\left(\tilde{\sigma} + \mathbf{b}_\theta(\mathbf{s})'\boldsymbol{\eta}_\sigma\right)$, $\tilde{\sigma} \sim N\left(\mu_\sigma, \sigma_\sigma^2{=}0.25\right)$, and $\boldsymbol{\eta}_\sigma \sim N_{r_\theta}\left(\mathbf{0}, \tau_\theta^2 \mathbf{I}_{r_\theta}\right)$. Note that we restrict the smoothness parameter $\upsilon(\cdot)$ to the interval $[0, 2]$, as "the data can rarely inform about smoothness of higher orders" (Banerjee *et al.*, 2008). The parameter τ_θ^2 determines how much $\theta(\cdot)$ is allowed to vary *a priori* over $\mathcal{D}$; we set $\tau_\theta^2 = (0.25)^2$ for all SV parameters (see Katzfuss and Cressie, 2012), inducing shrinkage towards stationarity for the covariance function C_P.

For $\mathbf{b}_\theta(\cdot)$ in (15), any choice of basis functions is possible. Assuming that the covariance parameters vary smoothly over space, we choose a relatively small number of power exponential correlation functions, $\mathbf{b}_\theta(\mathbf{s}) = \left(\exp\{-((\mathbf{s} - \mathbf{c}_1)/\lambda)^2\}, \ldots, \exp\left\{-((\mathbf{s} - \mathbf{c}_{r_\theta})/\lambda)^2\right\}\right)'$, with (relatively large) fixed scale parameter λ, and fixed centers $\mathbf{c}_1, \ldots, \mathbf{c}_{r_\theta}$. Specific choices depend on the domain $\mathcal{D}$ and are given in Section 4.

Restricting ourselves to $\mathcal{D} \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, for the remainder of this article, our choice for $\mathcal{T}$ in (9) is Kanter's function (Kanter, 1997):

$$\mathcal{T}(x) := (1 - x)\tfrac{\sin(2\pi x)}{2\pi x} + \tfrac{1 - \cos(2\pi x)}{2\pi^2 x}, \quad x \in (0, 1), \tag{16}$$

$\mathcal{T}(x) := 0$ for $x \geqslant 1$, and we set $\mathcal{T}(0) := 1$. The function $\mathcal{T}(\|\mathbf{h}\|)$ is positive-definite for $\mathbf{h} \in \mathbb{R}^3$, it is twice differentiable at the origin, and it minimizes the curvature at 0 within the class of all compactly supported and valid (in $\mathbb{R}^3$) correlation functions (Gneiting, 2002).

In summary, for fixed $\boldsymbol{\beta}$, $\mathcal{K}$, and $\boldsymbol{\theta}$, the covariance function of the true process $Y(\cdot)$ in (10) is

$$C_Y(\mathbf{s}_1, \mathbf{s}_2) = C_\nu(\mathbf{s}_1, \mathbf{s}_2) + \mathcal{T}(\|\mathbf{s}_1 - \mathbf{s}_2\|/L)\{C_P(\mathbf{s}_1, \mathbf{s}_2) - C_\nu(\mathbf{s}_1, \mathbf{s}_2)\}, \quad \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}, \tag{17}$$

where C_ν and $\mathcal{T}$ are given by (7) and (16), respectively. It follows immediately from Proposition 1 in Sang and Huang (2012) that this covariance function is positive definite. It is a close approximation to $C_P(\cdot, \cdot)$ for a large and dense set of knots, $\mathcal{K}$ (or for large L). Here, because $\mathcal{K}$ is random, (17) is more flexible than the parent covariance and hence preferable in many nonstationary real-world situations. Note that, because $\sigma(\cdot)$ is infinitely differentiable, $T(\cdot)$ is twice differentiable at the origin, and $\widetilde{\mathcal{M}}(\mathbf{s}, \mathbf{s} + \mathbf{h})$ is also at most twice differentiable for $\upsilon(\mathbf{s}) < 2$ (see also Paciorek and Schervish, 2006), the smoothness of $Y(\cdot)$ at location $\mathbf{s} \in \mathcal{D}$ is solely determined by $\upsilon(\mathbf{s})$ (for fixed $\boldsymbol{\beta}$, $\mathcal{K}$, and $\boldsymbol{\theta}$).

3. POSTERIOR INFERENCE

3.1. Summary of the model in vector notation

Integrating out $\boldsymbol{\eta}$ and $\delta(\cdot)$, the data, $\mathbf{Z} := (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$, are distributed as $\mathbf{Z}|\Omega \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_Z)$, where $\Omega := \{\boldsymbol{\beta}, \boldsymbol{\theta}, \mathcal{K}\}$, and the ith row of $\mathbf{X}$ is given by $\mathbf{x}(\mathbf{s}_i)'$. The data covariance matrix is

$$\boldsymbol{\Sigma}_Z := \mathrm{var}(\mathbf{Z}|\Omega) = \mathbf{B}\mathbf{W}\mathbf{B}' + \mathbf{V}, \tag{18}$$

where the ith row of the $n \times r$ matrix $\mathbf{B}$ is given by $\mathbf{b}(\mathbf{s}_i)'$ (see (7)), $\mathbf{W}$ is defined near (7), $\mathbf{V} := \mathbf{V}_\delta + \mathbf{V}_\epsilon$, $\mathbf{V}_\epsilon := \sigma_\epsilon^2 \mathbf{I}_n$, and $\mathbf{V}_\delta := (C_\delta(\mathbf{s}_i, \mathbf{s}_j))_{i,j=1,\ldots,n}$ is the sparse $n \times n$ covariance matrix of the vector $\boldsymbol{\delta} := (\delta(\mathbf{s}_1), \ldots, \delta(\mathbf{s}_n))'$ (see (9)).

In what is to follow, $[A]$ will denote the distribution or density of a generic random variable A, and $[A \mid \ldots]$ will denote the full conditional distribution of A (i.e., the distribution of A given the data and all parameters other than A in Ω). Further, let $N_k(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the probability density function of a k-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, evaluated at $\mathbf{a}$.

The densities of the full conditional distributions of the elements of Ω are all proportional to

$$[\mathbf{Z}, \Omega] = [\mathbf{Z}|\Omega][\Omega] = N_n(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_Z)[\boldsymbol{\beta}][\boldsymbol{\theta}][\mathcal{K}],$$

where $[\mathcal{K}] \propto 1$, $[\boldsymbol{\beta}] \propto 1$, and $[\boldsymbol{\theta}]$ is described near (15).

3.2. The reversible-jump Markov chain Monte Carlo algorithm

For posterior inference, we will employ a reversible jump Markov chain Monte Carlo (MCMC) algorithm (Green, 1995) based on a Gibbs sampler (Geman and Geman, 1984) with some adaptive Metropolis–Hastings steps (Metropolis *et al.*, 1953; Hastings, 1970; Haario *et al.*, 2001). We will emphasize dependence of $\boldsymbol{\Sigma}_Z$ on a set of parameters by placing the parameters in parentheses.

The MCMC sampler consists of the following steps:

1. Sample $\boldsymbol{\beta}$ from, $[\boldsymbol{\beta}| \ldots] = N_p\left(\boldsymbol{\beta}|\,(\mathbf{X}'\boldsymbol{\Sigma}_Z^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\boldsymbol{\Sigma}_Z^{-1}\mathbf{Z}, (\mathbf{X}'\boldsymbol{\Sigma}_Z^{-1}\mathbf{X})^{-1}\right)$.
2. Sample $\boldsymbol{\theta}$ using a Metropolis–Hastings step from, $[\boldsymbol{\theta}| \ldots] \propto [\boldsymbol{\theta}]\, N_n\,(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_Z(\boldsymbol{\theta}))$.
3. Sample a new set of knots from $[\mathcal{K}| \ldots]$, as follows. At each MCMC iteration, we propose one of three modifications to the current set of knots, each with probability $1/3$:
   (a) Add a knot: Draw a new knot, $\mathbf{k}_{r+1}$, from a uniform distribution on $\mathcal{D}$, and let $\mathcal{K}^* := \mathcal{K} \cup \{\mathbf{k}_{r+1}\}$ be the proposed set of knots, which now has size $r^* = r + 1$.
   (b) Delete a knot: Select one knot uniformly at random from $\mathcal{K}$; that is, draw $J \sim U(1, 2, \ldots, r)$. Then set $\mathcal{K}^* := \mathcal{K} \backslash \{\mathbf{k}_J\}$ and $r^* = r - 1$.
   (c) Moving a knot (a combination of (a) and (b)): First select a knot uniformly at random to be deleted, and then select a location uniformly on $\mathcal{D}$ at which to add a new one (i.e., where to move the old knot). This results in $\mathcal{K}^* := \{\mathbf{k}_{r+1}\} \cup \mathcal{K} \backslash \{\mathbf{k}_J\}$ and $r^* = r$.

   The reversible-jump acceptance probability (Green, 1995) for the proposed $\mathcal{K}^*$ can be shown to be equal to $\min\{1, \alpha\}$, where

   $$\alpha := \frac{N_n(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_Z(\mathcal{K}^*))}{N_n(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_Z(\mathcal{K}))} \frac{\mathcal{Q}(\mathcal{K}^*, \mathcal{K})}{\mathcal{Q}(\mathcal{K}, \mathcal{K}^*)}, \tag{19}$$

   and the proposal ratio is given by

   $$\frac{\mathcal{Q}_{\mathcal{K}}(\mathcal{K}^*, \mathcal{K})}{\mathcal{Q}_{\mathcal{K}}(\mathcal{K}, \mathcal{K}^*)} := \begin{cases} 1/(r+1), & r^* = r + 1 \\ r, & r^* = r - 1 \\ 1, & r^* = r. \end{cases} \tag{20}$$

   Note that for $r = 0$, deleting or moving a basis function is impossible, and so, in this case, we always propose to add a basis function. As a result, the proposal ratio in (20) is given by $1/3$ when $r = 0$.

There might be a concern that, for very large datasets, the data always favor a very large number of basis functions, unless there is strong penalization for large r through the prior distribution on $\mathcal{K}$. However, note that the acceptance probability (19) for a proposed set of knots, $\mathcal{K}^*$, is the product of the Bayes factor (of $\mathcal{K}^*$ versus $\mathcal{K}$) and a term depending only on the proposal distribution chosen for $\mathcal{K}^*$ (cf. Holmes and Mallick, 2000, App. I). This is reassuring, as "the Bayes factor functions as a fully automatic Occam's razor" (Kass and Raftery, 1995, p. 790), and so there is strong intuition that our flat prior, $[\mathcal{K}] \propto 1$, is sufficient and that no explicit penalty for large r is necessary.

3.3. Spatial prediction

In spatial statistics, the main interest is typically in making inference on the true process $Y(\cdot)$ at a set of prediction locations, $\left\{\mathbf{s}_1^P, \ldots, \mathbf{s}_{n_P}^P\right\}$, which might or might not include the set of observed locations. We write, $\mathbf{Y}^P = \mathbf{X}^P\boldsymbol{\beta} + \mathbf{B}^P\boldsymbol{\eta} + \boldsymbol{\delta}^P$, and so we need samples from

$$[\Omega, \boldsymbol{\eta}, \boldsymbol{\delta}^P|\mathbf{Z}] = [\Omega|\mathbf{Z}]\,[\boldsymbol{\eta}|\Omega, \mathbf{Z}]\,[\boldsymbol{\delta}^P|\boldsymbol{\eta}, \Omega, \mathbf{Z}], \tag{21}$$

where samples of the first term on the right-hand side were obtained in Section 3.2. Because it can be very computationally expensive, we only obtain samples of $\boldsymbol{\eta}$ and $\boldsymbol{\delta}^P$ for thinned MCMC iterations after convergence of the MCMC for Ω (see van Dyk and Park, 2008, for why this is valid). We have

$$\boldsymbol{\eta}|\Omega, \mathbf{Z} \sim N_r\left((\mathbf{B}'\mathbf{V}^{-1}\mathbf{B} + \mathbf{W}^{-1})^{-1}\mathbf{B}'\mathbf{V}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}), (\mathbf{B}'\mathbf{V}^{-1}\mathbf{B} + \mathbf{W}^{-1})^{-1}\right)$$

and

$$\boldsymbol{\delta}^P|\boldsymbol{\eta}, \Omega, \mathbf{Z} \sim N_{n_P}\left(\mathbf{V}_\delta^{P,O}\mathbf{V}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\boldsymbol{\eta}), \mathbf{V}_\delta^P - \mathbf{V}_\delta^{P,O}\mathbf{V}^{-1}\mathbf{V}_\delta^{P,O\prime}\right), \tag{22}$$

where $\mathbf{V}_{\delta}^{P} := \mathrm{var}(\boldsymbol{\delta}^{P})$, $\mathbf{V}_{\delta}^{P,O} := \mathrm{cov}(\boldsymbol{\delta}^{P}, \boldsymbol{\delta})$, and $\boldsymbol{\delta} := (\delta(\mathbf{s}_1), \ldots, \delta(\mathbf{s}_n))'$. After appropriate reordering, we write $\boldsymbol{\delta}^{P} = [\boldsymbol{\delta}', \boldsymbol{\delta}^{U\prime}]'$, where $\boldsymbol{\delta}^{U}$ denotes $\delta(\cdot)$ evaluated at all unobserved prediction locations. To avoid having to obtain $\mathbf{V}_{\delta}^{P,O}\mathbf{V}^{-1}\mathbf{V}_{\delta}^{P,O\prime}$ explicitly, we obtain a sample from (22) by calculating the quantity, $\check{\boldsymbol{\delta}}^{P} + \mathbf{V}_{\delta}^{P,O}\mathbf{V}^{-1}\left(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\boldsymbol{\eta} - \check{\boldsymbol{\delta}} - \check{\boldsymbol{\epsilon}}\right)$, where $\check{\boldsymbol{\delta}}^{P} := \left(\check{\boldsymbol{\delta}}', \check{\boldsymbol{\delta}}^{U\prime}\right)' \sim N_{n_P}\left(\mathbf{0}, \mathbf{V}_{\delta}^{P}\right)$ and $\check{\boldsymbol{\epsilon}} \sim N_n(\mathbf{0}, \mathbf{V}_{\epsilon})$ (cf. conditional simulation, Cressie, 1993, Section 3.6.2).

3.4. Computational issues

Note that $\boldsymbol{\Sigma}_Z$ in (18) is a dense $n \times n$ matrix of full rank n, and so naive calculation of its inverse, which appears in the MCMC updates, is computationally infeasible for large n. However, we can employ the Sherman–Morrison–Woodbury formula (Sherman and Morrison, 1950; Woodbury, 1950; Henderson and Searle, 1981) to obtain, $\boldsymbol{\Sigma}_Z^{-1} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{B}(\mathbf{W}^{-1} + \mathbf{B}'\mathbf{V}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{V}^{-1}$, and a similar formula gives $|\boldsymbol{\Sigma}_Z| = |\mathbf{V}||\mathbf{I}_r + \mathbf{W}\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}|$ (e.g., Cressie and Johannesson, 2008). Because the tapering range, L in (9), is fixed, the position of the non-zero elements of $\mathbf{V}$ is the same for all MCMC iterations. Hence, we order the locations to allow for efficient Cholesky decomposition of $\mathbf{V}$ (e.g., using the minimum-degree ordering) only once, at the beginning of the MCMC algorithm.

In general, the number of computations required for operations involving a sparse matrix depends on the number and locations of the non-zero elements (Gilbert *et al.*, 1992). Some numerical results in Furrer *et al.* (2006) indicate that the time required to compute the Cholesky decomposition of a tapered $n \times n$ covariance matrix increases roughly linearly with n (for fixed domain, fixed tapering length, and a regular sampling grid), which in turn indicates that the computational complexity of our algorithm is approximately of order n. Questions about theoretical computational complexity aside, in our experience, the majority of computation time at each of the MCMC iterations was actually spent on evaluating the modified Bessel function in (11) for each of the non-zero elements of the matrix $\mathbf{V}_{\delta}$ (and of $\mathbf{V}_{\delta}^{P}$ and $\mathbf{V}_{\delta}^{P,O}$ for iterations in which spatial predictions are obtained). We have considerable control over the speed of the MCMC algorithm through selection of the tapering range L in (9). For extremely massive datasets, we can set L to a very small value to achieve computational feasibility.

4. NUMERICAL MODEL COMPARISONS

In this section, we will compare our model with the model of Sang and Huang (2012), which represents the current state-of-the-art in terms of geostatistical approaches to the analysis of large spatial datasets. Sang and Huang (2012) showed that their model can result in better predictions and model fit than the predictive-process approach of Banerjee *et al.* (2008). Our approach can be viewed as an extension of the Sang and Huang (2012) model in terms of two components: random knots and the use of the nonstationary Matérn covariance of Section 2.4 as the parent covariance function. Therefore, our comparisons will examine the effects of two factors: random versus (a varying number of) fixed knots, and a nonstationary Matérn parent covariance (NPC) versus a stationary one (SPC). (SPC is a special case of our model obtained by setting $\tau_{\theta}^{2} = 0$ in (15).)

4.1. Simulation studies in one spatial dimension

For the following three simulation studies, the true process is assumed to exist on a one-dimensional domain, $\mathcal{D} = [1, 512]$, with potential measurement locations at $\{1, 2, \ldots, 512\}$.

In Simulation Study 1, we assumed that the true process $Y(\cdot)$ is a deterministic function:

$$Y(s) = 1 + \sin\left(2\pi\left(\frac{s-306}{512}\right)^2\right)\sin\left(20\pi\left(\frac{s-50}{512}\right)^2\right), \quad s \in \mathcal{D}. \tag{23}$$

This true process is shown in Figure 1.

On the basis of this true process, we created 100 datasets of observations by adding independent normal measurement error with variance $\sigma_{\epsilon}^{2} = \hat{\sigma}_{Y}^{2} \cdot 5\% = 0.004$, where $\hat{\sigma}_{Y}^{2} = 0.08$ is the empirical variance of $\{Y(1), \ldots, Y(512)\}$. To examine the medium-to-long-range prediction performance of the models, we created four test intervals, in which no data was observed (collectively referred to as missing by design, or MBD). These test intervals each have length 25 and begin at locations 70, 198, 326, 454, respectively. In addition, one third of the remaining locations (henceforth referred to as missing at random, or MAR) were selected at random at each iteration of the simulation study as unobserved test locations (to test short-range prediction performance near observed locations). The remaining 275 observed locations will be denoted OBS.

To ensure comparability of the results, we assumed the measurement-error variance to be known for all models. For each of the 100 simulated datasets, each of the models was run for 10,000 MCMC iterations (thinned by a factor of 10), the first 5000 of which were taken as burn-in (based on examination of trace plots in a pilot study). The tapering length in (9) was chosen as $L = 6.5$, resulting in about 2400 non-zero elements (less than nine per row) for $\mathbf{V}$ in (18). The prior distributions of the parameters of the parent covariance were as described in Section 2.4, with $\mu_{\sigma} = \log(\hat{\sigma}_{Y})$ and $\mu_{\gamma} = \log(3000)$. The spatial trend, $\mu(\cdot)$ in (3), consisted only of an intercept (i.e., $\mathbf{x}(\cdot) \equiv 1$).

For the random knots, the proposal distribution for new knots (as described in step 3 of Section 3.2) was a uniform distribution on $[-9, 522]$. As a pilot study, using our model showed that the posterior mean of r was around 11; we used two sets of fixed knots: the first consisted of eight evenly spaced knots between locations -10 and 522, and the second set consisted of 14 evenly spaced knots between -4 and 516. For the models with nonstationary parent covariance, we took $\mathbf{b}_{\theta}(\cdot)$ in (15) to be made up of four power exponential functions with scale parameter $\lambda = 74$, centered at locations 64, 192, 320, and 448, respectively. One set of observations, $\mathbf{Z}$, is shown in Figure 1, together with a summary of the corresponding results using our model. Very few knots are selected between locations 380 and 500, because the process fluctuates so quickly in that area that it can basically be picked up entirely by the tapered remainder component, $\delta(\cdot)$.

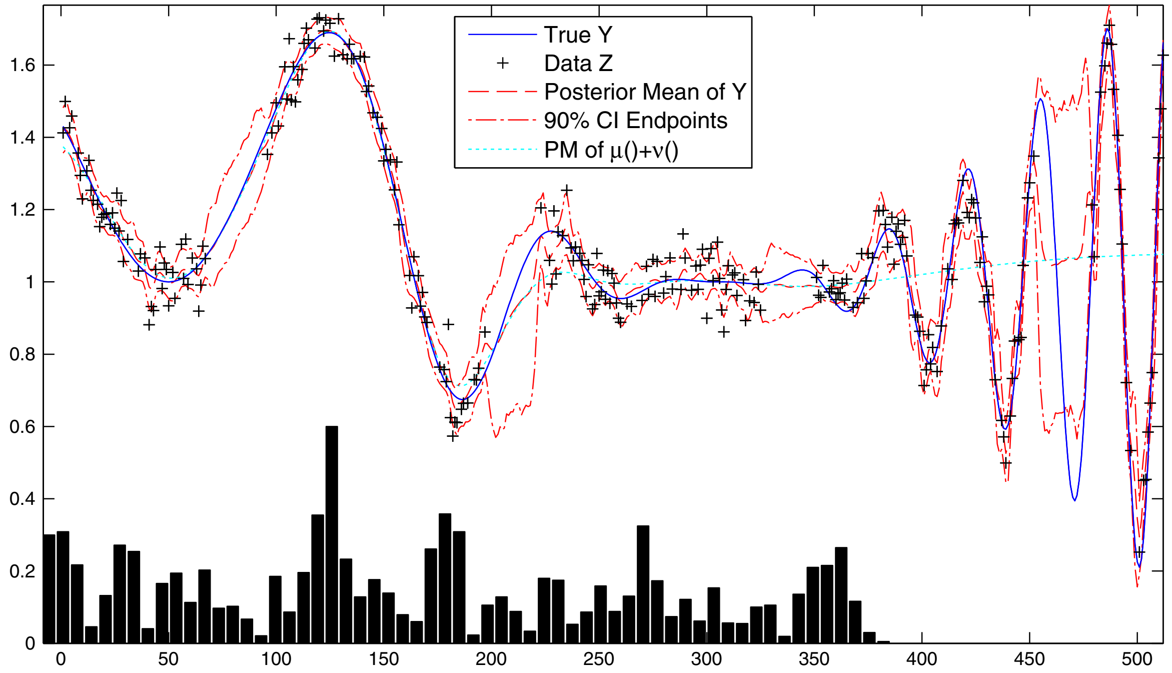

Figure 1. For Simulation Study 1, the true process $Y(\cdot)$ and one sample of the data $\mathbf{Z}$, together with the posterior mean and a point-wise posterior 90% credible interval of $Y(\cdot)$, the posterior mean of $\mu(\cdot)+\nu(\cdot)$ (i.e., without $\delta(\cdot)$), and the density of the knots (histogram at the bottom) using our model

Table 2. Results of Simulation Study 1

	Random knots		8 Fixed knots		14 Fixed knots		Full model	
Parent covariance	NPC	SPC	NPC	SPC	NPC	SPC	NPC	SPC
Time (min)	3.64	3.79	2.01	2.01	2.76	2.73	101.17	100.05
MSPE (ALL) ×100	1.00	1.02	1.19	1.25	1.38	1.56	1.48	1.57
MSPE (OBS) ×100	0.11	0.11	0.18	0.21	0.13	0.19	0.14	0.18
MSPE (MAR) ×100	0.24	0.25	0.51	0.57	0.36	0.48	0.23	0.28
MSPE (MBD) ×100	4.48	4.58	4.87	5.05	6.24	6.80	6.88	7.17
IS (ALL) ×100	26.71	28.45	33.16	47.28	45.65	72.84	62.94	80.60
IS (OBS) ×100	15.54	15.77	20.11	21.77	17.11	20.99	18.30	22.42
IS (MAR) ×100	21.64	21.93	33.01	38.56	27.16	36.52	23.92	29.77
IS (MBD) ×100	64.40	72.26	69.23	129.36	149.43	265.21	239.18	310.26
Posterior mean of r	10.59	11.24	(8)	(8)	(14)	(14)	(275)	(275)

NPC: nonstationary parent covariance; SPC: stationary parent covariance; MSPE: mean square prediction error; IS: interval score; Time: average time for each MCMC (averaged over the 100 simulated datasets); ALL: all 512 locations; OBS: the 275 observed locations; MAR: the 137 missing-at-random locations; MBD: the 100 missing-by-design locations.

To measure prediction accuracy of the models under consideration, we used the mean squared prediction error, the (average) squared difference between the true process $Y(\cdot)$, and the posterior mean for each of the models. To quantify the accuracy of the uncertainty estimation, we also calculated the interval score, which combines the width of a credible interval (here, 95% posterior credible intervals) with a penalty for not containing the true value (see Gneiting and Raftery, 2007, Section 6.2). The goal is for a small interval score. Both mean squared prediction error and interval score were averaged over the 100 simulated datasets and all 512 locations (ALL), and also averaged within each of the groups of locations described earlier (OBS, MAR, MBD).

The results for Simulation Study 1 are shown in Table 2. Two trends are evident in terms of both scores: using an NPC produced better predictions than using an SPC, and random knots resulted in better predictions when compared with fixed knots. The more fixed knots were used (we even included the full model, for which $r=n$), the closer the resulting models were to their respective parent processes (which are clearly the wrong models for $Y(\cdot)$ in (23)), and the worse the scores were for the test regions (MBD).

Throughout this article, we assumed that most real-world processes do not have covariances of simple parametric form. To examine the performance of our model in the unlikely event of encountering a process that does exhibit simple parametric covariance, we conducted two more simulation studies. In Simulation Studies 2 and 3, we sampled a new true process $Y(\cdot)$ 100 times each as a Gaussian process with constant mean 1 and the Matérn covariance function of (14). For Simulation Study 2, we chose

$$\begin{aligned} \sigma(s) &= 3\exp\left(\sin((1-|s/256-1|)\,2\pi)/2\right) \\ \gamma(s) &= 600\exp\left(-2\sin(s\,2\pi/256)\right)(s/256) \\ \upsilon(s) &= 3\Phi\left(-\sin(s\,2\pi/256)\right), \end{aligned} \tag{24}$$

and for Simulation Study 3, we used a stationary Matérn covariance with $\sigma(s) \equiv 3$, $\gamma(s) \equiv 600$, and $\upsilon(s) \equiv 1$. At each of the 100 iterations, we then simulated data, $\mathbf{Z}$, by adding a spatially independent measurement-error term with variance $\sigma_\epsilon^2 = 3^2 \cdot 5\% = 0.45$ at each observed location. The remaining setup was exactly the same as in Simulation Study 1, except that we chose $\mu_\sigma = \log(3)$ and $\mu_\gamma = \log(600)$.

In Simulation Study 2 (Table 3), the NPC models worked better than the corresponding SPC models (as expected, because the true $Y(\cdot)$ was nonstationary). Overall, random knots resulted in better predictions than fixed knots, especially for the (misspecified) SPC models.

For Simulation Study 3 (Table 4), NPC still worked slightly better than SPC, suggesting that there is no penalty in terms of predictive distributions for using the (more flexible) NPC model when the true process is stationary. The models with random knots again had the best results.

For all three simulation studies, the models with random knots resulted in longer computation times than the models with eight or 14 fixed knots.

4.2. Analysis of soil readings from a gamma-radiometer

We also compared the models of Section 4.1 using a large real-world spatial dataset. Viscarra Rossel *et al.* (2007) collected high-resolution soil information on Nowley farm in New South Wales, Australia. It is important to develop automated soil sensing for monitoring and precision agriculture, because conventional soil sampling is far too costly to be routinely used on a large scale.

Specifically, Viscarra Rossel *et al.* (2007) obtained 34,266 gamma-ray readings using a gamma-radiometer mounted on the front of a four-wheel-drive vehicle. After some preprocessing, they smoothed the data using "local kriging" and carried out a multivariate calibration of the hyperspectral gamma-ray data to predict soil properties. They showed that "kriging improved the signal-to-noise ratio of the gamma-ray spectra." We focus here on spatial prediction of the total radioactivity count, the integrated count over the 0.4–2.81 mega-electronvolt spectrum, given in units of counts per second. The total count has been shown to be closely associated with the clay content in the soil (Taylor *et al.*, 2002; Pracilio *et al.*, 2004). Previously, Cressie and Kang (2010) carried out an exploratory data analysis of total count and obtained spatial predictions using a spatial-random-effects model.

To assess prediction performance, we created a test region (called MBD) containing 409 observations. The test data were only used for model evaluation, and they were not available for model fitting. The remaining $n = 33,866$ measurements, together with the test region (MBD), are shown in the top left panel of Figure 2. The spatial domain was taken to be $\mathcal{D} := (223525, 225770) \times (6526400, 6527930)$ in Easting and Northing.

Table 3. Results of Simulation Study 2

	Random knots		8 Fixed knots		14 Fixed knots	
Parent covariance	NPC	SPC	NPC	SPC	NPC	SPC
Time (s)	184.12	192.42	111.22	110.38	152.10	150.19
MSPE (ALL)	1.80	1.89	2.02	2.14	1.95	2.07
MSPE (OBS)	0.23	0.28	0.26	0.35	0.24	0.34
MSPE (MAR)	1.66	1.84	1.69	1.73	1.61	1.70
MSPE (MBD)	6.30	6.38	7.28	7.61	7.12	7.31
IS (ALL)	4.76	5.83	4.97	7.28	4.82	7.46
IS (OBS)	2.28	2.69	2.40	2.93	2.30	2.89
IS (MAR)	5.65	7.74	5.45	8.44	5.08	8.43
IS (MBD)	10.37	11.87	11.39	17.68	11.38	18.67
Posterior mean of r	8.82	9.78	(8)	(8)	(14)	(14)

NPC: nonstationary parent covariance; SPC: stationary parent covariance; MSPE: mean square prediction error; IS: interval score; Time: average time for each MCMC (averaged over the 100 simulated datasets); ALL: all 512 locations; OBS: the 275 observed locations; MAR: the 137 missing-at-random locations; MBD: the 100 missing-by-design locations.

Table 4. Results of Simulation Study 3

	Random knots		8 Fixed knots		14 Fixed knots	
Parent covariance	NPC	SPC	NPC	SPC	NPC	SPC
Time (s)	193.68	196.75	108.87	109.51	149.12	148.85
MSPE (ALL)	1.11	1.12	1.56	1.58	1.24	1.25
MSPE (OBS)	0.20	0.20	0.24	0.24	0.22	0.22
MSPE (MAR)	0.40	0.40	0.55	0.57	0.47	0.48
MSPE (MBD)	4.58	4.62	6.57	6.67	5.08	5.10
IS (ALL)	4.12	4.20	5.07	5.16	4.68	4.71
IS (OBS)	2.13	2.14	2.30	2.31	2.25	2.26
IS (MAR)	3.12	3.14	3.62	3.67	3.38	3.38
IS (MBD)	10.93	11.33	14.69	15.04	13.13	13.27
Posterior mean of r	10.40	10.69	(8)	(8)	(14)	(14)

NPC: nonstationary parent covariance; SPC: stationary parent covariance; MSPE: mean square prediction error; IS: interval score; Time: average time for each MCMC (averaged over the 100 simulated datasets); ALL: all 512 locations; OBS: the 275 observed locations; MAR: the 137 missing-at-random locations; MBD: the 100 missing-by-design locations.

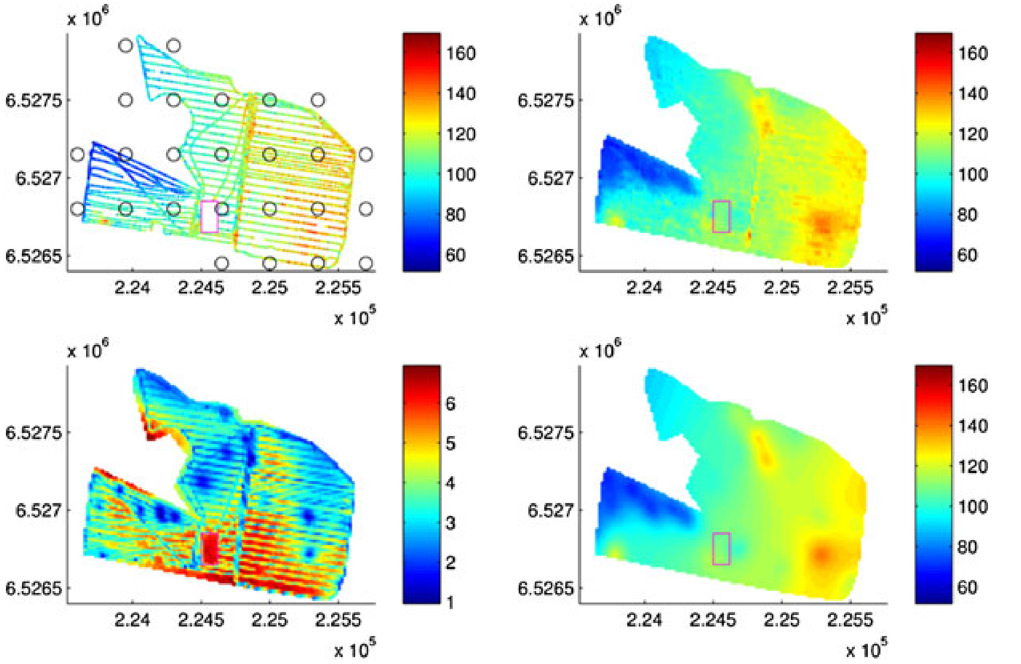

Figure 2. (Top left) Gamma emissions total count observations (small colored dots) and locations of the 25 basis-function centers for $\mathbf{b}_\theta(\cdot)$ (black circles). (Top right) Posterior mean of the true intensity using our model. (Bottom left) Posterior standard deviation of the true intensity. (Bottom right) Posterior mean of the smooth process without $\delta(\cdot)$ (see text). The test region (MBD) is represented by a pink rectangle. Color-scale units are counts per second; Easting and Northing are given in meters

Following Cressie and Kang (2010), we log-transformed the (shifted) data to obtain additive measurement error:

$$Z(\mathbf{s}_i) := \log(\mathrm{TC}(\mathbf{s}_i) + 160),\ i = 1, \ldots, n, \tag{25}$$

where TC denotes the total radioactivity count.

Cressie and Kang (2010) identified Easting and Northing as important trend terms, and so we set $\mathbf{x}(\mathbf{s}) := (1, \mathbf{s}')'$, where each location $\mathbf{s} \in \mathcal{D}$ is a two-dimensional vector consisting of Easting and Northing (in meters). The measurement-error variance (on the log scale) is known from another experiment to have a value of $\sigma_\epsilon^2 = 0.0016$ (Cressie and Kang, 2010). As the empirical variance of $\mathbf{Z}$ was calculated to

Table 5. Summary of the results of the soil data analysis

	Random knots		64 Fixed knots		144 Fixed knots	
Parent covariance	NPC	SPC	NPC	SPC	NPC	SPC
Time (h)	89.87	95.05	59.02	59.15	158.84	152.36
ASD (MBD) ×100	0.26	0.28	0.27	0.28	0.32	0.30
IS (MBD) ×100	26.96	28.13	27.39	28.48	29.41	31.38
Posterior mean of r	35.57	42.88	(64)	(64)	(144)	(144)

NPC: nonstationary parent covariance; SPC: stationary parent covariance; ASD: average squared distance; IS: interval score; Time: total time for the MCMC; MBD: missing-by-design (test region.)

be $\hat{\sigma}_Z^2 = 0.0026$ (after subtracting the trend as estimated by ordinary least squares), the signal-to-noise ratio is less than 2. Nonetheless, it is possible to distinguish signal from noise in many areas of the domain due to high sampling density (see top left panel of Figure 2).

We considered two equidistant grids of fixed knots on the domain $\mathcal{D}$, one with 64 and one with 144 locations. For the vector $\mathbf{b}_\theta(\cdot)$ in (15), we chose 25 power exponential functions with scale parameter 300 and centers shown in the top left panel of Figure 2. We chose a tapering length of $L = 35$ in (9), which resulted in roughly 150 non-zero elements per row for $\mathbf{V}$ in (18) (i.e., about 99.56% of the elements of $\mathbf{V}$ were zero). The prior distributions of the parent-covariance parameters were as described in Section 2.4, with $\mu_\sigma = \log(\hat{\sigma}_Y)$ and $\mu_\gamma = \log(577.76)$, where $\hat{\sigma}_Y := \sqrt{\hat{\sigma}_Z^2 - \sigma_\epsilon^2}$.

On an Intel Xeon X5560 machine with 94.5 GB RAM, we ran an MCMC for each of the models for 20,000 iterations, of which 10,000 were considered burn-in (based on examination of trace plots), and we only used every 10th of the remaining iterations for inference. We also obtained the posterior distribution of $Y(\cdot)$ at a grid of 5707 locations. In Figure 2, using our model, we show the posterior means (top right panel) and standard deviations (bottom left panel) of the (error-free) true intensity (TI) on the original scale, defined in analogy to the transformation (25) as $\mathrm{TI}(\mathbf{s}) := \exp\{Y(\mathbf{s})\} - 160$. We also show the posterior mean of $\exp\{\mu(\cdot) + \nu(\cdot)\} - 160$ (i.e., without $\delta(\cdot)$) in the bottom right panel of Figure 2.

The model comparison was carried out on the log-scale. We obtained samples from the posterior distribution of $Z(\mathbf{s}_j)$ at test location $\mathbf{s}_j$ as, $Z^{(k)}(\mathbf{s}_j) := Y^{(k)}(\mathbf{s}_j) + \epsilon^{(k)}(\mathbf{s}_j)$, where the $Y^{(k)}(\mathbf{s}_j)$ are posterior samples from $Y(\mathbf{s}_j)$, and $\epsilon^{(k)}(\mathbf{s}_j) \sim N(0, \sigma_\epsilon^2)$ is independent "measurement error." We then calculated the average squared distance (ASD) of the means of $\{Z^{(k)}(\mathbf{s}_j)\}$ to the test observations $Z(\mathbf{s}_j)$, and the interval score for 95% credible intervals for $Z(\mathbf{s}_j)$, for all models, averaged over the test locations.

The results are shown in Table 5. Random knots resulted in lower average squared distance and interval score than fixed knots. With the exception of average squared distance for the models with 144 fixed knots, NPC also improved over SPC. More knots resulted in less accurate predictive distributions.

On the basis of the examination of trace plots, mixing was somewhat slower for random knots than for fixed knots and slower for NPC than for SPC (both for the soil data here and for the simulated data in Section 4.1). Because the same number of MCMC iterations was used for all models, the computation times given in Tables 2–5 can be slightly misleading. However, as the focus in this article is not parameter estimation but on prediction, and predictive performance of the models was also assessed on the basis of an equal number of MCMC iterations, we feel that the comparison is fair.

5. CONCLUSIONS

In this article, our starting point was the Sang and Huang (2012) approach to analyzing large spatial datasets, which combines a low-rank predictive-process component with a tapered remainder component. To achieve enough flexibility for the nonstationary processes often encountered in real-world applications, we extended this model in two ways: first, the components in the model are parameterized on the basis of a nonstationary Matérn parent covariance function, in which the parameters vary spatially according to linear combinations of spatial basis functions. Second, for the low-rank component, which can be written as a linear combination of spatial basis functions, we make inference on the number, locations, and shapes of the basis functions. Posterior inference via reversible jump MCMC and related issues are described in detail.

The results of a simulation study (Section 4.1) and an analysis of a very large soil dataset (Section 4.2) indicate that the two extensions described previously can result in improved predictive distributions, especially in terms of quantifying prediction uncertainty. We show that for (typically nonstationary) real-world processes, it should often *not* be the goal to approximate a simple covariance model (e.g., the stationary Matérn covariance) as closely as possible. Results indicate that our model is sufficiently flexible to overcome a misspecified parent covariance, and its flexibility does not seem to result in a penalty in the unlikely event that the truth is, in fact, a simple stationary covariance (Simulation Study 3 in Section 4.1). Because of its adaptability, our model can be used to model highly nonstationary processes with varying levels of smoothness.

Acknowledgements

This research was supported by NASA under grant NNH08ZDA001N issued through the Advanced Information Systems Technology ROSES 2008 Solicitation, and by the Mathematics Center Heidelberg. I would like to thank Huiyan Sang, Emily Kang, the editor, associate editor, two anonymous referees, and especially Noel Cressie for helpful advice and comments. I am also grateful to James Taylor and Alex McBratney of the University of Sydney for making the Nowley soil dataset available. The collection of the data was directed by Professor McBratney and funded by the University of Sydney.

REFERENCES

Banerjee S, Carlin B, Gelfand AE. 2004. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall.

Banerjee S, Gelfand AE, Finley AO, Sang H. 2008. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B* **70**(4): 825–848.

Berliner LM, Wikle CK, Cressie N. 2000. Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *Journal of Climate* **13**(22): 3953–3968.

Bevilacqua M, Gaetan C, Mateu J, Porcu E. 2012. Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *Journal of the American Statistical Association* **107**(497): 268–280.

Calder CA. 2007. Dynamic factor process convolution models for multivariate space-time data with application to air quality assessment. *Environmental and Ecological Statistics* **14**(3): 229–247.

Cressie N. 1993. *Statistics for Spatial Data, Revised Edition*. John Wiley & Sons: New York, NY.

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

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

Cressie N, Kang EL. 2010. High-resolution digital soil mapping: kriging for very large datasets. In *Proximal Soil Sensing,* chapter 4, Viscarra-Rossel R, McBratney A, Minasny B (eds). Springer: Dordrecht, NL; 49–63.

Cressie N, Shi T, Kang EL. 2010. Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics* **19**(3): 724–745.

Curriero F, Lele S. 1999. A composite likelihood approach to semivariogram estimation. *Journal of Agricultural, Biological, and Environmental Statistics* **4**(1): 9–28.

Eidsvik J, Shaby BA, Reich BJ, Wheeler M, Niemi J. 2012. Estimation and prediction in spatial models with block composite likelihoods using parallel computing. *Journal of Computational and Graphical Statistics*, DOI: 10.1080/10618600.2012.76460, (to appear in print).

Finley AO, Sang H, Banerjee S, Gelfand AE. 2009. Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis* **53**(8): 2873–2884.

Furrer R, Genton MG, Nychka DW. 2006. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* **15**(3): 502–523.

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

Gilbert JR, Moler C, Schreiber R. 1992. Sparse matrices in MATLAB: design and implementation. *SIAM Journal on Matrix Analysis and Applications* **13**(1): 333–356.

Gneiting T. 2002. Compactly supported correlation functions. *Journal of Multivariate Analysis* **83**(2): 493–508.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**(477): 359–378.

Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation Bayesian model determination. *Biometrika* **82**(4): 711.

Guhaniyogi R, Finley AO, Banerjee S, Gelfand AE. 2011. Adaptive Gaussian predictive process models for large spatial datasets. *Environmetrics* **22**(8): 997–1007.

Haario H, Saksman E, Tamminen J. 2001. An adaptive Metropolis algorithm. *Bernoulli* **7**(2): 223–242.

Hastings W. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1): 97–109.

Henderson H, Searle S. 1981. On deriving the inverse of a sum of matrices. *SIAM Review* **23**(1): 53–60.

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

Holmes C, Mallick B. 2001. Bayesian regression with multivariate linear splines. *Journal of the Royal Statistical Society: Series B* **63**(1): 3–17.

Holmes C, Mallick BK. 2000. Bayesian wavelet networks for nonparametric regression. *IEEE Transactions on Neural Networks* **11**(1): 27–35.

Kang EL, Cressie N. 2011. Bayesian inference for the spatial random effects model. *Journal of the American Statistical Association* **106**(495): 972–983.

Kang EL, Liu D, Cressie N. 2009. Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models. *Computational Statistics & Data Analysis* **53**(8): 3016–3032.

Kanter M. 1997. Unimodal spectral windows. *Statistics & Probability Letters* **34**(4): 403–411.

Kass R, Raftery A. 1995. Bayes factors. *Journal of the American Statistical Association* **90**(430): 773–795.

Katzfuss M. 2011. Hierarchical spatial and spatio–temporal modeling of massive datasets, with application to global mapping of CO_2. *PhD Dissertation*, The Ohio State University, Columbus, OH.

Katzfuss M, Cressie N. 2011. Spatio–temporal smoothing and EM estimation for massive remote-sensing data sets. *Journal of Time Series Analysis* **32**(4): 430–446.

Katzfuss M, Cressie N. 2012. Bayesian hierarchical spatio–temporal smoothing for very large datasets. *Environmetrics* **23**(1): 94–107.

Kaufman C, Schervish M, Nychka DW. 2008. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association* **103**(484): 1545–1555.

Knuth K. 2005. Informed source separation: A Bayesian tutorial. In *European signal processing conference*, Sanjur B, Cetin E, Tekalp E, Kuruoglu E (eds). Antalya: Turkey.

Lemos RT, Sansó B. 2009. A spatio–temporal model for mean, anomaly, and trend fields of North Atlantic sea surface temperature. *Journal of the American Statistical Association* **104**(485): 5–18.

Lindgren F, Rue H, Lindström J. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B* **73**(4): 423–498.

Lindsay B. 1988. Composite likelihood methods. In *Statistical inference from stochastic processes*, Prabhu NU (ed.). RI. American Mathematical Society: Providence; 221–239.

Lopes HF, Salazar E, Gamerman D. 2008. Spatial dynamic factor analysis. *Bayesian Analysis* **3**(4): 759–792.

Mardia K, Goodall C, Redfern E, Alonso F. 1998. The kriged Kalman filter. *Test* **7**(2): 217–282.

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**(6): 1087–1092.

Paciorek C, Schervish M. 2006. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* **17**(5): 483–506.

Pracilio G, Smettem K, Harper R. 2004. New soil survey technologies to map landscape properties relevant to perennial plant performance. In *Salinity Solutions, Working with Science and Society*, Ridley A, Feikama P, Bennet S, Rogers M-J, Wilkinson R, Hirth J (eds). Bendigo, Victoria: Austalia. Proceedings of the Salinity Solutions Conference.

Sang H, Huang JZ. 2012. A full scale approximation of covariance functions. *Journal of the Royal Statistical Society, Series B* **74**(1): 111–132.

Sang H, Jun M, Huang JZ. 2011. Covariance approximation for large multivariate spatial datasets with an application to multiple climate model errors. *Annals of Applied Statistics* **5**(4): 2519–2548.

Shaby B, Ruppert D. 2012. Tapered covariance: Bayesian estimation and asymptotics. *Journal of Computational and Graphical Statistics* **21**(2): 433–452.

Sherman J, Morrison W. 1950. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics* **21**(1): 124–127.

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

Stein ML. 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer: New York, NY.

Stein ML. 2005. Nonstationary spatial covariance functions. *Technical Report Technical Report No. 21*, Center for Integrating Statistical and Environmental Science, The University of Chicago, Chicago, IL.

Stein ML. 2008. A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society* **37**(1): 3–10.

Taylor M, Smettem K, Pracilio G, Verboom W. 2002. Relationships between soil properties and high-resolution radiometrics, central eastern Wheatbelt, Western Australia. *Exploration Geophysics* **33**(2): 95–102.

van Dyk DA, Park T. 2008. Partially collapsed Gibbs samplers: theory and methods. *Journal of the American Statistical Association* **103**: 790–796.

Viscarra Rossel R, Taylor HJ, McBratney A. 2007. Multivariate calibration of hyperspectral γ-ray energy spectra for proximal soil sensing. *European Journal of Soil Science* **58**(1): 343–353.

Wikle CK. 2010. Low-rank representations for spatial processes. In *Handbook of Spatial Statistics*, Gelfand AE, Fuentes M, Guttorp P, Diggle P (eds). Chapman and Hall/CRC: Boca Raton, FL; 107–118.

Wikle CK, Cressie N. 1999. A dimension-reduced approach to space-time Kalman filtering. *Biometrika* **86**(4): 815–829.

Wikle CK, Milliff R, Nychka DW, Berliner LM. 2001. Spatiotemporal hierarchical Bayesian modeling: tropical ocean surface winds. *Journal of the American Statistical Association* **96**(454): 382–397.

Woodbury M. 1950. Inverting modified matrices. *Technical Report Memorandum Report 42*, Statistical Research Group, Princeton University, Princeton, NJ.

Xu B, Wikle CK, Fox N. 2005. A kernel-based spatio–temporal dynamical model for nowcasting radar precipitation. *Journal of the American Statistical Association* **100**(472): 1133–1144.