Published in final edited form as:

Annu Rev Stat Appl. 2017 March ; 4: 245–266. doi:10.1146/annurev-statistics-060116-054155.

Bayesian Modeling and Analysis of Geostatistical Data

Alan E. Gelfand¹ and Sudipto Banerjee²

¹Department of Statistical Science, Duke University, Durham, North Carolina 27708-0251

²Department of Biostatistics, University of California, Los Angeles, California 90095-1772

Abstract

The most prevalent spatial data setting is, arguably, that of so-called geostatistical data, data that arise as random variables observed at fixed spatial locations. Collection of such data in space and in time has grown enormously in the past two decades. With it has grown a substantial array of methods to analyze such data. Here, we attempt a review of a fully model-based perspective for such data analysis, the approach of hierarchical modeling fitted within a Bayesian framework. The benefit, as with hierarchical Bayesian modeling in general, is full and exact inference, with proper assessment of uncertainty. Geostatistical modeling includes univariate and multivariate data collection at sites, continuous and categorical data at sites, static and dynamic data at sites, and datasets over very large numbers of sites and long periods of time. Within the hierarchical modeling framework, we offer a review of the current state of the art in these settings.

Keywords

big spatial data; data assimilation; data fusion; Gaussian processes; integrated nested Laplace approximation; Markov chain Monte Carlo; multivariate spatial processes; spatiotemporal processes

1. INTRODUCTION

Spatial analysis has grown at a remarkable rate over the past two decades, largely fueled by the increased availability of inexpensive, high speed computing (as has been the case for many other areas). Such availability has enabled the collection of large spatial and spatiotemporal datasets across many fields, leading to widespread usage of sophisticated geographic information systems software, and the ability to investigate (fit and infer under) challenging, evermore appropriate and realistic models. As a result, spatial statistics has changed from a somewhat ad hoc field to one that is more model-driven.

Spatial data is usually viewed as one of three types: point-referenced data, areal unit data, and random spatial point pattern data. Here, we consider only the point-referenced case, that is, geostatistical data models. We thin an enormous literature to focus on hierarchical modeling, which we believe is the current and near future frontier for analyzing such data. In

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

general, geostatistical modeling can be viewed as a hierarchical specification, naturally leading to adoption of a Bayesian framework for inference and suitable Gibbs sampling/Markov chain Monte Carlo (MCMC) for model fitting. This is not surprising because, more generally, hierarchical modeling has emerged as the modeling paradigm for scientific work in the 21st century (Clark & Gelfand 2006) and for spatial statistics in particular.

This review will explore a wide range of geostatistical modeling problems. We begin by discussing the customary spatial linear geostatistical model including its fitting using MCMC along with fully Bayesian spatial prediction (kriging). Then, we turn to the spatial generalized linear model with the same objectives. Because multivariate measurements are often collected at locations, we consider multivariate geostatistical models. This requires specification of dependence both within and between locations, leading to cross-covariance functions. Model fitting can then proceed using MCMC. Kriging in the multivariate setting raises the issue of joint and conditional prediction. Also important in this regard is model-based imputation to handle missing data in the presence of spatial structure.

We then turn to spatial data fusion issues. Often, we have multiple data sources informing about say, environmental exposures. These data sources will usually be spatially misaligned owing to collection at different networks or at different spatial scales. How do we assimilate such datasets for improved spatial interpolation? How do we specify spatial models incorporating different spatial scales?

Next, we introduce time into the story, leading to spatiotemporal modeling (e.g., a monitoring station network collects data over time). Modeling challenges involving both continuous and discrete time arise, as well as the need to capture space-time dependence. On this framework, we can overlay multivariate data collected at locations over time.

We devote a substantial section to handling big data and finish with a brief subsection on two other challenges: efficient and user-friendly computation and general multivariate response with multiple covariates over space and time.

Hence, the format of the review is as follows. In Section 2 we discuss the univariate geostatistical model, and Section 3 extends it to the multivariate setting. Section 4 takes up data fusion, with Section 5 considering space-time modeling. Section 6 focuses on big data challenges, and Section 7 ends the article with some further challenges.

2. UNIVARIATE POINT-LEVEL MODELING

The basic geostatistical model specifies a response, $Y(\mathbf{s})$ at location $\mathbf{s}$ as

$$Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) + \varepsilon(\mathbf{s}). \quad (1)$$

Here, $\mathbf{x}(\mathbf{s})$ is a $p \times 1$ vector of spatially referenced covariates/predictors, with $\boldsymbol{\beta}$ a $p \times 1$ vector of regression coefficients. For example, if $Y(\mathbf{s})$ is the temperature at location $\mathbf{s}$, then $\mathbf{x}(\mathbf{s})$ might include elevation. The residual is partitioned into two pieces: One is spatial

($w(\mathbf{s})$), and one is nonspatial ($\varepsilon(\mathbf{s})$). Here, $w(\mathbf{s})$ is the geostatistical story, usually specified as a stationary Gaussian process (GP) with variance σ^2 ; $\varepsilon(\mathbf{s})$ is usually white noise, adding variance τ^2 , the so-called nugget effect (see, e.g., Cressie & Wikle 2011, Banerjee et al. 2014). τ^2 and σ^2 are variance components, and it is noteworthy that $\varepsilon(\mathbf{s})$ causes the $Y(\mathbf{s})$ surface to be everywhere discontinuous.

2.1. As a Hierarchical Model

Suppose we have data $Y(\mathbf{s}_i)$, $i = 1, \dots, n$, and let $\mathbf{Y} = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))^T$. We assume a GP for $w(\mathbf{s})$ with, say, a stationary covariance function, $\text{cov}(w(\mathbf{s}), w(\mathbf{s}')) = \sigma^2 \rho(\mathbf{s} - \mathbf{s}'; \phi)$, that is, σ^2 is the spatial variance above and ρ is a correlation function with parameter ϕ . With white noise for $\varepsilon(\mathbf{s})$, that is, independent and identically distributed (i.i.d.) $\varepsilon(\mathbf{s}) \sim N(0, \tau^2)$, we can marginalize over the $\{w(\mathbf{s}_i)\}$ to obtain the marginal covariance matrix, $\Sigma = \sigma^2 H(\phi) + \tau^2 I$ with $H(\phi)_{ij} = \text{corr}\{w(\mathbf{s}_i), w(\mathbf{s}_j)\}$, that is, the (i, j) th element of $H(\phi)$ is the correlation between $w(\mathbf{s}_i)$ and $w(\mathbf{s}_j)$. Σ is diagonally dominant and, therefore, well behaved with regard to inversion and determinant calculation needed for likelihood evaluation. That is, the resulting likelihood becomes:

$$\mathbf{Y}|\boldsymbol{\theta} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 H(\phi) + \tau^2 I), \quad (2)$$

where $\mathbf{X}$ is $n \times p$, collecting the $\mathbf{x}(\mathbf{s}_i)$ as row vectors, and $\boldsymbol{\theta}$ is the set of all of the parameters in the model.

In the Bayesian setting, independent priors are usually chosen for the different parameters, yielding $p(\boldsymbol{\theta}) = p(\boldsymbol{\beta}) p(\sigma^2) p(\tau^2) p(\phi)$. Here, illustratively, we are writing $\rho(\cdot)$ to be isotropic, that is, a function of $\|\mathbf{s} - \mathbf{s}'\|$ with decay parameter ϕ . Customary choices for priors are multivariate normal (or even flat) for $\boldsymbol{\beta}$ and inverse gamma for σ^2 and τ^2 . Specification of $p(\phi)$ depends upon choice of ρ ; usually Uniform or Gamma priors are selected. σ^2 , τ^2 and ϕ cannot have improper priors or else an improper posterior results (Berger et al. 2001).

It is useful to recast the above specification in a hierarchical fashion by considering a likelihood conditional on the spatial random effects $\mathbf{w} = (w(\mathbf{s}_1), \dots, w(\mathbf{s}_n))$. We have

- First stage: $\mathbf{Y}|\boldsymbol{\theta}, \mathbf{w} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{w}, \tau^2 I)$. That is, the $Y(\mathbf{s}_i)$ are conditionally independent given the $w(\mathbf{s}_i)$.
- Second stage: $\mathbf{w}|\sigma^2, \phi \sim N(\mathbf{0}, \sigma^2 H(\phi))$. Here, the GP serves as a second-stage process model.
- Third stage: priors on $(\boldsymbol{\beta}, \tau^2, \sigma^2, \phi)$. The third stage is often referred to as hyperprior specification.

The hierarchical model and the marginal model are inferentially equivalent. Fitting the marginal model is generally preferable (see, e.g., Banerjee et al. 2014). This hierarchical model form emerges as a special case of general multilevel process model which appears throughout this article, with richer scientific implications,

$$[\text{data}|\text{process, parameters}][\text{process}|\text{parameters}][\text{parameters}],$$

(see, e.g., Berliner 2000).

The number of scientific examples of such multilevel modeling has become substantial in the past twenty years. The book of Banerjee et al. (2014) presents a range of illustrations with associated references. These illustrations include modeling house prices using characteristics of the house and the neighborhood; modeling species presence/absence or abundance using environmental regressors; and joint modeling of exposure to environmental contaminants, explained using meteorological variables. In all cases, the strong presence of spatial dependence is demonstrated; the spatial variation σ^2 dominates the pure error contribution, τ^2 .

2.2. Classical and Gaussian Process Kriging

The problem of kriging (after D.G. Krige; see Krige 1951) is optimal spatial prediction: Given observations of a random field $\mathbf{Y} = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))'$, how do we best predict Y at a site $\mathbf{s}_0$ where it has not been observed? Ordinary kriging obtains the best linear predictor for $Y(\mathbf{s}_0)$ based on $\mathbf{Y}$ of the form $\sum_i \ell_i Y(\mathbf{s}_i) + \delta_0$ using squared error loss. Provided $E(Y(\mathbf{s} + \mathbf{h}) - Y(\mathbf{s})) = 0$ and the variogram $\gamma(\mathbf{h}) \equiv E(Y(\mathbf{s} + \mathbf{h}) - Y(\mathbf{s}))^2$ exists, this estimator can be obtained explicitly as a function of $\gamma(\mathbf{h})$. No other assumptions are required, but $\gamma(\mathbf{h})$ must be estimated.

In the fully model-based context presented in Section 2.1, we seek the function $h(\mathbf{Y})$ that minimizes the mean-squared prediction error, $E[(Y(\mathbf{s}_0) - h(\mathbf{Y}))^2 | \mathbf{Y}]$, which is $h(\mathbf{Y}) = E[Y(\mathbf{s}_0) | \mathbf{Y}]$. The posterior mean and variance for $Y(\mathbf{s}_0)$ have explicit forms:

$$E[Y(\mathbf{s}_0) | \mathbf{Y}] = \mathbf{x}(\mathbf{s}_0)^T \boldsymbol{\beta} + \boldsymbol{\gamma}^T \sum^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \quad (3)$$

$$\text{and var}[Y(\mathbf{s}_0) | \mathbf{Y}] = \sigma^2 + \tau^2 - \boldsymbol{\gamma}^T \sum^{-1} \boldsymbol{\gamma}, \quad (4)$$

where, under isotropy, $\boldsymbol{\gamma}^T = (\sigma^2 \rho(d_{01}; \phi), \dots, \sigma^2 \rho(d_{0n}; \phi))$ and $d_{0j} = \|\mathbf{s}_0 - \mathbf{s}_j\|$. The posterior mean is a linear predictor, but once the model parameters are estimated, this is no longer the case.

2.3. Fully Bayesian Kriging

Under the hierarchical model in Section 2.1, fully Bayesian kriging obtains the predictive distribution of $Y(\mathbf{s}_0)$ given $\mathbf{x}(\mathbf{s}_0)$ along with $\mathbf{Y}$ and $\mathbf{X}$. That is,

$$p(Y(\mathbf{s}_0) | \mathbf{x}(\mathbf{s}_0), \mathbf{Y}, \mathbf{X}) = \int p(Y(\mathbf{s}_0) | \mathbf{x}(\mathbf{s}_0), \mathbf{Y}, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathbf{Y}, \mathbf{X}) d\boldsymbol{\theta}, \quad (5)$$

where $p(Y(s_0)|\mathbf{x}(s_0), \mathbf{Y}, \boldsymbol{\theta})$ is a normal distribution arising from the joint multivariate normal distribution of $Y(s_0)$ and the original data $\mathbf{Y}$.

Simulation may be readily used to obtain estimates of Equation 5. Suppose we draw a posterior sample $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots, \boldsymbol{\theta}^{(B)}$ from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y}, X)$. Then Equation 5 may be approximated as a Monte Carlo mixture of the form

$$\hat{p}(Y(s_0)|\mathbf{x}(s_0), \mathbf{Y}, X) = \frac{1}{B} \sum_{b=1}^B p(Y(s_0)|\mathbf{x}(s_0), \mathbf{Y}, \boldsymbol{\theta}^{(b)}). \quad (6)$$

In practice, we use composition sampling to draw $Y(s_0)$, one for one for each $\boldsymbol{\theta}^{(b)}$, that is, $Y(s_0)^{(b)} \sim p(Y(s_0)|\mathbf{x}(s_0), \mathbf{Y}, \boldsymbol{\theta}^{(b)})$. The collection $\{Y(s_0)^{(1)}, Y(s_0)^{(2)}, \dots, Y(s_0)^{(B)}\}$ is a sample from the posterior predictive density, enabling a histogram or kernel density estimate to obtain an approximate plot of the density, bypassing the mixture calculation of Equation 6.

2.4. An Illustrative Example

We illustrate using a forest inventory dataset from the US Department of Agriculture Forest Service, Bartlett Experimental Forest, Bartlett, NH. This dataset is a part of the `spBayes` package in R and holds 1991 and 2002 forest inventory data for 437 plots. We use log-transformed total tree biomass as the outcome and regress it on five predictors: slope, elevation, and tasseled cap (a weighted sum of remote sensed readings) readings for brightness, greenness, and wetness components from spring, summer, and fall 2002 Landsat images. We perform fully Bayesian kriging as described in Section 2.3 using the `spLM` function in the R package `spBayes` available from the Comprehensive R Archive Network (CRAN) (<https://cran.r-project.org/web/packages/spBayes/index.html>). We present a brief analysis here. Further details on the dataset and the code may be found in Section 6.3 of Banerjee et al. 2014. The full dataset is available in `spBayes`.

We fit the Bayesian hierarchical model described in Section 2.1 using a flat prior for the regression coefficients $\boldsymbol{\beta}$, an exponential correlation function $\rho(\mathbf{s} - \mathbf{s}'; \phi) = \exp(-\phi \|\mathbf{s} - \mathbf{s}'\|)$, centered inverse-Gamma priors $\text{IG}(2, 0.08)$ and $\text{IG}(2, 0.02)$ for σ^2 and τ^2 , respectively, and a uniform prior $\text{U}(3/1500, 3/50)$ on ϕ suggesting an approximate range of between 50 and 1500 meters based upon the dimensions of the study domain. Once posterior samples for model parameters were obtained using standard MCMC algorithms, Bayesian kriging was carried out using Equations 5 and 6. Figure 1 presents the interpolated posterior predictive surface of the outcome, log metric tons of biomass over the study domain.

2.5. Avoiding Inappropriate Asymptotics

Inference for the one-dimensional time domain differs critically from that for the two-dimensional spatial domain: we have full order in the former, but only partial order in two (or more) dimensions. The implications of this remark are consequential. Large sample analysis for time series usually envisions an increasing time domain, letting time go to ∞ . By contrast, large sample analysis for geostatistical data usually envisions a fixed region

with more and more points filling in this domain (so-called infill asymptotics; see Stein 1999a,b; Loh 2005; Zhang & Zimmerman 2005).

When applying increasing domain asymptotic results, it is assumed that, as we collect more data, we can learn about temporal association at increasing distances in time. When applying infill asymptotic results, with a maximum distance fixed by the domain, we cannot learn about association at increasing distance. We will be able to do an increasingly better job with regard to spatial prediction at a given location. However, inference about process parameters will be bounded; Fisher information does not go to ∞ , Cramér-Rao lower bounds and asymptotic variances do not go to 0. Such concerns provide encouragement for the exact inference in a Bayesian framework under the implemented model, avoiding any asymptotic theory for inference. Of course, the fact that information is bounded implies that the data never overwhelms the prior, as is customarily assumed. There is no free lunch, and prior sensitivity analysis may be needed.

2.6. Spatial Generalized Linear Models

Often datasets preclude first-stage Gaussian modeling. In fact, $Y(\mathbf{s})$ need not be continuous. For example, $Y(\mathbf{s})$ can be a binary or count variable for observed locations or species data providing presence/absence or abundance, respectively. In this setting, we replace the Gaussian first stage with an exponential family member (perhaps with a dispersion parameter). We still will have a second-stage process (usually Gaussian) model (see Diggle et al. 1998). Now we have

- First stage: $Y(\mathbf{s}_i)$ are conditionally independent given $\boldsymbol{\beta}$ and $w(\mathbf{s}_i)$, so $f(Y(\mathbf{s}_i) | \boldsymbol{\beta}, w(\mathbf{s}_i), \gamma) = h(Y(\mathbf{s}_i), \gamma) \exp(\gamma [Y(\mathbf{s}_i)\boldsymbol{\eta}(\mathbf{s}_i) - \psi(\boldsymbol{\eta}(\mathbf{s}_i))])$, where $g(E(Y(\mathbf{s}_i))) = \boldsymbol{\eta}(\mathbf{s}_i) = \mathbf{x}^T(\mathbf{s}_i)\boldsymbol{\beta} + w(\mathbf{s}_i)$ (with $g(\cdot)$ is a suitable link function), and γ is a dispersion parameter.
- Second stage: $w(\mathbf{s})$ is a GP: $\mathbf{w} \sim N(\mathbf{0}, \sigma^2 H(\phi))$.
- Third stage: Hyperpriors for the model parameters are supplied.

Now, we do not have a stochastic process for $Y(\mathbf{s})$; rather, we have only a valid finite dimensional joint distribution. This causes no difficulty with regard to inference because we only fit the model for a finite number of points and we only interpolate at a finite number of points. It is not sensible to add pure error terms, $\varepsilon(\mathbf{s})$, to the link-transformed mean specification. They become redundant with the first-stage conditional independence. Also, the spatial random effects at the second-stage modeling encourage the means of the spatial variables at proximate locations to be similar to each other. For locations $\mathbf{s}$ and $\mathbf{s}'$, marginal spatial dependence is induced between $Y(\mathbf{s})$ and $Y(\mathbf{s}')$.

3. MULTIVARIATE GEOSTATISTICAL MODELING

A customary example of multivariate point-referenced spatial data arises from environmental monitoring stations, which might yield measurements on, for example, ozone, nitrous oxide, carbon monoxide, $\text{PM}_{2.5}$, or perhaps even species of $\text{PM}_{2.5}$. With such data we anticipate dependence between measurements at a particular location as well as across locations. Let $\mathbf{Y}(\mathbf{s})$ denote a $p \times 1$ vector of random variables at $\mathbf{s}$. We are interested in $\{\mathbf{Y}(\mathbf{s})$:

$\mathbf{s} \in D$). We envision p spatial surfaces with dependence between the surfaces, so we need a p -dimensional stochastic process. GPs seem a natural choice, but now, for n locations, we will need an np -dimensional multivariate normal.

The crucial object is the cross-covariance, $C(\mathbf{s}, \mathbf{s}') = \text{cov}(\mathbf{Y}(\mathbf{s}), \mathbf{Y}(\mathbf{s}'))$, a $p \times p$ matrix that need not be symmetric, that is, $\text{cov}(Y_j(\mathbf{s}), Y_{j'}(\mathbf{s}'))$ need not equal $\text{cov}(Y_{j'}(\mathbf{s}), Y_j(\mathbf{s}'))$. Also, $C(\mathbf{s}, \mathbf{s}')$ is not positive definite except in a limiting sense; $C(\mathbf{s}, \mathbf{s})$ is the covariance matrix associated with $\mathbf{Y}(\mathbf{s})$. With a p -variate mean specification, we will have the required np dimensional normal distribution.

There is substantial theoretical literature regarding creation of valid specifications for $C(\mathbf{s}, \mathbf{s}')$ (see, e.g., Chiles & Delfiner 1999). A simple constructive (hence, immediately valid) choice is the linear model of coregionalization, originally suggested for dimension reduction by Wackernagel (1994, 2003) and subsequently proposed for multivariate spatial modeling by Gelfand et al. (2004). Let $\mathbf{v}(\mathbf{s}) = A\mathbf{w}(\mathbf{s})$, where $\mathbf{w}(\mathbf{s}) = (w_1(\mathbf{s}), w_2(\mathbf{s}), \dots, w_p(\mathbf{s}))$ and A is any $p \times p$ nonsingular matrix. Here, the $w_j(\mathbf{s})$ are, say, p independent stationary spatial processes with respective correlation functions $\rho_j(\mathbf{s} - \mathbf{s}')$, $j = 1, 2, \dots, p$. The cross covariance matrix (with $\mathbf{a}_j$ denoting the columns of A) is $C(\mathbf{s} - \mathbf{s}') = \sum_{j=1}^p \rho_j(\mathbf{s} - \mathbf{s}') \mathbf{a}_j \mathbf{a}_j^T$. The linear transformation induces dependence between the processes and also provides a distinct covariance function for each component. The separable specification is the special case where there is a common $\rho(\mathbf{s} - \mathbf{s}')$ for all j . Then, $C(\mathbf{s} - \mathbf{s}') = \rho(\mathbf{s} - \mathbf{s}') T$ where $T = AA^T$.

Hence, we arrive at a general multivariate geostatistical model of coregionalization, $\mathbf{Y}(\mathbf{s}) = \boldsymbol{\mu}(\mathbf{s}) + \mathbf{v}(\mathbf{s}) + \boldsymbol{\varepsilon}(\mathbf{s})$ with $\boldsymbol{\varepsilon}(\mathbf{s}) \sim N(0, D_{\boldsymbol{\varepsilon}})$, $(D_{\boldsymbol{\varepsilon}})_{jj} = \tau_j^2$ and $\mathbf{v}(\mathbf{s}) = A\mathbf{w}(\mathbf{s})$, following the above specification. The $w_j(\mathbf{s})$ are mean 0 GPs with individual correlation functions. Usually, $\boldsymbol{\mu}(\mathbf{s})$ arises as a regression, that is, in the form $\mu_j(\mathbf{s}) = \mathbf{x}_j^T(\mathbf{s})\boldsymbol{\beta}_j$. Again, we have a hierarchical model conditional on $\mathbf{v}(\mathbf{s})$.

An attractive example that demonstrates the power of hierarchical modeling is the spatially varying coefficients model (Gelfand et al. 2003). Suppose we model a univariate observation $Y(\mathbf{s})$ at $\mathbf{s}$ as $Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T \boldsymbol{\beta}(\mathbf{s}) + \varepsilon(\mathbf{s})$. Here, the multivariate spatial process is for $\boldsymbol{\beta}(\mathbf{s})$. With $p = 2$ and $\mathbf{x}(\mathbf{s})$ having a column of 1s, we obtain $Y(\mathbf{s}) = \beta_0(\mathbf{s}) + \mathbf{x}(\mathbf{s})\beta_1(\mathbf{s}) + \varepsilon(\mathbf{s})$. We have a spatially varying intercept (like a spatial random effect) and a spatially varying slope. We need a bivariate spatial process because we expect the slope and intercept to be dependent at any location. We obtain a very rich class of nonlinear (but locally linear) models.

Remarkably, we can infer about the multivariate GP for $\boldsymbol{\beta}(\mathbf{s})$ while only observing the univariate $Y(\mathbf{s})$ process. Of course, with only an observation vector at n locations, there may not be much learning about the component processes of $\boldsymbol{\beta}(\mathbf{s})$. In practice, we often find an informative intercept surface but essentially flat slope surfaces.

3.1. Kernel Convolution Methods

Another approach to create multivariate geostatistical models is through moving averages (Ver Hoef & Barry 1998), also referred to as kernel convolution (Higdon 1998). In fact, this

is a special case of a general approach to create stationary processes (Yaglom 1987, chapter 8). Suppose $k_I(\cdot)$, $I = 1, \dots, p$ is a set of p square integrable kernel functions on $\mathcal{R}^2$ with $k_I(0) = 1$. Let $w(\mathbf{s})$ be a mean 0, variance 1 GP with correlation function ρ . Define the p -variate spatial process $\mathbf{v}(\mathbf{s})$ component-wise by $v_I(\mathbf{s}) = \sigma_I \int k_I(\mathbf{s} - \mathbf{t}) w(\mathbf{t}) d\mathbf{t}$, $I = 1, \dots, p$, $\sigma_I > 0$. Then, $\mathbf{v}(\mathbf{s})$ is a mean 0 multivariate GP with cross-covariance function $C(\mathbf{s}, \mathbf{s}')$, with

$$(C(\mathbf{s}, \mathbf{s}'))_{II'} = \sigma_I \sigma_{I'} \iint k_I(\mathbf{s} - \mathbf{t}) k_{I'}(\mathbf{s}' - \mathbf{t}') \rho(\mathbf{t} - \mathbf{t}') d\mathbf{t} d\mathbf{t}'. \quad (7)$$

By construction, $C(\mathbf{s}, \mathbf{s}')$ is valid. By change of variable, $(C(\mathbf{s}, \mathbf{s}'))_{II'}$ depends only on $\mathbf{s} - \mathbf{s}'$, that is, $\mathbf{v}(\mathbf{s})$ is a stationary process. If the k_I depend upon $\mathbf{s} - \mathbf{s}'$ only through $\|\mathbf{s} - \mathbf{s}'\|$ and ρ is isotropic, then $C(\mathbf{s} - \mathbf{s}')$ is isotropic. To work with this process requires discrete approximation, employing a finite set of knot locations. Specifically, we choose a finite set of locations $\mathbf{s}_1^*, \dots, \mathbf{s}_r^*$ and define $v_I(\mathbf{s}) = \sigma_I \sum_{j=1}^r k_I(\mathbf{s} - \mathbf{s}_j^*) w(\mathbf{s}_j^*)$. Now, $(C(\mathbf{s}, \mathbf{s}'))_{II'}$ is a double sum and, as a discrete approximation, the resulting process is no longer stationary.

3.2. Co-Kriging

In classical multivariate geostatistics, co-kriging refers to spatial prediction at a new location of either $\mathbf{Y}(\mathbf{s}_0)$ or a component $Y_j(\mathbf{s}_0)$. That is, we can predict the entire vector, $\mathbf{Y}(\mathbf{s}_0)$ given all of the observed vectors; we can predict its components given the observed vectors as well as values for the remaining components (collocated kriging). A detailed review can be found in Wackernagel (2003).

Similarly to Section 2.2, if we assume a multivariate Gaussian spatial process, we can obtain an explicit solution for the individual co-kriging problem. For simplicity, assume $\mathbf{Y}(\mathbf{s})$ is centered to have mean $\mathbf{0}$. The cross-covariance function provides $\Sigma_{\mathbf{Y}}$, the $np \times np$ covariance matrix for the data $\mathbf{Y} = (\mathbf{Y}(\mathbf{s}_1)^T, \mathbf{Y}(\mathbf{s}_2)^T, \dots, \mathbf{Y}(\mathbf{s}_n)^T)^T$. In addition, to predict, say, $Y(\mathbf{s}_1)$, it provides the $np \times 1$ vector, $\mathbf{C}_0$, which is blocked as vectors $\mathbf{C}_{0j}$, $j = 1, 2, \dots, n$ with k th element $c_{0j,k} = \text{cov}(Y_1(\mathbf{s}_0), Y_k(\mathbf{s}_j))$. Then, from the multivariate normal distribution of $(\mathbf{Y}, Y_1(\mathbf{s}_0))$, we obtain the co-kriging estimate,

$$E(Y_1(\mathbf{s}_0) | \mathbf{Y}) = \mathbf{C}_0^T \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y}. \quad (8)$$

The associated variance, $\text{var}(Y_1(\mathbf{s}_0) | \mathbf{Y})$, is also immediately available. Similarly, forms for collocated kriging can be developed; we omit details. Under the multivariate geostatistical model above, we can implement fully Bayesian kriging, analogous to Section 2.3 (see Banerjee et al. 2014 for details).

3.3. Regression in the Geostatistical Case

The regression problem of interest here is to learn about the conditional distribution for $Y(\mathbf{s}_0) | X(\mathbf{s}_0)$, not kriging for $Y(\mathbf{s}_0)$. For instance, we might seek the local relationship

between tree diameter and soil moisture. Why do we consider a bivariate process model here rather than specifying a geostatistical model as in Equation 1? Two reasons are as follows. First, we might be interested in an inverse regression problem, to learn about $X(s_0)$ for a given $Y(s_0)$. To do so requires modeling $X(s)$ to be random. Hence, if $X(s)$ is continuous, we need to introduce a process model for the X s as well as for the Y s, a bivariate (more generally, a multivariate) process model. Second, we may have missingness—equivalently, misalignment—in the data. That is, we have some X s without associated Y s and vice versa. This is particularly the case with tree diameter/soil moisture data. We would like to use all of the data rather than just the complete cases. We need a bivariate model for the missing data in order to do model-based imputation (e.g., Little & Rubin 2002).

First, assume no misalignment, and let $\mathbf{X} = (X(s_1), \dots, X(s_n))^T$ and $\mathbf{Y} = (Y(s_1), \dots, Y(s_n))^T$ be the measurements on the covariates and the response, respectively. We envision (perhaps after a suitable transformation) a bivariate Gaussian spatial process, $\mathbf{W}(s) = (X(s), Y(s))^T$ with mean $\boldsymbol{\mu}(s) = (\mu_X(s), \mu_Y(s))^T$ and, for simplicity, a separable cross-covariance function as defined above. So, the joint distribution of $Y(s)$ and $X(s)$ at location s is

$$\mathbf{W}(s) = \begin{pmatrix} X(s) \\ Y(s) \end{pmatrix} \sim N(\boldsymbol{\mu}(s), T). \quad (9)$$

With misalignment, let $\mathbf{X}$ be the vector of observed $X(s)$ s and $\mathbf{Y}$ be the vector of observed $Y(s)$ s. Let $\mathbf{X}'$ denote the vector of missing X observations with respect to the Y s and $\mathbf{Y}'$ be the vector of missing Y observations with respect to the X s. Then, in the preceding discussion, we can replace $\mathbf{X}$ and $\mathbf{Y}$ by the augmented vectors $\mathbf{X}_{aug} = (\mathbf{X}, \mathbf{X}')$ and $\mathbf{Y}_{aug} = (\mathbf{Y}, \mathbf{Y}')$. After permutation to line up the X s and Y s, they can be collected into a vector $\mathbf{W}_{aug}$. In the Bayesian model specification, $\mathbf{X}'$ and $\mathbf{Y}'$ are viewed as latent (unobserved) vectors. In implementing a Gibbs sampler for model fitting, we update the model parameters given $\mathbf{X}'$ and $\mathbf{Y}'$ (i.e., given $\mathbf{W}_{aug}$), and then update $(\mathbf{X}', \mathbf{Y}')$ given $\mathbf{X}$, $\mathbf{Y}$, and the model parameters.

For the regression problem, the distribution of interest is $f(E[Y(s_0)|X(s_0)] | X(s_0), \mathbf{Y}, \mathbf{X})$. For simplicity, suppose $\boldsymbol{\mu}(s) = (\mu_1, \mu_2)^T$ constant over sites. Then, from Equation 9, for the pair $(X(s), Y(s))$, $p(Y(s)|X(s), \beta_0, \beta_1, \sigma^2)$ is $N(\beta_0 + \beta_1 X(s), \sigma^2)$. That is, $E[Y(s)|X(s)] = \beta_0 + \beta_1 X(s)$, where

$$\beta_0 = \mu_2 - \frac{T_{12}}{T_{11}} \mu_1, \quad \beta_1 = \frac{T_{12}}{T_{11}}, \quad \text{and} \quad \sigma^2 = T_{22} - \frac{T_{12}^2}{T_{11}}. \quad (10)$$

So, given samples from the joint posterior distribution of $(\mu_1, \mu_2, T, \boldsymbol{\phi})$, we directly have samples from the posterior distributions for the parameters in Equation 10, and thus from the posterior distribution of $E[Y(s)|X(s)]$. Evidently, we can reverse the roles of X and Y to address the inverse problem. Also, extension to the multivariate case is clear.

Extension to the case of binary response in a geostatistical dataset proceeds as follows. Now, at each site, $Z(\mathbf{s})$ equals 0 or 1 according to whether we observed “failure” or “success” at that site. Thus, a realization of the process is called a two-color or binary map (DeOliveira 2000). Adopting the latent variable approach for probit modeling (as in, e.g., DeOliveira 2000), let $Y(\mathbf{s})$ be a latent spatial process associated with the sites and let $X(\mathbf{s})$ be a scalar covariate (again, extension to the multivariate case is clear). Let $Z(\mathbf{s}) = 1$ if and only if $Y(\mathbf{s}) > 0$. We envision our bivariate process $\mathbf{W}(\mathbf{s}) = (X(\mathbf{s}), Y(\mathbf{s}))^T$ distributed as in Equation 9, but where now $\boldsymbol{\mu}(\mathbf{s}) = (\mu_1, \mu_2 + \boldsymbol{\alpha}^T \mathbf{U}(\mathbf{s}))^T$ with $\mathbf{U}(\mathbf{s})$ a $p \times 1$ vector of fixed covariates. The conditional mean of $Y(\mathbf{s})$ given $X(\mathbf{s})$ shows that, relative to the regression coefficients, the conditional variance of $Y(\mathbf{s})$ given $X(\mathbf{s})$ is not identifiable. Thus, without loss of generality, we set $T_{22} = 1$ so that the T matrix has only two parameters. Finally, the probit spatial regression model is

$$P(Z(\mathbf{s})=1|x(\mathbf{s}), \mathbf{U}(\mathbf{s}), \boldsymbol{\alpha}, \mu_1, \mu_2, T_{11}, T_{12}) \\ = \Phi \left([\beta_0 + \beta_1 X(\mathbf{s}) + \boldsymbol{\alpha}^T \mathbf{U}(\mathbf{s})] / \sqrt{1 - \frac{T_{12}^2}{T_{11}}} \right). \quad (11)$$

Here, as in Equation 10, $\beta_0 = \mu_2 - (T_{12}/T_{11})\mu_1$, and $\beta_1 = T_{12}/T_{11}$.

4. DATA FUSION

Data fusion or assimilation (we use the terms interchangeably here) for spatial and spatiotemporal data can arise in various forms. There is a substantial literature on data assimilation for climate modeling (Kalnay 2003) as well as a growing literature on data fusion for environmental exposure assessment (the focus here). All of this work is embedded, in a sense, in the bigger world of data assimilation with computer model output. The overall goal of data fusion is to synthesize multiple data sources that are informing about spatial and spatiotemporal responses over the same region and the same time period. A typical example in an environmental setting would be modeling a contaminant such as $\text{PM}_{2.5}$. We might have monitoring data, satellite data, and computer model results. The anticipated benefit will be improved kriging and, in the temporal case, improved short term forecasting. It will almost always be the case that the data layers are misaligned in space and in time, that is, point-referenced for one, areal units for another, or hourly availability for one, daily collection for another. For illustration, we confine ourselves to the case of two data layers; one will be monitoring station data, that is, the geostatistical source, and the other will be output (rather than real data) from an environmental computer model for grid cells at some spatial resolution.

4.1. Fusion Modeling Using Stochastic Integration

The fusion approach proposed by Fuentes & Raftery (2005) builds upon earlier Bayesian melding work by Poole & Raftery (2000). It conceptualizes a true exposure surface and views the monitoring station data as well as the model output data as varying in a suitable way around the true surface. In particular, the average exposure in a grid cell A , denoted by $Z(A)$,

arises from a realization of the process surface as a so-called change of support problem (Gelfand et al. 2001) through the stochastic integral

$$Z(A) = \frac{1}{|A|} \int_A Z(s) ds, \quad (12)$$

where $|A|$ denotes the area of grid cell A and $Z(A)$ is called the block average. As an integral over a realization of a stochastic process, this is a random variable and is never available explicitly; we are not integrating a function. Fusion modeling, working with Equation 12, has been considered by, for example, Fuentes and Raftery (2005), as follows.

Let $Z(s)$ and $Y(s)$ denote an observed and the associated true exposure, respectively, at station s . The first model assumption is

$$Z(s) = Y(s) + \varepsilon(s) \quad (13)$$

where $\varepsilon(s) \sim N(0, \sigma_\varepsilon)$ represents the measurement error at location s . The true exposure process is assumed to be

$$Y(s) = \mu(s) + \eta(s), \quad (14)$$

where $\mu(s)$ provides the spatial mean surface, typically characterized by a trend, by elevation, etc. The error term $\eta(s)$ is a zero-mean GP.

The computer model output at areal unit scale is usually assumed to be biased and/or uncalibrated. Hence, the conceptual point level output from this model, denoted by $Q(s)$, is assumed to be related to the true surface as

$$Q(s) = a(s) + b(s)Y(s) + \delta(s) \quad (15)$$

where $a(s)$ and $b(s)$ denote, respectively, the local additive and multiplicative bias (recall the spatially varying coefficient model from Section 3). The error term, $\delta(s)$, is assumed to be a white noise process given by $N(0, \sigma_\delta)$. Then, because the computer model output is provided on a grid, $A_1, \dots, A_J$, the point-level process is converted to grid level by stochastic integration of Equation 15, that is, $Q(A_j) = \int_{A_j} a(s) ds + \int_{A_j} b(s) Y(s) ds + \int_{A_j} \delta(s) ds$. It has been observed that unstable model fitting (because of identifiability problems) accrues to the case where we have spatially varying $b(s)$ so $b(s) = b$ is adopted. Spatial prediction at a new location s_0 is done through the posterior predictive distribution $f(Y(s_0)|\mathbf{Z}, \mathbf{Q})$, where $\mathbf{Z}$ denotes all the station data and $\mathbf{Q}$ denotes all the grid-level computer output $Q(A_1), \dots, Q(A_J)$.

This fusion strategy becomes computationally infeasible when we have a large number of grid cells; each stochastic integral requires approximation, and the overall number of spatial random variables introduced becomes enormous (see the discussion of the “big n ” problem in Section 6 below). A dynamic implementation over many time periods becomes even more infeasible. In fact, although we have a very large number of grid cells, we have a relatively sparse number of monitoring sites. The emergent suggestion is that we should be scaling down rather than integrating up.

4.2. The Downscaler

The downscaler model was proposed by Berrocal et al. (2010). For ease of interpretation, we present it with regard to an ozone application. So, we consider two sources of ground-level ozone data (daily 8-hour maxima in ppb). The first one is geostatistical; it comes from a national air monitoring network. The second source of data in our study is the Community Multiscale Air Quality (CMAQ), a numerical model that estimates daily 8-hour maximum concentration of ozone. CMAQ model output is viewed as a tiled surface at grid cell resolution, say at 12-km squares, providing a tile height for each grid cell. To suggest magnitudes, for the eastern United States, we have 861 monitoring sites with 40,044 grid cells. We have many more grid cells than monitoring sites; the enormous amount of block averaging that would be needed to implement the fusion model of Fuentes & Raftery (2005) makes it computationally overwhelming.

For the downscaler, we let $Z(\mathbf{s})$ be the square root (for comfort with a normality assumption) of the observed ozone concentration at $\mathbf{s}$. We use $X(B)$ to denote the square root of the numerical model output over grid cell B . Each point $\mathbf{s}$ is associated with the CMAQ grid cell B in which it lies. We relate the observed data to the CMAQ output in the following way. For each $\mathbf{s}$ in B , we assume that

$$Z(\mathbf{s}) = \tilde{\beta}_0(\mathbf{s}) + \tilde{\beta}_1(\mathbf{s}) X(B) + \varepsilon(\mathbf{s}), \quad \varepsilon(\mathbf{s}) \stackrel{\text{ind}}{\sim} N(0, \tau^2), \quad (16)$$

where $\tilde{\beta}_0(\mathbf{s}) = \beta_0 + \beta_0(\mathbf{s})$, $\tilde{\beta}_1(\mathbf{s}) = \beta_1 + \beta_1(\mathbf{s})$, and $\varepsilon(\mathbf{s})$ is a white-noise process with nugget variance τ^2 . β_0 and β_1 represent the overall intercept and slope bias of the CMAQ model, and $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ are local adjustments to the additive and multiplicative bias, respectively. The spatially varying coefficients $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ are in turn modeled as bivariate mean-zero Gaussian spatial processes using the method of coregionalization, discussed in Section 3. Altogether, $\tilde{\beta}_0(\mathbf{s}) + \tilde{\beta}_1(\mathbf{s})X(B)$ plays the role of the true exposure $Y(\mathbf{s})$ in the previous subsection.

This model specification is simple but provides calibration at the local level and endows the spatial process $Z(\mathbf{s})$ with a flexible nonstationary covariance structure. It is much easier to fit than Bayesian melding because we eliminate the need to handle stochastic integrals. Moreover, to fit the model we only need to work with the responses associated with the set of monitoring sites, a much smaller number compared with the number of grid cells.

5. SPACE-TIME GEOSTATISTICAL MODELING

Spatiotemporal modeling has received dramatically increased attention in the past few years owing to the proliferation of datasets that are both spatially and temporally indexed, and the attendant need to understand them. For example, in studies of air pollution, we are interested not only in the spatial nature of a pollutant surface, but also in how this surface changes over time. When time is discretized to customary integer-spaced intervals, we may look at spatiotemporally indexed geostatistical data $Y(\mathbf{s}, t)$ in two ways. Writing $Y(\mathbf{s}, t) = Y_{\mathbf{s}}(t)$, we have a spatially varying time-series model. Writing $Y(\mathbf{s}, t) = Y_t(\mathbf{s})$, we have a temporally varying spatial model. For example, with climate data we may be interested in spatial patterns of temperature or precipitation at a given time, but also in dynamic patterns in weather. With real estate markets, we might be interested in how the single-family home sales market changes on a quarterly or annual basis. In this setting, we do not observe the same locations for each time period; the data are cross-sectional, rather than longitudinal. In the case of time series, we could regard the data as a multivariate measurement vector at each location and then attempt to employ multivariate spatial data models as in the previous section. With short series, this might be reasonable; with longer series, we would likely want to introduce aspects of usual time series modeling.

Missing data issues become more complicated in the spatiotemporal setting. Now we may encounter time points for which we lack spatial information, locations for which information is lacking for certain (possibly future) time points, or combinations thereof. Here the Bayesian hierarchical approach is particularly useful, because it not only helps organize our thinking about the model, but also properly attaches uncertainty for predictions that are farther from the observed data (in either space or time).

5.1. Geostatistical Modeling with Continuous Time

Modeling for geostatistical data with continuous time naturally extends that of Section 2. Let $Y(\mathbf{s}, t)$ denote the measurement at location $\mathbf{s}$ at time t . For continuous data assumed to be roughly normally distributed, we can write the general form

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + e(\mathbf{s}, t), \quad (17)$$

where $\mu(\mathbf{s}, t)$ denotes the mean structure and $e(\mathbf{s}, t)$ denotes the residual. If $\mathbf{x}(\mathbf{s}, t)$ is a vector of covariates associated with $Y(\mathbf{s}, t)$ then we might set $\mu(\mathbf{s}, t) = \mathbf{x}(\mathbf{s}, t)^T \boldsymbol{\beta}(\mathbf{s}, t)$. Note that this form allows spatiotemporally varying coefficients (in the spirit of Section 3), likely more general than we would want. $\boldsymbol{\beta}(\mathbf{s}, t) = \boldsymbol{\beta}$ is frequently adopted. Setting $\boldsymbol{\beta}(\mathbf{s}, t) = \boldsymbol{\beta}(\mathbf{s})$ yields spatially varying coefficients (again, following Section 3); setting $\boldsymbol{\beta}(\mathbf{s}, t) = \boldsymbol{\beta}(t)$ does not exist in the literature except when t is discretized as in section 5.3 below. Finally, $e(\mathbf{s}, t)$ would typically be rewritten as $w(\mathbf{s}, t) + \varepsilon(\mathbf{s}, t)$, where $\varepsilon(\mathbf{s}, t)$ is a Gaussian white-noise process and $w(\mathbf{s}, t)$ is a mean-zero spatiotemporal process.

We can therefore view Equation 17 as a hierarchical model with a conditionally independent first stage given $\{\mu(\mathbf{s}, t)\}$ and $\{w(\mathbf{s}, t)\}$. Then, in the spirit of Section 2.5, we can replace the

Gaussian first stage with another first-stage model (say, an exponential family model) and write $Y(\mathbf{s}, t) \sim f(y(\mathbf{s}, t) | \mu(\mathbf{s}, t), w(\mathbf{s}, t))$, where

$$f(y(\mathbf{s}, t) | \mu(\mathbf{s}, t), w(\mathbf{s}, t)) = b(y(\mathbf{s}, t)) \exp\{\gamma [\eta(\mathbf{s}, t)y(\mathbf{s}, t) - \chi(\eta(\mathbf{s}, t))]\}, \quad (18)$$

where γ is a positive dispersion parameter. In Equation 18, $g(\eta(\mathbf{s}, t)) = \mu(\mathbf{s}, t) + w(\mathbf{s}, t)$ for some link function g .

So, the challenge is to specify the spatiotemporal process $w(\mathbf{s}, t)$. Simple versions would specify $w(\mathbf{s}, t) = w(\mathbf{s}) + \alpha(t)$ (additively) or $w(\mathbf{s}, t) = w(\mathbf{s})\alpha(t)$ (multiplicatively) where $w(\mathbf{s})$ and $\alpha(t)$ are independent mean-zero GPs in space and time, respectively. More common is to provide a spatiotemporal process over $R^2 \times R^+$. Confining ourselves to GPs, we only need to specify a valid spatiotemporal covariance function. Here, “valid” means that for any set of locations and any set of time points, the covariance matrix for the resulting set of random variables is positive definite. An important point here is that it is not sensible to consider $(\mathbf{s}, t)$ as a three-dimensional vector and propose a valid correlation function on $\mathcal{R}^3$. This is because, regardless of scaling, distance in space has nothing to do with distance on the timescale.

As a result, a stationary spatiotemporal covariance specification is assumed to take the form $\text{cov}(w(\mathbf{s}, t), w(\mathbf{s}', t')) = c(\mathbf{s} - \mathbf{s}', t - t')$. [An isotropic form sets $\text{cov}(w(\mathbf{s}, t), w(\mathbf{s}', t')) = c(\|\mathbf{s} - \mathbf{s}'\|, |t - t'|)$.] A frequently used choice is the separable form

$$\text{cov}(w(\mathbf{s}, t), w(\mathbf{s}', t')) = \sigma^2 \rho^{(1)}(\mathbf{s} - \mathbf{s}'; \phi) \rho^{(2)}(t - t'; \psi), \quad (19)$$

where $\rho^{(1)}$ is a valid two-dimensional correlation function and $\rho^{(2)}$ is a valid one-dimensional correlation function (see, e.g., Mardia & Goodall 1993 and references therein). Equation 19 shows that dependence attenuates in a multiplicative manner across space and time.

Why is Equation 19 valid? For locations $\mathbf{s}_1, \dots, \mathbf{s}_I$ and times $t_1, \dots, t_J$ collecting the variables into a vector $\mathbf{w}_s^T = (w^T(\mathbf{s}_1), \dots, w^T(\mathbf{s}_I))$ where $\mathbf{w}(\mathbf{s}_i) = (w(\mathbf{s}_i, t_1), \dots, w(\mathbf{s}_i, t_J))^T$ the covariance matrix of $\mathbf{w}_s$ is

$$\sum_{\mathbf{w}_s} (\sigma^2, \phi, \psi) = \sigma^2 H_s(\phi) \otimes H_t(\psi), \quad (20)$$

where $\otimes$ again denotes the Kronecker product. In Equation 20, $H_s(\phi)$ is $I \times I$ with $(H_s(\phi))_{ii'} = \rho^{(1)}(\mathbf{s}_i - \mathbf{s}_{i'}; \phi)$, and $H_t(\psi)$ is $J \times J$ with $(H_t(\psi))_{jj'} = \rho^{(2)}(t_j - t_{j'}; \psi)$. Equation 20 clarifies that $\Sigma_{\mathbf{w}_s}$ is positive definite. So, $\mathbf{w}_s$ will follow an IJ -dimensional multivariate

normal with, in obvious notation, mean vector $\boldsymbol{\mu}_s(\boldsymbol{\beta})$ and the covariance matrix arising from Equation 20.

With associated data $\mathbf{Y}_s^T = (\mathbf{Y}^T(s_1), \dots, \mathbf{Y}^T(s_I))$ and given a prior for $\boldsymbol{\beta}$, σ^2 , $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$, the Bayesian model is completely specified. Simulation-based model fitting can be carried out similarly to the static spatial case by noting that the log-likelihood arising from $\mathbf{Y}_s$ is

$$-\frac{1}{2} \log \left| \sigma^2 H_s(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}) \right| - \frac{1}{2\sigma^2} (\mathbf{Y}_s - \boldsymbol{\mu}_s(\boldsymbol{\beta}))^T (H_s(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}))^{-1} (\mathbf{Y}_s - \boldsymbol{\mu}_s(\boldsymbol{\beta})).$$

However, $|\sigma^2 H_s(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi})| = (\sigma^2)^{IJ} |H_s(\boldsymbol{\phi})|^J |H_t(\boldsymbol{\psi})|^I$ and

$(H_s(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}))^{-1} = H_s^{-1}(\boldsymbol{\phi}) \otimes H_t^{-1}(\boldsymbol{\psi})$ by properties of Kronecker products. In other words, even though Equation 20 is $IJ \times IJ$, we need only the determinant and inverse for an $I \times I$ and a $J \times J$ matrix, expediting likelihood evaluation and hence Gibbs sampling.

Prediction is straightforward; see Banerjee et al. (2014) for details.

5.2. Nonseparable Spatiotemporal Models

In many applications we expect space-time interaction with regard to capturing the dependence structure in the data. That is, we expect spatial dependence to change over time or, equivalently, we expect temporal dependence to vary with spatial location. Examples include space time trends in the deposition of atmospheric pollutants, temporal evolution of spatial patterns in soil moisture content, spatiotemporal patterns of diseases and exposure to pollutants, and spatial characterization of population dynamics in ecology.

In this regard, the separable form for the spatiotemporal covariance function in Equation 19, though convenient for computation with attractive interpretation, limits the nature of space-time interaction. There is no space-time dependence.

Cressie & Huang (1999) introduce a flexible class of nonseparable stationary covariance functions that allow for space-time interaction. However, they work in the spectral domain and require that $c(\mathbf{s} - \mathbf{s}', t - t')$ can be computed explicitly. That is, the Fourier inversion can be obtained in closed-form, which occurs only in very special cases. Work by Gneiting (2002) adopts a similar approach but obtains very general classes of valid space-time models that do not rely on closed form Fourier inversions. One simple example is the class $c(\mathbf{s} - \mathbf{s}', t - t') = \sigma^2(|t - t'| + 1)^{-1} \exp(-\|\mathbf{s} - \mathbf{s}'\|(|t - t'| + 1)^{-\beta/2})$. Here, β is a space-time interaction parameter; $\beta = 0$ provides a separable specification. Stein (2005) also works in the spectral domain, providing a class of spectral densities whose resulting spatiotemporal covariance function is nonseparable with flexible analytic behavior. In particular, the spectral density is

$$\hat{c}(\mathbf{w}, v) \propto [c_1(\alpha_1^2 + \|\mathbf{w}\|^2)^{\alpha_1} + c_2(\alpha_2 + v^2)^{\alpha_2}]^{-\nu}.$$

Unfortunately, the associated covariance function cannot be computed explicitly; fast Fourier transforms offer the best computational prospects. Unlike Gneiting's class,

separability does not arise as a special or limiting case. For further discussion of nonseparable specifications, the reader is directed to chapter 23 of Gelfand et al. (2010).

5.3. Dynamic Spatiotemporal Models

Now, we turn to discrete time. Here, dynamic linear models, often referred to as state-space models, offer a versatile framework for fitting a range of time-structured models (West & Harrison 1997). We briefly recall the general dynamic linear modeling framework. Let $\mathbf{Y}_t$ be an $m \times 1$ vector of observables at time t . $\mathbf{Y}_t$ is related to a $p \times 1$ vector, $\boldsymbol{\theta}_t$, called the state vector, through a measurement equation. In general, the elements of $\boldsymbol{\theta}_t$ are not observable, but are generated by a first-order Markovian process, resulting in a transition equation. Therefore, we can describe the above framework as

$$\mathbf{Y}_t = F_t \boldsymbol{\theta}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N\left(\mathbf{0}, \sum_t^\varepsilon\right), \quad (21)$$

$$\boldsymbol{\theta}_t = G_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim N\left(\mathbf{0}, \sum_t^\eta\right), \quad (22)$$

where F_t and G_t are $m \times p$ and $p \times p$ matrices, respectively. In the measurement equation, $\boldsymbol{\varepsilon}_t$ is an $m \times 1$ vector of serially uncorrelated Gaussian variables with mean $\mathbf{0}$ and an $m \times m$ covariance matrix, $\sum_t^\varepsilon$. In the transition equation, $\boldsymbol{\eta}_t$ is a $p \times 1$ vector of serially uncorrelated zero-centered Gaussian disturbances and $\sum_t^\eta$ the corresponding $p \times p$ covariance matrix. The association structure can be computed explicitly across time, e.g., $\text{cov}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}) = G_t \text{var}(\boldsymbol{\theta}_{t-1})$ and $\text{cov}(\mathbf{Y}_t, \mathbf{Y}_{t-1}) = F_t G_t \text{var}(\boldsymbol{\theta}_{t-1}) F_t^T$.

F_t and G_t are referred to as system matrices, which are allowed to change over time. They may involve unknown parameters, but given the parameters, temporal evolution arises in a predetermined manner. The matrix F_t is usually specified by the design of the problem at hand, and G_t is specified through modeling assumptions. For example, $G_t = I_p$, the $p \times p$ identity matrix, provides a random walk for $\boldsymbol{\theta}_t$. Regardless, the system is linear, and for any time point t , $\mathbf{Y}_t$ can be expressed as a linear combination of the present $\boldsymbol{\varepsilon}_t$ and the present and past $\boldsymbol{\eta}_s$.

5.3.1. Formulation for spatiotemporal models—Consider a collection of sites $S = \{\mathbf{s}_1, \dots, \mathbf{s}_{N_S}\}$ and time points $T = \{t_1, \dots, t_{N_T}\}$, yielding observations $Y(\mathbf{s}, t)$ and covariate vectors $\mathbf{x}(\mathbf{s}, t)$, for every $(\mathbf{s}, t) \in S \times T$.

The response, $Y(\mathbf{s}, t)$, is first modeled through a measurement equation, which incorporates the measurement error, $\varepsilon(\mathbf{s}, t)$, as serially and spatially uncorrelated zero-centered Gaussian disturbances. The transition equation now involves the regression parameters (slopes) of the covariates. The slope vector, say $\tilde{\boldsymbol{\beta}}(\mathbf{s}, t)$, is decomposed into a purely temporal component, $\boldsymbol{\beta}_t$, and a spatiotemporal component, $\boldsymbol{\beta}(\mathbf{s}, t)$. Both of these are generated through transition

equations, capturing their Markovian dependence in time. Whereas the transition equation of the purely temporal component is as in usual state-space modeling, the spatiotemporal component is generated by a multivariate Gaussian spatial process. Thus, we may write the spatiotemporal modeling framework as

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + \varepsilon(\mathbf{s}, t); \varepsilon(\mathbf{s}, t) \stackrel{\text{ind}}{\sim} N(0, \sigma^2), \quad (23)$$

$$\mu(\mathbf{s}, t) = \mathbf{x}^T(\mathbf{s}, t) \tilde{\boldsymbol{\beta}}(\mathbf{s}, t), \quad (24)$$

$$\tilde{\boldsymbol{\beta}}(\mathbf{s}, t) = \boldsymbol{\beta}_t + \boldsymbol{\beta}(\mathbf{s}, t), \quad (25)$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \stackrel{\text{ind}}{\sim} N_p(\mathbf{0}, \Sigma_{\boldsymbol{\eta}}), \quad (26)$$

$$\text{and } \boldsymbol{\beta}(\mathbf{s}, t) = \boldsymbol{\beta}(\mathbf{s}, t-1) + \mathbf{w}(\mathbf{s}, t). \quad (27)$$

We introduce a linear model of coregionalization (as in Section 3) for $\mathbf{w}(\mathbf{s}, t)$, that is, $\mathbf{w}(\mathbf{s}, t) = A\mathbf{v}(\mathbf{s}, t)$, with $\mathbf{v}(\mathbf{s}, t) = (v_1(\mathbf{s}, t), \dots, v_p(\mathbf{s}, t))^T$ yielding $\Sigma_w = AA^T$. The $v_l(\mathbf{s}, t)$ are serially independent replications of a GP with unit variance and correlation function $\rho_l(\cdot; \phi_l)$, henceforth denoted by $\text{GP}(0, \rho_l(\cdot; \phi_l))$, for $l = 1, \dots, p$ and independent across l . Allowing a spatially varying coefficient $\boldsymbol{\beta}(\mathbf{s}, t)$ to be associated with $\mathbf{x}(\mathbf{s}, t)$ provides an arbitrarily rich explanatory relationship for the Y s with regard to the Y s.

We can compute the general association structure for the Y s under the above specification. For instance, $\text{cov}(Y(\mathbf{s}, t), Y(\mathbf{s}', t-1)) = \mathbf{x}^T(\mathbf{s}, t) \Sigma_{\tilde{\boldsymbol{\beta}}(\mathbf{s}, t), \tilde{\boldsymbol{\beta}}(\mathbf{s}', t-1)} \mathbf{x}(\mathbf{s}', t-1)$, where

$$\Sigma_{\tilde{\boldsymbol{\beta}}(\mathbf{s}, t), \tilde{\boldsymbol{\beta}}(\mathbf{s}', t-1)} = (t-1) \left(\Sigma_{\boldsymbol{\eta}} + \sum_{l=1}^p \rho_l(\mathbf{s} - \mathbf{s}'; \phi_l) \mathbf{a}_l \mathbf{a}_l^T \right), \text{ and}$$

$$\text{var}(Y(\mathbf{s}, t)) = \mathbf{x}^T(\mathbf{s}, t) t \left[\Sigma_{\boldsymbol{\eta}} + AA^T \right] \mathbf{x}(\mathbf{s}, t).$$

A Bayesian hierarchical model can be completed with prior specifications such as

$$\boldsymbol{\beta}_0 \sim N(\mathbf{m}_0, C_0) \text{ and } \boldsymbol{\beta}(\cdot, 0) \equiv \mathbf{0}, \quad (28)$$

$$\sum_{\eta} \sim IW(a_{\eta}, B_{\eta}), \sum_w \sim IW(a_w, B_w) \text{ and } \sigma_{\varepsilon}^2 \sim IG(a_{\varepsilon}, b_{\varepsilon}), \quad (29)$$

$$\mathbf{m}_0 \sim N(\mathbf{0}, \sum_0); \sum_0 = 10^5 \times I_p, \quad (30)$$

where B_{η} and B_w are $p \times p$ precision (hyperparameter) matrices for the inverted Wishart distribution. A Gibbs sampler for fitting hierarchical models of this sort is relatively straightforward (see Banerjee et al. 2014 for details).

6. THE “BIG N ” PROBLEM

Not surprisingly, as we have increasing computing power, we are collecting increasingly larger spatial and spatiotemporal datasets. We now look at monitoring environmental pollutants for a large spatial network at very high temporal resolution. We now study sea-surface temperature on a global scale at relatively high resolution. Datasets are now reaching order 10^6 or greater. Model fitting becomes more and more challenging.

In particular, implementing Gibbs sampling or other MCMC model fitting algorithms requires repeated evaluation of various full conditional density functions. In the case of hierarchical models built from random effects using GPs, this requires repeated evaluation of the likelihood and/or joint or conditional densities arising under the GP. With large n , computation associated with resulting $n \times n$ matrices can be unstable, and repeated computation (as for simulation-based model fitting) can be very slow, perhaps infeasible. Cholesky factorization for matrix inversion requires about $O(n^3/3)$ flops. This situation is informally termed “the big n problem.”

Extension to multivariate models with, say, p measurements at a location leads to $np \times np$ matrices. Extension to spatiotemporal models (say, spatial time series at T time points) leads to $nT \times nT$ matrices. Modeling strategies may simplify to, say, $Tn \times n$ matrices or an $n \times n$ and a $T \times T$ matrix, but the problem will persist if n is large.

A fairly recent discussion of approaches for tackling the big n problem is provided by Banerjee et al. (2014, chapter 12). Because the challenge under our geostatistical modeling is handling the $w(\mathbf{s})$ process, here we consider strategies for replacing a customary GP with more computationally feasible GPs. One approach is through dimension reduction leading to the predictive process (Banerjee et al. 2008, Finley et al. 2009) with extensions in Sang & Huang (2012) and Katzfuss (2016). A second is in introducing sparsity through likelihood approximations and extending to a valid process (Datta et al. 2016a,b).

6.1. Reduced Rank Models

A common way of dealing with large spatial datasets is to devise models that bring about dimension reduction (see, e.g., Wikle 2010 and references therein for a nice review of dimension reduction in hierarchical models). The idea is to replace the spatial process $w(\mathbf{s})$

with $\tilde{w}(\mathbf{s})$, where the latter is a reduced rank process. A reduced- or low-rank specification is typically based upon a representation in terms of the realizations of some latent process over a smaller set of locations called knots. To be precise,

$$\tilde{w}(\mathbf{s}) = \sum_{j=1}^m l(\mathbf{s}, \mathbf{s}_j^*) Z(\mathbf{s}_j^*), \quad (31)$$

where $Z(\mathbf{s})$ is a well-defined process. The surface/process realization for $\tilde{w}(\mathbf{s})$ is completely determined by the function $l(\cdot, \cdot)$ and the set of variables, $\{Z(\mathbf{s}_j^*), j = 1, 2, \dots, m\}$. The collection of $\mathbf{s}_j^*$ s are the knots. For a collection of locations, with an associated vector denoted by $\tilde{\mathbf{w}} = (\tilde{w}(\mathbf{s}_1), \tilde{w}(\mathbf{s}_2), \dots, \tilde{w}(\mathbf{s}_n))^T$, we write

$$\tilde{\mathbf{w}} = \mathbf{L}\mathbf{z}^*, \quad (32)$$

where $\mathbf{L}$ is the $n \times m$ matrix with (i, j) th element $l(\mathbf{s}_i, \mathbf{s}_j^*)$ and $\mathbf{z}^*$ is the $m \times 1$ vector with entries $Z(\mathbf{s}_j^*)$ with $m < n$.

Equation 32 immediately reveals dimension reduction; despite there being n $\tilde{w}(\mathbf{s}_i)$ s, we will only have to work with m $Z(\mathbf{s}_j^*)$ s. If $m \ll n$, the consequential dimension reduction is evident, and because we will write the model in terms of the Z s [with the $\tilde{w}$ s being deterministic from the Z s, given $l(\cdot, \cdot)$], the associated matrices we work with will be $m \times m$. Evidently, $\tilde{w}(\mathbf{s})$ as defined in Equation 31 spans only an m -dimensional space; we create an uncountable number of variables through a finite number of variables. So, whenever $n > m$, the joint distribution of $\tilde{\mathbf{w}}$ is singular. However, we do create a valid stochastic process. In particular, the valid covariance function is

$$\text{cov}(\tilde{w}(\mathbf{s}), \tilde{w}(\mathbf{s}')) = \mathbf{l}(\mathbf{s})^T \sum_{\mathbf{z}^*} \mathbf{l}(\mathbf{s}') \quad (33)$$

where $\mathbf{l}(\mathbf{s})$ is the $m \times 1$ vector with entries $l(\mathbf{s}_i, \mathbf{s}_j^*)$. From Equation 33, we see that, even if $Z(\mathbf{s})$ is stationary, the induced covariance function is not. However, if $\mathbf{z}^*$ is Gaussian, then $\tilde{w}(\mathbf{s})$ is a GP.

If the Z s are i.i.d. normal with mean 0 and variance σ^2 , that is, $Z(\mathbf{s})$ is a white noise process, Equation 33 simplifies to $\sigma^2 \mathbf{l}(\mathbf{s})^T \mathbf{l}(\mathbf{s}')$. This form appears in Barry & Ver Hoef (1996) and in Higdon (1998), the former calling it a moving average model, the latter calling it kernel convolution. In particular, a natural choice for l is a kernel function, say $K(\mathbf{s} - \mathbf{s}')$, which puts more weight on $\mathbf{s}'$ near $\mathbf{s}$. The kernel would have parameters (which induces a parametric covariance function) and might be spatially varying (Higdon 2002, Paciorek & Schervish 2006). The reduced rank form can be viewed as a discretization of a process specification of

the form $\tilde{w}(\mathbf{s}) = \int_{\mathcal{R}^2} K(\mathbf{s}-\mathbf{s}')Z(\mathbf{s}')d\mathbf{s}'$ (see Xia & Gelfand 2006 for discussion regarding this discrete approximation). Gaussian kernels are frequently used, though they lead to Gaussian covariance functions that typically yield process realizations too smooth to be satisfactory in practice (Stein 1999a, Paciorek & Schervish 2006). Moreover, the scope of processes that can be obtained through kernel convolution is limited; for instance, the widely used exponential covariance function does not arise from kernel convolution.

Instead, Banerjee et al. (2008) propose a class of models motivated by kriging ideas. Consider a set of knots $\mathcal{S}^* = \{\mathbf{s}_1^*, \dots, \mathbf{s}_m^*\}$, which may, but need not, be a subset of the entire collection of observed locations in $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$. Assume that $w(\mathbf{s}) \sim \text{GP}(0, C(\cdot; \boldsymbol{\theta}))$ and let $\mathbf{w}^*$ be a realization of $w(\mathbf{s})$ over $\mathcal{S}^*$. That is, $\mathbf{w}^*$ is $m \times 1$ with entries $w(\mathbf{s}_i^*)$ and $\mathbf{w}^* \sim \text{MVN}(\mathbf{0}, C^*(\boldsymbol{\theta}))$, where $C^*(\boldsymbol{\theta})$ is the associated $m \times m$ covariance matrix with entries $C(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})$.

The spatial interpolant (that leads to kriging) at a site $\mathbf{s}_0$ is given by

$$\tilde{w}(\mathbf{s}_0) = E[w(\mathbf{s}_0) | \mathbf{w}^*] = \mathbf{C}^T(\mathbf{s}_0; \boldsymbol{\theta}) \mathbf{C}^{*-1}(\boldsymbol{\theta}) \mathbf{w}^*, \quad (34)$$

where $\mathbf{C}(\mathbf{s}_0; \boldsymbol{\theta})$ is $m \times 1$ with entries $C(\mathbf{s}_0, \mathbf{s}_j^*; \boldsymbol{\theta})$. This single site interpolator defines a spatial process $\tilde{w}(\mathbf{s}) \sim \text{GP}(0, \tilde{C}(\cdot))$ with covariance function

$$\tilde{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \mathbf{C}^T(\mathbf{s}; \boldsymbol{\theta}) \mathbf{C}^{*-1}(\boldsymbol{\theta}) \mathbf{C}(\mathbf{s}', \boldsymbol{\theta}), \quad (35)$$

where $\mathbf{C}(\mathbf{s}; \boldsymbol{\theta}) = [C(\mathbf{s}, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^m$. $\tilde{w}(\mathbf{s})$ is referred to as the predictive process derived from the parent process $w(\mathbf{s})$. The realizations of $\tilde{w}(\mathbf{s})$ are precisely the kriged predictions conditional upon a realization of $w(\mathbf{s})$ over $\mathcal{S}^*$. The process is completely specified given the covariance function of the parent process and $\mathcal{S}^*$. From Equation 35, this process is nonstationary regardless of whether $w(\mathbf{s})$ is. The connection with Equation 31 is clear: We take $Z(\mathbf{s})$ to be the parent process and $\mathbf{l}(\mathbf{s}) = \mathbf{C}(\mathbf{s}; \boldsymbol{\theta})^T \mathbf{C}^{*-1}(\boldsymbol{\theta})$.

Therefore, every spatial process induces a predictive process model (in fact, arbitrarily many of them, according to the knots). The latter models project process realizations of the former to a lower-dimensional subspace, thereby reducing the computational burden. Recalling Equation 1, we replace $w(\mathbf{s})$ with $\tilde{w}(\mathbf{s})$ to obtain the predictive process model,

$$Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + \tilde{w}(\mathbf{s}) + \varepsilon(\mathbf{s}). \quad (36)$$

Because $\tilde{w}(\mathbf{s}) = \mathbf{C}^T(\mathbf{s}) \mathbf{C}^{*-1}(\boldsymbol{\theta}) \mathbf{w}^*$, $\tilde{w}(\mathbf{s})$ is a spatially varying linear transformation of $\mathbf{w}^*$. The dimension reduction is seen immediately. In fitting the model in Equation 36, the n

random effects $\{w(s_i), i = 1, 2, \dots, n\}$ are replaced with only the m random effects in $\mathbf{w}^*$; we can work with an m -dimensional joint distribution involving only $m \times m$ matrices.

Though the Gaussian predictive process is induced by a customary GP, it stands in itself as a process model for spatial random effects. Moreover, it is an optimal projection process in a Kullback-Leibler sense. Among all finite-dimensional posteriors for $w(\mathbf{s})$ over $\mathcal{S}$ given $w(\mathbf{s}^*)$ over $\mathcal{S}^*$, the predictive process minimizes the reverse Kullback-Leibler distance (Csató 2002, Banerjee et al. 2008). However, realizations of the predictive process are usually too smooth. If $C(\mathbf{s}; \boldsymbol{\theta})$ is infinitely differentiable whenever $\|\mathbf{s}\| > 0$, an assumption satisfied by most covariance functions in practice, then realizations are infinitely differentiable almost surely.

The predictive process deterministically interpolates over $\mathcal{S}^*$, that is, $\tilde{w}(\mathbf{s}_j^*) = w(\mathbf{s}_j^*)$. At the other extreme, suppose $\mathcal{S}$ and $\mathcal{S}^*$ are disjoint. Then $\mathbf{w}|\mathbf{w}^* \sim MVN(\mathbf{c}^T \mathbf{C}^{*-1} \mathbf{w}^*, \mathbf{C} - \mathbf{c}^T \mathbf{C}^{*-1} \mathbf{c})$ where $\mathbf{c}$ is the $m \times n$ matrix whose columns are the $\mathbf{C}(\mathbf{s}_i)$ and $\mathbf{C}$ is the $n \times n$ covariance matrix of $\mathbf{w}$. We can write $\mathbf{w} = \tilde{\mathbf{w}} + (\mathbf{w} - \tilde{\mathbf{w}})$, and the choice of $\mathbf{w}^*$ determines the (conditional) variability in the second term on the right side, that is, how close $\Sigma_{\mathbf{w}}$ is to $\Sigma_{\tilde{\mathbf{w}}}$. It also reveals that there will be less variability in the predictive process than in the parent process as n variables are determined by $m < n$ random variables. As a result, the residual or nugget variance tends to be overestimated. In fact, the following inequality holds for any fixed $\mathcal{S}^*$ and for any spatial process $w(\mathbf{s})$:

$$\text{var}\{w(\mathbf{s})\} = \text{var}\{E[w(\mathbf{s})|\mathbf{w}^*]\} + E\{\text{var}[w(\mathbf{s})|\mathbf{w}^*]\} \geq \text{var}\{E[w(\mathbf{s})|\mathbf{w}^*]\}, \quad (37)$$

which implies that $\text{var}\{w(\mathbf{s})\} \geq \text{var}\{\tilde{w}(\mathbf{s})\}$. If the process is Gaussian, then standard multivariate normal calculations yield the following closed form for the difference at any arbitrary location $\mathbf{s}$,

$$E\{\text{var}[w(\mathbf{s})|\mathbf{w}^*]\} = \mathbf{C}(\mathbf{s}, \mathbf{s}; \boldsymbol{\theta}_1) - \mathbf{C}(\mathbf{s}; \boldsymbol{\theta}_1)' \mathbf{C}^*(\boldsymbol{\theta}_1)^{-1} \mathbf{C}(\mathbf{s}, \boldsymbol{\theta}_1). \quad (38)$$

One simple remedy for the bias in the predictive process model (Finley et al. 2009) is to use a modified predictive process

$$\tilde{w}_\varepsilon(\mathbf{s}) = \tilde{w}(\mathbf{s}) + \tilde{\varepsilon}(\mathbf{s}), \quad (39)$$

where $\tilde{\varepsilon}(\mathbf{s}) \stackrel{\text{iid}}{\sim} N(0, \delta^2(\mathbf{s}))$ and $\tilde{\varepsilon}(\mathbf{s})$ is independent of $\tilde{w}(\mathbf{s})$. Now, the variance of $\tilde{w}_\varepsilon(\mathbf{s})$ equals that of the parent process $w(\mathbf{s})$. The remedy is computationally efficient—adding an independent space-varying nugget does not incur substantial computational expense.

Sang et al. (2011) and Sang & Huang (2012) note that this remedy may not capture small-scale dependence, that is, the off-diagonal terms, and so they propose a tapering covariance

adjustment to $\tilde{\epsilon}(\mathbf{s})$. Katzfuss (2013) has extended the tapering ideas in Sang et al. (2011) with regard to predictive process modeling. He specifies spatial error using two terms, combining a low-rank component to capture both medium-to-long-range dependence with a tapered residual component to capture local dependence. Nonstationary Matérn covariance functions are adopted to provide increased flexibility. Instead of tapering the residual process, Katzfuss (2016) uses a sequence of nested predictive processes to develop multiresolution approximations for massive spatial datasets.

6.2. The Nearest-Neighbor Gaussian Process

Rather than adopting a process that resides in an m -dimensional space, Datta et al. (2016a, b) offer an alternative computationally efficient strategy by constructing a sparsity-inducing spatial process, also induced by a given GP. In particular, they show how likelihood approximations using lower-dimensional conditional distributions will yield proper Gaussian densities with sparse precision matrices that correspond to the distribution of finite-dimensional realizations of a well-defined spatial process called the nearest-neighbor GP (NNGP). When n is very large, a small number m of knots in low-rank processes will tend to oversmooth and cannot be expected to adequately capture a spatial surface. Here, the NNGP clearly outperforms low-rank models. Furthermore, it is massively scalable and can be conveniently embedded within hierarchical models for analyzing spatial (or, in fact, space-time) data comprising more than 10^5 spatial (or space-time) points.

Let $w(\mathbf{s}) \sim \text{GP}(0, C(\cdot, \cdot | \boldsymbol{\theta}))$ denote a zero-centered GP with covariance function $C(\mathbf{s}, \mathbf{s}')$. Let $\mathcal{R} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_r\}$ be a fixed finite collection of locations in $\mathcal{D}$, which we call the reference set. If $\mathbf{w}_{\mathcal{R}} = (w(\mathbf{s}_1), w(\mathbf{s}_2), \dots, w(\mathbf{s}_r))^T$ then $\mathbf{w}_{\mathcal{R}} \sim N(\mathbf{0}, C_{\mathcal{R}}(\boldsymbol{\theta}))$, where $C_{\mathcal{R}}(\boldsymbol{\theta})$ is a positive definite $r \times r$ matrix with $C(\mathbf{s}_i, \mathbf{s}_j)$ as its (i, j) th element. The NNGP is a model that specifies a massively scalable GP in two steps. First, we specify a very computationally efficient multivariate Gaussian distribution that approximates the realizations of the process over the reference set. Subsequently we extend it to a process over the domain through kriging equations based on the locations in $\mathcal{R}$. These two steps are described below.

The probability density for the realizations $\mathbf{w}_{\mathcal{R}}$ is $N(\mathbf{0}, C_{\mathcal{R}}(\boldsymbol{\theta}))$. The first step seeks an approximation to this density. Following Vecchia (1988) and Stein et al. (2004), we write the joint density of $\mathbf{w}$ as $p(\mathbf{w}) = \prod_{i=1}^r p(w(\mathbf{s}_i) | \mathbf{w}_{H(\mathbf{s}_i)})$, where $H(\mathbf{s}_1)$ is the empty set, $H(\mathbf{s}_j) = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_{j-1}\}$ and $\mathbf{w}_{H(\mathbf{s}_j)} = \{w(\mathbf{s}_j) | \mathbf{s}' \in H(\mathbf{s}_j)\}$ for $2 \leq j \leq r$. We replace the conditioning sets $\mathbf{w}_{H(\mathbf{s}_j)}$ with $\mathbf{w}_{N(\mathbf{s}_j)}$, where $N(\mathbf{s}_j)$ is a set of m nearest neighbors (defined in terms of Euclidean distance) of $\mathbf{s}_j$ from the points in $H(\mathbf{s}_j)$ whenever $H(\mathbf{s}_j)$ has more than m points. Datta et al. (2016a) show that the resulting density is a normal density. Thus,

$$N(\mathbf{w}_{\mathcal{R}} | \mathbf{0}, C_{\mathcal{R}}(\boldsymbol{\theta})) = \prod_{i=1}^r p(w(\mathbf{s}_i) | \mathbf{w}_{H(\mathbf{s}_i)}) \approx \prod_{i=1}^r p(w(\mathbf{s}_i) | \mathbf{w}_{N(\mathbf{s}_i)}) = N(\mathbf{w}_{\mathcal{R}} | \mathbf{0}, K(\boldsymbol{\theta})). \quad (40)$$

The underlying idea behind the approximation in Equation 40 is to compress the conditioning sets from $H(\mathbf{s}_j)$ to $N(\mathbf{s}_j)$. This is analogous to constructing sparse Gaussian

graphical models or sparse Bayesian networks (see, e.g., Murphy 2012, chapter 10). It assumes conditional independence so that $p(w(s_i) | \mathbf{w}_{H(s_i)}) = p(w(s_i) | \mathbf{w}_{N(s_i)})$ for each s_i in the reference set. This results in a multivariate normal distribution with a sparse precision matrix $K^{-1}(\boldsymbol{\theta})$. Datta et al. (2016a) derive the exact structure of $K^{-1}(\boldsymbol{\theta})$ in closed form. It has at most $m(m+1)/2$ nonzero entries. Smaller values of m (number of neighbors) produce sparser $K^{-1}(\boldsymbol{\theta})$. Importantly, $\det(K(\boldsymbol{\theta}))$ is a product of r determinants of order at most $m \times m$. Again, with $m \ll r$, the computational savings, both in terms of storage and floating point operations, are evident.

The second step extends this approximation to a process. To do this, we first define $N(s)$, for any arbitrary site s , to be the set of m nearest neighbors of s among the sites in $\mathcal{R}$. Then, for any $s \notin \mathcal{R}$, we define

$$w(s) = \sum_{j=1}^r a_j(s) w(s_j) + \eta(s), \quad \eta(s_i) \stackrel{\text{ind}}{\sim} N(0, f(s)), \quad (41)$$

where $a_j(s) = 0$ whenever $s_j \notin N(s)$. This implies that the nonzero $a_j(s)$ s are the coefficients for kriging $w(s)$ on its neighbors in the reference set and $f(s)$ is the corresponding kriging variance (i.e., the variance of the kriging error). This development of the NNGP in terms of nearest-neighbor kriging is developed in Datta et al. (2016c). The covariance function corresponding to the NNGP is easy to derive and is supplied explicitly in Datta et al. (2016a). Although the resultant $K(\boldsymbol{\theta})$ is not invariant to the order in which we define $H(s_1) \subseteq H(s_2) \subseteq \dots \subseteq H(s_r)$, both Vecchia (1988) and Stein et al. (2004) assert that the approximation in Equation 40 is not sensitive to this ordering. This is corroborated by simulation experiments found in the supplement accompanying Datta et al. (2016a). Simulation experiments in Datta et al. (2016a,b) also seem to reveal that m can usually be taken to be very small compared with r ; there seems to be no inferential advantage to taking m to exceed 15, even for datasets comprising over 10^5 spatial locations.

For the approximation in Equation 40 to be effective and to ensure adequately representative neighbor sets, the size of the reference set, r , needs to be large enough to represent the spatial domain. However, this does not impede computations involving NNGP models because the storage and number of floating point operations are always linear in r . The reference set $\mathcal{R}$ can, in principle, be any finite set of locations in the study domain. A particularly convenient choice is to simply take $\mathcal{R}$ to be the set of observed locations in the dataset. Datta et al. (2016a) demonstrate through extensive simulation experiments and a real application that this simple choice seems to be very effective and that there seem to be no apparent benefits with alternate choices for $\mathcal{R}$. Because the NNGP is a proper GP, we can use it as a prior for the spatial random effects in any hierarchical model formulation. For example, if $\mathcal{R}$ is the set of observed locations, we may specify

$$p(\boldsymbol{\theta}, \tau, \beta) \times N(\mathbf{w}_{\mathcal{R}} | \mathbf{0}, K(\boldsymbol{\theta})) \times \prod_{i=1}^n N(y(\mathbf{s}_i) | \mathbf{x}^T(\mathbf{s}_i)\beta + w(\mathbf{s}_i), \tau^2),$$

where $y(\mathbf{s}_i)$ is the outcome measured at location $\mathbf{s}_i$, $\mathbf{x}(\mathbf{s}_i)$ is a vector of predictors, and τ^2 is the nugget variance corresponding to $y(\mathbf{s}_i)$. An efficient MCMC sampler using Gibbs steps and random-walk Metropolis steps described by Datta et al. (2016a) can now be used for updating $w(\mathbf{s})$, β , the covariance parameters $\boldsymbol{\theta}$, and the nugget τ^2 . Prediction at an arbitrary location $\mathbf{s} \notin \mathcal{R}$ is performed by sampling from the posterior predictive distribution

$$\int N(y(\mathbf{s}) | \mathbf{x}^T(\mathbf{s})\beta + w(\mathbf{s}), \tau^2) \times N(w(\mathbf{s}) | \mathbf{a}^T(\mathbf{s})\mathbf{w}_{N(\mathbf{s})}, f(\mathbf{s})) \times p(\mathbf{w}_{\mathcal{R}}, \beta, \tau^2, \boldsymbol{\theta} | \mathbf{y}_{\mathcal{R}}),$$

where $\mathbf{x}^T(\mathbf{s})$ is the corresponding matrix of predictors for location $\mathbf{s}$, and $\mathbf{a}(\mathbf{s})$ is a vector of the nonzero $a_j(\mathbf{s})$ s in (41). We refer the reader to Datta et al. (2016a,b) for details.

We conclude this section with a brief comparison among the full GP, a Gaussian predictive process, and the NNGP using a simulated Gaussian random field over 2,500 spatial locations. The details are provided by Datta et al. (2016a). Figure 2(a–c) presents posterior median estimates of the spatial random effects from a full GP, an NNGP model with $m = 10$, and a modified predictive process model with 64 knots. This figure shows that the NNGP model with $m = 10$ closely approximates the full GP model, whereas the reduced rank modified predictive process model based on 64 knots consequently smooths over small-scale patterns. This last observation highlights one of the major criticisms of reduced rank models (Stein 2014) and illustrates why these models often provide compromised predictive performance when the true surface has fine spatial resolution details. We direct the reader to Datta et al. (2016a) for further details and clarifications.

7. OTHER CHALLENGES

7.1. Efficient and User-Friendly Computation

Computational developments with regard to MCMC algorithms (see, e.g., Robert & Casella 2004) have contributed enormously to the popularity of Bayesian hierarchical models in a wide array of disciplines, and spatial modeling is no exception. However, the challenges for automated implementation of geostatistical model fitting and inference are substantial. First, expensive matrix computations are required that can become prohibitive with large datasets. Second, routines to fit unmarginalized models are less suited for direct updating using a Gibbs sampler in the BUGS (Bayesian Inference Using the Gibbs Sampler) paradigm (Finley et al. 2007) and result in slower convergence of the chains. Third, investigators often encounter multivariate spatial datasets with several spatially dependent outcomes, whose analysis requires multivariate spatial models that involve matrix computations that are poorly implemented in BUGS. These issues have, however, started to wane with the delivery of relatively simpler software packages in the R statistical computing environment via CRAN (<http://cran.r-project.org>). Several packages that automate Bayesian methods for point-referenced data and diagnose convergence of MCMC algorithms are easily available

from CRAN. Packages that fit Bayesian models include `geoR` (Ribeiro & Diggle 2001), `geoRglm` (Christensen & Ribeiro 2002), `spTimer` (Bakar & Sahu 2014), `spBayes` (Finley et al. 2015), `spate` (Sigrist et al. 2015), and `ramps` (Smith et al. 2008).

In terms of the hierarchical geostatistical models presented in this article, `spBayes` offers users a suite of Bayesian hierarchical models for Gaussian and non-Gaussian univariate and multivariate spatial data as well as dynamic Bayesian spatiotemporal models. It focuses upon performance issues for full Bayesian inference, sampler convergence rate, and efficiency using a collapsed Gibbs sampler. It decreases sampler run-time by avoiding expensive matrix computations, and increases scalability to large datasets by implementing predictive process models (Section 6). Beyond these general computational improvements for existing models, it analyzes data indexed in both space and time using a class of dynamic spatiotemporal models, and their predictive process counterparts, for settings where space is viewed as continuous and time is taken as discrete.

7.2. Multivariate Explanation, Multivariate Response

An illustration of a geostatistical problem that currently lacks a satisfying solution can be cast as a joint species distribution model (see, e.g., Pollock et al. 2014 and references therein). Illustrative data might consist of presence/absence (binary) or abundance (counts or percent coverage) for a large collection of species (say, order 10^3) over a large collection of spatial sites (again, say, order 10^3). At each site, a suite of environmental predictors is available. At any site there will be a large number of observed 0s. It is acknowledged that there is dependence between species with respect to co-occurrence or co-abundance and also, with regard to the response, that there will be spatial dependence over the sites. At present, very restrictive assumptions are needed to address this problem.

LITERATURE CITED

- Bakar KS, Sahu SK. `spTimer`: spatio-temporal Bayesian modeling using R. *J Stat Softw.* 2014; 63:1–32.
- Banerjee, S., Carlin, BP., Gelfand, AE. *Hierarchical Modeling and Analysis for Spatial Data.* 2. Boca Raton, FL: Chapman and Hall/CRC Press; 2014.
- Banerjee S, Gelfand AE, Finley AO, Sang H. Gaussian predictive process models for large spatial datasets. *J R Stat Soc B.* 2008; 70:825–48.
- Barry RP, Ver Hoef JM. Blackbox kriging: spatial prediction without specifying variogram models. *J Agric Biol Environ Stat.* 1996; 1:297–322.
- Berger JO, DeOliveira V, Sansó B. Objective Bayesian analysis of spatially correlated data. *J Am Stat Assoc.* 2001; 96:1361–74.
- Berliner LM. Hierarchical Bayesian modeling in the environmental sciences. *AStA Adv Stat Anal.* 2000; 84:141–53.
- Berrocal VJ, Gelfand AE, Holland DM. A spatio-temporal downscaler for outputs from numerical models. *J Agric Biol Environ Stat.* 2010; 15:176–97. [PubMed: 21113385]
- Chiles, JP., Delfiner, P. *Geostatistics: Modeling Spatial Uncertainty.* New York: Wiley; 1999.
- Christensen OF, Ribeiro PJ Jr. `geoRglm`: A package for generalised linear spatial models. *RNews.* 2002; 2:26–28.
- Clark JS, Gelfand AE. A future for models and data in ecology. *Trends Ecol Evol.* 2006; 21:375–80. [PubMed: 16815437]

- Cressie N, Huang H-C. Classes of nonseparable spatio-temporal stationary covariance functions. *J Am Stat Assoc.* 1999; 94:1330–40.
- Cressie, N., Wikle, CK. *Statistics for Spatiotemporal Data.* New York: Wiley; 2011.
- Csató, L. PhD thesis. Aston Univ; Birmingham, UK: 2002. Gaussian processes—iterative sparse approximations.
- Datta A, Banerjee S, Finley AO, Gelfand AE. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J Am Stat Assoc.* 2016a; 111:800–12.
- Datta A, Banerjee S, Finley AO, Gelfand AE. On nearest-neighbor Gaussian process models for massive spatial data. *WIREs Comput Stat.* 2016c; (8):162–71.
- Datta A, Banerjee S, Finley AO, Hamm NAS, Schaap M. Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *Ann Appl Stat.* 2016b; 10:1286–316.
- DeOliveira V. Bayesian prediction of clipped Gaussian random fields. *Comput Stat Data Anal.* 2000; 34:299–314.
- Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics (with discussion). *J R Stat Soc C.* 1998; 47:299–350.
- Finley AO, Banerjee S, Carlin BP. spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *J Stat Softw.* 2007; 19:4.
- Finley AO, Banerjee S, Gelfand AE. spBayes for large univariate and multivariate point-referenced spatiotemporal data models. *J Stat Softw.* 2015; 63:1–28.
- Finley AO, Sang H, Banerjee S, Gelfand AE. Improving the performance of predictive process modeling for large datasets. *Comput Stat Data Anal.* 2009; 53:2873–84. [PubMed: 20016667]
- Fuentes M, Raftery AE. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics.* 2005; 61:36–45. [PubMed: 15737076]
- Gelfand, AE., Diggle, P., Fuentes, M., Guttorp, P. *Handbook of Spatial Statistics.* Boca Raton, FL: CRC; 2010.
- Gelfand AE, Kim H-J, Sirmans CF, Banerjee S. Spatial modeling with spatially varying coefficient processes. *J Am Stat Assoc.* 2003; 98:387–96.
- Gelfand AE, Schmidt AM, Banerjee S, Sirmans CF. Nonstationary multivariate process modeling through spatially varying coregionalization (with discussion). *Test.* 2004; 13:263–312.
- Gelfand AE, Zhu L, Carlin BP. On the change of support problem for spatio-temporal data. *Biostatistics.* 2001; 2:31–45. [PubMed: 12933555]
- Gneiting T. Nonseparable, stationary covariance functions for space-time data. *J Am Stat Assoc.* 2002; 97:590–600.
- Higdon DM. A process-convolution approach to modeling temperatures in the north Atlantic Ocean. *J Environ Ecol Stat.* 1998; 5:173–90.
- Higdon, DM. Space and space-time modeling using process convolutions. In: Anderson, C. Barnett, V. Chatwin, PC., El-Shaarawi, AH., editors. *Quantitative Methods for Current Environmental Issues.* London: Springer-Verlag; 2002. p. 37–56.
- Kalnay, E. *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge, UK: Cambridge Univ. Press; 2003.
- Katzfuss M. Bayesian nonstationary spatial modeling for very large datasets. *Environmetrics.* 2013; 24:189–200.
- Katzfuss, M. A multi-resolution approximation for massive spatial datasets. *J Am Stat Assoc.* 2016. <http://dx.doi.org/10.1080/01621459.2015.1123632>
- Krige DG. A statistical approach to some basic mine valuation problems on the Witwatersrand. *J Chem Metall Min Soc S Afr.* 1951; 52:119–39.
- Little, RJA., Rubin, DB. *Statistical Analysis with Missing Data.* New York: Wiley-Interscience; 2002.
- Loh W-H. Fixed-domain asymptotics for a subclass of Matérn-type Gaussian random fields. *Ann Stat.* 2005; 33:2344–94.

- Mardia, KV., Goodall, C. Spatio-temporal analyses of multivariate environmental monitoring data. In: Patil, GP., Rao, CR., editors. *Multivariate Environmental Statistics*. Amsterdam: Elsevier; 1993. p. 347-86.
- Murphy, KP. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press; 2012.
- Paciorek CJ, Schervish M. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*. 2006; 17:483–506. [PubMed: 18163157]
- Pollock LJ, Tingley R, Morris WK, Golding N, O'Hara RB, et al. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods Ecol Evol*. 2014; 5:397–406.
- Poole D, Raftery AE. Inference for deterministic simulation models: the Bayesian melding approach. *J Am Stat Assoc*. 2000; 95:1244–55.
- Ribeiro PJ Jr, Diggle PJ. geoR: A package for geostatistical analysis. *R News*. 2001; 1/2:15–18.
- Robert, CP., Casella, G. *Monte Carlo Statistical Methods*. 2. New York: Springer-Verlag; 2004.
- Sang H, Huang JZ. A full scale approximation of covariance functions for large spatial data sets. *J R Stat Soc B*. 2012; 74:111–32.
- Sang H, Jun M, Huang JZ. Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *Ann Appl Stat*. 2011; 4:2519–48.
- Sigrist F, Kuensch HR, Stahel WA. spate: An R package for spatio-temporal modeling with a stochastic advection-diffusion process. *J Stat Softw*. 2015; 63:1–23.
- Smith BJ, Yan J, Cowles MK. Unified geostatistical modeling for data fusion and spatial heteroskedasticity with R package ramps. *J Stat Softw*. 2008; 25:1–21.
- Stein, ML. *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer-Verlag; 1999a.
- Stein ML. Predicting random fields with increasingly dense observations. *Ann Appl Probability*. 1999b; 9:242–73.
- Stein ML. Space-time covariance functions. *J Am Stat Assoc*. 2005; 100:310–21.
- Stein ML. Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Stat*. 2014; 8:1–19.
- Stein ML, Chi Z, Welty LJ. Approximating likelihoods for large spatial datasets. *J R Stat Soc B*. 2004; 66:275–96.
- Vecchia AV. Estimation and model identification for continuous spatial processes. *J R Stat Soc B*. 1988; 50:297–312.
- Ver Hoef JM, Barry RP. Constructing and fitting models for cokriging and multivariable spatial prediction. *J Stat Plann Inference*. 1998; 69:275–94.
- Wackernagel H. Cokriging versus kriging in regionalized multivariate data analysis. *Geoderma*. 1994; 62:83–92.
- Wackernagel, H. *Multivariate Geostatistics: An Introduction with Applications*. 3. New York: Springer-Verlag; 2003.
- West, M., Harrison, PJ. *Bayesian Forecasting and Dynamic Models*. 2. New York: Springer-Verlag; 1997.
- Wikle, CK. Low-rank representations for spatial processes. In: Gelfand, AE, Diggle, P, Fuentes, M., Guttorp, P., editors. *Handbook of Spatial Statistics*. Boca Raton, FL: Chapman and Hall/CRC; 2010. p. 107-18.
- Xia, G., Gelfand, AE. Tech Rep. ISDS, Duke Univ; 2006. Stationary process approximation for the analysis of large spatial datasets.
- Yaglom, AM. *Correlation Theory of Stationary and Related Random Functions: Volume I: Basic Results*. London: Springer; 1987.
- Zhang H, Zimmerman DL. Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*. 2005; 92:921–36.

Figure 1. Interpolated surface of observed log metric tons of biomass from the Bartlett Experimental Forest data.

Figure 2.

Interpolated posterior median estimates of spatial processes from three different models.

Abbreviations: GP, Gaussian process; GPP, Gaussian predictive process; m , number of neighbors fixed in the NNGP model; NNGP, nearest-neighbor GP.