Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets

Abhirup Datta, Sudipto Banerjee, Andrew O. Finley & Alan E. Gelfand

To cite this article: Abhirup Datta, Sudipto Banerjee, Andrew O. Finley & Alan E. Gelfand (2016) Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets, Journal of the American Statistical Association, 111:514, 800-812, DOI: 10.1080/01621459.2015.1044091

To link to this article: https://doi.org/10.1080/01621459.2015.1044091

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 514, 800–812, Theory and Methods
http://dx.doi.org/10.1080/01621459.2015.1044091

Taylor & Francis
Taylor & Francis Group

Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets

Abhirup Datta, Sudipto Banerjee, Andrew O. Finley, and Alan E. Gelfand

ABSTRACT

Spatial process models for analyzing geostatistical data entail computations that become prohibitive as the number of spatial locations become large. This article develops a class of highly scalable nearest-neighbor Gaussian process (NNGP) models to provide fully model-based inference for large geostatistical datasets. We establish that the NNGP is a well-defined spatial process providing legitimate finite-dimensional Gaussian densities with sparse precision matrices. We embed the NNGP as a sparsity-inducing prior within a rich hierarchical modeling framework and outline how computationally efficient Markov chain Monte Carlo (MCMC) algorithms can be executed without storing or decomposing large matrices. The floating point operations (flops) per iteration of this algorithm is linear in the number of spatial locations, thereby rendering substantial scalability. We illustrate the computational and inferential benefits of the NNGP over competing methods using simulation studies and also analyze forest biomass from a massive U.S. Forest Inventory dataset at a scale that precludes alternative dimension-reducing methods. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2014
Revised April 2015

KEYWORDS
Bayesian modeling; Gaussian process; Hierarchical models; Markov chain Monte Carlo; Nearest neighbors; Predictive process; Reduced-rank models; Sparse precision matrices; Spatial cross-covariance functions

1. Introduction

With the growing capabilities of Geographical Information Systems (GIS) and user-friendly software, statisticians today routinely encounter geographically referenced datasets containing a large number of irregularly located observations on multiple variables. This has, in turn, fueled considerable interest in statistical modeling for location-referenced spatial data; see, for example, the books by Stein (1999), Moller and Waagepetersen (2003), Schabenberger and Gotway (2004), and Cressie and Wikle (2011), and Banerjee, Carlin, and Gelfand (2014) for a variety of methods and applications. Spatial process models introduce spatial dependence between observations using an underlying random field, $\{w(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$, over a region of interest $\mathcal{D}$, which is endowed with a probability law that specifies the joint distribution for any finite set of random variables. For example, a zero-centered Gaussian process ensures that $\mathbf{w} = (w(\mathbf{s}_1), w(\mathbf{s}_2) \ldots, w(\mathbf{s}_n))' \sim N(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}))$, where $\mathbf{C}(\boldsymbol{\theta})$ is a family of covariance matrices, indexed by an unknown set of parameters $\boldsymbol{\theta}$. Such processes offer a rich modeling framework and are being widely deployed to help researchers comprehend complex spatial phenomena in the sciences. However, model fitting usually involves the inverse and determinant of $\mathbf{C}(\boldsymbol{\theta})$, which typically require $\sim n^3$ floating point operations (flops) and storage of the order of n^2. These become prohibitive when n is large and $\mathbf{C}(\boldsymbol{\theta})$ has no exploitable structure.

Broadly speaking, modeling large spatial datasets proceeds from either exploiting "low-rank" models or using sparsity. The former attempts to construct spatial processes on a lower-dimensional subspace (see, e.g., Higdon 2001; Kammann and Wand 2003; Rasmussen and Williams 2005; Stein 2007, 2008; Banerjee et al. 2008; Crainiceanu et al. 2008; Cressie and Johannesson 2008; Finley, Banerjee, and McRoberts 2009) by regressing the original (*parent*) process on its realizations over a smaller set of $r << n$ locations ("knots" or "centers"). The algorithmic cost for model fitting typically decreases from $O(n^3)$ to $O(nr^2 + r^3) \approx O(nr^2)$ flops since $n >> r$. However, when n is large, empirical investigations suggest that r must be fairly large to adequately approximate the parent process and the nr^2 flops become exorbitant (see Section 5.1). Furthermore, low-rank models perform poorly when neighboring observations are strongly correlated and the spatial signal dominates the noise (Stein 2014). Although bias-adjusted low-rank models tend to perform better (Finley, Banerjee, and McRoberts 2009; Banerjee et al. 2010; Sang and Huang 2012), they increase the computational burden.

Sparse methods include covariance tapering (see, e.g., Furrer, Genton, and Nychka 2006; Kaufman, Scheverish, and Nychka 2008; Du, Zhang, and Mandrekar 2009; Shaby and Ruppert 2012), which introduces sparsity in $\mathbf{C}(\boldsymbol{\theta})$ using compactly supported covariance functions. This is effective for parameter estimation and interpolation of the response ("kriging"), but it has not been fully developed or explored for more general inference on residual or latent processes. Introducing sparsity in $\mathbf{C}(\boldsymbol{\theta})^{-1}$ is prevalent in approximating Gaussian process likelihoods using Markov random fields (e.g., Rue and Held 2005), products of lower-dimensional conditional distributions (Vecchia 1988, 1992; Stein, Chi, and Welty 2004), or composite likelihoods (e.g., Bevilacqua and Gaetan 2014; Eidsvik et al. 2014). However, unlike low-rank processes, these do not, necessarily,

CONTACT Sudipto Banerjee sudipto@ucla.edu Department of Biostatistics, UCLA Fielding School of Public Health, Los Angeles, CA 90095-1772
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*

extend to new random variables at arbitrary locations. There may not be a corresponding process, which restricts inference to the estimation of spatial covariance parameters. Spatial prediction ("kriging") at arbitrary locations proceeds by imputing estimates into an interpolator derived from a different process model. This may not reflect accurate estimates of predictive uncertainty and is undesirable.

Our intended inferential contribution is to offer substantial scalability for fully process-based inference on underlying, perhaps completely unobserved, spatial processes. Moving from finite-dimensional sparse likelihoods to sparsity-inducing spatial processes can be complicated. We first introduce sparsity in finite-dimensional probability models using specified neighbor sets constructed from directed acyclic graphs. We use these sets to extend these finite-dimensional models to a valid spatial process over uncountable sets. We call this process a nearest-neighbor Gaussian process (NNGP). Its finite-dimensional realizations have sparse precision matrices available in closed form. While sparsity has been effectively exploited by Vecchia (1988), Stein, Chi, and Welty (2004), Emory (2009), Gramacy and Apley (2014), Gramacy, Niemi, and Weiss (2014), and Stroud, Stein, and Lysen (2014) for approximating expensive likelihoods cheaply, a fully process-based modeling and inferential framework has, hitherto, proven elusive. The NNGP fills this gap and enriches the inferential capabilities of existing methods by subsuming estimation of model parameters, prediction of outcomes, and interpolation of underlying processes into one highly scalable unifying framework.

To demonstrate its full inferential capabilities, we deploy the NNGP as a sparsity-inducing prior for spatial processes in a Bayesian framework. Unlike low-rank processes, the NNGP always specifies nondegenerate finite dimensional distributions making it a legitimate proper prior for random fields and is applicable to any class of distributions that support a spatial stochastic process. It can, therefore, model an underlying process that is never actually observed. The modeling provides structured dependence for random effects, for example, intercepts or coefficients, at a second stage of specification where the first stage need not be Gaussian. We cast a multivariate NNGP within a versatile spatially varying regression framework (Gelfand et al. 2003; Banerjee et al. 2008) and conveniently obtain entire posteriors for all model parameters as well as for the spatial processes at both observed and unobserved locations. Using a forestry example, we show how the NNGP delivers process-based inference for spatially varying regression models at a scale where even low-rank processes, let alone full Gaussian processes, are unimplementable even in high-performance computing environments.

Here is a brief outline. Section 2 formulates the NNGP using multivariate Gaussian processes. Section 3 outlines Bayesian estimation and prediction within a very flexible hierarchical modeling setup. Section 4 discusses alternative NNGP models and algorithms. Section 5 presents simulation studies to highlight the inferential benefits of the NNGP and also analyzes forest biomass from a massive USDA dataset. Finally, Section 6 concludes the article with a brief summary and pointers toward future work.

2. Nearest-Neighbor Gaussian Process

2.1 Gaussian Density on Sparse Directed Acyclic Graphs

We will consider a q-variate spatial process over $\Re^d$. Let $\mathbf{w}(\mathbf{s}) \sim GP(\mathbf{0}, \mathbf{C}(\cdot,\cdot \mid \boldsymbol{\theta}))$ denote a zero-centered q-variate Gaussian process, where $\mathbf{w}(\mathbf{s}) \in \Re^q$ for all $\mathbf{s} \in \mathcal{D} \subseteq \Re^d$. The process is completely specified by a valid cross-covariance function $\mathbf{C}(\cdot,\cdot \mid \boldsymbol{\theta})$, which maps a pair of locations $\mathbf{s}$ and $\mathbf{t}$ in $\mathcal{D} \times \mathcal{D}$ into a $q \times q$ real-valued matrix $\mathbf{C}(\mathbf{s},\mathbf{t})$ with entries $\text{cov}\{w_i(\mathbf{s}), w_j(\mathbf{t})\}$. Here, $\boldsymbol{\theta}$ denotes the parameters associated with the cross-covariance function. Let $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$ be a fixed collection of distinct locations in $\mathcal{D}$, which we call the *reference set*. So, $\mathbf{w}_\mathcal{S} \sim N(\mathbf{0}, \mathbf{C}_\mathcal{S}(\boldsymbol{\theta}))$, where $\mathbf{w}_\mathcal{S} = (\mathbf{w}(\mathbf{s}_1)', \mathbf{w}(\mathbf{s}_2)', \ldots, \mathbf{w}(\mathbf{s}_k)')'$ and $\mathbf{C}_\mathcal{S}(\boldsymbol{\theta})$ is a positive definite $qk \times qk$ block matrix with $\mathbf{C}(\mathbf{s}_i, \mathbf{s}_j)$ as its blocks. Henceforth, we write $\mathbf{C}_\mathcal{S}(\boldsymbol{\theta})$ as $\mathbf{C}_\mathcal{S}$, the dependence on $\boldsymbol{\theta}$ being implicit, with similar notation for all spatial covariance matrices.

The reference set $\mathcal{S}$ need not coincide with or be a part of the observed locations, so k need not equal n, although we later show that the observed locations are a convenient practical choice for $\mathcal{S}$. When k is large, parameter estimation becomes computationally cumbersome, perhaps even unfeasible, because it entails the inverse and determinant of $\mathbf{C}_\mathcal{S}$. Here, we benefit from expressing the joint density of $\mathbf{w}_\mathcal{S}$ as the product of conditional densities, that is,

$$p(\mathbf{w}_\mathcal{S}) = p(\mathbf{w}(\mathbf{s}_1))\; p(\mathbf{w}(\mathbf{s}_2) \mid \mathbf{w}(\mathbf{s}_1)) \\ \ldots\; p(\mathbf{w}(\mathbf{s}_k) \mid \mathbf{w}(\mathbf{s}_{k-1}), \ldots, \mathbf{w}(\mathbf{s}_1))\;, \tag{1}$$

and replacing the larger conditioning sets on the right-hand side of (1) with smaller, carefully chosen, conditioning sets of size at most m, where $m \ll k$ (see, e.g., Vecchia 1988; Stein, Chi, and Welty 2004; Gramacy and Apley 2014; Gramacy, Niemi, and Weiss 2014). So, for every $\mathbf{s}_i \in \mathcal{S}$, a smaller conditioning set $N(\mathbf{s}_i) \subset \mathcal{S} \setminus \{\mathbf{s}_i\}$ is used to construct

$$\tilde{p}(\mathbf{w}_\mathcal{S}) = \prod_{i=1}^{k} p(\mathbf{w}(\mathbf{s}_i) \mid \mathbf{w}_{N(\mathbf{s}_i)})\;, \tag{2}$$

where $\mathbf{w}_{N(\mathbf{s}_i)}$ is the vector formed by stacking the realizations of $\mathbf{w}(\mathbf{s})$ over $N(\mathbf{s}_i)$.

Let $N_\mathcal{S} = \{N(\mathbf{s}_i);\, i = 1, 2, \ldots, k\}$ be the collection of all conditioning sets over $\mathcal{S}$. We can view the pair $\{\mathcal{S}, N_\mathcal{S}\}$ as a directed graph $\mathcal{G}$ with $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$ being the set of nodes and $N_\mathcal{S}$ the set of directed edges. For every two nodes $\mathbf{s}_i$ and $\mathbf{s}_j$, we say $\mathbf{s}_j$ is a directed neighbor of $\mathbf{s}_i$ if there is a directed edge from $\mathbf{s}_i$ to $\mathbf{s}_j$. So, $N(\mathbf{s}_i)$ denotes the set of directed neighbors of $\mathbf{s}_i$ and is, henceforth, referred to as the "neighbor set" for $\mathbf{s}_i$. A "directed cycle" in a directed graph is a chain of nodes $\mathbf{s}_{i_1}, \mathbf{s}_{i_2}, \ldots, \mathbf{s}_{i_b}$ such that $\mathbf{s}_{i_1} = \mathbf{s}_{i_b}$ and there is a directed edge between $\mathbf{s}_{i_j}$ and $\mathbf{s}_{i_{j+1}}$ for every $j = 1, 2, \ldots, b-1$. A directed graph with no directed cycles is known as a "directed acyclic graph."

If $\mathcal{G}$ is a directed acyclic graph, then $\tilde{p}(\mathbf{w}_\mathcal{S})$, as defined above, is a proper multivariate joint density (see online Appendix A1 or Lauritzen (1996) for a similar result). Starting from a joint multivariate density $p(\mathbf{w}_\mathcal{S})$, we derive a new density $\tilde{p}(\mathbf{w}_\mathcal{S})$ using a directed acyclic graph $\mathcal{G}$. While this holds for any original density $p(\mathbf{w}_\mathcal{S})$, it is especially useful in our context, where $p(\mathbf{w}_\mathcal{S})$

is a multivariate Gaussian density and $\mathcal{G}$ is sufficiently sparse. To be precise, let $\mathbf{C}_{N(\mathbf{s}_i)}$ be the covariance matrix of $\mathbf{w}_{N(\mathbf{s}_i)}$ and let $\mathbf{C}_{\mathbf{s}_i,N(\mathbf{s}_i)}$ be the $q \times mq$ cross-covariance matrix between the random vectors $\mathbf{w}(\mathbf{s}_i)$ and $\mathbf{w}_{N(\mathbf{s}_i)}$. Standard distribution theory reveals

$$\tilde{p}(\mathbf{w}_{\mathcal{S}}) = \prod_{i=1}^{k} N(\mathbf{w}(\mathbf{s}_i) \mid \mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)}, \mathbf{F}_{\mathbf{s}_i}), \tag{3}$$

where $\mathbf{B}_{\mathbf{s}_i} = \mathbf{C}_{\mathbf{s}_i,N(\mathbf{s}_i)}\mathbf{C}^{-1}_{N(\mathbf{s}_i)}$ and $\mathbf{F}_{\mathbf{s}_i} = \mathbf{C}(\mathbf{s}_i,\mathbf{s}_i) - \mathbf{C}_{\mathbf{s}_i,N(\mathbf{s}_i)}\mathbf{C}^{-1}_{N(\mathbf{s}_i)}\mathbf{C}_{N(\mathbf{s}_i),\mathbf{s}_i}$. Appendix A2 (available online) shows that $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ in (3) is a multivariate Gaussian density with covariance matrix $\tilde{\mathbf{C}}_{\mathcal{S}}$, which, obviously, is different from $\mathbf{C}_{\mathcal{S}}$. Furthermore, if $N(\mathbf{s}_i)$ has at most m members for each $\mathbf{s}_i$ in $\mathcal{S}$, where $m \ll k$, then $\tilde{\mathbf{C}}^{-1}_{\mathcal{S}}$ is sparse with at most $km(m+1)q^2/2$ nonzero entries. Thus, for a very general class of neighboring sets, $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ defined in (2) is the joint density of a multivariate Gaussian distribution with a sparse precision matrix.

Turning to the neighbor sets, choosing $N(\mathbf{s}_i)$ to be any subset of $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{i-1}\}$ ensures an acyclic $\mathcal{G}$ and, hence, a valid probability density in (3). Several special cases exist in likelihood approximation contexts. For example, Vecchia (1988) and Stroud, Stein, and Lysen (2014) specified $N(\mathbf{s}_i)$ to be the m nearest neighbors of $\mathbf{s}_i$ among $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{i-1}$ with respect to Euclidean distance. Stein, Chi, and Welty (2004) considered nearest as well as farthest neighbors from $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{i-1}\}$. Gramacy and Apley (2014) offered greater flexibility in choosing $N(\mathbf{s}_i)$, but may require several approximations to be efficient.

All of the above choices depend upon an ordering of the locations. Spatial locations are not ordered naturally, so one imposes order by, for example, ordering on one of the coordinates. Of course, any other function of the coordinates can be used to impose order. However, the aforementioned authors have cogently demonstrated that the choice of the ordering has no discernible impact on the approximation of (1) by (3). Our own simulation experiments (see Appendix A5, available online) concur with these findings; inference based upon $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ is extremely robust to the ordering of the locations. This is not entirely surprising. Clearly, whatever order we choose in (1), $p(\mathbf{w}_{\mathcal{S}})$ produces the full joint density. Note that we reduce (1) to (2) based upon neighbor sets constructed with respect to the *specific* ordering in (1). A different ordering in (1) will produce a different set of neighbors for (2). Since $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ ultimately relies upon the information borrowed from the neighbors, its effectiveness is often determined by the number of neighbors we specify and *not* the specific ordering.

In the following section, we will extend the density $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ to a legitimate spatial process. We remark that our subsequent development holds true for any choice of $N(\mathbf{s}_i)$ that ensures an acyclic $\mathcal{G}$. In general, identifying a "best subset" of m locations for obtaining optimal predictions for $\mathbf{s}_i$ is a nonconvex optimization problem, which is difficult to implement and defeats our purpose of using smaller conditioning sets to ease computations. Nevertheless, we have found Vecchia's choice of m-nearest neighbors from $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{i-1}\}$ to be simple and to perform extremely well for a wide range of simulation experiments. In what ensues, this will be our choice for $N(\mathbf{s}_i)$ and the corresponding density $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ will be referred to as the "nearest neighbor" density of $\mathbf{w}_{\mathcal{S}}$.

2.2 Extension to a Gaussian Process

Let $\mathbf{u}$ be any location in $\mathcal{D}$ outside $\mathcal{S}$. Consistent with the definition of $N(\mathbf{s}_i)$, let $N(\mathbf{u})$ be the set of m-nearest neighbors of $\mathbf{u}$ in $\mathcal{S}$. Hence, for any finite set $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ such that $\mathcal{S} \cap \mathcal{U}$ is empty, we define the nearest neighbor density of $\mathbf{w}_{\mathcal{U}}$ conditional on $\mathbf{w}_{\mathcal{S}}$ as

$$\tilde{p}(\mathbf{w}_{\mathcal{U}} \mid \mathbf{w}_{\mathcal{S}}) = \prod_{i=1}^{r} p(\mathbf{w}(\mathbf{u}_i) \mid \mathbf{w}_{N(\mathbf{u}_i)}) \,. \tag{4}$$

This conditional density is akin to (2) except that all the neighbor sets are subsets of $\mathcal{S}$. This ensures a proper conditional density. Indeed (2) and (4) are sufficient to describe the joint density of *any* finite set over the domain $\mathcal{D}$. More precisely, if $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is *any* finite subset in $\mathcal{D}$, then, using (4) we obtain the density of $\mathbf{w}_{\mathcal{V}}$ as

$$\tilde{p}(\mathbf{w}_{\mathcal{V}}) = \int \tilde{p}(\mathbf{w}_{\mathcal{U}} \mid \mathbf{w}_{\mathcal{S}})\, \tilde{p}(\mathbf{w}_{\mathcal{S}}) \prod_{\{\mathbf{s}_i \in \mathcal{S}\setminus\mathcal{V}\}} d(\mathbf{w}(\mathbf{s}_i)) \quad \text{where } \mathcal{U} = \mathcal{V} \setminus \mathcal{S} \,. \tag{5}$$

If $\mathcal{U}$ is empty, then (4) implies that $\tilde{p}(\mathbf{w}_{\mathcal{U}} \mid \mathbf{w}_{\mathcal{S}}) = 1$ in (5). If $\mathcal{S} \setminus \mathcal{V}$ is empty, then the integration in (5) is not needed.

These probability densities, defined on finite topologies, conform to Kolmogorov's consistency criteria and, hence, correspond to a valid spatial process over $\mathcal{D}$ (see Appendix A3, available online). So, given any original (parent) spatial process and any *fixed* reference set $\mathcal{S}$, we can construct a new process over the domain $\mathcal{D}$ using a collection of neighbor sets in $\mathcal{S}$. We refer to this process as the "nearest neighbor process" derived from the original parent process. If the parent process is $\mathrm{GP}(\mathbf{0}, \mathbf{C}(\cdot,\cdot \mid \boldsymbol{\theta}))$, then

$$\tilde{p}(\mathbf{w}_{\mathcal{U}} \mid \mathbf{w}_{\mathcal{S}}) = \prod_{i=1}^{r} N(\mathbf{w}(\mathbf{u}_i) \mid \mathbf{B}_{\mathbf{u}_i}\mathbf{w}_{N(\mathbf{u}_i)}, \mathbf{F}_{\mathbf{u}_i}) = N(\mathbf{B}_{\mathcal{U}}\mathbf{w}_{\mathcal{S}}, \mathbf{F}_{\mathcal{U}}) \tag{6}$$

for any finite set $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ in $\mathcal{D}$ outside $\mathcal{S}$, where $\mathbf{B}_{\mathbf{u}_i}$ and $\mathbf{F}_{\mathbf{u}_i}$ are defined analogous to (3) based on the neighbor sets $N(\mathbf{u}_i)$, $\mathbf{F}_{\mathcal{U}} = \mathrm{diag}(\mathbf{F}_{\mathbf{u}_1}, \mathbf{F}_{\mathbf{u}_2}, \ldots, \mathbf{F}_{\mathbf{u}_r})$, and $\mathbf{B}_{\mathcal{U}}$ is a sparse $nq \times kq$ matrix with each row having at most mq nonzero entries (see Appendix A4, available online).

For any finite set $\mathcal{V}$ in $\mathcal{D}$, $\tilde{p}(\mathbf{w}_{\mathcal{V}})$ is the density of the realizations of a Gaussian process over $\mathcal{V}$ with cross-covariance function

$$\tilde{\mathbf{C}}(\mathbf{v}_1, \mathbf{v}_2; \boldsymbol{\theta}) = \begin{cases} \tilde{\mathbf{C}}_{\mathbf{s}_i,\mathbf{s}_j}\,, & \text{if } \mathbf{v}_1 = \mathbf{s}_i \text{ and } \mathbf{v}_2 = \mathbf{s}_j \text{ are both in } \mathcal{S}, \\ \mathbf{B}_{\mathbf{v}_1}\tilde{\mathbf{C}}_{N(\mathbf{v}_1),\mathbf{s}_j} & \text{if } \mathbf{v}_1 \notin \mathcal{S} \text{ and } \mathbf{v}_2 = \mathbf{s}_j \in \mathcal{S}, \\ \mathbf{B}_{\mathbf{v}_1}\tilde{\mathbf{C}}_{N(\mathbf{v}_1),N(\mathbf{v}_2)}\mathbf{B}'_{\mathbf{v}_2} + \delta_{(\mathbf{v}_1=\mathbf{v}_2)}\mathbf{F}_{\mathbf{v}_1}, & \text{if } \mathbf{v}_1 \text{ and } \mathbf{v}_2 \text{ are not in } \mathcal{S} \end{cases} \tag{7}$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are any two locations in $\mathcal{D}$, $\tilde{\mathbf{C}}_{A,B}$ denotes submatrices of $\tilde{\mathbf{C}}_{\mathcal{S}}$ indexed by the locations in the sets A and B, and $\delta_{(\mathbf{v}_1=\mathbf{v}_2)}$ is the Kronecker delta. Appendix A4 (available online) also shows that $\tilde{\mathbf{C}}(\mathbf{v}_1, \mathbf{v}_2 \mid \boldsymbol{\theta})$ is continuous for all pairs $(\mathbf{v}_1, \mathbf{v}_2)$ outside a set of Lebesgue measure zero.

This completes the construction of a well-defined *nearest neighbor Gaussian process*, $\mathrm{NNGP}(\mathbf{0}, \tilde{\mathbf{C}}(\cdot,\cdot \mid \boldsymbol{\theta}))$, derived from a *parent Gaussian process*, $\mathrm{GP}(\mathbf{0}, \mathbf{C}(\cdot,\cdot \mid \boldsymbol{\theta}))$. In the NNGP, the size of $\mathcal{S}$, that is, k, can be as large, or even larger than the size of the

dataset. The reduction in computational complexity is achieved through sparsity of the NNGP precision matrices. Unlike low-rank processes, the NNGP is *not* a degenerate process. It is a proper, sparsity-inducing Gaussian process, immediately available as a prior in hierarchical modeling, and, as we show in the next section, delivers massive computational benefits.

3. Bayesian Estimation and Implementation

3.1 A Hierarchical Model

Consider a vector of l dependent variables, say $\mathbf{y}(\mathbf{t})$, at location $\mathbf{t} \in \mathcal{D} \subseteq \Re^d$ in a spatially varying regression model,

$$\mathbf{y}(\mathbf{t}) = \mathbf{X}(\mathbf{t})'\boldsymbol{\beta} + \mathbf{Z}(\mathbf{t})'\mathbf{w}(\mathbf{t}) + \boldsymbol{\epsilon}(\mathbf{t}) \ , \tag{8}$$

where $\mathbf{X}(\mathbf{t})'$ is the $l \times p$ matrix of fixed spatially referenced predictors, $\mathbf{w}(\mathbf{t})$ is a $q \times 1$ spatial process forming the coefficients of the $l \times q$ fixed design matrix $\mathbf{Z}(\mathbf{t})'$, and $\boldsymbol{\epsilon}(\mathbf{t}) \overset{iid}{\sim} N(\mathbf{0}, \mathbf{D})$ is an $l \times 1$ white-noise process capturing measurement error or micro-scale variability with dispersion matrix $\mathbf{D}$, which we assume is diagonal with entries τ_j^2, $j = 1, 2, \ldots, l$. The matrix $\mathbf{X}(\mathbf{t})'$ is block diagonal with $p = \sum_{i=1}^{l} p_i$, where the $1 \times p_i$ vector $\mathbf{x}_i(\mathbf{t})'$, including perhaps an intercept, is the ith block for each $i = 1, 2, \ldots, l$. The model in (8) subsumes several specific spatial models. For instance, letting $q = l$ and $\mathbf{Z}(\mathbf{t})' = \mathbf{I}_{l\times l}$ leads to a multivariate spatial regression model, where $\mathbf{w}(\mathbf{t})$ acts as a *spatially varying intercept*. On the other hand, we could envision all coefficients to be spatially varying and set $q = p$ with $\mathbf{Z}(\mathbf{t})' = \mathbf{X}(\mathbf{t})'$.

For scalability, instead of a customary Gaussian process prior for $\mathbf{w}(t)$ in (8), we assume $\mathbf{w}(\mathbf{t}) \sim NNGP(\mathbf{0}, \tilde{\mathbf{C}}(\cdot, \cdot \mid \boldsymbol{\theta}))$ derived from the parent $GP(\mathbf{0}, \mathbf{C}(\cdot, \cdot \mid \boldsymbol{\theta}))$. Any valid isotropic cross-covariance function (see, e.g., Gelfand and Banerjee 2010) can be used to construct $\mathbf{C}(\cdot, \cdot \mid \boldsymbol{\theta})$. To elucidate, let $\mathcal{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n\}$ be the set of locations where the outcomes and predictors have been observed. This set may, but need not, intersect with the reference set $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$ for the NNGP. Without loss of generality, we split up $\mathcal{T}$ into $\mathcal{S}^*$ and $\mathcal{U}$, where $\mathcal{S}^* = \mathcal{S} \cap \mathcal{T} = \{\mathbf{s}_{i_1}, \mathbf{s}_{i_2}, \ldots, \mathbf{s}_{i_r}\}$ with $\mathbf{s}_{i_j} = \mathbf{t}_j$ for $j = 1, 2, \ldots, r$ and $\mathcal{U} = \mathcal{T} \setminus \mathcal{S} = \{\mathbf{t}_{r+1}, \mathbf{t}_{r+2}, \ldots, \mathbf{t}_n\}$. Since $\mathcal{S} \cup \mathcal{T} = \mathcal{S} \cup \mathcal{U}$, we can completely specify the realizations of the NNGP in terms of the realizations of the parent process over $\mathcal{S}$ and $\mathcal{U}$, hierarchically, as $\mathbf{w}_{\mathcal{U}} \mid \mathbf{w}_{\mathcal{S}} \sim N(\mathbf{B}_{\mathcal{U}}\mathbf{w}_{\mathcal{S}}, \mathbf{F}_{\mathcal{U}})$ and $\mathbf{w}_{\mathcal{S}} \sim N(\mathbf{0}, \tilde{\mathbf{C}}_{\mathcal{S}})$. For a full Bayesian specification, we further specify prior distributions on $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and the τ_j^2's. For example, with customary prior specifications, we obtain the joint distribution

$$p(\boldsymbol{\theta}) \times \prod_{j=1}^{l} IG(\tau_j^2 \mid a_{\tau_j}, b_{\tau_j}) \times N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \mathbf{V}_\beta) \times N(\mathbf{w}_{\mathcal{U}} \mid \mathbf{B}_{\mathcal{U}}\mathbf{w}_{\mathcal{S}}, \mathbf{F}_{\mathcal{U}})$$
$$\times N(\mathbf{w}_{\mathcal{S}} \mid \mathbf{0}, \tilde{\mathbf{C}}_{\mathcal{S}}) \times \prod_{i=1}^{n} N(\mathbf{y}(\mathbf{t}_i) \mid \mathbf{X}(\mathbf{t}_i)'\boldsymbol{\beta} + \mathbf{Z}(\mathbf{t}_i)'\mathbf{w}(\mathbf{t}_i), \mathbf{D}), \tag{9}$$

where $p(\boldsymbol{\theta})$ is the prior on $\boldsymbol{\theta}$ and $IG(\tau_j^2 \mid a_{\tau_j}, b_{\tau_j})$ denotes the inverse Gamma density.

3.2 Estimation and Prediction

To describe a Gibbs sampler for estimating (9), we define $\mathbf{y} = (\mathbf{y}(\mathbf{t}_1)', \mathbf{y}(\mathbf{t}_2)', \ldots, \mathbf{y}(\mathbf{t}_n)')'$, and $\mathbf{w}$ and $\boldsymbol{\epsilon}$ similarly. Also, we introduce $\mathbf{X} = [\mathbf{X}(\mathbf{t}_1) : \mathbf{X}(\mathbf{t}_2) : \ldots : \mathbf{X}(\mathbf{t}_n)]'$, $\mathbf{Z} = \text{diag}(\mathbf{Z}(\mathbf{t}_1)', \ldots, \mathbf{Z}(\mathbf{t}_n)')$, and $\mathbf{D}_n = \text{Cov}(\boldsymbol{\epsilon}) = \text{diag}(\mathbf{D}, \ldots, \mathbf{D})$. The full conditional distribution for $\boldsymbol{\beta}$ is $N(\mathbf{V}_\beta^*\boldsymbol{\mu}_\beta^*, \mathbf{V}_\beta^*)$, where $\mathbf{V}_\beta^* = (\mathbf{V}_\beta^{-1} + \mathbf{X}'\mathbf{D}_n^{-1}\mathbf{X})^{-1}$, $\boldsymbol{\mu}_\beta^* = (\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \mathbf{X}'\mathbf{D}_n^{-1}(\mathbf{y} - \mathbf{Z}\mathbf{w}))$. Inverse Gamma priors for the τ_j^2's leads to conjugate full conditional distribution $\text{IG}(a_{\tau_j} + \frac{n}{2}, b_{\tau_j} + \frac{1}{2}(\mathbf{y}_{*j} - \mathbf{X}_{*j}\boldsymbol{\beta} - \mathbf{Z}_{*j}\mathbf{w})'(\mathbf{y}_{*j} - \mathbf{X}_{*j}\boldsymbol{\beta} - \mathbf{Z}_{*j}\mathbf{w})$, where $\mathbf{y}_{*j}$ refers to the $n \times 1$ vector containing the jth coordinates of the $\mathbf{y}(\mathbf{t}_i)$'s, and $\mathbf{X}_{*j}$ and $\mathbf{Z}_{*j}$ are the corresponding fixed and spatial effect covariate matrices, respectively. For updating $\boldsymbol{\theta}$, we use a random walk Metropolis step with target density $p(\boldsymbol{\theta}) \times N(\mathbf{w}_{\mathcal{S}} \mid \mathbf{0}, \tilde{\mathbf{C}}_{\mathcal{S}}) \times N(\mathbf{w}_{\mathcal{U}} \mid \mathbf{B}_{\mathcal{U}}\mathbf{w}_{\mathcal{S}}, \mathbf{F}_{\mathcal{U}})$, where

$$\begin{aligned} N(\mathbf{w}_{\mathcal{S}} \mid \mathbf{0}, \tilde{\mathbf{C}}_{\mathcal{S}}) &= \textstyle\prod_{i=1}^{k} N(\mathbf{w}(\mathbf{s}_i) \mid \mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)}, \mathbf{F}_{\mathbf{s}_i}) \text{ and} \\ N(\mathbf{w}_{\mathcal{U}} \mid \mathbf{B}_{\mathcal{U}}\mathbf{w}_{\mathcal{S}}, \mathbf{F}_{\mathcal{U}}) &= \textstyle\prod_{i=r+1}^{n} N(\mathbf{w}(\mathbf{t}_i) \mid \mathbf{B}_{\mathbf{t}_i}\mathbf{w}_{N(\mathbf{t}_i)}, \mathbf{F}_{\mathbf{t}_i}). \end{aligned} \tag{10}$$

Each of the component densities under the product sign on the right-hand side of (10) can be evaluated without any n-dimensional matrix operations.

Since the components of $\mathbf{w}_{\mathcal{U}} \mid \mathbf{w}_{\mathcal{S}}$ are independent, we can update $\mathbf{w}(\mathbf{t}_i)$ from its full conditional $N(\mathbf{V}_{\mathbf{t}_i}\boldsymbol{\mu}_{\mathbf{t}_i}, \mathbf{V}_{\mathbf{t}_i})$ for $i = r+1, r+2, \ldots, n$ where $\mathbf{V}_{\mathbf{t}_i} = (\mathbf{Z}(\mathbf{t}_i)\mathbf{D}^{-1}\mathbf{Z}(\mathbf{t}_i)' + \mathbf{F}_{\mathbf{t}_i}^{-1})^{-1}$ and $\boldsymbol{\mu}_{\mathbf{t}_i} = \mathbf{Z}(\mathbf{t}_i)\mathbf{D}^{-1}(\mathbf{y}(\mathbf{t}_i) - \mathbf{X}(\mathbf{t}_i)'\boldsymbol{\beta}) + \mathbf{F}_{\mathbf{t}_i}^{-1}\mathbf{B}_{\mathbf{t}_i}\mathbf{w}_{N(\mathbf{t}_i)}$. Finally, we update the components of $\mathbf{w}_{\mathcal{S}}$ individually. For any two locations $\mathbf{s}$ and $\mathbf{t}$ in $\mathcal{D}$, if $\mathbf{s} \in N(\mathbf{t})$ and is the lth component of $N(\mathbf{t})$, that is, say $\mathbf{s} = N(\mathbf{t})(l)$, then define $\mathbf{B}_{\mathbf{t},\mathbf{s}}$ as the $l \times l$ submatrix formed by columns $(l-1)q+1, (l-1)q+2, \ldots, lq$ of $\mathbf{B}_{\mathbf{t}}$. Let $U(\mathbf{s}_i) = \{\mathbf{t} \in \mathcal{S} \cup \mathcal{T} \mid \mathbf{s}_i \in N(\mathbf{t})\}$ and for every $\mathbf{t} \in U(\mathbf{s}_i)$ define, $\mathbf{a}_{\mathbf{t},\mathbf{s}_i} = \mathbf{w}(\mathbf{t}) - \sum_{\mathbf{s}\in N(\mathbf{t}),\mathbf{s}\neq\mathbf{s}_i} \mathbf{B}_{\mathbf{t},\mathbf{s}}\mathbf{w}(\mathbf{s})$. Then, for $i = 1, 2, \ldots, k$, we have the full conditional $\mathbf{w}_{\mathbf{s}_i} \mid \cdot \sim N(\mathbf{V}_{\mathbf{s}_i}\boldsymbol{\mu}_{\mathbf{s}_i}, \mathbf{V}_{\mathbf{s}_i})$, where $\mathbf{V}_{\mathbf{s}_i} = (In(\mathbf{s}_i \in \mathcal{S}^*)\mathbf{Z}(\mathbf{s}_i)\mathbf{D}^{-1}\mathbf{Z}(\mathbf{s}_i)' + \mathbf{F}_{\mathbf{s}_i}^{-1} + \sum_{\mathbf{t}\in U(\mathbf{s}_i)} \mathbf{B}'_{\mathbf{t},\mathbf{s}_i}\mathbf{F}_{\mathbf{t}}^{-1}\mathbf{B}_{\mathbf{t},\mathbf{s}_i})^{-1}$, $\boldsymbol{\mu}_{\mathbf{s}_i} = In(\mathbf{s}_i \in \mathcal{S}^*)\mathbf{Z}(\mathbf{s}_i)\mathbf{D}^{-1}(\mathbf{y}(\mathbf{s}_i) - \mathbf{X}(\mathbf{s}_i)'\boldsymbol{\beta}) + \mathbf{F}_{\mathbf{s}_i}^{-1}\mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)} + \sum_{\mathbf{t}\in U(\mathbf{s}_i)} \mathbf{B}'_{\mathbf{t},\mathbf{s}_i}\mathbf{F}_{\mathbf{t}}^{-1}\mathbf{a}_{\mathbf{t},\mathbf{s}_i}$, and $In(\cdot)$ denotes the indicator function. Hence, the $\mathbf{w}$'s can also be updated without requiring storage or factorization of any $n \times n$ matrices.

Turning to predictions, let $\mathbf{t}$ be a new location where we intend to predict $\mathbf{y}(\mathbf{t})$ given $\mathbf{X}(\mathbf{t})$ and $\mathbf{Z}(\mathbf{t})$. The Gibbs sampler for estimation also generates the posterior samples $\mathbf{w}_{\mathcal{S}\cup\mathcal{T}} \mid \mathbf{y}$. So, if $\mathbf{t} \in \mathcal{S} \cup \mathcal{T}$, then we simply get samples of $\mathbf{y}(\mathbf{t}) \mid \mathbf{y}$ from $N(\mathbf{X}(\mathbf{t})'\boldsymbol{\beta} + \mathbf{Z}(\mathbf{t})'\mathbf{w}(\mathbf{t}), \mathbf{D})$. If $\mathbf{t}$ is outside $\mathcal{S} \cup \mathcal{T}$, then we generate samples of $\mathbf{w}(\mathbf{t})$ from its full conditional $N(\mathbf{B}_{\mathbf{t}}\mathbf{w}_{\mathcal{S}}, \mathbf{F}_{\mathbf{t}})$ and subsequently generate posterior samples of $\mathbf{y}(\mathbf{t}) \mid \mathbf{y}$ similar to the earlier case.

3.3 Computational Complexity

Implementing the NNGP model in Section 3.2 reveals that one entire pass of the Gibbs sampler can be completed without any large matrix operations. The only difference between (9) and a full geostatistical hierarchical model is that the spatial process is modeled as an NNGP prior as opposed to a standard GP. For comparisons, we offer rough estimates of the flop counts to generate $\boldsymbol{\theta}$ and $\mathbf{w}$ per iteration of the sampler. We express the computational complexity only in terms of the sample size

n, size of the reference set k, and the size of the neighbor sets m as other dimensions are assumed to be small. For all locations, $\mathbf{t} \in \mathcal{S} \cup \mathcal{T}$, $\mathbf{B_t}$, and $\mathbf{F_t}$ can be calculated using $O(m^3)$ flops. So, from (10) it is easy to see that $p(\boldsymbol{\theta} \mid \cdot)$ can be calculated using $O((n+k)m^3)$ flops. All subsequent calculations to generate a set of posterior samples for $\mathbf{w}$ and $\boldsymbol{\theta}$ require around $O((n+k)m^2)$ flops.

So, the total flop counts is of the order $(n+k)m^3$ and is, therefore, linear in the total number of locations in $\mathcal{S} \cup \mathcal{T}$. This ensures scalability of the NNGP to large datasets. Compare this with a full GP model with a dense correlation matrix, which requires $O(n^3)$ flops for updating $\mathbf{w}$ in each iteration. Simulation results in Section 5.1 and online Appendix A6 indicate that NNGP models with usually very small values of m ($\approx$10) provide inference almost indistinguishable to full geostatistical models. Therefore, for large n, this linear flop count is drastically less and linearity with respect to k ensures a feasible implementation even for $k \approx n$.

This offers substantial scalability over low-rank models where the computational cost is quadratic in the number of "knots," limiting the size of the set of knots. Also, the full geostatistical model requires storage of the $n \times n$ distance matrix, which can potentially exhaust storage resources for large datasets. An NNGP model only requires the distance matrix between neighbors for every location, thereby storing $n+k$ small matrices, each of order $m \times m$.

3.4 Model Comparison and Choice of $\mathcal{S}$ and m

As elaborated in Section 2, given any parent Gaussian process and *any* fixed reference set of locations $\mathcal{S}$, we can construct a valid NNGP. The resulting finite dimensional likelihoods of the NNGP depend upon the choice of the reference set $\mathcal{S}$ and the size of each $N(\mathbf{s}_i)$, that is, m. Choosing the reference set is similar to selecting the knots for a predictive process. Unlike the number of "knots" in low-rank models, the size of $\mathcal{S}$ does not thwart computational scalability. Since the flop count in an NNGP model only increases linearly with the size of $\mathcal{S}$, the number of locations in $\mathcal{S}$ can be large, with more flexible choices for $\mathcal{S}$.

Points over a grid across the entire domain seem to be a plausible choice for $\mathcal{S}$. For example, we can construct a large $\mathcal{S}$ using a dense grid to improve performance without adversely affecting computational costs. Another, perhaps even simpler, option for large datasets is to simply fix $\mathcal{S} = \mathcal{T}$, the set of observed locations. This choice reduces computational costs even further by avoiding additional sampling of $\mathbf{w}_\mathcal{U}$ in the Gibbs sampler. Our empirical investigations (see Section 5.1) reveal that choosing $\mathcal{S} = \mathcal{T}$ delivers inference almost indistinguishable from choosing $\mathcal{S}$ to be a grid over the domain for large datasets.

Stein, Chi, and Welty (2004) and Eidsvik et al. (2014) proposed using a sandwich variance estimator for evaluating the inferential abilities of neighbor-based pseudo-likelihoods. Shaby (2012) developed a post-sampling sandwich variance adjustment for posterior credible intervals of the parameters for quasi-Bayesian approaches using pseudo-likelihoods. However, the asymptotic results used to obtain the sandwich variance estimators are based on assumptions that are hard to verify in spatial settings with irregularly placed data points. Moreover, we view the NNGP as an independent model for fitting the data and not as an approximation to the original GP. Hence, we refrain from such sandwich variance adjustments. Instead, we can simply use any standard model comparison metrics such as deviance information criterion (DIC; Spiegelhalter et al. 2002), GPD score (Gelfand and Ghosh 1998), or root mean squared prediction error/root mean square error of coefficient of variation (RMSPE/RMSECV; Yeniay and Goktas 2002) to compare the performance of NNGP and any other candidate model. The same model comparison metrics are also used for selecting m. However, as we illustrate later in Section 5.1, usually a small value of m between 10 and 15 produces performance at par with the full geostatistical model. While larger m may be beneficial for massive datasets, perhaps under a different design scheme, it will be much smaller than the number of knots in low-rank models for comparable inference (see Section 5.1).

4. Alternate NNGP Models and Algorithms

4.1 Block Update of $\mathbf{w}_\mathcal{S}$ Using Sparse Cholesky

The Gibbs' sampling algorithm detailed in Section 3.2 is extremely efficient for large datasets with linear flop counts per iteration. However, it can sometimes experience slow convergence issues due to sequential updating of the elements in $\mathbf{w}_\mathcal{S}$. An alternative to sequential updating is to perform block updates of $\mathbf{w}_\mathcal{S}$. We choose $\mathcal{S} = \mathcal{T}$ so that $\mathbf{s}_i = \mathbf{t}_i$ for all $i = 1, 2, \ldots, n$ and we denote $\mathbf{w}_\mathcal{S} = \mathbf{w}_\mathcal{T}$ by $\mathbf{w}$. Then,

$$\mathbf{w}|\cdot \sim N(\mathbf{V}_\mathcal{S}\mathbf{Z}'\mathbf{D}_n^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \mathbf{V}_\mathcal{S}) \,, \quad \text{where } \mathbf{V}_\mathcal{S} = (\mathbf{Z}'\mathbf{D}_n^{-1}\mathbf{Z} + \tilde{\mathbf{C}}_\mathcal{S}^{-1})^{-1} \,. \tag{11}$$

Recall that $\tilde{\mathbf{C}}_\mathcal{S}^{-1}$ is sparse. Since $\mathbf{Z}$ and $\mathbf{D}_n$ are block diagonal, $\mathbf{V}_\mathcal{S}^{-1}$ retains the sparsity of $\tilde{\mathbf{C}}_\mathcal{S}^{-1}$. So, a sparse Cholesky factorization of $\mathbf{V}_\mathcal{S}^{-1}$ will efficiently produce the Cholesky factors of $\mathbf{V}_\mathcal{S}$. This will facilitate block updating of $\mathbf{w}$ in the Gibbs sampler.

4.2 NNGP Models for the Response

Another possible approach involves NNGP models for the response $\mathbf{y}(\mathbf{s})$. If $\mathbf{w}(\mathbf{s})$ is a Gaussian process, then so is $\mathbf{y}(\mathbf{s}) = \mathbf{Z}(\mathbf{s})'\mathbf{w}(\mathbf{s}) + \boldsymbol{\epsilon}$ (without loss of generality we assume $\beta = \mathbf{0}$). One can directly use the NNGP specification for $\mathbf{y}(\mathbf{s})$ instead of $\mathbf{w}(\mathbf{s})$. That is, we derive $\mathbf{y}(\mathbf{s}) \sim \text{NNGP}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}(\cdot,\cdot))$ from the parent Gaussian process $GP(\mathbf{0}, \boldsymbol{\Sigma}(\cdot,\cdot \mid \boldsymbol{\theta}))$. The Gibbs sampler analogous to Section 3 now enjoys the additional advantage of avoiding full conditionals for $\mathbf{w}$. This results in a Bayesian analogue for Vecchia (1988) and Stein, Chi, and Welty (2004) but precludes inference on the spatial residual surface $\mathbf{w}(\mathbf{s})$. Modeling $\mathbf{w}(\mathbf{s})$ provides additional insight into residual spatial contours and is often important in identifying lurking covariates or eliciting unexplained spatial patterns. Vecchia (1992) used the nearest neighbor approximation on a spatial model for observations ($\mathbf{y}$) with independent measurement error (nuggets) in addition to the usual spatial component ($\mathbf{w}$). However, it may not be possible to recover $\mathbf{w}$ using this approach. For example, a univariate stationary process $\mathbf{y}(\mathbf{s})$ with a nugget effect can be decomposed as $\mathbf{y}(\mathbf{s}) = \mathbf{w}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s})$ (letting $\boldsymbol{\beta} = \mathbf{0}$) for some $\mathbf{w}(\mathbf{s}) \sim \text{GP}(\mathbf{0}, \mathbf{C}(\cdot,\cdot \mid \boldsymbol{\theta}))$ and white-noise process

$\boldsymbol{\epsilon}(\mathbf{s})$. If $\mathbf{y} = \mathbf{w} + \boldsymbol{\epsilon}$, where $\mathbf{w} \sim N(\mathbf{0}, \mathbf{C})$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^2\mathbf{I}_n)$, then $\text{cov}(\mathbf{y}) = \mathbf{C} + \tau^2\mathbf{I} = \boldsymbol{\Sigma}$, all eigenvalues of $\boldsymbol{\Sigma}$ are greater than τ^2, and $\text{cov}(\mathbf{w} \mid \mathbf{y}) = \tau^2\mathbf{I}_n - \tau^4\boldsymbol{\Sigma}^{-1}$. For $\mathbf{y}(\mathbf{s}) \sim \text{NNGP}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}(\cdot,\cdot))$, however, the eigenvalues of $\tilde{\boldsymbol{\Sigma}}$ may be less than τ^2, so $\tau^2\mathbf{I}_n - \tau^4\tilde{\boldsymbol{\Sigma}}^{-1}$ need not be positive definite for every $\tau^2 > 0$ and $p(\mathbf{w} \mid \mathbf{y})$ is no longer well defined.

A different model is obtained by using an NNGP prior for $\mathbf{w}$, as in (9), and then integrating out $\mathbf{w}$. The resulting likelihood is $N(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_y)$, where $\boldsymbol{\Sigma}_y = \mathbf{Z}\tilde{\mathbf{C}}_{\mathcal{S}}\mathbf{Z}' + \mathbf{D}_n$ and the Bayesian specification is completed using priors on $\boldsymbol{\beta}$, τ_j^2's, and $\boldsymbol{\theta}$ as in (9). This model drastically reduces the number of variables in the Gibbs sampler, while preserving the nugget effect in the parent model. We can generate the full conditionals for the parameters in the marginalized model as follows: $\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\phi} \sim N((\mathbf{V}_\beta^{-1} + \mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\mathbf{X})^{-1}(\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\mathbf{y}), (\mathbf{V}_\beta^{-1} + \mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\mathbf{X})^{-1})$. It is difficult to factor out τ_j^2's from $\boldsymbol{\Sigma}_y^{-1}$, so conjugacy is lost with respect to any standard prior. Metropolis block updates for $\boldsymbol{\theta}$ are feasible for any tractable prior $p(\boldsymbol{\theta})$. This involves computing $\mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\mathbf{X}$, $\mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\mathbf{y}$, and $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Since $\boldsymbol{\Sigma}_y^{-1} = \mathbf{D}_n^{-1} - \mathbf{D}_n^{-1}\mathbf{Z}(\tilde{\mathbf{C}}_{\mathcal{S}}^{-1} + \mathbf{Z}'\mathbf{D}_n^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_n^{-1} = \mathbf{D}_n^{-1} - \mathbf{D}_n^{-1}\mathbf{Z}\mathbf{V}_{\mathcal{S}}\mathbf{Z}'\mathbf{D}_n^{-1}$, where $\mathbf{V}_{\mathcal{S}}$ is given by (11), a sparse Cholesky factorization of $\mathbf{V}_{\mathcal{S}}^{-1}$ will be beneficial. We draw posterior samples for $\mathbf{w}$ from $p(\mathbf{w} \mid \mathbf{y}) = \int p(\mathbf{w} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \{\tau_j^2\}, \mathbf{y})p(\boldsymbol{\theta}, \boldsymbol{\beta}, \{\tau_j^2\} \mid \mathbf{y})$ using composition sampling—we draw $\mathbf{w}^{(g)}$ from $p(\mathbf{w} \mid \boldsymbol{\theta}^{(g)}, \boldsymbol{\beta}^{(g)}, \{\tau_j^{2(g)}\}, \mathbf{y})$ one-for-one for each sampled parameter.

Using block updates for $\mathbf{w}_{\mathcal{S}}$ in (9) and fitting the marginalized version of (9) both require an efficient sparse Cholesky solver for $\mathbf{V}_{\mathcal{S}}^{-1}$. Note that computational expenses for most sparse Cholesky algorithms depend on the precise nature of the sparse structure (mostly on the bandwidth) of $\tilde{\mathbf{C}}_{\mathcal{S}}^{-1}$ (see, e.g., Davis 2006). The number of flops required for Gibbs sampling and prediction in this marginalized model depends upon the sparse structure of $\tilde{\mathbf{C}}_{\mathcal{S}}^{-1}$ and may, sometimes, heavily exceed the linear usage achieved by the unmarginalized model with individual updates for $\mathbf{w}_i$. Therefore, a prudent choice of the precise fitting algorithms should be based on the sparsity structure of $\tilde{\mathbf{C}}_{\mathcal{S}}^{-1}$ for the given dataset.

4.3 Spatiotemporal and GLM Versions

In spatiotemporal settings where we seek spatial interpolation at discrete time points (e.g., weekly, monthly, or yearly data), we write the response (possibly vector-valued) as $\mathbf{y}_t(\mathbf{s})$ and the random effects as $\mathbf{w}_t(\mathbf{s})$. Desired inference includes spatial interpolation for each time point. Spatial dynamic models incorporating the NNGP are easily formulated as below:

$$\begin{aligned}
\mathbf{y}_t(\mathbf{s}) &= \mathbf{X}_t(\mathbf{s})'\boldsymbol{\beta}_t + \mathbf{u}_t(s) + \boldsymbol{\epsilon}_t(\mathbf{s}),\ \boldsymbol{\epsilon}_t(\mathbf{s}) \overset{iid}{\sim} N(0, D)\\
\boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t,\ \boldsymbol{\eta}_t \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}_\eta),\ \boldsymbol{\beta}_0 \sim N(\mathbf{m}_0, \boldsymbol{\Sigma}_0)\\
\mathbf{u}_t(\mathbf{s}) &= \mathbf{u}_{t-1}(\mathbf{s}) + \mathbf{w}_t(\mathbf{s}),\ \mathbf{w}_t(\mathbf{s}) \overset{ind}{\sim} NNGP(\mathbf{0}, \tilde{\mathbf{C}}(\cdot,\cdot \mid \boldsymbol{\theta}_t)) \ . \quad (12)
\end{aligned}$$

Thus, one retains exactly the same structure of process-based spatial dynamic models, for example, as in Gelfand, Banerjee, and Gamerman (2005), and simply replaces the independent Gaussian process priors for $\mathbf{w}_t(\mathbf{s})$ with independent NNGPs to achieve computational tractability.

The above is illustrative of how attractive and extremely convenient the NNGP is for model building. One simply writes down the parent model and subsequently replaces the full GP with an NNGP. Being a well-defined process, the NNGP ensures a valid spatial dynamic model. Similarly NNGP versions of dynamic spatiotemporal Kalman-filtering (as, e.g., in Wikle and Cressie 1999) can be constructed.

Handling non-Gaussian (e.g., binary or count) data is also straightforward using spatial generalized linear models (GLMs; Diggle, Tawn, and Moyeed 1998; Lin et al. 2000; Kammann and Wand 2003; Banerjee, Carlin, and Gelfand 2014). Here, the NNGP provides structured dependence for random effects at the second stage. First, we replace $\text{E}[\mathbf{y}(\mathbf{t})]$ in (8) with $g(E(\mathbf{y}(\mathbf{t})))$, where $g(\cdot)$ is a suitable link function such that $\boldsymbol{\eta}(\mathbf{t}) = g(E(\mathbf{y}(\mathbf{t}))) = \mathbf{X}(\mathbf{t})'\boldsymbol{\beta} + \mathbf{Z}(\mathbf{t})'\mathbf{w}(\mathbf{t})$. In the second stage, we model the $\mathbf{w}(\mathbf{t})$ as an NNGP. The benefits of the algorithms in Sections 3.2 and 3.3 still hold, but some of the alternative algorithms in Section 4 may not apply. For example, we do obtain tractable marginalized likelihoods by integrating out the spatial effects.

5. Illustrations

We conduct simulation experiments and analyze a large forestry dataset. Additional simulation experiments are detailed in Appendices A5 through A9 (available online). Posterior inference for subsequent analysis were based upon three chains of 25,000 iterations (with a burn-in of 5000 iterations). All the samplers were programmed in C++ and leveraged Intel Math Kernel Library's (MKL) threaded BLAS and LAPACK routines for matrix computations on a Linux workstation with 384 GB of RAM and two Intel Nehalem quad-Xeon processors.

5.1 Simulation Experiment

We generated observations using 2500 locations within a unit square domain from the model (8) with $q = l = 1$ (univariate outcome), $p = 2$, $\mathbf{Z}(\mathbf{t})' = 1$ (scalar), the spatial covariance matrix $\mathbf{C}(\boldsymbol{\theta}) = \sigma^2\mathbf{R}(\boldsymbol{\phi})$, where $\mathbf{R}(\boldsymbol{\phi})$ is a $n \times n$ correlation matrix, and $\mathbf{D} = \tau^2$ (scalar). The model included an intercept and a covariate $\mathbf{x}_1$ drawn from $N(0, 1)$. The (i, j)th element of $\mathbf{R}(\boldsymbol{\phi})$ was calculated using the Matérn function

$$\rho(\mathbf{t}_i, \mathbf{t}_j; \boldsymbol{\phi}) = \frac{1}{2^{\nu-1}\Gamma(\nu)}(||\mathbf{t}_i - \mathbf{t}_j||\phi)^\nu \mathcal{K}_\nu(||\mathbf{t}_i - \mathbf{t}_j||\phi); \quad \phi > 0, \nu > 0, \quad (13)$$

where $||\mathbf{t}_i - \mathbf{t}_j||$ is the Euclidean distance between locations $\mathbf{t}_i$ and $\mathbf{t}_j$, $\boldsymbol{\phi} = (\phi, \nu)$ with ϕ controlling the decay in spatial correlation and ν controlling the process smoothness, Γ is the usual Gamma function, while $\mathcal{K}_\nu$ is a modified Bessel function of the second kind with order ν (Stein 1999). Evaluating the Gamma function for each matrix element within each iteration requires substantial computing time and can obscure differences in sampler run times; hence, we fixed ν at 0.5, which reduces (13) to the exponential correlation function. The first column in Table 1 gives the *true* values used to generate the responses. Figure 2(a) illustrates the $w(\mathbf{t})$ surface interpolated over the domain.

Table 1. Univariate synthetic data analysis parameter estimates and computing time in minutes for NNGP and full GP models. Parameter posterior summary 50 (2.5, 97.5) percentiles.

		NNGP ($\mathcal{S} \neq \mathcal{T}$)		NNGP ($\mathcal{S} = \mathcal{T}$)	
	True	$m = 10, k = 2000$	$m = 20, k = 2000$	$m = 10$	$m = 20$
β_0	1	0.99 (0.71, 1.48)	1.02 (0.73, 1.49)	1.00 (0.62, 1.31)	1.03 (0.65, 1.34)
β_1	5	5.00 (4.98, 5.03)	5.01 (4.98, 5.03)	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)
σ^2	1	1.09 (0.89, 1.49)	1.04 (0.85, 1.40)	0.96 (0.78, 1.23)	0.94 (0.77, 1.20)
τ^2	0.1	0.07 (0.04, 0.10)	0.07 (0.04, 0.10)	0.10 (0.08, 0.13)	0.10 (0.08, 0.13)
ϕ	12	11.81 (8.18, 15.02)	12.21 (8.83, 15.62)	12.93 (9.70, 16.77)	13.36 (9.99, 17.15)
p_D	—	1491.08	1478.61	1243.32	1249.57
DIC	—	1856.85	1901.57	2390.65	2377.51
G	—	33.67	35.68	77.84	76.40
P	—	253.03	259.13	340.40	337.88
D	—	286.70	294.82	418.24	414.28
RMSPE	—	1.22	1.22	1.2	1.2
95% CI cover %	—	97.2	97.2	97.6	97.6
95% CI width	—	2.19	2.18	2.13	2.12
Time	—	14.2	47.08	9.98	33.5
	True	Predictive process 64 knots	Full Gaussian process		
β_0	1	1.30 (0.54, 2.03)	1.03 (0.69, 1.34)		
β_1	5	5.03 (4.99, 5.06)	5.01 (4.99, 5.03)		
σ^2	1	1.29 (0.96, 2.00)	0.94 (0.76, 1.23)		
τ^2	0.1	0.08 (0.04, 0.13)	0.10 (0.08, 0.12)		
ϕ	12	**5.61 (3.48, 8.09)**	13.52 (9.92, 17.50)		
p_D	—	1258.27	1260.68		
DIC	—	13677.97	2364.80		
G	—	1075.63	74.80		
P	—	200.39	333.27		
D	—	1276.03	408.08		
RMSPE	—	1.68	1.2		
95% CI cover %	—	95.6	97.6		
95% CI width	—	2.97	2.12		
Time	—	43.36	560.31		

We then estimated the following models from the full data: (i) the full Gaussian process (*full GP*); (ii) the NNGP with $m = \{1, 2, \ldots, 25\}$ for $\mathcal{S} \neq \mathcal{T}$ and $\mathcal{S} = \mathcal{T}$; and (iii) a Gaussian predictive process (GPP) model (Banerjee et al. 2008) with 64 knots placed on a grid over the domain. For the NNGP with $\mathcal{S} \neq \mathcal{T}$, we considered 2000 randomly placed reference locations within the domain. The 64 knot GPP was chosen because its computing time was comparable to that of NNGP models. We used an efficient marginalized sampling algorithm for the Full GP and GPP models as implemented in the `spBayes` package in R (Finley, Banerjee, and Gelfand, in press). All the models were trained using 2000 of the 2500 observed locations, while the remaining 500 observations were withheld to assess predictive performance.

For all models, the intercept and slope regression parameters, β_0 and β_1, were given *flat* prior distributions. The variance components σ^2 and τ^2 were assigned inverse Gamma $IG(2, 1)$ and $IG(2, 0.1)$ priors, respectively, and the spatial decay ϕ received a uniform prior $U(3, 30)$, which corresponds to a spatial range between approximately 0.1 and 1 units.

Parameter estimates and performance metrics for the NNGP (with $m = 10$ and $m = 20$), GPP, and the full GP models are provided in Table 1. All model specifications produce similar posterior median and 95% credible intervals estimates, with the exception of ϕ in the 64 knot GPP model. Larger values of DIC and D suggest that the GPP model does not fit the data as well as the NNGP and full GP models. The NNGP $\mathcal{S} = \mathcal{T}$ models provide DIC, GPD scores that are comparable to those of the full GP model. These fit metrics suggest the NNGP $\mathcal{S} \neq \mathcal{T}$ models provide better fit to the data than that achieved by the full GP model, which is probably due to overfitting caused by a very large reference set $\mathcal{S}$. The last row in Table 1 shows computing times in minutes for one chain of 25,000 iterations reflecting on the enormous computational gains of NNGP models over full GP model.

Turning to out-of-sample predictions, the Full model's RMSPE and mean width between the upper and lower 95% posterior predictive credible interval is 1.2 and 2.12, respectively. As seen in Figure 1, comparable RMSPE and mean interval width for the NNGP $\mathcal{S} = \mathcal{T}$ model is achieved within $m \approx 10$. There is negligible difference between the predictive performances of the NNGP $\mathcal{S} \neq \mathcal{T}$ and $\mathcal{S} = \mathcal{T}$ models. Both the NNGP and full GP model have better predictive performance than the predictive process models when the number of knots is small, for example, 64. All models showed appropriate 95% credible interval coverage rates.

Figure 2(b)–2(f) illustrates the posterior median estimates of the spatial random effects from the Full GP, NNGP ($\mathcal{S} = \mathcal{T}$) with $m = 10$ and $m = 20$, NNGP ($\mathcal{S} \neq \mathcal{T}$) with $m = 10$, and GPP models. These surfaces can be compared to the *true* surface depicted in Figure 2(a). This comparison shows: (*i*) the NNGP models closely approximates the true surface and that estimated by the full GP model, and (*ii*) the reduced-rank predictive process model based on 64 knots greatly smooths over small-scale patterns. This last observation highlights one of the major criticisms of reduced-rank models (Stein 2014)

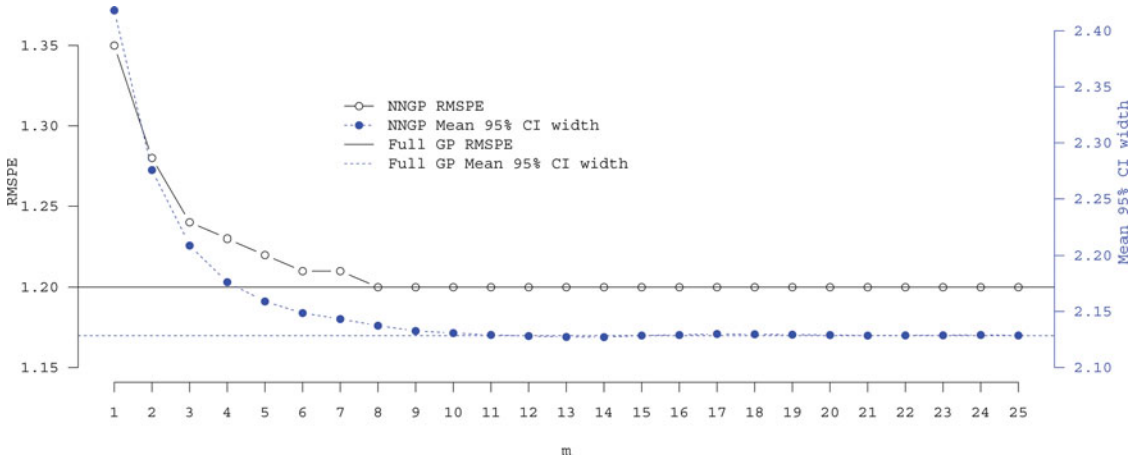


Figure 1. Choice of *m* in NNGP models: out-of-sample root mean squared prediction error (RMSPE) and mean width between the upper and lower 95% posterior predictive credible intervals for a range of *m* for the univariate synthetic data analysis.

and illustrates why these models often provide compromised predictive performance when the true surface has fine spatial resolution details. Overall, we see the clear computational advantage of the NNGP over the full GP model, and both inferential and computational advantage over the GPP model.

5.2 Forest Biomass Data Analysis

Information about the spatial distribution of forest biomass is needed to support global, regional, and local scale decisions, including assessment of current carbon stock and flux, bio-feedstock for emerging bio-economies, and impact of deforestation. In the United States, the Forest Inventory and Analysis (FIA) program of the USDA Forest Service collects the data needed to support these assessments. The program has established field plot centers in permanent locations using a sampling design that produces an equal probability sample (Bechtold and Patterson 2005). Field crews recorded stem measurements for all trees with diameter at breast height (DBH; 1.37 m above the forest floor) of 12.7 cm or greater. Given these data, established allometric equations were used to estimate each plot's forest biomass. For the subsequent analysis, plot biomass

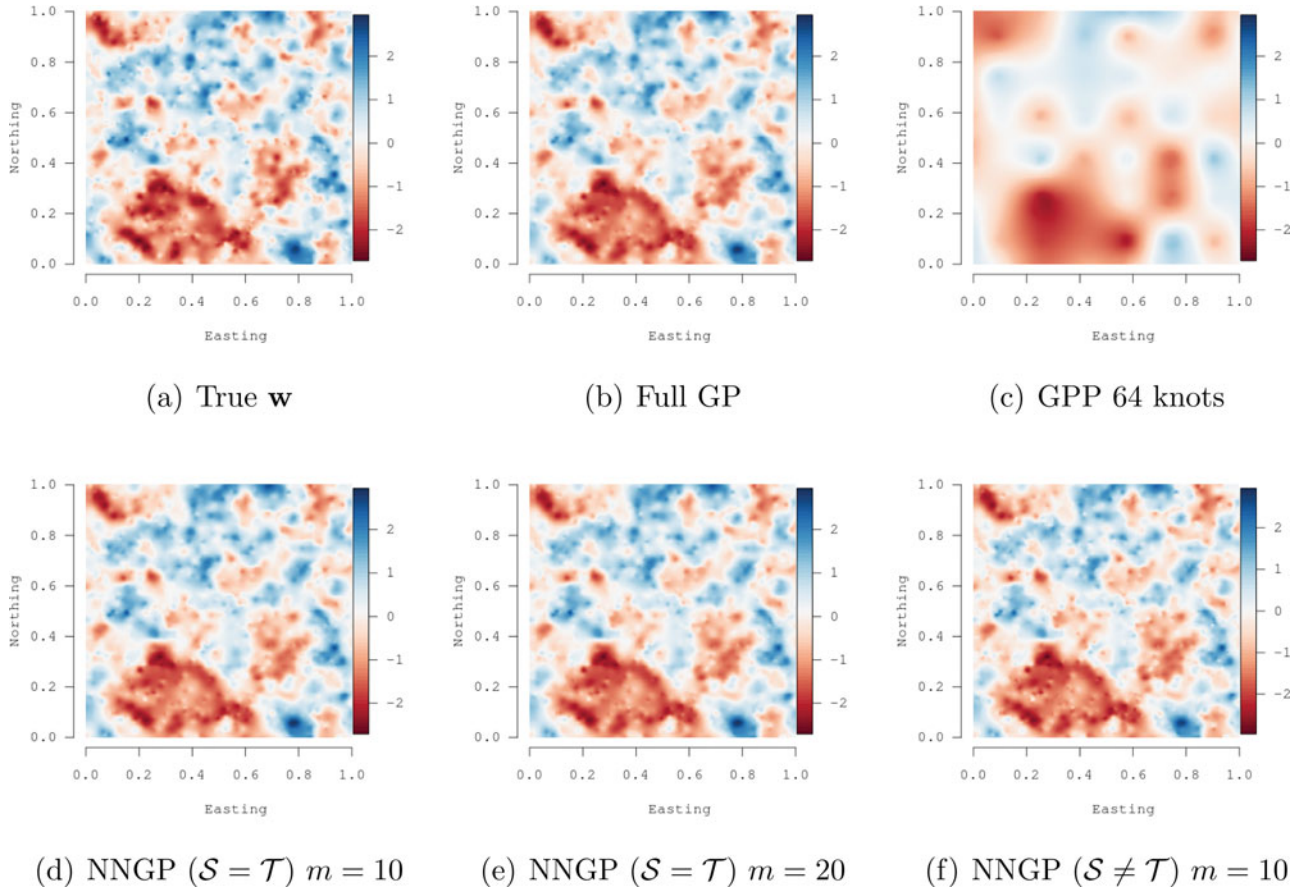


Figure 2. Univariate synthetic data analysis: interpolated surfaces of the true spatial random effects and posterior median estimates for different models.

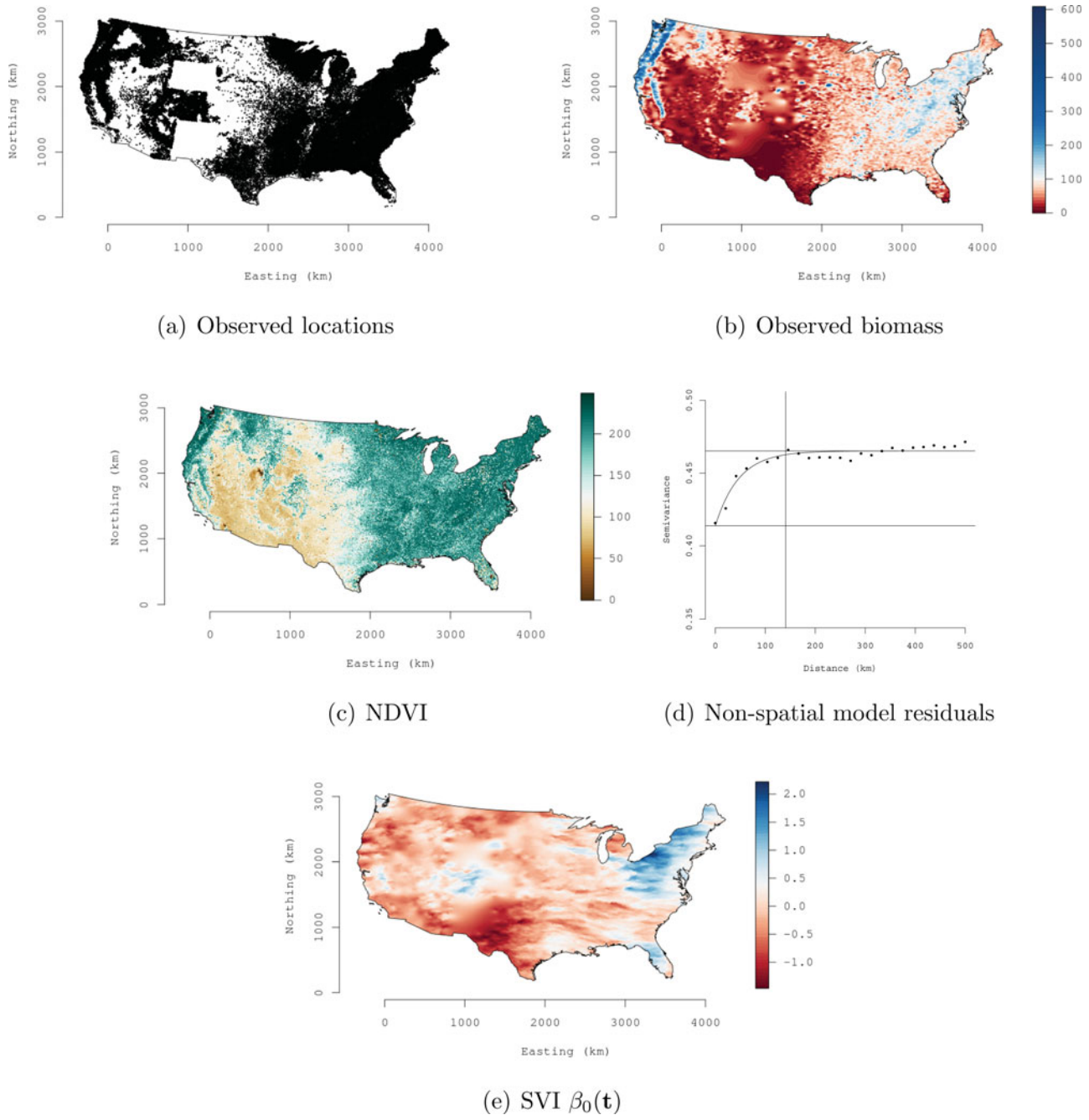


Figure 3. Forest biomass data analysis: (a) locations of observed biomass, (b) interpolated biomass response variable, (c) NDVI regression covariate, (d) variogram of non-spatial model residuals, and (e) surface of the SVI model random spatial effects posterior medians. Following our FIA data-sharing agreement, plot locations depicted in (a) have been "fuzzed" to hide the true coordinates.

was scaled to metric tons per ha then square root transformed. The transformation ensures that back transformation of subsequent predicted values have support greater than zero and helps to meet basic regression models assumptions.

Figure 3(a) illustrates the georeferenced forest inventory data consisting of 114, 371 forested FIA plots measured between 1999 and 2006 across the conterminous United States. The two blocks of missing observations in the Western and Southwestern United States correspond to Wyoming and New Mexico, which have not yet released FIA data. Figure 3(b) shows a deterministic interpolation of forest biomass observed on the FIA plots. Dark blue indicates high forest biomass, which is primarily seen in the Pacific Northwest, Western Coastal ranges, Eastern Appalachian Mountains, and in portions of New England. In contrast, dark red indicates regions where climate or land use limit vegetation growth.

A July 2006 Normalized Difference Vegetation Index (NDVI) image from the MODerate-resolution Imaging Spectroradiometer (MODIS; *http://glcf.umd.edu/data/ndvi*) sensor was used as a single predictor. NDVI is calculated from the visible and near-infrared light reflected by vegetation, and can be viewed as a measure of greenness. In this image, Figure 3(c), dark green corresponds to dense vegetation whereas brown identifies regions of sparse or no vegetation, for example, in the Southwest. NDVI is commonly used as a covariate in forest biomass regression models; see, for example, Zhang and Kondraguanta (2006).

Table 2. Forest biomass data analysis parameter estimates and computing time in hours for candidate models. Parameter posterior summary 50 (2.5, 97.5) percentiles.

	Nonspatial	NNGP Space-varying intercept	NNGP Space-varying coefficients
β_0	1.043 (1.02, 1.065)	1.44 (1.39, 1.48)	1.23 (1.20, 1.26)
β_{NDVI}	0.0093 (0.009, 0.0095)	0.0061 (0.0059, 0.0062)	0.0072 (0.0071, 0.0074)
σ^2	—	0.16 (0.15, 0.17)	—
$\mathbf{AA}'_{1,1}$	—	—	0.24 (0.23, 0.24)
$\mathbf{AA}'_{2,1}$	—	—	−0.00088 (−0.00093, −0.00083)
$\mathbf{AA}'_{2,2}$	—	—	0.0000052 (0.0000047, 0.0000056)
τ^2	0.52 (0.51, 0.52)	0.39 (0.39, 0.40)	0.39 (0.38, 0.40)
ϕ_1	—	0.016 (0.015, 0.016)	0.022 (0.021, 0.023)
ϕ_2	—	—	0.030 (0.029, 0.031)
ν_1	—	0.66 (0.64, 0.67)	0.92 (0.90, 0.93)
ν_2	—	—	0.92 (0.89, 0.93)
p_D	2.94	6526.95	4976.13
DIC	250137	224484.2	222845.1
G	59765.30	42551.08	43117.37
P	59667.15	47603.47	46946.49
D	119432.45	90154.55	90063.86
Time	—	14.53	41.35

Results from these and similar studies show a positive linear relationship between forest biomass and NDVI. The strength of this relationship, however, varies by forest tree species composition, age, canopy structure, and level of reflectance. We expect a space-varying relationship between biomass and NDVI, given tree species composition and disturbance regimes generally exhibit strong spatial dependence across forested landscapes.

The memory in our workstation was insufficient for storage of distance matrices required to fit a Full GP or GPP model. Subsequently, we explore the relationship between forest biomass and NDVI using a nonspatial model, an NNGP space-varying intercept (SVI) model (i.e., $q = l = 1$ and $\mathbf{Z}(\mathbf{t}) = 1$) in (8), and an NNGP spatially varying coefficients (SVC) regression model with $l = 1$, $q = p = 2$, and $\mathbf{Z}(\mathbf{t}) = \mathbf{X}(\mathbf{t})$ in (8). The reference sets for the NNGP models were again the observed locations and m was chosen to be 5 or 10. The parent process $\mathbf{w}(\mathbf{t})$ is a bivariate Gaussian process with an isotropic cross-covariance specification $\mathbf{C}(\mathbf{t}_i, \mathbf{t}_j \mid \boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\Gamma}(\boldsymbol{\phi})\mathbf{A}'$, where $\mathbf{A}$ is 2×2 lower-triangular with positive diagonal elements, $\boldsymbol{\Gamma}$ is 2×2 diagonal with $\rho(\mathbf{t}_i, \mathbf{t}_j; \boldsymbol{\phi}_b)$ (defined in (13)) as the bth diagonal entry, $b = 1, 2$, and $\boldsymbol{\phi}_b = (\phi_b, \nu_b)'$ (see, e.g., Gelfand and Banerjee 2010).

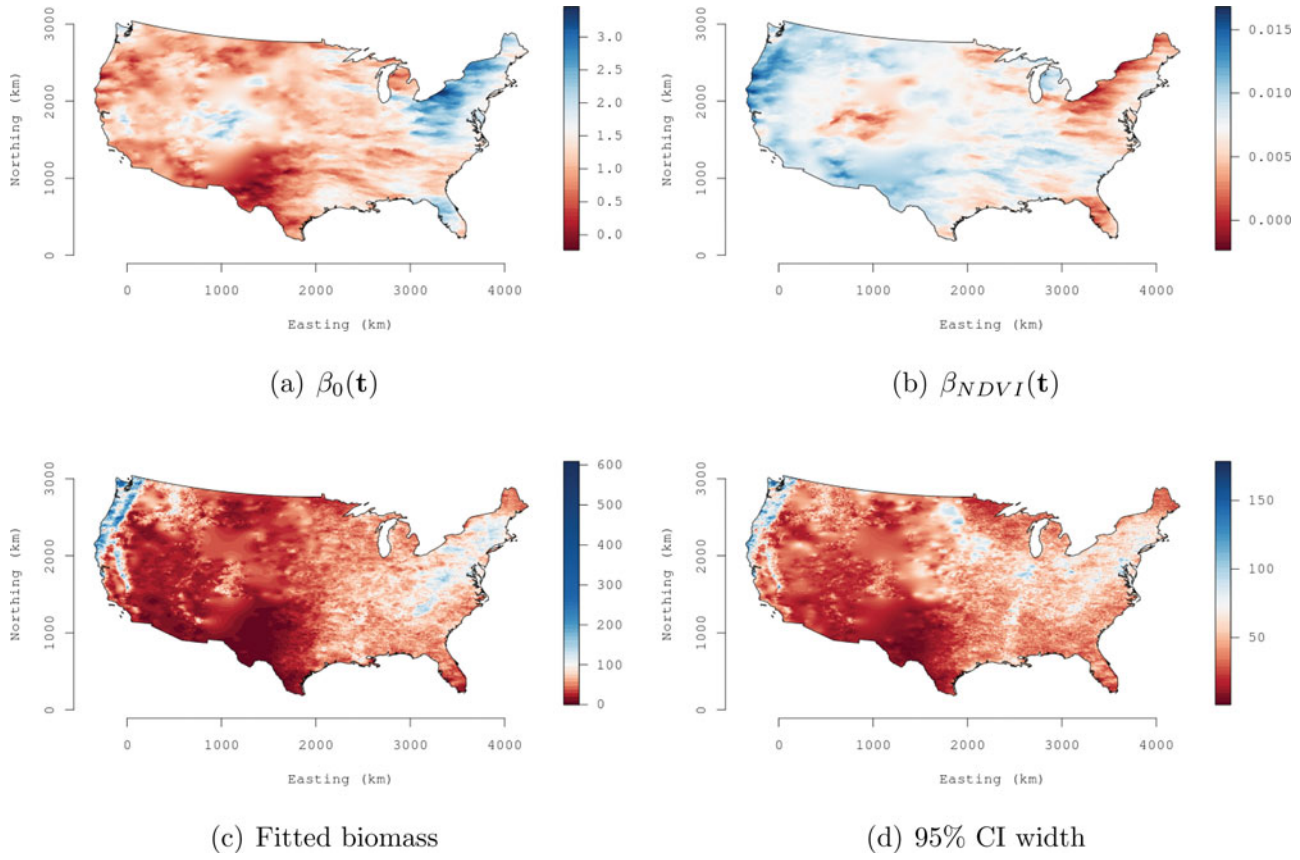
For all models, the intercept and slope regression parameters were given *flat* prior distributions. The variance components τ^2 and σ^2 were assigned inverse Gamma IG(2, 1) priors, the SVC model cross-covariance matrix $\mathbf{AA}'$ followed an inverse-Wishart IW(3, 0.1), and the Matérn spatial decay and smoothness parameters received uniform prior supports $U(0.01, 3)$ and $U(0.1, 2)$, respectively. These prior distributions on ϕ and ν correspond to support between approximately 0.5 and 537 km. Candidate models are assessed using the metrics described in Section 3.4, and inference drawn from mapped estimates of the regression coefficients and out-of-sample prediction.

Parameter estimates and performance metrics for NNGP with $m = 5$ are shown in Table 2. The corresponding numbers for $m = 10$ were similar. Relative to the spatial models, the nonspatial model has higher values of DIC and D, which suggests NDVI alone does not adequately capture the spatial structure of forest biomass. This observation is corroborated using a variogram fit to the nonspatial model's residuals; Figure 3(d). The variogram shows a nugget of ∼0.42, partial sill of ∼0.05, and range of ∼150 km. This residual spatial dependence is apparent when we map the SVI model spatial random effects as shown in Figure 3(e). This map, and the estimate of a nonnegligible spatial variance σ^2 in Table 2, suggests the addition of a spatial random effect was warranted and helps satisfy the model assumption of uncorrelated residuals.

The values of the SVC model's goodness-of-fit metrics suggest that allowing the NDVI regression coefficient to vary spatially improves model fit over that achieved by the SVI model. Figure 4(a) and 4(b) shows maps of posterior estimates for the spatially varying intercept and NDVI, respectively. The clear regional patterns seen in Figure 4(b) suggest the relationship between NDVI and biomass does vary spatially—with stronger positive regression coefficients in the Pacific Northwest and northern California areas. Forests in the Pacific Northwest and northern California are dominated by conifers and support the greatest range in biomass per unit area within the entire conterminous United States. The other strong regional pattern seen in Figure 4(b) is across western New England, where near zero regression coefficients suggest that NDVI is not as effective at discerning differences in forest biomass. This result is not surprising. For deciduous forests, NDVI can explain variability in low to moderate vegetation density. However, in high biomass deciduous forests, like those found across western New England, NDVI *saturates* and is no longer sensitive to changes in vegetation structure (Wang et al. 2005). Hence, we see a higher intercept in this region but lower slope coefficient on NDVI.

Figure 4(c) and 4(d) maps each location's posterior predictive median and the range between the upper and lower 95% credible interval, respectively, from the SVC model. Figure 4(c) shows strong correspondence with the deterministic interpolation of biomass in Figure 3(b). The prediction uncertainty in Figure 4(d) provides a realistic depiction of the model's ability to quantify forest biomass across the United States.

We also used prediction mean squared error (PMSE) to assess predictive performance. We fit the candidate models using 100,000 observations and withheld 14,371 for validation. PMSE for the nonspatial, SVI, and SVC models was 0.52, 0.41, and 0.42, respectively. Lower PMSE for the spatial models, versus the nonspatial model, corroborates the results from the model fit



(a) $\beta_0(\mathbf{t})$ (b) $\beta_{NDVI}(\mathbf{t})$

(c) Fitted biomass (d) 95% CI width

Figure 4. Forest biomass data analysis using SVC model: (1) posterior medians of the intercept, (b) NDVI regression coefficients, (c) median of biomass posterior predictive distribution, and (d) range between the upper and lower 95% percentiles of the posterior predictive distribution.

metrics and further supports the need for spatial random effects in the analysis.

6. Summary and Conclusions

We regard the NNGP as a highly scalable model, rather than a likelihood approximation, for large geostatistical datasets. It significantly outperforms competing low-rank processes such as the GPP, in terms of inferential performance and scalability. A reference set $\mathcal{S}$ and the resulting neighbor sets (of size m) define the NNGP. Larger m's would increase costs, but there is no apparent benefit to increasing m for larger datasets (see Appendix A6, available online). While some sensitivity to m and the choice of points in $\mathcal{S}$ is expected, our results indicate that inference is very robust with respect to $\mathcal{S}$ and very modest values of m (< 20) typically suffice. Larger reference sets may be needed for larger datasets, but its size does not thwart computations. In fact, the observed locations are a convenient choice for the reference set.

A potential concern with this choice is that if the observed locations have large gaps, then the resulting NNGP may be a poor approximation of the full Gaussian process. This arises from the fact that observations at locations outside the reference set are correlated via their respective neighbor sets and large gaps may imply two very near points have very different neighbor sets leading to low correlation. Our simulations in Appendix A7 (available online) indeed reveal that in such a situation, the NNGP covariance field is very flat at points in the gap. However, even with this choice of $\mathcal{S}$ the NNGP model performs at par with the full GP model as the latter also fails to provide strong information about observations located in large gaps. Of course, one can always choose a grid over the entire domain as $\mathcal{S}$ to construct an NNGP with covariance function similar to the full GP (see Figure A.5, available online). Another choice for $\mathcal{S}$ could be based upon configurations for treed Gaussian processes (Gramacy and Lee 2008).

Our simulation experiments revealed that estimation and kriging based on NNGP models closely emulate those from the true Matérn GP models, even for slow decaying covariances (see Appendix A8, available online). The Matérn covariance function is monotonically decreasing with distance and satisfies theoretical *screening* conditions, that is, the ability to predict accurately based on a few neighbors (Stein 2002). This, perhaps, explains the excellent performance of NNGP models with Matérn covariances. We also investigated the performance of NNGP models using a wave covariance function, which does not satisfy the screening conditions, in a setting where a significant proportion of nearest neighbors had negative correlation with the corresponding locations. The NNGP estimates were still close to the true model parameters and the kriged surface closely resembled the true surface (see Appendix A9, available online).

Most wave covariance functions (like the damped cosine or the cardinal sine function) produce covariance matrices with several small eigenvalues. The full GP model cannot be implemented for such models because the matrix inversion is numerically unstable. The NNGP model involves much smaller

matrix inversions and can be implemented in some cases (e.g., for the damped cosine model). However, for the cardinal sine covariance, the NNGP also faces numerical issues as even the small $m \times m$ covariance matrices are numerically unstable. Bias-adjusted low-rank GPs (Finley, Banerjee, and McRoberts 2009) possess a certain advantage in this aspect as the covariance matrix is guaranteed to have eigen values bounded away from zero, although stable computations will usually require full Cholesky decompositions.

Apart from being easily extensible to multivariate and spatiotemporal settings with discretized time, the NNGP can fuel interest in process-based modeling over graphs. Examples include networks, where data arising from nodes are posited to be similar to neighboring nodes. It also offers new modeling avenues and alternatives to the highly pervasive Markov random field models for analyzing regionally aggregated spatial data. Also, there is scope for innovation when space and time are jointly modeled as processes using spatiotemporal covariance functions. One will need to construct neighbor sets both in space and time and effective strategies, in terms of scalability and inference, will need to be explored. Comparisons with alternate approaches (see, e.g., Katzfuss and Cressie 2012) will also need to be made. Finally, a more comprehensive study on the alternate algorithms and parameterizations for faster Markov chain Monte Carlo convergence, including direct methods for executing sparse Cholesky factorizations (see Section 4), is being undertaken. More immediately, we plan to migrate our lower-level C++ code to the existing spBayes package (Finley, Banerjee, and Gelfand, in press) in the R statistical environment (*http://cran.r-project.org/web/packages/spBayes*) to facilitate wider user accessibility to NNGP models.

Supplementary Material

Supplementary material including detailed derivations of the properties of NNGP and several other simulation studies alluded to in this article are available in a separate file hosted on the journal website.

Acknowledgment

We thank the associate editor and anonymous reviewers for their suggestions. We also express our gratitude to Professors Michael Stein and Noel Cressie for discussions, which helped to enrich this work.

Funding

The work of the first three authors was partially supported by federal grants NSF/DMS 1106609 and NIH/NIGMS RC1-GM092400-01. The work of the second author was partially supported by NSF/DMS-1513654. The work of the third author was partially supported on by NSF grants EF-1137309, EF-1241874, EF-1253225, and DMS-1513481, as well as NASA Carbon Monitoring System grants, and the work of the fourth author was supported in part by NSF grant CM60934595.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data* (2nd ed), Boca Raton, FL: Chapman & Hall/CRC. [800,805]

Banerjee, S., Finley, A. O., Waldmann, P., and Ericcson, T. (2010), "Hierarchical Spatial Process Models for Multiple Traits in Large Genetic Trials," *Journal of the American Statistical Association*, 105, 506–521. [800]

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian Predictive Process Models for Large Spatial Datasets," *Journal of the Royal Statistical Society,* Series B, 70, 825–848. [800,801,806]

Bechtold, W. A., and Patterson, P. L. (2005), *The Enhanced Forest Inventory and Analysis National Sample Design and Estimation Procedures* (SRS-80), Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. [807]

Bevilacqua, M., and Gaetan, C. (2014), "Comparing Composite Likelihood Methods Based on Pairs for Spatial Gaussian Random Fields," *Statistics and Computing*, 25, 877–892. [800]

Crainiceanu, C. M., Diggle, P. J., and Rowlingson, B. (2008), "Bivariate Binomial Spatial Modeling of Loa Loa Prevalence in Tropical Africa," *Journal of the American Statistical Association*, 103, 21–37. [800]

Cressie, N. A.C., and Johannesson, G. (2008), "Fixed Rank Kriging for Very Large Data Sets," *Journal of the Royal Statistical Society,* Series B, 70, 209–226. [800]

Cressie, N. A.C., and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, Hoboken, NJ: Wiley. [800]

Davis, T. A. (2006), *Direct Methods for Sparse Linear Systems*, Philadelphia, PA: Society for Industrial and Applied Mathematics. [805]

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-Based Geostatistics" (with discussion), *Applied Statistics*, 47, 299–350. [805]

Du, J., Zhang, H., and Mandrekar, V. S. (2009), "Fixed-Domain Asymptotic Properties of Tapered Maximum Likelihood Estimators," *Annals of Statistics*, 37, 3330–3361. [800]

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014), "Estimation and Prediction in Spatial Models With Block Composite Likelihoods," *Journal of Computational and Graphical Statistics*, 23, 295–315. [800,804]

Emory, X. (2009), "The Kriging Update Equations and Their Application to the Selection of Neighboring Data," *Computational Geosciences*, 13, 269–280. [801]

Finley, A. O., Banerjee, S., and Gelfand, A. E. (2015), "spBayes for Large Univariate and Multivariate Point-Referenced Spatio-Temporal Data Models," *Journal of Statistical Software*, 63, 1–28. [806,811]

Finley, A. O., Banerjee, S., and McRoberts, R. E. (2009), "Hierarchical Spatial Models for Predicting Tree Species Assemblages Across Large Domains," *Annals of Applied Statistics*, 3, 1052–1079. [800,811]

Furrer, R., Genton, M. G., and Nychka, D. W. (2006), "Covariance Tapering for Interpolation of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 15, 503–523. [800]

Gelfand, A. E., and Banerjee, S. (2010), "Multivariate Spatial Process Models," in *Handbook of Spatial Statistics*, eds. A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, Boca Raton, FL: Chapman & Hall/CRC, pp. 495–516. [803,809]

Gelfand, A. E., Banerjee, S., and Gamerman, D. (2005), "Spatial Process Modelling for Univariate and Multivariate Dynamic Spatial Data," *Environmetrics*, 16, 465–479. [805]

Gelfand, A. E., and Ghosh, S. K. (1998), "Model Choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, 1–11. [804]

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003), "Spatial Modeling With Spatially Varying Coefficient Processes," *Journal of the American Statistical Association*, 98, 387–396. [801]

Gramacy, R. B., and Apley, D. W. (2014), "Local Gaussian Process Approximation for Large Computer Experiments." Available at *http://arxiv.org/abs/1303.0383*. [801,802]

Gramacy, R. B., and Lee, H. (2008), "Bayesian Treed Gaussian Process Models With an Application to Computer Experiments," *Journal of the American Statistical Association*, 103, 1119–1130. [810]

Gramacy, R. B., Niemi, J., and Weiss, R. M. (2014), "Massively Parallel Approximate Gaussian Process Regression." Available at *http://arxiv.org/abs/1310.5182*. [801]

Higdon, D. (2001), "Space and Space Time Modeling Using Process Convolutions," Technical Report, Institute of Statistics and Decision Sciences, Duke University, Durham. NC. [800]

Kammann, E. E., and Wand, M. P. (2003), "Geoadditive Models," *Applied Statistics*, 52, 1–18. [800,805]

Katzfuss, M., and Cressie, N. (2012), "Bayesian Hierarchical Spatio-Temporal Smoothing for Very Large Datasets," *Environmetrics*, 23, 94–107. [811]

Kaufman, C. G., Scheverish, M. J., and Nychka, D. W. (2008), "Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets," *Journal of the American Statistical Association*, 103, 1545–1555. [800]

Lauritzen, S. L. (1996), *Graphical Models*, Oxford, UK: Clarendon Press. [801]

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R., and Klein, B. (2000), "Smoothing Spline ANOVA Models for Large Data Sets With Bernoulli Observations and the Randomized GACV," *Annals of Statistics*, 28, 1570–1600. [805]

Moller, J., and Waagepetersen, R. P. (2003), *Statistical Inference and Simulation for Spatial Point Processes* (1st ed.), Boca Raton, FL: Chapman & Hall/CRC. [800]

Rasmussen, C. E., and Williams, C. K.I. (2005), *Gaussian Processes for Machine Learning* (1st ed), Cambridge, MA: The MIT Press. [800]

Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Boca Raton, FL: Chapman & Hall/CRC. [800]

Sang, H., and Huang, J. Z. (2012), "A Full Scale Approximation of Covariance Functions for Large Spatial Data Sets," *Journal of the Royal Statistical Society,* Series B, 74, 111–132. [800]

Schabenberger, O., and Gotway, C. A. (2004), *Statistical Methods for Spatial Data Analysis* (1st ed.), Boca Raton, FL: Chapman & Hall/CRC. [800]

Shaby, B. A. (2012), "The Open-Faced Sandwich Adjustment for MCMC Using Estimating Functions." Available at *http://arxiv.org/abs/1204.3687*. [804]

Shaby, B. A., and Ruppert, D. (2012), "Tapered Covariance: Bayesian Estimation and Asymptotics," *Journal of Computational and Graphical Statistics*, 21, 433–452. [800]

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society,* Series B, 64, 583–639. [804]

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging* (1st ed.), New York: Springer. [800,805]

——— (2002), "The Screening Effect in Kriging," *Annals of Statistics*, 30, 298–323. [810]

——— (2007), "Spatial Variation of Total Column Ozone on a Global Scale," *Annals of Applied Statistics*, 1, 191–210. [800]

——— (2008), "A Modeling Approach for Large Spatial Datasets," *Journal of the Korean Statistical Society*, 37, 3–10. [800]

——— (2014), "Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data," *Spatial Statistics*, 8, 1–19. [800,806]

Stein, M. L., Chi, Z., and Welty, L. J. (2004), "Approximating Likelihoods for Large Spatial Data Sets," *Journal of the Royal Statistical Society,* Series B, 66, 275–296. [800,801,802,804]

Stroud, J. R., Stein, M. L., and Lysen, S. (2014), "Bayesian and Maximum Likelihood Estimation for Gaussian Processes on an Incomplete Lattice." Available at *http://arxiv.org/abs/1402.4281*. [801,802]

Vecchia, A. V. (1988), "Estimation and Model Identification for Continuous Spatial Processes," *Journal of the Royal Statistical Society,* Series B, 50, 297–312. [800,801,802,804]

——— (1992), "A New Method of Prediction for Spatial Regression Models With Correlated Errors," *Journal of the Royal Statistical Society,* Series B, 54, 813–830. [800,804]

Wang, Q., Adiku, S., Tenhunen, J., and Granier, A. (2005), "On the Relationship of NDVI with Leaf Area Index in a Deciduous Forest Site," *Remote Sensing of Environment*, 94, 244–255. [809]

Wikle, C., and Cressie, N. A.C. (1999), "A Dimension-Reduced Approach to Space-Time Kalman Fltering," *Biometrika*, 86, 815–829. [805]

Yeniay, O., and Goktas, A. (2002), "A Comparison of Partial Least Squares Regression With Other Prediction Methods," *Hacettepe Journal of Mathematics and Statistics*, 31, 99–111. [804]

Zhang, X., and Kondraguanta, S. (2006), "Estimating Forest Biomass in the USA Using Generalized Allometric Models and MODIS Land Products," *Geophysical Research Letters*, 33, L09402. [808]