Meta-Kriging: Scalable Bayesian Modeling and Inference for Massive Spatial Datasets

Rajarshi Guhaniyogi and Sudipto Banerjee

June 13, 2017

Abstract

Spatial process models for analyzing geostatistical data entail computations that become prohibitive as the number of spatial locations becomes large. There is a burgeoning literature on approaches for analyzing large spatial datasets. In this article, we propose a divide-and-conquer strategy within the Bayesian paradigm. We partition the data into subsets, analyze each subset using a Bayesian spatial process model and then obtain approximate posterior inference for the entire dataset by optimally combining the individual posterior distributions from each subset. Importantly, as is often desired in spatial analysis, we offer full posterior predictive inference at arbitrary locations for the outcome as well as the residual spatial surface after accounting for spatially oriented predictors. We call this approach "Spatial Meta-Kriging" (SMK). We do not need to store the entire data in one processor, and this leads to superior scalability. We demonstrate SMK with various spatial regression models including Gaussian processes and tapered Gaussian processes. The approach is intuitive, easy to implement, and is supported by theoretical results presented in the Web Appendix. Empirical illustrations are provided using different simulation experiments and a geostatistical analysis of Pacific Ocean sea surface temperature data.

Key Words: Bayesian inference; Gaussian process models; low-rank models; M-posterior; posterior consistency; spatial process models; tapered Gaussian processes

1 Introduction

Increasing accessibility to computerized Geographical Information Systems (GIS) and related technologies have led to growing demands for analyzing massive spatially and temporally indexed databases on a variety of geographically-referenced outcomes. See, for example, the books by Gelfand et al. (2010), Cressie and Wikle (2015) and Banerjee et al. (2014) for a variety of methods for spatial data analysis. Gaussian processes are widely employed in spatial analysis, being especially attractive as a flexible and conveniently interpretable spatial interpolator acting as a stochastic surrogate for the underlying physical processes generating the data. Today, a primary challenge in geostatistics is the analysis of massive spatially-referenced data. This stems from the onerous Gaussian likelihood computations involving matrix factorizations (e.g., Cholesky) and determinant computations for large spatial covariance matrices that have no computationally exploitable structure. This is referred to as the "Big-N" problem in spatial statistics.

There is a burgeoning literature on the analysis of large spatial datasets which is too large to be comprehensively reviewed here. Briefly, these methods seek "dimension-reduction" by endowing the spatial covariance matrix either with a low-rank structure or with a sparse structure. Low-rank structures are usually derived from expressing the Gaussian process using basis functions such as fixed-rank kriging (Cressie and Johannesson (2008)), or predictive processes and variants thereof (e.g., Banerjee et al. (2008); Finley et al. (2009); Guhaniyogi et al. (2011); Sang and Huang (2012)) and multi-resolution approximations (e.g., Katzfuss (2016)). Wikle (2010) and Banerjee et al. (2014) provide more comprehensive reviews. Sparse structures intuit that spatial correlation between two distantly located observations is nearly zero, so little information is lost by assuming conditional independence given the intermediate locations. For example, covariance tapering (Furrer et al. (2012), Kaufman et al. (2008), Du et al. (2009), Shaby and Ruppert (2012)) uses compactly supported covariance functions to create sparse spatial covariance matrices that approximate the full covariance matrix. Alternately, one could introduce sparsity in the inverse covariance (precision) matrix using conditional independence assumptions or composite likelihoods (e.g., Vecchia (1988); Rue et al. (2009); Stein et al. (2004); Eidsvik et al. (2014); Datta et al. (2015); Stroud et al.

(2017); Guinness (2016)). In related literature pertaining to computer experiments, localized approximations of Gaussian process models are proposed, see for e.g. Gramacy and Apley (2015), Zhang et al. (2016) and Park and Apley (2017).

Some variants of dimension-reduction methods partition the spatial domain into subregions containing fewer spatial locations. Each of these subregions is modeled using Gaussian processes which are then hierarchically combined by borrowing information from across the subregions. Examples include non-stationary models (Banerjee et al. (2014)), multi-level and multi-resolution models (Gelfand et al. (2007); Nychka et al. (2015); Katzfuss (2016)) and the Bayesian Treed Gaussian Process models (Gramacy and Lee (2012)). These models usually achieve scalability by assuming block-independence at some level of the hierarchy, usually across subregions, but lose scalability when they borrow across sub-regions. Furthermore, the models and the inference are usually very sensitive to the specific partition adopted for the model.

Most existing methods for large spatial data are based upon approximations of the single Gaussian likelihood. Our current offering differs from these methods, and hence the aforementioned work, by focusing upon pooling posterior inference across a partition of data subsets. In some simple cases, for example with conjugate Bayesian linear regression models that we will revisit in a later section, one can exactly recover full posterior inference. However, such exact recovery is precluded for spatial and spatiotemporal process models and, more generally, for correlated data. Our objective is to develop a general approximation framework for obtaining the full posterior from posterior densities calculated over smaller subsets. The posteriors from various subsets (also known as "subset posteriors") are combined optimally to yield a single posterior distribution (the "meta-posterior") for the model parameters. Thus, we conduct a "meta-analysis" of the different datasets and also provide pooled posterior predictive inference for the spatial surface at arbitrary locations. We coin this as "spatial meta-kriging or SMK." To achieve this, we adapt the notion of a geometric median of a subset posterior (see, e.g., Minsker et al. (2014)). Unlike Minsker et al. (2014) who developed predictive models for independent data, we perform full Bayesian inference on each of the subsets using spatial process models. We obtain posterior samples for the process parameters and spatial random effects and derive the meta-posterior for the Bayesian

model. This approach can be used to considerably enhance the computational scalability of other Bayesian models for large spatial data. Once the post-burnin samples are stored for these models, sampling from the meta-posterior is extremely fast. For example, if it is feasible to estimate spatial process models to each subset of the data for n locations and one can run them on K subsets in parallel, then SMK will allow us to draw inference on nK locations. The values of n and K will depend upon the computational resources available and the model being fit to each dataset.

The manuscript follows this outline. In Section 2.1 we motivate the approach in conjugate non-spatial linear model. Our SMK approach will work with posterior samples from such models. Section 2.2 develops the framework for "Spatial Meta Kriging (SMK)" and discusses how to compute it. A detailed simulation study followed by a large data analysis is performed in Section 3 to justify usage of SMK for real data. Finally, Section 4 discusses what SMK achieves, and proposes a number of future directions to explore. Theoretical developments, including results on posterior consistency for the proposed SMK approach applied to Gaussian and tapered Gaussian process models are described in the Web Supplement.

2 Pooled Bayesian Inference

2.1 Conjugate Bayesian linear model

For some simple Bayesian models, one can exactly recover the posterior distributions of the parameters based upon quantities computed for subsets of the data. For example, consider the conjugate Bayesian Gaussian linear regression model

$$y = X\beta + \epsilon \; ; \quad \epsilon \sim N(0, \sigma^2 D) \; , \tag{1}$$

where y is an $N \times 1$ random vector of outcomes, X is a fixed $N \times p$ design matrix of explanatory variables, β is an unknown $p \times 1$ vector of slopes, D is a fixed $N \times N$ covariance matrix for y. This is extended to a Bayesian hierarchical model by assigning prior distributions $\beta \,|\, \sigma^2 \sim N(\mu_\beta, \sigma^2 V_\beta)$, and $\sigma^2 \sim IG(a, b)$. The joint posterior density $p(\beta, \sigma^2 \,|\, y)$ is available

in closed form as

$$p(\beta, \sigma^2 \,|\, y) = p(\sigma^2 \,|\, y) \times p(\beta \,|\, \sigma^2, y) \;, \tag{2}$$

where the marginal posterior density $p(\sigma^2 \,|\, y) = IG(\sigma^2 \,|\, a^*, b^*)$ and the conditional posterior density $p(\beta \,|\, \sigma^2, y) = N(\beta \,|\, Mm, \sigma^2 M)$ with $a^* = a + N/2$, $b^* = b + c/2$, $m = V_\beta^{-1}\mu_\beta + X^\top D^{-1} y$, $M^{-1} = V_\beta^{-1} + X^\top D^{-1} X$ and $c = \mu_\beta^\top V_\beta^{-1}\mu_\beta + y^\top D^{-1} y - m^\top M m$. Therefore, exact posterior inference can be carried out by first sampling σ^2 from $IG(a^*, b^*)$ and then sampling β from $N(Mm, \sigma^2 M)$ for each sampled value of σ^2. This results in samples from $p(\beta, \sigma^2 \,|\, y)$. Besides the fixed hyperparameters in the prior distributions, this exercise requires computing m, M and c.

Now consider a situation where N is large enough so that memory requirements for computing (1) is unfeasible. One possible resolution is to replace the likelihood in (1) with a composite likelihood that assumes independence across blocks formed by partitioning the data. We partition the $N \times 1$ vector y into K subvectors with y_k as the $n_k \times 1$ subvector forming the k-th subvector, where $\sum_{k=1}^K n_k = N$. Also, let X_k be the $n_k \times p$ matrix of predictors corresponding to y_k and let D_k be the marginal variance-covariance matrix for y_k. The conjugate Bayesian model with a block-independent composite likelihood assumes that

$$y_k = X_k\beta + \epsilon_k \;; \quad \epsilon_k \overset{ind}{\sim} N(0, \sigma^2 D_k) \;. \tag{3}$$

The Bayesian specification is completed by assiging priors to σ^2 and β as in (1). If we distribute the analysis to K different computing cores, where the k-th core fits the above model but only with the likelihood $N(y_k \,|\, X_k\beta, \sigma^2 D_k)$, then the quantities needed for sampling from the full $p(\beta, \sigma^2 \,|\, y)$ can be computed entirely using quantities obtained from the individual subsets of the data. For each $k = 1, 2, \ldots, K$ we independently compute $m_k = V_\beta^{-1}\mu_\beta + X_k^\top D_k^{-1} y_k$ and $M_k^{-1} = V_\beta^{-1} + X_k^\top D_k^{-1} X_k$ based upon the k-th subset of the data. We then combine them to obtain $m = \sum_{k=1}^K (m_k - (1-1/K)V_\beta^{-1}\mu_\beta)$ and $M^{-1} = \sum_{k=1}^K (M_k^{-1} - (1-1/K)V_\beta^{-1})$. Subsequently, we compute $c = \mu_\beta^\top V_\beta^{-1}\mu_\beta + \sum_{k=1}^K y_k^\top D_k^{-1} y_k - m^\top M m$. Therefore, sampling from the posterior distribution of β and σ^2 given the entire dataset can be achieved using quantities computed independently from each of the K smaller subsets of the

data. There is no need to interact between the subsets and one does not require to store or compute with large objects based upon the entire dataset.

This strategy will be efficient when the composite likelihood in (3) is a reasonable approximation for (1). In fact, for independent data modeled using D as an $N \times N$ identity matrix or a diagonal matrix, (1) and (3) are equivalent and the above method will exactly recover the inference from fitting the full model in (1) irrespective of how we partition the data. With correlated data, however, D is a non-diagonal correlation matrix and the analytical tractability above is lost. The composite likelihood in (3) is now only an approximation for (1) and will no longer be able to exactly recover the inference from (1). For finite sample inference, the performance of (3) may not be satisfactory and will depend upon a number of factors including how we partition the data. In the next section we discuss a computationally efficient algorithm to achieve accurate and robust inference by pooling posterior samples from the subsets of the data and subsequently apply this to spatially indexed data.

2.2 Pooled Bayesian inference for spatial models

Consider a customary spatial regression model given by

$$y(s) = x^{\top}(s)\beta + w(s) + \epsilon(s) \ , \tag{4}$$

where $x(s)$ is a $p \times 1$ vector of spatially referenced predictors, β is a $p \times 1$ vector of regression coefficients, $w(s)$ is a stochastic process capturing spatial dependence, while $\epsilon(s)$ captures variation at fine scales including those arising from measurement error. Customary specifications posit $w(s)$ is a zero-centered spatial Gaussian process with a covariance function $C_\theta(s, s')$ modeling $\text{cov}\{w(s), w(s')\}$ and $\epsilon(s)$ is a white-noise process independent of $w(s)$. Given a set of locations $\mathcal{S} = \{s_i : i = 1, 2, \ldots, N\}$ where $y(s)$ and $x(s)$ have been observed, the spatial regression in (4) is extended to a hierarchical linear mixed model framework

$$y = X\beta + w + \epsilon \ , \quad \epsilon \sim N(0, D(\theta)) \ , \tag{5}$$

where y, w and ϵ are $N \times 1$ vectors with elements $y(s_i)$, $w(s_i)$ and $\epsilon(s_i)$, respectively, X is the $N \times p$ matrix of regressors ($p < N$) with $x^{\top}(s_i)$ as its i-th row, $D(\theta)$ is an $N \times N$

covariance matrix corresponding to ϵ, $w \sim N(0, C(\theta))$, $C(\theta)$ is the $N \times N$ spatial covariance matrix with entries $C_\theta(s_i, s_j)$, $\beta \sim N(\mu_\beta, \Sigma_\beta)$ is the prior distribution for the slope vector, μ_β and Σ_β are assumed fixed, θ is a set of unknown parameters specifying the distributions for w and ϵ and is assigned a proper prior distribution $p(\theta)$. Note that here we do away with the conjugacy in Section 2.1, so Bayesian inference proceeds, customarily, by sampling $\Omega = \{\beta, \theta\}$ from (5) using Markov chain Monte Carlo (MCMC) methods (e.g., Robert and Casella (2009)).

Fitting the model in (5) entails matrix computations involving $C(\theta)$ and $D(\theta)$. While $D(\theta)$ is often specified as a diagonal (or sparse) matrix, e.g., $\tau^2 I$ which will arise by specifying $\epsilon(s) \overset{iid}{\sim} N(0, \tau^2)$, the spatial covariance matrix $C(\theta)$ is a dense $N \times N$ matrix. Irrespective of the specific parametrization or estimation algorithm, model fitting usually involves matrix decompositions for $C(\theta)$ requiring $\sim N^3$ floating point operations (flops) and $\sim N^2$ memory units in storage. These become prohibitive for large N since $C(\theta)$, in general, has no exploitable structure. Evidently, multivariate and spatial-temporal settings aggravate the situation.

Let the data be partitioned into $\{y_k, X_k\}$, for $k = 1, 2, \ldots, K$, where each y_k is $n_k \times 1$ and X_k is $n_k \times p$. Let $D_k(\theta)$ and $C_k(\theta)$ correspond to the k-th subset of the data. Assume that we are able to obtain posterior samples for $\Omega = \{\beta, \theta\}$ from (5) applied independently to each of K subsets of the data. To be specific, assume that $\Omega_k = \{\Omega_k^{(1)}, \Omega_k^{(2)}, \ldots, \Omega_k^{(M)}\}$ be a collection of M posterior samples from $p(\Omega \,|\, y_k)$. We refer to each $p(\Omega \,|\, y_k)$ as a "subset posterior". The subset posteriors are computed from subsets of the data and posterior inference from one subset will be substantially different from another. This is especially true for block-independent spatial models when the blocks (subsets) may not adequately represent the entire random field. One approach is to design partitions of the data that will ensure the block independent model is a good approximation to the full spatial model. This, howevever, may not be easily achieved and may be specific to the dataset. The meta-kriging approach we outline below should be more widely applicable. It attempts to combine, optimally and meaningfully, the subset posteriors to arrive at a legitimate probability density. We will refer to this as the "meta-posterior" and will tend to be more immune to the drawbacks of pooled inference using block-independent models.

Algorithm 1 Algorithm to compute Geometric Median (GM) of posterior distributions

a. **Initial Condition:** $\alpha_{\rho,k}^{(0)}(y) = \frac{1}{K}$.

b. `For` $m \geq 1$

   i. mth iterate of $\alpha_{\rho,k}^{(m)}(y)$ is given by $\alpha_{\rho,k}^{(m)}(y) = \frac{\|p_k - \pi^{*(m-1)}\|_\rho^{-1}}{\sum_{j=1}^K \|p_j - \pi^{*(m-1)}\|_\rho^{-1}}$.

   ii. mth iterate of π^* (denoted as $\pi^{*(m)}$) is given by $\pi^{*(m)} = \sum_{k=1}^K \alpha_{\rho,k}^{(m)}(\boldsymbol{y}) p_k$.
   Note that the posterior p_k is approximated by the corresponding empirical posterior $\frac{1}{M}\sum_{j=1}^M 1_{\boldsymbol{\Omega}_k^{(j)}}$ so that $\pi^{*(m-1)}$ is approximated by $\frac{1}{M}\sum_{k=1}^K \sum_{j=1}^M \alpha_{\rho,k}^{(m-1)}(\boldsymbol{y}) 1_{\boldsymbol{\Omega}_k^{(j)}}$.

c. **Stopping Condition:** Iteration proceeds until $\|\pi^{*(m)} - \pi^{*(m-1)}\|_\rho < \epsilon$, where ϵ is a user-specified tolerance level.

Our approach relies upon the unique *Geometric Median* (GM) of the subset posteriors (Minsker (2015) and Minsker et al. (2014)). Assume that the individual posterior densities $p_k \equiv p(\Omega \,|\, y_k)$ reside on a Banach space $\mathcal{H}$ equipped with norm $\|\cdot\|$. The GM is defined as

$$\pi^*(\cdot \,|\, y) = \arg\min_{\pi \in \mathcal{H}} \sum_{k=1}^K \|p_k - \pi\|_\rho \;, \tag{6}$$

where $y = (y_1^\top, y_2^\top, \ldots, y_K^\top)^\top$. The norm quantifies the distance between any two posterior densities $\pi_1(\cdot)$ and $\pi_2(\cdot)$ as $\|\pi_1 - \pi_2\|_\rho = \| \int \rho(\Omega, \cdot) d(\pi_1 - \pi_2)(\Omega) \|$, where $\rho(\cdot)$ is a positive-definite kernel function. In what follows, we assume $\rho(z_1, z_2) = \exp(-\|z_1 - z_2\|^2)$.

The GM is unique. Further, the geometric median lies in the convex hull of the individual posteriors so $\pi^*(\Omega \,|\, y)$ is a legitimate probability density. Specifically, $\pi^*(\Omega \,|\, y) = \sum_{k=1}^K \alpha_{\rho,k}(y) p_k$,$\sum_{k=1}^K \alpha_{\rho,k}(y) = 1$, each $\alpha_{\rho,k}(y)$ being a function of ρ, y, so that $\int_\Omega \pi^*(\Omega \,|\, y) d\Omega = 1$.

Computing the GM $\pi^* \equiv \pi^*(\Omega \,|\, y)$ is achieved by the popular Weiszfeld's iterative algorithm that estimates $\alpha_{\rho,k}(y)$ from the subset posteriors p_k for each $k = 1, 2, \ldots, K$. To further elucidate, we use a well known result that the GM π^* satisfies $\pi^* = \frac{\sum_{k=1}^K \|p_k - \pi^*\|_\rho^{-1} p_k}{\sum_{k=1}^K \|p_k - \pi^*\|_\rho^{-1}}$, so that $\alpha_{\rho,k}(y) = \frac{\|p_k - \pi^*\|_\rho^{-1}}{\sum_{j=1}^K \|p_k - \pi^*\|_\rho^{-1}}$. Since there is no apparent closed form solution for $\alpha_{\rho,k}(y)$ satisfying this equation, we resort to the Weiszfeld iterative algorithm outlined in Algorithm 1 (Minsker et al., 2014).

A closed form expression for $\|p_k - \pi^{*(m-1)}\|_\rho$ is easily obtained by referring to the formula

$$\|\pi_1 - \pi_2\|_\rho = \sum_{i=1}^{M_1}\sum_{j=1}^{M_1}\gamma_{1i}\gamma_{1j}\rho(z_{1i}, z_{1j}) + \sum_{i=1}^{M_2}\sum_{j=1}^{M_2}\gamma_{2i}\gamma_{2j}\rho(z_{2i}, z_{2j}) - 2\sum_{i=1}^{M_1}\sum_{j=1}^{M_2}\gamma_{1i}\gamma_{2j}\rho(z_{1i}, z_{2j}), \tag{7}$$

where $\pi_1 = \sum_{i=1}^{M_1}\gamma_{1i}1_{z_{1i}}$ and $\pi_2 = \sum_{i=1}^{M_2}\gamma_{2i}1_{z_{2i}}$. z_{1i}, z_{2i}'s are dummy variables representing atoms of Ω, $1_{z_{1i}}, 1_{z_{2i}}$ are indicator functions at z_{1i}, z_{2i} respectively. Weiszfeld's algorithm yields the geometric median of points lying on a separable Banach space.

In the Web Appendix we show that for a large sample $\pi^*(\cdot \,|\, y)$ provides an "optimal" approximation of the full posterior distribution in certain restrictive settings. It is, therefore, reasonable to approximate the posterior predictive distribution $p(y(s_0) \,|\, y)$ by the subset posterior predictive distributions $p(y(s_0) \,|\, y_k)$. Let $\{y(s_0)^{(j,k)}\}_{j=1}^M$, $k = 1, \ldots, K$, be samples obtained from the posterior predictive distribution $p(y(s_0) \,|\, y_k)$ for the k-th subset posterior. Then,

$$p(y(s_0) \,|\, y) \approx \sum_{k=1}^{K}\alpha_{\rho,k}(y)p(y(s_0) \,|\, y_k) = \sum_{k=1}^{K}\alpha_{\rho,k}(y)\int p(y(s_0) \,|\, \Omega, y_k)p(\Omega \,|\, y_k)d\Omega \ ,$$

Therefore, the empirical posterior predictive distribution of the meta posterior is given by $\sum_{k=1}^{K}\sum_{j=1}^{M}\frac{\alpha_{\rho,k}(\boldsymbol{y})}{M}1_{y(s_0)^{(j,k)}}$, from which the posterior predictive median and the 95% posterior predictive interval for the unobserved $y(s_0)$ are readily available.

Regarding inference for the spatial process $w(\cdot)$ at arbitrary location $\boldsymbol{s}_0$, we use posterior samples $\{w(s_0)^{(j,k)}\}_{j=1}^M$ from the k-th subset posterior $p(w(s_0) \,|\, y_k)$ for each $k = 1, \ldots, K$. Again, an approximation for $p(w(s_0) \,|\, y)$ is readily available through the meta posterior

$$p(w(s_0) \,|\, y) \approx \sum_{k=1}^{K}\alpha_{\rho,k}(y)p(w(s_0) \,|\, y_k) = \sum_{k=1}^{K}\alpha_{\rho,k}(y)\int p(w(s_0) \,|\, \Omega, y_k)p(\Omega \,|\, y_k)d\Omega \ .$$

Approximate posterior sampling from $p(w(s_0) \,|\, y)$ then proceeds by drawing samples from the empirical approximation given by $\sum_{k=1}^{K}\sum_{j=1}^{M}\frac{\alpha_{\rho,k}(y)}{M}1_{w(s_0)^{(j,k)}}$. Obtaining the approximate posterior median and 95% credible interval for $w(s_0)$ are now easily achieved.

3 Illustrations

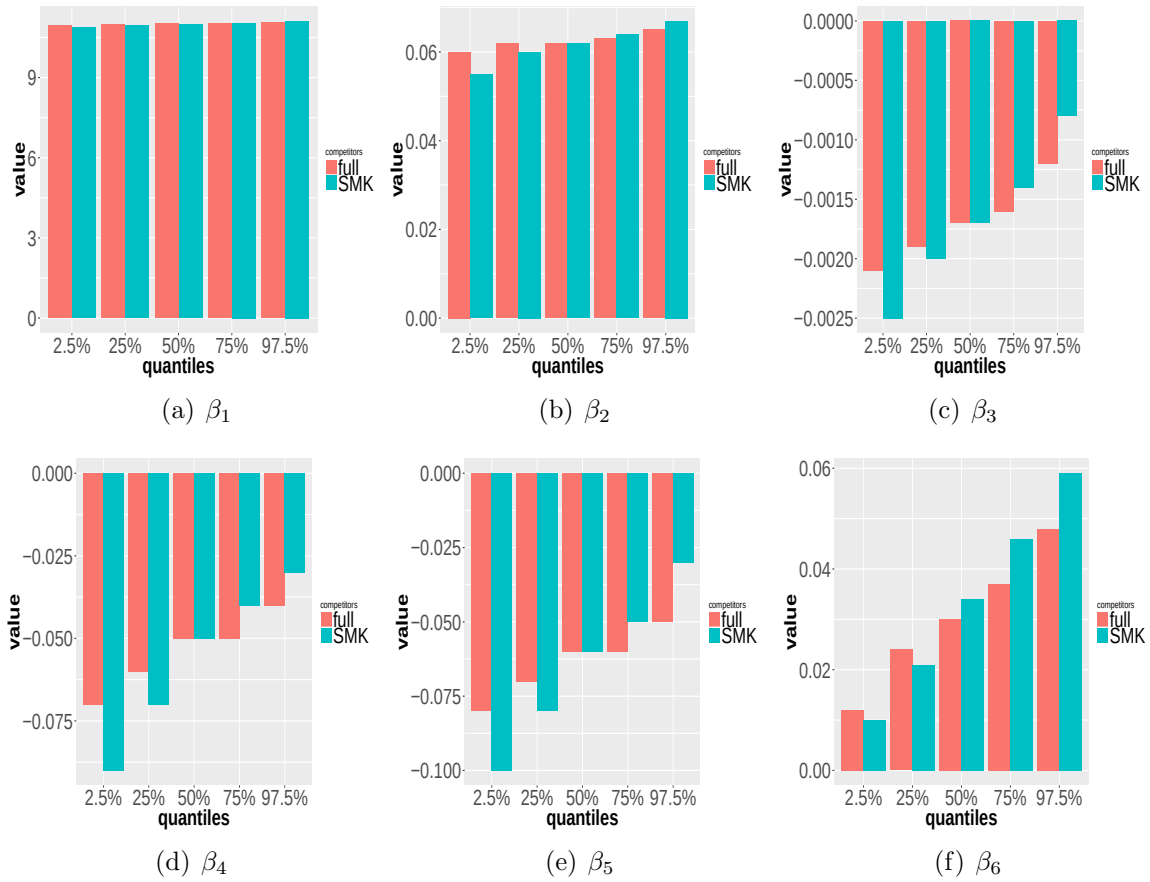
3.1 Illustrating Weiszfeld's algorithm for the conjugate Bayesian linear model

As a first step to study the meta posterior, we turn our attention to the non-spatial conjugate Bayesian linear model (1). As described in Section 2.1, conjugate Bayesian linear models (1) yield the joint posterior distribution of $\{\beta, \sigma^2\}$ in closed-form and is easy to sample from. It is, therefore, instructive to see the accuracy of the approximation offered by the meta posterior of β in comparison with the exact posterior distribution of β. This section presents such an analysis by fitting both the full posterior and the meta posterior from (1) on `FORMGMT` data from the `spBayes` package. The `FORMGMT` dataset contains information on a response and 6 predictors at 1342 locations. To evaluate the meta posterior, this dataset is divided randomly into 6 subsets approximately of the same size. Weiszfeld's algorithm is then applied to the subset posteriors to obtain an empirical approximation of the meta posterior for each component of β. Figure 1 demonstrates the accuracy of the meta posterior by presenting the $2.5\%, 25\%, 50\%, 75\%$ and 97.5% quantiles for each component of β from the meta posterior and the exact full posterior. The figure shows that the quantiles of β from the meta and the exact full posterior are very similar. A similar story is told by the meta posterior of σ^2. Performance of the meta posterior in non-spatial models are convincing enough to propel careful implementation of SMK on more general spatial process models. The next few sections lay them out in detail.

3.2 Simulation experiments

We use synthetic datasets to assess model performance with regard to learning about process parameters, interpolating the unobserved residual spatial surface and predicting at new locations. Though SMK potentially adapts to any spatial regression model, we confine ourselves to studying SMK for (i) Gaussian process based geostatistical models (GP) and (ii) the tapered Gaussian process (TGP) or Gaussian process with compactly supported correlation functions. Further theoretical results are presented in the Web Appendix.

This section presents two simulation studies. *Simulation 1* is presented for moderately



(a) β_1 (b) β_2 (c) β_3

(d) β_4 (e) β_5 (f) β_6

Figure 1: Posterior quantiles of full posterior and meta posterior: (a) β_1; (b) β_2; (c) β_3; (d) β_4; (e) β_5, and (f) β_6.

large datasets with $3,500$ locations, while *Simulation 2* presents a study with $41,000$ locations. The moderate size in Simulation 1 allows full Bayesian implementation of the full Gaussian process (GP) model (without approximation) and the tapered Gaussian process (TGP) for comparison with SMK approximations to the full GP (SMK-GP). For both simulations, data are generated from a standard Gaussian process model with the `RandomFields` package. Additionally, *Simulation 2* presents a case study with data generated from the tapered Gaussian process model. For this simulation, inference from SMK with tapered Gaussian process (SMK-TGP) model fitted to each subset is studied to assess how good an approximation to the full TGP is offered by SMK-TGP.

Competitors: As competitors to SMK, we employ

(a) locally approximated Gaussian process models (laGP) (Gramacy and Apley, 2015). This

is a state-of-the-art procedure in computer emulations and is not designed to provide full scale Bayesian inference. However, predictive point estimates with associated standard errors can readily be obtained from laGP. They are used to compare predictive inference including point estimates and uncertainties between SMK and laGP. Gramacy and Apley (2015) mention that laGP often outperforms nearest neighbor methods. Thus, in the absence of easily implementable `R` package/publicly available codes for nearest neighbor methods, comparison with laGP serves as a reasonable indicator. The `laGP` package in CRAN (`https://cran.r-project.org/web/packages/laGP/index.html`) is used to implement laGP.

(b) Multiresolution Kriging based on Markov random fields (LatticeKrig) (Nychka et al., 2015). The `LatticeKrig` package hosted on CRAN (`https://cran.r-project.org/web/packages/LatticeKrig/index.html`) offers frequentist implementations of LatticeKrig. Similar to laGP, predictive point estimates with associated standard errors can readily be obtained from LatticeKrig. We often refer to LatticeKrig as LK.

(c) Block independent pooled spatial models, referred to as BISP. BISP is a two stage procedure. In the first stage, similar to SMK, one fits a spatial model independently on K exhaustive and mutually exclusive subsets of data. In the second stage, weighted inference is drawn based upon subset posteriors and weights $1/K$ corresponding to each subset posterior. For fair comparison between BISP and SMK, two models are fitted under the same subset partitioning scheme.

In *Simulation 1* with moderately large number of data locations, we could also implement a full Gaussian process without any approximation and the full tapered Gaussian process with compactly supported correlation function as competitors to assess the accuracy of the approximation offered by SMK-GP. However, in *Simulation 2* with $41,000$ locations, full Bayesian inference for the full Gaussian process is prohibitive and is not considered. Moreover, full Bayesian inference on tapered Gaussian processes also comes with a lot of computational expense, primarily due to computing the determinant of the covariance matrix in each iteration. Therefore, the tapered Gaussian process model is also omitted from the bigger simulation study. We also implement Treed GP (Treed-GP) with the `tgp` package in

R for *Simulation 1* (not shown) and find that the Treed-GP's inferential performance is less than the full GP. Treed-GP is found to be computationally prohibitive for *Simulation 2*.

We consider a *parallel implementation* of the SMK over multiple cores. The entire analysis implementing parallelization is carried out in R with the `doParallel` and `foreach` packages on a Unix workstation with 64 cores. All the interpolated spatial surfaces are obtained using the R package `MBA`. All predictive inferences are based upon 25 simulated datasets.

3.3 Simulation 1

Simulation 1 is performed under moderately large sample sizes to accommodate the full GP model. We generate $3,500$ observations within a unit square domain from the classical geostatistical model with likelihood $y \sim N(\beta_0, V_y(\theta))$, $V_y(\theta) = \{\kappa(s_i, s_j)\}_{i,j=1}^n + \tau^2 I$, $\theta = \{\sigma^2, \tau^2, \phi, \nu\}$. For this article we will only use the exponential covariance function $\kappa(s_i, s_j) = \sigma^2 \exp(-\phi \|s_i - s_j\|)$, where $\theta = \{\sigma^2, \tau^2, \phi\}$ which arises from the popular Matérn class with the smoothness parameter $\nu = 1/2$ (see, e.g., Stein (2012)).

To fit GP models in every subset, we assign a noninformative prior to β_0. τ^2 and σ^2 are assigned a $IG(2, 1)$ priors (mean is 1). The spatial decay parameter ϕ is assigned a $U(0.3, 300)$ which corresponds to a slow decay in spatial correlation and a strong spatial signal in the simulated data, given that the maximum distance between any two observations is 1.4.

One important ingredient of the SMK is partitioning the dataset into subsets. Consequently, we have explored different partitioning schemes for the dataset to assess their impact on the inference. For example, we have investigated the SMK by partitioning the domain into disjoint sub-domains followed by choosing each subset consisting of observations from these sub-domains. This, however, is inefficient because many sub-domains are not representative of the full dataset and there is the risk of incorrectly estimating the process parameters from the sub-domains. This scheme also does not work well for block-independent models. The Treed-GP model attempts to circumvent this problem by averaging over the partitions. This improves inferential performance with regard to estimation but posterior predictive inference at arbitrary locations is still complicated.

The GM, hence SMK, tends to be more robust to partitioning schemes and one need not average over partitions. However, simply partitioning the data according to sub-regions may

still be unwise. Instead, we adopt a *random partitioning* scheme that proceeds as follows

- Draw $\mathcal{S}_1$, the first subset, randomly from the full data (denoted by $\mathcal{S}$).
- For $k = 2, ..., K$, draw $\mathcal{S}_k$, the kth subset, randomly from $\mathcal{S} - (\cup_{i=1}^{k-1}\mathcal{S}_i)$.

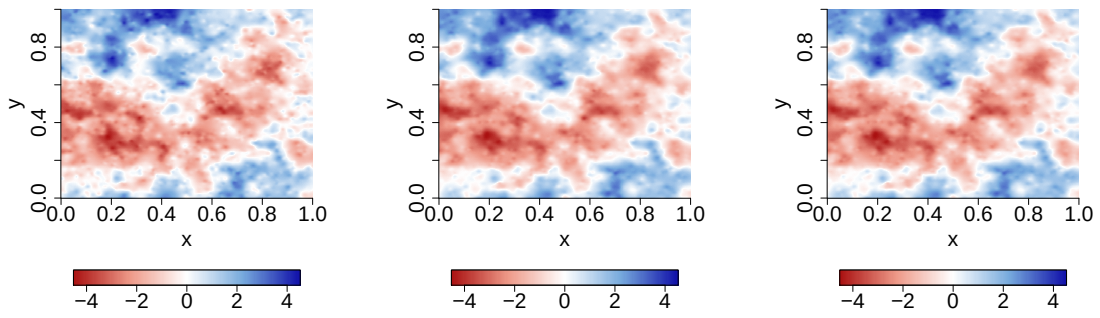
The *random partitioning* scheme facilitates each subset to be a reasonable representative of the entire domain, so that each subset posterior acts as a "weak learner" of the full posterior. Our investigations also reveal that k-means clustering of locations into subsets often leads to inferior inference than SMK fitted on subsets constructed with the random partitioning scheme. In fact, SMK seems to work best when each subset posterior is a noisy approximation of the full posterior. Subset posteriors estimated using random subsets is one way to find noisy approximations to the full posterior. A more sophisticated approach would divide the domain and choose representative samples from each sub-domain in every subset. Few simulations show indistinguishable performance of this sophisticated partitioning scheme with the random partitioning scheme. Hereafter, we stick to the random partitioning scheme for all simulation studies and real data analysis.

To demonstrate the SMK for various choices of the number of subsets (K) under the random partitioning scheme, we experiment with $K = 3, 6$ and 10 subsets of the data with $n = 1000, 500$ and 300 observations in each subset respectively.

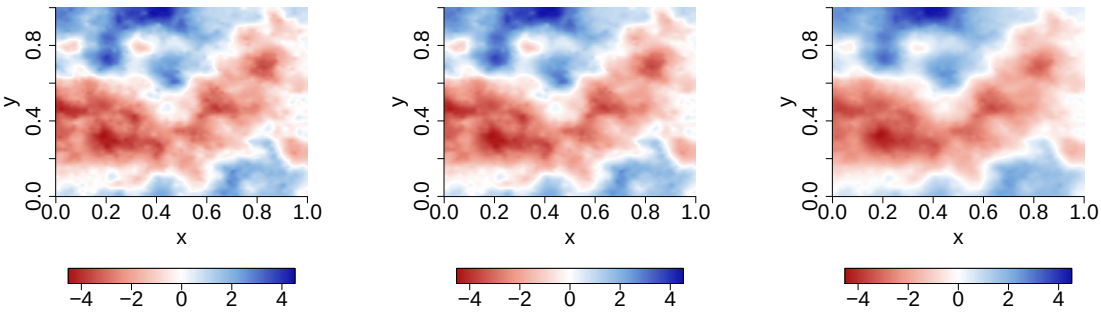
Table 1 shows point estimates of the parameters along with their 95% credible intervals for a representative simulation. While true values of β_0 and τ^2 are always contained within the 95% credible intervals for all of the parameters, σ^2 and ϕ estimates in SMK-GP are also found to be consistent with the full Gaussian process. As expected, computation time for SMK-GP is much smaller than both GP and TGP.

In terms of surface interpolation, Figure 2 shows, not surprisingly, that the performance of SMK-GP improves by reducing the number of subsets. Clearly, in surface interpolation, the full Gaussian process sets the benchmark. It is observed that the estimated spatial surface from SMK-GP with $K = 3$ subsets is indistinguishable from the surface obtained using the full Gaussian process, barring some negligible smoothing effects.

Predictive performance of the different approaches are compared using mean squared prediction error (MSPE), length and coverage of 95% predictive intervals. Figure 3 demon-



(a) True surface (b) full GP (c) Tapered GP



(d) $K = 3$ (e) $K = 6$ (f) $K = 10$

Figure 2: Residual spatial surface for: (a) synthetic spatial random effect generated using 3,000 observations; (b) full Gaussian process; (c) tapered Gaussian process; (d) estimated spatial random effects for meta posterior with $K = 3$; (e) estimated spatial random effects for meta posterior with $K = 6$, and (f) estimated spatial random effects for meta posterior with $K = 10$.

Table 1: Parameter credible intervals, 50 (2.5 97.5) percentiles of all the parameters for SMK-GP, full GP and tapered GP (TGP). SMK with Gaussian process is fitted for $K = 3, 6, 10$ number of subsets.

				SMK-GP		
Parameter	True value	GP	TGP	3	6	10
β_0	1	1.32 (-2.63,3.93)	0.24 (0.14,0.34)	1.21 (-1.12,2.38)	1.27 (-3.09, 4.27)	1.23 (-3.07, 5.56)
τ^2	0.10	0.10 (0.09,0.11)	0.11 (0.10,0.12)	0.10 (0.08, 0.11)	0.11 (0.07, 0.16)	0.13 (0.06, 0.20)
σ^2	4	6.47 (3.73, 11.47)	0.63 (0.56,0.69)	6.58 (3.84, 11.38)	9.02 (6.49, 14.60)	9.56 (7.05, 15.40)
ϕ	3	1.56 (1.35, 2.86)	0.49 (0.38,0.84)	1.21 (0.61, 2.43)	0.68 (0.36, 1.12)	0.65 (0.35, 1.02)
time (in min)	–	680.40	518.86	49.29	29.45	12.65

strates similar coverage with narrower predictive interval for SMK-GP with $K = 3$ compared to $K = 10$. It is also observed that naively combining subset posterior inferences using BISP

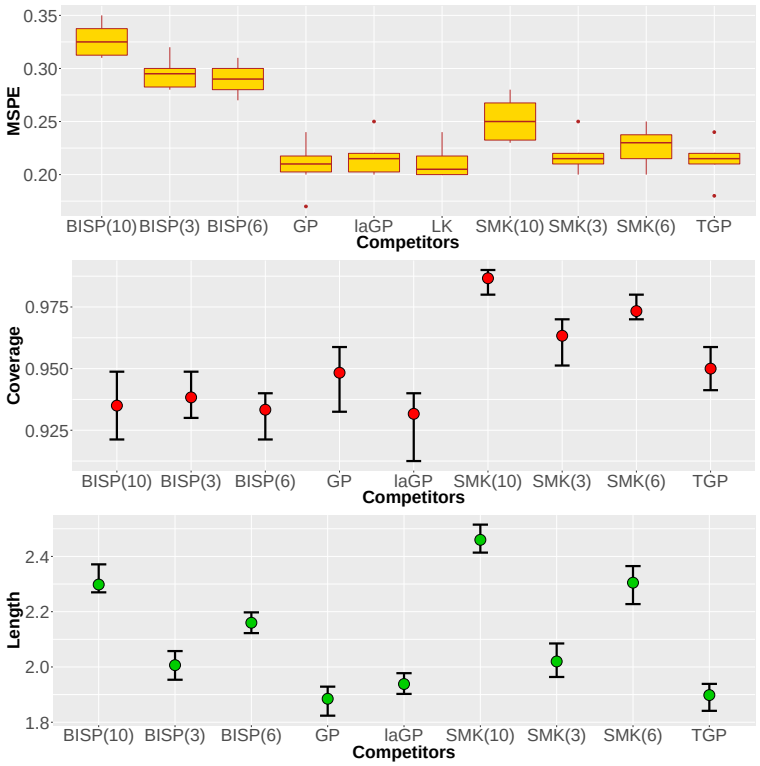


Figure 3: Plot at the top indicates boxplot of mean squared prediction error for all competitors over 25 replications. Second and third plots show coverage and length of 95% predictive intervals for the competitors over the same replications. LatticeKrig shows extreme under-coverage compared to the others and is not presented alongside the others.

leads to significantly higher MSPE. In terms of MSPE, SMK-GP demonstrates almost indistinguishable performance with full GP, TGP and other approaches such as laGP and LK. Additionally, SMK-GP exhibit slightly higher predictive coverage with slightly wider prediction intervals than the full GP, TGP and laGP. On the other hand, LK shows severe under-coverage (not shown) with narrower predictive intervals. In fact, the average length and coverage of 95% predictive intervals for LK is given by 0.63 and 0.52 respectively. The under-coverage of LK is perhaps caused due to using asymptotic predictive interval.

Finally, we demonstrate through Figure 4 that for all three cases ($K = 3, 6, 10$), substantial reduction in MSPE is achieved by the meta posteriors as compared to subset posteriors. This is expected since the meta posterior is centered closer to the full un-approximated posterior than the individual subset posteriors. Simulation 1 thus presents a convincing case about the ability of SMK-GP to act as a computationally convenient approximation to the

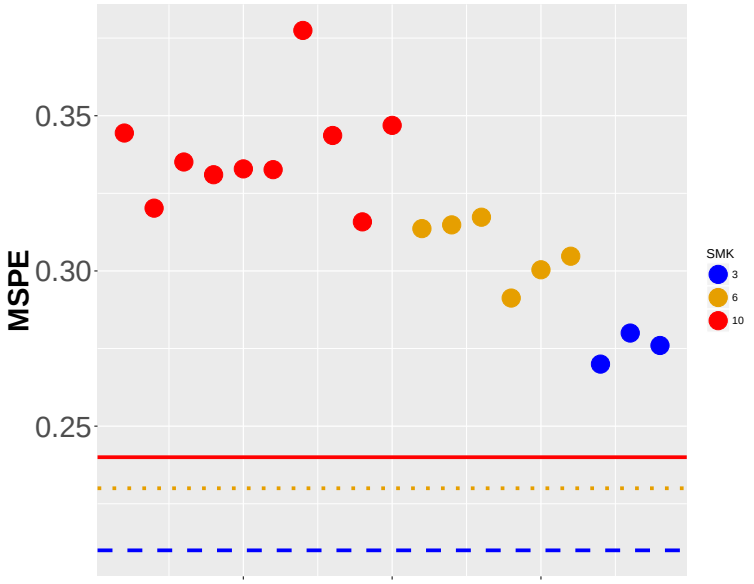


Figure 4: Plots the MSPE calculated from subsets for one representative simulation. Blue, golden and red colors represent subset MSPE values for $K = 3, 6, 10$ subsets respectively. Corresponding MSPEs from meta posteriors are provided in the solid, dotted and dashed lines respectively.

full GP. The next section strengthens our argument further using simulations with much larger sample sizes.

3.4 Large Simulation Studies

3.4.1 Simulation 2

While Simulation 1 compares SMK-GP with the full Gaussian process, we shall ultimately be interested in assessing the performance of SMK-GP in large data settings that prohibit fitting full Gaussian processes. Accordingly, Simulation 2 generates $41,000$ observations from the Gaussian processes with an exponential correlation kernel, of which $N = 40,000$ are used for model fitting, and the rest for prediction. Following the general SMK algorithm, training data with N samples are equally divided into K non-overlapping subsets with Gaussian process models fitted to each subset. Choice of prior distributions on $\phi, \tau^2, \sigma^2, \beta_0$ are kept similar to Simulation 1.

To study the performance of SMK-GP with respect to the number of subsets, SMK-GP's architecture is employed with $K = 20, 25, 40$. Table 2 presents the posterior median along

with 95% credible intervals for all parameters for a representative simulation. SMK-GP delivers accurate point estimates of parameters with 95% credible intervals containing the true parameter values except ϕ. This is not entirely unexpected, given that ϕ is weakly identifiable. Also, unsuprisingly, the credible intervals are a little wider for $K = 40$ than $K = 20$. The range parameter shows a little underestimation which is not entirely surprising as ϕ is weakly identifiable. Most importantly, SMK-GP approximation to the full GP is able to deliver full Bayesian inference for $40,000$ observations within a few hours, which otherwise would have taken a month for the full GP without the SMK approximation.

Table 2: Parameter credible intervals, 50 (2.5 97.5) percentiles for all the parameters. SMK with Gaussian process is fitted for $K = 20, 25, 40$ number of subsets.

		SMK (Gaussian Process)		
Parameter	True value	20	25	40
β_0	1	0.28 (-1.96,2.39)	0.28 (-2.02, 2.33)	0.29 (-1.99, 2.70)
τ^2	0.10	0.09 (0.08, 0.10)	0.09 (0.08, 0.11)	0.09 (0.08, 0.11)
σ^2	2	1.45 (0.65, 2.33)	1.43 (0.62, 2.49)	1.51 (0.63, 3.12)
ϕ	3	1.35 (1.31, 1.64)	1.36 (1.26, 1.70)	1.37 (1.32, 1.80)
time (in min)	–	260.40	216	64.8

A comprehensive study of predictive inference for SMK-GP along with BISP, laGP and LK is presented in Figures 5. Consistent with our earlier findings, SMK-GP performs significantly better than BISP with regard to MSPE. The laGP and LK approaches perform little better for point prediction although SMK-GP is competitive. As discussed before, SMK-GP credible intervals tend to be slightly wider than laGP resulting in marginally higher coverage. BISP exhibits marginally lower coverage while LK suffers from severely lower coverage. Similar to Simulation 1, subset posteriors are found to provide significantly higher MSPE than the meta-posterior, but we omit this analysis here.

3.4.2 SMK on Tapered Gaussian Process

We now turn to SMK on Gaussian processes with compactly supported correlation functions, popularly referred to as tapered Gaussian processes. Investigating the computationally convenient SMK approximation of the tapered GP (TGP), referred to as SMK-TGP, is significant for multiple reasons. Employing Gaussian processes with compactly supported correlation functions is common practice in many real life applications pertaining to the environmental and geological sciences. Additionally, Kaufman et al. (2008) argue that Gaussian

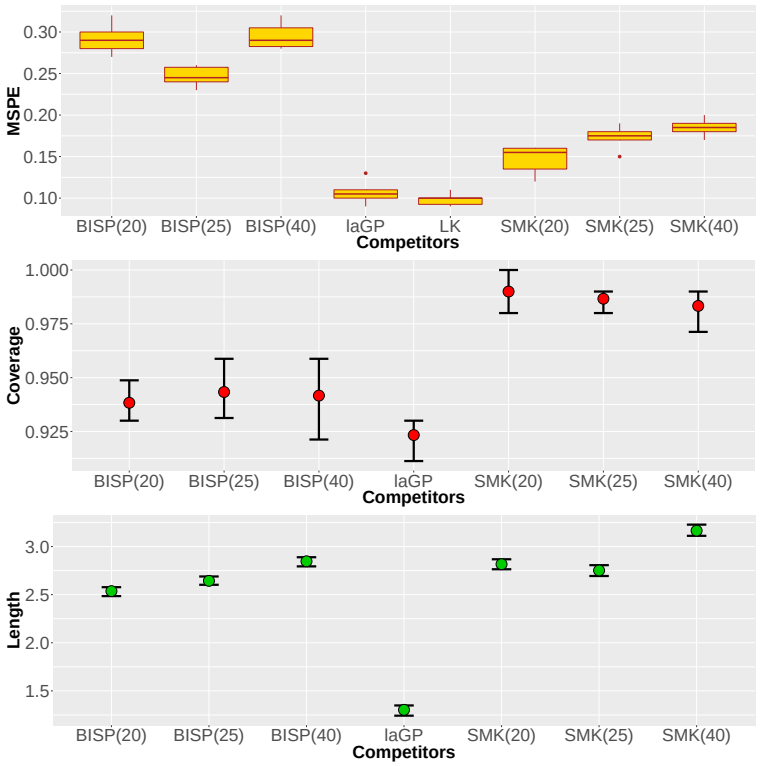


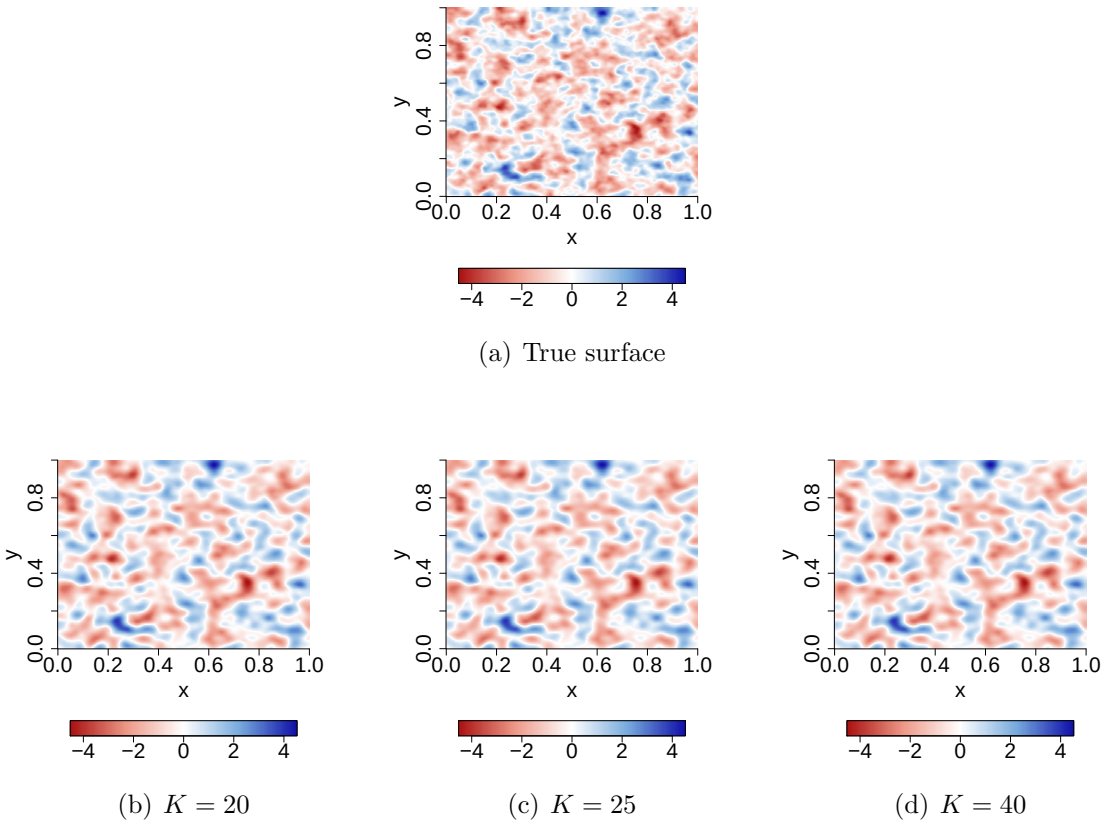
Figure 5: Plot at the top indicates boxplot of mean squared prediction error for all competitors over 25 replications. Second and third plots show coverage and length of 95% predictive intervals for the competitors over the same replications. LatticeKrig shows extreme undercoverage compared to the others and is not presented alongside the others.

processes specified with the Matérn class of covariance functions (see Stein (2012)) can be well approximated by a certain class of computationally convenient alternative Gaussian processes with compactly supported correlation functions. Such an edge in terms of computation for TGP over GP disappears for large sample sizes primarily due to evaluating determinants of large $N \times N$ covariance matrices. Therefore, it is important to investigate if a fast approximation to the TGP can emerge from the spatial meta kriging approach.

Table 3: Parameter credible intervals, 50 (2.5 97.5) percentiles for all the parameters. SMK with Gaussian process is fitted for $K = 20, 25, 40$ number of subsets.

		SMK (Tapered Gaussian Process)		
Parameter	True value	20	25	40
β_0	1	0.91 (0.74,1.08)	0.90 (0.73, 1.08)	0.91 (0.72, 1.09)
τ^2	0.10	0.09 (0.06, 0.12)	0.09 (0.08, 0.11)	0.10 (0.06, 0.14)
σ^2	2	1.87 (1.67, 2.10)	1.43 (0.62, 2.49)	1.88 (1.66, 2.13)
ϕ	3	3.38 (1.88, 5.40)	1.36 (1.26, 1.70)	3.23 (1.34, 5.55)
time (in min)	–	218.72	178.36	52.42

We fit a tapered Gaussian process in data subsets in different processors and combine subset posteriors using Algorithm 1 to compute the meta posterior. To carry out posterior inference in each subset, prior distributions similar to section 3.4.1 are assigned to the parameters of interest $\{\beta_0, \phi, \tau^2, \sigma^2\}$. As a tapering kernel, the popularly used Wendland tapering kernel (Wendland, 2004), $\kappa_\delta(s, s') = \left(1 - \frac{||s-s'||}{\delta}\right)_+^4 \left(1 + 4\frac{||s-s'||}{\delta}\right)$ is employed, where δ is a tuning parameter that controls the sparsity of the covariance matrix and is chosen depending on the computational architecture available to the user. For our analysis, $\delta = 0.1$ is chosen, which yields $\sim$ on an average 8% nonzero entries in the dispersion matrix in each subset.



(a) True surface

(b) $K = 20$ (c) $K = 25$ (d) $K = 40$

Figure 6: Residual spatial surface for: (a) synthetic spatial random effect generated using 40,000 observations using TGP; estimated spatial random effects for meta posterior with TGP fitted in each subset for (b) $K = 20$; (c) $K = 25$, and (d) $K = 40$.

Similar to Section 3.4.1, we find that the meta posterior for tapered GP demonstrates $\sim$ 30-40% improvement in MSPE over subset posteriors. The range of MSPE for subset

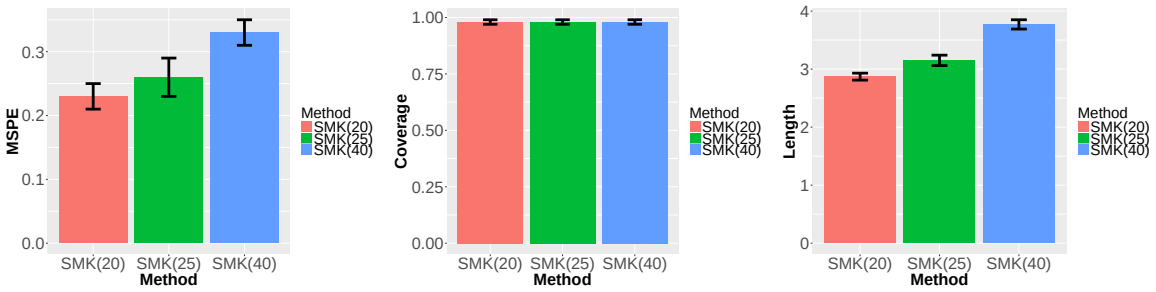


Figure 7: Figures present MSPE, length and coverage of 95% predictive intervals for SMK-TGP with different number of subsets. SMK(20), SMK(25) and SMK(40) stand for SMK-TGP with $20, 25, 40$ subsets respectively.

posteriors is (0.37,0.44), (0.43,0.52) and (0.46,0.55) for $K = 20, 25, 40$, respectively. This indicates that the meta posterior for tapered Gaussian process concentrates significantly better than subset posteriors. Parameter estimates along with their 95% credible intervals are presented in Table 3. Clearly, all parameters are correctly estimated with their 95% credible intervals covering the truth. Additionally, Figure 6 shows that the residual spatial surfaces for SMK-TGP quite accurately reconstruct the true spatial surface. Interestingly, even with increasing K, surface interpolation for SMK-TGP deteriorates minimally. One explanation might arise from the fact that the data generated from tapered GP has minimal long range dependence that facilitates better performance of SMK-TGP. Similar to Section 3.4.1, SMK-TGP achieves proper well calibrated prediction by maintaining predictive uncertainty little over 95% as observed in Figure 7.

Computation time

Further, to study the effect of the number of subsets on the predictive performance of the meta posterior and the computational advantages they fetch, computation times for SMK-GP are provided for every simulation study. With parallel implementation of subset GPs in different processors, one needs to evaluate the subset likelihood for the Metropolis step in every processor. The metropolis step in every processor requires a Cholesky decomposition of an $\left(\frac{N}{K}\right) \times \left(\frac{N}{K}\right)$ matrix involving $O(\left(\frac{N}{K}\right)^3)$ flop counts. It is a well known fact that the computational complexity of GP regression per iteration is dominated by this term. Thus, with a parallel implementation of the algorithm, the computational complexity is given by $O(\left(\frac{N}{K}\right)^3)$. Even if the entire computation is performed in one processor, the computational

complexity for the entire data is given by $O(K\left(\frac{N}{K}\right)^3)$. Clearly, the number of subsets K plays a central role in controlling the computational complexity. Ideally, the choice of K is decided depending upon the computational architecture so as to keep the computation fast without losing much performance accuracy. Depending on the available computational resources, the natural idea would be to vary K slowly with N, i.e. $K \sim N^c$, for some $0 < c < 1$. This leads to a computational complexity of $O(N^{3-3c})$ for each subset. Additionally, SMK-GP frees the storage of the $N \times N$ covariance matrix in the memory and requires storage of K $\frac{N}{K} \times \frac{N}{K}$ matrices. Indeed with $K = N^c$, SMK-GP reduces storage complexity from $O(N^2)$ down to $O(N^{2-2c})$. Finally, it is to be mentioned that the computational complexity of the SMK framework is dependent upon the computational complexity of the model fitted to each subset. Complexity of computation in each subset can be substantially mitigated by fitting a fast nearest neighbor or a multiscale approach to each subset. In fact, SMK framework applied to such models may dramatically reduce the computational complexity, even to the point of being sub-linear in N. We propose to pursue this in a future article.

3.5 Analysis of Sea Surface Temperature Data

An important ecological issues concerning our planet is climate change. It is generally accepted that the earth's climate will change in response to radiative forces induced by the changes in atmospheric gases, cloud temperature, sea surface temperature, water vapor, aerosol (liquid and solid particles suspended in the air), among others. Developing conceptual and predictive global climate models to accurately assess climate and potential climate changes are of major interest in recent years. Of particular interest is the collection of sea surface temperature data (in Centigrade). This is important for tropical cyclogenesis as well as for studying the formation of sea breezes and sea fog and for calibrating measurements from weather satellites. For a long time, sea surface temperature data from ocean samples has been collected by voluntary observing ships, buoys, military and scientific cruises. In the early days, interest resided mainly in the mean climatological state of the ocean so as to understand the flow and distribution of water streams. As climatological research started to emerge, another important requirement became quantifying the variability around the mean in spatial and temporal scales. A number of articles have appeared in order to address this

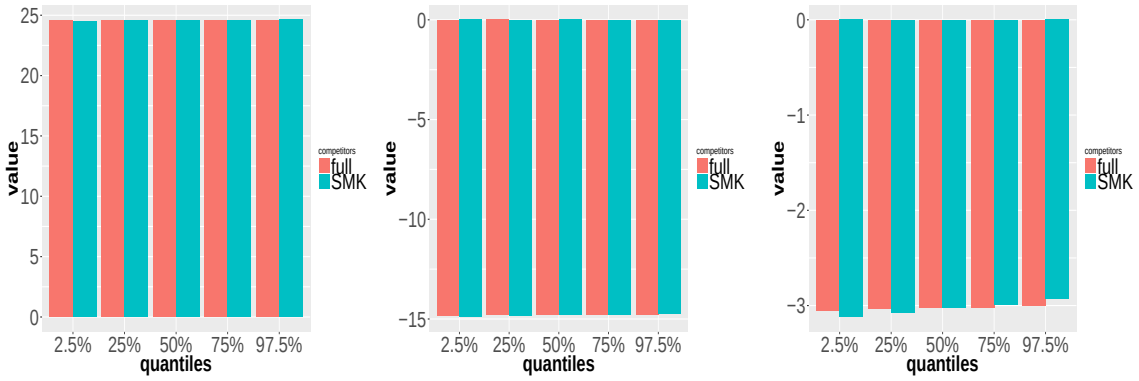
issue in the recent years, see e.g. Higdon (1998), Lemos and Sansó (2009), Lemos and Sansó (2006), Berliner et al. (2000).

In this article, we consider the problem of capturing the spatial trend and characterizing anomaly (uncertainty) in the sea surface temperature (SST) in the West coast of mainland USA, Canada and Alaska, between $30^0 - 60^0$ N. latitude and $122^0 - 152^0$ W. longitude. The dataset has been obtained from NODC World Ocean Database 2016 and we use data collected in the month of October for all the spatial locations. Note that SMK implemented with Gaussian process does not possess any temporal component, and so data collected in the same month across the domain is used for the analysis. We perform screening of the data to ensure quality control and then choose a random subset of $120,000$ spatial observations over the domain of interest. Out of the total observations, about 98%, i.e $N = 117,600$ observations are used for model fitting and the rest are used for prediction. The domain of interest is large enough to allow considerable spatial variation in SST from north to south and provides an important first step in extending these models for the analysis of global scale SST database.

The plot of the sea surface is provided in Figure 9(b). As expected, the plot reveals a clear trend of decrease in the sea surface temperature with increasing latitude. Thus, sea surface temperature data possesses inherent directional anisotropy that makes fitting ordinary Gaussian process model with stationary covariance kernel unreasonable. Consequently, we add latitude and longitude as linear predictors to each subset while fitting the SMK-GP. To justify our approach, a non-spatial model with latitude and longitude as linear predictors is fitted and surface plots of ordinary least square (OLS) residuals are presented in Figure 9(c). No clear anisotropic pattern emerges from Figure 9(c). Further, the empirical semivariogram (see Figure 9(a)) of the OLS residuals confirms nearly isotropic behavior of the spatial covariance function.

For spatial GP models, the full posterior distribution can not be obtained in closed form. Thus before fitting the full spatial SMK-GP, we turn our attention to the non-spatial conjugate Bayesian linear model (1) that yields closed form joint posterior distributions for (β, σ^2) belonging to the NIG family of distributions. It is instructive to see the accuracy of approximation offered by meta posterior of β in comparison with the exact posterior distribution

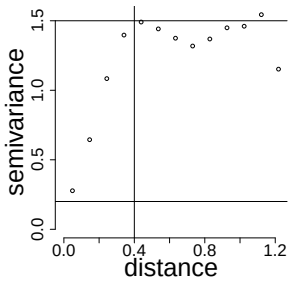
of β from the non-spatial NIG model on this dataset. Similar to Section 3.1, the sea surface temperature dataset is divided into 40 subsets and exact posterior quantiles from each component of β is plotted (see Figure 8) with corresponding posterior quantiles from the meta posterior. The quantiles from the exact and meta posterior are found to be indistinguishable for any practical purpose. Our investigation reveals that 25%, 50% and 75% quantiles of SMK almost always coincide with the corresponding quantiles of full posterior even with large number of subsets. However, the tail of meta posterior is little more spread out than the tail of the full posterior with large number of subsets. Quantiles in both extremes tend to match for full and meta posterior as the number of subsets decreases. Analysis on non-spatial model justifies usage of meta posterior as a computationally convenient approximation to the full posterior. Next we move to the more complex spatial analysis of the data and judge performance of meta posterior when the Gaussian process model is fitted to each subset.



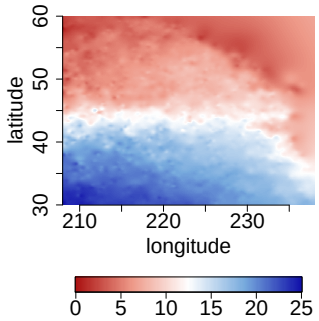
(a) β_0 (b) β_1 (c) β_2

Figure 8: Posterior quantiles of full posterior and meta posterior: (a) β_0; (b) β_1; (c) β_2.

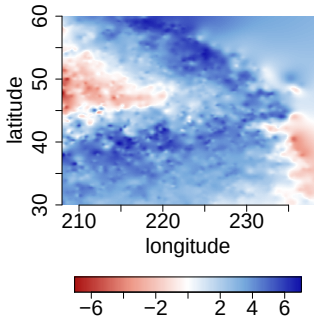
Memory in our workstation was insufficient to store the $N \times N$ distance matrix to run the full GP model and the tapered Gaussian process model. Other popular methods such as the treed Gaussian process on the full data takes a long time to run and is deemed impractical as a competitor for the dataset of interest. Subsequently, we fit SMK-GP and BISP for various choices of K. Additionally, laGP and LatticeKrig are fitted as competitors of SMK to study predictive inference. For brevity, results of SMK-GP are presented for $K = 40$ and $K = 60$ subsets.



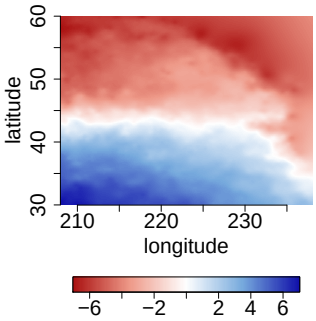
(a) Empirical semivariogram



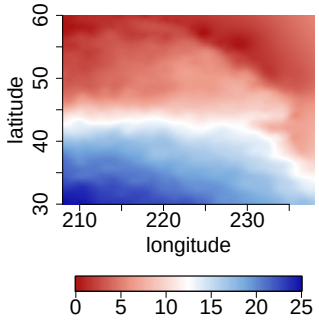
(b) Sea surface temperature



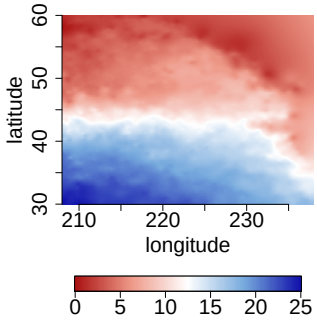
(c) OLS residual



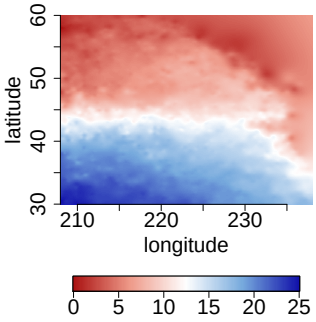
(d) SMK predicted surface ($K = 40$)



(e) SMK predicted surface ($K = 60$)



(f) laGP predicted surface



(g) LatticeKrig predicted surface

Figure 9: Figure 9(b) shows the plot of the sea surface temperature data. Estimated OLS residual from the non-spatial model is presented in Figure 9(c). Figure 9(a) presents the empirical semivariogram for OLS residuals. Figures 9(d) and 9(e) show estimated residual spatial surfaces for SMK-GP fitted with $K = 40$ and $K = 60$ respectively. Figures 9(f), 9(g) present interpolated surfaces for laGP and LatticeKrig respectively. x and y axes in every figure represent Longitude and Latitude in the scale of [0,360] and [0,90] respectively.

All spatial locations are transformed to lie in $[0,1] \times [0,1]$ intervals for our analysis. For all the competing models, the intercept is assigned a flat prior and τ^2 and σ^2 are assigned an IG(2,1) prior. The spatial range parameter is assigned a $U(0.3, 300)$ prior that ensures huge support given that the transformed coordinates belong to $[0,1] \times [0,1]$ domain. Parameter estimates along with their estimated 95% credible intervals for SMK with $K = 40$ and 60 are presented in Table 4. Both of them yield high estimates of the signal to noise ratio $\frac{\sigma^2}{\tau^2}$, which suggests a sophisticated spatial model to capture looming spatial dependence of the sea surface temperature.

Table 4: Parameter credible intervals, 50 (2.5 97.5) percentiles for all the parameters. SMK with Gaussian process is fitted for $K = 40, 60$ number of subsets

	SMK (Gaussian Process)	
Parameter	$K = 40$	$K = 60$
β_0	23.98 (18.66, 28.27)	23.93 (19.04, 28.80)
β_1	-4.95 (-9.03,0.16)	-4.84 (-8.75,0.10)
β_2	-14.33 (-17.41,-10.17)	-14.18 (-17.09,-11.32)
τ^2	0.11 (0.08,0.13)	0.09 (0.07, 0.12)
σ^2	12.33 (8.39, 16.95)	10.56 (7.24, 15.98)
ϕ	0.37 (0.30,0.91)	0.38 (0.30, 0.88)
time (in min)	643.2	213

Table 5: Mean squared prediction error (MSPE), length and coverage of 95% predictive intervals of SMK-GP ($K = 40, 60$), BISP ($K = 40, 60$), laGP and LatticeKrig.

	SMK-GP ($K = 60$)	SMK-GP ($K = 40$)	BISP ($K = 60$)	BISP ($K = 40$)	laGP	LatticeKrig
MSPE	0.13	0.11	0.16	0.15	0.10	0.09
Length of 95% PI	2.31	1.70	1.65	1.58	1.18	0.41
Coverage of 95% PI	0.97	0.97	0.93	0.93	0.94	0.62

Predictive power of the proposed architecture, along with the other approaches, is assessed based on MSPE, coverage and length of 95% predictive intervals. The non-spatial model, SMK-GP with $K = 60$ and SMK-GP with $K = 40$ yield MSPE $1.31, 0.13$ and 0.11 respectively. Such dramatic improvement in MSPE for spatial models (shown in Table 4) corroborate the strong spatial story inherent in the data. Further, Table 5 demonstrates about 30% improvement in terms of mean squared prediction error for SMK over BISP. Additionally, BISP suffers from a little under-coverage, presumably due to simplifying the

correlation structure among different subsets that fails to capture complex spatial association. laGP and SMK-GP demonstrate almost indistinguishable performance in terms of MSPE but vary in characterizing predictive uncertainty. LatticeKrig turns out to be the superior performer in terms of point prediction, but suffers heavily in characterizing predictive uncertainty. Overall, SMK-GP positions itself as a competitive performer in predictive inference. Predictive surfaces in Figure 9 further corroborate this fact. Importantly, the in-built parallel structure in SMK leads to full Bayesian inference and prediction in approximately 4 hours and 10 hours (with parallel implementation) for $K = 60$ and 40 partitions respectively. Fitting SMK-GP beyond $K = 40$ unnecessarily exacerbates computational burden with minimal improvement of inferential and predictive performance.

4 Conclusion and Future Work

This article has developed a practical approximation to Bayesian spatial inference for "big-N" problems. We propose dividing big data into multiple subsets, carrying out independent inference in each subset followed by combining inference from all subsets. The entire procedure is "trivially parallelizable", offers rapid computation for big data and also eliminates the need to store the entire dataset in one processor. Given the dramatic computational and storage gains, we expected to pay some price in terms of spatial inference, but were pleasantly surprised that this price seemed to be acceptably small in most cases. Further, SMK formulation provides a generic "divide and conquer" algorithm that is potentially useful for any choice of the spatial model for data subsets. As a first step to launch SMK, this article implements SMK algorithm on stationary Gaussian process and tapered Gaussian process models. We demonstrate competitive predictive performance of SMK-GP with state-of-the-art models. Unlike many other state-of-the-art models, SMK-GP provides full scale Bayesian inference and that too within manageable time. The potential of SMK-GP or SMK-TGP are best understood by acknowledging the fact that these are fast and accurate approximations of stationary GP or TGP for big data.

SMK opens a number of future research directions. It remains to investigate the quality of approximation "meta posterior" offers to full posterior obtained from a non-stationary spatial process model. A potential concern with the current specification of SMK is that if

the spatial random process generating the data has substantial non-stationary local behavior, SMK might miss a few local features of the spatial process. This is due to the fact that most of the subset posteriors miss some important local behavior of the spatial random process if samples are sparsely drawn from each subset. Of course, one can increase the number of data points in each subset to improve performance, compromising a bit on the scalability. We are currently developing an extension of SMK that is able to provide approximation to a few non-stationary moderately scalable spatial models to deliver full scale Bayesian inference from a dataset with 3 million observations and having significant local behavior. The extension also allows us to match a few prefixed quantiles of parameters in the "meta posterior" with the corresponding quantiles of the full posterior. Thus, uncertainty characterization is found to be more exact with the proposed extension to SMK.

It is also of natural interest to simply extend the proposed spatial SMK to large scale spatio-temporal analysis. We intend to explore the possibility of scalable inference in two different situations: (a) when spatio-temporal interpolation is sought at discrete time-points (e.g., monthly or yearly data), and (b) when spatiotemporal interpolation is sought at arbitrary locations and timepoints. It remains an important question as to how one should partition spatio-temporal data for SMK to capture both spatial and temporal associations.

References

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

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

Berliner, L. M., C. K. Wikle, and N. Cressie (2000). Long-lead prediction of pacific ssts via bayesian dynamic modeling. *Journal of climate 13*(22), 3953–3968.

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

Cressie, N. and C. K. Wikle (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Datta, A., S. Banerjee, A. O. Finley, and A. E. Gelfand (2015). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association* (just-accepted), 00–00.

Du, J., H. Zhang, V. Mandrekar, et al. (2009). Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *the Annals of Statistics 37*(6A), 3330–3361.

Eidsvik, J., B. A. Shaby, B. J. Reich, M. Wheeler, and J. Niemi (2014). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics 23*(2), 295–315.

Finley, A. O., S. Banerjee, P. Waldmann, and T. Ericsson (2009). Hierarchical spatial modeling of additive and dominance genetic variance for large spatial trial datasets. *Biometrics 65*(2), 441–451.

Furrer, R., M. G. Genton, and D. Nychka (2012). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*.

Gelfand, A. E., S. Banerjee, C. Sirmans, Y. Tu, and S. E. Ong (2007). Multilevel modeling using spatial processes: Application to the singapore housing market. *Computational Statistics & Data Analysis 51*(7), 3567–3579.

Gelfand, A. E., P. Diggle, P. Guttorp, and M. Fuentes (2010). *Handbook of spatial statistics*. CRC press.

Gramacy, R. B. and D. W. Apley (2015). Local gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics 24*(2), 561–578.

Gramacy, R. B. and H. K. Lee (2012). Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*.

Guhaniyogi, R., A. O. Finley, S. Banerjee, and A. E. Gelfand (2011). Adaptive gaussian predictive process models for large spatial datasets. *Environmetrics 22*(8), 997–1007.

Guinness, J. (2016). Permutation methods for sharpening gaussian process approximations. *arXiv preprint arXiv:1609.05372*.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics 5*(2), 173–190.

Katzfuss, M. (2016). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association* (just-accepted).

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

Lemos, R. T. and B. Sansó (2006). Spatio-temporal variability of ocean temperature in the portugal current system. *Journal of Geophysical Research: Oceans 111*(C4).

Lemos, R. T. and B. Sansó (2009). A spatio-temporal model for mean, anomaly, and trend fields of north atlantic sea surface temperature. *Journal of the American Statistical Association 104*(485), 5–18.

Minsker, S. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli 21*(4), 2308–2335.

Minsker, S., S. Srivastava, L. Lin, and D. B. Dunson (2014). Robust and scalable bayes via a median of subset posterior measures. *arXiv preprint arXiv:1403.2660*.

Nychka, D., S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain (2015). A multiresolution gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics 24*(2), 579–599.

Park, C. and D. Apley (2017). Patchwork kriging for large-scale gaussian process regression. *arXiv preprint arXiv:1701.06655*.

Robert, C. and G. Casella (2009). *Introducing Monte Carlo Methods with R*. Springer Science & Business Media.

Rue, H., S. Martino, and N. Chopin (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology) 71*(2), 319–392.

Sang, H. and J. Z. Huang (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(1), 111–132.

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

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Stein, M. L., Z. Chi, and L. J. Welty (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(2), 275–296.

Stroud, J. R., M. L. Stein, and S. Lysen (2017). Bayesian and maximum likelihood estimation for gaussian processes on an incomplete lattice. *Journal of Computational and Graphical Statistics 26*(1), 108–120.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 297–312.

Wendland, H. (2004). *Scattered data approximation*, Volume 17. Cambridge university press.

Wikle, C. K. (2010). Low-rank representations for spatial processes. *Handbook of Spatial Statistics*, 107–118.

Zhang, R., C. D. Lin, and P. Ranjan (2016). Local gaussian process model for large-scale dynamic computer experiments. *arXiv preprint arXiv:1611.09488*.