Submitted to Statistical Science

Distributed Bayesian Inference in Massive Spatial Data

Rajarshi Guhaniyogi[*]
Department of Statistics, Texas A & M University, College Station, Texas, U.S.A.

Cheng Li[†]
Department of Statistics and Data Science, National University of Singapore, Singapore

Terrance Savitsky[‡]
U.S. Bureau of Labor Statistics, Washington D.C., U.S.A.

Sanvesh Srivastava[§]
Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, Iowa, U.S.A.

Abstract. Gaussian process (GP) regression is computationally expensive in spatial applications involving massive data. Various methods address this limitation, including a small number of Bayesian methods based on distributed computations (or the divide-and-conquer strategy). Focusing on the latter literature, we achieve three main goals. First, we develop an extensible Bayesian framework for distributed spatial GP regression that embeds many popular methods. The proposed framework has three steps that partition the entire data into many subsets, apply a readily available Bayesian spatial process model in parallel on all the subsets, and combine the posterior distributions estimated on all the subsets into a pseudo posterior distribution that conditions on the entire data. The combined pseudo posterior distribution replaces the full data posterior distribution in prediction and inference problems. Demonstrating our framework's generality, we extend posterior computations for (non-distributed) spatial process models with a stationary full-rank and a nonstationary low-rank GP priors to the distributed setting. Second, we contrast the empirical performance of popular distributed approaches with some widely used non-distributed alternatives and highlight their relative advantages and shortcomings. Third, we provide theoretical support for our numerical observations and show that the Bayes L_2-risks of the combined posterior distributions obtained from a subclass of the divide-and-conquer methods achieves the near-optimal convergence rate in estimating the true spatial surface with various types of covariance functions. Additionally, we provide upper bounds on the number of subsets to achieve these near-optimal rates.

Key words and phrases: Distributed Bayesian inference, Gaussian process, low-rank Gaussian process, massive spatial data, Wasserstein barycenter.

1. INTRODUCTION

A fundamental challenge in geostatistics is the analysis of massive spatially-referenced data. Such

[*]rajguhaniyogi@tamu.edu
[†]stalic@nus.edu.sg
[‡]savitsky.terrance@bls.gov
[§]sanvesh-srivastava@uiowa.edu Corresponding author.

data sets provide scientists with an unprecedented opportunity to hypothesize and test complex theories, see for example Cressie and Wikle (2011), Banerjee et al. (2014). This has led to the development of complex and flexible GP-based models that are computationally intractable for a large number of spatial locations, denoted as n, due to the $O(n^3)$ computational cost and the $O(n^2)$ storage cost. An overwhelming number of methods exists to address this issue that develop either efficient alternatives to the GP model or efficient approximations of the likelihood. We broadly refer to these approaches as the *non-distributed* methods. An emerging class of Bayesian methods addresses this problem using distributed computations, where the scalability of an existing, possibly non-distributed, spatial GP regression model is enhanced multiple folds by suitably distributing the computations and storage of data subsets across many machines. This article proposes a novel class of distributed Bayesian framework for process-based geostatistical models that contains many popular approaches, presents a comparative study of important approaches within this class, and contrasts their performance with representative non-distributed methods.

1.1 Non-distributed Methods for GP Modeling of Massive Spatial Data

Efficient GP-based models for massive spatial data have received extensive attention due to their great practical importance (Heaton et al., 2019). A common idea in GP-based modeling is to seek dimension-reduction by endowing the spatial covariance matrix either with a low-rank or a sparse structure. Low-rank structures on the spatial covariance matrix are the most widely used tool for efficient spatial computation. They represent the spatial surface using r *apriori* chosen basis functions with associated computational complexity of $O(nr^2+r^3)$ (Cressie and Johannesson, 2008, Banerjee et al., 2008, Finley et al., 2009, Guhaniyogi et al., 2011, Wikle, 2010); however, a major shortcoming of the above methods is that a small (r/n)-ratio yields inaccurate GP approximations, resulting in the propensity to oversmooth the spatial data (Stein, 2014, Simpson et al., 2012).

A specific form of sparse structure, called covariance tapering, uses compactly supported covariance functions to create sparse spatial covariance matrices that approximate the full covariance matrix (Kaufman et al., 2008, Furrer et al., 2006, Daley et al., 2015, Bevilacqua et al., 2022). Covariance tapering still requires expensive determinant evaluation of the massive covariance matrix, and the choice of the taper range can be difficult for spatial data over irregularly spaced locations (Anderes et al., 2013). An alternative approach is to introduce sparsity in the inverse covariance (precision) matrix of the GP likelihoods using products of lower dimensional conditional distributions (Vecchia, 1988, Rue et al., 2009, Stein et al., 2004), or via composite likelihoods (Eidsvik et al., 2014, Bai et al., 2012, Bevilacqua and Gaetan, 2015). Extending these ideas, recent approaches introduce sparsity in the inverse covariance (precision) matrix of process realizations and hence enable "kriging" at arbitrary locations (Datta et al., 2016, Guinness, 2018, Finley et al., 2019). In related literature on computer experiments, localized approximations of GP models are proposed; see, for example, Gramacy and Apley (2015), Gramacy and Haaland (2016). These methods scale well with the sample size and are able to capture local spatial variations.

The remaining variants of dimension-reduction methods combine the benefits of low-rank and sparse covariance functions. Examples include non-stationary models (Banerjee et al., 2014) and multi-resolution models (Nychka et al., 2015, Katzfuss, 2017, Guinness, 2021, Katzfuss and Guinness, 2021, Guhaniyogi and Sanso, 2020). Multi-resolution models are difficult to implement and lack large sample theoretical guarantees, but they successfully capture spatial variation at multiple scales and are computationally efficient. The GP with Matérn covariance can be viewed as the solution of a stochastic partial differential equation. This observation has motivated GP approximations (Lindgren et al., 2011, Bolin and Lindgren, 2013), including a recent extension to multivariate non-Gaussian models with marginal Matérn covariance functions (Bolin and Wallin, 2020). This class of methods work well for Matérn covariance functions but are inapplicable in scaling GP with low-rank or non-stationary covariance functions.

1.2 Distributed Bayes

Rooted in the divide-and-conquer technique, the distributed Bayesian methods do not belong to any of the classes of methods in Section 1.1. They instead fit an existing model on different data subsets exploiting the distributed computing architecture. The results from the subsets are combined using an aggregation algorithm. These methods were first proposed in machine learning, including Consensus Monte Carlo (Scott et al., 2016), the Weierstrass sampler (Wang and Dunson, 2013), the semiparametric density product (Neiswanger et al., 2014), the median posterior (Minsker et al., 2014) and the Wasserstein posterior (Srivastava et al., 2015). Most of these methods are developed only for independent data. Recently, distributed Bayes has been applied to a variety of statistical problems in both modeling and computation, such as density estimation (Su, 2020), modeling of multivariate binary data (Mehrotra et al., 2021), sequential Monte Carlo (Lindsten et al., 2017), random partition trees (Wang et al., 2015), etc. For GP models, Zhang and Williamson (2019) proposes to combine GP fitted on different data subsets via an importance-sampled mixture-of-experts model. Theoretical results on distributed GP inference have been developed recently (Cheng and Shang, 2017, Szabo and van Zanten, 2019, Shang et al., 2019). Nevertheless, these theoretical works and applications have mainly focused on univariate domains for nonparametric regression and have not considered the GP-based models used in spatial applications such as GP with Matérn covariance on a spatial domain.

On the spatial front, Barbian and Assunção (2017) propose combining point estimates of spatial parameters obtained from different subsets, but they do not provide combined inference on the spatial processes or predictions. Similarly, Heaton et al. (2017) partition the spatial domain and assume independence between the data in different partitions. Guhaniyogi and Banerjee (2018, 2019) propose the idea of "meta-posterior," a computationally efficient approximation to the full data posterior. This approach does not assume independence across data blocks and enables accurate prediction with uncertainty (Heaton et al., 2019); however, Guhaniyogi and Banerjee (2018) does not offer any theoretical guidance on choosing the number of subsets for optimal inference on the spatial surface.

Instead of developing a new spatial GP regression model, we describe a general class of three-step distributed Bayesian approaches for extending an existing process-based geostatistical model, which includes a number of important special cases. To implement the general approach, the n spatial locations are divided into k subsets such that each subset has representative data samples from all regions of the spatial domain with the jth subset containing m_j data samples. Second, posterior computations are implemented in parallel on the k subsets using any chosen spatial process model after raising the model likelihood to a power of n/m_j in the jth subset. The pseudo posterior distribution obtained using the modified likelihood is called the "subset pseudo posterior distribution". Since jth subset pseudo posterior distribution conditions on (m_j/n)-fraction of the full data, the modification of the likelihood by raising it to the power of n/m_j ensures that variance of each subset pseudo posterior is of the same order (as a function of n) as that of the full data posterior distribution. Third, the k subset pseudo posterior distributions are combined into a single pseudo probability distribution, called the combined pseudo posterior, that conditions on the full data and replaces the computationally expensive full data posterior distribution for prediction and inference. Our distributed framework leverages existing spatial GP regression models and enhances their scalability by embedding them within the three-step framework. For example, Section 3.1 embeds full-rank and low-rank spatial GP regression models within the distributed framework and Section 3.3 discusses various methods for combining the k subset pseudo posteriors.

The proposed framework builds on the recent works that combine the subset pseudo posterior distributions through their geometric centers (e.g., mean, median) and guarantee wide applicability under general assumptions (Minsker et al., 2014, Srivastava et al., 2015, Li et al., 2017, Minsker et al., 2017, Savitsky and Srivastava, 2018, Srivastava et al., 2018, Minsker, 2019, Wang and Srivastava, 2021). The theory and practice of such distributed approaches are limited to parametric models. In contrast, the framework proposed here is tuned for accurate and computationally efficient

posterior inference in nonparametric Bayesian models based on GP priors. In particular, we develop a new approach to modify the likelihood for computing the subset pseudo posterior distribution of an unknown function, an infinite-dimensional parameter, that subsumes the parametric distributed methods. We offer novel theoretical results on the convergence rate of the combined pseudo posterior to the true function. Finally, we also provide guidance on choosing k depending on the covariance function and n such that the combined pseudo posterior maintains near minimax optimal performance as $n \to \infty$. The proposed distributed framework delivers principled Bayesian inference and predictions without any restrictive data- or model-specific assumptions, such as the independence between data subsets or independence between blocks of parameters.

A related focus of this article is to present a comparative study of the proposed class of distributed approaches with important non-distributed approaches for modeling massive spatial data. We illustrate the application of the distributed framework for enhancing the scalability of spatial models with a low-rank non-stationary GP prior called the modified predictive process (MPP) prior (Finley et al., 2009). This prior is commonly used for estimating spatial surfaces in applications with massive sample size, but it struggles to provide accurate inference in a manageable time beyond (approximately) 10^4 observations. We embed MPP within our distributed framework and scale it to spatial applications of much bigger sizes and assess its performance relative to other distributed and state-of-the-art non-distributed alternatives for efficient spatial GP modeling. Unfortunately, there is no theoretical guarantee for convergence of the Markov chain to its stationary distribution, where MCMC samples are drawn from the subset pseudo posteriors with an MPP prior on spatial effects; however, we find strong empirical evidence for it and propose to develop the theoretical support elsewhere.

2. BAYESIAN INFERENCE IN GP-BASED SPATIAL MODELS

Consider the model for the data observed at location $\mathbf{s}$ in a compact domain $\mathcal{D}$,

$$y(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T \boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{1}$$

where $y(\mathbf{s})$ and $\mathbf{x}(\mathbf{s})$ are the response and a $p \times 1$ predictor vector respectively at $\mathbf{s}$, $\boldsymbol{\beta}$ is a $p \times 1$ predictor coefficient, $w(\mathbf{s})$ is the value of an unknown spatial function $w(\cdot)$ at $\mathbf{s}$, and $\epsilon(\mathbf{s})$ is the value of a white-noise process $\epsilon(\cdot)$ at $\mathbf{s}$, which is independent of $w(\cdot)$. The Bayesian implementation of the model in (1) customarily assumes that (a) $\boldsymbol{\beta}$ apriori follows $\mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ and (b) $w(\cdot)$ and $\epsilon(\cdot)$ apriori follow mean 0 GPs with covariance functions $C_{\boldsymbol{\alpha}}(\mathbf{s}_1, \mathbf{s}_2)$ and $D_{\boldsymbol{\alpha}}(\mathbf{s}_1, \mathbf{s}_2)$ that model $\mathrm{cov}\{w(\mathbf{s}_1), w(\mathbf{s}_2)\}$ and $\mathrm{cov}\{\epsilon(\mathbf{s}_1), \epsilon(\mathbf{s}_2)\}$, respectively, where $\boldsymbol{\alpha}$ are the process parameters indexing the two families of covariance functions and $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}$; therefore, the parameters are $\boldsymbol{\Omega} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$. The training data consists of predictors and responses observed at n spatial locations, denoted as $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$.

Standard Markov chain Monte Carlo (MCMC) algorithms exist for performing posterior inference on $\boldsymbol{\Omega}$ and $w(\cdot)$ at a set of locations $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_l^*\}$, where $\mathcal{S}^* \cap \mathcal{S} = \emptyset$, and for predicting $y(\mathbf{s}^*)$ for any $\mathbf{s}^* \in \mathcal{S}^*$ (Banerjee et al., 2014). Given $\mathcal{S}$, the prior assumptions on $w(\cdot)$ and $\epsilon(\cdot)$ imply that $\mathbf{w}^T = \{w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n)\}$ and $\boldsymbol{\epsilon}^T = \{\epsilon(\mathbf{s}_1), \ldots, \epsilon(\mathbf{s}_n)\}$ are independent and follow $N\{\mathbf{0}, \mathbf{C}(\boldsymbol{\alpha})\}$ and $N\{\mathbf{0}, \mathbf{D}(\boldsymbol{\alpha})\}$, respectively, with the (i,j)th entries of $\mathbf{C}(\boldsymbol{\alpha})$ and $\mathbf{D}(\boldsymbol{\alpha})$ are $C_{\boldsymbol{\alpha}}(\mathbf{s}_i, \mathbf{s}_j)$ and $D_{\boldsymbol{\alpha}}(\mathbf{s}_i, \mathbf{s}_j)$, respectively. The hierarchy in (1) is completed by assuming that $\boldsymbol{\alpha}$ apriori follows a distribution with density $\pi(\boldsymbol{\alpha})$. The MCMC algorithm for sampling $\boldsymbol{\Omega}$, $\mathbf{w}^{*T} = \{w(\mathbf{s}_1^*), \ldots, w(\mathbf{s}_l^*)\}$, and $\mathbf{y}^{*T} = \{y(\mathbf{s}_1^*), \ldots, y(\mathbf{s}_l^*)\}$ cycle through the following three steps until sufficient MCMC samples are drawn post convergence:

1. Integrate over $\mathbf{w}$ in (1) and
   (a) sample $\boldsymbol{\beta}$ given $\mathbf{y}, \mathbf{X}, \boldsymbol{\alpha}$ from $N(\mathbf{m}_{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}})$, where

$$\mathbf{V}_{\boldsymbol{\beta}} = \left\{ \mathbf{X}^T \mathbf{V}(\boldsymbol{\alpha})^{-1} \mathbf{X} + \Sigma_{\boldsymbol{\beta}}^{-1} \right\}^{-1}, \tag{2}$$
$$\mathbf{m}_{\boldsymbol{\beta}} = \mathbf{V}_{\boldsymbol{\beta}} \left\{ \mathbf{X}^T \mathbf{V}(\boldsymbol{\alpha})^{-1} \mathbf{y} + \Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} \right\},$$

   $\mathbf{X} = [\mathbf{x}(\mathbf{s}_1) : \cdots : \mathbf{x}(\mathbf{s}_n)]^T$ is the $n \times p$ matrix of predictors, with $p < n$, $\mathbf{V}(\boldsymbol{\alpha}) = \mathbf{C}(\boldsymbol{\alpha}) + \mathbf{D}(\boldsymbol{\alpha})$; and
   (b) sample $\boldsymbol{\alpha}$ given $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}$ using the Metropolis-Hastings algorithm with a normal random walk proposal.

2. Sample $\mathbf{w}^*$ given $\mathbf{y}, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ from $N(\mathbf{m}_*, \mathbf{V}_*)$, where

$$\begin{aligned} \mathbf{V}_* &= \mathbf{C}_{*,*}(\boldsymbol{\alpha}) - \mathbf{C}_*(\boldsymbol{\alpha})\,\mathbf{V}(\boldsymbol{\alpha})^{-1}\,\mathbf{C}_*(\boldsymbol{\alpha})^T, \\ \mathbf{m}_* &= \mathbf{C}_*(\boldsymbol{\alpha})\,\mathbf{V}(\boldsymbol{\alpha})^{-1}(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta}), \end{aligned} \tag{3}$$

$\mathbf{C}_*(\boldsymbol{\alpha})$ and $\mathbf{C}_{*,*}(\boldsymbol{\alpha})$ are $l \times n$ and $l \times l$ matrices, respectively, and the (i,j)th entries of $\mathbf{C}_{*,*}(\boldsymbol{\alpha})$ and $\mathbf{C}_*(\boldsymbol{\alpha})$ are $C_{\boldsymbol{\alpha}}(\mathbf{s}_i^*, \mathbf{s}_j^*)$ and $C_{\boldsymbol{\alpha}}(\mathbf{s}_i^*, \mathbf{s}_j)$, respectively.

3. Sample $\mathbf{y}^*$ given $\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}^*$ from $N\{\mathbf{X}^*\boldsymbol{\beta} + \mathbf{w}^*, \mathbf{D}(\boldsymbol{\alpha})\}$, where $\mathbf{X}^{*T} = [\mathbf{x}(\mathbf{s}_1^*) : \cdots : \mathbf{x}(\mathbf{s}_l^*)]$.

Many spatial models can be formulated in terms of (1) by assuming different forms of $C_{\boldsymbol{\alpha}}(\mathbf{s}_1, \mathbf{s}_2)$ and $D_{\boldsymbol{\alpha}}(\mathbf{s}_1, \mathbf{s}_2)$; see Banerjee et al. (2014) and supplementary material for details on the MCMC algorithm. Irrespective of the form of $\mathbf{D}(\boldsymbol{\alpha})$, if no additional assumptions are made on the structure of $\mathbf{C}(\boldsymbol{\alpha})$, then the three steps require $O(n^3)$ flops in computation and $O(n^2)$ memory units in storage in every MCMC iteration. Spatial models with this form of posterior computations are based on a *full-rank* GP prior, which are infeasible to compute for big data.

There are methods which either impose a low-rank structure or a sparse structure on $\mathbf{C}(\boldsymbol{\alpha})$ to address this computational issue (Banerjee et al., 2014). Methods with a low-rank structure on $\mathbf{C}(\boldsymbol{\alpha})$ expresses $\mathbf{C}(\boldsymbol{\alpha})$ in terms of $r \ll n$ basis functions, in turn inducing a *low-rank* GP prior. Again, a class of sparse structure uses compactly supported covariance functions to create $\mathbf{C}(\boldsymbol{\alpha})$ with overwhelming zero entries (Kaufman et al., 2008, Furrer et al., 2006), where as another variety of sparse structure imposes a Markov random field model on the joint distribution of $\mathbf{y}$ (Vecchia, 1988, Rue et al., 2009, Stein et al., 2004) or $\mathbf{w}$ (Datta et al., 2016, Guinness, 2018). We use the MPP prior as a representative example of this broad class of computationally efficient methods. Let $\mathcal{S}^{(0)} = \{\mathbf{s}_1^{(0)}, ..., \mathbf{s}_r^{(0)}\}$ be a set of r locations, known as the "knots," which may or may not intersect with $\mathcal{S}$. Let $\mathbf{c}(\mathbf{s}, \mathcal{S}^{(0)}) = \{C_{\boldsymbol{\alpha}}(\mathbf{s}, \mathbf{s}_1^{(0)}), \ldots, C_{\boldsymbol{\alpha}}(\mathbf{s}, \mathbf{s}_r^{(0)})\}^T$ be an $r \times 1$ vector and $\mathbf{C}(\mathcal{S}^{(0)})$ be an $r \times r$ matrix whose (i,j)th entry is $C_{\boldsymbol{\alpha}}(\mathbf{s}_i^{(0)}, \mathbf{s}_j^{(0)})$. Using $\mathbf{c}(\mathbf{s}_1, \mathcal{S}^{(0)}), \ldots, \mathbf{c}(\mathbf{s}_n, \mathcal{S}^{(0)})$ and $\mathbf{C}(\mathcal{S}^{(0)})$, define the diagonal matrix $\boldsymbol{\delta} = \text{diag}\{\delta(\mathbf{s}_1), \ldots, \delta(\mathbf{s}_n)\}$ with $\delta(\mathbf{s}_i) = C_{\boldsymbol{\alpha}}(\mathbf{s}_i, \mathbf{s}_i) - \mathbf{c}^T(\mathbf{s}_i, \mathcal{S}^{(0)})\,\mathbf{C}(\mathcal{S}^{(0)})^{-1}\,\mathbf{c}(\mathbf{s}_i, \mathcal{S}^{(0)})$, $i = 1, \ldots, n$. Let $\mathbf{1}(\mathbf{a} = \mathbf{b}) = 1$ if $\mathbf{a} = \mathbf{b}$ and 0 otherwise. Then, MPP is a GP with covariance function

$$\begin{aligned} \tilde{C}_{\boldsymbol{\alpha}}(\mathbf{s}_1, \mathbf{s}_2) = \mathbf{c}^T(\mathbf{s}_1, \mathcal{S}^{(0)})\,\mathbf{C}(\mathcal{S}^{(0)})^{-1}\,\mathbf{c}(\mathbf{s}_2, \mathcal{S}^{(0)}) \\ + \delta(\mathbf{s}_1)\,\mathbf{1}(\mathbf{s}_1 = \mathbf{s}_2), \end{aligned} \tag{4}$$

where $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}$, $\tilde{C}_{\boldsymbol{\alpha}}(\mathbf{s}_1, \mathbf{s}_2)$ depends on the covariance function of the parent GP and the selected r knots, which define $\mathbf{C}(\mathcal{S}^{(0)})$, $\mathbf{c}^T(\mathbf{s}_1, \mathcal{S}^{(0)})$, and $\mathbf{c}^T(\mathbf{s}_2, \mathcal{S}^{(0)})$. We have used a $\tilde{}$ in (4) to distinguish the covariance function of a low-rank GP prior from that of its parent full-rank GP. If $\tilde{\mathbf{C}}(\boldsymbol{\alpha})$ is a matrix with (i,j)th entry $\tilde{C}_{\boldsymbol{\alpha}}(\mathbf{s}_i, \mathbf{s}_j)$, then the posterior computations using MPP, a low-rank GP prior, replace $\mathbf{C}(\boldsymbol{\alpha})$ by $\tilde{\mathbf{C}}(\boldsymbol{\alpha})$ in the steps 1(a), 1(b), and 2. The (low) rank r structure imposed by $\mathbf{C}(\mathcal{S}^{(0)})$ implies that $\tilde{\mathbf{C}}(\boldsymbol{\alpha})^{-1}$ computation requires $O(nr^2)$ flops using the Woodbury formula (Harville, 1997); however, massive spatial data require that $r = O(\sqrt{n})$, leading to the computational inefficiency of low-rank methods.

The next section discusses a general three-step distributed framework to scale the posterior computations in spatial GP regression models with full-rank and low-rank GP priors. Briefly, the first and second steps divide the full data and fit a low-rank or full-rank spatial GP regression model on each data subset after modifying the subset likelihood, respectively, and the third step combines draws from the all the subset pseudo posteriors. We discuss a few popular alternatives for combining draws from the subset pseudo posteriors and offer novel convergence rate results for an important subclass of combination approaches.

3. DISTRIBUTED FRAMEWORK FOR BAYESIAN INFERENCE IN SPATIAL REGRESSION MODELS

3.1 First Step: Partitioning of Spatial Locations

We partition the n spatial locations into k non-overlapping subsets. The default partitioning scheme is to randomly allocate the locations into k possibly non-overlapping subsets (referred to as the random partitioning scheme hereon) to ensure that each subset has representative data samples from all subregions of the domain. We provide discussion on the choice of k later.

Let $\mathcal{S}_j = \{\mathbf{s}_{j1}, \ldots, \mathbf{s}_{jm_j}\}$ denote the set of m_j spatial locations in subset j $(j = 1, \ldots, k)$. Conceptually, a spatial location can belong to multiple subsets, though for this work we have assumed disjoint subsets, so that $\sum_{j=1}^{k} m_j = n$ and $\cup_{j=1}^{k} \mathcal{S}_j = \mathcal{S}$, where $\mathbf{s}_{ji} = \mathbf{s}_{i'}$ for some $\mathbf{s}_{i'} \in \mathcal{S}$ and for every $i = 1, \ldots, m_j$ and $j = 1, \ldots, k$. Denote the data in the jth partition as $\{\mathbf{y}_j, \mathbf{X}_j\}$ $(j = 1, \ldots, k)$, where $\mathbf{y}_j = \{y(\mathbf{s}_{j1}), \ldots, y(\mathbf{s}_{jm_j})\}^T$ is a $m_j \times 1$ vector and $\mathbf{X}_j = [\mathbf{x}(\mathbf{s}_{j1}) : \cdots : \mathbf{x}(\mathbf{s}_{jm_j})]^T$ is a $m_j \times p$ matrix of predictors corresponding to the spatial locations in $\mathcal{S}_j$ with $p < m_j$. In modern grid or cluster computing environments, all the machines in the network have similar computational power, so the performances of distributed Bayesian methods are optimized by choosing similar values of $m_1, \ldots, m_k$.

One can choose more sophisticated partitioning schemes than random partitioning. For example, it is possible to cluster the data based on centroid clustering (Knorr-Held and Raßer, 2000) or hierarchical clustering based on spatial gradients (Anderson et al., 2014, Heaton et al., 2017), and then construct subsets such that each subsets contains representative data samples from each cluster. Detailed exploration later shows that even random partitioning leads to desirable inference in the various simulation settings and in the sea surface data example. Perhaps more sophisticated blocking methods may provide further improvement in the cases where spatial locations are drawn based on specific designs; for example, sophisticated partitioning schemes have inferential benefits when a sub-domain shows substantial local behavior compared to the others (Guhaniyogi and Sanso, 2020), or sampled locations are chosen based on a specific survey design. Since they are atypical examples in the spatial context, we will pursue them elsewhere.

The univariate spatial GP regression model for any location $\mathbf{s}_{ji} \in \mathcal{S}_j \subset \mathcal{D}$ is

$$y(\mathbf{s}_{ji}) = \mathbf{x}(\mathbf{s}_{ji})^T \boldsymbol{\beta} + w(\mathbf{s}_{ji}) + \epsilon(\mathbf{s}_{ji}), \quad i = 1, \ldots, m_j. \tag{5}$$

Let $\mathbf{w}_j^T = \{w(\mathbf{s}_{j1}), \ldots, w(\mathbf{s}_{jm_j})\}$ and $\boldsymbol{\epsilon}_j^T = \{\epsilon(\mathbf{s}_{j1}), \ldots, \epsilon(\mathbf{s}_{jm_j})\}$ be the realizations of GP $w(\cdot)$ and white-noise process $\epsilon(\cdot)$, respectively, in the jth subset. After marginalizing over $\mathbf{w}_j$ in the GP-based model for the jth subset, the likelihood of $\boldsymbol{\Omega} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ is given by $\ell_j(\boldsymbol{\Omega}) = N\{\mathbf{y}_j \mid \mathbf{X}_j \boldsymbol{\beta}, \mathbf{V}_j(\boldsymbol{\alpha})\}$, where $\mathbf{V}_j(\boldsymbol{\alpha}) = \mathbf{C}_j(\boldsymbol{\alpha}) + \mathbf{D}_j(\boldsymbol{\alpha})$ and $\mathbf{V}_j(\boldsymbol{\alpha}) = \tilde{\mathbf{C}}_j(\boldsymbol{\alpha}) + \mathbf{D}_j(\boldsymbol{\alpha})$ for full-rank and low-rank GP priors, respectively, and $\mathbf{C}_j(\boldsymbol{\alpha}), \tilde{\mathbf{C}}_j(\boldsymbol{\alpha}), \mathbf{D}_j(\boldsymbol{\alpha})$ are obtained by extending the definitions of $\mathbf{C}(\boldsymbol{\alpha}), \tilde{\mathbf{C}}(\boldsymbol{\alpha}), \mathbf{D}(\boldsymbol{\alpha})$ to the jth subset. The likelihood of $\mathbf{w}_j$ given $\mathbf{y}_j$, $\mathbf{X}_j$, and $\boldsymbol{\Omega}$ is $\ell_j(\mathbf{w}_j) = N\{\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta} \mid \mathbf{w}_j, \mathbf{D}_j(\boldsymbol{\alpha})\}$. The likelihoods in $\ell_j(\boldsymbol{\Omega})$ and $\ell_j(\mathbf{w}_j)$ yield the posterior distributions for $\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{w}^*$, $\mathbf{y}^*$ ($\mathbf{w}^*$ and $\mathbf{y}^*$ have already been defined in the second paragraph of Section 2) based on full-rank or low-rank GP priors and are called jth subset pseudo posterior distributions.

3.2 Second Step: Sampling From Subset Pseudo Posterior Distributions

We define subset pseudo posterior distributions by modifying the likelihoods in $\ell_j(\boldsymbol{\Omega})$ and $\ell_j(\mathbf{w}_j)$. More precisely, the density of the jth subset pseudo posterior distribution of $\boldsymbol{\Omega}$ is given by

$$\pi_{m_j}(\boldsymbol{\Omega} \mid \mathbf{y}_j) = \frac{\{\ell_j(\boldsymbol{\Omega})\}^{n/m_j} \pi(\boldsymbol{\Omega})}{\int \{\ell_j(\boldsymbol{\Omega})\}^{n/m_j} \pi(\boldsymbol{\Omega}) d\boldsymbol{\Omega}}, \tag{6}$$

where we assume that $\int \{\ell_j(\boldsymbol{\Omega})\}^{n/m_j} \pi(\boldsymbol{\Omega}) d\boldsymbol{\Omega} < \infty$, and the subscript '$m_j$' denotes that the density conditions on m_j data samples in the jth subset. The modification of likelihood to yield the subset pseudo posterior density in (6) is called *stochastic approximation* (Minsker et al., 2014). Raising the likelihood to the power of n/m_j is equivalent to replicating every $y(\mathbf{s}_{ji})$ n/m_j times $(i = 1, \ldots, m_j)$, so stochastic approximation accounts for the fact that the jth subset pseudo posterior distribution conditions on a (m_j/n)-fraction of the full data and ensures that its variance is of the same order (as a function of n) as that of the full data posterior distribution. Unlike parametric models, stochastic approximation in spatial regression models has not been studied previously in the literature.

We address this gap using the proposed stochastic approximation in (6). The full conditional densities of jth subset pseudo posterior distributions for prediction and inference follow from their full data counterparts. The jth full conditional densities of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ in the GP-based models are

$$\pi_{m_j}(\boldsymbol{\beta} \mid \mathbf{y}_j, \boldsymbol{\alpha}) = \frac{\{\ell_j(\boldsymbol{\Omega})\}^{n/m_j} \pi(\boldsymbol{\beta})}{\int \{\ell_j(\boldsymbol{\Omega})\}^{n/m_j} \pi(\boldsymbol{\beta}) d\boldsymbol{\beta}},$$

$$\pi_{m_j}(\boldsymbol{\alpha} \mid \mathbf{y}_j, \boldsymbol{\beta}) = \frac{\{\ell_j(\boldsymbol{\Omega})\}^{n/m_j} \pi(\boldsymbol{\alpha})}{\int \{\ell_j(\boldsymbol{\Omega})\}^{n/m_j} \pi(\boldsymbol{\alpha}) d\boldsymbol{\alpha}},$$

where $\pi(\boldsymbol{\beta}) = N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, $\pi(\boldsymbol{\alpha})$ is the prior density of $\boldsymbol{\alpha}$, and we assume that $\int\{\ell_j(\boldsymbol{\Omega})\}^{n/m_j}\pi(\boldsymbol{\beta})d\,\boldsymbol{\beta}$ and $\int\{\ell_j(\boldsymbol{\Omega})\}^{n/m_j}\pi(\boldsymbol{\alpha})d\,\boldsymbol{\alpha}$ respectively are finite. The jth full conditional densities of $\mathbf{y}^*$ and $\mathbf{w}^*$ are calculated after modifying the likelihood of $\mathbf{w}_j$ using stochastic approximation. Given $\mathbf{y}_j$, $\mathbf{X}_j$, and $\boldsymbol{\Omega}$, straightforward calculation yields that the jth subset pseudo posterior predictive density of $\mathbf{w}^*$ is $\pi_{m_j}(\mathbf{w}^* \mid \mathbf{y}_j, \boldsymbol{\Omega}) = N(\mathbf{w}^* \mid \mathbf{m}_{j*}, \mathbf{V}_{j*})$, with

$$\mathbf{V}_{j*} = \mathbf{C}_{*,*}(\boldsymbol{\alpha}) - \mathbf{C}_{*j}(\boldsymbol{\alpha})\,\mathbf{V}_j(\boldsymbol{\alpha})^{-1}\,\mathbf{C}_{*j}(\boldsymbol{\alpha})^T,$$
$$\mathbf{m}_{j*} = \mathbf{C}_{*j}(\boldsymbol{\alpha})\,\mathbf{V}_j(\boldsymbol{\alpha})^{-1}(\mathbf{y}_j - \mathbf{X}_j\,\boldsymbol{\beta}),$$

where $\mathbf{V}_j(\boldsymbol{\alpha}) = \mathbf{C}_j(\boldsymbol{\alpha}) + (n/m_j)^{-1}\,\mathbf{D}_j(\boldsymbol{\alpha})$ and $\mathbf{V}_j(\boldsymbol{\alpha}) = \tilde{\mathbf{C}}_j(\boldsymbol{\alpha}) + (n/m_j)^{-1}\,\mathbf{D}_j(\boldsymbol{\alpha})$ for full-rank and low-rank GP priors, respectively, and $\mathbf{C}_{*,*}(\boldsymbol{\alpha}), \mathbf{C}_{*j}(\boldsymbol{\alpha})$ are $l \times l$, $l \times m_j$ matrices obtained by extending the definition in (3) to subset j for full-rank and low-rank GP priors with covariance functions $C_{\boldsymbol{\alpha}}(\cdot,\cdot)$ and $\tilde{C}_{\boldsymbol{\alpha}}(\cdot,\cdot)$, respectively. We note that the stochastic approximation exponent, n/m_j, scales $\mathbf{D}_j(\boldsymbol{\alpha})$ in $\mathbf{V}_j(\boldsymbol{\alpha})$ so that the uncertainty in subset and full data posterior distributions are of the same order (as a function of n). The jth subset pseudo posterior predictive density of $\mathbf{y}^*$ given the MCMC samples of $\mathbf{w}^*$ and $\boldsymbol{\Omega}$ in the jth subset is $N\{\mathbf{y}^* \mid \mathbf{X}^*\,\boldsymbol{\beta} + \mathbf{w}^*, \mathbf{D}_j(\boldsymbol{\alpha})\}$.

We specialize the sampling algorithm (Steps 1–3) introduced in Section 2 to subset j ($j = 1, \ldots, k$), sampling $\{\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{y}^*, \mathbf{w}^*\}$ in each subset across multiple MCMC iterations; see supplementary material for subset pseudo posterior sampling algorithms in the full-rank and low-rank GP priors. The computational complexity of jth subset pseudo posterior computations follows from their full data counterparts if we replace n by m_j. Specifically, the computational complexities for sampling a subset pseudo posterior are $O(m^3)$ and $O(mr^2)$ flops per iteration if the model in (5) uses a full-rank or a low-rank GP prior, respectively, where $m = \max_j m_j$. Performing subset pseudo posterior computations in parallel across k servers also alleviates the need to store large covariance matrices. We hereon refer to subset pseudo posterior as subset posterior.

Our second step in the distributed framework resembles some existing methods based on the composite likelihood (Varin et al., 2011); for example, Chandler and Bate (2007) and Ribatet et al. (2012) construct pseudo likelihood to replace the full data likelihood, where the pseudo likelihood attempts to capture important features of the full data likelihood while offering computational efficiency. In the context of geostatistical modeling with GP or its variants, for computational efficiency, the pseudo likelihood will naturally be based on independence of data blocks at some level. To make up for the incorrect asymptotic distribution of the posterior distribution due to the incorrect independence assumption, they propose a number of adjustments in the composite log likelihood (e.g., the margin adjustment and the curvature adjustment). Similar to these approaches, the likelihood adjustment in each subset for the second step of our general distributed approach is also born out of consideration to scale the asymptotic variance of subset posteriors to the same order as the asymptotic variance of the full posterior; however, unlike these composite likelihood approaches, the distributed approaches we focus on do not assume any restrictive structure (e.g., block independence) in the data likelihood. In fact, there is no guarantee that the induced data likelihood that leads to the combined pseudo posterior for any distributed method assumes any block independence form. Moreover, Savitsky and Srivastava (2018) represents an example of embedding a composite likelihood in a distributed setup that computes the Wasserstein barycenter. Likewise, we believe that most of these "flexible" composite likelihoods can be used in extensions of the distributed framework for subset sampling in models where the true likelihood is unavailable or expensive to compute.

3.3 Third Step: Combination of Subset Posterior Distributions

We now discuss strategies for combining subset posteriors to construct a "combined pseudo posterior", which is used as an computationally efficient alternative to the full data posterior. The combination strategies discussed here include representative approaches used for the distributed Bayesian inference in independent data, but they have not been studied empirically or theoretically for correlated spatial data setting. Specifically, we compare the following combination schemes with our approach: (i) Consensus Monte Carlo (CMC); (ii) Double Parallel Monte Carlo (DPMC); and (iii) combination

through the Wasserstein barycenter.

3.3.1 Consensus Monte Carlo (CMC) For a scalar or vector parameter of interest θ, Consensus Monte Carlo (CMC) (Scott et al., 2016) draws samples from an approximation of the full posterior. In our setting, θ can be taken as $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\mathbf{w}^*$, $\mathbf{y}^*$, their individual components, or any functional of these parameters. Let $\left\{\theta_1^{(j)}, \ldots, \theta_T^{(j)}\right\}$ denote the T posterior samples of θ generated from subset j post convergence. Based on the Bernstein-von Mises (BvM) theorem, Scott et al. (2016) proposed to use the weighted average $\sum_{j=1}^{k} w_j \theta_i^{(j)}$, $i = 1, \ldots, T$ to approximate T samples from the full data posterior, where the BvM theorem says that the full data posterior tends to a normal distribution centered around the true parameter value as n grows and w_j is the inverse of the empirical covariance matrix of $\left\{\theta_1^{(j)}, \ldots, \theta_T^{(j)}\right\}$. This algorithm is exact when the samples are independent and each subset posterior is Gaussian, but this assumption is rarely satisfied in spatial applications.

3.3.2 Double Parallel Monte Carlo (DPMC) Following the notation for CMC, let θ be the parameter of interest. Denote the average of θ draws on the subset j as $\overline{\theta}^{(j)} = (\theta_1^{(j)} + \ldots + \theta_T^{(j)})/T$ $(j = 1, \ldots, k)$ and $\overline{\theta} = (\overline{\theta}^{(1)} + \ldots + \overline{\theta}^{(k)})/k$ be their average. DPMC (Xue and Liang, 2019) re-centers each of the subset posteriors to $\overline{\theta}$ and then uses the mixture of re-centered subset posteriors, given by $\frac{1}{k}\sum_{j=1}^{k} \pi_{m_j}(\theta - \overline{\theta} + \overline{\theta}^{(j)} | \, \mathbf{y}_j)$, to approximate the full data posterior. Following the implementation of DPMC in the context of independent data, we simply transform $\theta_t^{(j)}$ to $\overline{\theta} + (\theta_t^{(j)} - \overline{\theta}^{(j)})$ $(t = 1, \ldots, T;\ j = 1, \ldots, k)$ and treat them as draws from the combined posterior distribution.

3.3.3 Combining subset posteriors using Wasserstein barycenter This combination algorithm relies on the notion of Wasserstein barycenter (Srivastava et al., 2015). If $\nu_1, \ldots, \nu_k$ are the k subset posterior distributions of θ, then the combined pseudo posterior $\overline{\nu}$ is the Wasserstein barycenter defined as

$$\overline{\nu} = \underset{\nu \in \mathcal{P}_2(\Theta)}{\operatorname{argmin}} \frac{1}{k} \sum_{j=1}^{k} W_2^2(\nu, \nu_j), \tag{7}$$

$$W_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\Theta \times \Theta} \|x - y\|^2 \, d\pi(x, y),$$

where $\|\cdot\|$ is a metric on the parameter space Θ, $\mathcal{P}(\Theta)$ be the space of all probability measures on Θ, $\mathcal{P}_2(\Theta) = \{\mu \in \mathcal{P}(\Theta) : \int_\Theta \|\theta - \theta_0\|^2 \mu(d\theta) < \infty\}$, $W_2(\mu, \nu)$ is the Wasserstein distance between $\mu, \nu \in \mathcal{P}_2(\Theta)$, and $\Pi(\mu, \nu)$ is the space of all joint distributions of $\Theta \times \Theta$ with μ, ν as marginals. It is known that $\overline{\nu}$ exists and is unique (Agueh and Carlier, 2011).

In practice, ν_j is replaced by its empirical approximation obtained using the θ draws from subset j. A variety of efficient algorithms are available to provide an empirical approximation of $\overline{\nu}$ $(j = 1, \ldots, k)$ (Cuturi and Doucet, 2014). This approach for combining subset posteriors leads to the combined pseudo posterior referred to as the Wasserstein posterior (WASP), which is preferred over several other combination methods for independent data (Srivastava et al., 2018); for example, directly averaging over many subset posterior densities with different means can usually result in an undesirable multimodal pseudo posterior distribution, but the WASP does not have this problem; see Figure 1 in Srivastava et al. (2018). Besides, the WASP does not rely on the asymptotic normality of the subset posterior distributions as in other approaches, such as the CMC.

3.3.4 Computing the WASP with constraints Computing the WASP is inefficient if k is large, so $\overline{\nu}$ is computed with additional constraints (Srivastava and Xu, 2021). One such approach constrains θ to be a one-dimensional functional of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\mathbf{w}^*$, or $\mathbf{y}^*$. For a scalar parameter, the Wasserstein barycenter of θ can be easily obtained by averaging empirical subset posterior quantiles (Li et al., 2017). We refer to this approach as distributed kriging (DISK) and the combined pseudo posterior is called as the DISK posterior. Let ν and ν_j be the full posterior and jth subset posterior distribution of θ, and $\overline{\nu}$ be the Wasserstein barycenter of $\nu_1, \ldots, \nu_k$ as in (7). For any $q \in (0, 1)$, let $\hat{\nu}_j^q$ be the qth empirical quantile of ν_j based on the MCMC samples from ν_j, and $\hat{\overline{\nu}}^q$ be the qth quantile of the empirical version of $\overline{\nu}$. Then, $\hat{\overline{\nu}}^q$ can be computed as

$$\hat{\overline{\nu}}^q = \frac{1}{k} \sum_{j=1}^{k} \hat{\nu}_j^q, \quad q = \xi, 2\xi, \ldots, 1 - \xi, \tag{8}$$

where ξ is the grid-size of the quantiles. If the ξ-grid is fine enough, then the θ draws from the marginal DISK distribution are obtained by inverting the empirical distribution function supported on the quantile estimates (Li et al., 2017). In practice, the primary interest often lies in the posterior distribution of some one-dimensional functional of θ; therefore, the univariate WASP obtained by averaging quantiles in (8) accomplishes this with great generality and convenient implementation. Our simulation studies in Section 4 investigate if the multi-variate combination approaches in CMC, DPMC, or WASP lead to any notable improvement over the univariate quantile combination in (8).

The choice of the grid size is mainly determined by the Monte Carlo approximation error of each subset posterior. In general, the Monte Carlo approximation error to subset posteriors can be measured in terms of the size of MCMC samples (say T). This error is evaluated by taking T to infinity and differs from the statistical error, where n tends to infinity. In the context of distributed Bayesian inference for independent data, Theorem 3 in the supplementary material of Li et al. (2017) has shown that the Monte Carlo error is usually in some polynomial order of T such as $O(T^{-1/2})$ and $O(T^{-1/4})$ depending on the distance measure and is independent of the statistical error defined in terms of n. Following this intuition, in application, we usually draw at least 10^4 MCMC samples for each subset posterior and use all of them to construct the quantiles.

A key feature of the combination scheme for the four distributed approaches is that given the subset posterior MCMC samples, the combination step is agnostic to the choice of a model. Specifically, given MCMC samples from the k subset posterior distributions, (8) remains the same for models based on a full-rank GP prior, a low-rank GP prior, such as MPP, or any other model described in Section 1.1. Since the combination step over k subsets takes $O(k)$ flops for all four combination schemes and $k < n$, the total time for computing the empirical quantile estimates of the combined pseudo posterior in inference or prediction requires $O(k) + O(m^3)$ and $O(k)+O(rm^2)$ flops in models based on full-rank and low-rank GP priors, respectively. Assuming that we have abundant computational resources, k is chosen large enough so that $O(m^3)$ computations are feasible. This would enable applications of the proposed distributed framework in models based on both full-rank and low-rank GP priors in massive n settings.

3.4 Bayes L_2-Risk: Bias-Variance Decomposition and Convergence Rates

In the distributed Bayesian setup, it is already known that when the data are independent and identically distributed (i.i.d.), the combined posterior distribution using the Wasserstein barycenter of subset posteriors approximates the full data posterior distribution at a near optimal parametric rate, under certain conditions as $n, k, m_1, \ldots, m_k \to \infty$ (Li et al., 2017, Srivastava et al., 2018); however, in models based on spatial process, data are not i.i.d. and inference on the infinite dimensional true spatial surface is of primary importance. Few formal theoretical results are available in this nonparametric distributed Bayes setup. The recent work (Szabo and van Zanten, 2019) has shown that combination using Wasserstein barycenter has optimal Bayes risk and adapts to the smoothness of $w_0(\cdot)$, the true but unknown $w(\cdot)$, in the Gaussian white noise model, which is a special case of (1) with additional smoothness assumptions on $w_0(\cdot)$.

We mainly focus on the theoretical properties of the DISK posterior of the mean surface $\mathbf{x}(\cdot)^T\boldsymbol{\beta}+w(\cdot)$, and our theoretical framework can be possibly extended to the other three combination schemes described in Section 3.3. For ease of presentation, we assume that $m_1 = \cdots = m_k = m$ and $k = n/m$. Determining the appropriate order for k in terms of n is one of the key issues for all distributed statistical methods. Our theory below reveals that the number of subsets k cannot increase too fast with n, or equivalently, the subset size m cannot be too small, mainly because a small subset size m will result in larger random errors in the estimation from subset posterior distributions.

We formally explain the model setup for our theory development. Suppose that the data generation process follows the model (1) with the true parameter value $\boldsymbol{\Omega}_0 = (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ and the true spatial surface $w_0(\cdot)$. We focus on the Bayes L_2-risk of the DISK predictive posterior for the mean function in (1); that is, $\mathbf{x}(\mathbf{s}^*)^T\boldsymbol{\beta}+w(\mathbf{s}^*)$ for any testing location $\mathbf{s}^* \in \mathcal{S}$. To ease the complexity of our theory, we first

present two theorems below for the simplified model

$$y(\mathbf{s}_i) = w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \quad \epsilon(\mathbf{s}_i) \sim N\left(0, \tau^2\right), \quad w(\cdot) \sim \mathrm{GP}\{0, \lambda_n^{-1} C_{\boldsymbol{\alpha}}(\cdot,\cdot)\}, \tag{9}$$

for $i = 1, \ldots, n$. Compared to the spatial model (1), the model (9) does not contain the regression term $\mathbf{x}(\mathbf{s})^T \boldsymbol{\beta}$; however, our theory includes this regression term later by modifying the covariance function; see Corollary 3.3 below. The tuning parameter λ_n is a user-chosen deterministic sequence that depends on n. In real applications, one can simply set $\lambda_n = 1$, but one can also choose λ_n such that the posterior convergence rate is minimax optimal; see Theorem 3.2 below and the discussions therein.

We introduce some theoretical definitions used in stating our results. Let $\boldsymbol{\alpha}_0$ be the true kernel parameter. Let $\mathbb{P}_{\mathbf{s}}$ be a design distribution of $\mathbf{s}$ over $\mathcal{D}$, $L_2(\mathbb{P}_{\mathbf{s}})$ be the L_2 space under $\mathbb{P}_{\mathbf{s}}$, the inner product in $L_2(\mathbb{P}_{\mathbf{s}})$ is defined as $\langle f, g\rangle_{L_2(\mathbb{P}_{\mathbf{s}})} = \mathbb{E}_{\mathbb{P}_{\mathbf{s}}}(fg)$ for any $f, g \in L_2(\mathbb{P}_{\mathbf{s}})$ where $\mathbb{E}_{\mathbb{P}_{\mathbf{s}}}(\cdot)$ represents an expectation taken with with respect to the distribution, $\mathbb{P}_{\mathbf{s}}$. For any $f \in L_2(\mathbb{P}_{\mathbf{s}})$ and $\mathbf{s} \in \mathcal{D}$, define the linear operator $(T_{\boldsymbol{\alpha}_0} f)(\mathbf{s}) = \int_{\mathcal{D}} C_{\boldsymbol{\alpha}_0}(\mathbf{s}, \mathbf{s}') f(\mathbf{s}') d\,\mathbb{P}_{\mathbf{s}}(\mathbf{s}')$. According to the Mercer's theorem, there exists an orthonormal basis $\{\varphi_i(\mathbf{s})\}_{i=1}^{\infty}$ in $L_2(\mathbb{P}_{\mathbf{s}})$, such that $C_{\boldsymbol{\alpha}_0}(\mathbf{s}, \mathbf{s}') = \sum_{i=1}^{\infty} \mu_i \varphi_i(\mathbf{s}) \varphi_i(\mathbf{s}')$, where $\mu_1 \geq \mu_2 \geq \ldots \geq 0$ are the eigenvalues and $\{\varphi_i(\mathbf{s})\}_{i=1}^{\infty}$ are the eigenfunctions of $T_{\boldsymbol{\alpha}_0}$. The trace of the kernel $C_{\boldsymbol{\alpha}_0}$ is defined as $\mathrm{tr}(C_{\boldsymbol{\alpha}_0}) = \sum_{i=1}^{\infty} \mu_i$. Any $f \in L_2(\mathbb{P}_{\mathbf{s}})$ has the series expansion $f(\mathbf{s}) = \sum_{i=1}^{\infty} \theta_i \varphi_i(\mathbf{s})$, where $\theta_i = \langle f, \varphi_i \rangle_{L_2(\mathbb{P}_{\mathbf{s}})}$. The reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ attached to $C_{\boldsymbol{\alpha}_0}$ is the space of all functions $f \in L_2(\mathbb{P}_{\mathbf{s}})$ such that the $\mathbb{H}$-norm $\|f\|_{\mathbb{H}}^2 = \sum_{i=1}^{\infty} \theta_i^2/\mu_i < \infty$. The RKHS $\mathbb{H}$ is the completion of the linear space of functions defined as $\sum_{i=1}^{I} a_i C_{\boldsymbol{\alpha}_0}(\mathbf{s}_i, \cdot)$, where I is a positive integer, $\mathbf{s}_i \in \mathcal{D}$, and $a_i \in \mathbb{R}$ $(i = 1, \ldots, I)$; see van der Vaart and van Zanten (2008) for more details on RKHS.

We impose the following assumptions.

A.1 (Sampling) The locations $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and $\mathbf{s}^*$ are independently drawn from the same sampling distribution $\mathbb{P}_{\mathbf{s}}$. $\mathcal{S}_1, \ldots, \mathcal{S}_k$ is a random disjoint partition of $\mathcal{S}$, each with size $m = n/k$.

A.2 (True model) The true function w_0 is an element of the RKHS $\mathbb{H}$ attached to the kernel $C_{\boldsymbol{\alpha}_0}$. At a location $\mathbf{s}$, the observation is $y(\mathbf{s}) = w_0(\mathbf{s}) + \epsilon(\mathbf{s})$, where $\epsilon(\mathbf{s})$ is a white noise process with the true variance $\tau_0^2 < \infty$.

A.3 (Trace class kernel) $\mathrm{tr}(C_{\boldsymbol{\alpha}_0}) < \infty$.

A.4 (Moment condition) There are positive constants ρ and $q > 4$ such that $\mathbb{E}_{\mathbb{P}_{\mathbf{s}}}\{\varphi_i^{2q}(\mathbf{s})\} \leq \rho^{2q}$ for every $i \in \mathbb{N}$.

The random partition in A.1 guarantees that each individual subset $\mathcal{S}_j$ $(j = 1, \ldots, k)$ is a random sample from $\mathbb{P}_{\mathbf{s}}$. In general, the RKHS $\mathbb{H}$ in A.2 can be a smaller space relative to the support of the GP prior. While we use $w_0 \in \mathbb{H}$ in A.2 mainly for technical simplicity, this assumption can be possibly relaxed by considering sieves with increasing $\mathbb{H}$-norms, similar to Assumption B′ and Theorem 2 in Zhang et al. (2015). Furthermore, A.2 only requires that the true unknown error distribution to have a finite variance. Although we fit the data using the normal error in model (9), our theory below allows this error distribution to be misspecified and not normal; therefore, our posterior convergence rate results also hold for heavy-tailed error distributions such as t_4, which are more general than van der Vaart and van Zanten (2011) whose techniques fully depend on the normal error assumption. In A.3, $\mathrm{tr}(C_{\boldsymbol{\alpha}})$ measures the size of the covariance function and imposes conditions on the regularity of functions that DISK can learn. A.4 on the eigenfunctions controls the error in approximating $C_{\boldsymbol{\alpha}_0}(\mathbf{s}, \mathbf{s}')$ by a finite sum, similar to Assumption A in Zhang et al. (2015).

We first consider the case where both the error variance τ^2 and the kernel parameter $\boldsymbol{\alpha}$ are fixed and known, similar to van der Vaart and van Zanten (2011). We extend our results to a special case where τ^2 is assigned a prior with bounded support in Corollary 1.1 of the supplementary material.

A.5 (Fixed parameters) $\boldsymbol{\alpha}$ and τ^2 are fixed at their true values $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$, $\tau^2 = \tau_0^2$.

We begin by examining the Bayes L_2-risk of the DISK posterior for estimating w_0 in (9). Let $\overline{w}(\mathbf{s}^*)$ be a random variable that follows the DISK posterior for estimating $w_0(\mathbf{s}^*)$. Let $\mathbb{E}_{\mathbf{s}^*}$, $\mathbb{E}_{\mathcal{S}}$, and $\mathbb{E}_{\mathbf{y}, \overline{w}(\mathbf{s}^*)|\mathcal{S}, \mathbf{s}^*}$ respectively be the expectations with respect to the distributions of $\mathbf{s}^*$, $\mathcal{S}$, and $\{\mathbf{y}, \overline{w}(\mathbf{s}^*)\}$ given $\mathcal{S}, \mathbf{s}^*$. Given the random partition assumption in A.1, each individual subset $\mathcal{S}_j$ $(j = 1, \ldots, k)$ is a random sample from $\mathbb{P}_{\mathbf{s}}$. By A.5, we can drop the subscript "0" in $\boldsymbol{\alpha}_0$ and τ_0^2. Then, $\overline{w}(\mathbf{s}^*)$ given $\mathbf{y}, \mathcal{S}, \mathbf{s}^*$ has the den-

sity $N(\overline{m}, \overline{v})$, where

$$\overline{m} = \frac{1}{k}\sum_{j=1}^{k} \mathbf{c}_{j,*}^T (\mathbf{C}_{j,j} + \tfrac{\tau^2\lambda_n}{k}\mathbf{I})^{-1}\mathbf{y}_j,$$

$$\overline{v}^{1/2} = \frac{1}{k}\sum_{j=1}^{k} v_j^{1/2}, \tag{10}$$

$$v_j = \lambda_n^{-1}\left\{c_{*,*} - \mathbf{c}_{j,*}^T(\mathbf{C}_{j,j} + \tfrac{\tau^2\lambda_n}{k}\mathbf{I})^{-1}\mathbf{c}_{j,*}\right\},$$

$\mathbf{c}_{j,*}^T = [\operatorname{cov}\{w(\mathbf{s}_{j1}), w(\mathbf{s}^*)\}, \ldots, \operatorname{cov}\{w(\mathbf{s}_{jm}), w(\mathbf{s}^*)\}]$, and $c_{*,*} = \operatorname{cov}\{w(\mathbf{s}^*), w(\mathbf{s}^*)\}$. The Bayes L_2-risk of DISK in estimating w_0 is $\mathbb{E}_{\mathbf{s}^*}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S},\mathbf{s}^*}\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\}^2$. This risk can be used to quantify how quickly the DISK posterior concentrates around the unknown true surface $w_0(\cdot)$ as the total sample size n increases to infinity. The convergence rate of this Bayes L_2-risk towards zero also gives the posterior contraction rate of the DISK posterior defined in the same way as in Bayesian nonparametrics, such as van der Vaart and van Zanten (2011, Theorem 2). When the parameters τ^2 and $\boldsymbol{\alpha}$ are fixed and known, it is straightforward to show (see the proof of Theorem 3.1 in the supplementary material) that this Bayes L_2-risk can be decomposed into the squared bias, the variance of subset posterior means, and the variance of DISK posterior terms as

$$\begin{aligned}
&\text{bias}^2 = \mathbb{E}_{\mathbf{s}^*}\mathbb{E}_{\mathcal{S}}\{\mathbf{c}_*^T(k\mathbf{L} + \tau^2\lambda_n\mathbf{I})^{-1}\mathbf{w}_0 - w_0(\mathbf{s}^*)\}^2,\\
&\text{var}_{\text{mean}} = \tau^2\,\mathbb{E}_{\mathbf{s}^*}\mathbb{E}_{\mathcal{S}}\left\{\mathbf{c}_*^T(k\mathbf{L} + \tau^2\lambda_n\mathbf{I})^{-2}\mathbf{c}_*\right\},\\
&\text{var}_{\text{DISK}} = \mathbb{E}_{\mathbf{s}^*}\mathbb{E}_{\mathcal{S}}\{\overline{v}(\mathbf{s}^*)\},
\end{aligned} \tag{11}$$

where $\overline{v}(\mathbf{s}^*) = \mathbb{E}_{\mathbf{y}|\mathcal{S}}[\operatorname{var}\{\overline{w}(\mathbf{s}^*) \mid \mathbf{y}\}]$, $\mathbf{c}_*^T = (\mathbf{c}_{1,*}^T, \ldots, \mathbf{c}_{k,*}^T)$, $\mathbf{w}_{0j} = \{w_0(\mathbf{s}_{j1}), \ldots, w_0(\mathbf{s}_{jk})\}$ for $j = 1, \ldots, k$, $\mathbf{w}_0^T = (\mathbf{w}_{01}, \ldots, \mathbf{w}_{0k})$, and $\mathbf{L}$ is a block-diagonal matrix with $\mathbf{C}_{1,1}, \ldots, \mathbf{C}_{k,k}$ along the diagonal. The next theorem provides theoretical upper bounds for each of the three terms in (11).

Theorem 3.1 *If Assumptions A.1–A.5 hold, then*

$$\begin{aligned}
&\text{Bayes } L_2 \text{ risk}\\
&= \mathbb{E}_{\mathbf{s}^*}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S},\mathbf{s}^*}\{\overline{w}(\mathbf{s}_*) - w_0(\mathbf{s}_*)\}^2\\
&= \text{bias}^2 + \text{var}_{\text{mean}} + \text{var}_{\text{DISK}},
\end{aligned} \tag{12}$$

$$\begin{aligned}
\text{bias}^2 &\le \frac{8\tau^2\lambda_n}{n}\|w_0\|_{\mathbb{H}}^2\\
&\quad + \|w_0\|_{\mathbb{H}}^2 \inf_{d\in\mathbb{N}}\left[\frac{8n}{\tau^2\lambda_n}\rho^4\operatorname{tr}(C_{\boldsymbol{\alpha}})\operatorname{tr}(C_{\boldsymbol{\alpha}}^d)\right.\\
&\quad \left. + \mu_1 R(m,n,d,q)\right],\\
\text{var}_{\text{mean}} &\le \left(\frac{2n}{k\lambda_n} + \frac{4\|w_0\|_{\mathbb{H}}^2}{k}\right)\inf_{d\in\mathbb{N}}\left[\mu_{d+1}\right.\\
&\quad \left. + \frac{12n}{\tau^2\lambda_n}\rho^4\operatorname{tr}(C_{\boldsymbol{\alpha}})\operatorname{tr}(C_{\boldsymbol{\alpha}}^d) + R(m,n,d,q)\right]\\
&\quad + \frac{12\tau^2\lambda_n}{kn}\|w_0\|_{\mathbb{H}}^2 + 12\frac{\tau^2}{n}\gamma\left(\frac{\tau^2\lambda_n}{n}\right),\\
\text{var}_{\text{DISK}} &\le 3\frac{\tau^2}{n}\gamma\left(\frac{\tau^2\lambda_n}{n}\right)\\
&\quad + \inf_{d\in\mathbb{N}}\left[\left\{\frac{4n}{\tau^2\lambda_n^2}\operatorname{tr}(C_{\boldsymbol{\alpha}}) + \frac{1}{\lambda_n}\right\}\operatorname{tr}(C_{\boldsymbol{\alpha}}^d)\right.\\
&\quad \left. + \lambda_n^{-1}\operatorname{tr}(C_{\boldsymbol{\alpha}})R(m,n,d,q)\right],
\end{aligned}$$

where $\mathbb{N}$ is the set of all positive integers, A is a global positive constant that does not depend on any of the quantities here, and

$$\begin{aligned}
b(m,d,q) &= \max\left(\sqrt{\max(q,\log d)}, \frac{\max(q,\log d)}{m^{1/2-1/q}}\right),\\
R(m,n,d,q) &= \left\{\frac{A\rho^2 b(m,d,q)\gamma(\tau^2\lambda_n/n)}{\sqrt{m}}\right\}^q,\\
\gamma(a) &= \sum_{i=1}^{\infty}\frac{\mu_i}{\mu_i + a} \text{ for } a > 0, \quad \operatorname{tr}(C_{\boldsymbol{\alpha}}^d) = \sum_{i=d+1}^{\infty}\mu_i.
\end{aligned}$$

These upper bounds are similar to the bounds obtained in Theorem 1 of Zhang et al. (2015) for the frequentist distributed estimator in kernel ridge regression. Although the upper bounds in (12) appear very complicated and involve many terms, the dominant term among them is $\frac{\tau^2}{n}\gamma\left(\frac{\tau^2\lambda_n}{n}\right)$, where the function $\gamma(\cdot)$ is related to the "effective dimensionality" of the covariance function $C_{\boldsymbol{\alpha}}$ (Zhang, 2005). This term determines how fast the Bayes L_2-risk converges to zero, as long as k is chosen to be some proper order of n such that all the other terms in the upper bounds of (12) can be made negligible compared to $\frac{\tau^2}{n}\gamma\left(\frac{\tau^2\lambda_n}{n}\right)$. In particular, the term $R(m,n,d,q)$ that quantifies the random error and appears in the infimums in all three upper bounds of

(12) generally decreases with m and increases with k; therefore, to ensure the dominance of $\frac{\tau^2}{n}\gamma\left(\frac{\tau^2\lambda_n}{n}\right)$, k cannot increase too fast with n; see Theorem 3.2 below.

In contrast to the frequentist literature such as Zhang et al. (2015), a significant difference in our Theorem 3.1 is that our risk bounds involve two different variance terms. Our analysis naturally introduces the variance term var_{DISK} that corresponds to the variance of the DISK posterior distribution, while frequentist kernel ridge regression only finds a point estimate of w_0 and thus does not include this variance term. Each of the three upper bounds in Theorem 3.1 can be made close to zero as n increases to ∞ and k is chosen to grow at an appropriate rate depending on n. The next theorem finds the appropriate order for k in terms of n, such that the DISK posterior achieves nearly minimax optimal rates in its Bayes L_2-risk (12), for three types of commonly used covariance functions/kernels, (i) degenerate kernels, (ii) kernels with exponentially decaying eigenvalues, and (iii) kernels with polynomially decaying eigenvalues. The kernel $C_{\boldsymbol{\alpha}}$ is a degenerate kernel of rank d^* if there is some constant positive integer d^* such that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_{d^*} > 0$ and $\mu_{d^*+1} = \mu_{d^*+2} = \ldots = \mu_\infty = 0$.

Theorem 3.2 *If Assumptions A.1–A.5 hold, then as* $n \to \infty$,

(i) if $C_{\boldsymbol{\alpha}}$ *is a degenerate kernel of rank* d^*, $\lambda_n = 1$, *and* $k \leq cn^{\frac{q-4}{q-2}}/(\log n)^{\frac{2q}{q-2}}$ *for some constant* $c > 0$, *then the Bayes* L_2*-risk of DISK posterior satisfies* $\mathbb{E}_{\mathbf{s}^*}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S},\mathbf{s}^*}\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\}^2 = O\left(n^{-1}\right)$;

(ii) if $\mu_i \leq c_{1\mu}\exp\left(-c_{2\mu}i^{\kappa}\right)$ *for some constants* $c_{1\mu} > 0$, $c_{2\mu} > 0$, $\kappa > 0$ *and all* $i \in \mathbb{N}$, $\lambda_n = 1$, *and* $k \leq cn^{\frac{q-4}{q-2}}/(\log n)^{\frac{2(q\kappa+q-1)}{\kappa(q-2)}}$ *for some constant* $c > 0$, *then the Bayes* L_2*-risk of DISK posterior satisfies* $\mathbb{E}_{\mathbf{s}^*}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S},\mathbf{s}^*}\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\}^2 = O\left\{(\log n)^{1/\kappa}/n\right\}$;

(iii) if $\mu_i \leq c_\mu i^{-2\eta}$ *for some constants* $c_\mu > 0, \eta > \frac{q-1}{q-4}$ *and all* $i \in \mathbb{N}$, $\lambda_n = 1$, *and* $k \leq cn^{\frac{(q-4)\eta-(q-1)}{(q-2)\eta}}/(\log n)^{\frac{2q}{q-2}}$ *for some constant* $c > 0$, *then the Bayes* L_2*-risk of DISK posterior satisfies* $\mathbb{E}_{\mathbf{s}^*}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S},\mathbf{s}^*}\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\}^2 = O\left(n^{-\frac{2\eta-1}{2\eta}}\right)$; *and*

(iv) if $\mu_i \leq c_\mu i^{-2\eta}$ *for some constants* $c_\mu > 0$, $\eta > \frac{q-1}{q-4}$ *and all* $i \in \mathbb{N}$, $\lambda_n = c_1 n^{1/(2\eta+1)}$, *and* $k \leq c_2 n^{\frac{(2\eta-1)q-8\eta}{(q-2)(2\eta+1)}}/(\log n)^{\frac{2q}{q-2}}$ *for some positive constants* c_1, c_2, *then the Bayes* L_2*-risk of DISK posterior satisfies* $\mathbb{E}_{\mathbf{s}^*}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S},\mathbf{s}^*}\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\}^2 = O\left(n^{-\frac{2\eta}{2\eta+1}}\right)$.

In Theorem 3.2, the space of w_0 is the RKHS $\mathbb{H}$ attached to $C_{\boldsymbol{\alpha}}$ by Assumption A.2. In Case (i), the RKHS of $C_{\boldsymbol{\alpha}}$ is a d^*-dimensional space of functions. For example, the covariance functions in subset of regressors approximation (Quiñonero-Candela and Rasmussen, 2005) and predictive process (Banerjee et al., 2008) are both degenerate with their ranks equaling the number of inducing variables and knots, respectively. One example of Case (ii) is the squared exponential kernel, which is popular in machine learning. The squared exponential kernel defined on $\mathbb{R}$ with $\mathbb{P}_{\mathbf{s}}$ being a Gaussian measure has exponentially decaying eigenvalues (Zhu et al., 1998), and it RKHS only contains functions with infinite smoothness. The rate of decay of the L_2-risks in Case (i) and Case (ii) with $\kappa = 2$ are known to be minimax optimal (Raskutti et al., 2012, Yang et al., 2017).

Cases (iii) and (iv) apply to the class of kernels with polynomially decaying eigenvalues. For example, consider the Matérn covariance function $C_{\sigma^2,\phi,\nu}(\mathbf{s},\mathbf{s}') = \sigma^2\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\phi\|\,\mathbf{s}-\mathbf{s}'\,\|\right)^{\nu}\mathcal{K}_\nu\left(\phi\|\,\mathbf{s}-\mathbf{s}'\,\|\right)$, where $\mathbf{s},\mathbf{s}' \in \mathcal{D} \subset \mathbb{R}^d$, $\sigma^2 > 0$, $\phi > 0$, $\boldsymbol{\alpha} = (\sigma^2,\phi)$, $\nu \geq d/2$ is known, $\Gamma(\cdot)$ is the gamma function, and $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of the second kind. Then the RKHS of $C_{\sigma^2,\phi,\nu}(\mathbf{s},\mathbf{s}')$ defined on a compact domain $\mathcal{D}$ with Lipschitz boundary is norm equivalent to the Sobolev space with order $\nu + d/2$ (Wendland 2005, Corollary 10.48). Furthermore, when $\mathbb{P}_{\mathbf{s}}$ is the uniform distribution on $\mathcal{D}$, the eigenvalues of Matérn kernels decay as $\mu_i \leq c_\mu i^{-2\nu/d}$ for all $i \in \mathbb{N}$, such that $\eta = \nu/d$ in Cases (iii) and (iv) (Santin and Schaback 2016, Theorem 6). In the special case of $\nu = 1/2$ and $d = 1$, $C_{\sigma^2,\phi,1/2}(\mathbf{s},\mathbf{s}') = \sigma^2\exp\left(-\phi\|\,\mathbf{s}-\mathbf{s}'\,\|\right)$ is the exponential kernel, whose eigenfunctions are bounded sine and cosine functions, so (A.4) is also satisfied with $q = +\infty$ (Van Trees 2001, Section 3.4.1). It is unknown whether the eigenfunctions of Matérn kernels can be uniformly bounded for general ν and d.

When $\eta = \nu/d$ in Cases (iii) and (iv), the rate

$O\big(n^{-\frac{2\nu-d}{2\nu}}\big)$ for the Bayes L_2-risk in Case (iii) is not minimax optimal for estimating functions in the Sobolev space of order $\nu + d/2$, whereas the faster rate $O\big(n^{-\frac{2\nu}{2\nu+d}}\big)$ in Case (iv) is minimax optimal. This is because (iv) has used the additional optimal tuning parameter $\lambda_n = c_1 n^{\nu/(2\nu+d)}$, while setting $\lambda_n = 1$ is sub-optimal in this case. The use of a tuning parameter to achieve optimal convergence is common in Gaussian process regression and kernel ridge regression (Zhang et al., 2015, Yang et al., 2017). Although van der Vaart and van Zanten (2011) have shown the minimax optimal posterior convergence rates for the Matérn kernel without using tuning parameters, their proof only works when the true error distribution of $\epsilon(\mathbf{s})$ is sub-Gaussian. In comparison, our Assumption A.1 only requires that $\epsilon(\mathbf{s})$ has a finite variance without the normality assumption, which is more general and allows the model (9) to be misspecified in the error distribution.

For the conditions on k, in the case when $q = +\infty$, the upper bounds on k in (i), (ii), (iii), and (iv) reduce to $k = O\{n/(\log n)^2\}$, $k = O\{n/(\log n)^{2/\kappa}\}$, $k = O\{n^{\frac{\eta-1}{\eta}}/(\log n)^2\}$, and $k = O\{n^{\frac{2\eta-1}{2\eta+1}}/(\log n)^2\}$, respectively. The convergence rate results in Theorem 3.2 hold as long as k does not grow too fast with n.

We can generalize the results in Theorems 3.1 and 3.2 to the model (1). Besides A.1–A.4, we further make the following assumption on $\mathbf{x}(\cdot)$ and the prior on $\boldsymbol{\beta}$:

B.1 All p components of $\mathbf{x}(\cdot)$ are non-random functions in $\mathcal{S}$. The prior on $\boldsymbol{\beta}$ is $N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$ and it is independent of the prior on $w(\cdot)$, which is $\mathrm{GP}\{0, C_{\boldsymbol{\alpha}}(\cdot,\cdot)\}$.

By the normality and joint independence in Assumption B.1, it is straightforward to show that the mean function $\mathbf{x}(\mathbf{s})^T\boldsymbol{\beta} + w(\mathbf{s})$ has a GP prior $\mathrm{GP}\big\{\mathbf{x}(\cdot)^T\boldsymbol{\mu}_{\boldsymbol{\beta}}, \check{C}_{\boldsymbol{\alpha}}(\cdot,\cdot)\big\}$, where the modified covariance function $\check{C}_{\boldsymbol{\alpha}}$ is given by

$$\begin{aligned}&\check{C}_{\boldsymbol{\alpha}}(\mathbf{s}_1,\mathbf{s}_2)\\ &= \mathrm{cov}\left\{\mathbf{x}(\mathbf{s}_1)^T\boldsymbol{\beta} + w(\mathbf{s}_1),\ \ \mathbf{x}(\mathbf{s}_2)^T\boldsymbol{\beta} + w(\mathbf{s}_2)\right\}\\ &= \mathbf{x}(\mathbf{s}_1)^T\Sigma_{\boldsymbol{\beta}}\,\mathbf{x}(\mathbf{s}_2) + C_{\boldsymbol{\alpha}}(\mathbf{s}_1,\mathbf{s}_2),\end{aligned} \tag{13}$$

for any $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{S}$. With this modified covariance function, we have the following corollary:

Corollary 3.3 *If Assumption B.1 holds, Assumptions A.1–A.5 hold with all $C_{\boldsymbol{\alpha}}$ replaced by $\check{C}_{\boldsymbol{\alpha}}$ in (13), and $\mu_{\boldsymbol{\beta}} = \mathbf{0}$, the conclusions of Theorems 3.1 and 3.2 hold for the Bayes L_2-risk of the mean surface $\mathbf{x}(\cdot)^T\boldsymbol{\beta} + w(\cdot)$ in the model (1).*

4. EXPERIMENTS

4.1 Simulation Setup

This section presents a comparative study of important non-distributed and distributed approaches on large spatial data based on the performance in learning the process parameters, interpolating the unobserved spatial surface, and predicting the response at new locations. Two simulation studies and a real data analysis are presented. The first simulation (*Simulation 1*) generates the data from a spatial linear model, where the spatial process is simulated from a GP with an exponential covariance function, leading to a fairly rough (nowhere differentiable) spatial surface. Following Gramacy and Apley (2015), we use an analytic function with local features to simulate the data in the second simulation (*Simulation 2*). The number of locations in the two simulations is moderately large with $n = 10,000$. Our real data analysis is based on a large data subset of sea surface temperature data with $n = 1,000,000$ locations. For the two simulations and in the real data analysis, the response at $(n + l)$ locations is modeled as

$$y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + w(\mathbf{s}_i) + \epsilon_i, \tag{14}$$

$\epsilon_i \sim N(0, \tau^2)$, $\mathbf{s}_i \in \mathcal{D} \subset \mathbb{R}^2$ for $i = 1, \ldots, n+l$, where $\mathcal{D}$ is the spatial domain, $y(\mathbf{s}_i)$, $x(\mathbf{s}_i)$, $w(\mathbf{s}_i)$, and ϵ_i are the response, covariate, spatial process, and idiosyncratic error values at the location $\mathbf{s}_i$, β_0 is the intercept, β_1 models the covariate effect, and l is the number of new locations where surface interpolation and prediction are sought.

A number of popular and state-of-the-art non-distributed Bayesian and non-Bayesian spatial models are compared with a few important distributed Bayesian approaches in the two simulations and in the real data analysis. Among non-distributed Bayesian and non-Bayesian methods, we fit: (i) Integrated nested Laplace approximation (INLA) using the `INLA` package in R (Illian et al., 2012); (ii) LatticeKrig (Nychka et al., 2015) using the

TABLE 1

The errors in estimating the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1), \sigma^2, \phi, \tau^2$ in Simulation 1. The parameter estimates for the Bayesian methods $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$, $\hat{\sigma}^2, \hat{\phi}$, $\hat{\tau}^2$ are defined as the posterior medians of their respective MCMC samples and their true values are $\boldsymbol{\beta}_0 = (1, 2)$, $\sigma_0^2 = 1, \phi_0 = 4$ and $\tau_0^2 = 0.1$. The entries in the table are averaged across 10 simulation replications.

	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$	$\lvert\hat{\sigma}^2 - \sigma_0^2\rvert$	$\lvert\hat{\phi} - \phi_0\rvert$	$\lvert\hat{\tau}^2 - \tau_0^2\rvert$
INLA	0.21	-	-	-
LaGP	0.08	-	-	-
NNGP ($m = 10$)	0.11	0.07	0.37	0.00
NNGP ($m = 20$)	0.12	0.09	0.51	0.00
NNGP ($m = 30$)	0.11	0.11	0.58	0.00
LatticeKrig	0.11	0.09	1.59	0.06
GpGp	0.08	0.11	0.64	0.01
Vecchia ($m = 10$)	0.10	0.11	0.51	0.01
Vecchia ($m = 20$)	0.10	0.10	0.55	0.01
Vecchia ($m = 30$)	0.10	0.38	1.13	0.01
MPP ($r = 200$)	0.35	0.23	1.98	0.17
MPP ($r = 400$)	0.19	0.09	1.88	0.07
	Random Partitioning			
DISK ($r = 200$, $k = 10$)	0.09	0.11	0.64	0.01
DISK ($r = 400$, $k = 10$)	0.09	0.11	0.64	0.01
DISK ($r = 200$, $k = 20$)	0.10	0.12	0.66	0.02
DISK ($r = 400$, $k = 20$)	0.10	0.12	0.66	0.02
	Grid-Based Partitioning			
DISK ($r = 200$, $k = 10$)	0.09	0.12	0.62	0.01
DISK ($r = 400$, $k = 10$)	0.09	0.12	0.62	0.01
DISK ($r = 200$, $k = 20$)	0.10	0.12	0.63	0.01
DISK ($r = 400$, $k = 20$)	0.10	0.12	0.64	0.01

`LatticeKrig` package in R with 3 resolutions (Nychka et al., 2016); (iii) modified predictive process (MPP) using the `spBayes` package in R with the full data; (iv) nearest neighbor Gaussian process (NNGP) using the `spNNGP` package in R with the number of nearest neighbors m set to be 10, 20, and 30 (Datta et al., 2016); (v) locally approximated Gaussian process (laGP) using the `laGP` package in R (Gramacy and Apley, 2015); (vi) Vecchia's approximation using the `GPvecchia` package in R with the number of nearest neighbors m set to be 10, 20, and 30 (Katzfuss and Guinness, 2021); (vii) Fisher Scoring of Vecchia's Approximation using the GpGp (Guinness, 2021).

In fitting (i), (ii), (iv), (v), (vi), (vii), we assume an exponential correlation in the random field given by $\text{cov}\{w(\mathbf{s}), w(\mathbf{s}')\} = \sigma^2 e^{-\phi\|\mathbf{s}-\mathbf{s}'\|}$, $\mathbf{s}, \mathbf{s}' \in \mathcal{D}$. To fit MPP for (iii), the MPP prior on $w(\cdot)$ is fitted with rank $r = 200, 400$ in Simulations 1, 2 and with $r = 400, 600$ in the real data analysis, where r knots are selected randomly from $\mathcal{D}$. For Bayesian model fitting, we apply a flat prior on (β_0, β_1), a IG$(2, 0.1)$ prior on τ^2, an IG$(2, 2)$ prior on σ^2 and a uniform prior on ϕ, where IG(a, b) is the Inverse-Gamma distribution with mean $b/(a-1)$.

The non-distributed approaches are compared with distributed Bayesian methods for model-free subset posterior aggregation discussed in Section 3 of this article. They are (viii) CMC (Scott et al. (2016)); (ix) DPMC (Xue and Liang (2019)); (x) WASP (Srivastava et al. (2015)); (xi) DISK (with $\xi = 10^{-4}$), for our exposition. Identical priors, covariance functions, ranks, and knots are used for the non-distributed process models and their distributed counterparts for a fair comparison. We emphasize that the distributed methods do not compete with the non-distributed methods in (i)-(vii). Instead, each of them can be potentially embedded in the second step of any of the distributed methods for improved performance because the distributed approaches are not model-specific. More importantly, MPP is not considered to be the state-of-the-art, so it is instructive to investigate the competitiveness of (viii)-(xi) with MPP fitted on each subset.

In the interest of space, we present the performance comparison between distributed and non-distributed approaches only, and similar comparisons between CMC, DISK, DPMC and WASP are presented in the supplementary material. Because DISK shows better or similar performance as its distributed competitors in all simulations, we only present results from DISK with the non-distributed methods in the main text. Notably, DISK combines one-dimensional marginals of subset posteriors, but

DPMC and WASP aggregate subset posteriors of multivariate parameters; therefore, similar performances of DISK, DPMC, and WASP in the supplementary material shows that combining subset posteriors of univariate parameters does not lead to any significant loss in inference or predictions.

Any distributed method has two important choices: (A) the value of k and (B) the construction of subsets. We choose k in our experiments based on two broad guidelines: (a) available computational resources and (b) the subset size to draw reliable inference on the spatial surface with data subsets. To assess (b), we plot the histograms or density estimates of subset posterior draws of representative parameters and see if they are very far from each other. If so, the subset posteriors fail to provide a noisy approximation of the full data posterior, resulting in inaccuracy of the combined pseudo posterior for a distributed approach. Empirically, we also propose computing the pairwise Wasserstein or total variation distance between the subset posteriors of representative parameters. If the average of these distances is much larger than the average distance between the combined and subset posterior distributions, then the combined pseudo posterior provides a poor approximation of the full data posterior. Assuming that the fitted model can reasonably capture variation of the data, these checks would imply that one has to fit a distributed approach with a smaller value of k.

Regarding (B), we present performance of the distributed approaches when data subsets are constructed (a) under a random partitioning scheme and (b) under a grid partitioning scheme. Random partitioning scheme randomly partitions the data into subsets. In contrast, grid partitioning scheme partitions the domain into a number of sub-domains and creates each subset with representative samples from each sub-domain. All tables in the main article and in supplementary material show results from both partitioning schemes.

We run all the experiments on an Oracle Grid Engine cluster with 2.6GHz 16 core compute nodes. The non-distributed methods (INLA, LatticeKrig, MPP, NNGP, laGP, GPvecchia, and GpGp) and the distributed methods (DISK, DPMC, CMC, and WASP) are allotted memory resources of 64GB and 16GB, respectively. Every MCMC algorithm runs for 10,000 iterations, out of which the first 5,000 MCMC samples are discarded as burn-ins and the rest of the chain is thinned by collecting every fifth MCMC sample. We also refer to Section 5 of the supplementary material that presents comparison between effective sample size of model parameters averaged over all subsets to the effective sample size of model parameters from the full data posterior in simulations. We compare the quality of prediction and estimation of spatial surface at predictive locations $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_l^*\}$. If $w(\mathbf{s}_{i'}^*)$ and $y(\mathbf{s}_{i'}^*)$ are the value of the spatial surface and response at $\mathbf{s}_{i'}^* \in \mathcal{S}^*$, then the estimation and prediction errors are defined as

$$\text{Est Err}^2 = \frac{1}{l}\sum_{i'=1}^{l}\{\hat{w}(\mathbf{s}_{i'}^*) - w(\mathbf{s}_{i'}^*)\}^2, \tag{15}$$

$$\text{Pred Err}^2 = \frac{1}{l}\sum_{i'=1}^{l}\{\hat{y}(\mathbf{s}_{i'}^*) - y(\mathbf{s}_{i'}^*)\}^2,$$

where $\hat{w}(\mathbf{s}_{i'}^*)$ and $\hat{y}(\mathbf{s}_{i'}^*)$ denote the point estimates of $w(\mathbf{s}_{i'}^*)$ and $y(\mathbf{s}_{i'}^*)$ obtained using any distributed or non-distributed methods. For sampling-based methods, we set $\hat{w}(\mathbf{s}_{i'}^*)$ and $\hat{y}(\mathbf{s}_{i'}^*)$ to be the medians of posterior MCMC samples for $w(\mathbf{s}_{i'}^*)$ and $y(\mathbf{s}_{i'}^*)$, respectively, for $i' = 1, \ldots, l$. We also estimate the point-wise 95% credible or confidence intervals (CIs) of $w(\mathbf{s}_{i'}^*)$ and predictive intervals (PIs) of $y(\mathbf{s}_{i'}^*)$ for every $\mathbf{s}_{i'} \in \mathcal{S}^*$ and compare the CI and PI coverages and lengths for every method. Finally, we compare the performance of all the methods for parameter estimation using the posterior medians and the 95% CIs. Posterior medians are reported instead of posterior means as point estimators since they are easily estimated for the DISK combined posterior following equation (8).

4.2 Simulation 1: Spatial Linear Model Based On GP

Our first simulation generates data using the spatial linear model in (14). We set $\mathcal{D} = [-2, 2] \times [-2, 2] \subset \mathbb{R}^2$, $n = 10,000$, $l = 500$ and uniformly draw $(n + l)$ spatial locations $\mathbf{s}_i = (s_{i1}, s_{i2})$ in $\mathcal{D}$ $(i = 1, \ldots, n + l)$. The spatial surface $w(\cdot)$ at the $(n + l)$ locations, $\{w(\mathbf{s}_1), \ldots, w(\mathbf{s}_{n+l})\}$, is simulated from $\text{GP}(0, \sigma^2 \exp\{-\phi \| \mathbf{s} - \mathbf{s}' \|)\}$, where $\mathbf{s}, \mathbf{s}' \in \mathcal{D}$, $\phi = 4$, and $\sigma^2 = 1$. The covariance function ensures the generated spatial surface is continuous every-

TABLE 2

The estimates of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1), \sigma^2, \phi, \tau^2$ and their 95% marginal credible intervals (CIs) in Simulation 1. The parameter estimates for the Bayesian methods $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$, $\hat{\sigma}^2, \hat{\phi}$, $\hat{\tau}^2$ are defined as the posterior medians of their respective MCMC samples. The parameter estimates and upper and lower quantiles of 95% CIs are averaged over 10 simulation replications; '-' indicates that the uncertainty estimates are not provided by the software or the competitor.

	β_0	β_1	σ^2	ϕ	τ^2
Truth	1.00	2.00	1.00	4.00	0.10
	Parameter Estimates				
INLA	1.00	2.00	-	-	-
laGP	1.01	2.00	-	-	-
NNGP ($m = 10$)	1.02	2.00	0.99	4.00	0.10
NNGP ($m = 20$)	0.98	2.00	0.94	4.30	0.10
NNGP ($m = 30$)	0.99	2.00	0.94	4.34	0.10
LatticeKrig	1.01	2.00	0.93	2.42	0.16
GpGp	0.99	2.00	0.92	4.43	0.11
Vecchia ($m = 10$)	0.99	2.00	0.94	3.93	0.09
Vecchia ($m = 20$)	0.99	2.00	0.95	3.93	0.09
Vecchia ($m = 30$)	1.00	2.00	1.10	3.68	0.09
MPP ($r = 200$)	1.26	2.00	0.77	2.02	0.27
MPP ($r = 400$)	1.08	2.00	0.99	2.14	0.17
DISK ($r = 200$, $k = 10$)	1.00	2.00	0.92	4.35	0.11
DISK ($r = 400$, $k = 10$)	1.00	2.00	0.92	4.35	0.11
DISK ($r = 200$, $k = 20$)	1.00	2.00	0.91	4.38	0.11
DISK ($r = 400$, $k = 20$)	1.00	2.00	0.91	4.38	0.11
	95% Credible Intervals				
INLA	(0.26, 1.73)	(1.98, 2.02)	-	-	-
laGP	(0.99, 1.03)	(1.98, 2.02)	-	-	-
NNGP ($m = 10$)	(0.87, 1.15)	(1.99, 2.01)	(0.86, 1.24)	(3.15, 4.70)	(0.09, 0.11)
NNGP ($m = 20$)	(0.85, 1.13)	(1.99, 2.01)	(0.82, 1.14)	(3.46, 4.95)	(0.09, 0.11)
NNGP ($m = 30$)	(0.86, 1.12)	(1.99, 2.01)	(0.81, 1.11)	(3.62, 5.03)	(0.09, 0.11)
LatticeKrig	-	-	-	-	-
GpGp	(0.75, 1.23)	(1.99, 2.01)	-	-	-
Vecchia ($m = 10$)	-	-	-	-	-
Vecchia ($m = 20$)	-	-	-	-	-
Vecchia ($m = 30$)	-	-	-	-	-
MPP ($r = 200$)	(1.06, 1.26)	(1.98, 2.00)	(0.70, 0.85)	(2.01, 2.07)	(0.24, 0.30)
MPP ($r = 400$)	(0.76, 1.08)	(1.99, 2.00)	(0.91, 1.08)	(2.07, 2.26)	(0.15, 0.19)
DISK ($r = 200$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(4.00, 4.69)	(0.09, 0.12)
DISK ($r = 400$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(4.00, 4.69)	(0.09, 0.12)
DISK ($r = 200$, $k = 20$)	(0.94, 1.06)	(1.98, 2.01)	(0.86, 0.96)	(4.07, 4.67)	(0.09, 0.13)
DISK ($r = 400$, $k = 20$)	(0.94, 1.06)	(1.99, 2.01)	(0.86, 0.96)	(4.07, 4.68)	(0.09, 0.13)

where but differentiable nowhere, which is a more familiar simulation scenario in the spatial context. Setting $\beta_0 = 1$, $\beta_1 = 2$, and $\tau^2 = 0.1$, we simulate the responses at $(n + l)$ locations using (14). The three-step distributed frameworks are applied using the low-rank MPP priors with $k = 10$ and $k = 20$, having average subset sizes 1000 and 500, respectively. We replicate this simulation ten times.

DISK with MPP prior, NNGP, and GPvecchia have similar performance in parameter estimation (Tables 1 and 2). The parameter estimates obtained using DISK are very close to their true values and the estimation errors are very similar to those of NNGP and non-Bayesian methods based on the Vecchia approximation, including GpGp and GPvecchia. The 95% credible intervals of β_0, β_1, τ^2 in DISK cover the true values and their lower and upper quantiles are very similar to those of NNGP. DISK underestimates σ^2 and overestimates ϕ slightly. Both results are the impacts of parent MPP prior, which also shows less accurate estimation of the posterior distribution of σ^2 and ϕ for the two choices of the number of knots r. More importantly, the impacts the choice of r on parameter estimation are less severe in the distributed methods compared to that in its parent MPP prior. The CIs are not available from GPvecchia, LatticeKrig and laGP, so that the cells corresponding these methods are kept blank in Table 2.

Despite the discrepancy in parameter estimates, the correlation function estimates obtained using the combined posteriors from distributed competitors (DISK pseudo posterior being a representative) are very close to those obtained using NNGP and

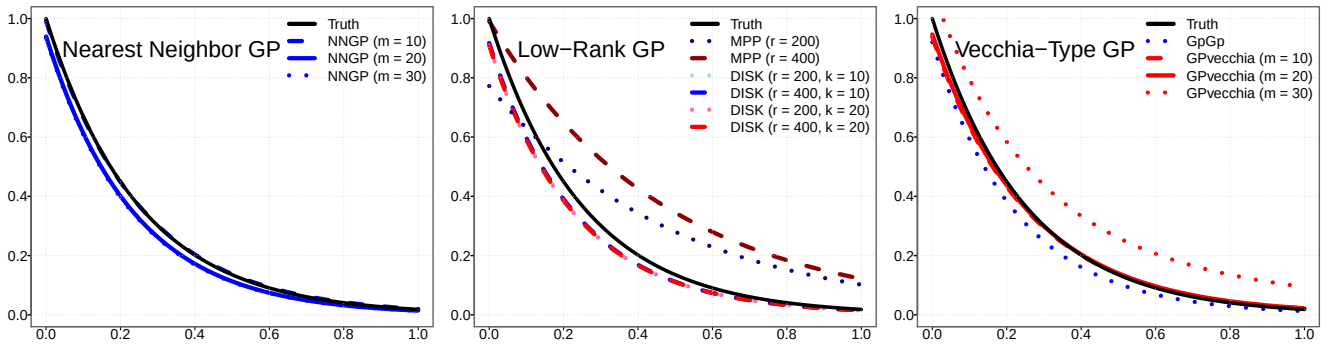

Fig 1: Estimated covariance function using three types of GP priors on the spatial surface. The true covariance function is $\text{cov}\{w(\mathbf{s}_i), w(\mathbf{s}_j)\} = \exp(-4\| \mathbf{s}_i - \mathbf{s}_j \|_2)$.

GPvecchia (Figure 1). Similar to the observations of Sang and Huang (2012), there is considerable discrepancy between the estimated and true correlation functions when the MPP prior is used. In contrast, for the same choices of r as its parent MPP prior, DISK's estimate of the correlation function is much closer to the truth and is insensitive to the choice of $k = 10, 20$. DISK estimates are similar to those obtained using Vecchia-type approximation, except when the number of nearest neighbor is 30 and the GPvecchia-based estimate of the correlation function has a significant positive bias.

The predictive performance of the representative distributed competitor DISK is little inferior to that of NNGP. NNGP, MPP, and DISK have close to nominal predictive coverage, but the PIs of NNGP have smaller lengths for every choice of nearest neighbor. The PI coverage values and lengths of MPP and DISK are similar and stable for the different choices of r and k. PIs in GPvecchia have the smallest length and their coverage values are smaller than the nominal value for all the three choices of nearest neighbor. Focusing on spatial surface interpolation, the estimation error of DISK is smaller than that of MPP for both choices of r when $k = 10$ and is slightly larger when $k = 20$ and $r = 400$. Similarly, MPP's coverage of the spatial surface is smaller than the nominal value when $r = 200$, but DISK shows better coverage than its parent MPP prior for both choices of k. Consequently, the lengths of DISK's credible intervals are slightly larger than those obtained using its parent MPP prior.

In summary, the distributed methods are competitive with state-of-the-art non-distributed methods NNGP and GPvecchhia in inference on the spatial surface and predictions, respectively. laGP is the only non-distributed competing method that yields comprehensively better inferential and predictive performance than all distributed methods, but it is not designed to provide estimates for the σ^2, ϕ, and τ^2. LatticeKrig has a very similar point estimation, but inferior uncertainty quantification compared to GpGp and GPvecchia. INLA underperforms in surface interpolation and prediction. Supplementary material shows comparative performance of distributed competitors and also ensures that stochastic approximation does not impact the mixing of the Markov chains on the subsets. The model free nature of the distributed methods also allows us to fit a nearest neighbor approach, including NNGP, on each subset to improve inference and expedite computations by multiple folds. Finally, the results show that the random partitioning scheme yields little better point estimation with similar uncertainty quantification compared to a more sophisticated grid partitioning scheme.

4.3 Simulation 2: Spatial Linear Model Based On Analytic Spatial Surface

Our second simulation generates data by setting $w(\cdot)$ in (14) to be an analytic function. For any $s \in [-2, 2]$, define the function $f_0(s) = e^{-(s-1)^2} + e^{-0.8(s+1)^2} - 0.05\sin\{8(s + 0.1)\}$ and set $w(\mathbf{s}_i) = -f_0(s_{i1})f_0(s_{i2})$. Although the function $w(\cdot)$ simulated in this way is theoretically infinitely smooth, the response surface simulated from (14) exhibits complex local behavior, which is challenging to capture using spatial process-based models as we

TABLE 3

Inference on the values of spatial surface and response at the locations in $\mathcal{S}_*$ in Simulation 1. The estimation and prediction errors are defined in (15) and coverage and credible intervals are calculated pointwise for the locations in $\mathcal{S}_*$. The entries in the table are averaged over 10 simulation replications; '-' indicates that the estimates are not provided by the software or the competitor.

	Est Err	Pred Err	95% CI Coverage		95% CI Length	
	GP	Y	GP	Y	GP	Y
INLA	-	0.90	-	0.80	-	0.17
laGP	0.20	0.28	0.98	0.95	2.06	1.04
NNGP ($m = 10$)	0.38	0.47	0.93	0.95	1.39	1.84
NNGP ($m = 20$)	0.38	0.47	0.93	0.95	1.38	1.81
NNGP ($m = 30$)	0.38	0.47	0.92	0.95	1.37	1.82
LatticeKrig	0.38	0.47	-	0.73	-	1.08
GpGp	-	0.47	-	-	-	-
Vecchia ($m = 10$)	-	0.47	-	0.87	-	1.43
Vecchia ($m = 20$)	-	0.47	-	0.86	-	1.41
Vecchia ($m = 30$)	-	0.47	-	0.86	-	1.41
MPP ($r = 200$)	0.73	0.59	0.93	0.95	3.05	3.02
MPP ($r = 400$)	0.43	0.47	0.96	0.95	2.76	2.67
	Random Partitioning					
DISK ($r = 200$, $k = 10$)	0.55	0.64	0.97	0.97	3.20	3.45
DISK ($r = 400$, $k = 10$)	0.42	0.51	0.97	0.97	2.88	3.15
DISK ($r = 200$, $k = 20$)	0.58	0.67	0.97	0.97	3.25	3.51
DISK ($r = 400$, $k = 20$)	0.46	0.55	0.97	0.97	2.98	3.25
	Grid-Based Partitioning					
DISK ($r = 200$, $k = 10$)	0.75	0.80	0.97	0.97	3.45	3.45
DISK ($r = 400$, $k = 10$)	0.65	0.72	0.97	0.97	3.15	3.15
DISK ($r = 200$, $k = 20$)	0.76	0.82	0.97	0.97	3.51	3.51
DISK ($r = 400$, $k = 20$)	0.68	0.74	0.97	0.97	3.26	3.26

demonstrate later. We set $\beta_0 = 1$, $\beta_1 = 0$, and $\tau^2 = 0.01$, use the same values of the spatial domain, k, and r as used in the previous simulation, and replicate this simulation 10 times.

The parameter estimation results in this simulation are similar to those in Simulation 1 with one important exception in inference on β_0 (Tables 4 and 5). All the methods except GpGp show excellent performance in estimating τ^2; however, NNGP, GPvecchia, and MPP estimate β_0 with a significant bias. 95% credible intervals of β_0 computed from DISK has better coverage properties than those of NNGP. Unlike our observation in the previous section, all the methods underestimate τ^2 slightly, and the 95% credible intervals of NNGP, MPP prior, and DISK fail to cover the true value. Similar to the previous simulation results, DISK performs better than its parent MPP prior for both choices of r.

The predictive and inferential performance of distributed methods in this simulation are also very similar to those in Simulation 1. The prediction error, PI coverage, and PI length of all the methods except GPvecchia are fairly similar and are close to the nominal value. The PI length of GPvecchia is the smallest, but its coverage values are critically low for all choices of nearest neighbor; that is, GPvecchia has a relatively inferior performance for estimating spatial surfaces that are not simulated from a GP. The PI coverage values of distributed method DISK is a little higher than those of NNGP and MPP priors while the PI lengths of DISK are very close to those of MPP and NNGP priors. A noticeable feature of our comparison is that the distributed methods improve the performance of their parent MPP prior when $r = 200$. In this case, the CI coverage values of distributed methods for both choices of k are greater the nominal value, whereas the parent MPP prior has fails to cover the spatial surface. Intuitively, for most competitors in this simulation the estimation of fixed and random effects are mixed up, whereas the overall mean effect is estimated correctly by all competitors.

As in Simulation 1, INLA still underperforms in surface interpolation and prediction, and laGP maintains its superior predictive and inferential performance, especially because it is tuned for inference in such analytic surfaces with many local features (Gramacy and Apley, 2015). LatticeKrig also offers excellent performance and it outperforms the distributed methods in terms of surface estimation and

TABLE 4

The errors in estimating the parameters β, τ^2 in Simulation 2. The parameter estimates for the Bayesian methods $\hat{\beta}$, $\hat{\tau}^2$ are defined as the posterior medians of their respective MCMC samples and $\beta_0 = 1$ and $\tau_0^2 = 0.01$. The entries in the table are averaged across 10 simulation replications.

	$\|\hat{\beta} - \beta_0\|$	$\|\hat{\tau}^2 - \tau_0^2\|$
INLA	0.18	-
LaGP	-	-
NNGP ($m = 10$)	0.84	0.03
NNGP ($m = 20$)	0.84	0.03
NNGP ($m = 30$)	0.84	0.03
LatticeKrig	-	0.01
GpGp	0.31	0.39
Vecchia ($m = 10$)	0.85	0.01
Vecchia ($m = 20$)	0.85	0.01
Vecchia ($m = 30$)	0.85	0.01
MPP ($r = 200$)	0.75	0.05
MPP ($r = 400$)	0.48	0.04
	Random Partitioning	
DISK ($r = 200$, $k = 10$)	0.18	0.04
DISK ($r = 400$, $k = 10$)	0.13	0.04
DISK ($r = 200$, $k = 20$)	0.18	0.04
DISK ($r = 400$, $k = 20$)	0.13	0.04
	Grid-Based Partitioning	
DISK ($r = 200$, $k = 10$)	0.03	0.09
DISK ($r = 400$, $k = 10$)	0.03	0.09
DISK ($r = 200$, $k = 20$)	0.02	0.09
DISK ($r = 400$, $k = 20$)	0.02	0.09

prediction. Simulation 2 shows that grid based partitioning yields better point estimation for β_0, but inferior point estimation for τ_0^2 (Table 4). This leads to little better surface estimation for random partitioning scheme than grid-based partitioning scheme, but practically indistinguishable predictive performance as demonstrated in Table 6. We conclude that the distributed methods are promising tools even when the spatial surface is not simulated from a GP.

4.4 Real Data Analysis: Sea Surface Temperature Data

A description of the evolution and dynamics of the SST is a key component of the study of the Earth's climate. SST data (in centigrade) from ocean samples have been collected by voluntary observing ships, buoys, and military and scientific cruises for decades. During the last 20 years or so, the SST database has been complemented by regular streams of remotely sensed observations from satellite orbiting the earth. A careful quantification of variability of SST data is important for climatological research, which includes determining the formation of sea breezes and sea fog and calibrating measurements from weather satellites (Di Lorenzo et al., 2008). A number of articles have appeared to address this issue in recent years; see Berliner et al. (2000), Lemos and Sansó (2009), Wikle and Holan (2011), Hazra and Huser (2021).

We consider the problem of capturing the spatial trend and characterizing the uncertainties in the SST in the west coast of mainland U.S.A., Canada, and Alaska between 40°–65° north latitudes and 100°–180° west longitudes. The data is obtained from NODC World Ocean Database (`https://www.nodc.noaa.gov/OC5/WOD/pr_wod.html`) and the entire data corresponds to sea surface temperature measured by remote sensing satellites on 16th August 2016. All data locations are distinct and there is no time replicate; therefore, we can practically ignore the temporal variation of sea surface temperature for our analysis. After screening the data for quality control, we choose a random subset of $1,000,800$ spatial observations over the selected domain. From these observations, we randomly select 10^6 observations as training data and the remaining observations are used to compare the performance of distributed and non-distributed competitors. We replicate this setup ten times. The selected domain 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 analyzing global-scale SST database.

The SST data in the selected domain shows a

TABLE 5

The estimates of parameters $\beta, \sigma^2, \phi, \tau^2$ and their 95% marginal credible intervals (CIs) in Simulation 2. The parameter estimates for the Bayesian methods $\hat{\beta}$, $\hat{\sigma}^2, \hat{\phi}$, $\hat{\tau}^2$ are defined as the posterior medians of their respective MCMC samples. The parameter estimates and upper and lower quantiles of 95% CIs are averaged over 10 simulation replications; '-' indicates that the uncertainty estimates are not provided by the software or the competitor.

	β	σ^2	ϕ	τ^2
Truth	1.00	-	-	0.01
	Parameter Estimates			
INLA	0.8161	-	-	-
laGP	-	-	-	-
NNGP ($m = 10$)	0.2897	0.1933	0.1075	0.0091
NNGP ($m = 20$)	0.3002	0.1660	0.1059	0.0092
NNGP ($m = 30$)	0.2892	0.1557	0.1058	0.0093
LatticeKrig	-	-	0.0842	0.0099
GpGp	1.0346	0.0669	0.2643	0.1620
Vecchia ($m = 10$)	0.2792	0.4063	0.7796	0.0099
Vecchia ($m = 20$)	0.2792	0.2904	0.9479	0.0099
Vecchia ($m = 30$)	0.2792	0.2746	0.9587	0.0099
MPP ($r = 200$)	1.5634	0.1535	0.1185	0.0077
MPP ($r = 400$)	1.2333	0.1586	0.1200	0.0080
DISK ($r = 200$, $k = 10$)	1.0322	0.2133	0.1196	0.0087
DISK ($r = 400$, $k = 10$)	0.9830	0.2185	0.1402	0.0082
DISK ($r = 200$, $k = 20$)	1.0328	0.2133	0.1194	0.0087
DISK ($r = 400$, $k = 20$)	0.9822	0.2185	0.1402	0.0082
	95% Credible Intervals			
INLA	(0.5320, 1.2108)	-	-	-
laGP	-	-	-	-
NNGP ($m = 10$)	(0.2678, 0.3143)	(0.1568, 0.2223)	(0.1010, 0.1339)	(0.0088, 0.0094)
NNGP ($m = 20$)	(0.2801, 0.3226)	(0.1361, 0.1906)	(0.1009, 0.1279)	(0.0089, 0.0095)
NNGP ($m = 30$)	(0.2660, 0.3103)	(0.1293, 0.1794)	(0.1009, 0.1284)	(0.0090, 0.0095)
LatticeKrig	-	-	-	-
GpGp	(0.7090, 1.3601)	-	-	-
Vecchia ($m = 10$)	-	-	-	-
Vecchia ($m = 20$)	-	-	-	-
Vecchia ($m = 30$)	-	-	-	-
MPP ($r = 200$)	(0.9931, 2.1464)	(0.1307, 0.1760)	(0.1104, 0.1327)	(0.0073, 0.0081)
MPP ($r = 400$)	(0.6130, 1.8412)	(0.1269, 0.1876)	(0.1096, 0.1480)	(0.0076, 0.0084)
DISK ($r = 200$, $k = 10$)	(0.7961, 1.2722)	(0.1783, 0.2418)	(0.1088, 0.1439)	(0.0084, 0.0091)
DISK ($r = 400$, $k = 10$)	(0.8180, 1.1582)	(0.1743, 0.2589)	(0.1192, 0.1773)	(0.0079, 0.0086)
DISK ($r = 200$, $k = 20$)	(0.7987, 1.2719)	(0.1781, 0.2417)	(0.1087, 0.1434)	(0.0084, 0.0091)
DISK ($r = 400$, $k = 20$)	(0.8172, 1.1568)	(0.1721, 0.2588)	(0.1190, 0.1806)	(0.0079, 0.0086)

clear decreasing trend in SST with increasing latitude (Figure 2). Based on this observation, we add latitude as a linear predictor in the univariate spatial regression model (14) to explain the long-range directional variability in the SST. Similar to Simulation 1 and 2, Section 4.4 in the supplementary material shows that among distributed competitors DISK shows identical or little better performance than the other distributed approaches for the sea surface data. Thus, we only present results from DISK in this section due to space constraint considering it as a representative distributed competitor. The detailed performance comparison of all distributed competitors in the real data can be found in Section 4.4 of the supplementary material. To fit distributed competitors, we set $k = 300$, which results in subsets of approximately 3300 locations. Since each subset has larger sample size than the simulation studies, we increase the number of knots in each subset for model fitting and use MPP priors with 400 and 600 knots, respectively, in each subset. All the non-distributed competitors except laGP fail to produce results due to numerical issues. Specifically, GPvecchia and GpGp fail after 8 and 21 iterations with an error in `vecchia_Linv` function, INLA fails with an error in `GMRFLib_factorise_sparse_matrix_TAUCS` function, spNNGP fails an error in the `dpotrf` function, and MPP fails from memory bottlenecks. Due to the lack of ground truth for estimating $w(\mathbf{s}^*)$, we compare the DISK and laGP in terms of their inference on $\mathbf{\Omega}$ and prediction of $y(\mathbf{s}^*)$ for $\mathbf{s}^* \in \mathcal{S}^*$ in terms of MSPE and the length and coverage of 95% posterior PIs.

DISK provides inference on the covariance func-

TABLE 6

Inference on the values of spatial surface and response at the locations in $\mathcal{S}_*$ in Simulation 2. The estimation and prediction errors are defined in (15) and coverage and credible intervals are calculated pointwise for the locations in $\mathcal{S}_*$. The entries in the table are averaged over 10 simulation replications; '-' indicates that the estimates are not provided by the software or the competitor.

	Est Err	Pred Err	95% CI Coverage		95% CI Length	
	GP	Y	GP	Y	GP	Y
INLA	-	0.1552	-	0.0755	-	0.0268
laGP	0.0004	0.0103	1.0000	0.9456	0.3890	0.3902
NNGP ($m = 10$)	0.5058	0.0104	0.0000	0.9439	0.1496	0.3949
NNGP ($m = 20$)	0.4908	0.0103	0.0000	0.9456	0.1392	0.3938
NNGP ($m = 30$)	0.5103	0.0103	0.0000	0.9479	0.1388	0.3969
LatticeKrig	0.0002	0.0101	0.9867	0.9463	-	0.3901
GpGp	-	0.0103	-	-	-	-
Vecchia ($m = 10$)	-	0.0106	-	0.3559	-	0.0951
Vecchia ($m = 20$)	-	0.0103	-	0.2815	-	0.0728
Vecchia ($m = 30$)	-	0.0102	-	0.2612	-	0.0674
MPP ($r = 200$)	0.3732	0.0105	0.0000	0.9498	0.4061	0.4061
MPP ($r = 400$)	0.0623	0.0102	0.2946	0.9477	0.3976	0.3976
	Random Partitioning					
DISK ($r = 200$, $k = 10$)	0.0017	0.1035	1.0000	0.9696	0.5388	0.4449
DISK ($r = 400$, $k = 10$)	0.0009	0.1026	1.0000	0.9724	0.4477	0.4578
DISK ($r = 200$, $k = 20$)	0.0015	0.1041	1.0000	0.9646	0.5211	0.4248
DISK ($r = 400$, $k = 20$)	0.0007	0.1031	1.0000	0.9672	0.4253	0.4359
	Grid-Based Partitioning					
DISK ($r = 200$, $k = 10$)	0.0394	0.1036	1.0000	0.9660	0.4452	0.4452
DISK ($r = 400$, $k = 10$)	0.0368	0.1028	1.0000	0.9594	0.4249	0.4249
DISK ($r = 200$, $k = 20$)	0.0304	0.1040	1.0000	0.9700	0.4590	0.4590
DISK ($r = 400$, $k = 20$)	0.0268	0.1030	1.0000	0.9642	0.4371	0.4371

tion, including credible intervals for σ^2, ϕ, and τ^2, which are unavailable in laGP. The 50%, 2.5%, and 97.5% quantiles of the posterior distributions for $\mathbf{\Omega}$, $w(\mathbf{s}^*)$ and $y(\mathbf{s}^*)$ for every $\mathbf{s}^* \in \mathcal{S}^*$ are used for estimation and uncertainty quantification. We observe significantly higher estimation of spatial variability than non-spatial variability from DISK indicating local spatial variation in SST. Importantly, the point estimate of β_1 is negative and its 95% CI does not include zero, which confirms that SST decreases as latitude increases. For every $\mathbf{s}^* \in \mathcal{S}^*$, laGP's and DISK's estimates of $w(\mathbf{s}^*)$ and $y(\mathbf{s}^*)$ agree closely (Figures 2 and 3 and Table 7). The pointwise predictive coverages of laGP and DISK match their nominal levels; however, the 95% posterior PIs of DISK are wider than those of laGP because DISK accounts for uncertainty due to the error term and unknown parameters (Figure 2 and Table 7). As a whole, SST data analysis reinforces our findings on the importance of distributed Bayesian methods as computationally efficient and flexible tools for full Bayesian inference.

5. DISCUSSION

This article presents a comparative study of a class of distributed Bayesian and popular non-distributed methods that are tuned for spatial GP regression in massive data settings. As part of our exposition, we have demonstrated through simulated and real data analyses that distributed Bayesian methods compare well with state-of-the-art non-distributed methods. Motivated by the promising empirical performance, we provide theoretical support for our numerical results. In particular, under commonly assumed regularity conditions, we have provided explicit upper bound on the number of subsets k depending on the analytic properties of the spatial surface so that the Bayes L_2-risk of the combined pseudo posterior for a subclass of distributed methods is nearly minimax optimal. Additional empirical and theoretical results in the supplementary material shed light on the relative empirical performances of different distributed Bayesian methods in simulations and in the real data analyses.

The simplicity and generality of distributed frameworks enable scaling of any spatial model. For example, recent applications have confirmed that the NNGP prior requires modifications if scalability is

TABLE 7

Parametric inference and prediction in SST data. DISK uses MPP-based modeling with $r = 400, 600$ on $k = 300$ subsets. For parametric inference posterior medians are provided along with The 95% credible intervals (CIs) in the parentheses, where available. Similarly mean squared prediction errors (MSPEs) along with length and coverage of 95% predictive intervals (PIs) are presented, where available. The upper and lower quantiles of 95% CIs and PIs are averaged over 10 simulation replications; '-' indicates that the parameter estimate or prediction is not provided by the software or the competitor

	β_0	β_1	σ^2	ϕ	τ^2
	Parameter Estimate				
laGP	32.98	-0.37	-	-	-
DISK ($r = 400$)	32.33	-0.32	11.82	0.04	0.18
DISK ($r = 600$)	32.33	-0.32	11.85	0.04	0.18
	95% Credible Interval				
laGP	-	-	-	-	-
DISK ($r = 400$)	(31.72, 32.93)	(-0.33, -0.31)	(11.24, 12.43)	(0.0373, 0.0412)	(0.18, 0.19)
DISK ($r = 600$)	(31.72, 32.93)	(-0.33, -0.31)	(11.25, 12.45)	(0.0372, 0.0413)	(0.18, 0.19)
	Predictions				
	MSPE	95% PI Coverage	95% PI Length		
laGP	0.24	0.95	1.35		
DISK ($r = 400$)	0.43	0.95	2.65		
DISK ($r = 600$)	0.36	0.95	2.34		

desired for even a few millions of locations (Finley et al., 2019). While computing subset posteriors with MCMC algorithm, we have tacitly assumed that the MCMC chain converges to the subset posterior. While there is no theoretical result to support this, we find enough empirical evidence regarding convergence of the MCMC chain for each subset posterior. We plan to explore this issue theoretically in a future work. In future, we also aim to scale ordinary NNGP and other multiscale approaches to tens of millions of locations with distributed frameworks.

We have focused on developing the distributed framework for spatial modeling due to the motivating applications from massive geostatistical data. The distributed frameworks, however, are applicable to any mixed effects model where the random effects are assigned a GP prior, which includes Bayesian nonparametric regression using GP prior. We plan to explore more general applications in the future with high dimensional covariates.

ACKNOWLEDGEMENTS

The four authors contributed equally to this work. Rajarshi Guhaniyogi's and Sanvesh Srivastava's research are partially supported by from Office of Naval Research award no. N00014-18-1-2741 and National Science Foundation DMS-2220840/1854667. Cheng Li's research is supported by Singapore Ministry of Education Academic Research Funds Tier 1 Grants R155000172133 and R155000201114.

REFERENCES

Agueh, M. and G. Carlier (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis 43*(2), 904–924.

Anderes, E., R. Huser, D. Nychka, and M. Coram (2013). Nonstationary positive definite tapering on the plane. *Journal of Computational and Graphical Statistics 22*(4), 848–865.

Anderson, C., D. Lee, and N. Dean (2014). Identifying clusters in Bayesian disease mapping. *Biostatistics 15*(3), 457–469.

Bai, Y., P. X.-K. Song, and T. Raghunathan (2012). Joint composite estimating functions in spatiotemporal models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(5), 799–824.

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.

Barbian, M. H. and R. M. Assunção (2017). Spatial subsemble estimator for large geostatistical data. *Spatial Statistics 22*, 68–88.

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.

Bevilacqua, M., C. Caamaño-Carrillo, and E. Porcu (2022). Unifying compactly supported and matérn covariance functions in spatial statistics. *Journal of Multivariate Analysis 189*, 104949.

Bevilacqua, M. and C. Gaetan (2015). Comparing composite likelihood methods based on pairs for spatial Gaussian random fields. *Statistics and Computing 25*(5), 877–892.

Bolin, D. and F. Lindgren (2013). A comparison between markov approximations and other methods for large spatial data sets. *Computational Statistics & Data Analysis 61*, 7–21.

Bolin, D. and J. Wallin (2020). Multivariate type G Matérn stochastic partial differential equation random fields. *Jour-*

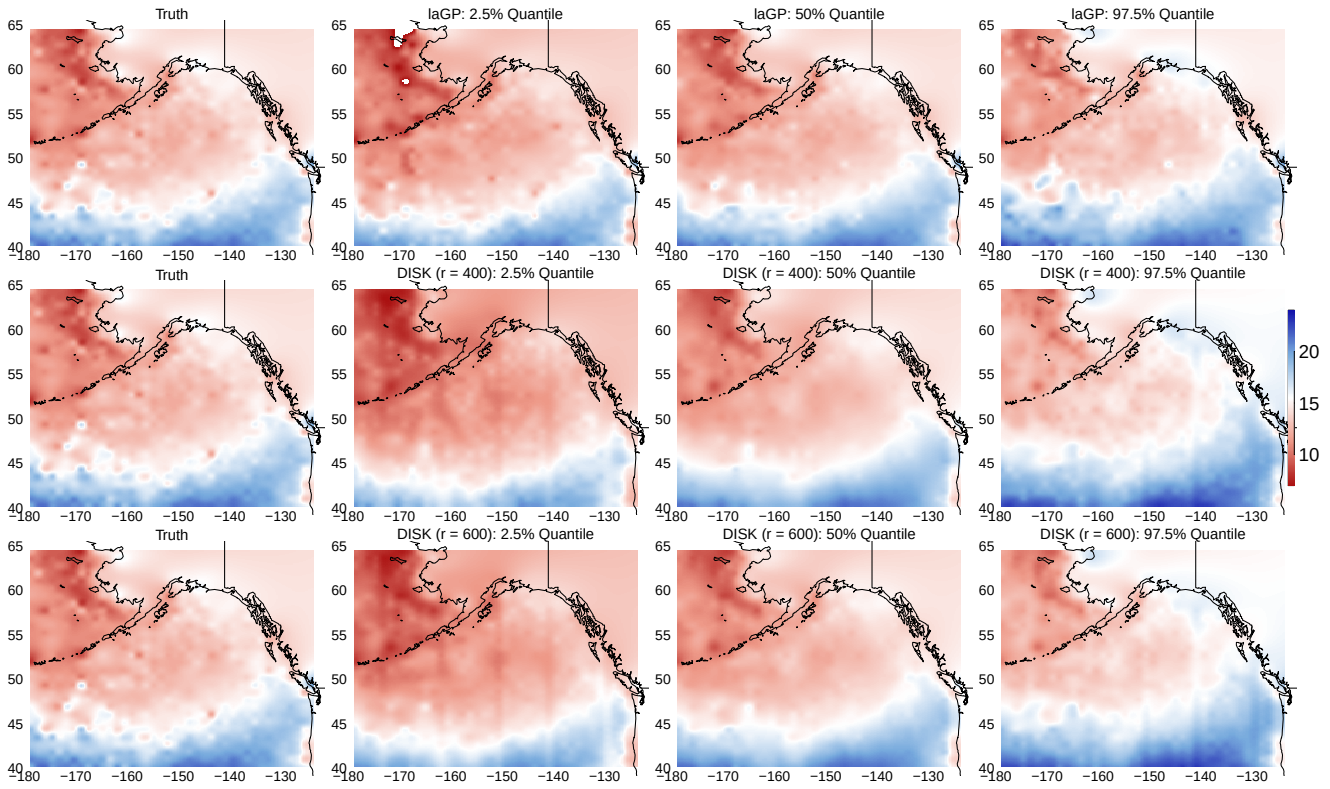

Fig 2: Prediction of sea surface temperatures at the locations in $\mathcal{S}^*$. Negative longitude means degree west from Greenwich. DISK uses MPP-based modeling with $r = 400,600$ on $k = 300$ subsets and laGP uses the 'nn' method. The 2.5%, 50%, and 97.5% quantile surfaces, respectively, represent pointwise quantiles of the posterior distribution for $y(\mathbf{s}^*)$ for every $\mathbf{s}^* \in \mathcal{S}^*$.

nal of the Royal Statistical Society: Series B (Statistical Methodology) 82(1), 215–239.

Chandler, R. E. and S. Bate (2007). Inference for clustered data using the independence loglikelihood. *Biometrika 94*(1), 167–183.

Cheng, G. and Z. Shang (2017). Computational limits of divide-and-conquer method. *Journal of Machine Learning Research 18*, 1–37.

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. Wikle (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ.

Cuturi, M. and A. Doucet (2014). Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning, JMLR W&CP*, Volume 32.

Daley, D. J., E. Porcu, and M. Bevilacqua (2015). Classes of compactly supported covariance functions for multivariate random fields. *Stochastic Environmental Research and Risk Assessment 29*(4), 1249–1263.

Datta, A., S. Banerjee, A. O. Finley, and A. E. Gelfand (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association 111*(514), 800–812.

Di Lorenzo, E., N. Schneider, K. Cobb, P. Franks, K. Chhak, A. Miller, J. McWilliams, S. Bograd, H. Arango, E. Curchitser, et al. (2008). North Pacific gyre oscillation links ocean climate and ecosystem change. *Geophysical Research Letters 35*(8).

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., A. Datta, B. D. Cook, D. C. Morton, H. E. Andersen, and S. Banerjee (2019). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *Journal of Computational and Graphical Statistics 28*(2), 401–414.

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

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

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 B. Haaland (2016). Speeding up neighborhood search in local Gaussian process prediction. *Technometrics 58*(3), 294–303.

Guhaniyogi, R. and S. Banerjee (2018). Meta-kriging: Scalable Bayesian modeling and inference for massive spatial datasets. *Technimetrics 60*(4), 430–444.

Guhaniyogi, R. and S. Banerjee (2019). Multivariate spatial meta kriging. *Statistics & probability letters 144*, 3–8.

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.

Guhaniyogi, R. and B. Sanso (2020). Large multiscale spatial modeling using tree shrinkage priors. *Statistica Sinica 30*(4), 2023–2050.

Guinness, J. (2018). Permutation methods for sharpening Gaussian process approximations. *Technometrics 60*(4), 415–429.

Guinness, J. (2021). Gaussian process learning via Fisher scoring of Vecchia's approximation. *Statistics and Computing 31*(3), Article:25.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*, Volume 1. Springer.

Hazra, A. and R. Huser (2021). Estimating high-resolution red sea surface temperature hotspots, using a low-rank semiparametric spatial model. *The Annals of Applied Statistics 15*(2), 572–596.

Heaton, M. J., W. F. Christensen, and M. A. Terres (2017). Nonstationary Gaussian process models using spatial hierarchical clustering from finite differences. *Technometrics 59*(1), 93–101.

Heaton, M. J., A. Datta, A. Finley, R. Furrer, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, F. Lindgren, D. W. Nychka, F. Sun, and A. Zammit-Mangion (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics 24*, 398–425.

Illian, J. B., S. H. Sørbye, and H. Rue (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics*, 1499–1530.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association 112*(517).

Katzfuss, M. and J. Guinness (2021). A general framework for Vecchia approximations of Gaussian processes. *Statistical Science 36*(1), 124–141.

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.

Knorr-Held, L. and G. Raßer (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics 56*(1), 13–21.

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.

Li, C., S. Srivastava, and D. B. Dunson (2017). Simple, scalable and accurate posterior interval estimation. *Biometrika 104*(3), 665–680.

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

Lindsten, F., A. M. Johansen, C. A. Naesseth, B. Kirkpatrick, T. B. Schön, J. Aston, and A. Bouchard-Côté (2017). Divide-and-conquer with sequential Monte Carlo. *Journal of Computational and Graphical Statistics 26*(2), 445–458.

Mehrotra, S., H. Brantley, J. Westman, L. Bangerter, and A. Maity (2021). Divide-and-Conquer MCMC for Multivariate Binary Data. *arXiv preprint arXiv:2102.09008*.

Minsker, S. (2019). Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics 13*(2), 5213–5252.

Minsker, S., S. Srivastava, L. Lin, and D. Dunson (2014). Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1656–1664.

Minsker, S., S. Srivastava, L. Lin, and D. B. Dunson (2017). Robust and scalable Bayes via a median of subset posterior measures. *The Journal of Machine Learning Research 18*(1), 4488–4527.

Neiswanger, W., C. Wang, and E. Xing (2014). Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the 30th International Conference on Uncertainty in Artificial Intelligence*, pp. 623–632.

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.

Nychka, D., D. Hammerling, S. Sain, and N. Lenssen (2016). LatticeKrig: Multiresolution kriging based on Markov random fields. R package version 6.4.

Quiñonero-Candela, J. and C. E. Rasmussen (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research 6*(Dec), 1939–1959.

Raskutti, G., M. J. Wainwright, and B. Yu (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research 13*(Feb), 389–427.

Ribatet, M., D. Cooley, and A. C. Davison (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 813–845.

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.

Santin, G. and R. Schaback (2016). Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics 42*(4), 973–993.

Savitsky, T. D. and S. Srivastava (2018). Scalable Bayes un-

der informative sampling. *Scandinavian Journal of Statistics 45*(3), 534–556.

Scott, S. L., A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch (2016). Bayes and big data: the consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management 11*(2), 78–88.

Shang, Z., B. Hao, and G. Cheng (2019). Nonparametric Bayesian aggregation for massive data. *Journal of Machine Learning Research 20*, 1–81.

Simpson, D., F. Lindgren, and H. Rue (2012). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics 23*(1), 65–74.

Srivastava, S., V. Cevher, Q. Dinh, and D. Dunson (2015). WASP: Scalable Bayes via barycenters of subset posteriors. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 912–920.

Srivastava, S., C. Li, and D. B. Dunson (2018). Scalable Bayes via barycenter in Wasserstein space. *Journal of Machine Learning Research 19*, 1–35.

Srivastava, S. and Y. Xu (2021). Distributed Bayesian inference in linear mixed-effects models. *Journal of Computational and Graphical Statistics 30*(3), 594–611.

Stein, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics 8*, 1–19.

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.

Su, Y. (2020). A divide and conquer algorithm of Bayesian density estimation. *arXiv preprint arXiv:2002.07094*.

Szabo, B. and H. van Zanten (2019). An asymptotic analysis of distributed nonparametric methods. *Journal of Machine Learning Research 20*, 1–30.

van der Vaart, A. and H. van Zanten (2011). Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research 12*, 2095–2119.

van der Vaart, A. W. and J. H. van Zanten (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pp. 200–222. Institute of Mathematical Statistics.

Van Trees, H. L. (2001). *Detection, Estimation, and Modulation Theory*. John Wiley & Sons.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica*, 5–42.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 50*(2), 297–312.

Wang, C. and S. Srivastava (2021). Divide-and-Conquer Bayesian inference in hidden Markov models. *arXiv preprint arXiv:2105.14395*.

Wang, X. and D. B. Dunson (2013). Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*.

Wang, X., F. Guo, K. A. Heller, and D. B. Dunson (2015). Parallelizing MCMC with random partition trees. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 28.

Wendland, H. (2005). *Scattered Data Approximation*. Cambridge University Press.

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

Wikle, C. K. and S. H. Holan (2011). Polynomial nonlinear spatio-temporal integro-difference equation models. *Journal of Time Series Analysis 32*(4), 339–350.

Xue, J. and F. Liang (2019). Double-parallel Monte Carlo for Bayesian analysis of big data. *Statistics and Computing 29*(1), 23–32.

Yang, Y., A. Bhattacharya, and D. Pati (2017). Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv preprint arXiv:1708.04753*.

Yang, Y., M. Pilanci, M. J. Wainwright, et al. (2017). Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics 45*(3), 991–1023.

Zhang, M. M. and S. A. Williamson (2019). Embarrassingly parallel inference for Gaussian processes. *Journal of Machine Learning Research 20*, 1–26.

Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Computation 17*(9), 2077–2098.

Zhang, Y., J. C. Duchi, and M. J. Wainwright (2015). Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research 16*, 3299–3340.

Zhu, H., C. K. I. Williams, R. J. Rohwer, and M. Morciniec (1998). Gaussian regression and optimal finite dimensional linear models. In C. M. Bishop (Ed.), *Neural Networks and Machine Learning*.

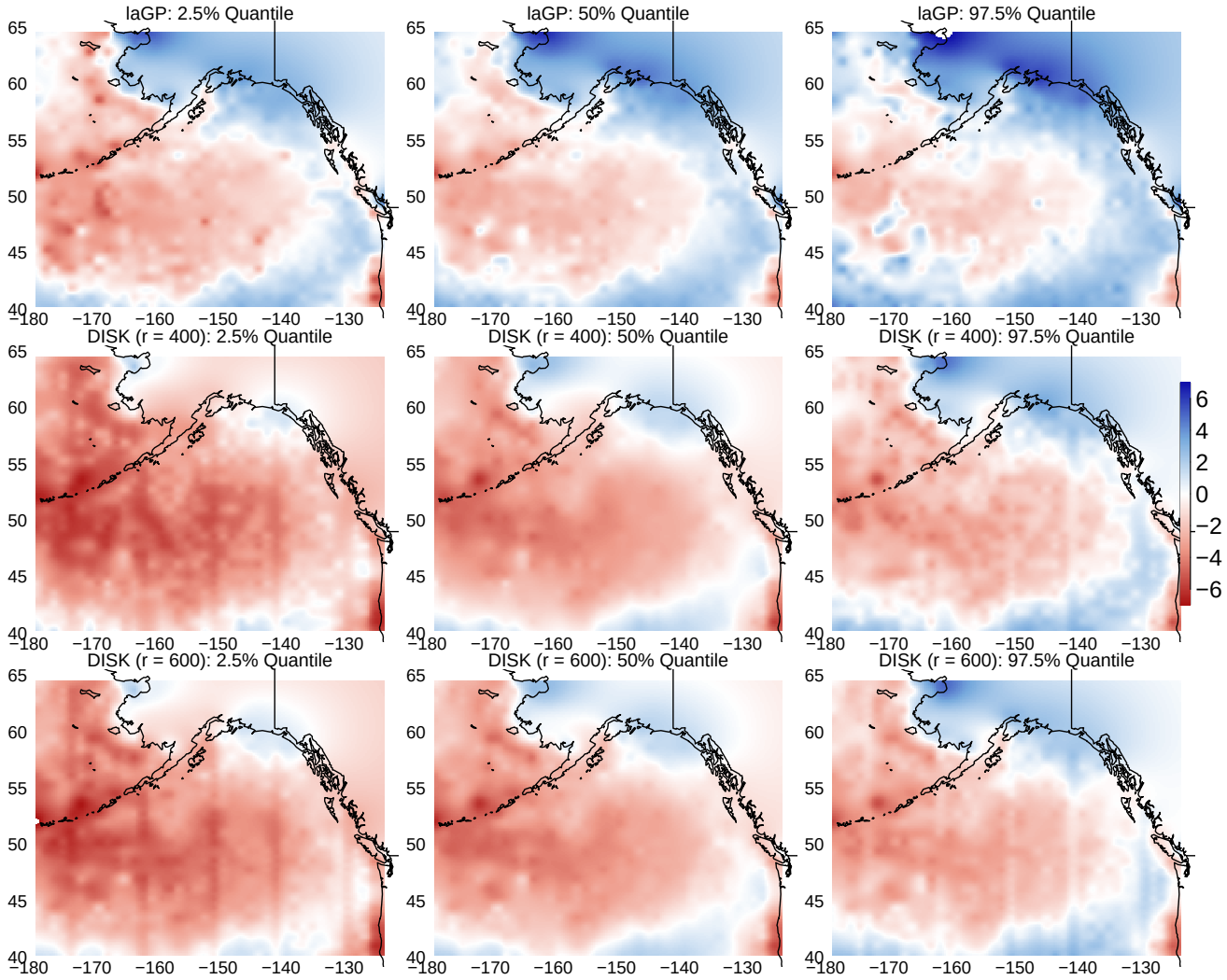

Fig 3: Interpolated spatial surface w at the locations in $\mathcal{S}^*$. Negative longitude means degree west from Greenwich. DISK uses MPP-based modeling with $r = 400, 600$ on $k = 300$ subsets and laGP uses the 'nn' method. The 2.5%, 50%, and 97.5% quantile surfaces, respectively, represent pointwise quantiles of the posterior distribution for $w(\mathbf{s}^*)$ for every $\mathbf{s}^* \in \mathcal{S}^*$.