The Canadian Journal of Statistics
Vol. 00, No. 00, 2020, Pages 000–000
La revue canadienne de statistique

A sequential split-and-conquer approach for the analysis of big dependent data in computer experiments

Chengrui LI[1], Ying HUNG[2*], and Minge XIE[2]

[1]*Bloomberg LP, New York, NY, U.S.A.*
[2]*Department of Statistics, Rutgers, the State University of New Jersey, Piscataway, NJ, U.S.A.*

Key words and phrases: Computer experiment; confidence distribution; divide-conquer-combine method; predictive distribution.

MSC 2010: Primary 62P30.

Abstract: Massive correlated data with many inputs are often generated from computer experiments to study complex systems. The Gaussian process (GP) model is a widely used tool for the analysis of computer experiments. Although GPs provide a simple and effective approximation to computer experiments, two critical issues remain unresolved. One is the computational issue in GP estimation and prediction where intensive manipulations of a large correlation matrix are required. For a large sample size and with a large number of variables, this task is often unstable or infeasible. The other issue is how to improve the naive plug-in predictive distribution which is known to underestimate the uncertainty. In this article, we introduce a unified framework that can tackle both issues simultaneously. It consists of a sequential split-and-conquer procedure, an information combining technique using confidence distributions (CD), and a frequentist predictive distribution based on the combined CD. It is shown that the proposed method maintains the same asymptotic efficiency as the conventional likelihood inference under mild conditions, but dramatically reduces the computation in both estimation and prediction. The predictive distribution contains comprehensive information for inference and provides a better quantification of predictive uncertainty as compared with the plug-in approach. Simulations are conducted to compare the estimation and prediction accuracy with some existing methods, and the computational advantage of the proposed method is also illustrated. The proposed method is demonstrated by a real data example based on tens of thousands of computer experiments generated from a computational fluid dynamic simulator. *The Canadian Journal of Statistics* 00: 000–000; 2020 © 2020 Statistical Society of Canada

Résumé: s expériences informatiques génèrent souvent des données corrélées massives avec de nombreuses entrées pour étudier des systèmes complexes. Les processus gaussiens (PG) sont largement utilisés comme outil pour leur analyse. Même si les PG offrent une approximation simple et efficace aux expériences informatiques, ils présentent deux problèmes critiques non résolus. Le premier se trouve au niveau computationnel dans l'estimation et les prévisions des PG qui nécessitent d'intenses manipulations de grandes matrices de corrélation. Pour une taille d'échantillon élevée et un grand nombre de variables, cette tâ che devient souvent instable, voire infaisable. L'autre problème réside dans l'amélioration de l'approche naïve de substitution de la distribution prédictive qui conduit à une sous-estimation de l'incertitude. Les auteurs introduisent un cadre unifié qui peut régler ces deux problèmes simultanément. Il repose sur une procédure séquentielle de type diviser pour régner, une technique de combinaison de l'information utilisant les distributions de confiance (DC) et une distribution prédictive fréquentiste basée sur des DC

Additional Supporting Information may be found in the online version of this article at the publisher's website.

*Author to whom correspondence may be addressed.
E-mail: yhung@stat.rutgers.edu

combinées. Les auteurs montrent que la méthode proposée conserve la même efficacité asymptotique que la vraisemblance conventionnelle sous des hypothèses raisonnables tout en réduisant substantiellement la quantité de calcul nécessaire pour l'estimation et la prévision. La distribution prédictive comporte une information complète pour l'inférence et une meilleure quantification de l'incertitude de prévision en comparaison de l'approche de substitution. Les auteurs présentent des simulations comparant la justesse de l'estimation et des prévisions par rapport aux méthodes existantes. Ils illustrent également l'avantage computationnel de leur approche. Ils démontrent finalement l'usage de leur méthode en analysant des données réelles de dizaines de milliers d'expériences informatiques provenant d'un simulateur numérique portant sur la dynamique des fluides. *La revue canadienne de statistique* 00: 000–000; 2020 © 2020 Société statistique du Canada

1. INTRODUCTION

A computer experiment refers to the study of a real system using mathematical models. It has been widely used as an alternative to physical experiments, especially for studying complex systems. In recent years, there has been a growing interest in analysing computer experiments by Gaussian process (GP) models. GP models are simpler than real systems, but they still effectively provide key summary information for mathematical systems (Sacks et al., 1989). Different from the conventional applications of GPs in spatial statistics, analysis of computer experiments is more complex and often involves more variables which creates new challenges in estimation and prediction (Santner, Williams & Notz, 2003; Fang, Li & Sudjianto, 2006).

There are two critical issues in modelling computer experiments by GPs. First, given sample size n, the estimation and prediction heavily involve manipulations of the n-by-n correlation matrix, which require $O(n^3)$ computations and often result in singularities for large n (Kaufman, Schervish & Nychka, 2008). This issue has been recognized in the literature, and the proposed approaches can be characterized broadly as either changing the model to a low-rank model or approximating the likelihood by imposing a sparsity constraint on the correlation matrix such as tapering (Kaufman, Schervish & Nychka, 2008; Stein, 2013) and compact support (Gneiting, 2002; Stein, 2008; Kaufman et al., 2011). Examples of the former include Rue & Tjelmeland (2002), Rue & Held (2005), Cressie & Johannesson (2008), Banerjee et al. (2008), Wikle (2010), Chang et al. (2014); while approximation approaches include Stein, Chi & Welty (2004), Snelson & Ghahramani (2005), Furrer, Genton & Nychka (2006), Fuentes (2007), Kaufman, Schervish & Nychka (2008), Eidsvik et al. (2014), Gramacy & Apley (2015), Nychka et al. (2015), Zhang, Lin & Ranjan (2018), Park & Apley (2018), Katzfuss (2017) and Sung et al. (2019). The computational difficulty and numerical instability of GP are critical in analysing computer experiments due to the large amount of unknown correlation parameters involved.

The second issue is how to accurately quantify the uncertainty in GP modelling. It is well known that the predictive distributions constructed by substituting the true parameters by their estimators, often called plug-in predictive distributions, underestimate the uncertainty (Santner, Williams & Notz, 2003, p. 98). However, they are still widely used due to the lack of computationally efficient alternatives. Alternative approaches, such as bootstrap predictive distributions (Sjöstedt-de Luna, 2003; Santner, Williams & Notz, 2003) and Bayesian procedures (Kennedy & O'Hagan, 2001; Schmidt & O'Hagan, 2003) provide better quantification of uncertainty, but they require intensive computation and typically are infeasible for high-dimensional problems (Datta et al., 2016).

Although numerous methods have been proposed to address these issues, to the best of our knowledge, they are developed for solving one of the issues. So our goal is to introduce a unified framework based on GP models which can address both issues simultaneously. This framework is called sequential split-conquer-combine (SSCC), which consists of a sequential split-and-conquer procedure, an information combining technique using confidence distributions (CDs), and a predictive distribution obtained based on combined CDs (Singh, Xie & Strawderman,

2005; Yang et al., 2014; Liu, Liu & Xie, 2015; Schweder & Hjort, 2016). CDs are a frequentist analogue of Bayesian posteriors. The CD-based approach does not require additional assumptions on the prior, yet enjoys the flexibility of Bayesian approaches. We extend the CD development to combine information from dependent sub-datasets and make prediction based on large spatial data. Theoretical developments are provided to guarantee the statistical performance, including consistency, coverage and efficiency, in both estimation and prediction of the proposed method.

The sequential split-and-conquer procedure reduces the computational complexity by splitting the data into smaller subsets and allowing estimations to be performed on the subsets individually. Although similar ideas of data splitting are discussed in the literature (e.g., Stein, 2013; Chen & Xie, 2014; Mackey, Talwalkar & Jordan, 2015), information from individual subsets is often assumed to be independent, which is invalid for GPs. In contrast, the proposed sequential split-and-conquer procedure takes into account the data dependency among subsets by carefully updating information sequentially from neighbourhood sets one at a time so that the data information and thus the estimation efficiency is preserved. After splitting the data into subsets, individual information from each subset is combined using a CD technique to get a combined CD, which provides not only an efficient overall estimate but also a flexible tool for inference. In addition, a predictive distribution is constructed based on the combined CD and it leads to an easy-to-compute GP prediction method. Apart from the computational reduction, the proposed framework provides combined estimates and predictions which are asymptotically equivalent to the conventional ones under very mild conditions. Furthermore, it provides comprehensive information for statistical inference and a better quantification of predictive uncertainty as compared with the plug-in approach.

The remainder of this paper is organized as follows. In Section 1.1, we introduce the commonly used GP models. The unified framework is introduced in Section 2. In Section 3, the prediction procedure and its uncertainty quantification are discussed. Simulations are presented to demonstrate the performance of the proposed framework in Section 4. In Section 5, the proposed approach is applied to a data centre thermal management study. A summary and concluding remarks are given in Section 6.

1.1. GP Models and Mathematical Notations

A GP model can be written as

$$y(\boldsymbol{x}) = \mu(\boldsymbol{x}) + Z(\boldsymbol{x}), \tag{1}$$

where $y \in \mathbb{R}$ is the output, $\boldsymbol{x} \in \mathbb{R}^p$ is the input in a compact set, $\mu(\boldsymbol{x})$ is the mean function assumed to be $\mu(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$ with unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^p$. Compared with $\mu(\boldsymbol{x}) = 0$, known as simple Kriging (Stein, 1999), the mean function $\mu(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$ can address non-stationarity issues, provide better interpretability and improve prediction accuracy (Hung, 2011). $Z(\boldsymbol{x})$ is a GP with mean zero and covariance $\mathrm{Cov}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma^2 \phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$, where $\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$ is the correlation function and $\boldsymbol{\theta}$ is a vector of unknown correlation parameters. Various correlation functions have been considered in the literature, but we focus on a popular choice in computer experiments, a product form of power exponential functions (Sacks et al., 1989; Gramacy & Apley, 2015; Kaufman et al., 2011):

$$\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) = \prod_{k=1}^{p} R_k(|x_{ik} - x_{jk}|) = \prod_{k=1}^{p} \exp(-\theta_k |x_{ik} - x_{jk}|^{\alpha_k}), \tag{2}$$

where $0 < \alpha_k \leq 2$ is a tuning parameter and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ with $\theta_k \geq 0$ for all k. Since the correlation parameters θ_k's are not constrained to be equal, the model can handle different signals in each input dimension and thus (2) is particularly attractive to the analysis of computer experiments. Note that in (2), given θ_k, the correlation decreases with the increase of $|x_{ik} - x_{jk}|$

and the correlation is zero if $|x_{ik} - x_{jk}| = \infty$. Due to the product correlation structure, the resulting $Cov(x_i, x_j)$ is zero as long as one of the dimension has zero correlation.

Given n realizations $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and the corresponding inputs $X = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top = (V_1, \ldots, V_p)$, where the vector V_k contains the values of the kth input variable, the joint log-likelihood function for (1) can be written as

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma) = -\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^\top \Sigma^{-1}(\boldsymbol{\theta})(\boldsymbol{y} - X\boldsymbol{\beta}) - \frac{1}{2}\log|\Sigma(\boldsymbol{\theta})| - \frac{n}{2}\log(\sigma^2). \tag{3}$$

Here, $\Sigma(\boldsymbol{\theta})$ is the $n \times n$ correlation matrix with the ijth element equal to $\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$. For each given $\boldsymbol{\theta}$, the maximum likelihood estimates (MLEs) of $\boldsymbol{\beta}$ and σ can be obtained by

$$\widehat{\boldsymbol{\beta}} = (X^\top \Sigma^{-1}(\boldsymbol{\theta})X)^{-1} X^\top \Sigma^{-1}(\boldsymbol{\theta})\boldsymbol{y} \quad \text{and} \quad \widehat{\sigma}^2 = (\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})^\top \Sigma^{-1}(\boldsymbol{\theta})(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})/n.$$

By maximizing the logarithm of the profile likelihood, the MLE of $\boldsymbol{\theta}$ can be obtained by

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{n \log(\widehat{\sigma}^2) + \log|\Sigma^{-1}(\boldsymbol{\theta})|\}. \tag{4}$$

For the estimation of correlation parameters $\boldsymbol{\theta}$, there are some likelihood-based alternatives, including the restricted maximum likelihood (Irvine, Gitelman & Hoeting, 2007) and robust approaches based on a penalized likelihood (Li & Sudjianto, 2005). In this paper, we focus on the study of MLEs but the results can be further extended to the likelihood-based alternatives.

When the parameters are known, the conditional distribution of y_0 at a new input $\boldsymbol{x}_0$, given the observations $\boldsymbol{y}$, is normal with mean $m_0(\boldsymbol{\beta}, \boldsymbol{\theta})$ and variance $v_0(\boldsymbol{\beta}, \boldsymbol{\theta})$, where

$$m_0(\boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{x}_0^\top \boldsymbol{\beta} + \gamma(\boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{\theta})(\boldsymbol{y} - X\boldsymbol{\beta}) \quad \text{and} \tag{5}$$

$$v_0(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sigma^2(1 - \gamma(\boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{\theta})\gamma(\boldsymbol{\theta})), \tag{6}$$

and $\gamma(\boldsymbol{\theta})$ is an $n \times 1$ vector with the ith element equal to $\phi(\boldsymbol{x}_i, \boldsymbol{x}_0; \boldsymbol{\theta})$. In practice, when the parameters are unknown, the conventional plug-in approach constructs a predictive distribution by replacing the true parameters by their MLEs. Therefore, the (estimated) plug-in predictive distribution is normally distributed with mean $m_0(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ and variance $v_0(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$.

Calculating the MLEs in (1.1) and (4) and also the GP predictors in (6) is computationally intensive because the calculation requires manipulations of an $n \times n$ correlation matrix Σ, such as Σ^{-1} and $|\Sigma|$, which are numerically highly unstable for moderate sample sizes with a larger p and infeasible for large sample sizes. In addition, ignoring the parameter uncertainty in the construction of plug-in predictive distribution clearly leads to an underestimation of predictive uncertainty.

To reduce computational complexity of GP, a commonly used approach is the sparse matrix technique which introduces zeros into the correlation matrix (Pissanetzky, 1984; Barry & Pace, 1999). Methods along this line, such as compactly supported correlation functions and covariance tapering, have received increasing attention in the literature (Gneiting, 2002; Furrer, Genton & Nychka, 2006; Kaufman, Schervish & Nychka, 2008; Kaufman et al., 2011; Bickel & Levina, 2008; Stein, 2008; 2013; Chu, Zhu & Wang, 2011). The compactly supported correlation function introduces zeros into the correlation matrix by assuming

$$R_k(|x_{ik} - x_{jk}|) := 0, \quad \text{if } |x_{ik} - x_{jk}| \geq \tau_k, \tag{7}$$

for $\tau_k \geq 0$ and $k = 1, \ldots, p$. The tuning parameters τ_k are called the range parameters. Another commonly used approach is covariance tapering in which the covariance matrix is multiplied by a tapering function defined by a single range parameter. For these tapering-type of methods, the resulting estimates can display sizable bias when the range parameter is relatively smaller than the truth (Kaufman et al., 2011). Therefore, larger values of τ_k are preferred for the purpose of estimation despite that it leads to a significant increase of computational complexity.

2. SEQUENTIAL SPLIT-CONQUER-COMBINE FRAMEWORK

2.1. Overview

We introduce in this section a unified SSCC framework to tackle the aforementioned computing and prediction uncertainty issues simultaneously. A diagram describing the SSCC framework is shown in Figure 1 and it consists of two stages. The first stage is called *split-and-conquer* and the second stage is called *combine*. The idea behind the *split-and-conquer* is to split the data into m smaller subsets, $\mathbb{D}_1, \ldots, \mathbb{D}_m$, which are disjoint but correlated, and then sequentially update the data by removing the dependency. So the resulting new subsets, $\mathbb{D}_1^*, \ldots, \mathbb{D}_m^*$, are mutually independent. Detailed procedures on how to split the data are given in Section 2.2. Let ψ be the collection of parameters in model (1). In the *combine* stage, individual estimates from each subset, denoted by $\hat{\psi}_1, \ldots, \hat{\psi}_m$, are combined via CDs (Singh, Xie & Strawderman, 2005; Schweder & Hjort, 2016) and the combined estimate is denoted by $\hat{\psi}_c$.

In Section 2.2, the two stages are discussed for the estimation of $\boldsymbol{\beta}$, which is a simplified case assuming σ and the correlation parameters $\boldsymbol{\theta}$ are known. A general procedure for the estimation of all the unknown parameters is given in Section 2.3.

2.2. Estimation of β When θ and σ are Known

We begin by illustrating the SSCC framework using a simple case where $\boldsymbol{\theta}$ and σ are known.

Stage 1: sequentially split-and-conquer

A key idea of the proposed framework is to reduce computational complexity by splitting the data into smaller subsets and allowing the estimation to be performed separately within each smaller dataset. This concept is attractive and is discussed under various settings, including spatial–temporal models (Stein, 2013), matrix factorization in machine learning (Mackey, Talwalkar & Jordan, 2015), linear models (Lin & Xi, 2011; Schifano et al., 2016; Song & Liang,

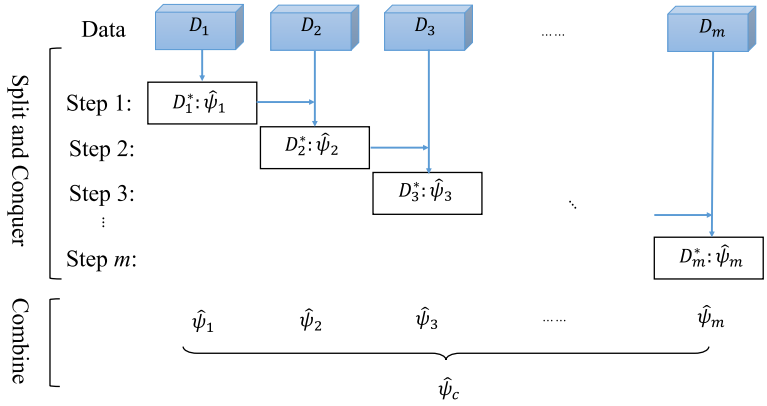

FIGURE 1: Diagram of the SSCC procedure.

2015) and penalized regressions (Chen & Xie, 2014; Tang, Zhou & Song, 2016). However, most of the existing methods handle the splitted subsets separately and that they do not take into account any dependency between subsets, which is crucial in the setting of GP models. If the dependence is not accounted for, it will lead to a significant loss of efficiency in estimation. Therefore, a sequential updating procedure is introduced to incorporate the dependency between neighbouring subsets.

In the proposed framework, the full data $\boldsymbol{y}$ is first split into m disjoint subsets, $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$, according to the values of one of the input variables. Without loss of generality, we denote this variable by the first variable V_1. Theoretically, the results developed in this paper are valid regardless of the choice of V_1 (as long as, for the chosen variable, the correlation parameter $\theta_k > 0$). In practice, we suggest conducting a preliminary regression analysis to select the most significant variable as V_1. Some detailed procedures on how to choose this variable are illustrated in Section 5. Although not a necessary condition, to simplify our notations, we assume that the range of each input variable is divided into equally spaced intervals and each setting $\boldsymbol{x}$ is a point chosen from the regular grids. The number of subsets, m is defined by $m = \lfloor M_1/\tau \rfloor$, where $M_1 = \max(V_1) - \min(V_1)$ is the range of the first variable and where τ is a tuning parameter closely connected to the range parameter in tapering. After sorting the full data according to their values of V_1, the disjoint subsets $\boldsymbol{y}_a$, where $a = 1, \ldots, m-1$, are obtained by the data with $V_1 \in [\min(V_1) + (a-1)\tau, \min(V_1) + a\tau)$ and $\boldsymbol{y}_m$ are obtained by the data with $V_1 \in [\min(V_1) + (m-1)\tau, \max(V_1)]$. The size of each subset $\boldsymbol{y}_a$ is denoted by n_a and $\sum_{a=1}^{m} n_a = n$. The n_a values are assumed to be of the same order so that each subset of the data has sufficient information to obtain an accurate estimation. This assumption is valid if the data are collected from space-filling designs (Santner, Williams & Notz (2003), Chapter 5).

After rearranging the data by variable V_1 and splitting, the covariance matrix Σ can then be decomposed into corresponding blocks indicating the within-subset correlations and between-subset correlations:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \cdots & \Sigma_{1m} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} & \cdots & \Sigma_{2m} \\ & \ddots & \ddots & \ddots & \\ \Sigma_{m1} & \Sigma_{m2} & \cdots & \Sigma_{m(m-1)} & \Sigma_{mm} \end{pmatrix}_{n \times n},$$

where Σ_{aa}, $a = 1, \ldots, m$, captures the correlation within subset $\boldsymbol{y}_a$, and where Σ_{ab} is a block matrix capturing the correlations between subsets $\boldsymbol{y}_a$ and $\boldsymbol{y}_b$, A block-wise tapering/thresholding is applied in which the correlation matrix Σ is approximated by Σ_t as follows:

$$\Sigma_t = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & O_{13} & \cdots & O_{1m} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} & \cdots & O_{2m} \\ & \ddots & \ddots & \ddots & \\ O_{m1} & O_{m2} & \cdots & \Sigma_{m(m-1)} & \Sigma_{mm} \end{pmatrix}_{n \times n}, \tag{8}$$

where O_{ab}'s are n_a-by-n_b matrices with all zeros. That is, Σ_{ab} is set to $O_{ab} = 0$, if $|a-b| \geq 2$. Equivalently, the correlation between two observations that are not in the same or neighbouring block(s) is set to be 0. By replacing Σ by Σ_t in the log-likelihood function (3), we have the approximate log-likelihood function denoted by l_t. This approximation is accurate and performs well when $\Sigma_{ab} \approx 0$ holds for all $|a-b| \geq 2$.

We would like to remark that the correlation matrix Σ_t brings in sparsity by introducing zeros to the correlation matrix if the data are neither within the same subset nor in the neighbouring subsets. Our method with tapering by blocks has three advantages over the existing tapering-type of approaches. First, the sparsity assumption is only related to one variable, while the typical

compactly supported correlation and tapering methods require the sparsity assumption on all the variables as defined in (7). Second, the full correlation information between any two neighbouring subsets is maintained in our Σ_t, while only partial information is maintained in the usual multiple-variable tapering-type of correlation function. For example, in the tapering approach discussed by Stein (2013), the off-diagonal matrices $\Sigma_{a,a+1}$ and $\Sigma_{a+1,a}$ are assumed to be zeros and none of the correlation information between neighbourhood are maintained. Also, suppose that x_{i1}, for $i = 1, \ldots, n$ are equally spaced. Then, the tapering assumption in (7) with $k = 1$ (i.e., $R_1(|x_{i1} - x_{j1}|) = 0$, if $|x_{i1} - x_{j1}| > \tau$, for range parameter τ) corresponds to the assumption that the off-diagonal matrices $\Sigma_{a,a+1}$ and $\Sigma_{a+1,a}$ are lower and upper triangular matrices instead of full matrices. Third, by working on the subsets of data, it is computationally affordable for the proposed framework to use a larger τ, compared with the typical tapering-type of methods. This leads to a smaller loss of information and therefore provides a higher estimation efficiency and accuracy.

We also would like to comment that (8) is a function of θ and it is not guaranteed to be positive definite for all values of θ because it is a band matrix. In the search for the MLE for θ, some values of θ can lead to negative correlation matrices. Techniques introduced in the literature, such as Cai & Zhou (2420), can be applied to ensure a positive semi-definite correlation matrix, but the implementations are often computationally intensive. In our numerical analysis, we discard those θ values (often a small percentage) and only search within those values of θ that lead to positive definite correlation matrices.

We now transform $\boldsymbol{y}$ to $\boldsymbol{y}^* = (\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_m^*)$ by a sequential method to preserve the correlation information in the neighbourhood blocks. Specifically, we sequentially update each subset as follows:

$$\boldsymbol{y}_a^* = \boldsymbol{y}_a - L_{a(a-1)}\boldsymbol{y}_{a-1}^*, \tag{9}$$

where $L_{(a+1)a} = \Sigma_{(a+1)a}D_a^{-1}$, $D_a = \Sigma_{aa} - L_{a(a-1)}D_{(a-1)}L_{a(a-1)}^\top$, and where $L_{(a+1)a}$ and D_a's are solved iteratively by initialing $D_1 = \Sigma_{11}$. The update of $\boldsymbol{y}_a^*$ depends only on $\boldsymbol{y}_a$ and $\boldsymbol{y}_{a-1}^*$, which are small subsets; thus it involves only up to $(n_{a-1} + n_a)$ data points in this step and therefore is easy to compute. This transformation is guided by the *block LDL-decomposition* (Fang, 2011). The next lemma states that the subsets $\boldsymbol{y}_a^*$ are mutually independent. All the proofs are given in the Supplementary Material.

Lemma 1. *After transformation, the covariance within each subset $\boldsymbol{y}_a^*$ is D_a and any two subsets are mutually independent. That is, $\boldsymbol{y}^* = (\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_m^*)$ has covariance matrix D, where*

$$D = \begin{pmatrix} D_1 & \ldots & O \\ & \ddots & \\ O & \ldots & D_m \end{pmatrix}, \; L = \begin{pmatrix} I & O & \ldots & O \\ L_{21} & I & \ldots & O \\ \vdots & \vdots & \ddots & \vdots \\ L_{m1} & L_{m2} & \cdots & I \end{pmatrix},$$

and $LDL^\top = \Sigma_t$, and where I's are identity matrices.

We complete the split-and-conquer stage by analysing the individual subset data $\boldsymbol{y}_a^*$, for $a = 1, \ldots, m$. For each individual subset $\boldsymbol{y}_a^*$, the log-likelihood function of $\boldsymbol{y}_a^*$ can be written as

$$l_t^{(a)}(\boldsymbol{\beta}) = -\frac{1}{2}\log|D_a| - \frac{1}{2\sigma^2}(C_a\boldsymbol{\beta} - \boldsymbol{y}_a^*)^\top D_a^{-1}(C_a\boldsymbol{\beta} - \boldsymbol{y}_a^*),$$

where $n_a \times p$ matrix $C_a = X_a + \sum_{b=1}^{a-1} B_{ab}X_b$, $n_a \times n_b$ matrix $B_{ab} = \prod_{k=b+1}^{a}(-L_{k(k-1)})$, and $n_a \times p$ matrix X_a is the design matrix corresponding to $\boldsymbol{y}_a^*$. By maximizing $l_t^{(a)}(\boldsymbol{\beta})$, we have the MLE of

$\boldsymbol{\beta}$ estimated from the ath subset as

$$\widehat{\boldsymbol{\beta}}_a = \arg\max_{\boldsymbol{\beta}} l_t^{(a)}(\boldsymbol{\beta}) = (C_a^\top D_a^{-1} C_a)^{-1} C_a^\top D_a^{-1} \boldsymbol{y}_a^*. \tag{10}$$

Since $\widehat{\boldsymbol{\beta}}_a$ is linear in $\boldsymbol{y}_a^*$ thus linear in $\boldsymbol{y}$ and $\mathrm{Cov}(\boldsymbol{y}_a^*) = D_a$ by Lemma 1, we have $S_a^{-1/2}(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}) \sim N(\mathbf{0}, \mathbf{I})$, where $S_a = \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_a) = \sigma^2 (C_a^\top D_a^{-1} C_a)^{-1}$.

Stage 2: information combining via CDs

CD refers to any sample-dependent distribution function that can represent confidence intervals or regions of all levels for a parameter of interest (e.g., Xie & Singh (2013); Schweder & Hjort (2016)). Singh, Xie & Strawderman (2005) and Xie, Singh & Strawderman (2011) introduce a general framework to combine information based on CDs for a univariate parameter, which can subsume almost all information combining methods used in current practice. Liu, Liu & Xie (2015) and Yang et al. (2014) extend the development to combine CDs of shared parameter vectors from independent studies. However, the existing developments of combining information based on CDs are for independent studies (or sub-datasets). The research in this paper is the first effort to utilize the CD concept to combine information from dependent datasets (sub-datasets split from partial data). In addition, this CD-based development also allows us to utilize and devise a split-conquer and effective CD-based prediction approach, to be discussed in Section 3, to handle prediction problems.

From (10), a resulting CD for $\boldsymbol{\beta}$ in the ath subset, expressed in its density form, is

$$h_a(\boldsymbol{\beta}) \propto \exp\left[-\frac{1}{2\sigma^2}(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta})^\top S_a^{-1}(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta})\right].$$

That is, $N(\widehat{\boldsymbol{\beta}}_a, S_a)$ is a multivariate normal CD for $\boldsymbol{\beta}$; see Singh, Xie & Strawderman (2007) and Liu, Liu & Xie (2015) for the formal definition of multivariate normal CD. Then, following Liu, Liu & Xie (2015) and also relevant discussions in section 4 of Singh, Xie & Strawderman (2005), a combined point estimator of $\boldsymbol{\beta}$ can be obtained by

$$\widehat{\boldsymbol{\beta}}_c = \arg\max_{\boldsymbol{\beta}} \prod_{a=1}^{m} h_a(\boldsymbol{\beta}). \tag{11}$$

By a direct calculation, we have an explicit expression that $\widehat{\boldsymbol{\beta}}_c = (\sum W_a)^{-1}(\sum W_a \widehat{\boldsymbol{\beta}}_a)$, where $W_a = C_a^\top D_a^{-1} C_a$ is the weight matrix. Furthermore, the covariance of $\widehat{\boldsymbol{\beta}}_c$ is $S_c = \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_c) = (\sum W_a)^{-1}(\sum W_a S_a W_a)(\sum W_a)^{-1} = \sigma^2(\sum W_a)^{-1} = \sigma^2 (X^\top \Sigma_t^{-1} X)^{-1}$ and $S_c^{-1/2}(\widehat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}) \sim N(\mathbf{0}, \mathbf{I})$. Again, by the definition of Singh, Xie & Strawderman (2007) and Liu, Liu & Xie (2015), $N(\widehat{\boldsymbol{\beta}}_c, S_c)$ is a multivariate normal CD for $\boldsymbol{\beta}$. We call $N(\widehat{\boldsymbol{\beta}}_c, S_c)$ a *combined CD*, and it is a function on the space of $\boldsymbol{\beta}$ and depends on the data in all subsets. Following Xie & Singh (2013) and Schweder & Hjort (2016), statistical inference, such as constructing confidence intervals/regions of $\boldsymbol{\beta}$ or calculating p-values, can be easily obtained from the combined CD.

The following theorem shows that $\widehat{\boldsymbol{\beta}}_c$ is asymptotically equivalent to $\widehat{\boldsymbol{\beta}}$, the MLE obtained based on (1.1) without splitting the data. A proof is provided in the Supplementary Material.

Theorem 1. *Under the regularity conditions B in the Supplementary Material, the combined estimator $\widehat{\boldsymbol{\beta}}_c$ is a consistent estimator of $\boldsymbol{\beta}$ and has the following asymptotic distribution:*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}) \xrightarrow{D} N(0, S),$$

as $n \to \infty$, where $S = n\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_{mle}) = n\sigma^2 (X^\top \Sigma^{-1} X)^{-1}$.

2.3. Estimation of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and σ

We illustrate in this section the SSCC framework in a general setting in which both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are unknown. The computation is more demanding as compared with the estimation of $\boldsymbol{\beta}$ because the MLEs can be obtained only by maximizing the likelihood (4) without a closed form expression and the maximization involves intensive operations of large correlation matrices. Therefore, a computationally efficient estimation procedure is even more critical. We extend the procedure of Section 2.2 to the situation where $\boldsymbol{\theta}$ is also unknown. The idea is to obtain the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ by updating $\boldsymbol{\beta}|\boldsymbol{\theta}$ and $\boldsymbol{\theta}|\boldsymbol{\beta}$ iteratively. Here we describe one of the iterations with details and the detailed algorithm is given in the Supplementary Material.

Starting from an estimate of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}^{(t-1)}$, $\boldsymbol{\beta}^{(t)}$ can be estimated by the combined estimator $\widehat{\boldsymbol{\beta}}_c$ given in (11) with $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t-1)}$. Given $\boldsymbol{\beta}^{(t)}$, a two-step procedure that is the analogue to the one in Section 2.2 is implemented to obtain $\boldsymbol{\theta}^{(t)}$. In Step 1, based on the same splitting $(\boldsymbol{y}_1, \ldots \boldsymbol{y}_m)$, the sequential updating (9) is modified by

$$\boldsymbol{y}_a^*(\boldsymbol{\theta}) = \boldsymbol{y}_a - L_{a(a-1)}(\boldsymbol{\theta})\boldsymbol{y}_{a-1}^*(\boldsymbol{\theta}),$$

where

$$\begin{aligned} L_{a(a-1)}(\boldsymbol{\theta}) &= \Sigma_{a(a-1)}(\boldsymbol{\theta})D_{a-1}^{-1}(\boldsymbol{\theta}), \\ D_a(\boldsymbol{\theta}) &= \Sigma_{aa}(\boldsymbol{\theta}) - L_{a(a-1)}(\boldsymbol{\theta})D_{a-1}(\boldsymbol{\theta})L_{a(a-1)}^{\top}(\boldsymbol{\theta}). \end{aligned}$$

In Step 2, the closed form expression of MLE in (10) is replaced by maximizing the likelihood

$$l_t^{(a)}(\boldsymbol{\theta}|\boldsymbol{\beta}^{(t)}) = -\frac{1}{2}\log|D_a(\boldsymbol{\theta})| - \frac{1}{2\sigma^2}(\boldsymbol{y}_a^*(\boldsymbol{\theta}) - C_a(\boldsymbol{\theta})\boldsymbol{\beta}^{(t)})^{\top}D_a^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_a^*(\boldsymbol{\theta}) - C_a(\boldsymbol{\theta})\boldsymbol{\beta}^{(t)}),$$

where $C_a(\boldsymbol{\theta}) = X_a + \sum_{b=1}^{a-1} B_{ab}(\boldsymbol{\theta})X_b$, $B_{ab}(\boldsymbol{\theta}) = \prod_{k=b+1}^{a}(-L_{k(k-1)}(\boldsymbol{\theta}))$ and X_a is the design matrix for $\boldsymbol{y}_a$. Note that the calculation of log-likelihood $l_t^{(a)}$ depends only on the current subset $\boldsymbol{y}_a^*$, previous subset $\boldsymbol{y}_{a-1}^*$, and the correlation between these two subsets and thus it is still easy to compute. It is also clear that $l_t = \sum_{a=1}^{m} l_t^{(a)}$. The estimate of $\boldsymbol{\theta}$ from individual subset $\boldsymbol{y}_a^*$ is denoted by

$$\widehat{\boldsymbol{\theta}}_a = \arg\min_{\boldsymbol{\theta}} l_t^{(a)}(\boldsymbol{\theta}|\boldsymbol{\beta})$$

and the combined estimate for $\boldsymbol{\theta}$ can be calculated by

$$\widehat{\boldsymbol{\theta}}_c = \left(\sum_{a=1}^{m} R_a^{-1}\right)^{-1}\left(\sum_{a=1}^{m} R_a^{-1}\widehat{\boldsymbol{\theta}}_a\right),$$

where $R_a = -H_a^{-1}(\widehat{\boldsymbol{\theta}}_a)$ and $H_a(\cdot)$ is the ath Hessian matrix derived from $l_t^{(a)}(\boldsymbol{\theta}|\boldsymbol{\beta})$. Therefore, given $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\theta}^{(t)}$ is updated by the combined estimator, that is, $\boldsymbol{\theta}^{(t)} = \widehat{\boldsymbol{\theta}}_c$. Following Singh, Xie & Strawderman (2007) and Liu, Liu & Xie (2015) and similar to Section 2.2, an individual CD from the ath block is $N(\widehat{\boldsymbol{\theta}}_a, R_a)$ and the combined CD is $N(\widehat{\boldsymbol{\theta}}_c, S_c)$, where $S_c = (\sum_{a=1}^{m} R_a^{-1})^{-1}$. In addition, after updating $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\theta}^{(t)}$, σ is estimated by

$$\sigma^{2(t)} = (\boldsymbol{y} - X\boldsymbol{\beta}^{(t)})^{\top}\Sigma_t^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}^{(t)})/n = \sum_{a=1}^{m} \boldsymbol{\varepsilon}_a^{*\top}D_a^{-1}\boldsymbol{\varepsilon}_a^*/n,$$

where $\Sigma_t = \Sigma_t(\boldsymbol{\theta}^{(t)}), D_1 = \Sigma_{11}, \varepsilon_1^* = \boldsymbol{y}_1 - X_1\boldsymbol{\beta}^{(t)}$ and, for $a = 2, \ldots, m$, we have $\varepsilon_a = \boldsymbol{y}_a - X_a\boldsymbol{\beta}^{(t)}$, $L_{a(a-1)} = \Sigma_{a(a-1)}D_{a-1}^{-1}, \varepsilon_a^* = \varepsilon_a - L_{a(a-1)}\varepsilon_{a-1}^*, D_a = \Sigma_{aa} - L_{a(a-1)}D_{a-1}L_{a(a-1)}^\top$.

Despite the significant computational reduction, the combined estimators maintain desirable asymptotic properties as the original MLE. This result is shown in the next theorem with a proof given in the Supplementary Material.

Theorem 2. *Under the regularity conditions in the Supplementary Material, the combined estimator $\widehat{\boldsymbol{\lambda}}_c = (\widehat{\boldsymbol{\beta}}_c, \widehat{\boldsymbol{\theta}}_c)$ is a consistent estimator of $\boldsymbol{\lambda} = (\boldsymbol{\beta}, \boldsymbol{\theta})$ and asymptotically as efficient as MLE $\widehat{\boldsymbol{\lambda}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ obtained from (1.1).*

Kaufman, Schervish & Nychka (2008) point out that estimation based on tapering can be biased. In fact, this can be an issue in most of the tapering-type of methods including the compactly supported correlations and the current method. This is because, for example, when $\boldsymbol{\beta} = 0$, we have $E\left\{\frac{\partial\{-l_t(\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}}\right\} = E\left\{\frac{1}{2}\mathrm{tr}(\Sigma_t^{-1}\Sigma_t') + \frac{1}{2}\boldsymbol{y}'\Sigma_t^1\boldsymbol{y}\right\} = \frac{1}{2}\mathrm{tr}\left(\Sigma_t^{-1}\Sigma_t'\right) - \mathrm{tr}(\Sigma_t^{-1}\Sigma_t'\Sigma_t^{-1}\Sigma)$ $= \frac{1}{2}\mathrm{tr}(\Sigma_t^1(\Sigma - \Sigma_t)) \neq 0$, where $l_t(\cdot)$ denotes the log-likelihood function by compactly supported correlation, $\Sigma_t' = \partial\Sigma_t/\partial\boldsymbol{\theta}$, $\Sigma_t^1 = \partial\Sigma_t^{-1}/\partial\boldsymbol{\theta} = -\Sigma_t^{-1}\Sigma_t'\Sigma_t^{-1}$. In our case, this bias is diminished asymptotically, as stated in the following corollary whose proof can be found in the Supplementary Material.

Corollary 1. *Under the assumptions of Theorem 1, the combined estimator is asymptotically unbiased.*

3. PREDICTION AND UNCERTAINTY QUANTIFICATION

A CD-based predictive distribution is introduced in this section. It has two advantages. First, it is constructed based on a modified GP predictor which overcomes the computational difficulty often encountered in the conventional approach (6) yet maintains the same asymptotic efficiency. Second, it provides comprehensive information for statistical inference and a better quantification of prediction uncertainty as compared with the plug-in approach.

Based on the sequential split-and-conquer procedure and the combined estimates obtained from Section 2.3, we propose to approximate the GP predictive mean and variance (6) by $m_1(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $v_1(\boldsymbol{\beta}, \boldsymbol{\theta})$ as follows:

$$m_1(\boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{x}_0^\top\boldsymbol{\beta} + \sum_{a=1}^{m}\gamma_a^*(\boldsymbol{\theta})^\top D_a^{-1}(\boldsymbol{\theta})\boldsymbol{y}_a^* + \sum_{a=1}^{m}\gamma_a^*(\boldsymbol{\theta})^\top D_a^{-1}(\boldsymbol{\theta})C_a(\boldsymbol{\theta})\boldsymbol{\beta}, \tag{12}$$

$$v_1(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sigma^2\left(1 - \sum_{a=1}^{m}\gamma_a^*(\boldsymbol{\theta})^\top D_a^{-1}(\boldsymbol{\theta})\gamma_a^*(\boldsymbol{\theta})\right), \tag{13}$$

where $\gamma_a^*(\boldsymbol{\theta}) = \gamma_a(\boldsymbol{\theta}) + L_{a(a-1)}(\boldsymbol{\theta})\gamma_{a-1}^*(\boldsymbol{\theta})$, $\gamma_a(\boldsymbol{\theta})$ is the $n_a \times 1$ vector with ith element equals to $\phi(||\boldsymbol{x}_i - \boldsymbol{x}_0||; \boldsymbol{\theta})$ where $i = \sum_{b=1}^{a-1} n_b + 1, \cdots, \sum_{b=1}^{a} n_b$. These two estimates enjoy the computational efficiency because their calculation involves only a small correlation matrix with size $n_a \times n_a$, $a = 1, \ldots, m$. The new predictive mean (12) and variance (13) have the following asymptotic properties.

Theorem 3. *Under the regularity conditions in the Supplementary Material, we have $|m_1(\boldsymbol{\beta}, \boldsymbol{\theta}) - m_0(\boldsymbol{\beta}, \boldsymbol{\theta})| \to 0$ and $|v_1(\boldsymbol{\beta}, \boldsymbol{\theta}) - v_0(\boldsymbol{\beta}, \boldsymbol{\theta})| \to 0$ in probability as $n \to \infty$.*

Theorem 3 shows that the new predictive mean (12) and variance (13) are asymptotically equivalent to the conventional ones. This implies that the new predictor still enjoys the

interpolation property asymptotically (i.e., the predictive variance for the observed data is asymptotically zero), which is the most desirable feature of GPs in computer experiment modelling (Santner, Williams & Notz (2003)).

To provide a better alternative to the plug-in predictive distribution, we construct a CD-based predictive distribution which captures the parameter uncertainty using the combined CDs. The CD-based predictive distribution is introduced by Shen, Liu & Xie (2018) under a general setting. Here we extend the idea to GP models and construct a CD-based predictive distribution which is not only more accurate but also easy to compute.

A CD-based predictive distribution function is defined by:

$$Q(y_0; \boldsymbol{y}) = \int_{\lambda \in \Theta} G_\lambda(y_0) dF_c(\lambda; \boldsymbol{y}), \tag{14}$$

where $G_\lambda(y_0)$ is the cumulative distribution function (CDF) of the predictive distribution with known $\lambda = (\boldsymbol{\beta}, \boldsymbol{\theta})$, that is, a normal distribution with mean m_1 and variance v_1 given in (12) and (13); and $F_c(\cdot; \boldsymbol{y})$ is the CDF of $N(\widehat{\lambda}_c, S_c^\lambda)$, the CD of λ. Here, $\widehat{\lambda}_c = (\widehat{\boldsymbol{\beta}}_c, \widehat{\boldsymbol{\theta}}_c)$ is the combined estimator of λ, variance matrix $S_c^\lambda = \mathrm{Var}(\widehat{\lambda}_c)$ equals to the corresponding Hessian matrix calculated from the log-likelihood. The CD-based predictive distribution is closely related to the Bayesian predictive distribution and the bootstrap predictive distribution as discussed by Shen, Liu & Xie (2018) and Schweder & Hjort (2016). As in Theorem 4 of Shen, Liu & Xie (2018), it can be shown that the CD-based predictive distribution outperforms the plug-in approach, measured by the average Kullback−Leibler distance to the true predictive distribution.

To implement the predictive distribution formulated in (14), we use the following Monte Carlo algorithm which is simple yet broadly applicable.

Monte Carlo prediction algorithm: Obtain S simulated copies of y_0 from $Q(\cdot; \boldsymbol{y})$, denoted by $y_0^{(s)}$, for $s = 1, \ldots, B$, by iteratively performing the following two steps.

1. Simulate a random variable $\lambda^{(s)} | \boldsymbol{y} \sim N(\widehat{\lambda}_c, S_c^\lambda)$.
2. Obtain $y_0^{(s)} | \lambda^{(s)} \sim N(p_1(\lambda^{(s)}), v_1(\lambda^{(s)}))$.

These B copies of y_0 can be used to approximate the predictive distribution in (14). Note that $(\widehat{\lambda}_c, S_c^\lambda)$ has already been computed by the SSCC method in the estimation step and $(p_1(\lambda^{(t)}), v_1(\lambda^{(t)}))$ for a given $\boldsymbol{x}_0$ needs to be computed by the SSCC method only once. The Monte Carlo algorithm for a predictive distribution is simple and fast to carry out.

4. SIMULATION

Simulation studies are conducted to examine the performance, including estimation and prediction, of the proposed framework. Two types of data are generated in the studies, one is simulated from a underlying GP model and the other is a computer experiment simulated from an actual computer model called the Bohachevsky function (Surjanovic & Bingham (2017), https://www.sfu.ca/~ssurjano/). All simulations are carried out by a machine with a quad-core CPU @ 3.50GHz, 12GB RAM under R 3.3.1 in Windows 10.

To demonstrate the estimation performance, we compare the proposed combined estimator with the regular MLE and those obtained by the compactly supported correlation Kaufman et al. (2011), denoted by "Compact." We consider $\boldsymbol{x} \in [0, 1]^4$ with three different sample sizes, $n = 1{,}000$, 1,500 and 2,000, in which $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are unknown. Although these sample sizes are relatively small, they have pushed our machine to its limits to compute the regular MLE. This simulation study serves the purpose of providing a comprehensive comparison across all three

methods. For each sample size, we randomly divide each variable into equally spaced intervals. For example, when $n = 1{,}000$, we select 8 equally spaced design points for the first variable, 5 equally spaced points for the rest of the three variables. Therefore, we have $1{,}000 = 8 \times 5 \times 5 \times 5$ design points in total. In the first simulation, the responses are simulated from a GP model with $\boldsymbol{\beta} = (2, 3, 1, 2, 1.5)$ and

$$\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) = \prod_{k=1}^{4} \exp(-\theta_k |x_{ik} - x_{jk}|),$$

where $\boldsymbol{\theta} = (15, 1.5, 2, 3)$ and $\sigma^2 = 1$ is assumed to be known. To implement the SSCC framework, we assume $\tau = 0.2$, so that the number of blocks is $m = \lfloor M_1/\tau \rfloor = \lfloor 1/0.2 \rfloor = 5$. To emphasize the estimation performance of the parameters, we specify $\alpha = 1$ for the three methods without further tuning. Similar performance and comparison results (perhaps on a different magnitude) are expected for a different tuning constant α, for example, say $\alpha = 1.5$ or 2. The "Compact" method is implemented by (7), which involves one threshold parameter for each dimension. Some tuning is performed to maintain a reasonable estimation accuracy for "Compact" and the resulting setting is $\tau_1 = \tau = 0.2$ and $\tau_k = 1$, for $k = 2, \ldots, p$. For each sample size, we repeat the simulation 100 times and report the mean, standard deviation and the computing times denoted by CT, in Table 1 in the Supplementary Material.

Based on the numerical results, the estimation performance of the proposed estimator is comparable with the other two estimators. This observation is consistent with the theoretical results. In terms of the computing time, the proposed method provides a significant reduction as compared with the other two methods, especially for large sample sizes. Specifically, comparing with the regular MLE and the compactly supported correlation approach, the computing time is reduced by more than 86% by the proposed combined estimator for all three different sample sizes and this reduction increases with sample sizes.

To illustrate the performance of the proposed predictive distribution, we implement the *Monte Carlo prediction algorithm* (Section 3) to construct predictive distributions for several untried points following the previous settings with sample size $n = 2{,}000$. We focus on examining the predictive performance by changing the setting of the most important variable, because this is of interest in many applications including the real data analysis in Section 5. Four untried settings are assumed by varying the settings of the most active variables, that is, changing the setting of the first variable to be 0.2, 0.4, 0.6 and 0.8 respectively. The setting of the other three variables is fixed to be $(0.43, 0.5, 1)$. Based on the estimators in Table 1 in the Supplementary Material, we have $\widehat{\lambda}_c = (2.08, 2.90, 1.04, 1.98, 1.49, 14.79, 1.49, 2.00, 3.00)$. Here, we construct the CD-based confidence distribution according to 1,000 copies of y_0 generated by the Monte Carlo prediction algorithm, denoted by $y_0^{(1)}, \ldots, y_0^{(1000)}$. Figure 1 in the Supplementary Material shows the corresponding histograms of y_0 for the four untried settings. The red dashed lines are the mean function calculated by the true parameters. The CD-based predictive distribution not only contains the information of the predictive mean but also provides a flexible way to construct predictive intervals with any level of frequentist coverage probability.

In the next study, the objective is to understand the computing cost of the proposed framework, including estimation and prediction, and make comparisons with existing alternatives. We compared seven methods, including (1) the Bayesian Tree Gaussian Process via `btgp` package in R (Gramacy & Lee, 2008), (2) local Gaussian process via `laGP` package in R (Gramacy & Apley, 2015), (3) an experimental design based subsampling approach proposed by Zhao, Amemiya & Hung (2018), denoted by LHD, with $1/4$ of the full data randomly sampled, (4) the proposed method denoted by "SSCC+MC," (5) the SSCC estimated with plug-in prediction denoted by "SSCC+plugin," (6) the MLE with Monte Carlo predictive distribution denoted by "MLE-MC" and (7) the MLE with plug-in predictive distribution denoted by "MLE+plug-in."

The first method (btgp) is a Bayesian version of GP proposed by Gramacy & Lee (2008). The second one is a local GP method, denoted by laGP, is introduced by (Gramacy & Apley, 2015). Although the idea of laGP using local data points in GP modelling is quite different from the proposed method, it is an efficient alternative to address the computational issue especially with a large number of inputs. laGP is implemented by the R package with the initial number of nearest neighbours set to be 6 and the total size of local designs set to be 100. For prediction, a default method of laGP which minimizes the predictive variance is used.

The design matrix X is sampled based on Latin hypercube designs in $[0,1]^2$ with sample size $n = 1{,}000$. Here, we use $p = 2$, instead of a relatively a larger p in a typical computer experiment, so that the seven methods can all be implemented without much complications. The data is generated by GP as before with $\beta = (1, 1.5, 2)$ and $\theta = (10, 15)$. We assume that $\tau = 0.2$ and thus $m = \lfloor M/\tau \rfloor = 5$. For those involving Monte Carlo samples, the number of Monte Carlo iterations is set to be 500 due to computational constraints.

The results are summarized by the 90% predictive/posterior intervals at four randomly selected untried settings, $(0.029, 0.841)$, $(0.334, 0.587)$, $(0.035, 0.144)$, and $(0.829, 0.943)$, in Figure 2 of the Supplementary Material and the computing time is reported in Table 2 in the Supplementary Material. It is shown in Figure 2 that the predictive intervals produced by the Monte Carlo methods are in general wider than the plug-in methods and they are comparable to the posterior interval created by Bayesian tree GP. The laGP is the most computationally efficient one among all the methods, followed by LHD. However, both laGP and LHD create larger uncertainties in prediction because the estimation and prediction are obtained based on a small subset of the data. The SSCC+plug-in is the fastest among the five methods, except laGP, and provides a reasonable coverage of the predictive uncertainty. The time for SSCC+MC, MLE+MC and btgp is comparable because the computing time associated with these methods is mostly dominated by the Monte Carlo iterations.

To demonstrate the performance with larger sample sizes, we compare SSCC with laGP. The design is generated by Latin hypercube with sample size 10,000 in 6 dimensions, and the data are simulated from GP with $\beta = (1, 1.5, 2.1, 2, 3, 4, 5)$ and $\theta = (10, 12, 14, 15, 16, 20)$. For laGP, we consider two settings. One is the default setting which minimizes the predictive variance, denoted by laGP-ALC, and the other is MSPE which minimizes the mean-squared prediction error denoted by laGP-MSPE. The number of nearest neighbour locations for initialization is six and the total size of local designs is 1,000. For SSCC, we assume $\tau = 0.2$ and $m = 5$. For four randomly selected untried settings, the comparisons are summarized by the 90% predictive interval in Figure 3 of the Supplementary Material and the computing time is reported in Table 3 there. Although the two laGP methods are computationally faster than SSCC, they produce larger prediction uncertainty in general.

In the second type of simulation, we conduct a computer experiment using the Bohachevsky function, which can be written as

$$f(\boldsymbol{x}) = x_1^2 + 2x_2^2 - 0.3\cos(3\pi x_1) - 0.4\cos(4\pi x_2) + 0.7,$$

where $x_1, x_2 \in [-100, 100]$. The design is a Latin hypercube sample with sample size 1,000 for training and 20 for testing. We assume that $\tau = 0.2$ and thus $m = \lfloor M/\tau \rfloor = 5$. For those involving Monte Carlo samples, including btgp, SSCC+MC, and MLE+MC, the number of Monte Carlo iterations is set to be 100. Similar to the first simulation, seven methods are compared and the root mean squared prediction errors are summarized in Table 4 in the Supplementary Material. Again, the proposed SSCC approaches are comparable to the performance using MLEs, and the LHD-based approach seems to be similar to SSCC in this case but laGP has a much larger prediction error.

5. DATA CENTRE THERMAL MANAGEMENT

A data centre is a computing infrastructure facility that houses large amounts of information technology equipment used to process, store, and transmit digital information. The objective is to model the thermal distribution in a data centre and the final goal is to design a data centre with an efficient heat-removal mechanism. For a data centre thermal study, physical experiments are not always feasible because some settings are highly dangerous and expensive to perform. Therefore, computer experiments based on computational fluid dynamics (CFD) are widely used (Lopez & Hamann, 2011).

There are 26,820 temperature outputs generated from the CFD simulator based on an irregular grid over a 9-dimensional space. The nine variables are listed in Table 5 in the Supplementary Material. The first six variables control the cooling mechanism, including four computer room air conditioning (CRAC) units with different flow rates ($V_1, \ldots, V_4$), the overall room temperature setting (V_5), and the perforated floor tiles with different percentage of open areas (V_6). The last three variables are the spatial location, x-axis, y-axis, and height, (V_7 to V_9).

To implement the proposed method, an important step in practice is to determine the first variable. Although any variable can potentially be used, there are some choices that we found efficient in practice. Ideally, a variable that follows the tapering assumption is a desirable choice. In other words, for this particular variable, the correlations between pairs of responses with larger distances are nearly zero, and therefore little information is lost by assuming them to be conditionally independent as described by (8). We examine this assumption by checking "within-variable" correlations for each variable. Specifically, using the variable "height" as an example, there are 18 (say, $h_1, h_2, \ldots, h_{18}$) different height values that are equally spaced. We randomly select 1,000 data points at each level and calculate $\binom{18}{2} = 153$ pairwise correlations (i.e., the sample correlation of those 1,000 y's with height = h_i and those 1,000 y's with height = h_i). We then average the pairwise correlations with that same value of h_i—h_j (we use d(height) to represent the value of the difference). Figure 4 in the Supplementary Material is a scatter plot of the average pairwise correlations versus their corresponding d(height) values. Such "within-variable" correlation plots can be obtained for other variables as well. In the cases where the variables are continuous, their values should be discretized first. In Figure 4, the three variables with the fastest decaying correlations are given. Among these variables, "height" shows the fastest decay compared with the rest, and therefore is chosen as the first variable. The dataset is then divided according to "height." We first normalized the 18 equally spaced levels of height to $[0, 1]$ and set $\tau = 2/17$, which guarantees that each block consists of two different levels of height and that there are 9 blocks in total.

The proposed method is applied to the full data $n = 26860$ and two smaller subsets, $n = 1{,}800$ and $n = 3600$, to compare the performance with the original MLE and the compact support correlation, provided the same threshold settings as in Section 4. Estimation results are summarized in Table 5 in the Supplementary Material, where "-" indicates no result available. For $n = 1{,}800$, we are able to calculate the estimators for the three approaches. The results show that, with a similar estimation performance, the proposed combined estimator reduces the computing time by more than 98% compared with the other two methods. For $n = 3600$ and the full data, $n = 26{,}860$, the original MLE and the compactly supported correlation approach cannot be carried out due to computational and/or memory limitation. Note that the compact support approach in Kaufman et al. (2011) handles a larger training data by creating a correlation matrix with high sparsity using a relatively small τ, which is a strong assumption. In contrast, the proposed method relaxes this assumption by allowing a larger τ. In this example, the compact support approach is carried out by using the same τ as in the proposed SSCC which may provide insufficient sparsity for the method to be computationally feasible.

Based on the estimation results in Table 5 in the Supplementary Material, we can construct the CD-based predictive distribution for some untried settings, which is a crucial step in finding

an efficient cooling mechanism in a data centre. The prediction performance is first illustrated by predicting the heat map in the data centre by varying the most active variable, height, with the control variables assumed to be: CRAC unit flow rate 6,500, unit 2 flow rate 6,500, unit 3 flow rate 2,750, unit 4 flow rate 2,750, room temperature 71° F and tile percentage 75%. Figure 5 in the Supplementary Material presents the CD-based predictive heat map at four different heights, that is, 0, 2.25, 4.25, 6.75. From the heat maps, it is shown that on average, temperature increases with height which agrees with thermal dynamics in general. Apart from the predictive heat map, the CD-based predictive distribution can be used to construct confidence intervals with any level of frequentist coverage probability. It also provides valuable insights of the thermal distribution in the data centre. For example, Figure 6 in the Supplementary Material shows the predictive distributions for four randomly selected untried settings at location x-axis = 23.5, y-axis = 14.5, with four different heights. At height = 2.25 given other settings, the confidence that the temperature will be below 66° F is 99.4%; at height = 6.75, the confidence that the temperatures fall into the interval (74, 77) is 84.2%.

Based on four randomly sampled untried settings in the data centre, the predictive performances of SSCC-MC and SSCC-Plugin are compared with laGP which appears to be the most computationally efficient approach with massive data. The results are summarized by Figure 7 in the Supplementary Material. As shown in the figure, the predictive uncertainty of laGP is much larger than the two SSCC results which is consistent with the observations in the simulation studies. This observation is also consistent with the finding in 10-fold cross-validation when $n = 1{,}800$. For the 10-fold cross-validation, the RMSPEs and the standard deviations for both SSCC-MC and laGP are summarized in Table 6 in the Supplementary Material.

We further compare the CD-based predictive distribution with the plug-in approach when the MLE is available, that is, $n = 1{,}800$. Based on the same untried setting with zero height, Figure 8 in the Supplementary Material shows in the black curve the empirical predictive density obtained by the combined CD and in the red dotted curve the corresponding plug-in predictive density. It appears that the plug-in approach slightly underestimates the predictive uncertainty and this underestimation is expected to be larger when the sample size gets smaller. So the empirical result shows that, apart from computational reduction, the CD-based predictive distribution provides a better quantification of predictive uncertainty as compared with the traditional plug-in approach.

6. SUMMARY AND CONCLUDING REMARKS

We propose a unified SSCC framework, called SSCC, to tackle two open problems in the analysis of computer experiments using GP models: the computational difficulty and the underestimation of prediction uncertainty. This framework consists of a sequential split-and-conquer procedure, information combining using CDs, and a predictive distribution obtained by combined CD. Under mild conditions, the new estimators and predictors are shown to be asymptotically equivalent to the conventional ones using full data, while the computing time is significantly reduced. A Monte Carlo algorithm is introduced to construct the CD-based predictive distribution which provides rich information for inference and a better quantification of prediction uncertainty compared to with the plug-in approach.

The asymptotic properties discussed in this paper is based on increasing domain (Mardia & Marshall, 1984) asymptotics where more and more data are collected in increasing domains while the sampling density stays constant. There is another framework called the fixed-domain asymptotics (Stein, 1999), where data are collected by sampling more and more densely in a fixed domain. It is shown by Zhang & Zimmerman (2005) that, given quite different behaviour under the two frameworks in a general setting, their approximation quality performs about equally well under certain assumptions. Therefore, although results given here are based on an increasing domain, they provide some insights about the proposed estimators in both frameworks.

The proposed framework shares some similarities with the composite likelihood (Eidsvik et al., 2014), but there are also noted differences. The composite likelihood generally has two approximations. First, a composite (pairwise-block) likelihood function is used to approximate the full likelihood. Second, the tapering method in Eidsvik et al. (2014), blocks that are not immediate neighbours are assumed to be independent. Both approximations lead to loss of estimation efficiency. The proposed SSCC approach only uses one approximation: a tapering-type method to control weak dependence among the spatial responses. We further provide a theoretical condition on τ under which the tapering approximation is sufficiently accurate. As a result, we can provide a set of clean-cut conditions to show that our SSCC estimator is asymptotically efficient and asymptotically equivalent to the full likelihood MLE. Furthermore, we provide a unified theory to quantify the performance of the joint estimates as well as prediction, which is not available in Eidsvik et al. (2014).

Finally, we would like to comment that GP models provide a simple and convenient tool to analyse expensive computer experiments. The actual correlation among the data generated form a computer experiment may not follow the correlation pattern specified by a GP model, although we often expect GP models to provide useful summaries of the key information. Specifically, when a likelihood approach is used, the likelihood estimating equation can be viewed as a quasi-likelihood estimating equation when the model is misspecified. As long as the estimating equation is Fisher consistent, we expect to have consistent summaries of the key information such as the first or second moments, and so on. The proposed SSCC provides equivalent results as the likelihood approach without any additional assumptions that otherwise are typically required by a bootstrap or a Bayesian approach. Similar to likelihood approaches, we expect our outcome to be more robust and resistant to a potential misspecification of the correlation pattern by GP than the bootstrap or Bayesian method. Further exploration of big dependent computer experiment data beyond a GP model is a future research topic. As pointed out by one of the referees, the prediction performance of the proposed method is particularly useful for computer experiments because the prediction efficiency and accuracy are essential to several major issues in computer experiments, including calibration and uncertainty quantification. On the other hand, the estimation efficiency and accuracy of GP models may be of significant interest in spatial statistics.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the constructive advice from the associate editor and the referees. This work was supported by NSF-DMS-1812048, NSF-DMS-1737857, NSF-DMS-1660477, and NSF HDR TRIPODS award CCF-1934924.

Bibliography

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

Barry, R. P. & Pace, R. K. (1999). Monte Carlo estimates of the log determinant of large sparse matrices. *Linear Algebra and its Applications*, 289, 41–54.

Bickel, P. & Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics*, 36, 199–227.

Cai, T. T. & Zhou, H. H. (2420). Optimal rates of convergence for sparse covariance matrix estimation. *Annals of Statistics*, 40, 2389.

Chang, W., Haran, M., Olson, R., & Keller, K. (2014). Fast dimension-reduced climate model calibration and the effect of data aggregation. *The Annals of Applied Statistics*, 8, 649–673.

Chen, X. & Xie, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24, 1655–1684.

Chu, T., Zhu, J., & Wang, H. (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *Annals of Statistics*, 39, 2607–2625.

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

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

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

Fang, H. (2011). Stability Analysis of Block LDL^T Factorizations for Symmetric Indefinite Matrices. *IMA Journal of Numerical Analysis*, 528–555.

Fang, K. -T., Li, R., & Sudjianto, A. (2006). *Design and Modeling for Computer Experiments*. Chapman and Hall/CRC Press.

Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association*, 102, 321–331.

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

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97, 590–600.

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

Gramacy, R. B. & Lee, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103, 1119–1130.

Hung, Y. (2011). Penalized blind Kriging in computer experiments. *Statistica Sinica*, 21, 1171–1190.

Irvine, K. M., Gitelman, A. I., & Hoeting, J. A. (2007). Spatial designs and properties of spatial correlation: effects on covariance estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, 12, 450–469.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112, 201–214.

Kaufman, C., Bingham, D., Habib, S., Heitmann, K., & Frieman, J. A. (2011). Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *Annals of Applied Statistics*, 5, 2470–2492.

Kaufman, C., Schervish, M., & Nychka, D. (2008). Covariance tapering for likelihood-based estimation in large spatial datasets. *Journal of the American Statistical Assocication*, 103, 1545–1555.

Kennedy, M. C. & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society*, 63, 425–464.

Li, R. & Sudjianto, A. (2005). Analysis of computer experiments using penalized likelihood. *Technometrics*, 47, 111–120.

Lin, N. & Xi, R. (2011). Aggregated estimating equation estimation. *Statistics and its Interface*, 4, 73–83.

Liu, D., Liu, R., & Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association*, 110, 326–340.

Lopez, V. & Hamann, H. F. (2011). Heat transfer modeling in data centers. *International Journal of Heat and Mass Transfer*, 54, 5306–5318.

Mackey, L., Talwalkar, A., & Jordan, M. I. (2015). Distributed matrix completion and robust factorization. *Journal of Machine Learning Research*, 16, 913–960.

Mardia, K. V. & Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71, 135–146.

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

Park, C. & Apley, D. (2018). Parchwork Kriging for large-scale Gaussian process regression. *Journal of Machine Learning Research*, 19, 1–43.

Pissanetzky, S. (1984). *Sparse Matrix Technology-electronic edition*. Academic Press.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. CRC Press.

Rue, H. & Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian journal of Statistics*, 29, 31–49.

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and Analysis of Computer Experiments. *Statistical Science*, 4, 409–423.

Santner, T. J., Williams, B. J., & Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer.

Schifano, E. D., Wu, J., Wang, C., Yan, J., & Chen, M. -H. (2016). Online updating of statistical inference in the big data setting. *Technometrics*, 58, 393–403.

Schmidt, A. M. & O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 743–758.

Schweder, T. & Hjort, N. (2016). *Confidence, Likelihood and Probability*. Cambridge University Press, Cambridge.

Shen, J., Liu, R. Y., & Xie, M. -g. (2018). Prediction with confidence—a general framework for predictive inference. *Journal of Statistical Planning and Inference*, 195, 126–140.

Singh, K., Xie, M., & Strawderman, W. E. (2007). Confidence distribution (CD): distribution estimator of a parameter. *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond. IMS Lecture Notes-Monograph Series*, Vol. 45, 132–150.

Singh, K., Xie, M., & Strawderman, W. E. (2005). Combining information from independent sources through confidence distributions. *The Annals of Statistics*, 33, 159–183.

Sjöstedt-de Luna, S. (2003). The bootstrap and Kriging prediction intervals. *Scandinavian Journal of Statistics*, 30, 175–192.

Snelson, E. & Ghahramani, Z. (2005). Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, 1257–1264.

Song, Q. & Liang, F. (2015). A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 947–972.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer.

Stein, M. L. (2008). A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society*, 37, 3–10.

Stein, M. L. (2013). Statistical properties of covariance tapers. *Journal of Computational and Graphical Statistics*, 22, 866–885.

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

Sung, C. -L., Wang, W., Plumlee, M., & Haaland, B. (2019). Multiresolution functional ANOVA for large-scale, many-input computer experiments. *Journal of the American Statistical Association*, 1–23.

Surjanovic, S. & Bingham, D. (2017). *Virtual Library of Simulation Experiments: Test Functions and Datasets*. https://www.sfu.ca/~ssurjano/.

Tang, L., Zhou, L., & Song, P. X. -K. (2016). Method of divide-and-combine in regularised generalised linear models for big data. arXiv preprint arXiv:1611.06208.

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

Xie, M. & Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter (with discussions). *International Statistical Review*, 81, 3–39.

Xie, M., Singh, K., & Strawderman, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, 106, 320–333.

Yang, G., Liu, D., Liu, R. Y., Xie, M., & Hoaglin, D. (2014). A confidence distribution approach for an efficient network meta-analysis. *Statistical Methodology*, 20, 105–125.

Zhang, H. & Zimmerman, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, 92, 921–936.

Zhang, R., Lin, C. D., & Ranjan, P. (2018). Local Gaussian process model for large-scale dynamic computer experiments. *Journal of Computational and Graphical Statistics*, 27, 798–807.

Zhao, Y., Amemiya, Y., & Hung, Y. (2018). Efficient Gaussian process modeling using experimental design-based subagging. *Statistica Sinica*, 28, 1459–1479.

Received 25 October 2018
Accepted 01 March 2020