Received: 27 December 2022 | Revised: 10 November 2023 | Accepted: 1 February 2024

DOI: 10.1002/env.2844

RESEARCH ARTICLE

WILEY

Recursive nearest neighbor co-kriging models for big multi-fidelity spatial data sets

Si Cheng[1] | **Bledar A. Konomi**[1] | **Georgios Karagiannis**[2] | **Emily L. Kang**[1]

[1]Division of Statistics and Data Sciences, Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio, USA

[2]Department of Mathematical Sciences, Durham University, Durham, UK

Correspondence
Bledar A. Konomi, Division of Statistics and Data Sciences, Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH, USA.
Email: alex.konomi@uc.edu

Funding information
National Science Foundation, Grant/Award Number: DMS-2053668; Simons Foundation's Collaboration Award, Grant/Award Numbers: #317298, #712755

Abstract

Big datasets are gathered daily from different remote sensing platforms. Recently, statistical co-kriging models, with the help of scalable techniques, have been able to combine such datasets by using spatially varying bias corrections. The associated Bayesian inference for these models is usually facilitated via Markov chain Monte Carlo (MCMC) methods which present (sometimes prohibitively) slow mixing and convergence because they require the simulation of high-dimensional random effect vectors from their posteriors given large datasets. To enable fast inference in big data spatial problems, we propose the recursive nearest neighbor co-kriging (RNNC) model. Based on this model, we develop two computationally efficient inferential procedures: (a) the collapsed RNNC which reduces the posterior sampling space by integrating out the latent processes, and (b) the conjugate RNNC, an MCMC free inference which significantly reduces the computational time without sacrificing prediction accuracy. An important highlight of conjugate RNNC is that it enables fast inference in massive multifidelity data sets by avoiding expensive integration algorithms. The efficient computational and good predictive performances of our proposed algorithms are demonstrated on benchmark examples and the analysis of the High-resolution Infrared Radiation Sounder data gathered from two NOAA polar orbiting satellites in which we managed to reduce the computational time from multiple hours to just a few minutes.

KEYWORDS

nearest neighbor Gaussian process, recursive co-kriging, remote sensing

1 | INTRODUCTION

Global geophysical information is measured daily by numerous satellite sensors. Due to aging and exposure to the harsh environment of space the satellite sensors degrade over time, resulting in decreased performance reliability. Decreased performance may affect data measurement accuracy (Goldberg, 2011). In addition, newer satellites with technologically more advanced sensors provide information of higher fidelity than older sensors. These discrepancies in sensor performance have created the need to develop efficient methods to analyse daily global remote sensing measurements with

varying fidelity. Here, our work is motivated by a data set produced from the high-resolution infrared radiation sounder (HIRS), which provides hundred of thousands of measurements from multiple satellite platforms daily.

Multiple methods in remote sensing have been developed to assess satellite sensor performance and consistency (Chander et al., 2013; National Research Council, 2004; Xiong et al., 2010). These methods do not account for spatial correlation and oversimplify the relationship between sensors. Furthermore, statistical methods to analyse these data sets which account for spatial correlation pose challenges due to the multifideltiy presence as well as the size and computationally intensive procedures. Nguyen et al. (2012, 2017) have proposed data fusion techniques to model multivariate spatial data at potentially different spatial resolutions based on fixed ranked kriging (Cressie & Johannesson, 2008). The accuracy of this approach relies on the number of basis functions and can only capture large scale variation of the covariance function. When the data sets are dense, strongly correlated, and the noise effect is sufficiently small, the low rank kriging techniques have difficulty accounting for small scale variation (Stein, 2014).

Many statistical methods for large spatially correlated data sets have been developed over the past two decade. For instance, we distinguish, the low-rank approximation methods (Banerjee et al., 2008; Cressie & Johannesson, 2008), approximate likelihood methods (Gramacy & Apley, 2015; Stein et al., 2004), covariance tapering methods (Du et al., 2009; Furrer et al., 2006; Kaufman et al., 2008), sparse structures (Datta et al., 2016; Lindgren et al., 2011; Ma & Kang, 2020; Michele Peruzzi & Finley, 2022; Nychka et al., 2015), lower dimensional conditional distributions (Datta et al., 2016; Katzfuss & Guinness, 2021; Stein et al., 2004; Vecchia, 1988), and multiple-scale approximation (Abdulah et al., 2023; Katzfuss, 2017; Sang & Huang, 2012; Shirota et al., 2023). All these methods have been developed for data sets obtained from the same source or instrument which translates to a single fidelity data source. However, their extension to multi-fidelity data sets is not straightforward.

Autoregressive co-kriging models (Kennedy & O'Hagan, 2000; Le Gratiet, 2013; Qian et al., 2005), originally built for computer simulation problems, can be used for the analysis of multiple fidelity remote sensing observations with spatially nested structure and no random error. Nested design in the multifidelity setting means the design points at the higher fidelity levels are subsets of the lower fidelity ones. Konomi and Karagiannis (2021) and Ma et al. (2022) relaxed the nested design requirements by properly introducing an imputation mechanism. However, the aforesaid methods rely on Gaussian process models and are computationally impossible for big data problems. For cases when the observed space can be expressed as a tensor product Konomi et al. (2023) uses a separable covariance function within the co-kriging model to improve the computational efficiency. For large data sets which are irregularly positioned over space and contaminated with random error, Cheng et al. (2021) proposed the nearest neighbour co-kriging Gaussian process (NNCGP) to embed nearest neighbor Gaussian process (NNGP; Datta et al., 2016) into an autoregressive co-kriging model to make computations possible. NNCGP achieves this by using imputation ideas into the latent variables to construct a nested reference set of multiple NNGP levels. Although NNCGP makes the analysis of big multi-fidelity data sets computationally possible, its computational speed depends on an expensive iterative MCMC procedure which makes it impractical for analysing daily large data sets.

To overcome the iterative MCMC procedure, we propose a recursive formulation based on the latent variable of the NNCGP model following similar ideas with Le Gratiet and Garnier (2014), who proposed the recursive formulation directly into the observations. Based on this new formulation, which we call recursive nearest neighbors co-kriging (RNNC), we are able to build a nearest neighbors co-kriging model with T levels by building T conditionally independent NNGPs. This enables the development of two alternative inferential procedures which aim to reduce high-dimensional parametric space, improve convergence, and reduce computational time in comparison to the NNCGP. Both proposed procedures are able to address applications for large non-nested and irregular spatial data sets from different platforms and with varying quality. The first proposed procedure, called Collapsed RNNC, reduces the MCMC posterior sampling space by integrating out the spatial latent variables. Based on the collapsed RNNC, we propose an MCMC free procedure to speed up Bayesian inference. We build an algorithm which sequentially decomposes the parametric space into conditionally independent parts for each fidelity level where we can apply a K-fold optimization method. Each sequential step can be viewed as collapsed NNGP (Finley et al., 2019) where the bases functions of the Gaussian process mean are determined at the previous step. We name this second inferential procedure conjugate RNNC. We note that the MCMC free procedure proposed in (Finley et al., 2019) cannot be applied directly in the NNCGP model because the computational complexity of the K-fold cross-validation method depends on the dimension of the parametric space. Our simulation study and our analysis of the HERS data set shows that the proposed conjugate RNNC procedure reduces the computational time notably without significantly sacrificing prediction accuracy over the existing NNCGP approach.

The layout of the paper is as follows. In Section 2, we introduce the high-resolution infrared radiation sounder data studied in this work. In Section 3, we review the NNCGP model. In Section 4, we introduce the proposed RNNC model. In Section 4.1, we integrate out the latent variables from the model and design an MCMC algorithm for this model. In Section 4.2, we design an MCMC free approach tailored to the proposed RNNC model that facilitates parametric and predictive inference. In Section 5, we investigate the performance of the proposed procedure. Specifically in Section 5.1 we introduce two simulation studies and in Section 5.2 we implement the proposed method for the analysis of data sets from two satellites, NOAA-14 and NOAA-15. Finally, in Section 6 we give a summary and conclusion.

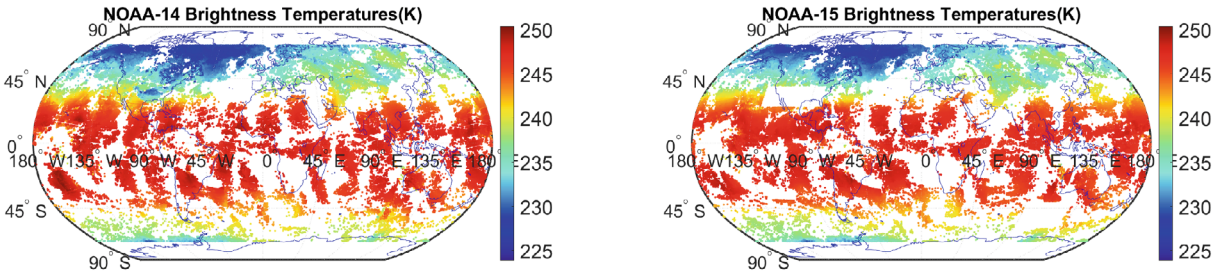
2 | HIGH-RESOLUTION INFRARED RADIATION SOUNDER DATA

Satellite soundings have been providing measurements of the Earth's atmosphere, oceans, land, and ice since the 1970s to support the study of global climate system dynamics. Long term observations from past and current environmental satellites are widely used in developing climate data records (CDR) (National Research Council, 2004). HIRS mission objectives include observations of atmospheric temperature, water vapor, specific humidity, sea surface temperature, cloud cover, and total column ozone. The HIRS instrument is comprised of 20 channels, including 12 longwave channels, seven shortwave channels, and one visible channel. The dataset being considered in this study is limb-corrected HIRS swath data as brightness temperatures (Jackson et al., 2003). The data is stored as daily files, where each daily file records approximately 120,000 geolocated observations. The current archive includes data from NOAA-5 through NOAA-17 along with Metop-02, covering the time period of 1978–2017. In all, this data archive is more than 2 TB, with an average daily file size of about 82 MB. The HIRS CRD faces some common challenges regarding the consistency and accuracy over time, due to degradation of sensors and intersatellite discrepancies. Furthermore, there is missing information caused by atmospheric conditions such as thick cloud cover.

We examine HIRS Channel 5 observations from a single day, March 1, 2001, as illustrated in Figure 1. On this day, we may exploit a period of temporal overlap in the NOAA POES series where two satellites captured measurements: NOAA-14 and NOAA-15. The HIRS sensors on these two satellites have similar technical designs which allow us to ignore the spectral and spatial footprint differences. NOAA-14 became operational in December 1994 while NOAA-15 became operational in October 1998. The spatial resolution footprint for both satellites is approximately 10 km at nadir. Given the sensor age difference, it is reasonable to consider that the instruments on-board NOAA-15 are in better condition than those of NOAA-14 and hence provide more accurate data. Therefore, we treat observations from NOAA-14 as a low fidelity dataset, and those from NOAA-15 as a high fidelity dataset.

3 | NEAREST NEIGHBOR CO-KRIGING GAUSSIAN PROCESS

Let $y_t(\mathbf{s})$ denote the output function at the spatial location s at fidelity level $t = 1, \ldots, T$ in a system with fidelity T levels. The fidelity level index t runs from the least accurate to the most accurate one. Let $z_t(\mathbf{s})$ denote the observed output at location s. We specify the co-kriging model as:



(a) Observations of NOAA 14 (b) Training data of NOAA 15

FIGURE 1 NOAA-14 brightness temperatures observation data-set, NOAA-15 brightness temperatures training data-set for channel 5 on March 1, 2001. (a) Observations of NOAA 14. (b) Training data of NOAA 15.

$$
\begin{aligned}
z_t(\mathbf{s}) &= y_t(\mathbf{s}) + \epsilon_t, \\
y_t(\mathbf{s}) &= \zeta_{t-1}(\mathbf{s}) y_{t-1}(\mathbf{s}) + \delta_t(\mathbf{s}), \\
\delta_t(\mathbf{s}) &= \mathbf{h}_t^T(\mathbf{s})\boldsymbol{\beta}_t + w_t(\mathbf{s}),
\end{aligned} \tag{1}
$$

where $z_t(\mathbf{s})$ is contaminated by additive random noise $\epsilon_t \sim N(0, \tau_t^2)$ for $t = 2, \ldots, T$, and $y_1(\mathbf{s}) = \mathbf{h}_1^T(\mathbf{s})\boldsymbol{\beta_1} + w_1(\mathbf{s})$ is the noiseless output. Here, $\zeta_{t-1}(\mathbf{s})$ and $\delta_t(\mathbf{s})$ represent the scale and additive discrepancies between systems with fidelity levels t and $t-1$, $\mathbf{h}_t(\cdot)$ is a vector of preselected bases functions, and $\boldsymbol{\beta}_t$ is a vector of coefficients at fidelity level t. The latent random function $w_t(\mathbf{s})$ is modeled as a Gaussian process, mutually independent for different t; that is, $w_t(\cdot) \sim GP(0, C_t(\cdot,\cdot;\boldsymbol{\theta}_t))$ where $C_t(\cdot,\cdot;\boldsymbol{\theta}_t)$ is a covariance function with covariance parameters $\boldsymbol{\theta}_t$ at fidelity level t. Any well defined covariance function can be used $C_t(\mathbf{s},\mathbf{s}'|\boldsymbol{\theta}_t) = \sigma_t^2 R(\mathbf{s},\mathbf{s}'|\boldsymbol{\phi}_t)$, where $\boldsymbol{\theta}_t = \{\sigma_t^2, \boldsymbol{\phi}_t\}$. This indicates that discrepancy term $\delta_t(\mathbf{s})$ is a Gaussian process. Finally, the unknown scale discrepancy function $\zeta_{t-1}(\mathbf{s})$ is modeled as a basis expansion $\zeta_{t-1}(\mathbf{s}|\boldsymbol{\gamma}_{t-1}) = \mathbf{g}_{t-1}(\mathbf{s})^T\boldsymbol{\gamma}_{t-1}$ (usually low degree), where $\mathbf{g}_t(\mathbf{s})$ is a vector of polynomial basis functions and $\{\boldsymbol{\gamma}_{t-1}\}$ is a vector of random coefficients, for $t = 2, \ldots, T$.

Let us assume the system is observed at n_t locations at fidelity level t. Let $\mathbf{S}_t = \{\mathbf{s}_{t,1}, \ldots, \mathbf{s}_{t,n_t}\}$ be the set of n_t observed locations, let $\mathbf{w}_t = w_t(\mathbf{S}_t) = \{w_t(\mathbf{s}_{t,1}), \ldots, w_t(\mathbf{s}_{t,n_t})\}$ the latent spatial random effect vector at fidelity level t, and let $\mathbf{Z}_t = z_t(\mathbf{S}_t) = \{z_t(\mathbf{s}_{t,1}), \ldots, z_t(\mathbf{s}_{t,n_t})\}$ represent the observed output at fidelity level t. If data $\{\mathbf{Z}_t\}$ are observed in non-nested locations across the fidelity levels, the calculation of the likelihood requires $\mathcal{O}((\sum_{t=1}^T n_t)^3)$ flops to invert the covariance matrix of the observations (denoted by $\boldsymbol{\Lambda}$) and additional $\mathcal{O}((\sum_{t=1}^T n_t)^2)$ memory to store it as explained in (Konomi & Karagiannis, 2021). To reduce the computational complexity, Cheng et al. (2021) proposed NNCGP which assigns conditionally independent NNGP models within a nested reference set. For Bayesian inference, Cheng et al. (2021) proposed a Gibbs sampler taking advantage of the nearest neighbor structure at each fidelity level. However, this sampler is based on updating a conditionally independent high dimensional latent variable which could cause slow convergence and high autocorrelation (Liu et al., 1994). The slow convergence can significantly increase the number of the Gibbs sampler iterations I. The overall computational cost of NNCGP for m neighbours and non-nested spatial locations is $\mathcal{O}(I \times (\sum_{t=1}^T n_t)m^3)$ floating point operations (flops). Also, for a fixed computational budget, the produced Monte Carlo estimates may be sensitive to the initial values of the MCMC sampler. Despite reducing the computational complexity to linear for every MCMC iteration, the number of the iterations can significantly increase computational cost. This simple observation makes the existing NNCGP too expensive for the vast majority of real remote sensing applications.

4 | RECURSIVE NEAREST NEIGHBOR CO-KRIGING MODEL

Improvement in the convergence of MCMC can be achieved by integrating out the latent variable $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_S)$ from the Bayesian hierarchical NNCGP model which allows dimension reduction in the sampling space and the involved posterior distributions. However, integrating out the latent variables $\mathbf{w}$ in the NNCGP model is not feasible under non-nested designs. This is because the posterior distribution of latent variable of the lower fidelity is affected by the likelihood of higher fidelity. To make possible the integration of latent variables $\mathbf{w}$, we propose a recursive formulation for the co-kriging model by using ideas similar to (Le Gratiet & Garnier, 2014). Precisely, our proposed recursive nearest neighbors co-kriging (RNNC) has the following hierarchical structure:

$$
\begin{aligned}
z_t(\mathbf{s}) &= y_t(\mathbf{s}) + \epsilon_t \\
y_t(\mathbf{s}) &= \zeta_{t-1}(\mathbf{s})\hat{y}_{t-1}(\mathbf{s}) + \delta_t(\mathbf{s}), \\
\delta_t(\mathbf{s}) &= \mathbf{h}_t^T(\mathbf{s})\boldsymbol{\beta}_t + w_t(\mathbf{s}),
\end{aligned} \tag{2}
$$

where $\delta_t(\mathbf{s})$ is a Gaussian process as before and $\hat{y}_{t-1}(\mathbf{s})$ is a Gaussian process with distribution $[\hat{y}_{t-1}(\mathbf{s})|\mathbf{Z}_{t-1}, \hat{y}_{t-2}(\mathbf{s}), \boldsymbol{\theta}_{t-1}, \boldsymbol{\beta}_{t-1}]$. Essentially, we express $y_t(\mathbf{s})$ (the Gaussian process response at level t) as a function of the Gaussian process $y_{t-1}(\mathbf{s})$ conditioned by the values $\mathbf{Z}^{(t-1)} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{t-1})$. For computational efficiency, we assume NNGP independent priors for $w_t(\mathbf{s})$, $t = 1, \ldots, T$. Based on the NNGP priors, the conditional distribution can be computed for all types of reference sets. So, based on the recursive representation we can relax the nested condition on the NNCGP nested reference set. Specifically,

$$
\hat{y}_{t-1}(\mathbf{s})|\mathbf{Z}_{t-1}, \hat{y}_{t-2}(\mathbf{s}), \boldsymbol{\theta}_{t-1}, \boldsymbol{\beta}_{t-1} \sim N(\zeta_{t-2}\hat{y}_{t-2}(\mathbf{s}) + \mathbf{h}_{t-1}^T(\mathbf{s})\boldsymbol{\beta}_{t-1} + V_{t-1,\mathbf{s}}\mu_{t-1,\mathbf{s}}, V_{t-1,\mathbf{s}}), \tag{3}
$$

with, $\mu_{t-1,\mathbf{s}} = V^{-1}_{t-1,\mathbf{s}}\mathbf{B}_{t-1,\mathbf{s}}\left[z_{t-1}(N_{t-1}(\mathbf{s})) - \mathbf{h}^T_{t-1}(N_{t-1}(\mathbf{s}))\boldsymbol{\beta}_{t-1} - \zeta_{t-2}(N_{t-1}(\mathbf{s}))\circ\hat{y}_{t-2}(N_{t-1}(\mathbf{s}))\right]$, $\mathbf{B}_{t,\mathbf{s}} = C^T_{\mathbf{s},N_t(\mathbf{s})}C^{-1}_{N_t\mathbf{s}}$, and $V_{t,\mathbf{s}} = C(\mathbf{s},\mathbf{s}) - C^T_{\mathbf{s},N_t(\mathbf{s})}C^{-1}_{N_t\mathbf{s}}C_{\mathbf{s},N_t(\mathbf{s})}$. The $\circ$ represents the Hadamard product between two matrices.

Using the Markovian property of the co-kriging model (O'Hagan, 1998), the joint likelihood of the proposed model in (2) can be factorized as a product of likelihoods at different fidelity levels conditional on $\hat{y}_{t-1}(\mathbf{S}_t) = \{\hat{y}_{t-1}(\mathbf{s}_{t,1}), \ldots, \hat{y}_{t-1}(\mathbf{s}_{t,n_t})\}$ for $t = 2, \ldots, T$ and prior $\mathbf{w}_t$ for $t = 1, \ldots, T$, that is,

$$\begin{aligned} L(\boldsymbol{Z}_{1:T}|\cdot) &= p(\mathbf{Z}_1|\mathbf{w}_1, \boldsymbol{\beta}_1, \tau_1)\prod_{t=2}^{T} p(\mathbf{Z}_t|\mathbf{w}_t, \boldsymbol{\beta}_t, \hat{y}_{t-1}(\mathbf{S}_t), \boldsymbol{\gamma}_{t-1}, \tau_t) \\ &= N(\mathbf{Z}_1|\mathbf{h}_1(\mathbf{S}_1)\boldsymbol{\beta}_1 + \mathbf{w}_1, \tau_1\mathbf{I})\prod_{t=2}^{T} N(\mathbf{Z}_t|\zeta_{t-1}(\mathbf{S}_t)\circ\hat{y}_{t-1}(\mathbf{S}_t) + \mathbf{h}_t(\mathbf{S}_t)\boldsymbol{\beta}_t + \mathbf{w}_t, \tau_t\mathbf{I}), \end{aligned} \tag{4}$$

where $(\cdot)$ denotes all the parameters associated with the model. This representation makes it possible to integrate out the latent variable $\mathbf{w}_t$ independently for each fidelity level $t = 1, \ldots, T$.

4.1 | Collapsed recursive nearest neighbor co-kriging model

We represent the multivariate Gaussian latent variable $\mathbf{w}_t(\mathbf{S}_t)$ as a linear model:

$$\begin{aligned} w_t(\mathbf{s}_{t,1}) &= 0 + \eta_{t,1}, \\ w_t(\mathbf{s}_{t,i}) &= a_{t,i,1}w_t(\mathbf{s}_{t,1}) + a_{t,i,2}w_t(\mathbf{s}_{t,2}) + \cdots + a_{t,i,i-1}w_t(\mathbf{s}_{t,i-1}) + \eta_{t,i}, \quad \text{for } i = 2, \ldots, n_t \end{aligned}$$

for $t = 1, \ldots, T$. We set $\eta_{t,i} \sim N(0, d_{t,i,i})$ independently for all t, i, $d_{t,1,1} = \delta(w_{t,1})$ and $d_{t,i,i} = \delta(w_{t,i}|\{w_{t,j}; j < i\})$ for $i = 2, \ldots, n_t$ and $t = 1, \ldots, T$. In a matrix form we can write $\mathbf{w}_t(\mathbf{S}_t) = \mathbf{A}_t\mathbf{w}_t(\mathbf{S}_t) + \boldsymbol{\eta}_t$, where $\mathbf{A}_t$ is an $n \times n$ strictly lower-triangular matrix and $\boldsymbol{\eta}_t \sim N(0, \mathbf{D})$ and $\mathbf{D}$ is diagonal. Based on the structure of $\mathbf{A}_t$, we can write the covariance of each level as $\mathbf{C}_t(\boldsymbol{\theta}_t) = (\mathbf{I}_t - \mathbf{A}_t)^{-1}\mathbf{D}_t(\mathbf{I}_t - \mathbf{A}_t)^{-T}$. The NNGP prior constructs a sparse strictly lower triangular matrix $\mathbf{A}$ with no more than m (where $m \ll n$) non-zero entries in each row resulting in an approximation of the covariance matrix $\mathbf{C}_t$. So the approximated inverse $\tilde{\mathbf{C}}^{-1}_t(\boldsymbol{\theta}_t) = (\mathbf{I}_t - \mathbf{A}_t)\mathbf{D}^{-1}_t(\mathbf{I}_t - \mathbf{A}_t)^T$ is a sparse matrix and can be computed based on $\mathcal{O}(n_t m^3)$ operations.

We call the integrated version of the above model collapsed RNNC model. Specifically, after integrating out $\mathbf{w}_t$ the proposed RNNC model can be written as:

$$\begin{aligned} z_1(\mathbf{S}_1)|\boldsymbol{\beta}_1, \boldsymbol{\theta}_1, \tau_1 &\sim N(\mathbf{h}^T_1(\mathbf{S}_1)\boldsymbol{\beta}_1, \tilde{\Lambda}_1(\mathbf{S}_1, \boldsymbol{\theta}_1, \tau_1)), \\ z_t(\mathbf{S}_t)|\boldsymbol{\beta}_t, \boldsymbol{\theta}_t, \tau_t, \zeta_t(\mathbf{S}_t), \hat{y}_{t-1}(\mathbf{S}_t) &\sim N(\zeta_t(\mathbf{S}_t)\circ\hat{y}_{t-1}(\mathbf{S}_t) + \mathbf{h}^T_t(\mathbf{S}_t)\boldsymbol{\beta}_t, \tilde{\boldsymbol{\Lambda}}_t), \end{aligned} \tag{5}$$

for $t = 2, \ldots, T$, where $\tilde{\Lambda}_t(\boldsymbol{\theta}_t, \tau_t) = \tilde{C}_t(\boldsymbol{\theta}_t) + \tau^2_t\mathbf{I} = \sigma^2_t\tilde{\mathbf{R}}_t(\boldsymbol{\phi}_t) + \tau^2_t\mathbf{I}$ is the covariance matrix of the observations, $\tilde{C}_t(\boldsymbol{\theta}_t)$ is the sparse covariance matrix with parameters $\boldsymbol{\theta}_t = \{\sigma^2_t, \boldsymbol{\phi}_t\}$ and τ^2_t is the variance of the error ϵ_t at level t. By applying Sherman-Morrison-Woodbury formula, the inverse and determinant of $\tilde{\boldsymbol{\Lambda}}$ get the computationally convenient form

$$\begin{aligned} \tilde{\boldsymbol{\Lambda}}^{-1}_t &= \tau^{-2}_t\mathbf{I} - \tau^{-4}_t(\tilde{\mathbf{C}}_t(\boldsymbol{\theta}_t)^{-1} + \tau^{-2}_t\mathbf{I})^{-1}, \\ \det(\tilde{\boldsymbol{\Lambda}}_t) &= \tau^{2n}_t \det(\tilde{\mathbf{C}}_t(\boldsymbol{\theta}_t)) \det(\tilde{\mathbf{C}}_t(\boldsymbol{\theta}_t)^{-1} + \tau^{-2}_t\mathbf{I}). \end{aligned}$$

For simplicity let us denote $\boldsymbol{\Theta}_t = (\boldsymbol{\beta}_t, \gamma_t, \boldsymbol{\theta}_t, \tau_t)$. Based on this representation, the joint posterior approximation of all the unknowns is:

$$\begin{aligned} p(\boldsymbol{\Theta}_{1:T}, \hat{\mathbf{y}}_{1:T-1}(\mathbf{S}^*_{2:T})|\mathbf{Z}_{1:T}) &= p(\boldsymbol{\Theta}_1|\mathbf{Z}_1)\prod_{t=2}^{T} p(\boldsymbol{\Theta}_t, \hat{\mathbf{y}}_1(\mathbf{S}^*_t)|\mathbf{Z}_t) \\ &\propto p(\boldsymbol{\Theta}_1)\tilde{L}(\mathbf{Z}_1|\boldsymbol{\Theta}_1)\prod_{t=2}^{T} p(\boldsymbol{\Theta}_t)\tilde{L}(\mathbf{Z}_t|\boldsymbol{\Theta}_t, \hat{\mathbf{y}}_{t-1}(\mathbf{S}_t))\tilde{p}(\hat{\mathbf{y}}_{t-1}(\mathbf{S}^*_t)|\cdot), \end{aligned} \tag{6}$$

where $\tilde{L}(\mathbf{Z}_t|\boldsymbol{\Theta}_t, \hat{\mathbf{y}}_{t-1}(\mathbf{S}_t))$ is the approximated likelihood using the sparse representation and $\tilde{p}(\hat{\mathbf{y}}_{t-1}(\mathbf{S}_t^*)|\cdot)$ the nearest neighbor Gaussian process prediction at locations $\mathbf{S}_t^* = \bigcup_{i=t+1}^{T}\mathbf{S}_i \setminus \mathbf{S}_t = \{s_{t,1}^*, \ldots, s_{t,n_t^*}^*\}$ as a set of knots of fidelity level t. This contains the observed locations that are not in the t^{th} level but in the higher fidelity levels. The $(\cdot)$ represents the parameters and data necessary to produce the prediction distortion at level $t-1$. Note that the prediction probability can be excluded for cases with hierarchically nested structure for the spatial locations. For each level, a Gibbs sampler can be employed to fascilitate inference based on the conditional representation $p(\boldsymbol{\Theta}_T|\hat{\mathbf{y}}_{t-1}(\mathbf{S}_t), \mathbf{Z}_T)$ and $p(\hat{\mathbf{y}}_{t-1}(\mathbf{S}_t^*)|\boldsymbol{\Theta}_T, \mathbf{Z}_T)$ which is given in Equation (3).

For $\boldsymbol{\Theta}_t = (\boldsymbol{\beta}_t, \gamma_t, \theta_t, \tau_t)$, by assigning independent conjugate prior $\boldsymbol{\beta}_t \sim N(\boldsymbol{\mu}_{\beta_t}, \mathbf{V}_{\beta_t})$ and $\gamma_t \sim N(\mu_{\gamma_t}, \mathbf{V}_{\gamma_t})$, we achieve explicit forms of the conditional distribution for parameters $\boldsymbol{\beta}_t$ and γ_t as:

$$\boldsymbol{\beta}_t|\mathbf{Z}_t, \hat{\mathbf{y}}_{t-1}(\mathbf{S}_t), \theta_t, \boldsymbol{\gamma}_t\tau_t^2 \sim N(\mathbf{V}_{\beta_t}^*\boldsymbol{\mu}_{\beta_t}^*, \mathbf{V}_{\beta_t}^*), \tag{7}$$

$$\boldsymbol{\gamma}_t|\mathbf{Z}_t, \hat{\mathbf{y}}_{t-1}(\mathbf{S}_t), \theta_t, \boldsymbol{\beta}_t, \tau_t^2 \sim N(\mathbf{V}_{\gamma_t}^*\boldsymbol{\mu}_{\gamma_t}^*, \mathbf{V}_{\gamma_t}^*), \tag{8}$$

where $\mathbf{V}_{\beta_t}^*\boldsymbol{\mu}_{\beta_t}^*, \mathbf{V}_{\gamma_t}^*, \boldsymbol{\mu}_{\gamma_t}^*$ are given in Appendix C. Finally, for each fidelity level t, we use a Metropolis-Hastings (MH) algorithm targeting the distribution $p(\theta_t, \tau_t^2|\mathbf{Z}_t, \hat{\mathbf{y}}_{t-1}(\mathbf{S}_t), \boldsymbol{\beta}_t)$ to carry out the inference.

In the special case of hierarchical nested structure for the spatial locations, we can avoid sampling from $p(\hat{\mathbf{y}}_{t-1}(\mathbf{S}_t^*)|\boldsymbol{\Theta}_T, \mathbf{Z}_T)$ using the observed locations $z_{t-1}(\mathbf{S}_t^*)$. We can prove that the mean and variance of the predictive distribution at level T of the collapsed RNNC is the same as the mean and variance of the predictive distribution of the NNCGP. The proof is very similar to Le Gratiet and Garnier (2014) in the sense that we just need to substitute the GP priors with the NNGP priors and add the nugget effect in each level.

4.2 | Conjugate recursive nearest neighbor co-kriging model

Both NNCGP and collapsed RNNC models rely on the MCMC inference which can be practically prohibitive when analyzing thousands or millions of spatial data sets. Following recent work by Finley et al. (2019), we propose a MCMC free procedure to achieve exact Bayesian inference at a more practical time. Because the computational efficiency of the estimation procedure in Finley et al. (2019) is sensitive to the number of parameters, it cannot be applied directly to our model. We utilize the RNNC model conditionally independent posterior representation to decompose the parametric space into smaller different groups based on the fidelity levels. To make MCMC free inference possible, we re-parameterize the covariance function of the collapsed recursive co-kriging model as $\tilde{\boldsymbol{\Lambda}}_t(\theta_t, \tilde{\tau}_t^2) = \sigma_t^2\tilde{\boldsymbol{\Sigma}}_t$, where $\tilde{\boldsymbol{\Sigma}}_t = \tilde{\mathbf{R}}_t + \tilde{\tau}_t^2\mathbf{I}$, $\tilde{\mathbf{R}}_t$ is the nearest-neighbor approximation correlation matrix, and $\tilde{\tau}_t^2 = \frac{\tau_t^2}{\sigma_t^2}$. To avoid the computational bottleneck due to the MCMC, we propose to make fast estimation $(\boldsymbol{\phi}_t, \tilde{\tau}_t^2)$ through a cross-validation approach for each level as well as use the prediction means of $\mathbf{y}_{t-1}(\mathbf{S}_t)$ based on the estimated values. We estimate $\hat{y}_t(\mathbf{S}_t^*)$ by the posterior mean $\bar{\hat{y}}_t(\mathbf{S}_t^*) = \mathbf{1}_{t>1}(t)\mathbf{g}_{t-1}^T(\mathbf{S}_t^*)\hat{\boldsymbol{\gamma}}_{t-1}\hat{y}_{t-1}(\mathbf{S}_t^*) + \mathbf{h}_t^T(\mathbf{S}_t^*)\hat{\boldsymbol{\beta}}_t + V_{t,\mathbf{S}_t^*}\mu_{t,\mathbf{S}_t^*}$. In the case that we have nested locations, $y_t(\mathbf{s}_u)$ for a location $\mathbf{s}_u \in \mathbf{S}_{t-1}$ is estimated with an empirical approach $\bar{\hat{y}}_{t-1}(\mathbf{s}_u)$ as $z_{t-1}(\mathbf{s}_u)$ and its variance is equal to the variance of the nugget effect. Given $\boldsymbol{\phi}_t$, $\tilde{\tau}_t^2$ and $\hat{y}_t(\mathbf{S}_t)$, the covariance matrix $\tilde{\boldsymbol{\Sigma}}_t$ can be calculated analytically.

For computational convenience, we assign an independent conjugate prior for the parameters of each level such as $p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T, \sigma_1^2, \ldots, \sigma_T^2, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_T) = \prod_{t=1}^{T}p(\boldsymbol{\beta}_t)p(\sigma_t^2)p(\boldsymbol{\gamma}_t)$ such as $\boldsymbol{\beta}_t \sim N(\boldsymbol{\mu}_{\beta_t}, \sigma_t^2\mathbf{V}_{\beta_t})$, $\sigma_t^2 \sim IG(a_t, b_t)$, and $\boldsymbol{\gamma}_t \sim N(\boldsymbol{\mu}_{\gamma_t}, \sigma_{t+1}^2\mathbf{V}_{\gamma_t})$. Based on this specifications, the posterior density function can be separated for each level t such as:

$$\begin{aligned} p(\boldsymbol{\beta}_t, \boldsymbol{\gamma}_{t-1}, \sigma_t^2|\mathbf{Z}_t, \hat{y}_{t-1}(\mathbf{S}_t)) \propto\ & IG(\sigma_t^2|a_t, b_t)N(\boldsymbol{\beta}_t|\boldsymbol{\mu}_{\beta_t}, \sigma_t^2\mathbf{V}_{\beta_t})N(\boldsymbol{\gamma}_{t-1}|\boldsymbol{\mu}_{\gamma_{t-1}}, \sigma_t^2\mathbf{V}_{\gamma_{t-1}}) \\ & \times N(\mathbf{Z}_t|\zeta_{t-1}(\mathbf{S}_t)\circ\hat{y}_{t-1}(\mathbf{S}_t) + \mathbf{h}_t^T\boldsymbol{\beta}_t, \sigma_t^2\tilde{\boldsymbol{\Sigma}}_t). \end{aligned} \tag{9}$$

We can compute the full conditional density function of $\boldsymbol{\gamma}_{t-1}$, $\boldsymbol{\beta}_t$, and σ_t^2 as

$$\boldsymbol{\gamma}_{t-1}|\boldsymbol{\beta}_t, \sigma_t^2, \mathbf{Z}_t, \hat{y}_{t-1}(\mathbf{S}_t)) \sim N(\boldsymbol{\gamma}_{t-1}|\tilde{\mathbf{V}}_{\gamma_{t-1}}\tilde{\boldsymbol{\mu}}_{\gamma_{t-1}}, \sigma^2\tilde{\mathbf{V}}_{\gamma_{t-1}}), \tag{10}$$

$$\boldsymbol{\beta}_t|\sigma_t^2, \mathbf{Z}_t, \hat{y}_{t-1}(\mathbf{S}_t) \sim N(\boldsymbol{\beta}_t|\tilde{\mathbf{V}}_{\beta_t}\tilde{\boldsymbol{\mu}}_{\beta_t}, \sigma_t^2\tilde{\mathbf{V}}_{\beta_t}) \tag{11}$$

$$\sigma_t^2|\mathbf{Z}_t, \hat{y}_{t-1}(\mathbf{S}_t) \sim IG(\sigma_t^2|a_t^*, b_t^*) \tag{12}$$

Algorithm 1. The algorithm steps for the MCMC free conjugate RNNC procedure. MCMC free posterior sampling for multi-fidelity level system with T levels.

step 1 Start from fidelity level 1($t = 1$), construct a set L_t that contains l_t number of candidates of parameters $\boldsymbol{\phi}_t$ and $\tilde{\tau}_t^2$.

step 2 Choose a $(\phi_t, \tilde{\tau}_t{}^2)$ from L_t. Split the data set of fidelity level t into K folds.

step 3 Remove k^{th} fold of data set $\mathbf{S}_t$, denoteas $\mathbf{S}_{t,k}$, then estimate $\sigma_t^2|\mathbf{Z}_t, \hat{y}_{t-1}(\mathbf{S}_t)$ with the posterior mean $\hat{\sigma}_t^2 = \frac{b_t^*}{a_t^*-1}$ of (12). Estimate $\boldsymbol{\beta}_t|\sigma_t^2, \mathbf{Z}_t, \hat{y}_{t-1}(\mathbf{S}_t)$ with the posterior mean $\hat{\boldsymbol{\beta}}_t = \tilde{\mathbf{V}}_{\beta_t}\tilde{\boldsymbol{\mu}}_{\beta_t}$ of (11). Estimate $\boldsymbol{\gamma}_{t-1}|\boldsymbol{\beta}_t, \sigma_t^2, \mathbf{Z}_t, \hat{y}_{t-1}(\mathbf{S}_t)$ with the posterior mean $\hat{\boldsymbol{\gamma}}_{t-1} = \tilde{\mathbf{V}}_{\gamma_{t-1}}\tilde{\boldsymbol{\mu}}_{\gamma_{t-1}}$ of (10).

step 4 Predict test data set $z_t(\mathbf{S}_{t,k})$ by posterior mean

$$\hat{z}_t(\mathbf{S}_{t,k}) = \mathbf{1}_{t>1}(t)\mathbf{g}_{t-1}^T(\mathbf{S}_{t,k})\hat{\boldsymbol{\gamma}}_{t-1}\hat{y}_{t-1}(\mathbf{S}_{t,k}) + \mathbf{h}_t^T(\mathbf{S}_{t,k})\hat{\boldsymbol{\beta}}_t + V_{t,\mathbf{S}_{t,k}}\mu_{t,\mathbf{S}_{t,k}}.$$

step 5 Repeat steps 3–4 over all K folds, calculate the average root mean square prediction error(RMSPE) by

$$\text{RMSPE} = \frac{\sum_{k=1}^{K}[\sum_{\mathbf{s}=\mathbf{S}_{t,k}}(z_t(\mathbf{s}) - \hat{z}_t(\mathbf{s}))^2/n_k]}{K}.$$

step 6 Repeat steps 2–5 over all values in candidate set L_t, choose the value of $\hat{\boldsymbol{\phi}}_t$ and $\hat{\tilde{\tau}}_t^2$ that minimizes the RMSPE. Repeat step 3 on full data set $\mathbf{S}_t$ by fixing $\boldsymbol{\phi}_t = \hat{\boldsymbol{\phi}}_t$, $\boldsymbol{\sigma}_t^2 = \hat{\boldsymbol{\sigma}}_t^2$. Estimate $\hat{y}_t(\mathbf{S}_t^*)$ by posterior mean

$$\bar{\hat{y}}_t(\mathbf{S}_t^*) = \mathbf{1}_{t>1}(t)\mathbf{g}_{t-1}^T(\mathbf{S}_t^*)\hat{\boldsymbol{\gamma}}_{t-1}\hat{y}_{t-1}(\mathbf{S}_t^*) + \mathbf{h}_t^T(\mathbf{S}_t^*)\hat{\boldsymbol{\beta}}_t + V_{t,\mathbf{S}_t^*}\mu_{t,\mathbf{S}_t^*}$$

step 7 For a new input location $\mathbf{s}_p$, predict $y_t(\mathbf{s}_p)$ by posterior:

$$\hat{y}_{t-1}(\mathbf{s})|\mathbf{Z}_{t-1}, \hat{y}_{t-2}(\mathbf{s})) \sim N(\zeta_{t-2}\hat{y}_{t-2}(\mathbf{s}) + \mathbf{h}_{t-1}^T(\mathbf{s})\boldsymbol{\beta}_{t-1} + V_{t-1,\mathbf{s}}\mu_{t-1,\mathbf{s}}, V_{t-1,\mathbf{s}})$$

Find a confidence interval based on the quantiles of the above distributions.

step 8 Repeat steps 1–7 over all T fidelity levels.

were $\tilde{\mathbf{V}}_{\gamma_{t-1}}, \tilde{\boldsymbol{\mu}}_{\gamma_{t-1}}, \tilde{\mathbf{V}}_{\beta_t}\tilde{\boldsymbol{\mu}}_{\beta_t}, \sigma_t^2\tilde{\mathbf{V}}_{\beta_t}, a_t^*$, and b_t^* are given analytically in Appendix D. Note that for $t = 1$, $\boldsymbol{\gamma}_0$ and $y_0(\mathbf{S}_t)$ do not exist. Also the conditional posterior density function of $\boldsymbol{\beta}_1$ and σ_1^2 are slightly different as explained in Appendix D.

A K-fold cross-validation method is used for the selection of optimal values for the parameters $\boldsymbol{\phi}_t$ and $\tilde{\tau}_t^2$ at level t that provide best prediction performance for the model, from a group of candidates. The criteria for choosing $\boldsymbol{\phi}_t$ and $\tilde{\tau}_t^2$ can be the root mean square prediction error (RMSPE) over the K folds of data set. The geolocated observations of the t fidelity are partitioned into K equal size subsets. Then, one of the subsets is used as a test set and the others are used for training. The procedure is repeated K times such that each subset is used once as a test set. The computational complexity of these procedures is reduced significantly from the use of the NNGP priors in the recursive co-kriging model. The estimation, tuning and prediction procedure of conjugate RNNC model are given in Algorithm 1. Similar to NNCGP model, the conjugate RNNC model analyzes the data set of each fidelity level sequentially from the lowest level to the highest. For each single fidelity level t, the conjugate RNNC model is able to run in parallel for tuning the parameter $\boldsymbol{\phi}_t$ and $\tilde{\tau}_t^2$ using a K fold cross validation procedure. Step 3, for given values of $(\boldsymbol{\phi}_t, \tilde{\tau}_t^2)$, to estimate $(\hat{\boldsymbol{\beta}}_t, \hat{\sigma}_t^2)$ requires $\mathcal{O}(n_t m^3 + n_t m p_t^2)$ floating point operations (flops) where p_t is the dimension of $\boldsymbol{\beta}_t$. Step 6, to predict at new locations $\mathbf{S}_t^*$ requires $\mathcal{O}(n_t^* m^3)$ flops where n_t^* is the dimension of $\mathbf{S}_t^*$. Step 7, to predict at a new location requires $\mathcal{O}(m^3)$ flops. When we use parallel computing within a fidelity level for parameters $(\boldsymbol{\phi}_t, \tilde{\tau}_t^2)$ the computation in step 3 becomes extremely fast. The conjugate RNNC model provides an empirical estimation of spatial effect parameter $\boldsymbol{\phi}_t$ and noise parameter $\tilde{\tau}_t{}^2$ within a given resolution. We note that the proposed MCMC free inference can be viewed as a sequential optimization technique which splits the parametric space into several lower dimension components where we can apply conditional independent conjugate NNGP models.

5 | SYNTHETIC DATA EXAMPLE AND REAL DATA ANALYSIS

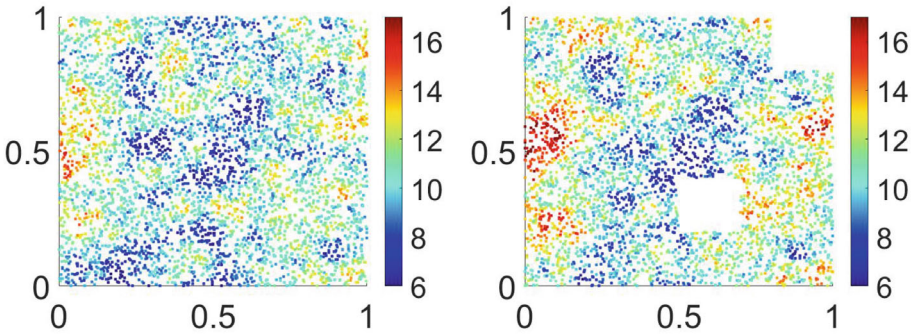
We study the performance of our proposed procedures, the conjugate RNNC and the collapsed RNNC, as well as compare their performances with that of the sequential NNCGP. The empirical study is based on one synthetic data set example with nested and one with non-nested input data sets. Also we use a real satellite data set application. As measures of performance we use the root mean squared prediction errors (RMSPE), coverage probability of the 95% equal tail credible interval (CVG (95%)), average length of the 95% equal tail credible interval (ALCI (95%)), and continuous rank probability score (CRPS) (Gneiting & Raftery, 2007). Details of these measurements are given in Appendix E. The simulations were performed in MATLAB R2018a, on a personal computer with specifications (intelR i7-3770 3.4GHz Processor, RAM 8.00GB, MS Windows 64bit).We have also included a simulation study with four levels of fidelity in the supplementary materials.

5.1 | Simulation study

We consider a two-fidelity level system represented by the hierarchical statistical model (1) defined on a two dimensional unit square domain with univariate observation data sets for both $\mathbf{Z}_1$ and $\mathbf{Z}_2$. Let the design matrix be $\mathbf{h}(\mathbf{S}_t) = \mathbb{1}$, the autoregressive coefficient function be an unknown constant $\zeta_1(\mathbf{s}) = \gamma_1$, and exponential covariance functions. We generate two synthetic data sets for the above statistical model. The true values of the parameters are listed in Tables 1 and 2. The data sets on the nested spatial locations consists of observations $\mathbf{Z}_1$ and $\mathbf{Z}_2$ from 100×100 grids $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively. The data sets, shown in Figure 2a,b are based on a fully non-nested input where the low fidelity observations $\mathbf{Z}_1$ and the high fidelity observations $\mathbf{Z}_2$ are generated at irregularly located at point in sets $\mathbf{S}_1$ and $\mathbf{S}_2$ of size 5000, while $\mathbf{S}_1 \cap \mathbf{S}_2 = \emptyset$. In all data sets, a few small square regions from $\mathbf{Z}_2$ are treated as a testing data-set, and the rest of $\mathbf{Z}_2$ and $\mathbf{Z}_1$ are treated as training data sets. The testing regions for the non-nested input can be seen as white boxes in Figure 2b.

Regarding the Bayesian inference, we compared the sequential NNCGP model, with the proposed collapsed RNNC model and that with the conjugate RNNC model, on both nested and non-nested data sets. We assigned similar non-informative priors for all the four models. We assign independent conjugate prior on parameters $\beta_1 \sim N(0, 1000)$, $\beta_2 \sim N(0, 1000)$, and scale parameter γ_1. We assign independent inverse gamma prior on spatial variance parameters $\sigma_1^2 \sim IG(2, 1)$, and $\sigma_2^2 \sim IG(2, 1)$ and on the noise parameters $\tau_1^2 \sim IG(2, 1)$, $\tau_2^2 \sim IG(2, 1)$. We also assign uniform prior on the range parameters $\phi_1 \sim U(0, 25)$, and $\phi_2 \sim U(0, 25)$. For the collapsed RNNC model, we run Markov chain Monte Carlo (MCMC) samplers for 35,000 iterations where the first 5000 iterations are discarded as a burn-in, and convergence of the MCMC sampler was diagnosed from the individual trace plots. The RMSPE with a five-fold cross-validation was used for the conjugate RNNC model. We select $(\phi_t, \tilde{\tau}_t^2)$ on a grid, for ϕ_t from the range $[0.1, 25]$, and for $\tilde{\tau}_t^2$ from the range $[0.0005, 0.4]$. No significant differences were observed when we used three-fold cross-validation and seven-fold cross-validation approach.

In Tables 1 and 2, we report the Monte Carlo estimates of the posterior means and the associated 95% marginal credible intervals of the unknown parameters using the two different NNCGP based procedures: sequential NNCGP, collapsed



(a) Non-nested low fidelity observations (b) Non-nested high fidelity observations

FIGURE 2 Observations for two fidelity level structure of non-nested observed input space. White boxes indicate the testing regions. (a) Non-nested low fidelity observations. (b) Non-nested high fidelity observations.

TABLE 1 The estimation of parameters in nested input dataset, using sequential NNCGP, collapsed RNNC and conjugate RNNC models.

		Nested data-set				
	True values	Sequential NNCGP		Collapsed RNNC		Conjugate RNNC
β_1	10	10.29	(9.93,10.57)	9.96	(9.60,10.32)	10.02
β_2	1	0.77	(0.59,1.04)	0.87	(0.59,1.13)	0.82
σ_1^2	4	3.55	(2.77,4.38)	3.46	(2.96,4.27)	3.15
σ_2^2	1	0.81	(0.27, 2.05)	0.98	(0.43, 1.88)	0.79
$1/\phi_1$	10	10.42	(8.15,13.47)	10.50	(8.59,13.90)	12.1
$1/\phi_2$	10	14.96	(3.37, 20.29)	15.69	(5.92, 19.98)	19.6
γ_1	1	0.99	(0.98,1.00)	0.99	(0.98,1.00)	0.99
τ_1^2	0.1	0.12	(0.10,0.14)	0.15	(0.10,0.19)	0.12
τ_2^2	0.05	0.07	(0.03,0.11)	0.10	(0.04,0.19)	0.16
m	10	—	—	—	—	—

TABLE 2 The estimation of parameters in non-nested input dataset using sequential NNCGP, collapsed RNNC and conjugate RNNC models.

		Non-nested data-set				
	True values	Sequential NNCGP		Collapsed RNNC		Conjugate RNNC
β_1	10	9.71	(9.36, 10.16)	9.97	(9.52,10.41)	9.71
β_2	1	0.87	(0.39,1.36)	1.23	(0.24,2.19)	1.27
σ_1^2	4	3.51	(2.71,4.52)	3.28	(3.02,3.72)	3.84
σ_2^2	1	1.05	(0.18,2.31)	1.00	(0.64, 1.49)	1.19
$1/\phi_1$	10	10.77	(8.07,13.91)	13.23	(9.93,15.85)	6.50
$1/\phi_2$	10	12.61	(3.93,24.07)	16.01	(12.84, 19.92)	20.00
γ_1	1	0.99	(0.98,1.05)	0.97	(0.94,0.99)	0.96
τ_1^2	0.1	0.13	(0.10,0.15)	0.10	(0.07,0.15)	0.11
τ_2^2	0.05	0.16	(0.04,0.23)	0.18	(0.03,0.29)	0.10
m	10	—	—	—	—	—

RNNC, along with the posterior mean and tuned values of parameters using conjugate RNNC, with $m = 10$. There is no significant difference in the estimation of parameters for all MCMC based models (NNCGP and collapsed RNNC) and the true values of the parameters are successfully included in the 95% marginal credible intervals. The introduction of latent interpolants may have caused a small overestimation of τ_2^2 for all models. Instead, the conjugate RNNC is underestimating the variance of the nugget for the second fidelity level. The parameter estimations can be improved with a semi-nested or nested structure between the observed locations of the fidelity levels, and it is also shown for the auto-regressive co-kriging model in Konomi and Karagiannis (2021). The conjugate RNNC model has similar performance on estimating the mean of the parameters compared to the NNCGP and RNNC models. However, it does not provide uncertainties regarding these estimations.

In Tables 3 and 4, we report standard performance measures (defined in Appendix E) for the sequential NNCGP, collapsed RNNC and conjugate RNNC with $m = 10$ number of neighbors. All performance measures indicate that the collapsed RNNC model has similar predictive ability with the sequential NNCGP model. The conjugate RNNC model produced RMSPE value that is 10% larger than other NNCGP and collapsed RNNC models, but it is still significantly smaller than the RMSPE values from single level NNGP and combined NNGP. The tables also show that the running time of the collapsed RNNC model is not different from the sequential NNCGP model, this is consistent with our previous

TABLE 3 Performance measures for the predictive ability of the sequential NNCGP model, collapsed RNNC model and conjugate RNNC model.

	Nested data-set		
	Sequential NNCGP	**Collapsed RNNC**	**Conjugate RNNC**
RMSPE	0.63	0.69	0.72
NSME	0.77	0.75	0.71
CRPS	0.45	0.44	0.41
CVG (95%)	0.91	0.88	0.93
ALCI (95%)	1.93	1.92	2.54
Time (h)	4.5	4.7	0.08

TABLE 4 Performance measures for the predictive ability of the Sequential NNCGP model, collapsed RNNC model and conjugate RNNC model, in non-nested input.

	Non-nested data-set		
	Sequential NNCGP	**Collapsed RNNC**	**Conjugate RNNC**
RMSPE	0.93	0.92	1.07
NSME	0.75	0.78	0.75
CRPS	0.61	0.61	0.65
CVG (95%)	0.93	0.97	0.95
ALCI (95%)	1.49	1.76	2.49
Time (h)	4.1	4.3	0.05

discussion since the two procedures have the same computational complexity. We observe that the conjugate RNNC model has extremely smaller running time compared to sequential NNCGP models, since the inference for conjugate RNNC model requires the same amount of running time as one iteration in collapsed RNNC model. It is worth pointing out that the cross validation process and tuning process in Algorithm 1 are independent from each other, which makes conjugate RNNC model benefit from parallel computation environments and greatly reduce computational time.

Figures 2 and 3 provide the non-nested synthetic observations and the prediction plots from sequential NNCGP, collapsed RNNC, conjugate RNNC, combined NNGP and single level NNGP models. We observed that for the testing regions the NNCGP models provides similar prediction surfaces and all NNCGP models has the better presentation of patterns in prediction surface comparing to single level NNGP model.

5.2 | Application to high-resolution infrared radiation sounder data

We model our data based on the two-fidelity level conjugate RNNC model and on the two-fidelity level sequential NNCGP model. Moreover, we provide comparisons with the single level NNGP model and combined NNGP model. We consider a linear model for the mean of the Gaussian processes, in $y_1(\cdot)$ and $\delta_2(\cdot)$, with linear basis function representation $\mathbf{h}(s_t)$ and coefficients $\boldsymbol{\beta}_t = \{\beta_{0,t}, \beta_{1,t}, \beta_{2,t}\}^T$. We consider the scalar discrepancy $\zeta(\mathbf{s})$ to be unknown constant and equal to γ. The number of nearest neighbors m is set to 10, and the spatial process $\mathbf{w}_t$ is considered to have a diagonal anisotropic exponential covariance function.

We assign independent normal distribution priors with zero mean and large variances for $\beta_{0,t}, \beta_{1,t}, \beta_{2,t}$ and γ. We assign independent uniform prior distributions $U(0, 1000)$ to the range correlation parameters $(\phi_{t,1}, \phi_{t,2})$ for $t = 1, 2$. Also, we assign independent $IG(2, 1)$ prior distributions for the variance parameters σ_t^2 and τ_t^2. For the Bayesian inference of the sequential NNCGP, we run the MCMC sampler with of 35,000 iterations where the first 5000 iterations are discarded as a burn-in. For the Bayesian inference of conjugate RNNC, we consider using posterior means as the estimated values for

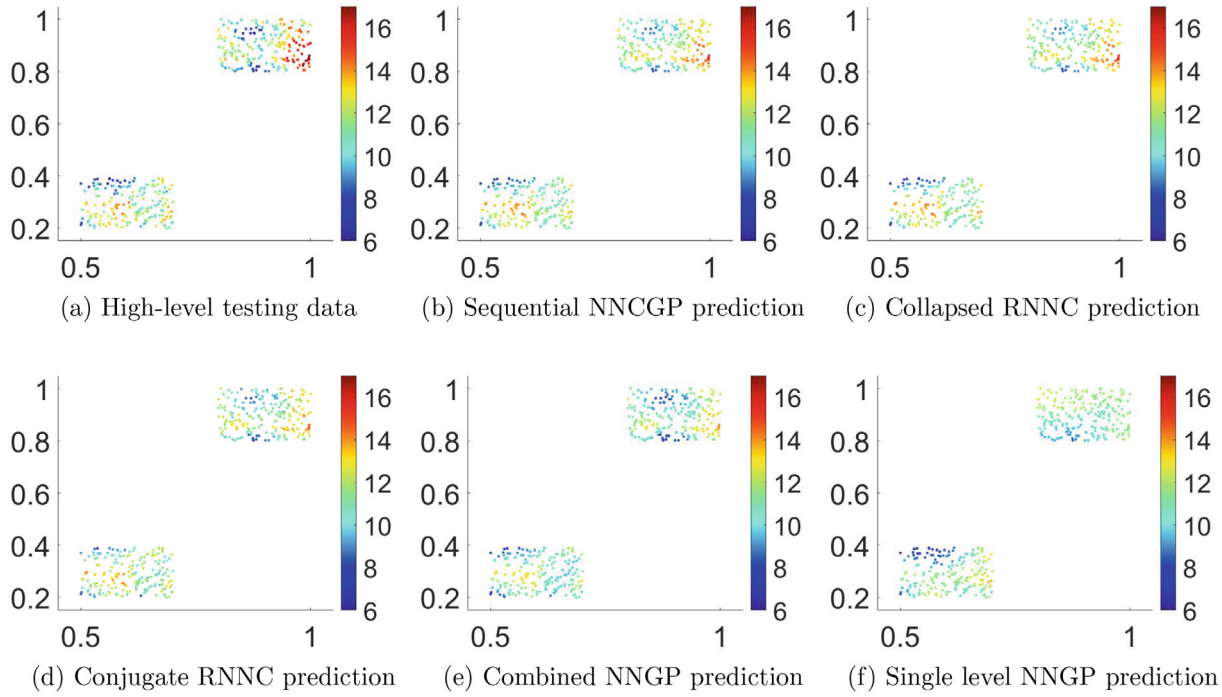


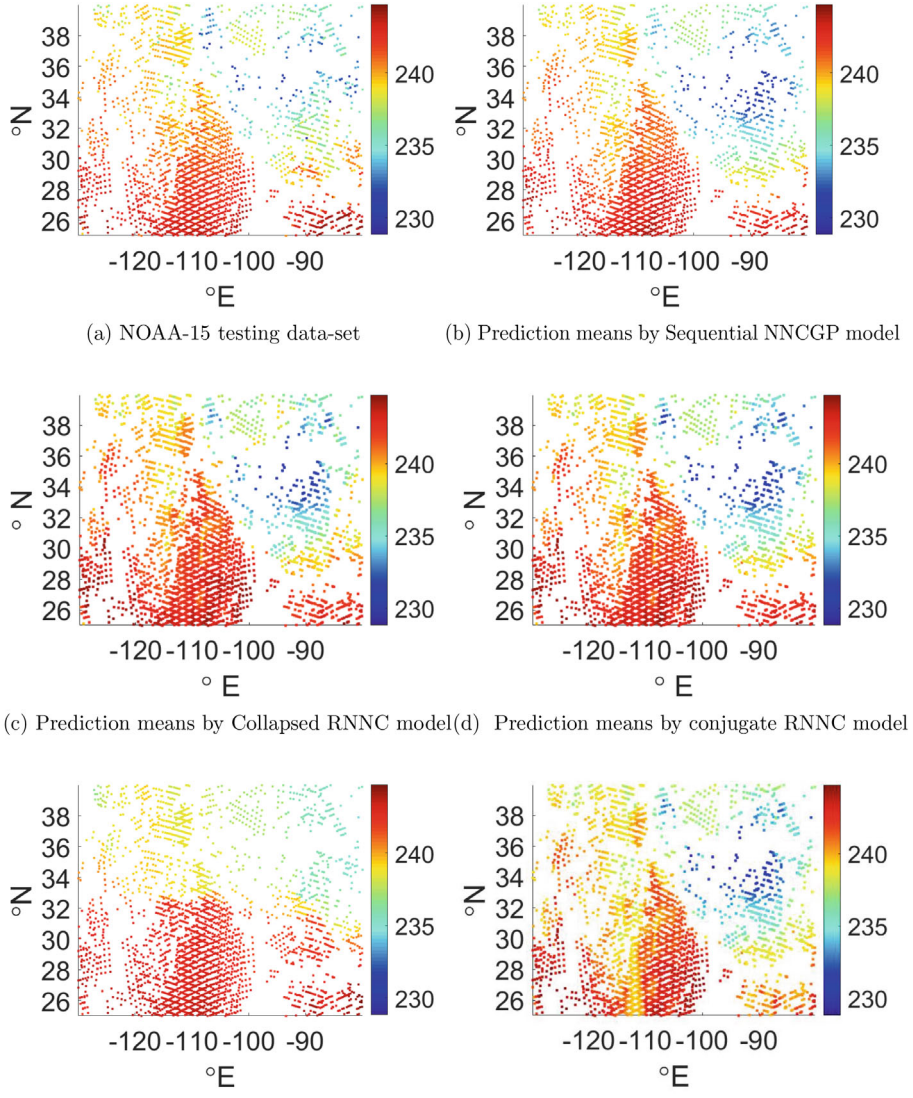
FIGURE 3 Non-nested input observations with two fidelity level structure. Original testing data (a) along with predictions of the high fidelity level data-set by (b) sequential NNCGP, (c) collapsed RNNC, (d) conjugate RNNC, (e) combined NNCGP and (f) single level NNCGP. (a) High-level testing data. (b) Sequential NNCGP prediction. (c) Collapsed RNNC prediction. (d) Conjugate RNNC prediction. (e) Combined NNGP prediction. (f) Single level NNGP prediction.

parameters $\beta_{0,t}, \beta_{1,t}, \beta_{2,t}, \sigma_t^2$ and γ; we also use posterior means as the imputation values for latent process $\tilde{\mathbf{y}}_t$ and for the prediction values of $z(s_p)$ at location $s_p \notin \tilde{\mathbf{S}}_t$.

The prediction performance metrics of the four different methods are given in Table 5. Compared to the single level NNGP model and combined NNGP model, the sequential NNCGP model and conjugate RNNC model produced a 20%–30% smaller RMSPE and their NSME is closer to 1. The sequential NNCGP model and collapsed RNNC model also produced larger CVG and smaller ALCI than the single level NNGP model and combined NNGP model. The result suggests that the NNCGP and RNNC models have a substantial improvement in terms of predictive accuracy in real data analysis too. In the prediction plots (Figure 4) of the testing data of NOAA-15, we observe that RNNC models are more capable of capturing the pattern of the testing data than single level NNGP model and combined NNGP model. This is reasonable because the observations from NOAA-14 have provided information of the testing region, and comparing to combined NNGP model, the NNCGP and RNNC models are capable of modeling the discrepancy of observations from different satellites. In the non-nested structure, the computational complexity of the single level NNGP model is $\mathcal{O}(n_2 m^3)$

TABLE 5 Performance measures for the predictive ability of sequential NNCGP, single level NNGP, combined NNGP, collapsed RNNC and conjugate RNNC models in NOAA 14 and NOAA 15 HIRS instrument data analysis.

	Model				
	Sequential NNCGP	**Single level NNGP**	**Combined NNGP**	**Collapsed RNNC**	**Conjugate RNNC**
RMSPE	1.20	1.82	1.68	1.21	1.36
NSME	0.84	0.55	0.67	0.85	0.82
CRPS	0.70	1.65	0.93	0.68	0.75
CVG (95%)	0.93	0.84	0.92	0.94	0.94
ALCI (95%)	3.09	4.21	5.79	3.16	4.39
Time (h)	38	20	32	40	0.3



(a) NOAA-15 testing data-set (b) Prediction means by Sequential NNCGP model

(c) Prediction means by Collapsed RNNC model (d) Prediction means by conjugate RNNC model

(e) Prediction means by single level NNGP model (f) Prediction means by combined NNGP model

FIGURE 4 Predictions of NOAA-15 brightness temperatures (K) testing data-set by (b) sequential NNCGP, (c) collapsed RNNC, (d) conjugate RNNC, (e) single level NNGP and (f) combined NNGP models. (a) NOAA-15 testing data-set. (b) Prediction means by sequential NNCGP model. (c) Prediction means by collapsed RNNC model. (d) Prediction means by conjugate RNNC model. (e) Prediction means by single level NNGP model. (f) Prediction means by combined NNGP model.

and that of NNCGP model is $\mathcal{O}((n_1 + n_2)m^3)$, for an MCMC iteration. However, the whole computational complexity of the conjugate RNNC model is $\mathcal{O}((n_1 + n_2)m^3)$ with parallel computational environment, which makes it remarkably computationally efficient without losing significant prediction accuracy. This is consistent with the running times of the models shown in Table 5.

We apply the MCMC free conjugate RNNC model for gap-filling predictions based upon a discrete global grid. We chose to use 1° latitude by 1.25° longitude ($1° \times 1.25°$) pixels as grids with global spatial coverage from $-70°$ to 70°N. By applying the NNCGP model, we predict gridded NOAA-15 brightness temperature data on the center of the grids, based on the NOAA-14 and NOAA-15 swath-based spatial support. The prediction plot (Figure 5) illustrates the ability of the MCMC free conjugate RNNC model to handle large irregularly spaced data sets and produce a gap-filled composite gridded dataset. The resulting global image of the brightness temperature is practically the same as the sequential NNCGP.

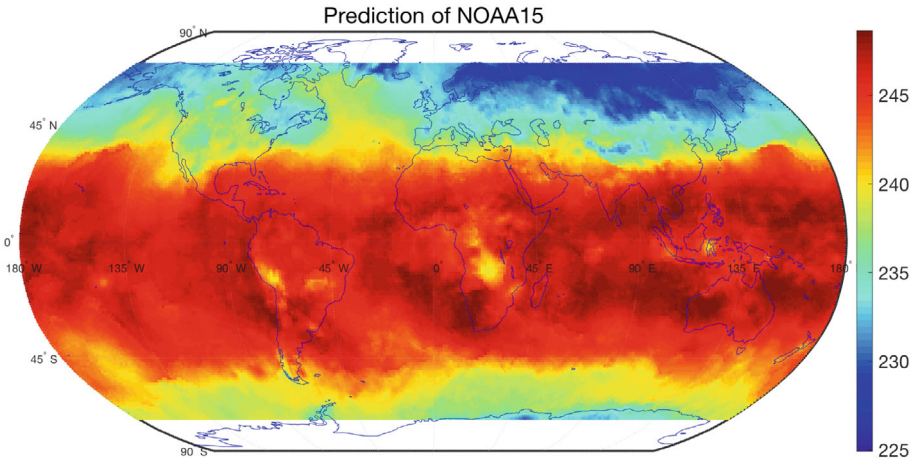


FIGURE 5 The global prediction brightness temperature values of NOAA 15 using the MCMC free conjugate model.

6 | SUMMARY AND CONCLUSIONS

We have proposed a new computationally efficient co-kriging method, the recursive nearest neighbor Autoregressive Co-Kriging (RNNC) model, for the analysis of large and multi-fidelity spatial data sets. In particular, we proposed two computationally efficient inferential procedures: (a) the collapsed RNNC, and (b) the conjugate RNNC. Regarding the collapsed RNNC, we integrate out the latent variables of the RNNC model which enables the factorisation of the likelihood into terms involving smaller and sparse covariance matrices within each level. Then, a prediction focused approximation is applied to the aforesaid model to further speed up the computation. The cross-validation using grid search on a two or three dimensional space is a computationally feasible method to estimate the hyperparameters. Regarding the proposed conjugate RNNC, it is MCMC free and at most computationally linear in the total number of all spatial locations of all fidelity levels. We compared the proposed collapsed RNNC and conjugate RNNC with NNCGP in a simulation study and a real data application of intersatellite calibration. We observed that similar to NNCGP, the collapsed and conjugate RNNC were also able to improve the accuracy of the prediction for the HIRS brightness temperatures from the NOAA-15 polar-orbiting satellite by incorporating information from an older version of the same HIRS sensor on board the polar orbiting satellite NOAA-14. The RNNC can be viewed as a modularization approach to NNCGP model in Bayesian statistics Bayarri et al. (2009) where the analysis is done in steps rather than jointly.

The proposed procedures can be used for a variety of large multi-fidelity data sets in remote sensing with overlapping areas of observed locations. A natural extension of our model can be done based on a recently proposed sparse plus Low-rank Gaussian Process (SPLGP) (Shirota et al., 2023) who used a combination of Gaussian predictive process and NNGP in an MCMC-free framework. A natural choice for introducing the non-stationarity in the conjugate RNNC is to use non-dynamic partition methods such as (Matthew J. Heaton et al., 2017; Konomi et al., 2019). Moreover, we can use more complex Vechias approximations (Guinness, 2018; Katzfuss et al., 2020; Stein et al., 2004; Vecchia, 1988), similar to the NNGP, where the ordering of the data is more complicated but results in a better approximation. These Vechias approximation techniques of ordering can be applied naturally in the proposed RNNC model, however, they are out of the scope of this paper and will be investigated in future work. Next steps will include extending the proposed method in the multivariate setting by using ideas from parallel partial autoregressive co-kriging (Ma et al., 2022) and NNGP spatial factor models (Taylor-Rodriguez et al., 2018). Spherical covariance function can be used for global data set analysis (Guinness & Fuentes, 2016), however their extension to anisotropic representation is not straightforward. Still, work needs to be done in developing new strategies for tuning the hyperparameters in more complex covariance functions with multiple parameters within a fidelity level.

ACKNOWLEDGMENTS

The research of Konomi and Kang was supported in part by National Science Foundation grant NSF DMS-2053668 and the Taft Research Center at the University of Cincinnati. Kang was also supported in part by Simons Foundation's Collaboration Award (#317298 and #712755).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Bledar A. Konomi https://orcid.org/0000-0003-2020-8493

REFERENCES

Abdulah, S., Li, Y., Cao, J., Ltaief, H., Keyes, D. E., Genton, M. G., & Sun, Y. (2023). Large-scale environmental data science with exageostatr. *Environmetrics*, *34*, e2770 https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2770

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.

Bayarri, M. J., Berger, J. O., & Liu, F. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, *4*, 119–150. https://doi.org/10.1214/09-BA404

Chander, G., Hewison, T., Fox, N., Wu, X., Xiong, X., & Blackwell, W. (2013). Overview of intercalibration of satellite instruments. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(3), 1056–1080.

Cheng, S., Konomi, B. A., Matthews, J. L., Karagiannis, G., & Kang, E. L. (2021). Hierarchical bayesian nearest neighbor co-kriging gaussian process models; an application to intersatellite calibration. *Spatial Statistics*, *44*, 100516 https://www.sciencedirect.com/science/article/pii/S2211675321000269

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.

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

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., & Banerjee, S. (2019). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *Journal of Computational and Graphical Statistics*, *28*, 401–414. https://doi.org/10.1080/10618600.2018.1537924

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., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.

Goldberg, M. (2011). The global space-based inter-calibration systems. *Bulletin of the American Meteorological Society*, *92*, 467–475.

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

Guinness, J. (2018). Permutation and grouping methods for sharpening Gaussian process approximations. *Technometrics*, *60*, 415–429. https://doi.org/10.1080/00401706.2018.1437476

Guinness, J., & Fuentes, M. (2016). Isotropic covariance functions on spheres: Some properties and modeling considerations. *Journal of Multivariate Analysis*, *143*, 143–152. https://www.sciencedirect.com/science/article/pii/S0047259X15002109

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

Jackson, D., Wylie, D., & Bates, J. (2003). The hirs pathfinder radiance data set (1979–2001). Proc. of the 12th conference on satellite meteorology and oceanography, Long Beach, CA, USA, 10-13 February 2003, vol. 5805 of LNCS. Springer.

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

Katzfuss, M., & Guinness, J. (2021). A general framework for Vecchia approximations of Gaussian processes. *Statistical Science*, *36*, 124–141. https://doi.org/10.1214/19-STS755

Katzfuss, M., Guinness, J., Gong, W., & Zilber, D. (2020). Vecchia approximations of gaussian-process predictions. *Journal of Agricultural, Biological and Environmental Statistics*, *25*, 383–414.

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

Kennedy, M. C., & O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, *87*, 1–13.

Konomi, B. A., Hanandeh, A. A., Ma, P., & Kang, E. L. (2019). Computationally efficient nonstationary nearest-neighbor Gaussian process models using data-driven techniques. *Environmetrics*, *30*, e2571 https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2571

Konomi, B. A., Kang, E. L., Almomani, A., & Hobbs, J. (2023). Bayesian latent variable co-kriging model in remote sensing for quality flagged observations. *Journal of Agricultural, Biological and Environmental Statistics*, *4*, 119–150. https://doi.org/10.1007/s13253-023-00530-9

Konomi, B. A., & Karagiannis, G. (2021). Bayesian analysis of multifidelity computer models with local features and nonnested experimental designs: Application to the wrf model. *Technometrics*, *63*, 510–522. https://doi.org/10.1080/00401706.2020.1855253

Le Gratiet, L. (2013). Bayesian analysis of hierarchical multifidelity codes. *SIAM/ASA Journal on Uncertainty Quantification*, *1*, 244–269.

Le Gratiet, L., & Garnier, J. (2014). Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, *4*, 365–386.

Lindgren, F., Rue, H., & Lindström, J. (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*, 423–498.

Liu, J. S., Wong, W. H., & Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, *81*, 27–40. https://doi.org/10.1093/biomet/81.1.27

Ma, P., & Kang, E. L. (2020). A fused Gaussian process model for very large spatial data. *Journal of Computational and Graphical Statistics*, *29*, 479–489.

Ma, P., Karagiannis, G., Konomi, B. A., Asher, T. G., Toro, G. R., & Cox, A. T. (2022). Multifidelity computer model emulation with high-dimensional output: An application to storm surge. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 71*, 861–883. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12558

Michele Peruzzi, S. B., & Finley, A. O. (2022). Highly scalable bayesian geostatistical modeling via meshed Gaussian processes on partitioned domains. *Journal of the American Statistical Association*, *117*, 969–982.

National Research Council. (2004). *Climate data records from environmental satellites: Interim report*. The National Academies Press https://www.nap.edu/catalog/10944/climate-data-records-from-environmental-satellites-interim-report

Nguyen, H., Cressie, N., & Braverman, A. (2012). Spatial statistical data fusion for remote sensing applications. *Journal of the American Statistical Association*, *107*, 1004–1018.

Nguyen, H., Cressie, N., & Braverman, A. (2017). Multivariate spatial data fusion for very large remote sensing datasets. *Remote Sensing*, *9*, 142.

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.

O'Hagan, A. (1998). A Markov property for covariance structures. *Statistics Research Report*, *98*, 510.

Qian, Z., Seepersad, C. C., Joseph, V. R., Allen, J. K., & Jeff Wu, C. F. (2005). Building surrogate models based on detailed and approximate simulations. *Journal of Mechanical Design*, *128*, 668–677. https://doi.org/10.1115/1.2179459

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

Shirota, S., Finley, A. O., Cook, B. D., & Banerjee, S. (2023). Conjugate sparse plus low rank models for efficient bayesian interpolation of large spatial data. *Environmetrics*, *34*, e2748 https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2748

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

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.

Taylor-Rodriguez, D., Finley, A. O., Datta, A., Babcock, C., Andersen, H.-E., Cook, B. D., Morton, D. C., & Banerjee, S. (2018). Spatial factor models for high-dimensional and large spatial data: An application in forest variable mapping. arXiv preprint arXiv:1801.02078.

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

Xiong, X., Cao, C., & Chander, G. (2010). An overview of sensor calibration inter-comparison and applications. *Frontiers of Earth Science in China*, *4*, 237–252.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Cheng, S., Konomi, B. A., Karagiannis, G., & Kang, E. L. (2024). Recursive nearest neighbor co-kriging models for big multi-fidelity spatial data sets. *Environmetrics*, *35*(4), e2844. https://doi.org/10.1002/env.2844

APPENDIX A. NNGP SPECIFICATIONS

The posterior distribution of

$$\begin{aligned}\tilde{p}(\mathbf{w}_t|\cdot) &\propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n_t}\left\{w_t(\mathbf{s}_{t,i}) - \mathbf{B}_{t,\mathbf{s}_{t,i}}\mathbf{w}_{t,N_t(\mathbf{s}_{t,i})}\right\}^T F^{-1}_{t,\mathbf{s}_{t,i}}\left\{w_t(\mathbf{s}_{t,i}) - \mathbf{B}_{t,\mathbf{s}_{t,i}}\mathbf{w}_{t,N_t(\mathbf{s}_{t,i})}\right\}\right] \\ &= \exp\left(-\frac{1}{2}\mathbf{w}_t^T\mathbf{B}_t^T\mathbf{F}_t^{-1}\mathbf{B}_t\mathbf{w}_t\right),\end{aligned} \tag{A1}$$

where $\mathbf{F}_t = \text{diag}(F_{t,\mathbf{s}_{t,1}}, F_{t,\mathbf{s}_{t,2}}, \dots, F_{t,\mathbf{s}_{t,n_t}})$, $\mathbf{B}_t = \left(\mathbf{B}_{t,1}^T, \mathbf{B}_{t,2}^T, \dots, \mathbf{B}_{t,n_t}^T\right)^T$, and for each element in $\mathbf{B}_t$, we have $\mathbf{B}_{t,i} = \left(\mathbf{B}_{t,s_{t,i},1}^T, \mathbf{B}_{t,s_{t,i},2}^T, \dots, \mathbf{B}_{t,s_{t,i},n_t}^T\right)^T$ and

$$\mathbf{B}_{t,s_{t,i}j} = \begin{cases} 1, & \text{if } i = j, \\ -\mathbf{B}_{t,s_{t,i}}[,k], & \text{if } s_{t,j} \text{ is the } k^{th} \text{ element in } N_t(s_{t,i}), \\ 0, & \text{Others.} \end{cases} \tag{A2}$$

APPENDIX B. MEAN AND VARIANCE SPECIFICATIONS

The mean vector $\boldsymbol{\mu} = \left(\mu_1(\mathbf{s}_{1,1}), \dots, \mu_1(\mathbf{s}_{1,n_1}), \dots, \mu_T(\mathbf{s}_{T,n_T})\right)$ is

$$\begin{aligned} \mu_t(\mathbf{s}_{t,k}) = \mathbf{1}_{\{t>1\}}(t) \sum_{i=1}^{t-1} \left\{ \prod_{j=i}^{t-1} \zeta_j(\mathbf{s}_{t,k}) \right\} \left\{ \mathbf{h}_i^T(\mathbf{s}_{t,k})\boldsymbol{\beta}_i + \mathbf{1}_{\{\mathbf{s}_{t,k} \in \mathbf{S}_i\}}(\mathbf{s}_{t,k}) w_i(\mathbf{s}_{t,k}) \right\} \\ + \mathbf{h}_t^T(\mathbf{s}_{t,k})\boldsymbol{\beta}_t + w_t(\mathbf{s}_{t,k}), \end{aligned} \tag{B1}$$

for $t = 1, \dots T$, $i = 1, \dots, n_t$. $\mathbf{1}_{\{\cdot\}}(\cdot)$ is the indicator function, and covariance matrix $\boldsymbol{\Lambda}$ is a block matrix with blocks $\Lambda^{(1,1)}, \dots, \Lambda^{(1,T)}, \dots, \Lambda^{(T,T)}$, and the size of $\boldsymbol{\Lambda}$ is $\sum_{t=1}^T n_t \times \sum_{t=1}^T n_t$. The $\Lambda^{(t,t)}$ components are calculated as:

$$\begin{aligned} \Lambda_{k,l}^{(t,t)} = \text{cov}(z_t(\mathbf{s}_{t,k}), z_t(\mathbf{s}_{t,l})|\cdot) = \sum_{i=1}^{t-1} \mathbf{1}_{\{\mathbf{s}_{t,k}, \mathbf{s}_{t,l} \notin \mathbf{S}_i\}}(\mathbf{s}_{t,k}, \mathbf{s}_{t,l}) \left\{ \prod_{j=i}^{t-1} \zeta_j(\mathbf{s}_{t,k})^T \zeta_j(\mathbf{s}_{t,l}) \right\} C_i(\mathbf{s}_{t,k}, \mathbf{s}_{t,l}|\boldsymbol{\theta}_i) \\ + \mathbf{1}_{\mathbf{s}_{t,k}=\mathbf{s}_{t,l}}(\mathbf{s}_{t,k}, \mathbf{s}_{t,l})\tau_t^2, \end{aligned}$$

for t and $t' = 1, \dots, T$; $k = 1, \dots, n_t$; $l = 1, \dots, n_{t'}$, and

$$\begin{aligned} \Lambda_{k,l}^{(t,t\prime)} = \text{cov}(z_t(s_{t,k}), z_{t\prime}(\mathbf{s}_{t\prime,l})|\cdot) = \sum_{i=1}^{\min(t,t\prime)-1} \mathbf{1}_{\{\mathbf{s}_{t,k}, \mathbf{s}_{t\prime,l} \notin \mathbf{S}_i\}}(\mathbf{s}_{t,k}, \mathbf{s}_{t\prime,l}) \left\{ \prod_{j=i}^{\min(t,t\prime)-1} \zeta_j(\mathbf{s}_{t,k})^T \zeta_j(\mathbf{s}_{t\prime,l}) \right\} \\ \times C_i(\mathbf{s}_{t,k\mathbf{s}_{t\prime,l}}|\boldsymbol{\theta}_i) + \mathbf{1}_{\{\mathbf{s}_{t,k}, \mathbf{s}_{t\prime,l} \notin \mathbf{S}_{\min(t,t\prime)}\}}(\mathbf{s}_{t,k}, \mathbf{s}_{t\prime,l}) C_{\min(t,t\prime)}(\mathbf{s}_{t,k}, \mathbf{s}_{t\prime,l}|\boldsymbol{\theta}_{\min(t,t\prime)}), \end{aligned} \tag{B2}$$

for $t \neq t'$, $\Lambda^{(t,t\prime)}$.

APPENDIX C. GIBBS SAMPLER

$$\begin{aligned} \mathbf{V}_{\beta_t}^* &= (\mathbf{h}_t(\mathbf{S}_t)\tilde{\Lambda}_t(\mathbf{S}_t, \boldsymbol{\theta}_t)^{-1}\mathbf{h}_t^T(\mathbf{S}_t) + \mathbf{V}_{\beta_t}^{-1})^{-1}, \\ \boldsymbol{\mu}_{\beta_t}^* &= \mathbf{V}_{\beta_t}^{-1}\boldsymbol{\mu}_{\beta_t} + \mathbf{h}_t(\mathbf{S}_t)\tilde{\Lambda}_t(\mathbf{S}_t, \boldsymbol{\theta}_t)^{-1}(z_t(\mathbf{S}_t) - \zeta_{t-1}(\mathbf{S}_t)\hat{y}_{t-1}(\mathbf{S}_t)). \end{aligned} \tag{C1}$$

$$\begin{aligned} \mathbf{V}_{\gamma_t}^* &= \left[(\mathbf{g}_t^T(\mathbf{S}_{t+1})\hat{y}_t(\mathbf{S}_{t+1}))^T \tilde{\Lambda}_{t+1}(\mathbf{S}_{t+1}, \boldsymbol{\theta}_{t+1}, \tau_{t+1})^{-1}(\mathbf{g}_t^T(\mathbf{S}_{t+1})\hat{y}_t(\mathbf{S}_{t+1})) + \mathbf{V}_{\gamma_t}^{-1}\right]^{-1}, \\ \boldsymbol{\mu}_{\gamma_t}^* &= \mathbf{V}_{\gamma_t}^{-1}\boldsymbol{\mu}_{\gamma_t} + (\mathbf{g}_t^T(\mathbf{S}_{t+1})\hat{y}_t(\mathbf{S}_{t+1}))^T \tilde{\Lambda}_{t+1}(\mathbf{S}_{t+1}, \boldsymbol{\theta}_{t+1}, \tau_{t+1})^{-1}(\mathbf{Z}_{t+1} - \mathbf{h}_t^T(\mathbf{S}_{t+1})\boldsymbol{\beta}_{t+1}). \end{aligned} \tag{C2}$$

APPENDIX D. CONJUGATE CONDITIONAL PROBABILITIES

we derive the posterior distribution as

$$\begin{aligned} p(\boldsymbol{\beta}_t, \boldsymbol{\gamma}_{t-1}, \sigma_t^2|\mathbf{Z}_t, \hat{y}_{t-1}(\mathbf{S}_t)) &\propto IG(\sigma_t^2|a_t, b_t) N(\boldsymbol{\beta}_t|\boldsymbol{\mu}_{\beta_t}, \sigma_t^2\mathbf{V}_{\beta_t}) N(\boldsymbol{\gamma}_{t-1}|\boldsymbol{\mu}_{\gamma_{t-1}}, \sigma_t^2\mathbf{V}_{\gamma_{t-1}}) \\ &\times N(\mathbf{Z}_t|\zeta_{t-1}(\mathbf{S}_t) \circ \hat{y}_{t-1}(\mathbf{S}_t) + \mathbf{h}_t^T\boldsymbol{\beta}_t, \sigma_t^2\tilde{\boldsymbol{\Sigma}}_t) \\ &\propto p(\sigma_t^2|\mathbf{Z}_t, \hat{y}_{t-1}(\mathbf{S}_t)) p(\boldsymbol{\beta}_t|\sigma_t^2, \mathbf{Z}_t, \hat{y}_{t-1}(\mathbf{S}_t)) p(\boldsymbol{\gamma}_{t-1}|\boldsymbol{\beta}_t, \sigma_t^2, \mathbf{Z}_t, \hat{y}_{t-1}(\mathbf{S}_t)) \end{aligned}$$

$$\propto (\sigma_t^2)^{a_t+0.5n_t}\exp\left(-\frac{1}{2\sigma_t^2}(\boldsymbol{\beta}_t-\boldsymbol{\mu}_{\beta_t})^T\mathbf{V}_{\beta_t}^{-1}(\boldsymbol{\beta}_t-\boldsymbol{\mu}_{\beta_t})\right)$$
$$\times\exp\left(-\frac{1}{2\sigma_t^2}(\boldsymbol{\gamma}_{t-1}-\boldsymbol{\mu}_{\gamma_{t-1}})^T\mathbf{V}_{\gamma_{t-1}}^{-1}(\boldsymbol{\gamma}_{t-1}-\boldsymbol{\mu}_{\gamma_{t-1}})\right)$$
$$\times\exp\left(-\frac{1}{2\sigma_t^2}(\mathbf{Z}_t-\mathbf{g}^T(\mathbf{S}_t)\boldsymbol{\gamma}_{t-1}\hat{y}_{t-1}(\mathbf{S}_t)-\mathbf{h}_t^T(\mathbf{S}_t)\boldsymbol{\beta}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}(\mathbf{Z}_t-\mathbf{g}^T(\mathbf{S}_t)\boldsymbol{\gamma}_{t-1}\hat{y}_{t-1}(\mathbf{S}_t)-\mathbf{h}_t^T(\mathbf{S}_t)\boldsymbol{\beta}_t)\right).$$

The full conditional density function of $\boldsymbol{\gamma}_{t-1}$ is

$$\begin{aligned}p(\boldsymbol{\gamma}_{t-1}|\boldsymbol{\beta}_t,\sigma_t^2,\mathbf{Z}_t,\hat{y}_{t-1}(\mathbf{S}_t))&\propto\exp\Bigg(-\frac{1}{2\sigma_t^2}\Big[\mathbf{g}(\mathbf{S}_t)\boldsymbol{\gamma}_{t-1}^T\hat{y}_{t-1}(\mathbf{S}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{g}^T(\mathbf{S}_t)\boldsymbol{\gamma}_{t-1}\hat{y}_{t-1}(\mathbf{S}_t)\\&\quad-2(\mathbf{Z}_t-\mathbf{h}_t^T(\mathbf{S}_t)\boldsymbol{\beta}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{g}^T(\mathbf{S}_t)\boldsymbol{\gamma}_{t-1}\hat{y}_{t-1}(\mathbf{S}_t)\Big]\Bigg)\\&\propto N(\boldsymbol{\gamma}_{t-1}|\tilde{\mathbf{V}}_{\gamma_{t-1}}\tilde{\boldsymbol{\mu}}_{\gamma_{t-1}},\sigma^2\tilde{\mathbf{V}}_{\gamma_{t-1}}),\\\tilde{\boldsymbol{\mu}}_{\gamma_{t-1}}&=\mathbf{V}_{\gamma_{t-1}}^{-1}\boldsymbol{\mu}_{\gamma_{t-1}}+\mathbf{g}(\mathbf{S}_t)\hat{y}_{t-1}(\mathbf{S}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}(\mathbf{Z}_t-\mathbf{h}_t^T(\mathbf{S}_t)\boldsymbol{\beta}_t),\\\tilde{\mathbf{V}}_{\gamma_{t-1}}&=\left(\mathbf{V}_{\gamma_{t-1}}^{-1}+\mathbf{g}(\mathbf{S}_t)\hat{y}_{t-1}(\mathbf{S}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\hat{y}_{t-1}(\mathbf{S}_t)\mathbf{g}^T(\mathbf{S}_t)\right)^{-1}.\end{aligned}\tag{D1}$$

After integrate $\boldsymbol{\gamma}_{t-1}$ out, the conditional posterior density function of $\boldsymbol{\beta}_t$ is

$$\begin{aligned}p(\boldsymbol{\beta}_t|\sigma_t^2,\mathbf{Z}_t,\hat{y}_{t-1}(\mathbf{S}_t))&\propto\exp\left(-\frac{1}{2\sigma_t^2}[(\mathbf{h}_t^T(\mathbf{S}_t)\boldsymbol{\beta}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}(\mathbf{h}_t^T(\mathbf{S}_t)\boldsymbol{\beta}_t)-2\mathbf{Z}_t^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{h}_t^T(\mathbf{S}_t)\boldsymbol{\beta}_t]\right)\\&\quad\times\exp\left(-\frac{1}{2\sigma_t^2}(\boldsymbol{\beta}_t-\boldsymbol{\mu}_{\beta_t})^T\mathbf{V}_{\beta_t}^{-1}(\boldsymbol{\beta}_t-\boldsymbol{\mu}_{\beta_t})\right)\exp\left(\frac{1}{2\sigma_t^2}\tilde{\boldsymbol{\mu}}_{\gamma_{t-1}}^T\tilde{\mathbf{V}}_{\gamma_{t-1}}\tilde{\boldsymbol{\mu}}_{\gamma_{t-1}}\right),\\&\propto N(\boldsymbol{\beta}_t|\tilde{\mathbf{V}}_{\beta_t}\tilde{\boldsymbol{\mu}}_{\beta_t},\sigma_t^2\tilde{\mathbf{V}}_{\beta_t}),\\\tilde{\boldsymbol{\mu}}_{\beta_t}&=\mathbf{V}_{\beta_t}^{-1}\boldsymbol{\mu}_{\beta_t}+\mathbf{h}_t(\mathbf{S}_t)\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{Z}_t-(\mathbf{g}(\mathbf{S}_t)y_{t-1}(\mathbf{S}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{h}_t^T(\mathbf{S}_t))^T\tilde{\mathbf{V}}_{\gamma_{t-1}}(\mathbf{V}_{\gamma_{t-1}}^{-1}\boldsymbol{\mu}_{\gamma_{t-1}}\\&\quad+\mathbf{g}(\mathbf{S}_t)\hat{y}_{t-1}(\mathbf{S}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{Z}_t),\\\tilde{\mathbf{V}}_{\beta_t}&=\left(\mathbf{V}_{\beta_t}^{-1}+\mathbf{h}(\mathbf{S}_t)\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{h}(\mathbf{S}_t)^T-(\mathbf{g}(\mathbf{S}_t)\hat{y}_{t-1}(\mathbf{S}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{h}_t^T(\mathbf{S}_t))^T\tilde{\mathbf{V}}_{\gamma_{t-1}}(\mathbf{g}(\mathbf{S}_t)y_{t-1}(\mathbf{S}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{h}_t^T(\mathbf{S}_t))\right)^{-1}.\end{aligned}\tag{D2}$$

The marginalized posterior density function of σ_t is

$$\begin{aligned}&p(\sigma_t^2|\mathbf{Z}_t,\hat{y}_{t-1}(\mathbf{S}_t))\\&\quad\propto\sigma_t^{-a_t-0.5n_t}\exp\Bigg(-\frac{1}{2\sigma_t^2}\Big[2b_t+\mathbf{Z}_t^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{Z}_t+\boldsymbol{\mu}_{\beta_t}^T\mathbf{V}_{\beta_t}^{-1}\boldsymbol{\mu}_{\beta_t}+\boldsymbol{\mu}_{\gamma_{t-1}}^T\mathbf{V}_{\gamma_{t-1}}^{-1}\boldsymbol{\mu}_{\gamma_{t-1}}-\tilde{\boldsymbol{\mu}}_{\beta_t}^T\tilde{\mathbf{V}}_{\beta_t}\tilde{\boldsymbol{\mu}}_{\beta_t}\\&\qquad-(\mathbf{V}_{\gamma_{t-1}}^{-1}\boldsymbol{\mu}_{\gamma_{t-1}}+\mathbf{g}(\mathbf{S}_t)\hat{y}_{t-1}(\mathbf{S}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{Z}_t)^T\tilde{\mathbf{V}}_{\gamma_{t-1}}(\mathbf{V}_{\gamma_{t-1}}^{-1}\boldsymbol{\mu}_{\gamma_{t-1}}+\mathbf{g}(\mathbf{S}_t)\hat{y}_{t-1}(\mathbf{S}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{Z}_t)\Big]\Bigg),\\&\sigma_t^2|\mathbf{Z}_t,\hat{y}_{t-1}(\mathbf{S}_t)\sim IG(\sigma_t^2|a_t^*,b_t^*),\\&a_t^*=a_t+n_t/2,\\&b_t^*=b_t+0.5\Big(\mathbf{Z}_t^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{Z}_t+\boldsymbol{\mu}_{\beta_t}^T\mathbf{V}_{\beta_t}^{-1}\boldsymbol{\mu}_{\beta_t}+\boldsymbol{\mu}_{\gamma_{t-1}}^T\mathbf{V}_{\gamma_{t-1}}^{-1}\boldsymbol{\mu}_{\gamma_{t-1}}-\tilde{\boldsymbol{\mu}}_{\beta_t}^T\tilde{\mathbf{V}}_{\beta_t}\tilde{\boldsymbol{\mu}}_{\beta_t}\\&\qquad-(\mathbf{V}_{\gamma_{t-1}}^{-1}\boldsymbol{\mu}_{\gamma_{t-1}}+\mathbf{g}(\mathbf{S}_t)\hat{y}_{t-1}(\mathbf{S}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{Z}_t)^T\tilde{\mathbf{V}}_{\gamma_{t-1}}(\mathbf{V}_{\gamma_{t-1}}^{-1}\boldsymbol{\mu}_{\gamma_{t-1}}+\mathbf{g}(\mathbf{S}_t)\hat{y}_{t-1}(\mathbf{S}_t)^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{Z}_t)\Big).\end{aligned}\tag{D3}$$

the conditional posterior density function of $\boldsymbol{\beta}_1$ is

$$\begin{aligned}\boldsymbol{\beta}_1|\sigma_1^2, \mathbf{Z}_1 &\sim N(\boldsymbol{\beta}_1|\tilde{\mathbf{V}}_{\beta_1}\tilde{\boldsymbol{\mu}}_{\beta_1}, \sigma_1^2\tilde{\mathbf{V}}_{\beta_1}),\\ \tilde{\boldsymbol{\mu}}_{\beta_1} &= \mathbf{V}_{\beta_1}^{-1}\boldsymbol{\mu}_{\beta_1} + \mathbf{h}_1(\mathbf{S}_1)\tilde{\boldsymbol{\Sigma}}_1^{-1}\mathbf{Z}_1,\\ \tilde{\mathbf{V}}_{\beta_1} &= \left(\mathbf{V}_{\beta_1}^{-1} + \mathbf{h}(\mathbf{S}_1)\tilde{\boldsymbol{\Sigma}}_1^{-1}\mathbf{h}(\mathbf{S}_1)^T\right)^{-1},\end{aligned} \tag{D4}$$

and the marginal posterior density function of σ_1^2 is

$$\begin{aligned}\sigma_1^2|\mathbf{Z}_1 &\sim IG(\sigma_1^2|a_1^*, b_1^*),\\ a_1^* &= a_1 + n_1/2,\\ b_1^* &= b_1 + 0.5\left(\mathbf{Z}_1^T\tilde{\boldsymbol{\Sigma}}_t^{-1}\mathbf{Z}_1 + \boldsymbol{\mu}_{\beta_1}^T\mathbf{V}_{\beta_1}^{-1}\boldsymbol{\mu}_{\beta_1} - \tilde{\boldsymbol{\mu}}_{\beta_1}^T\tilde{\mathbf{V}}_{\beta_1}\tilde{\boldsymbol{\mu}}_{\beta_1}\right).\end{aligned} \tag{D5}$$

APPENDIX E. PERFORMANCE METRICS

In the empirical comparisons, we used the following performance metrics:

1. Root mean square prediction error (RMSPE) is defined as

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{\text{pred}} - y_i^{\text{obs}})^2}$$

   where y^{obs} is the observed value in test data-set and y_i^{pred} is the predicted value from the model. It measures the accuracy of the prediction from model. Smaller values of RMSPE indicate more a accurate model.
2. Nash-Sutcliffe model efficiency coefficient (NSME) is defined as:

$$\text{NSME} = 1 - \frac{\sum_{i=1}^{n}(y_i^{\text{pred}} - y_i^{\text{obs}})^2}{\sum_{i=1}^{n}(y_i^{\text{obs}} - \bar{y}^{\text{obs}})^2}$$

   where y^{obs} is the observed value in test data-set and y_i^{pred} is the predicted value from the model. NSME gives the relative magnitude of the residual variance from data and the model variance. NSME values closer to 1 indicate that the model has a better predictive performance.
3. 95% CVG is the coverage probability of 95% equal tail prediction interval. 95% CVG values closer to 0.95 indicate better prediction performance for the model.
4. 95% ALCI is average length of 95% equal tail prediction interval. Smaller 95% ALCI values indicate better prediction performance for the model.
5. Continuous Ranked Probability Score (CRPS), which is defined as

$$\text{CRPS}(G, y) = \int (G - H_y)^2,$$

   where y is a scalar quantity that needs to forecast, and y admits an underlying distribution that is described by CDF F; G is a CDF that is chosen by the forecaster in order to predict F. H is a unit step function and $H_y(x)$ indicates a centered Heaviside function $H(x - y)$. CRPS is negatively oriented and the lower scores imply better performance.