Hierarchical Bayesian Nearest Neighbor Co-Kriging Gaussian Process Models; an Application to Intersatellite Calibration

Si Cheng[1], Bledar A. Konomi * [1], Jessica L. Matthews[2], Georgios Karagiannis[3], and Emily L. Kang[1]

[1]University of Cincinnati
[2]North Carolina State University, Cooperative Institute for Satellite Earth System Studies (CISESS)
[3]Durham University

May 11, 2021

Abstract

Recent advancements in remote sensing technology and the increasing size of satellite constellations allow for massive geophysical information to be gathered daily on a global scale by numerous platforms of different fidelity. The auto-regressive co-kriging model provides a suitable framework for the analysis of such data sets as it is able to account for cross-dependencies among different fidelity satellite outputs. However, its implementation in multifidelity large spatial data sets is practically infeasible because the computational complexity increases cubically with the total number of observations. In this paper, we propose a nearest neighbor co-kriging Gaussian process (GP) that couples the auto-regressive model and nearest neighbor GP by using augmentation ideas. Our model reduces the computational complexity to be linear with the total number of spatially observed locations. The spatial random effects of the nearest neighbor GP are augmented in a manner which allows the specification of semi-conjugate priors. This facilitates the design of an efficient MCMC sampler involving mostly direct sampling updates. The good predictive performance of the proposed method is demonstrated in a simulation study. We use the proposed method to analyze High-resolution Infrared Radiation Sounder data gathered from two NOAA polar orbiting satellites.

Keywords: Augmented hierarchically nested design, Autoregressive Co-kriging, Nearest neighbor Gaussian process, Remote sensing.

*Corresponding author: Bledar A. Konomi (alex.konomi@uc.edu)

1 Introduction

Due to the advancement of remote sensing technology and the growing size of satellite constellations, it has become increasingly common for geophysical information to be measured by numerous platforms at similar times and locations. Aging and exposure to the harsh environment of space results in sensor degradation over the satellite's lifetime causing a decrease on performance reliability. This results in inaccuracy of the data as a true measure for long term trend analysis (Goldberg, 2011). Generally, newer satellites with more advanced sensors provide information of higher fidelity than older models. Different platforms often collect large amounts of observations with varying fidelity for spatial areas that may or may not overlap or have the same spatial footprint. Over the years multiple methods have been developed with the overarching goal of enabling the intercomparison amongst satellite platforms including: use of ground-based observations and leveraging temporally stable targets (e.g. moon, desert sites, deep convective clouds) to assess satellite sensor performance and consistency (Chander et al., 2013; Xiong et al., 2010; National Research Council, 2004). But to date, these methods fail to account for differing fidelity levels between satellite platforms with statistical rigor.

As a specific example of how the remote sensing community currently manages this challenge, we consider the strategy applied to the high-resolution infrared radiation sounder (HIRS) which provides measurements from multiple satellite platforms. As described in Cao et al. (2004, 2005), the intersatellite calibration of HIRS sensors is based on calculating differences between all the near-nadir overlapping points from two satellites within a period of time. These differences are assessed in separated 10-degree brightness temperature bins as an aggregated mean, with no spatial or temporal dependencies. The current intersatellite calibration of the observations is simply a least square bias correction term based on a linear relationship of the differences in brightness temperature in the adjacent satellite (assumed to

have the higher fidelity). However, this naive approach ignores spatial dependency by assigning the same bias correction across the spatial domain, which can possibly yield misleading results. A single composite feature, which includes adequate information from multiple data sources in space, is preferred for statistical inference.

In geostatistics, co-kriging is a suitable framework for analysis of spatially correlated random processes (Davis and Greenes, 1983; Aboufirassi and Mariño, 1984; Ver Hoef and Cressie, 1993; Furrer and Genton, 2011; Genton and Kleiber, 2015). Complex cross-covariance functions can lead to infeasible computational complexity, even for moderate amounts of data. To address this issue, Kennedy and O'Hagan (2000) proposed an autoregressive co-kriging model which is simple, but yet flexible, to model complex dependency structures. The autoregressive co-kriging framework has gained popularity in computer experiments (Qian and Wu, 2008; Han et al., 2010; Le Gratiet, 2013; Koziel et al., 2014) due to its computational convenience. Its framework fits well with the multi-sensor geographical information system, since the hierarchy is established based on age and technology of the sensors. Most of the computational benefits of autoregressive co-kriging are lost when the multi-fidelity data are not observed in a hierarchically nested structure. Multi-sensor geographical information systems are usually observed irregularly in space and are hierarchically non-nested. Recently, Konomi and Karagiannis (2021) proposed a Bayesian augmented hierarchical co-kriging procedure which makes the analysis of partially-nested and/or non-nested structures possible with feasible computational cost by splitting the augmented likelihood into conditionally independent parts. Despite this simplification, the method cannot be applied directly to large data sets. Each conditional component of the likelihood requires evaluation of the determinant as well as inversion of a large co-variance matrix.

Recently, statistical methods for large spatial data sets have received much attention. Many of the most popular techniques rely on low-rank approximation (Banerjee et al., 2008; Cressie and Johannesson, 2008), approximate likelihood methods (Stein et al., 2004; Gramacy

and Apley, 2015), covariance tapering methods (Furrer et al., 2006; Kaufman et al., 2008; Du et al., 2009), sparse structures (Lindgren et al., 2011; Nychka et al., 2015; Datta et al., 2016), multiple-scale approximation (Sang and Huang, 2012; Katzfuss, 2016), and lower dimensional conditional distributions (Vecchia, 1988; Stein et al., 2004; Datta et al., 2016; Katzfuss and Guinness, 2021). A number of these methods have been generalized to handle large data from multiple sources. For example, Nguyen et al. (2012, 2017) have proposed data fusion techniques based on fixed ranked kriging (Cressie and 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, low rank kriging techniques have difficulty accounting for small scale variation (Stein, 2014). More recently, Taylor-Rodriguez et al. (2018) embedded the nearest-neighbor Gaussian process (NNGP) into a spatial factor model and used NNGP to model the resulting independent GP processes. This method assumes that data sets from different sources follow an overlapping structure, limiting its use for real applications.

In this paper, we propose a new computationally efficient autoregressive co-kriging method based on the nearest neighbor Gaussian process (NNGP), which is called the nearest neighbor Co-kriging Gaussian process (NNCGP). The proposed method is applicable to large non-nested and irregular spatial data sets from different platforms having varying quality. NNCGP utilizes an approximate imputation procedure based on a nested reference set to address large data sets of non-nested observations. This formulation allows the evaluation of the likelihood and predictions with low computational cost, as well as allows the specification of conditional conjugate priors. Compared to the aforementioned models, it exhibits both computational efficiency and flexibility. This method enables the analysis of high-resolution infrared radiation sounder (HIRS) data sets gathered daily from two polar orbiting satellite series (POES) of the National Oceanic and Atmospheric Administration (NOAA). We show that the proposed method is both more accurate and computationally more efficient

4

for these type of data sets. Furthermore, the NNCGP model shows significant improvement in prediction accuracy over the existing NNGP approach, based on our simulation study and the analysis for the real data application.

The layout of the paper is as follows. In Section 2, we introduce a spatial representation of the autoregressive co-kriging model. In Section 3, we introduce our proposed NNCGP as an extension of the existing autoregressive co-kriging model. In Section 4, we design an MCMC approach tailored to the proposed NNCGP model that facilitates parametric and predictive inference. In Section 5, we investigate the performance of the proposed procedure on a toy example. In Section 6, we apply the proposed method to HIRS data sets from two satellites, NOAA-14 and NOAA-15. Finally, we summarize our findings in Section 7.

2   Spatial Co-kriging Gaussian Process

We consider T platforms which gather spatial observations with similar footprint. We assume that: (a) observations from different platforms are correlated, (b) platform t provides a more accurate representation of the ground truth than platform $t-1$, and (c) prior belief about observations from a platform can be modeled by a Gaussian process. We refer to the observations of platform t as the observations of fidelity level t. Let $y_t(s)$ denote the output at the spatial location s at fidelity level $t = 1, ..., T$. Here, the fidelity level index t runs from the least to most accurate platform. We consider that the observed output $z_t(s)$ at location s is contaminated by additive random noise $\epsilon_t(\boldsymbol{s}) \sim N(0, \tau_t)$ with unknown variance τ_t, and $y_t(\boldsymbol{s})$ depends on the fidelity level output $y_{t-1}(\boldsymbol{s})$ via an autoregressive co-kriging model. Specifically, we model the observation $z_t(\boldsymbol{s})$ as:

$$z_t(\boldsymbol{s}) = y_t(\boldsymbol{s}) + \epsilon_t(\boldsymbol{s})$$

$$y_t(\boldsymbol{s}) = \zeta_{t-1}(\boldsymbol{s})y_{t-1}(\boldsymbol{s}) + \delta_t(\boldsymbol{s}), \tag{2.1}$$

$$\delta_t(\boldsymbol{s}) = \mathbf{h}_t^T(\boldsymbol{s})\boldsymbol{\beta}_t + w_t(s),$$

for $t = 2, \ldots, T$, and $y_1(\boldsymbol{s}) = \mathbf{h}_1^T(\boldsymbol{s})\boldsymbol{\beta_1} + w_1(\boldsymbol{s})$. Here, $\zeta_{t-1}(\boldsymbol{s})$ and $\delta_t(\boldsymbol{s})$ represent the scale and additive discrepancies between the output of platforms with fidelity levels t and $t-1$, and $\epsilon(\boldsymbol{s})$ are uncorrelated pure error terms with variance τ_t^2. Moreover, $\mathbf{h}_t(\cdot)$ is a vector of basis functions and $\boldsymbol{\beta}_t$ is a vector of coefficients at fidelity level t. We model, a priori, $w_t(\boldsymbol{s})$ as Gaussian processes, mutually independent for different t; i.e. $w_t(\cdot) \sim GP(0, C_t(\cdot, \cdot, \boldsymbol{\theta}_t))$ where $C_t(\cdot, \cdot, \boldsymbol{\theta}_t)$ is a covariance function with parameters $\boldsymbol{\theta}_t$ at fidelity level t. This implies that $y_1, \delta_2, \ldots, \delta_T$ are a priory mutually independent Gaussian processes. The unknown scale discrepancy function $\zeta_{t-1}(\boldsymbol{s})$ is modeled as a basis expansion $\zeta_{t-1}(\boldsymbol{s}|\boldsymbol{\gamma}_{t-1}) = \mathbf{g}_{t-1}(\boldsymbol{s})^T\boldsymbol{\gamma}_{t-1}$ (usually with low degree), where $\mathbf{g}_t(\boldsymbol{s})$ is a vector of basis functions and $\{\boldsymbol{\gamma}_{t-1}\}$ is a vector of coefficients for the scale discrepancies, for $t = 2, \ldots, T$.

The statistical model in (2.1) is different from the co-kriging model of Kennedy and O'Hagan (2000), which was developed for the analysis of deterministic computer models, because it accounts for a nugget effect through $\epsilon(\boldsymbol{s})$. The introduction of a nugget effect can play an important role by accounting for measurement errors in spatial statistics as well as modeling the error in stochastic computer models (Baker et al., 2020). The benefits of considering a nugget effect in spatial data models has been previously noticed by Cressie (1993) and Stein (1999). Gramacy and Lee (2012) argued that the use of a nugget can also mitigate poor fitting when there is deviation from the GP model assumptions. Finally, Kennedy and O'Hagan (2000) originally used a constant scalar discrepancy based on stationarity arguments. We use a more general polynomial format for the scalar discrepancy (Qian et al., 2005) for model flexibility and to improve predictions when needed. To avoid identifiability issues, it is recommended to use low degree basis expansions for both the additive and scalar discrepancy.

For each level of fidelity, we choose a product exponential covariance function: $C_t(\boldsymbol{s}, \boldsymbol{s}'|\boldsymbol{\theta}_t) = \sigma_t^2 \exp\left(-\sum_{i=1}^d \frac{|s_i - s_i'|}{\phi_{t,i}}\right)$, where $\boldsymbol{\theta}_t = \{\sigma_t^2, \boldsymbol{\phi}_t\} = \{\sigma_t^2, \phi_{t,1}, \phi_{t,2}, \ldots, \phi_{t,d}\}$, σ_t^2 is the variance parameter and $\phi_{t,i}$ control the spatial dependence strength in $\mathbb{R}^d$ at fidelity level t and direction

i. The product covariance function is equivalent to a diagonal anisotropic covariance function. More intricate covariance functions, such as the stationary ones from the Matérn family (Cressie, 1993; Stein, 1999; Banerjee et al., 2014) or the non-stationary ones of (Paciorek and Schervish, 2006; Konomi et al., 2014) can also be used in this model.

Let's assume the system is observed at n_t locations at each fidelity level t. Let $\mathbf{S}_t = \{\boldsymbol{s}_{t,1}, \ldots, \boldsymbol{s}_{t,n_t}\}$ be the set of n_t observed locations and $\mathbf{Z}_t = z_t(\mathbf{S}_t) = \{z_t(\boldsymbol{s}_{t,1}), \ldots, z_t(\boldsymbol{s}_{t,n_t})\}$ represent the observed outputs at fidelity level t. The joint sampling distribution of the observations at all levels $\mathbf{Z}_{1:T} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_T\}$ is Gaussian, hence the likelihood $L(\mathbf{Z}_{1:T}|\boldsymbol{\theta}_{1:T}, \boldsymbol{\beta}_{1:T}, \boldsymbol{\gamma}_{1:T-1}, \tau_{1:T})$ is a multivariate Normal density function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Lambda}$ that cannot easily be computed. Specifically, if the data are observed in non-nested locations for each fidelity level, the calculation of the likelihood requires $\mathcal{O}((\sum_{t=1}^{T} n_t)^3)$ flops to invert the covariance matrix $\boldsymbol{\Lambda}$ and an additional $\mathcal{O}((\sum_{t=1}^{T} n_t)^2)$ memory to store it as explained in Konomi and Karagiannis (2021). Thus the likelihood evaluation is computationally costly, if not practically impossible, when n_t is large. For instance, in our application, for each satellite we have $\sim 10^5$ observations at non-nested locations making the practical implementation impossible.

3 Nearest Neighbor Co-kriging Gaussian Process

To deal with the computational complexity of the co-kriging model, we propose to apply a set of independent nearest-neighbor Gaussian process (NNGP) priors (Datta et al., 2016) at the spatial process of each level of fidelity. NNGP is a fully dimensional GP with sparse representation in the precision matrix of the spatial process. Let $\mathbf{w}_t = w_t(\mathbf{S}_t) = \{w_t(\boldsymbol{s}_{t,1}), \ldots, w_t(\boldsymbol{s}_{t,n_t})\}$ denote the vector of the spatial process over the observed locations $\mathbf{S}_t$ at fidelity level t. Based on the independent assumptions in (2.1), the joint density of

$\mathbf{w}_{1:T}$ can be written as the product of conditional Normal densities:

$$p(\mathbf{w}_{1:T}|\boldsymbol{\theta}_{1:T}) = \prod_{t=1}^{T} p(\mathbf{w}_t|\boldsymbol{\theta}_t) = \prod_{t=1}^{T} \prod_{i=1}^{n_t} p(w_t(\boldsymbol{s}_{t,i})|\mathbf{w}_{t,<i}), \qquad (3.1)$$

where $p(\mathbf{w}_t|\boldsymbol{\theta}_t) = N(\mathbf{w}_t|0, \mathbf{C}(\boldsymbol{\theta}_t))$, $\prod_{i=1}^{n_t} p(w_t(\boldsymbol{s}_{t,i})|\mathbf{w}_{t,<i})$ is the conditional representation of the joint distribution of $\mathbf{w}_t$, $\mathbf{w}_{t,<i} = \{w_t(s_{t,1}), w_t(\boldsymbol{s}_{t,2}), \ldots, w_t(\boldsymbol{s}_{t,i-1})\}$ for $2 \le i \le n_t$, and $\mathbf{w}_{t,<1} = \emptyset$. We specify a multivariate Gaussian distribution over a fixed set of points in the domain, to which we refer to as the *reference set*. For simplicity and computational efficiency, the reference set is chosen to coincide with the set of observed locations $\mathbf{S}_t$. As demonstrated in (Datta et al., 2016), based on the reference set, we can extend this finite-dimensional multivariate Normal distribution to a stochastic process over the domain.

Let $N_t(\boldsymbol{s}_{t,i})$ be a subset of locations from $\mathbf{S}_{t,<i} = \{\boldsymbol{s}_{t,1}, \boldsymbol{s}_{t,2}, \ldots, \boldsymbol{s}_{t,i-1}\}$. $N_t(\boldsymbol{s}_{t,i})$ is constructed by choosing at most m "nearest neighbors" of location $\boldsymbol{s}_{t,i}$ from $\mathbf{S}_{t,<i}$ such that:

$$N_t(\boldsymbol{s}_{t,i}) = \begin{cases} \emptyset & , \text{ for } i = 1, \\ \{\boldsymbol{s}_{t,1}, \boldsymbol{s}_{t,2}, \ldots, \boldsymbol{s}_{t,i-1}\} & , \text{ for } 2 \le i \le m, \\ m \text{ nearest neighbors among} \{\boldsymbol{s}_{t,1}, \boldsymbol{s}_{t,2}, \ldots, \boldsymbol{s}_{t,i-1}\} & , \text{ for } i > m. \end{cases}$$

Given the above specification of nearest neighbors, and its ordering mechanism, the density $p(\mathbf{w}_t|\boldsymbol{\theta}_t)$ is approximated by $\tilde{p}(\mathbf{w}_t|\boldsymbol{\theta}_t) = \prod_{i=1}^{n_t} p(w_t(\boldsymbol{s}_{t,i})|\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i})})$. It can be shown that $w_t(\boldsymbol{s}_{t,i})|\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i})} \sim N(\mathbf{B}_{t,\boldsymbol{s}_{t,i}}\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i})}, F_{t,\boldsymbol{s}_{t,i}})$, where $\mathbf{B}_{t,\boldsymbol{s}_{t,i}} = \mathbf{C}^T_{\boldsymbol{s}_{t,i},N_t(\boldsymbol{s}_{t,i})}\mathbf{C}^{-1}_{N_t(\boldsymbol{s}_{t,i})}$, $F_{t,\boldsymbol{s}_{t,i}} = \mathbf{C}(\boldsymbol{s}_{t,i}, \boldsymbol{s}_{t,i}) - \mathbf{C}^T_{\boldsymbol{s}_{t,i},N_t(\boldsymbol{s}_{t,i})}\mathbf{C}^{-1}_{N_t(\boldsymbol{s}_{t,i})}\mathbf{C}_{\boldsymbol{s}_{t,i},N_t(\boldsymbol{s}_{t,i})}$, $\mathbf{C}_{\boldsymbol{s}_{t,i},N_t(\boldsymbol{s}_{t,i})}$ is the covariance matrix of $w_t(\boldsymbol{s}_{t,i})$ and $\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i})}$, and $\mathbf{C}_{N_t(\boldsymbol{s}_{t,i})}$ is the covariance matrix of $\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i})}$. Thus the nearest neighbor density $\tilde{p}(\mathbf{w}_t|\boldsymbol{\theta}_t)$ is Normal with mean zero and covariance $\tilde{\mathbf{C}}(\boldsymbol{\theta}_t)$, where $\tilde{\mathbf{C}}^{-1}(\boldsymbol{\theta}_t)$ is a sparse matrix with at most $\frac{1}{2}n_t m(m+1)$ non-zero elements (Appendix A).

With NNGP prior specification for $\mathbf{w}_t$ and general prior formulation for $\boldsymbol{\Theta}_t = (\boldsymbol{\phi}_t, \boldsymbol{\sigma}_t^2, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_{t-1}, \tau_t)$, the posterior distribution is

$$p(\boldsymbol{\Theta}_{1:T}, \mathbf{w}_{1:T}|\mathbf{Z}_{1:T}) \propto L(\mathbf{Z}_{1:T}|\boldsymbol{\theta}_{1:T}, \boldsymbol{\beta}_{1:T}, \boldsymbol{\gamma}_{1:T-1}, \tau_{1:T}, \mathbf{w}_{1:T}) \prod_{t=1}^{T} p(\boldsymbol{\Theta}_t)\tilde{p}(\mathbf{w}_t|\boldsymbol{\theta}_t) \qquad (3.2)$$

while the likelihood kernel is a multivariate Normal density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Lambda}$, which are defined in Appendix B. The covariance matrix $\boldsymbol{\Lambda}$ is not sparse since the cross covariance in $\mathbf{S}_{t'} \backslash \mathbf{S}_t$ for $t' > t$ is not zero. So the likelihood, conditional on the nearest neighbor spatial random effects, cannot be simplified unless $\mathbf{S}_{t'} \backslash \mathbf{S}_t = \emptyset$ (or equivalently $\mathbf{S}_{t'} \subseteq \mathbf{S}_t$, that is the observation locations are in a nested hierarchical structure). Thus, the direct implementation of NNGP on $\mathbf{w}_{1:T}$ when observed locations are not fully nested (such that $\mathbf{S}_t \backslash \mathbf{S}_{t'} = \emptyset$ for $t > t'$) may still lead to infeasible computational complexity.

To overcome the computational issue, we introduce new evaluations of the spatial process $w_t(\cdot)$ for each level. The new evaluations of $w_t(\cdot)$ are done in such a way that the reference set of level t is nested within the reference set of level $t+1$. Choosing this fully nested structure allows the likelihood and the posterior to be factorized into T conditionally independent parts. Moreover, each of the conditionally independent parts of the factorized likelihood has a diagonal covariance matrix similar to NNGP. We call this new computationally efficient procedure the Nearest Neighbor Co-kriging Gaussian Process (NNCGP). NNCGP utilizes the computational advantages of both the auto-regressive co-kriging model and the NNGP. The proposed NNCGP is a well-defined process derived from a parent co-kriging Gaussian process. For any fidelity level t and finite set $V \in D$, $\tilde{p}(w_{t,V})$ is the density of the realizations of a Gaussian process over V.

Consider observed data sets $\{\mathbf{Z}_t, \mathbf{S}_t\}$, with the corresponding spatial process vectors and output vectors $\mathbf{y}_t = y_t(\mathbf{S}_t) = \{y_t(\boldsymbol{s}_{t,1}), \ldots, y_t(\boldsymbol{s}_{t,n_t})\}$. Set $\mathbf{S}_t^* = \bigcup\limits_{i=t+1}^{T} \mathbf{S}_i \backslash \mathbf{S}_t = \{s_{t,1}^*, \ldots, s_{t,n_t^*}^*\}$ as an additional reference set of fidelity level t, which contains the observed locations that are not in the tth level but are in higher fidelity levels. Denote $\mathbf{w}_t^* = \{w_t(\boldsymbol{s}_{t,1}^*), \ldots, w_t(\boldsymbol{s}_{t,n_t^*}^*)\}$ as the spatial interpolants with corresponding $\mathbf{y}_t^*$. By construction, S_t^* is the smallest collection of sets of spatial locations required to be added to the original observations S_t in order to obtain hierarchically nested locations. Consequently, S_1^* consists of all locations observed in higher fidelity levels but not at the first fidelity level and $S_T^* = \emptyset$. Let $\tilde{\mathbf{w}}_t = \mathbf{w}_t^* \cup \mathbf{w}_t$,
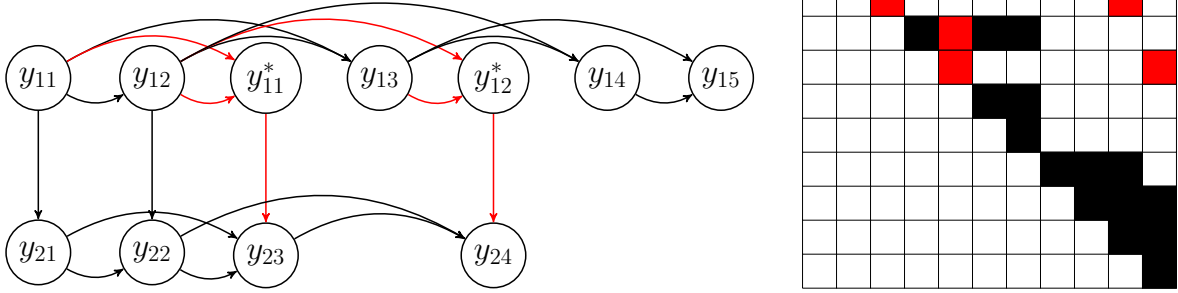
9

Graphical representation of NNCGP

Figure 1: Toy example of the NNCGP for two fidelity levels (T=2) with $n_1 = 5$, $n_2 = 4$, $n_1^* = 2$. Left: directed acyclic graphical representation of the noiseless part of the NNCGP model. Right: covariance matrix sparsity of the model (white squares represent zero cross-covariance).

$\tilde{\mathbf{y}}_t = y_t(\mathbf{S}_t^*) \cup y_t(\mathbf{S}_t)$, $\tilde{\boldsymbol{S}}_t = \mathbf{S}_t^* \cup \mathbf{S}_t$, and $\tilde{n}_t = n_t + n_t^*$. Thus the complete set of observed locations $\tilde{\boldsymbol{S}}_t$ and $\mathbf{S}_{t'}$ have a nested hierarchical structure with $\mathbf{S}_{t'} \subseteq \tilde{\boldsymbol{S}}_t$ when $t' \geq t$. By sequentially adding $\mathbf{w}_t^*$ to each level, we can construct a fully nested hierarchical model. Figure 1 illustrates the proposed procedure in a directed acyclic graph (DAG) representation of a toy example with two fidelity levels.

Using the Markovian property of the co-kriging model, the joint likelihood can be factorized as a product of likelihoods from different fidelity levels conditional on augmented spatial interpolants, i.e.:

$$L(\mathbf{Z}_{1:T}|\cdot) = p(\mathbf{Z}_1|\mathbf{w}_1, \boldsymbol{\beta}_1, \tau_1) \prod_{t=2}^{T} p(\mathbf{Z}_t|\mathbf{w}_t, \boldsymbol{\beta}_t, y_{t-1}(\mathbf{S}_t), \boldsymbol{\gamma}_{t-1}, \tau_t)$$

$$= N(\mathbf{Z}_1|\mathbf{h}_1(\mathbf{S}_1)\boldsymbol{\beta}_1 + \mathbf{w}_1, \tau_1\mathbf{I}) \prod_{t=2}^{T} N(\mathbf{Z}_t|\zeta_{t-1}(\mathbf{S}_t) \circ y_{t-1}(\mathbf{S}_t) + \delta_t(\mathbf{S}_t), \tau_t\mathbf{I}), \quad (3.3)$$

where $\circ$ is the Hadamard production symbol and $y_{t-1}(\mathbf{S}_t) \subset \tilde{\mathbf{y}}_{t-1}$. Based on the NNGP prior on the $\mathbf{w}_{1:T}$ described above, we can write the joint prior distribution of $\tilde{\mathbf{w}}_{1:T}$ as:

$$\tilde{p}(\tilde{\mathbf{w}}_{1:T}|\boldsymbol{\theta}_{1:T}) = \prod_{t=1}^{T} \tilde{p}(\mathbf{w}_t^*|\mathbf{w}_t, \boldsymbol{\theta}_t)\tilde{p}(\mathbf{w}_t|\boldsymbol{\theta}_t). \quad (3.4)$$

Given the above representation of the likelihood and prior, the joint posterior density function

10

of NNCGP for a T level system is:

$$p(\mathbf{\Theta}_{1:T}, \tilde{\mathbf{w}}_{1:T}|\mathbf{Z}_{1:T}) \propto p(\mathbf{\Theta}_1)\tilde{p}(\mathbf{w}_1|\boldsymbol{\theta}_1)\tilde{p}(\mathbf{w}_1^*|\mathbf{w}_1, \boldsymbol{\theta}_1)N(\mathbf{Z}_1|\mathbf{h}_1(\mathbf{S}_1)\boldsymbol{\beta}_1 + \mathbf{w}_1, \tau_1\mathbf{I})$$

$$\times \prod_{t=2}^{T} \{p(\mathbf{\Theta}_t)\tilde{p}(\mathbf{w}_t|\boldsymbol{\theta}_t)\tilde{p}(\mathbf{w}_t^*|\mathbf{w}_t, \boldsymbol{\theta}_t)N(\mathbf{Z}_t|\zeta_{t-1}(\mathbf{S}_t) \circ y_{t-1}(\mathbf{S}_t) + \delta_t(\mathbf{S}_t), \tau_t\mathbf{I})\} . \quad (3.5)$$

The computational complexity of implementing the NNCGP model is dominated by the evaluation and storage of T sparse matrices $(\tilde{\mathbf{C}}_1^{-1}(\boldsymbol{\theta}_1), \ldots, \tilde{\mathbf{C}}_T^{-1}(\boldsymbol{\theta}_T))$. Thus, the joint posterior distribution of the NNCGP model can be calculated using $\mathcal{O}(\sum_{t=1}^{T} \tilde{n}_t m^3)$ flops and needs $\mathcal{O}(\sum_{t=1}^{T} \tilde{n}_t m^2)$ dynamic memory storage. Introducing $\mathbf{w}_{1:T}^*$ reduces the computational complexity as well as enables the specification of semi-conjugate priors for $(\boldsymbol{\sigma}_t^2, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_{t-1}, \tau_t)$ which facilitates tractability of posterior marginals and conditionals, as we explain below. An alternative approach is to use a common reference set for all of the fidelity levels. However, this may result into a more expensive procedure since each level will have the same computational complexity as the first level $\mathcal{O}(T\tilde{n}_1 m^3)$.

4   Bayesian Inference

In this section, we present the MCMC sampler for the inference of parameters $\mathbf{\Theta}_{1:T}$ for a T level NNCGP with observations $\mathbf{Z}_{1:T}$ and spatial location input sets $\mathbf{S}_{1:T}$. We also present the prediction procedure for output $\mathbf{Z}_t(\boldsymbol{s}_p)$ at an unobserved location $\boldsymbol{s}_p$ for any specified fidelity level t.

NNCGP model representation allows us to construct an efficient MCMC sampler to facilitate parameter and prediction inference. Since the components of $\mathbf{w}_t^*|\mathbf{w}_t$ are independent, we can update $\mathbf{w}_t^*$ individually. For locations $\boldsymbol{s}_u \in \mathbf{S}_t^*$, the full conditional posterior distribution of $w_t(\boldsymbol{s}_u) \sim N(V_t^*(\boldsymbol{s}_u)\mu_t^*(\boldsymbol{s}_u), V_t^*(\boldsymbol{s}_u))$ with $V_t^*(\boldsymbol{s}_u)$ and $\mu_t^*(\boldsymbol{s}_u)$ are specified in Equation C3 in the Appendix. The introduction of the spatial interpolant $\mathbf{w}_{t-1}^*$ provides the full conditional

posterior distribution of a spatial random process $\mathbf{w}_t$ as

$$w_t(\boldsymbol{s}_u)|\boldsymbol{\beta}_t, \boldsymbol{\theta}_t, \tau_t, \mathbf{Z}_t, \tilde{\mathbf{y}}_{t-1}, \boldsymbol{\gamma}_{t-1} \sim N(V_{\mathbf{w}_t}(\boldsymbol{s}_u)\mu_{\mathbf{w}_t}(\boldsymbol{s}_u), V_{\mathbf{w}_t}(\boldsymbol{s}_u))$$

where

$$V_{\mathbf{w}_t}(\boldsymbol{s}_u) = \left(F_{t,\boldsymbol{s}_u}^{-1} + \tau_t^{-2}\right)^{-1},$$

$$\mu_{\mathbf{w}_t}(\boldsymbol{s}_u) = \tau_t^{-2}[z_t(\boldsymbol{s}_u) - \mathbf{h}_t^T(\boldsymbol{s}_u)\boldsymbol{\beta}_t - \zeta_{t-1}(\boldsymbol{s}_u)y_{t-1}(\boldsymbol{s}_u)] + F_{t,\boldsymbol{s}_u}^{-1}\mathbf{B}_{t,\boldsymbol{s}_u}\mathbf{w}_{t,N_t(\boldsymbol{s}_u)}, \qquad (4.1)$$

for $t = 2, \ldots, T$, $\boldsymbol{s}_u \in \mathbf{S}_t$. The full conditional distribution of $\mathbf{w}_1$ is similar to the univariate case of the full conditional distribution of the spatial process (Datta et al., 2016), see Equation D.1 in the Appendix.

To take full advantage of the posterior representation in (3.5), we chose independent prior distributions for parameters at different levels such as:

$$p(\boldsymbol{\Theta}_{1:T}) = \prod_{t=1}^{T} p(\sigma_t)p(\boldsymbol{\phi}_t)p(\boldsymbol{\beta}_t)p(\tau_t)p(\boldsymbol{\gamma}_{t-1}). \qquad (4.2)$$

The above prior representation coupled with (3.5) results in T separate conditional parts for the posterior. To facilitate further computations, we assign conditional conjugate priors: $\sigma_t^2 \sim IG(a_t, b_t)$, $\boldsymbol{\beta}_t \sim N(\boldsymbol{\mu}_{\beta_t}, \mathbf{V}_{\beta_t})$, $\tau_t^2 \sim IG(c_t, d_t)$ for $t = 1, 2, \ldots, T$ and $\boldsymbol{\gamma}_{t-1} \sim N(\mu_{\gamma_{t-1}}, V_{\gamma_{t-1}})$ for $t = 2, 3, \ldots, T$, which leads to standard full conditional posteriors

$$\boldsymbol{\beta}_t|\mathbf{w}_t, \tilde{\mathbf{y}}_{t-1}, \boldsymbol{\gamma}_{t-1}, \tau_t, \mathbf{Z}_t \sim N(\mathbf{V}_{\beta_t}^*\boldsymbol{\mu}_{\beta_t}^*, \mathbf{V}_{\beta_t}^*),$$

$$\boldsymbol{\gamma}_t|\tilde{\mathbf{y}}_t, \boldsymbol{\beta}_{t+1}, \tau_{t+1}, \mathbf{Z}_{t+1} \sim N(\mathbf{V}_{\gamma_t}^*\boldsymbol{\mu}_{\gamma_t}^*, \mathbf{V}_{\gamma_t}^*),$$

$$\sigma_t^2|\tilde{\mathbf{w}}_t, \boldsymbol{\phi}_t \sim IG(a_{\sigma_t}^*, b_{\sigma_t}^*),$$

$$\tau_t^2|\boldsymbol{\beta}_t, \mathbf{w}_t, \tilde{\mathbf{y}}_{t-1}, \mathbf{Z}_t, \boldsymbol{\gamma}_{t-1} \sim IG(a_{\tau_t}, b_{\tau_t}), \qquad (4.3)$$

where the parameters are specified in Equations D.2-D.6 of the Appendix. For the range parameter $p(\phi_{t,j})$, we chose the bounded prior $p(\phi_{t,j}), \sim U(0, l_{t,j})$ to avoid numerical instabilities, where $l_{t,j}$ is defined from the researcher and is usually associated with the maximum distance in the jth direction. The conditional posterior distribution for $\boldsymbol{\phi_t}$ in NNCGP model

12

is

$$p(\boldsymbol{\phi}_t|\tilde{\mathbf{w}}_t, \sigma_t^2) \propto p(\boldsymbol{\phi}_t)|\tilde{\mathbf{C}}(\boldsymbol{\theta}_t)|^{-1/2}\exp\left\{-\frac{1}{2}\tilde{\mathbf{w}}_t^T\tilde{\mathbf{C}}^{-1}(\boldsymbol{\theta}_t)\tilde{\mathbf{w}}_t\right\}, \tag{4.4}$$

where $\boldsymbol{\phi}_t$ appears in the sparse cross covariance matrix $\tilde{\mathbf{C}}(\boldsymbol{\theta}_t)$ and it cannot be sampled directly. The Metropolis-Hasting algorithm (Hastings, 1970) can be used to update $\boldsymbol{\phi}_t$ in the full conditional distribution.

For a new input location $s_p \notin \tilde{\boldsymbol{S}}_t$, the prediction process is to generate $z_t(s_p)$ based on its predictive distribution. Subsequently, we generate $w_{t'}(s_p)$ independently for each level from sampler $w_{t'}(\boldsymbol{s}_p) \sim N(V_{t',\boldsymbol{s}_p}\mu_{t',\boldsymbol{s}_p}, V_{t',\boldsymbol{s}_p})$ for $t' = 1, ..., t$; where $V_{t',\boldsymbol{s}_p}$, $\mu_{t',\boldsymbol{s}_p}$ are specified in (D.7), while $y_{t'}(\boldsymbol{s}_p)$ are generated by $y_{t'}(\boldsymbol{s}_p) = \zeta_{t'-1}(\boldsymbol{s}_p)y_{t'-1}(\boldsymbol{s}_p) + \mathbf{h}_{t'}(\boldsymbol{s}_p)\boldsymbol{\beta}_{t'} + w_{t'}(\boldsymbol{s}_p)$. The $z_t(\boldsymbol{s}_p)$ is generated by the MCMC sampler $z_t(\boldsymbol{s}_p)|\cdots \sim N(y_t(\boldsymbol{s}_p), \tau_t)$.

5 Synthetic data example

This section conducts a simulation study to evaluate the performance of the proposed NNCGP model in comparison to the NNGP model using the highest fidelity level data only (denoted as the single level NNGP model) and using both fidelity level data sets combined as a single data-set (denoted as the combined NNGP model). Details about the metrics used for the comparison can be found in Appendix E. The simulations were performed in MATLAB R2018a, on a computer with specifications (intelR i7-3770 3.4GHz Processor, RAM 8.00GB, MS Windows 64bit).

We consider a two-fidelity level system in a two dimensional unit square domain, parameterized as an auto-regressive co-kriging Gaussian process as specified in (2.1). For simplicity, the mean of $y_1(s)$, the mean of the additive discrepancy $\delta_2(s)$, and the scalar discrepancy $\zeta_1(s)$ are assumed to be constant. The covariance function of $y_1(s)$ and $\delta_2(s)$ are assumed to be exponential. The true values of the model parameters are listed in the first column

of Table 1. Based on the above statistical model, we generated $5,000$ observations for each fidelity level $\mathbf{Z}_1$ and $\mathbf{Z}_2$ at randomly selected locations $\mathbf{S}_1$ and $\mathbf{S}_2$ such that $\mathbf{S}_1 \cap \mathbf{S}_2 = \emptyset$. The $10,000$ generated observations are shown in Figures 2(a-c). To assess predictive performance, we randomly selected two small square subregions for testing from the high fidelity level dataset (Figure 2b).

For the Bayesian inference of NNCGP on the unknown parameters β_1, β_2 and γ_1, we assigned independent Normal prior distributions with zero mean and large variances. We used inverse Gamma priors for the spatial and noise variances $\sigma_t^2 \sim IG(2,1)$ and $\tau_t^2 \sim IG(2,1)$, respectively. The range correlation parameters ϕ_1 and ϕ_2 each used a uniform prior $U(0,100)$. Similar non-informative priors were used for both the single level NNGP as well as the combined NNGP model. For all three models, we ran the Markov chain Monte Carlo (MCMC) sampler as described in Section 4 with $40,000$ iterations where the first $5,000$ iterations were discarded as burn-in. The convergence of the MCMC sampler for each parameter was assessed from their associated trace plots.

In Table 1, we report the Monte Carlo estimates of the posterior means and the associated 95% marginal credible intervals of the unknown parameters using the three different NNGP based procedures with $m = 10$ neighbors. All but τ_1^2 true values of the parameters are successfully included in the 95% marginal credible intervals. The introduction of spatial interpolants may have caused a slight over estimation of τ_2, however, the true values of the nugget variances are successfully captured in the 95% marginal credible intervals. The uncertainty in the parameter estimations can be improved with a semi-nested or nested structure between the observed locations for the different fidelity levels, as shown for the auto-regressive co-kriging model in Konomi and Karagiannis (2021).

In Table 2, we report standard performance measures (defined in Appendix E) for the proposed NNCGP, single level NNGP, and combined NNGP models with $m = 10$ neighbors. All performance measures indicate that NNCGP has better predictive ability than the single

	True	Single level NNGP		Combined NNGP		NNCGP	
				Model			
β_1	10	10.82	(9.84, 11.27)	10.15	(9.57, 10.71)	9.71	(9.36, 10.16)
β_2	1	-	-	-	-	0.87	(0.39, 1.36)
σ_1^2	4	3.79	(2.97, 5.19)	4.89	(3.65, 6.27)	3.51	(2.71, 4.52)
σ_2^2	1	-	-	-	-	1.05	(0.18, 2.31)
$1/\phi_1$	10	13.29	(9.33, 17.51)	8.75	(6.49, 11.94)	10.77	(8.07, 13.91)
$1/\phi_2$	10	-	-	-	-	12.61	(3.93, 24.07)
γ_1	1	-	-	-	-	0.995	(0.983, 1.051)
τ_1^2	0.1	0.138	(0.115, 0.183)	0.478	(0.451, 0.508)	0.125	(0.097, 0.148)
τ_2^2	0.05	-	-	-	-	0.158	(0.041, 0.232)
m	10	-	-	-	-	-	-

Table 1: Unknown parameters are in the 1st column; their true values in the 2nd column; their Bayesian point estimates and marginal credible intervals for the single level NNGP, combined NNGP, and NNCGP models are in the 3rd, 4th and 5th columns, respectively. The level 2 data set is used in the single level NNGP model, while both the level 1 and 2 data sets are used in the combined NNGP model and are treated as following a one level system. The estimated parameters by the single level NNGP and combined NNGP models are treated as level 1 parameters.

level NNGP and combined NNGP. NNCGP produces a significantly smaller effective number of model parameters (PD) and Deviance Information Criterion (DIC) than the single level and combined NNGP, which suggests that NNCGP provides a better fit when complexity is considered. The root mean square prediction error (RMSPE) produced by NNCGP is approximately 40 - 50% smaller than that of the single level NNGP and 20 - 30% smaller than that of the combined NNGP. The Nash-Sutcliffe model efficiency coefficient (NSME) of NNCGP is closer to 1 than that of both other methods, which suggests that NNCGP provides a substantial improvement in the prediction.

In Figure 2 we observe that, for the testing regions, the NNCGP has more accurately captured the roughness and sharp changes in the response surface while it also provides a better representation of the patterns in the prediction surface. Applying NNGP directly to the high-fidelity dataset provides a smoother prediction surface due to the lack of the information from the low-fidelity dataset; while it fails to produce reliable predictions in the

	Model		
	Single level NNGP	Combined NNGP	NNCGP
RMSPE	2.1325	1.5202	1.0987
NSME	-1.1349	0.0888	0.5108
CVG(95%)	0.8216	0.7136	0.9573
ALCI(95%)	5.2024	3.0856	3.2265
PD	12706	5883	2544
DIC	18136	9659	5551
Time(Hour)	1.4	2.9	4.1

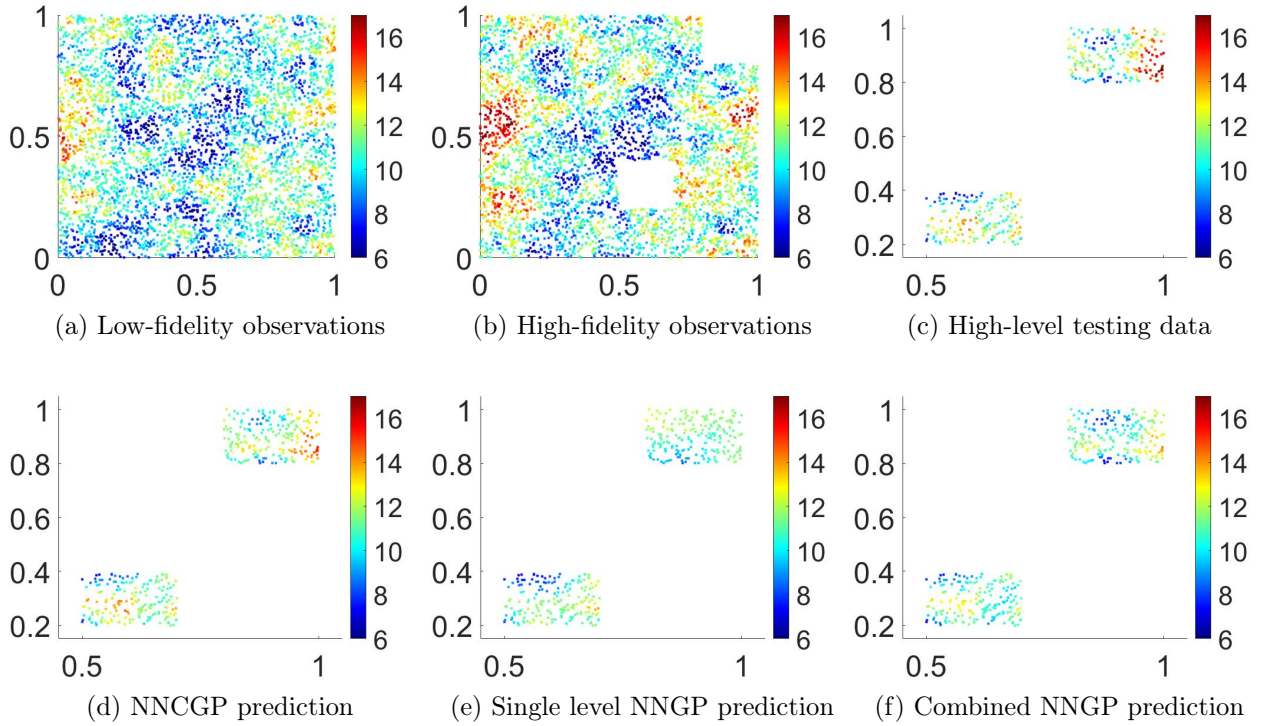Table 2: Performance measures for the predictive ability of the NNCGP model, single level NNGP model and combined NNGP model. (Definitions are given in Appendix E.)

(a) Low-fidelity observations   (b) High-fidelity observations   (c) High-level testing data

(d) NNCGP prediction   (e) Single level NNGP prediction   (f) Combined NNGP prediction

Figure 2: Non-nested observations with a two (low- and high-) fidelity level structure. White boxes in (b) indicate the testing regions. Original testing data (c) along with predictions of the high fidelity level data-set by the (d) multifidelty Nearest Neighbor Co-Krigging Gaussian Process (NNCGP), (e) single level nearest neighbor Gaussian process (NNGP), and (f) the combined nearest neighbor Gaussian process (NNGP) under the non-nested structure.

blank regions. Applying NNGP to the combined dataset with both high- and low- fidelity levels (combined NNGP) also provides an unreliable prediction surface that is similar to the observations in low level regions. Modeling the scalar and additive discrepancies between different levels helps improve predictions. Moreover, NNCGP has produced a CVG closer to 0.95 and a 95% ALCI smaller than that of the single level NNGP and combined NNGP (Table 1). This indicates that NNCGP produces more accurate predictions with a higher probability to cover the true values in narrower credible intervals.

To test the sensitivity of the proposed NNCGP method to the number of neighbors m, we compare the RMSPEs of the three different methods for $m = \{1, 2, 3, 4, 5, 10, 15\}$, and 20 in Figure 3(a). We use the same prior specifications and computational strategies as described above. In terms of prediction accuracy, the NNCGP outperforms both the single level NNGP and the combined NNGP for all m. The decrease of the RMSPE is smaller as m becomes greater than 10. The computational time is longer for the NNCGP than both the single level NNGP and combined NNGP (Figure 3(b)). The NNCGP uses data from both levels and also expands the reference set to ensure the nested structure between different levels. The single NNGP model only uses data from the high-fidelity level. The combined NNGP uses the same dataset as the NNCGP. However, the reference set of the combined NNGP is equal to the reference set of the first fidelity level of NNCGP. It is worth mentioning that if the observation locations are nested or semi-nested the computational complexity of the NNCGP can be further reduced since $\tilde{n}_1 < (n_1 + n_2)$.

6   Application to intersatellite calibration

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). We examine here one
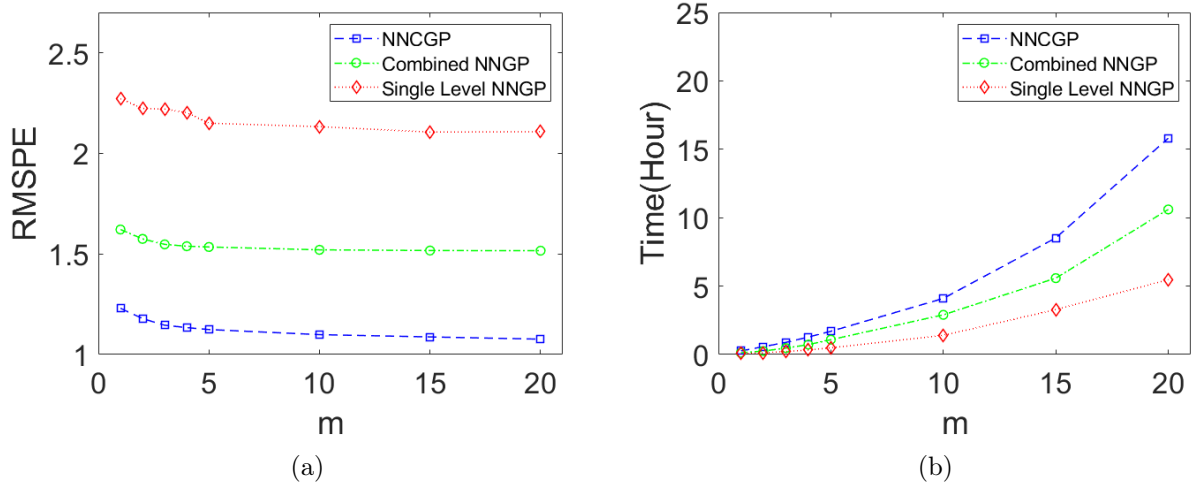
Figure 3: Sensitivity analysis to number of nearest neighbors m: a) Root mean square prediction error (RMSPE) and b) Running time for the NNCGP, single level NNGP, and combined NNGP models for a range of m values over a two fidelity level non-nested synthetic dataset.

instrument in particular, the high-resolution infrared radiation sounder (HIRS) instrument that has been taking measurements since 1978 on board the National Oceanic and Atmospheric Administration (NOAA) polar orbiting satellite series (POES) and the meteorological operational satellite program (Metop) series operated by the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT). This series of more than a dozen satellites currently constitutes over 40 years of HIRS observations, and this unique longevity is valuable to characterize climatological trends. Examples of essential climate variables derived from HIRS measurements include long-term records of temperature and humidity profiles (Shi et al., 2016; Matthews and Shi, 2019).

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 twenty channels, including twelve longwave channels, seven shortwave channels, and one visible channel. Among the longwave channels, Channels 1 to 7 are in the carbon dioxide (CO_2) absorption band to measure atmospheric temperatures from near-

surface to stratosphere, Channel 8 is a window channel for surface temperature observation and cloud detection, Channel 9 is an ozone channel, and Channels 10–12 are for water vapor signals at the near-surface, mid-troposphere, and upper troposphere, respectively. There have been several versions of the instruments where there is a notable change in spatial resolution. In particular, for the HIRS/2 instrument, with observations from the late 1970s to mid-2007, the spatial footprint is approximately 20 km. HIRS/3, with observations from 1998 to mid-2014, has a spatial footprint of approximately 18 km. The currently operational version, HIRS/4, improved the spatial resolution to approximately 10 km at nadir with observations beginning in 2005. 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 data record faces some common challenges when developing CDRs from the time series. Specifically, there are consistency and accuracy issues due to degradation of sensors and intersatellite discrepancies. Furthermore, there is missing information caused by atmospheric conditions such as thick cloud cover. As early as 1991, to address some of these challenges, the co-kriging technique has been applied to remotely sensed data sets (Bhatti et al., 1991). As an improvement to these techniques, we consider using the NNCGP model as a method for intersatellite calibration, data imputation, and data prediction.

We examine HIRS Channel 5 observations from a single day, March 1, 2001, as illustrated in Figure 4. 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
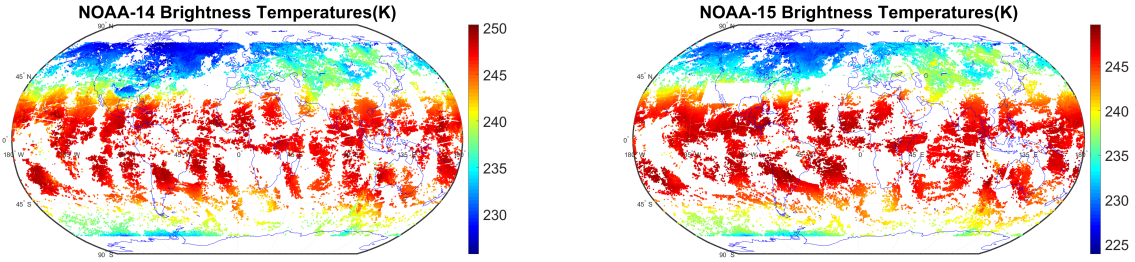
(a) Observations of NOAA 14   (b) Training data of NOAA 15

Figure 4:   Brightness temperature observations by HIRS Channel 5 on March 1, 2001 by (a) NOAA-14 and (b) NOAA-15.

while NOAA-15 became operational in October 1998. 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. Therefore, we treat observations from NOAA-14 as a dataset of low fidelity level, and those from NOAA-15 as a dataset of high fidelity level. A small region of observations from NOAA-15 are treated as testing data, and the remainder of the NOAA-15 observations are treated as training data.

We model our data based on the two-fidelity level NNCGP model as described in Sections 3 & 4. We consider a linear model for the mean of the Gaussian processes in $y_1(\cdot)$ and $\delta_2(\cdot)$ with a 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 $\zeta(s)$ to be an 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 as described in Section 3.

We assign independent normally distributed 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

20

NNCGP, we run the MCMC sampler described in Section 4 with $35,000$ iterations where the first $5,000$ iterations are discarded as a burn-in.

| | Model | | |
	NNCGP	Single level NNGP	Combined NNGP
RMSPE	1.2044	1.8153	1.6772
NSME	0.8439	0.5499	0.6726
CVG(95%)	0.9255	0.8350	0.9197
ALCI(95%)	3.094	4.214	5.778
Time(Hour)	38	20	32

Table 3: Performance measures for the predictive ability of the NNCGP model, single level NNGP model and combined NNGP models for Channel 5 NOAA-14 and NOAA-15 observations on March 1, 2001. (Definitions are given in Appendix E.)

The prediction performance metrics of the three different methods are given in Table 3. Compared to the single level NNGP and combined NNGP models, the NNCGP model produced an approximately 30% smaller RMSPE and its NSME is closer to 1. The NNCGP also produced a larger CVG and a smaller ALCI than the single level and combined NNGP models. The results suggest that the NNCGP model provides a substantial improvement in terms of predictive accuracy in real data analysis. The NOAA-15 testing data shown in Figure 5 shows that the NNCGP model is more capable of capturing the spatial patterns of the testing data than either a single level or a combined NNGP model. Unlike the single level NNGP, the proposed NNCGP uses additional information from NOAA-14. Compared to a combined NNGP, the proposed NNGCP benefits from modeling the discrepancy of observations from the different satellites. With the fully non-nested structure, the computational complexity of the single level NNGP model is $\mathcal{O}(n_2 m^3)$ and for the NNCGP model is $\mathcal{O}((n_1 + n_2)m^3)$; this is consistent with the running times of the models shown in Table 3.

Baseline observations of brightness temperature are used as inputs for so-called remote sensing retrieval algorithms wherein thematic climate variables (e.g. precipitation rates, cloud cover, surface temperature, etc.) are derived. These retrieval algorithms are typically highly nonlinear, so a small change in the input brightness temperature value can have

(a) NOAA-15 testing data-set

(b) Prediction means by NNCGP model

(c) Prediction means by single level NNGP model

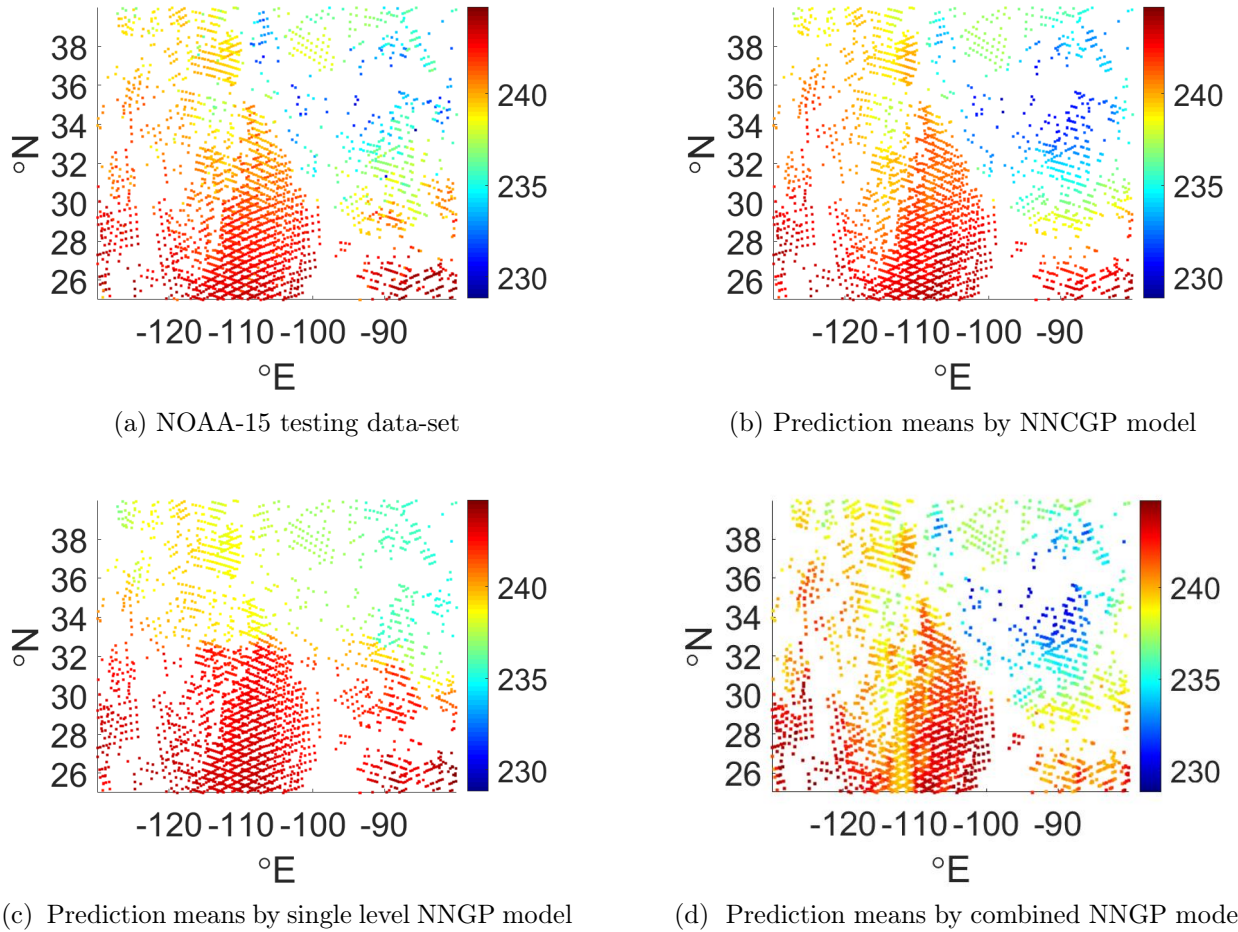(d) Prediction means by combined NNGP model

Figure 5: Predictions of NOAA-15 Brightness Temperatures(K) testing data-set by NNCGP, single level NNGP and combined NNGP under fully non-nested experimental design.

a large impact on the value of derived climate variables. The significant improvement in the prediction accuracy to the brightness temperatures provided by the NNCGP model is therefore critical to the downstream climate variables because of this sensitivity. So although the computation by the NNCGP model is costly compared to the single level NNGP model, it is still worthwhile to apply the NNCGP model.

We applied the NNCGP model to gap-fill predictions based upon a discrete global grid. We chose to use 1° latitude by 1.25° longitude (1° × 1.25°) pixels as a grid structure with near-global spatial coverage from $-70°$ to 70°N. By applying the NNCGP model, we pre-
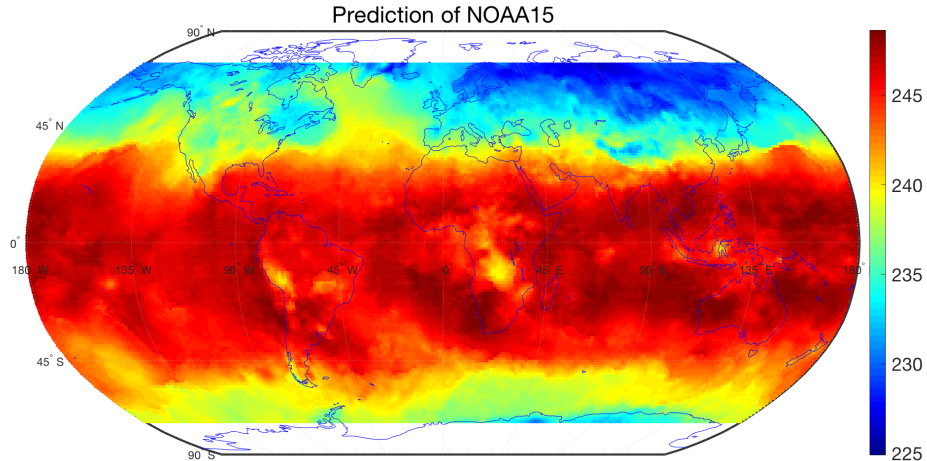
Figure 6: Gap-filled global brightness temperature predictions of NOAA-15 data on a regular grid.

dict gridded NOAA-15 brightness temperature data on the center of the grids, based on the NOAA-14 and NOAA-15 swath-based spatial supports. The prediction plot (Figure 6) illustrates the ability of the NNCGP model to handle large, irregularly spaced data sets and produce a gap-filled composite gridded data-set.

7   Summary and conclusions

In this manuscript we have proposed a new, computationally efficient Nearest Neighbor Autoregressive Co-Kriging Gaussian process (NNCGP) method for the analysis of large irregularly spaced and multi-fidelity spatial data. The proposed NNCGP method extends the scope of the classical auto-regressive co-kriging models (Kennedy and O'Hagan, 2000; Konomi and Karagiannis, 2021) to deal with large data sets. To deal with this complexity, we used independent NNGP priors at each level in the auto-regressive co-kriging model where the neighboors are simply defined within each level. However, when the observed locations are hierarchically non-nested, the likelihood does not simplify; this makes the computational complexity of this approach infeasible. To overcome this issue, we augment the spatial

random effects such that they can form a suitable nested structure. The augmentation of the spatial random effects facilitates the computation of the likelihood of NNCGP. This is because it enables the factorization of the likelihood into terms with smaller covariance matrices and gains the computational efficiency provided by the nearest neighbors within each level. The proposed method is at most computationally linear in the total number of all spatial locations of all fidelity levels. In our simulations, we observed that computations were faster when the observed locations at higher fidelity were nested within those at lower fidelity levels. Moreover, the nested design of the reference sets allows the assignment of semi-conjugate priors for the majority of the parameters. Based on these specifications, we develop efficient and independent MCMC block updates for Bayesian inference. As in the original NNGP paper (Datta et al., 2016), our results indicate that inference is very robust with respect to values of neighbors.

We compared the proposed NNCGP with NNGP in the single level of highest-fidelity with a simulation study and a real data application of intersatellite calibration. We observed that the NNCGP was able to improve the prediction accuracy of HIRS brightness temperatures from the NOAA-15 polar-orbiting satellite by incorporating information from an older version of the same HIRS sensor onboard the polar-orbiting satellite NOAA-14. Beyond HIRS, the proposed methodology can be used for a variety of large multi-fidelity level remotely sensed data sets with overlapping observations from sensors of similar design. Furthermore, we propose that the proposed methodology can be used in a wide range of applications in physical science and engineering when multiple computer models with large simulation runs are available.

Several extensions of the proposed NNCGP method can be pursued in future work. Based on the resulting conditional posterior distributions, the Bayesian inference can be derived in parallel for each fidelity level similarly to how is done for the NNGP (Datta et al., 2016). In this case, if parallel computing is available, the computational complexity of the approach

can be reduced up to $\mathcal{O}(\tilde{n}_1 m^3)$. Another possible extension is to accelerate our inferential method with a reference or conjugate NNCGP based on ideas given in Finley et al. (2019). The proposed model may also be extended in the multivariate setting by using parallel partial autoregressive co-kriging (Ma et al., 2019). Finally, the proposed NNCGP model can be extended to spatial-temporal settings with discretized time steps, as both autoregressive structures and the NNGP approach are capable of incorporating temporal dependence.

Acknowledgments

References

Aboufirassi, M. and Mariño, M. A. (1984), "Cokriging of aquifer transmissivities from field measurements of transmissivity and specific capacity," *Journal of the International Association for Mathematical Geology*, 16, 19–35.

Baker, E., Barbillon, P., Fadikar, A., Gramacy, R. B., Herbei, R., Higdon, D., Huang, J., Johnson, L. R., Ma, P., Mondal, A., Pires, B., Sacks, J., and Sokolov, V. (2020), "Analyzing Stochastic Computer Models: A Review with Opportunities," .

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, New York: Chapman and Hall/CRC, second edition.

Banerjee, S., Gelfand, A. E., Finley, A. O., and 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.

Bhatti, A., Mulla, D., and Frazier, B. (1991), "Estimation of soil properties and wheat yields on complex eroded hills using geostatistics and Thematic Mapper images," *Remote Sensing of Environment*, 37, 181–191.

Cao, C., Weinreb, M., and Xu, H. (2004), "Predicting simultaneous nadir overpasses among polar-orbiting meteorological satellites for the intersatellite calibration of radiometers," *Journal of Atmospheric and Oceanic Technology*, 21, 537–542.

Cao, C., Xu, H., Sullivan, J., McMillin, L., Ciren, P., and Hou, Y.-T. (2005), "Intersatellite radiance biases for the High-Resolution Infrared Radiation Sounders (HIRS) on board NOAA-15,-16, and-17 from simultaneous nadir observations," *Journal of Atmospheric and Oceanic Technology*, 22, 381–395.

Chander, G., Hewison, T., Fox, N., Wu, X., Xiong, X., and Blackwell, W. (2013), "Overview of Intercalibration of Satellite Instruments," *IEEE Transactions on Geoscience and Remote Sensing*, 51:3, 1056–1080.

Cressie, N. (1993), *Statistics for Spatial Data*, New York: John Wiley & Sons, revised edition.

Cressie, N. and 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., and Gelfand, A. E. (2016), "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets," *Journal of the American Statistical Association*, 111, 800–812.

Davis, B. M. and Greenes, K. A. (1983), "Estimation using spatially distributed multivariate data: an example with coal quality," *Journal of the International Association for Mathematical Geology*, 15, 287–300.

Du, J., Zhang, H., Mandrekar, V., et al. (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., and Banerjee, S. (2019), "Efficient algorithms for Bayesian nearest neighbor Gaussian processes," *Journal of Computational and Graphical Statistics*, 1–14.

Furrer, R. and Genton, M. G. (2011), "Aggregation-cokriging for highly multivariate spatial data," *Biometrika*, 98, 615–631.

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

Genton, M. G. and Kleiber, W. (2015), "Cross-covariance functions for multivariate geostatistics," *Statistical Science*, 147–163.

Goldberg, M., e. a. (2011), "The Global Space-Based Inter-Calibration Systems," *Bull. Am. Meteorol. Soc.*, 92, 467–475.

Gramacy, R. B. and Apley, D. W. (2015), "Local Gaussian Process Approximation for Large Computer Experiments," *Journal of Computational and Graphical Statistics*, 24, 561–578.

Gramacy, R. B. and Lee, H. K. H. (2012), "Cases for the nugget in modeling computer experiments," *Statistics and Computing*, 22, 713–722.

Han, Z.-H., Zimmermann, R., and Goretz, S. (2010), "A new cokriging method for variable-fidelity surrogate modeling of aerodynamic data," in *48th AIAA Aerospace sciences meeting including the new horizons forum and Aerospace exposition*.

Hastings, W. K. (1970), "Monte Carlo sampling methods using Markov chains and their

applications," .

Jackson, D., Wylie, D., and Bates, J. (2003), "The HIRS pathfinder radiance data set (1979–2001)," in *Proc. of the 12th Conference on Satellite Meteorology and Oceanography, Long Beach, CA, USA, 10-13 February 2003*, volume 5805 of *LNCS*, Springer.

Katzfuss, M. (2016), "A multi-resolution approximation for massive spatial datasets," *Journal of the American Statistical Association*. DOI:10.1080/01621459.2015.1123632.

Katzfuss, M. and Guinness, J. (2021), "A General Framework for Vecchia Approximations of Gaussian Processes," *Statistical Science*, 36, 124 – 141, URL https://doi.org/10.1214/19-STS755.

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

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

Konomi, B. A. and Karagiannis, G. (2021), "Bayesian Analysis of Multifidelity Computer Models With Local Features and Nonnested Experimental Designs: Application to the WRF Model," *Technometrics*, 0, 1–13, URL https://doi.org/10.1080/00401706.2020.1855253.

Konomi, B. A., Sang, H., and Mallick, B. K. (2014), "Adaptive Bayesian Nonstationary Modeling for Large Spatial Datasets Using Covariance Approximations," *Journal of Computational and Graphical Statistics*, 23, 802–829, URL https://doi.org/10.1080/10618600.2013.812872.

Koziel, S., Bekasiewicz, A., Couckuyt, I., and Dhaene, T. (2014), "Efficient multi-objective simulation-driven antenna design using co-kriging," *IEEE Transactions on Antennas and Propagation*, 62, 5900–5905.

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

Lindgren, F., Rue, H., and 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.

Ma, P., Karagiannis, G., Konomi, B. A., Asher, T. G., Toro, G. R., and Cox, A. T. (2019), "Multifidelity Computer Model Emulation with High-Dimensional Output: An Application to Storm Surge," .

Matthews, J. L. and Shi, L. (2019), "Intercomparisons of Long-Term Atmospheric Temperature and Humidity Profile Retrievals," *Remote Sensing*, 11, 853.

National Research Council (2004), *Climate Data Records from Environmental Satellites: Interim Report*, Washington, DC: The National Academies Press, URL https://www.nap.edu/catalog/10944/

`climate-data-records-from-environmental-satellites-interim-report`.

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

— (2017), "Multivariate spatial data fusion for very large remote sensing datasets," *Remote Sensing*, 9, 142.

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

Paciorek, C. and Schervish, M. (2006), "Spatial modelling using a new class of nonstationary covariance functions," *Environmetrics*, 17, 483–506.

Qian, P. Z. and Wu, C. J. (2008), "Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments," *Technometrics*, 50, 192–204.

Qian, Z., Seepersad, C. C., Joseph, V. R., Allen, J. K., and Jeff Wu, C. F. (2005), "Building Surrogate Models Based on Detailed and Approximate Simulations," *Journal of Mechanical Design*, 128, 668–677, URL `https://doi.org/10.1115/1.2179459`.

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

Shi, L., Matthews, J., Ho, S.-p., Yang, Q., and Bates, J. (2016), "Algorithm development of temperature and humidity profile retrievals for long-term HIRS observations," *Remote Sensing*, 8, 280.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging. 2nd edition*, New York: Springer.

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

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

Taylor-Rodriguez, D., Finley, A. O., Datta, A., Babcock, C., Andersen, H.-E., Cook, B. D., Morton, D. C., and 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.

Ver Hoef, J. M. and Cressie, N. (1993), "Multivariable spatial prediction," *Mathematical Geology*, 25, 219–240.

Xiong, X., Cao, C., and Chander, G. (2010), "An overview of sensor calibration inter-

comparison and applications," *Frontiers of Earth Science in China*, 4, 237–252.

Appendix

A NNGP specifications

The posterior distribution of

$$\tilde{p}(\mathbf{w}_t|\cdot) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n_t}\left\{w_t(s_{t,i}) - \mathbf{B}_{t,s_{t,i}}\mathbf{w}_{t,N_t(s_{t,i})}\right\}^T F_{t,s_{t,i}}^{-1}\left\{w_t(s_{t,i}) - \mathbf{B}_{t,s_{t,i}}\mathbf{w}_{t,N_t(s_{t,i})}\right\}\right]$$

$$= \exp\left(-\frac{1}{2}\mathbf{w}_t^T\mathbf{B}_t^T\mathbf{F}_t^{-1}\mathbf{B}_t\mathbf{w}_t\right), \tag{A.1}$$

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

$$\mathbf{B}_{t,s_{t,i},j} = \begin{cases} 1, & \text{if } i = j, \\ -\mathbf{B}_{t,s_{t,i}}[,k], & \text{if } s_{t,j} \text{ is the } k^{th} \text{ element in } N_t(s_{t,i}), \\ 0, & \text{Others.} \end{cases} \tag{A.2}$$

B Mean and Variance Specifications

The mean vector $\boldsymbol{\mu} = (\mu_1(\boldsymbol{s}_{1,1}), \ldots, \mu_1(\boldsymbol{s}_{1,n_1}), \ldots, \mu_T(s_{T,n_T}))$ is

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

$$+ \mathbf{h}_t^T(s_{t,k})\boldsymbol{\beta}_t + w_t(\boldsymbol{s}_{t,k}), \tag{B.1}$$

for $t = 1, \ldots T$, $i = 1, \ldots, n_t$. $\mathbf{1}_{\{\cdot\}}(\cdot)$ is the indicator function, and covariance matrix $\boldsymbol{\Lambda}$ is a block matrix with blocks $\Lambda^{(1,1)}, \ldots, \Lambda^{(1,T)}, \ldots, \Lambda^{(T,T)}$, 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:

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

$$+ \mathbf{1}_{\boldsymbol{s}_{t,k}=\boldsymbol{s}_{t,l}}(\boldsymbol{s}_{t,k},\boldsymbol{s}_{t,l})\tau_t^2,$$

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

$$\Lambda_{k,l}^{(t,t')} = \mathrm{cov}(z_t(\boldsymbol{s}_{t,k}), z_{t'}(\boldsymbol{s}_{t',l})|\cdot) = \sum_{i=1}^{\min(t,t')-1} \mathbf{1}_{\{\boldsymbol{s}_{t,k}, \boldsymbol{s}_{t',l} \notin \mathbf{S}_i\}}(\boldsymbol{s}_{t,k}, \boldsymbol{s}_{t',l}) \left\{ \prod_{j=i}^{\min(t,t')-1} \zeta_j(\boldsymbol{s}_{t,k})^T \zeta_j(\boldsymbol{s}_{t',l}) \right\}$$

$$\times C_i(\boldsymbol{s}_{t,k}, \boldsymbol{s}_{t',l}|\boldsymbol{\theta}_i) + \mathbf{1}_{\{\boldsymbol{s}_{t,k}, \boldsymbol{s}_{t',l} \notin \mathbf{S}_{\min(t,t')}\}}(\boldsymbol{s}_{t,k}, \boldsymbol{s}_{t',l}) C_{\min(t,t')}(\boldsymbol{s}_{t,k}, \boldsymbol{s}_{t',l}|\boldsymbol{\theta}_{\min(t,t')}),$$

$$\tag{B.2}$$

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

C   Alternative form of joint posterior distribution

The joint posterior density function of NNCGP for a T level system is:

$$p(\boldsymbol{\Theta}_{1:T}, \tilde{\mathbf{w}}_{1:T}|\mathbf{Z}_{1:T}) = \prod_{t=1}^{T} \left[p(\boldsymbol{\Theta}_t)\tilde{p}(\mathbf{w}_t|\boldsymbol{\theta}_t)\tilde{p}(\mathbf{w}_t^*|\mathbf{w}_t, \boldsymbol{\theta}_t) \right] L(\mathbf{Z}_{1:T}|\boldsymbol{\beta}_{1:T}, \tau_{1:T}^2, \boldsymbol{\gamma}_{1:T-1}, \tilde{\mathbf{w}}_{1:T}), \quad \text{(C.1)}$$

The conditional joint likelihood is

$$L(\mathbf{Z}_{1:T}|\boldsymbol{\beta}_{1:T}, \tau_{1:T}^2, \boldsymbol{\gamma}_{1:T-1}, \tilde{\mathbf{w}}_{1:T}) = \prod_{t=1}^{T}\prod_{i=1}^{n_t} N\left(z_t(\boldsymbol{s}_{t,i})|\mu_t(\boldsymbol{s}_{t,i}), \tau_t \right),$$

$$\mu_t(\boldsymbol{s}_{t,i}) = \mathbb{1}_{\{t \neq 1\}}(t) \times \sum_{j=1}^{t-1} \left\{ [\prod_{k=j}^{t-1} \zeta_k(\boldsymbol{s}_{t,i})] \times [\mathbf{h}_j(\boldsymbol{s}_{t,i})^T\boldsymbol{\beta}_j + w_j(\boldsymbol{s}_{t,i})] \right\} + \mathbf{h}_t(\boldsymbol{s}_{t,i})^T\boldsymbol{\beta}_t + w_t(\boldsymbol{s}_{t,i}).$$

$$\tag{C.2}$$

We assume $\mathbf{w}_t^*|\mathbf{w}_t$ are independent from each other so they can be updated individually. For locations $\boldsymbol{s}_u \in \mathbf{S}_t^*$, the full conditional posterior distribution of $w_t(\boldsymbol{s}_u) \sim N(V_t^*(\boldsymbol{s}_u)\mu_t^*(\boldsymbol{s}_u), V_t^*(\boldsymbol{s}_u))$ with

$$V_t^*(\boldsymbol{s}_u) = \Big[\sum_{q=t+1}^{T}(\prod_{i=t}^{q-1} I_{\{\boldsymbol{U}_{t,q}\}}(\boldsymbol{s}_u)\zeta_i^2(\boldsymbol{s}_u)\tau_q^{-2}) + F_{t,\boldsymbol{s}_u}^{-1} \Big]^{-1},$$

$$\mu_t^*(\boldsymbol{s}_u) = \sum_{q=t+1}^{T}\prod_{i=t}^{q-1} I_{\{\boldsymbol{U}_{t,q}\}}(\boldsymbol{s}_u)\zeta_i(\boldsymbol{s}_u)\tau_q^{-2}\Big(z_q(\boldsymbol{s}_u) - [\zeta_{q-1}(\boldsymbol{s}_u)y_{q-1}(\boldsymbol{s}_u) - \prod_{i=t}^{q-1} \zeta_i(\boldsymbol{s}_u)w_t(\boldsymbol{s}_u)] \Big)$$

$$+ F_{t,\boldsymbol{s}_u}^{-1}\mathbf{B}_{t,\boldsymbol{s}_u}\mathbf{w}_{t,N_t(\boldsymbol{s}_u)}, \quad t = 1, 2, \ldots, T-1, \tag{C.3}$$

where $I_{\{\boldsymbol{U}\}}(\boldsymbol{s}_u)$ is an indicator function with value 1 for location $\boldsymbol{s}_u \in \boldsymbol{U}$ and 0 for $\boldsymbol{s}_u \notin \boldsymbol{U}$, $\boldsymbol{U}_{A,B} = \mathbf{S}_A \cap \mathbf{S}_B$.

D Gibbs Sampler

The full conditional distribution of $\mathbf{w}_1$ is

$$w_1(\boldsymbol{s}_u)|\boldsymbol{\Theta}_1, \mathbf{Z}_1 \sim N(V_{\mathbf{w}_1}(\boldsymbol{s}_u)\mu_{\mathbf{w}_1}(\boldsymbol{s}_u), V_{\mathbf{w}_1}(\boldsymbol{s}_u)),$$
$$V_{\mathbf{w}_1}(\boldsymbol{s}_u) = \left(F_{1,\boldsymbol{s}_u}^{-1} + \tau_1^{-2}\right)^{-1},$$
$$\mu_{\mathbf{w}_1}(\boldsymbol{s}_u) = \tau_1^{-2}[z_1(\boldsymbol{s}_u) - \mathbf{h}_1^T(\boldsymbol{s}_u)\boldsymbol{\beta}_1] + F_{1,\boldsymbol{s}_u}^{-1}\mathbf{B}_{1,\boldsymbol{s}_u}\mathbf{w}_{1,N_1(\boldsymbol{s}_u)}, \tag{D.1}$$

for $s_u \in \mathbf{S}_1$.

With the specification of priors, the posterior distributions of the parameters are:

$$\boldsymbol{\beta}_t|\mathbf{w}_t, \tilde{\mathbf{y}}_{t-1}, \boldsymbol{\gamma}_{t-1}, \tau_t, \mathbf{Z}_t \sim N(\mathbf{V}_{\boldsymbol{\beta}_t}^*\boldsymbol{\mu}_{\boldsymbol{\beta}_t}^*, \mathbf{V}_{\boldsymbol{\beta}_t}^*),$$
$$\boldsymbol{\gamma}_t|\tilde{\mathbf{y}}_t, \boldsymbol{\beta}_{t+1}, \tau_{t+1}, \mathbf{Z}_{t+1} \sim N(\mathbf{V}_{\boldsymbol{\gamma}_t}^*\boldsymbol{\mu}_{\boldsymbol{\gamma}_t}^*, \mathbf{V}_{\boldsymbol{\gamma}_t}^*),$$
$$\sigma_t^2|\tilde{\mathbf{w}}_t, \boldsymbol{\phi}_t \sim IG(a_{\sigma_t}^*, b_{\sigma_t}^*),$$
$$\tau_t^2|\boldsymbol{\beta}_t, \mathbf{w}_t, \tilde{\mathbf{y}}_{t-1}, \mathbf{Z}_t, \boldsymbol{\gamma}_{t-1} \sim IG(a_{\tau_t}, b_{\tau_t}), \tag{D.2}$$

For $\boldsymbol{\beta}_t$, we have:

$$p(\boldsymbol{\beta}_t|\cdot) \propto N(\boldsymbol{\beta}_t|\boldsymbol{\mu}_{\boldsymbol{\beta}_t}, \mathbf{V}_{\boldsymbol{\beta}_t}) \times N(\mathbf{Z}_t|\mathbf{1}_{t>1}(t)\zeta_{t-1} \circ y_{t-1}(\mathbf{S}_t) + \boldsymbol{\delta}_t, \tau_t\mathbf{I})$$

$$\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_{\boldsymbol{\beta}_t})^T\mathbf{V}_{\boldsymbol{\beta}_t}^{-1}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_{\boldsymbol{\beta}_t})\right\} \times$$

$$\exp\left[-\frac{1}{2\tau_t^2}\left\{\mathbf{Z}_t - \mathbf{1}_{t>1}(t)\zeta_{t-1} \circ y_{t-1}(\mathbf{S}_t) - \boldsymbol{\delta}_t\right\}^T\left\{\mathbf{Z}_t - \mathbf{1}_{t>1}(t)\zeta_{t-1} \circ y_{t-1}(\mathbf{S}_t) - \boldsymbol{\delta}_t\right\}\right]$$

$$\propto \exp\left\{-\frac{1}{2}\left\{\mathbf{V}_{\boldsymbol{\beta}_t}^{-1} + \frac{1}{\tau_t^2}\mathbf{h}_t\mathbf{h}_t^T\right\}\boldsymbol{\beta}_t^T\boldsymbol{\beta}_t + \right.$$

$$\left.\left[\boldsymbol{\mu}_{\boldsymbol{\beta}_t}^T\mathbf{V}_{\boldsymbol{\beta}_t}^{-1} + \frac{1}{\tau_t^2}(\mathbf{Z}_t - \mathbf{1}_{t>1}(t)\zeta_{t-1} \circ y_{t-1}(\mathbf{S}_t) - \mathbf{w}_t)^T\mathbf{h}_t^T\right]\boldsymbol{\beta}_t\right\},$$

and we have:

$$\boldsymbol{\beta}_t|\mathbf{w}_t, \tilde{\mathbf{y}}_{t-1}, \boldsymbol{\gamma}_{t-1}, \tau_t, \mathbf{Z}_t \sim N(\mathbf{V}_{\boldsymbol{\beta}_t}^*\boldsymbol{\mu}_{\boldsymbol{\beta}_t}^*, \mathbf{V}_{\boldsymbol{\beta}_t}^*),$$

$$\boldsymbol{\mu}_{\boldsymbol{\beta}_t}^* = \boldsymbol{\mu}_{\boldsymbol{\beta}_t}^T\mathbf{V}_{\boldsymbol{\beta}_t}^{-1} + \frac{1}{\tau_t^2}(\mathbf{Z}_t - \mathbf{1}_{t>1}(t)\zeta_{t-1} \circ y_{t-1}(\mathbf{S}_t) - \mathbf{w}_t)^T\mathbf{h}_t^T,$$

$$\mathbf{V}_{\boldsymbol{\beta}_t}^* = \left(\mathbf{V}_{\boldsymbol{\beta}_t}^{-1} + \frac{1}{\tau_t^2}\mathbf{h}_t\mathbf{h}_t^T\right)^{-1}, \tag{D.3}$$

where $\mathbf{1}_{t>1}(t)$ is an indicator function equals 1 for $t > 1$, otherwise 0. The full conditional distribution for parameter σ_t^2 is:

$$p(\sigma_t^2|a_t, b_t, w_t(\tilde{\mathbf{S}}_t), \boldsymbol{\phi}_t) \propto IG(a_t, b_t) \times \tilde{p}(\mathbf{w}_t) \times \tilde{p}(w_t(\mathbf{S}_t^*)|\mathbf{w}_t).$$

From nearest neighbor Gaussian process approach, we have

$$\tilde{p}(\mathbf{w}_t) = \prod_{i=1}^{n_t} p(w_t(\boldsymbol{s}_{t,i})|\mathbf{w}_{t,N_t \boldsymbol{s}_{t,i}}),$$

$$w_t(\boldsymbol{s}_{t,i})|\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i})} \sim N(\mathbf{B}_{\boldsymbol{s}_{t,i}}\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i})}, F_{\boldsymbol{s}_{t,i}}),$$

where $\mathbf{B}_{\boldsymbol{s}_{t,i}} = \mathbf{C}_{\boldsymbol{s}_{t,i},N(\boldsymbol{s}_{t,i})}^{T}\mathbf{C}_{N(\boldsymbol{s}_{t,i})}^{-1}$, $\mathbf{F}_{\boldsymbol{s}_{t,i}} = \mathbf{C}_{\boldsymbol{s}_{t,i},s_{t,i}} - \mathbf{C}_{\boldsymbol{s}_{t,i},N(\boldsymbol{s}_{t,i})}^{T}\mathbf{C}_{N(\boldsymbol{s}_{t,i})}^{-1}\mathbf{C}_{\boldsymbol{s}_{t,i},N(\boldsymbol{s}_{t,i})}$, here $\mathbf{C}$ is the covariance matrix. Denote $\mathbf{C}_t(.,.|\sigma_t^2,\phi_t) = \sigma_t^2\boldsymbol{\rho}_t(.,.|\phi_t)$, we have:

$$\mathbf{F}_{\boldsymbol{s}_{t,i}} = \sigma_t^2(\boldsymbol{\rho}(\boldsymbol{s}_{t,i},\boldsymbol{s}_{t,i}) - \boldsymbol{\rho}_{\boldsymbol{s}_{t,i},N(\boldsymbol{s}_{t,i})}\boldsymbol{\rho}_{N(\boldsymbol{s}_{t,i})}^{-1}\boldsymbol{\rho}_{\boldsymbol{s}_{t,i},N(\boldsymbol{s}_{t,i})}) = \sigma_t^2\tilde{\mathbf{F}}_{\boldsymbol{s}_{t,i}},$$

and the following full conditional distribution for σ_t^2 is:

$$p(\sigma_t^2|a_t,b_t,\tilde{\mathbf{w}}_t,\boldsymbol{\phi}_t) \propto (\sigma_t^2)^{-a_t-1}\exp(-\frac{b_t}{\sigma_t^2}) \times \tilde{p}(\mathbf{w}_t) \times \tilde{p}(w_t(\mathbf{S}_t^*)|\mathbf{w}_t)$$

$$\propto (\sigma_t^2)^{-a_t-1-\frac{1}{2}(n_t+n_t^*)}\exp\Big[-\frac{1}{2\sigma_t^2}\sum_{i=1}^{n_t}\big\{w_t(\boldsymbol{s}_{t,i}) - \mathbf{B}_{\boldsymbol{s}_{t,i}}\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i})}\big\}^T (\tilde{F}_{\boldsymbol{s}_{t,i}})^{-1}\big\{w_t(s_{t,i}) - \mathbf{B}_{\boldsymbol{s}_{t,i}}\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i})}\big\}$$

$$-\frac{1}{2\sigma_t^2}\sum_{i=1}^{n_t^*}\big\{w_t(s_{t,i}^*) - \mathbf{B}_{\boldsymbol{s}_{t,i}^*}\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i}^*)}\big\}^T (\tilde{F}_{\boldsymbol{s}_{t,i}^*})^{-1}\big\{w_t(s_{t,i}^*) - \mathbf{B}_{\boldsymbol{s}_{t,i}^*}\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i}^*)}\big\} - \frac{b_t}{\sigma_t^2}\Big],$$

which is a $IG(a_t^*,b_t^*)$ distribution and

$$a_{\sigma_t}^* = a_t + \frac{1}{2}(n_t + n_t^*),$$

$$b_{\sigma_t}^* = b_t + \frac{1}{2}\sum_{i=1}^{n_t}\big\{w_t(\boldsymbol{s}_{t,i}) - \mathbf{B}_{\boldsymbol{s}_{t,i}}\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i})}\big\}^T (\tilde{F}_{\boldsymbol{s}_{t,i}})^{-1}\big\{w_t(\boldsymbol{s}_{t,i}) - \mathbf{B}_{\boldsymbol{s}_{t,i}}\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i})}\big\}$$

$$+ \frac{1}{2}\sum_{i=1}^{n_t^*}\big\{w_t(\boldsymbol{s}_{t,i}^*) - \mathbf{B}_{\boldsymbol{s}_{t,i}^*}\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i}^*)}\big\}^T (\tilde{F}_{\boldsymbol{s}_{t,i}^*})^{-1}\big\{w_t(\boldsymbol{s}_{t,i}^*) - \mathbf{B}_{\boldsymbol{s}_{t,i}^*}\mathbf{w}_{t,N_t(\boldsymbol{s}_{t,i}^*)}\big\}. \qquad \text{(D.4)}$$

For τ_t, we have the full conditional distribution for each level:

$$p(\tau_t|\cdot) = IG(\tau_t|c_t,d_t) \times \prod_{i=1}^{n_t} N(\mathbf{Z}_t|\cdot).$$

This structure gives us the inverse gamma distribution with:

$$p(\tau_t|\cdot) \sim IG(a_{\tau_t}^*,b_{\tau_t}^*)$$

$$a_{\tau_t}^* = c_t + \frac{1}{2}n_t,$$

$$b^*_{\tau_t} = d_t + \frac{1}{2}\sum_{u\in\mathbf{S}_t}(z_t(\boldsymbol{s}_u) - \zeta_{t-1}(\boldsymbol{s}_u)y_{t-1}(\boldsymbol{s}_u) - \delta_t(\boldsymbol{s}_u))^2. \tag{D.5}$$

For $\boldsymbol{\gamma}_t$, we have:

$$p(\boldsymbol{\gamma}_t|\cdot) \propto N(\boldsymbol{\gamma}_t|\boldsymbol{\mu}_{\boldsymbol{\gamma}_t}, \mathbf{V}_{\boldsymbol{\gamma}_t}) \times N(\mathbf{Z}_{t+1}|\mathbf{g}_t^T\boldsymbol{\gamma}_t y_t(\mathbf{S}_{t+1}) + \mathbf{h}_{t+1}\boldsymbol{\beta}_{t+1} + \mathbf{w}_{t+1}, \tau_{t+1}\mathbf{I})$$

$$\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma}_t - \boldsymbol{\mu}_{\boldsymbol{\gamma}_t})^T\mathbf{V}_{\boldsymbol{\gamma}_t}^{-1}(\boldsymbol{\gamma}_t - \boldsymbol{\mu}_{\boldsymbol{\gamma}_t})\right\}\times$$

$$\exp\left[-\frac{1}{2\tau_{t+1}^2}\left\{\mathbf{Z}_{t+1} - \mathbf{g}_t^T\boldsymbol{\gamma}_t y_t(\mathbf{S}_{t+1}) - \boldsymbol{\delta}_{t+1}\right\}^T\left\{\mathbf{Z}_{t+1} - \mathbf{g}_t^T\boldsymbol{\gamma}_t y_t(\mathbf{S}_{t+1}) - \boldsymbol{\delta}_{t+1}\right\}\right]$$

$$\propto \exp\left[-\frac{1}{2}\boldsymbol{\gamma}_t^T\mathbf{V}_{\boldsymbol{\gamma}_t}^{-1}\boldsymbol{\gamma}_t - \frac{1}{2\tau_{t+1}^2}y_t(\mathbf{S}_{t+1})^T y_t(\mathbf{S}_{t+1})(\mathbf{g}_t^T\boldsymbol{\gamma}_t)^T(\mathbf{g}_t^T\boldsymbol{\gamma}_t)+\right.$$

$$\left.\left\{\boldsymbol{\mu}_{\boldsymbol{\gamma}_t}^T\mathbf{V}_{\boldsymbol{\gamma}_t}^{-1} + \frac{1}{\tau_{t+1}^2}\left[(\mathbf{Z}_{t+1} - \boldsymbol{\delta}_{t+1})^T y_t(\mathbf{S}_{t+1})\right]\mathbf{g}_t^T\right\}\boldsymbol{\gamma}_t\right]$$

so that:

$$\boldsymbol{\gamma}_t|\tilde{\mathbf{y}}_t, \boldsymbol{\beta}_{t+1}, \tau_{t+1}, \mathbf{Z}_{t+1} \sim N(\mathbf{V}_{\boldsymbol{\gamma}_t}^*\boldsymbol{\mu}_{\boldsymbol{\gamma}_t}^*, \mathbf{V}_{\boldsymbol{\gamma}_t}^*),$$

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}_t}^* = \boldsymbol{\mu}_{\boldsymbol{\gamma}_t}^T\mathbf{V}_{\boldsymbol{\gamma}_t}^{-1} + \frac{1}{\tau_{t+1}^2}\left[(\mathbf{Z}_{t+1} - \boldsymbol{\delta}_{t+1})^T y_t(\mathbf{S}_{t+1})\right]\mathbf{g}_t^T,$$

$$\mathbf{V}_{\boldsymbol{\gamma}_t}^* = \left(\mathbf{V}_{\boldsymbol{\gamma}_t}^{-1} + \frac{1}{\tau_{t+1}^2}y_t(\mathbf{S}_{t+1})^T y_t(\mathbf{S}_{t+1})\mathbf{g}_t\mathbf{g}_t^T\right)^{-1}. \tag{D.6}$$

For a new input location $s_p \notin \tilde{\boldsymbol{S}}_t$, we have the predictive distribution of $w_t(\boldsymbol{s}_p)$:

$$w_t(\boldsymbol{s}_p) \sim N(V_{t,\boldsymbol{s}_p}\mu_{t,\boldsymbol{s}_p}, V_{t,\boldsymbol{s}_p}),$$

$$V_{t,\boldsymbol{s}_p} = (\tau_t^{-2} + \tilde{F}_{t,\boldsymbol{s}_p}^{-1})^{-1},$$

$$\mu_{t,\boldsymbol{s}_p} = \tau_t^{-2}[z_t(\boldsymbol{s}_p) - \mathbf{h}_t^T(\boldsymbol{s}_p)\boldsymbol{\beta}_t - \zeta_{t-1}(\boldsymbol{s}_p)y_{t-1}(\boldsymbol{s}_p)] + \tilde{F}_{t,\boldsymbol{s}_p}^{-1}\tilde{\mathbf{B}}_{t,\boldsymbol{s}_p}\tilde{\mathbf{w}}_{t,\tilde{N}_t(\boldsymbol{s}_p)}, \quad t = 1, 2, \ldots, T-1, \tag{D.7}$$

with $\tilde{\mathbf{B}}_{t,\boldsymbol{s}_p} = \mathbf{C}_{\boldsymbol{s}_p,\tilde{N}_t(\boldsymbol{s}_p)}^T\mathbf{C}_{\tilde{N}_t(\boldsymbol{s}_p)}^{-1}$, $\tilde{F}_{t,\boldsymbol{s}_p} = \mathbf{C}(\boldsymbol{s}_p, \boldsymbol{s}_p) - \mathbf{C}_{\boldsymbol{s}_p,\tilde{N}_t(\boldsymbol{s}_p)}^T\mathbf{C}_{\tilde{N}_t(\boldsymbol{s}_p)}^{-1}\mathbf{C}_{\boldsymbol{s}_p,\tilde{N}_t(\boldsymbol{s}_p)}$, $\tilde{N}_t(\boldsymbol{s}_p)$ is the m nearest neighbors in $\tilde{\boldsymbol{S}}_{t,<p}$, and $\tilde{\mathbf{w}}_{t,\tilde{N}_t(\boldsymbol{s}_p)}$ is the corresponding nearest neighbor subset of $\tilde{\mathbf{w}}_t$.

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. Deviance Information Criterion (DIC) and the effective number of parameters of the model(p_D) are defined as:

$$D(\theta) = -2\log(p(y|\theta)) + C,$$
$$p_D = \overline{D(\theta)} - D(\bar{\theta}),$$
$$\text{DIC} = p_D + \overline{D(\theta)}$$

It is used in Bayesian model selection. Models with smaller DIC and p_D are preferable.