Journal of the Indian Statistical Association
Vol.59 No. 2, 2021

# Modeling Large Multivariate Spatial Data with a Multivariate Fused Gaussian Process

**Emily L. Kang**

*University of Cincinnati*

**Miaoqi Li**

*Wells Fargo*

**Kerry Cawse-Nicholson**

*Jet Propulsion Laboratory, California Institute of Technology*

**Amy Braverman**

*Jet Propulsion Laboratory, California Institute of Technology*



**Abstract**

Large multivariate spatial data sets are common in environmental and climate sciences. This article proposes a flexible multivariate spatial statistical model for such data. Built upon Ma and Kang (2020), we model multivariate spatial processes with an additive form having two components that induce spatial dependence and a relationship between variables: One component is low-rank, and the other is multivariate spatial conditional autoregressive (CAR) structure. The resulting model not only allows for efficient computation of parameter estimates and spatial predictions, but is also flexible enough to describe potentially nonstationary and asymmetric spatial covariance and cross-covariance structures. We call the proposed model the multivariate fused Gaussian process (MFGP) model, and we investigate its performance through an extensive simulation study and a real-data example. The results show that, by borrowing information from complementary data, MFGP provides substantially improved spatial predictions compared to univariate models. We also demonstrate

Received : month year

that MFGP outperforms a multivariate model with only a low-rank component, or a multivariate CAR model with a separable covariance matrix. Supplementary Materials for this article, including the source code and results from additional numerical studies are also available.

**Key Words** : *Basis functions; cross-covariance function; Gaussian Markov random fields; Gaussian process; multivariate geostatistics; spatial prediction*



## 1 Introduction

Multivariate spatial data are ubiquitous in environmental and climate sciences. For example, remote sensing instruments provide observations of multiple geophysical processes interacting with each other [30]. Monitoring stations provide in-situ observations of many variables related to the environment and air quality [16]. It is now common to obtain massive spatial data sets that cover very large geographical regions or even the globe, often at very high spatial resolutions. The prevalence and societal importance of these large multivariate spatial data sets demands development of computationally efficient statistical models to analyze them.

Modeling spatial dependence structure for multivariate spatial processes is challenging, as it requires models that flexibly capture not only the spatial dependence within each variable, but also complicated relationships between variables, as defined through cross-covariance functions. Cross-covariance functions are usually difficult to specify as they must be nonnegative definite. [16, 14] present various ways to construct valid cross-covariances, including formulating a cross-covariance function from valid univariate covariance functions, and the linear model of coregionalization (LMC). [23] introduce the notion of spectral coherence for multivariate spatial processes, and discuss how some commonly used parametric cross-covariance functions can result in very different properties using this method. Alternatively, [9] suggest a conditional approach to build a multivariate spatial model that guarantees the validity of the resulting cross-covariance and is not necessarily

stationary or symmetric.

Although the aforementioned methods have been used in many studies, adapting them for large multivariate spatial data sets is complicated and nontrivial. Many developments in spatial statistics in the past decade have focused on models for univariate spatial processes to tackle the "big $n$" problem with large or massive data, including low-rank methods such as fixed rank kriging (FRK, [8]) and the predictive process [3], approximation methods that result in sparse matrices and thus efficient computation such as Lattice krig [32], the nearest neighbor Gaussian process (NNGP; [11]), the Vecchia approximation [22] and the meshed Gaussian process [33], and variations based on them including the full-scale approximation (FSA; [35]), the multi-resolution approximation (MRA; [20]), and the fused Gaussian process (FGP; [27]). Some of these methods have been extended to model multivariate Gaussian processes. [36] combine FSA and the LMC approach to model multivariate spatial data. [42] build upon the NNGP to formulate a Bayesian hierarchical model for large multivariate spatial data. [24] adopt the basis function representation in Lattice Krig for multivariate spatial processes. [30, 31] build low-rank statistical models based on the FRK approach to fuse large multivariate spatial and spatio-temporal data, which we refer to as multivariate FRK (MFRK) in this paper. [17] propose nonparametric spectral methods combined with LMC to efficiently estimate stationary multivariate spatial spectra from gridded data. Many of these computationally efficient models for multivariate spatial processes rely on the assumption of a specific parametric cross-covariance function, which is often stationary and symmetric, such as the multivariate Matérn cross-covariance models [16, 2].

Motivated by the FGP approach for univariate spatial processes [27], we propose a model for large multivariate spatial data. This model consists of two additive components: one component is in the low-rank basis-function representation, as in FRK, and doesn't require the assumption of a specific

parametric cross-covariance function. The other component is defined through the conditional approach suggested in [9], and uses multivariate spatial conditional autoregressive (CAR) models as building blocks to induce sparse matrices. For univariate spatial processes, [27] have shown that by adding the CAR-model component to the low-rank one, the resulting univariate FGP model substantially improves the predictive performance compared to one that uses the low-rank component alone, as in FRK. Unlike many methods assuming a specific parametric form of a stationary covariance function known up to a few parameters, FGP is flexible enough to provide good predictions even when the data present a nonstationary dependence structure. In this paper, we extend FGP to the context of multivariate modeling and call the resulting model the multivariate fused Gaussian process (MFGP). We will demonstrate that the MFGP model inherits the modelling flexibility and inferential benefits of FGP, and provides superior prediction performance without assuming stationarity or symmetry of the cross-covariance function.

The remainder of this article is organized as follows. Section 2 presents the MFGP model and discusses relevant model specifications and related methods in the literature. In Section 3, we give the derivation of likelihood-based inference, including parameter estimation and spatial prediction. An extensive simulation study is described in Section 4 to demonstrate the robustness of MFGP's predictive performance. In Section 5, we apply MFGP to large multivariate environmental data sets from an uncertainty quantification study in remote sensing. We conclude in Section 6 with a brief summary and discussion of possible future work, and proof of a proposition related to MFGP in Section 7. Additional numerical results and source code are available in the Supplementary Materials.

# 2 The Multivariate Fused Gaussian Process (MFGP) Model

In this section, we describe the MFGP model for $q$-variate spatial processes. We begin with the bivariate spatial process with $q = 2$, and then explain how it can be extended for $q > 2$.

## 2.1 Model Specification

Let $\mathbf{Y}(\mathbf{s}) = (Y_1(\mathbf{s}), \ldots, Y_q(\mathbf{s}))'$ and $\{\mathbf{Y}(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$ denote a hidden $q$-variate spatial process over spatial domain $\mathcal{D}$, where $\mathcal{D} \subset \mathcal{R}^d$. We are interested in making inferences on this hidden process from observations $\{Z_i(\mathbf{s}) : i = 1, \ldots, q, \mathbf{s} \in \mathcal{D}\}$, which include measurement errors:

$$Z_i(\mathbf{s}) = Y_i(\mathbf{s}) + \epsilon_i(\mathbf{s}); \quad i = 1, \ldots, q; \quad \mathbf{s} \in \mathcal{D}, \tag{2.1}$$

where $\{\epsilon_i(\cdot) : i = 1, \ldots, q\}$ represent independent Gaussian white noise with mean zero and variance $\sigma^2_{\epsilon,i}$, and for which we allow heterogeneous measurement-error variances across variables. As pointed out in [8, 30, 31], the variance parameters $\sigma^2_{\epsilon,i}$ can be inferred from validation data or instrument specification in remote sensing. If they are unknown, we can estimate $\sigma^2_{\epsilon,i}$ by fitting empirical semivariograms near the origin [19].

To model the hidden $q$-variate spatial process $\mathbf{Y}(\mathbf{s})$, we adopt the setup in [30] and assume:

$$Y_i(\mathbf{s}) = \mu_i(\mathbf{s}) + \nu_i(\mathbf{s}), \quad \mathbf{s} \in \mathcal{D}, \tag{2.2}$$

where $\mu_i(\cdot)$ denotes the spatial trend for the $i$th variable. In this paper, we model it as $\mu_i(\mathbf{s}) = \mathbf{X}_i(\mathbf{s})'\boldsymbol{\beta}_i$, where $\boldsymbol{\beta}_i$ is a $p_i$-dimensional vector of unknown coefficients for $p_i$ known covariates, $\mathbf{X}_i(\mathbf{s}) \equiv (X_i^1(\mathbf{s}), \ldots, X_i^{p_i}(\mathbf{s}))'$, for $i = 1, \ldots, q$. For the second term on the right-hand-side of (2.2), we follow the fused Gaussian process (FGP; [27]) structure and assume:

$$\nu_i(\mathbf{s}) = \mathbf{S}_i(\mathbf{s})'\boldsymbol{\eta}_i + \mathbf{A}(\mathbf{s})'\boldsymbol{\xi}_i, \tag{2.3}$$

where $\mathbf{S}_i(\mathbf{s}) \equiv (S_{i1}(\mathbf{s}), \ldots, S_{ir_i}(\mathbf{s}))'$ is defined through a set of $r_i$ known spatial basis functions $\{S_{ij}(\cdot) : j = 1, \ldots, r_i\}$, and $\boldsymbol{\eta}_i$ is the corresponding $r_i$-dimensional zero-mean Gaussian random vector with $var(\boldsymbol{\eta}_i) = \mathbf{K}_i$, for $i = 1, \ldots, q$. The term $\mathbf{S}_i(\mathbf{s})'\boldsymbol{\eta}_i$ is called the *low-rank component* in [27], as it follows the low-rank basis-function representation in FRK [8]. Multi-resolution local bisquare functions are suggested for these basis functions [8, 41], but other types of functions such as wavelets and cubic B-splines can also be used [39, 4, 7, 40].

The second term in (2.3) is called the Gaussian-graphical-model (GGM) component in FGP [27]. Similar to [30, 31, 27], we first assume that the spatial domain $\mathcal{D}$ is made up of a set of $N$ pre-specified and non-overlapping basic areal units (BAUs), $\mathcal{D} \equiv \mathcal{B}_1 \bigcup \mathcal{B}_2 \bigcup \cdots \bigcup \mathcal{B}_N$ and $\mathcal{B}_i \bigcap \mathcal{B}_j = \emptyset$ for $1 \leq i < j \leq N$. Let $\mathbf{s}_i$ denote the centroid of the $i$th BAU $\mathcal{B}_i$, for $i = 1, \ldots, N$. In practice, these BAUs can be specified based on the finest spatial resolution of scientific interest. Then, $\boldsymbol{\xi}_i$ is an $N$-dimensional Gaussian random vector corresponding to these $N$ BAUs for the $i$th variable, $i = 1, \ldots, q$. The $N$-dimensional vector $\mathbf{A}(\mathbf{s}) \equiv (A_1(\mathbf{s}), \ldots, A_N(\mathbf{s}))'$ maps a spatial location $\mathbf{s}$ to the corresponding BAU with $A_j(\mathbf{s}) = \mathbb{1}_{\mathbf{s} \in \mathcal{B}_j}$ for $j = 1, \ldots, N$ and $i = 1, \ldots, q$, where $\mathbb{1}_{\mathbf{s} \in \mathcal{B}_j}$ is the indicator function equal to 1 if $\mathbf{s}$ is in the $j$th BAU $\mathcal{B}_j$ and 0 otherwise. If we need to interpolate between the BAUs, we can specify $\mathbf{A}(\cdot)$ to be piecewise linear basis functions relating $\mathbf{s}$ to the centroids $\{\mathbf{s}_i : i = 1, \ldots, N\}$ as suggested in [26]. [27] assume that the $N$-dimensional random vector $\boldsymbol{\xi}_i$ can be represented as an undirected Gaussian graphical model, and treat the spatial conditional autoregressive (CAR) model as a special case in numerical examples.

To generalize the FGP model for multivariate spatial processes, we maintain the assumption in [27] that the two components $\{\boldsymbol{\eta}_i\}$ and $\{\boldsymbol{\xi}_i\}$ are independent but will introduce dependence across different variables, $i = 1, \ldots, q$. Specifically, for the low-rank component, we assume $Cov(\boldsymbol{\eta}_i, \boldsymbol{\eta}_j) = \mathbf{K}_{ij}$, for $i, j = 1, \ldots, q$, and $i \neq j$. As in [30], we don't assume a

specific parametric form for the $r_i \times r_j$ cross covariance matrix $\mathbf{K}_{ij}$, but choose to estimate it directly using the expectation-maximization (EM) algorithm described in Section 3. We expect this semiparametric form to inherit the flexibility shown in [27] for describing nonstationary spatial dependence within and between distinct variables. For the GGM component, we need to generalize the CAR model to the multivariate context. [18] proposed a class of multivariate spatial conditional autoregressive models by generalizing the model in [29], and specifying simpler conditional and marginal subsidiary spatial models. Alternatively, [9] suggest a conditional approach for multivariate modeling. In this paper, we combine these two methods to model $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_q\}$.

We first describe this model for the bivariate case with $q = 2$ and explain in Section 2.2 further the specification of the model when $q > 2$. With $q = 2$, we assume:

$$\boldsymbol{\xi}_1 \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_1) \tag{2.4}$$

$$\boldsymbol{\xi}_2|\boldsymbol{\xi}_1 \sim MVN(\mathbf{P}_{2,1}\boldsymbol{\xi}_1, \boldsymbol{\Sigma}_2), \tag{2.5}$$

where $MVN(\boldsymbol{\mu}, \mathbf{B})$ denotes the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{B}$. Both distributions in (2.4) and (2.5) are assumed to be spatial conditional autoregressive (CAR) models where the precision matrices are $\mathbf{Q}_i \equiv \boldsymbol{\Sigma}_i^{-1} = \tau_i^{-2}(\mathbf{I} - \gamma_i \mathbf{H})$; $i = 1, 2$. Here, $\tau_i^2 > 0$ is a scale parameter, representing the conditional variance; $\gamma_i$ is interpreted as the strength of spatial dependence; and $\mathbf{H} \equiv (h_{ij})$ is a known $N \times N$ sparse proximity matrix with zero diagonal elements. In the numerical examples in this paper, $\mathbf{H}$ is constructed based on the first-order neighborhood structure. In (2.5), $\mathbf{P}_{2,1}$ is an $N \times N$ matrix representing how the conditional mean $E(\boldsymbol{\xi}_2|\boldsymbol{\xi}_1)$ is related to the elements in $\boldsymbol{\xi}_1$. Following [18], we assume a parsimonious parametric form $\mathbf{P}_{2,1} = \alpha_{2,1,0}\mathbf{I} + \alpha_{2,1,1}\mathbf{H}$, where $\alpha_{2,1,0}$ and $\alpha_{2,1,1}$ are two parameters describing how the conditional mean is related to the element in $\boldsymbol{\xi}_1$ in the same BAU, and those at neighboring BAUs,

respectively.

Define $\boldsymbol{\xi} \equiv (\boldsymbol{\xi}_1', \boldsymbol{\xi}_2')'$. It is straightforward to show that $\boldsymbol{\xi} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\xi)$, with

$$\boldsymbol{\Sigma}_\xi^{-1} = \begin{pmatrix} \mathbf{Q}_1 + \mathbf{P}_{2,1}'\mathbf{Q}_2\mathbf{P}_{2,1} & -\mathbf{P}_{2,1}'\mathbf{Q}_2 \\ -\mathbf{Q}_2\mathbf{P}_{2,1} & \mathbf{Q}_2 \end{pmatrix} := \mathbf{M}.$$

Further, we can rewrite the precision matrix $\mathbf{M}$ as follows:

$$\mathbf{M} = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{P}_{2,1}'\mathbf{Q}_2 \\ \mathbf{0} & -\mathbf{Q}_2 \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{P}_{2,1} & -\mathbf{I} \end{pmatrix}.$$

The determinant of $\mathbf{M}$ can be easily obtained by calculating the determinants of two $N \times N$ sparse matrices instead of dealing with a $(2N) \times (2N)$ matrix, i.e., $|\mathbf{M}| = |\mathbf{Q}_1| \times |\mathbf{Q}_2|$.

Let $\mathbf{Y}_i = (Y_i(\mathbf{s}_1), \ldots, Y_i(\mathbf{s}_N))'$ for $i = 1, 2$, and $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')'$, representing the bivariate process at all the $N$ BAUs. Based on the models in (2.2), (2.3), (2.4), and (2.5), we have the following model for the bivariate process:

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{pmatrix}. \tag{2.6}$$

For simplicity, we rewrite equation (2.6) as follows:

$$\underset{(2N)\times 1}{\mathbf{Y}} = \underset{(2N)\times(p_1+p_2)}{\mathbf{X}} \underset{(p_1+p_2)\times 1}{\boldsymbol{\beta}} + \underset{(2N)\times(r_1+r_2)}{\mathbf{S}} \underset{(r_1+r_2)\times 1}{\boldsymbol{\eta}} + \underset{(2N)\times 1}{\boldsymbol{\xi}}, \tag{2.7}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$; $\mathbf{S} = blockdiag(\mathbf{S}_1, \mathbf{S}_2)$; $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2')' \sim MVN(\mathbf{0}, \mathbf{K})$ with

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_1 & \mathbf{K}_{12} \\ \mathbf{K}_{12}' & \mathbf{K}_2 \end{pmatrix}.$$

We call this model the *multivariate fused Gaussian process* (MFGP). The spatial covariance and cross-covariances within and between $Y_1(\cdot)$ and $Y_2(\cdot)$

can be derived analogously:

$$\begin{aligned}
\mathrm{Cov}\{Y_1(\mathbf{s}_i), Y_1(\mathbf{s}_j)\} &= \mathbf{S}_1(\mathbf{s}_i)\mathbf{K}_1\mathbf{S}_1(\mathbf{s}_j)' + [\tau_1^2(\mathbf{I}-\gamma_1\mathbf{H})^{-1}]_{ij}; && (2.8)\\
\mathrm{Cov}\{Y_2(\mathbf{s}_i), Y_2(\mathbf{s}_j)\} &= \mathbf{S}_2(\mathbf{s}_i)\mathbf{K}_2\mathbf{S}_2(\mathbf{s}_j)' + [\tau_1^2(\alpha_{2,1,0}\mathbf{I}+\alpha_{2,1,1}\mathbf{H})(\mathbf{I}-\gamma_1\mathbf{H})^{-1} \\
&\qquad (\alpha_{2,1,0}\mathbf{I}+\alpha_{2,1,1}\mathbf{H}) + \tau_2^2(\mathbf{I}-\gamma_2\mathbf{H})^{-1}]_{ij}; && (2.9)\\
\mathrm{Cov}\{Y_1(\mathbf{s}_i), Y_2(\mathbf{s}_j)\} &= \mathbf{S}_1(\mathbf{s}_i)\mathbf{K}_{12}\mathbf{S}_2(\mathbf{s}_j)' + [\tau_1^2(\mathbf{I}-\gamma_1\mathbf{H})^{-1}(\alpha_{2,1,0}\mathbf{I}+\alpha_{2,1,1}\mathbf{H})]_{ij}, && (2.10)
\end{aligned}$$

where $[\mathbf{B}]_{ij}$ represents the $(i,j)$th element in the matrix $\mathbf{B}$. From (2.8), (2.9), and (2.10) above, we can see that the resulting cross-covariance is not necessarily stationary or symmetric. Furthermore, we can obtain the covariance matrix of $\mathbf{Y}$:

$$\begin{aligned}
\boldsymbol{\Sigma}_Y &= \mathrm{Cov}(\mathbf{Y}) = \mathbf{SKS}' + \boldsymbol{\Sigma}_\xi \\
&= \begin{pmatrix} \mathbf{S}_1\mathbf{K}_1\mathbf{S}_1' + \boldsymbol{\Sigma}_1 & \mathbf{S}_1\mathbf{K}_{12}\mathbf{S}_2' + \boldsymbol{\Sigma}_1\mathbf{P}_{2,1}' \\ \mathbf{S}_2\mathbf{K}_{12}'\mathbf{S}_1' + \mathbf{P}_{2,1}\boldsymbol{\Sigma}_1 & \mathbf{S}_2\mathbf{K}_2\mathbf{S}_2' + \mathbf{P}_{2,1}\boldsymbol{\Sigma}_1\mathbf{P}_{2,1}' + \boldsymbol{\Sigma}_2 \end{pmatrix}. \quad (2.11)
\end{aligned}$$

In practice, we may not observe data at all $N$ BAUs; we may only have observations at some of them. Suppose we have observations from the bivariate process $\mathbf{Z} \equiv (\mathbf{Z}_1', \mathbf{Z}_2')'$ with $\mathbf{Z}_i = (Z(\mathbf{s}_{o,1}^i), \ldots, Z(\mathbf{s}_{o,n_i}^i))'$, where $\mathbf{s}_{o,i}^j$ denotes the $i$th observation location for the $j$th variable, for $i = 1, \ldots, n_j$ and $j = 1, 2$. Let $\mathbf{A}$ denote the $(n_1 + n_2) \times (2N)$ incident matrix to relate the these $(n_1 + n_2)$ observation locations to the $N$ BAUs:

$$\mathbf{A} = \begin{pmatrix} \underset{n_1 \times N}{\mathbf{A}_1} & \mathbf{0} \\ \mathbf{0} & \underset{n_2 \times N}{\mathbf{A}_2} \end{pmatrix},$$

where the $j$th row of $\mathbf{A}_i$ is the vector $\mathbf{A}(\mathbf{s}_{o,j}^i)'$. Recall that the $l$th element in the $N$-dimensional vector $\mathbf{A}(\mathbf{s})$ is equal to 1 if $\mathbf{s}$ is in the $l$th BAU and 0 otherwise. Note that $\mathbf{A}_1$ and $\mathbf{A}_2$ do not need to be the same, thus the observations from the two variables are not necessarily aligned. Combining this with the model for $\mathbf{Y}$ in (2.7), we then have the model for the data vector $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{AY} + \boldsymbol{\epsilon} = \mathbf{AX}\boldsymbol{\beta} + \mathbf{AS}\boldsymbol{\eta} + \mathbf{A}\boldsymbol{\xi} + \boldsymbol{\epsilon}, \quad (2.12)$$

where $\boldsymbol{\epsilon} \equiv (\boldsymbol{\epsilon}_1', \boldsymbol{\epsilon}_2')'$ represents the measurement errors distributed as $MVN(\mathbf{0}, \mathbf{V}_\epsilon)$ with $\mathbf{V}_\epsilon \equiv blockdiag(\sigma^2_{\epsilon,1}\mathbf{I}, \sigma^2_{\epsilon,2}\mathbf{I})$. Therefore, we have $\mathbf{Z} \sim MVN(\mathbf{AX}\boldsymbol{\beta}, \boldsymbol{\Sigma}_Z)$, where $\boldsymbol{\Sigma}_Z = \mathbf{A}\boldsymbol{\Sigma}_Y\mathbf{A}' + \mathbf{V}_\epsilon$. This completes the model of the data vector. The unknown parameters are $\{\boldsymbol{\beta}, \mathbf{K}, \gamma_1, \tau_1^2, \alpha_{2,1,0}, \alpha_{2,1,1}, \gamma_2, \tau_2^2\}$. We discuss how to estimate these parameters and infer the hidden process $\mathbf{Y}(\mathbf{s})$ in Section 3.

## 2.2 MFGP with $q > 2$

We explain how the MFGP model is formulated when $q > 2$. For the trend term, we assume $\mu_i(\cdot) = \mathbf{X}_i(\cdot)'\boldsymbol{\beta}_i$ for the $i$th variable, for $i = 1, \ldots, q$. The matrix $\mathbf{X}$ in (2.7) will become a $(qN) \times (\sum_{i=1}^q p_i)$ block diagonal matrix $\mathbf{X} = blockdiag(\mathbf{X}_1, \ldots, \mathbf{X}_q)$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_q')'$. For the low-rank component, extension to more than two variables is straightforward: We have $\mathbf{S} = blockdiag(\mathbf{S}_1, \ldots, \mathbf{S}_q)$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_q')'$, and the matrix $\mathbf{K} = var(\boldsymbol{\eta})$ is formulated with the blocks, $\mathbf{K}_{ij} = cov(\boldsymbol{\eta}_i, \boldsymbol{\eta}_j)$ and $\mathbf{K}_i = var(\boldsymbol{\eta}_i)$ for $i, j = 1, \ldots, q$ and $i \neq j$. For the GGM component, we use the conditional approach to build the model for $\boldsymbol{\xi} = (\boldsymbol{\xi}_1', \ldots, \boldsymbol{\xi}_q')'$. For example, when $q = 3$, we have:

$$\boldsymbol{\xi}_1 \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_1) \tag{2.13}$$

$$\boldsymbol{\xi}_2|\boldsymbol{\xi}_1 \sim MVN(\mathbf{P}_{2,1}\boldsymbol{\xi}_1, \boldsymbol{\Sigma}_2), \tag{2.14}$$

$$\boldsymbol{\xi}_3|\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \sim MVN(\mathbf{P}_{3,1}\boldsymbol{\xi}_1 + \mathbf{P}_{3,2}\boldsymbol{\xi}_2, \boldsymbol{\Sigma}_3), \tag{2.15}$$

where we assume $\mathbf{P}_{i,j} = \alpha_{i,j,0}\mathbf{I} + \alpha_{i,j,1}\mathbf{H}$ for $i = 2, 3$ and $j = 1, \ldots, i-1$; the precision matrices $\mathbf{Q}_i \equiv \boldsymbol{\Sigma}_i^{-1} = \tau_i^{-2}(\mathbf{I} - \gamma_i\mathbf{H})$, for $i = 1, 2, 3$. It can be shown that the precision matrix of $\boldsymbol{\xi}$ can be written as,

$$\begin{aligned}
\mathbf{M} &= \begin{pmatrix} \mathbf{Q}_1 + \mathbf{P}_{2,1}'\mathbf{Q}_2\mathbf{P}_{2,1} + \mathbf{P}_{3,1}'\mathbf{Q}_3\mathbf{P}_{3,1} & \mathbf{P}_{3,1}'\mathbf{Q}_3\mathbf{P}_{3,2} - \mathbf{P}_{2,1}'\mathbf{Q}_2 & -\mathbf{P}_{3,1}'\mathbf{Q}_3 \\ (\mathbf{P}_{3,1}'\mathbf{Q}_3\mathbf{P}_{3,2} - \mathbf{P}_{2,1}'\mathbf{Q}_2)' & \mathbf{Q}_2 + \mathbf{P}_{3,2}'\mathbf{Q}_3\mathbf{P}_{3,2} & -\mathbf{P}_{3,2}'\mathbf{Q}_3 \\ -(\mathbf{P}_{3,1}'\mathbf{Q}_3)' & -(\mathbf{P}_{3,2}'\mathbf{Q}_3)' & \mathbf{Q}_3 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{Q}_1 & -\mathbf{P}_{2,1}' & \mathbf{P}_{3,1}' \\ \mathbf{0} & \mathbf{I} & \mathbf{P}_{3,2}' \\ \mathbf{0} & \mathbf{0} & -\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ -\mathbf{Q}_2'\mathbf{P}_{2,1} & \mathbf{Q}_2 & \mathbf{0} \\ \mathbf{Q}_3'\mathbf{P}_{3,1} & \mathbf{Q}_3'\mathbf{P}_{3,2} & -\mathbf{Q}_3 \end{pmatrix},
\end{aligned}$$

and so $|\mathbf{M}| = |\mathbf{Q}_1| \cdot |\mathbf{Q}_2| \cdot |\mathbf{Q}_3|$.

When $q > 3$, we model $\boldsymbol{\xi}$ in a manner similar to (2.13), (2.14), and (2.15). We also derive the following proposition whose proof is given in Section 7.

**Proposition 2.1** *Consider a $q$-variate spatial processes. Recall that $\boldsymbol{M}$ denotes the $(qN) \times (qN)$ precision matrix of $\boldsymbol{\xi} = (\boldsymbol{\xi}_1', \ldots, \boldsymbol{\xi}_q')'$. Under the model specification of MFGP described above, we have:*

$$|\boldsymbol{M}| = |\boldsymbol{Q}_1| \cdot |\boldsymbol{Q}_2| \cdot \ \cdots \ \cdot |\boldsymbol{Q}_q|.$$

When we use the MFGP model for a $q$-variate spatial process, the unknown parameters are $\boldsymbol{\beta}$, $\mathbf{K}$, $\{\gamma_i : i = 1, \ldots, q\}$, $\{\tau_i^2 : i = 1, \ldots, q\}$, $\{\alpha_{i,j,k} : i = 2, \ldots, q; j = 1, \ldots, i-1; k = 0, 1\}$. Therefore, as $q$ increases, the number of parameters increase as $O(q^2)$. When $q$ is large, the $q$-variate spatial processes is called a *highly-multivariate spatial processes* [12, 25]. The original parameterization of MFGP can result in a large number of parameters for such highly-multivariate spatial processes, and this may pose computational difficulties for parameter estimation. In Section 6 we discuss possible ways to extend the MFGP model for highly-multivariate spatial processes in view of this.

## 2.3 Alternative Model Specifications and Related Existing Methods

In specifying the distribution of $\boldsymbol{\xi} = (\boldsymbol{\xi}_1', \ldots, \boldsymbol{\xi}_q')'$ in the MFGP model, it is also possible to use the multivariate CAR model suggested in [13] and [5], as an alternative to the specification above. When $q = 2$ this model is,

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{pmatrix} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\xi) \text{ with } \boldsymbol{\Sigma}_\xi = \boldsymbol{\Gamma} \otimes \mathbf{Q}^{-1}, \tag{2.16}$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \text{ and } \mathbf{Q} = \tau^{-2}(\mathbf{I} - \gamma \mathbf{H}).$$

Here, the $(2N) \times (2N)$ covariance matrix $\boldsymbol{\Sigma}_\xi$ takes a separable form as the Kronecker product of a $2 \times 2$ correlation matrix, with $\rho$ representing the

correlation across different variables, and an $N \times N$ spatial covariance matrix from a univariate CAR model. The quantity $\gamma$ is the spatial dependence parameter, and $\tau^2$ is the conditional variance. Compared to the model in (2.4) and (2.5), (2.16) contains fewer parameters, but assumes the same correlation across variables at all spatial locations (i.e., the same $\rho$ at all locations), and the same spatial dependence structure for all variables (i.e., $\mathbf{Q}^{-1}$ as the same covariance matrix for all $\boldsymbol{\xi}_i$'s, $i = 1, \ldots, q$). This may not be realistic in practice. In this paper, we call the resulting model of $\mathbf{Y}$ based on this multivariate CAR model for $\boldsymbol{\xi}$ the separable fused Gaussian process (SFGP).

The MFGP model is closely related to the data fusion model in [30, 31], which is built upon the framework of the fixed rank kriging [8]. However, there the model assumes that elements in $\boldsymbol{\xi}_i$ are Gaussian white noise. MFGP allows them to have spatial dependence both within and between $\boldsymbol{\xi}_i$, $i = 1, \ldots, q$. In numerical examples presented in Section 4 and Section 5, we include results based on the original, FRK-based data fusion model in our comparisons, and call it MFRK for short, in this paper.

Our MFGP model is an extension of the univariate FGP model presented in [27]. One may reasonably ask whether MFGP performs better than simple application of univariate FGP to all variables independently. Call the latter *independent FGP* (IFGP). In the numerical examples in Section 4 we show that MFGP does in fact provide improved predictions compared to IFGP by borrowing strength across variables.

## 3 Inference

We adapt the Expectation-Maximization (EM) algorithm used in [21, 27] to estimate parameters in MFGP. Without loss of generality, we describe this EM algorithm and how to make spatial prediction with $q = 2$, but the extension to $q > 2$ is straightforward.

For the bivariate MFGP with $q = 2$, the unknown parameters are $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \mathbf{K}, \gamma_1, \gamma_2, \tau_1^2, \tau_2^2, \alpha_{2,1,0}, \alpha_{2,1,1}\}$. We treat $\boldsymbol{\eta}$ as latent variables and devise the EM algorithm to minimize the twice-negative-marginal-log-likelihood function of the data, $\mathbf{Z}$:

$$-2\ln L(\boldsymbol{\theta}, \mathbf{Z}) = \ln|\boldsymbol{\Sigma}_Z| + (\mathbf{Z} - \mathbf{A}\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_Z^{-1}(\mathbf{Z} - \mathbf{A}\mathbf{X}\boldsymbol{\beta}) + constant. \quad (3.1)$$

The conditional distribution of $\boldsymbol{\eta}$ given data $\mathbf{Z}$ can be shown to be a multivariate normal distribution with conditional mean given in (3.2) and conditional variance-covariance matrix given in (3.3):

$$\boldsymbol{\mu}_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}} \equiv \mathbb{E}(\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\theta}) = (\mathbf{A}\mathbf{S}\mathbf{K})'\boldsymbol{\Sigma}_Z^{-1}(\mathbf{Z} - \mathbf{A}\mathbf{X}\boldsymbol{\beta}), \quad (3.2)$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}} \equiv \mathrm{Var}(\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\theta}) = \mathbf{K} - (\mathbf{A}\mathbf{S}\mathbf{K})'\boldsymbol{\Sigma}_Z^{-1}(\mathbf{A}\mathbf{S}\mathbf{K}). \quad (3.3)$$

Furthermore, based on the assumptions in the MFGP model, it is straightforward to show that $\mathbf{Z}|\boldsymbol{\eta} \sim MVN(\mathbf{A}(\mathbf{X}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\eta}), \mathbf{A}(\mathbf{M}^{-1} + \mathbf{V}_\epsilon)\mathbf{A}')$. The corresponding twice-negative-complete-log-likelihood function is,

$$\begin{aligned} -2\ln L_c(\boldsymbol{\eta}, \mathbf{Z}) &= \ln|\mathbf{D}^{-1}| + (\mathbf{Z} - \mathbf{A}\mathbf{X}\boldsymbol{\beta} - \mathbf{A}\mathbf{S}\boldsymbol{\eta})'\mathbf{D}(\mathbf{Z} - \mathbf{A}\mathbf{X}\boldsymbol{\beta} - \mathbf{A}\mathbf{S}\boldsymbol{\eta}) \\ &\quad + \ln|\mathbf{K}| + \boldsymbol{\eta}'\mathbf{K}^{-1}\boldsymbol{\eta}, \end{aligned} \quad (3.4)$$

where $\mathbf{D} = (\mathbf{A}\mathbf{M}^{-1}\mathbf{A}' + \mathbf{V}_\epsilon)^{-1}$.

In the expectation step (E-step), we derive the expected value of $-2\ln L_c(\boldsymbol{\eta}, \mathbf{Z})$ with respect to the conditional distribution of latent variables $\boldsymbol{\eta}$ given data and all other of parameters. It is denoted by $-2Q(\boldsymbol{\theta}; \boldsymbol{\theta}_l)$ and given in (3.5), where $\boldsymbol{\theta}_l$ denotes the parameter estimates in the $l$th iteration of the EM algorithm. In the maximization step (M-step), we update $\boldsymbol{\theta}$ by maximizing $Q(\boldsymbol{\theta}; \boldsymbol{\theta}_l)$, or, equivalently, minimizing $-2Q(\boldsymbol{\theta}; \boldsymbol{\theta}_l)$ with respect to $\boldsymbol{\theta}$. Specifically, it is straightforward to show that $\widehat{\boldsymbol{\beta}}_l$ and $\widehat{\mathbf{K}}_l$ can be updated in closed form, as shown in (3.6) and (3.7), respectively. To update the other parameters, $\{\tau_i^2\}_{i=1}^2$, $\{\gamma_i\}_{i=1}^2$, and $\{\alpha_{2,1,0}, \alpha_{2,1,1}\}$, we need to minimize $-2Q(\boldsymbol{\theta}; \boldsymbol{\theta}_l)$, or equivalently, minimize the function $f(\cdot)$ given in

(3.8), numerically. We use the function `fmincon` in Matlab to perform this minimization in the numerical examples shown later in this paper.

$$\begin{aligned} -2Q(\boldsymbol{\theta};\boldsymbol{\theta}_l) &= E_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}_l}\left[-2\ln L_c(\boldsymbol{\eta},\mathbf{Z})\right] \\ &= \ln|\mathbf{D}^{-1}| + \ln|\mathbf{K}| + (\mathbf{Z}-\mathbf{AX}\boldsymbol{\beta})'\mathbf{D}(\mathbf{Z}-\mathbf{AX}\boldsymbol{\beta}) \\ &\quad - 2(\mathbf{Z}-\mathbf{AX}\boldsymbol{\beta})'\mathbf{DAS}\boldsymbol{\mu}_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}_l} + tr\left\{\left[(\mathbf{AS})'\mathbf{D}(\mathbf{AS})+\mathbf{K}^{-1}\right]\boldsymbol{\Sigma}_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}_l}\right\} \\ &\quad + \boldsymbol{\mu}'_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}_l}\left[(\mathbf{AS})'\mathbf{D}(\mathbf{AS})+\mathbf{K}^{-1}\right]\boldsymbol{\mu}_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}_l} \end{aligned} \tag{3.5}$$

$$\widehat{\boldsymbol{\beta}}_{l+1} = \left[(\mathbf{AX})'\mathbf{D}(\mathbf{AX})\right]^{-1}(\mathbf{AX})'\mathbf{D}\left[\mathbf{Z}-(\mathbf{AS})\boldsymbol{\mu}_{\boldsymbol{\eta}|\boldsymbol{\theta}_l,\mathbf{Z}}\right] \tag{3.6}$$

$$\widehat{\mathbf{K}}_{l+1} = \boldsymbol{\Sigma}_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}_l} + \boldsymbol{\mu}_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}_l}\boldsymbol{\mu}'_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}_l} \tag{3.7}$$

$$\begin{aligned} & f(\tau_1^2,\tau_2^2,\gamma_1,\gamma_2,\alpha_{2,1,0},\alpha_{2,1,1}) \\ = \ & \ln|\mathbf{D}^{-1}| + \left[\mathbf{Z}-(\mathbf{AX})\widehat{\boldsymbol{\beta}}_{l+1}\right]'\mathbf{D}\left[\mathbf{Z}-(\mathbf{AX})\widehat{\boldsymbol{\beta}}_{l+1}\right] \\ & - 2\left[\mathbf{Z}-(\mathbf{AX})\widehat{\boldsymbol{\beta}}_{l+1}\right]'\mathbf{DAS}\boldsymbol{\mu}_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}_l} + tr\left\{(\mathbf{AS})'\mathbf{D}(\mathbf{AS})\boldsymbol{\Sigma}_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}_l}\right\} \\ & + \boldsymbol{\mu}'_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}_l}\left[(\mathbf{AS})'\mathbf{D}(\mathbf{AS})\right]\boldsymbol{\mu}_{\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\theta}_l}. \end{aligned} \tag{3.8}$$

Recall that the $(n_1+n_2)\times(n_1+n_2)$ matrix $\mathbf{D}=(\mathbf{AM}^{-1}\mathbf{A}'+\mathbf{V}_\epsilon)^{-1}$. In the EM algorithm we need to evaluate $|\mathbf{D}^{-1}|$ and calculate $\mathbf{D}$. When $n_1+n_2$ is large, this cannot be done directly. Instead, we apply Sylvester's determinant identity [1] and the Sherman-Morrison-Woodbury formula [8]:

$$|\mathbf{D}^{-1}| = |\mathbf{M}+\mathbf{A}'\mathbf{V}_\epsilon^{-1}\mathbf{A}|\cdot|\mathbf{M}^{-1}|\cdot|\mathbf{V}_\epsilon|, \tag{3.9}$$

$$\mathbf{D} = \mathbf{V}_\epsilon^{-1} - \mathbf{V}_\epsilon^{-1}\mathbf{A}[\mathbf{M}+\mathbf{A}'\mathbf{V}_\epsilon^{-1}\mathbf{A}]^{-1}\mathbf{A}'\mathbf{V}_\epsilon^{-1}. \tag{3.10}$$

Hence, evaluation of (3.8), (3.9), and (3.10) involves solving systems of linear equation or calculating $\mathbf{x}_1 = [\mathbf{M}+\mathbf{A}'\mathbf{V}_\epsilon^{-1}\mathbf{A}]^{-1}\mathbf{a}_1$ and calculating $|\mathbf{M}+\mathbf{A}'\mathbf{V}_\epsilon^{-1}\mathbf{A}|$, where $\mathbf{a}_1$ denotes a $2N$-dimensional vector, and $\mathbf{M}+\mathbf{A}'\mathbf{V}_\epsilon^{-1}\mathbf{A}$ is a sparse $(2N)\times(2N)$ matrix. Therefore, both (3.9) and (3.10) can be calculated efficiently. Furthermore, note that to evaluate $|\mathbf{M}^{-1}|$ in (3.9), we use the result from Proposition 2.1 and thus have $|\mathbf{M}^{-1}| = 1/|\mathbf{M}| = 1/(|\mathbf{Q}_1|\cdot|\mathbf{Q}_2|)$.

Also, note that the matrix $\mathbf{\Sigma}_Z^{-1}$ is needed in (3.2) and (3.3). To calculate it efficiently, we use the Sherman-Morrison-Woodbury formula again:

$$\begin{aligned}\mathbf{\Sigma}_Z^{-1} &= [\mathbf{D}^{-1} + (\mathbf{AS})\mathbf{K}(\mathbf{AS})']^{-1} \\ &= \mathbf{D} - \mathbf{D}(\mathbf{AS})[\mathbf{K}^{-1} + (\mathbf{AS})'\mathbf{D}(\mathbf{AS})]^{-1}(\mathbf{AS})'\mathbf{D}. \qquad (3.11)\end{aligned}$$

Notice that (3.11) only involves solving systems of linear equation $\mathbf{x}_2 = [\mathbf{K}^{-1} + (\mathbf{AS})'\mathbf{D}(\mathbf{AS})]^{-1}\mathbf{a}_2$, where $\mathbf{K}^{-1} + (\mathbf{AS})'\mathbf{D}(\mathbf{AS})$ has dimension $(r_1 + r_2) \times (r_1 + r_2)$, and $\mathbf{a}_2$ denotes an $(r_1 + r_2)$-dimensional vector. Recall that $r_1$ and $r_2$ are the numbers of basis functions in the low-rank component, which are small or only moderate at worst.

Suppose that we would like to predict $Y_i(\cdot)$ at $m_i$ prediction locations $\{\mathbf{s}_{i1}^P, \ldots, \mathbf{s}_{im_i}^P\}$ for $i = 1, 2$. Define $\mathbf{Y}_i^P = (Y_i(\mathbf{s}_{i1}^P), \ldots, Y_i(\mathbf{s}_{im_i}^P))'$ for $i = 1, 2$, and $\mathbf{Y}^P = (\mathbf{Y}_1^{P\prime}, \mathbf{Y}_2^{P\prime})'$. We use $\mathbf{A}_i^P$ to denote the $m_i \times N$ matrix relating the $m_i$ prediction locations to BAUs, for $i = 1, 2$, and further define $\mathbf{A}^P = blockdiag(\mathbf{A}_1^P, \mathbf{A}_2^P)$. It is straightforward to show that conditioning on the data $\mathbf{Z}$ and parameters $\boldsymbol{\theta}$, $\mathbf{Y}^P|\mathbf{Z}$ is given by:

$$\mathbf{Y}^P|\mathbf{Z} \sim MVN(\mathbf{A}^P\mathbf{X}\boldsymbol{\beta} + \mathbf{A}^P\mathbf{S}\boldsymbol{\mu}_{\boldsymbol{\eta}|\mathbf{Z}} + \mathbf{A}^P\boldsymbol{\mu}_{\boldsymbol{\xi}|\mathbf{Z}}\ ,\ \mathbf{\Sigma}_{\mathbf{Y}^P|\mathbf{Z}}),$$

where $\boldsymbol{\mu}_{\boldsymbol{\eta}|\mathbf{Z}}$ is given in (3.2),

$$\boldsymbol{\mu}_{\boldsymbol{\xi}|\mathbf{Z}} = \mathbb{E}(\boldsymbol{\xi}|\mathbf{Z}) = (\mathbf{A}\mathbf{\Sigma}_{\boldsymbol{\xi}})'\mathbf{\Sigma}_Z^{-1}(\mathbf{Z} - \mathbf{AX}\boldsymbol{\beta}),$$

and

$$\begin{aligned}\mathbf{\Sigma}_{\mathbf{Y}^P|\mathbf{Z}} &= (\mathbf{A}^P\mathbf{S})\mathbf{\Sigma}_{\boldsymbol{\eta}|\mathbf{Z}}(\mathbf{A}^P\mathbf{S})' \\ &\quad + \mathbf{A}^P\mathbf{\Sigma}_{\boldsymbol{\xi}|\mathbf{Z}}(\mathbf{A}^P)' + (\mathbf{A}^P\mathbf{S})\mathbf{\Sigma}_{\boldsymbol{\eta},\boldsymbol{\xi}|\mathbf{Z}}(\mathbf{A}^P)' + \left[(\mathbf{A}^P\mathbf{S})\mathbf{\Sigma}_{\boldsymbol{\eta},\boldsymbol{\xi}|\mathbf{Z}}(\mathbf{A}^P)'\right]'.\end{aligned}$$

Here, $\mathbf{\Sigma}_{\boldsymbol{\eta}|\mathbf{Z}}$ is given in (3.3),

$$\mathbf{\Sigma}_{\boldsymbol{\xi}|\mathbf{Z}} = \mathrm{Var}(\boldsymbol{\xi}|\mathbf{Z}) = \mathbf{\Sigma}_{\boldsymbol{\xi}} - (\mathbf{A}\mathbf{\Sigma}_{\boldsymbol{\xi}})'\mathbf{\Sigma}_Z^{-1}(\mathbf{A}\mathbf{\Sigma}_{\boldsymbol{\xi}}),$$

and

$$\mathbf{\Sigma}_{\boldsymbol{\eta},\boldsymbol{\xi}|\mathbf{Z}} = \mathrm{Cov}(\boldsymbol{\eta}, \boldsymbol{\xi}|\mathbf{Z}) = -(\mathbf{ASK})'\mathbf{\Sigma}_Z^{-1}(\mathbf{A}\mathbf{\Sigma}_{\boldsymbol{\xi}}).$$

We end this section with remarks on the computational complexity of parameter estimation and prediction with MFGP. In the EM algorithm described above, we need to calculate the log-determinant of sparse matrices $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{M} + \mathbf{A}\mathbf{V}_\epsilon\mathbf{A}'$, which has computational complexity $\mathcal{O}(N^{1.5})$, $\mathcal{O}(N^{1.5})$ and $\mathcal{O}((2N)^{1.5})$, respectively, as discussed by [34]. When we calculate $\mathbf{\Sigma}_Z^{-1}$, we need to solve systems of linear equations involving sparse matrices, with computational complexity no more than $\mathcal{O}(2N)$. The calculation of $[\mathbf{K}^{-1} + (\mathbf{AS})'\mathbf{D}(\mathbf{AS})]^{-1}\mathbf{a}_1$ has computational complexity $\mathcal{O}((r_1+r_2)^3)$, but the number of basis functions, $r_1$ and $r_2$, are much smaller than $N$. The EM algorithm also involves sparse matrix multiplication such as $\mathbf{AS}$ and $\mathbf{P}'_{2,1}\mathbf{Q}_2$, whose computational complexity is at most $\mathcal{O}(2N(r_1+r_2)$ and $\mathcal{O}(nnz(\mathbf{P}_{2,1})\ nnz(\mathbf{Q}_2)/N)$, respectively, where $nnz(\mathbf{B})$ denotes the number of non-zero elements in the matrix $\mathbf{B}$. Note that $\mathbf{P}_{2,1}$ and $\mathbf{Q}_2$ are both sparse matrices. Therefore, the overall computational cost for these calculations is $\mathcal{O}(nnz(\mathbf{P}_{2,1})\ nnz(\mathbf{Q}_2)/N + N(r_1 + r_2) + (r_1 + r_2)^3)$. Note also that we need to perform numerical optimization in the M-step, whose computational cost is hard to quantify. As for memory cost, we need to store sparse matrices $\mathbf{A}$, $\mathbf{S}$, $\mathbf{AS}$ and $\mathbf{H}$, which is $\mathcal{O}(N(r_1 + r_2))$ at most. $\mathbf{M}$ should also be stored and occupies $\mathcal{O}(N)$ of memory. The EM algorithm requires storage of the Cholesky factors of $N \times N$ sparse matrices, which will be $\mathcal{O}(N \ln(N))$, at most, after suitable sparse matrix reordering. Therefore, the overall memory cost is $\mathcal{O}(N(r_1 + r_2) + N \ln(N))$. Although the EM algorithm for parameter estimation can not be implemented in parallel computational environments directly, it is possible to carry out some components in parallel. For example, we can calculate spatial predictions and associated prediction standard errors in parallel. Choosing appropriate initial values in the EM algorithm can also accelerate convergence [27]. For parameters $\{\alpha_{2,1,0}, \alpha_{2,1,1}\}$, one way to set their initial values is to first use detail residuals which is calculated as the original data minus the trend fitted via least squares as approximations of $\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2\}$ and then fit a simple

regression model using these approximations. The fitted intercept and slope can then be used as the initial values of $\{\alpha_{2,1,0}, \alpha_{2,1,1}\}$. This is how we set the initial values of the EM algorithm in all numerical examples in Sections 4 and 5.

# 4 Simulation Examples

We present an extensive simulation study to demonstrate the performance of MFGP. Specifically, we consider two scenarios where stationary and nonstationary bivariate spatial data are simulated. In all numerical examples, the MFGP model is implemented in Matlab and the Matlab function `fmincon` is used for numerical optimization in the EM algorithm. We also implement IFGP, MFRK, and SFGP described in Section 2.3 and compare their performance with that of MFGP. Code for simulating and analyzing data in the simulation examples is available at `https://github.com/li2mq/MFGP-Bcode`.

## 4.1 Scenario 1: Performance under a Stationary Cross-Covariance Function

We present analyses with simulated data from a bivariate Matérn cross-covariance function [16, 37]. We consider a spatial domain $\mathcal{D} = [0, 20] \times [0, 20] \subset \mathbb{R}^2$, from which BAUs are defined by regularly discretizing $\mathcal{D}$ to a $50 \times 50$ grid, resulting a total of $N = 50 \times 50 = 2,500$ BAUs. We simulate a bivariate spatial process $\mathbf{Y}(\cdot) = (Y_1(\cdot), Y_2(\cdot))'$ over $\mathcal{D}$. Specifically, we assume zero mean for both variables, and the cross-covariance follows the form,

$$\text{Cov}\{Y_i(\mathbf{s}), Y_j(\mathbf{u})\} \quad = \quad \sigma_{ij}^2 \mathcal{M}(\mathbf{s}, \mathbf{u} | \nu_{ij}, a) \qquad \text{for } i, j = 1, 2, \quad (4.1)$$

where $\mathcal{M}(\cdot, \cdot)$ is the Matérn correlation function [16]:

$$\mathcal{M}(\mathbf{s}, \mathbf{u} \mid \nu, a) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(a\|\mathbf{s} - \mathbf{u}\|\right)^{\nu} \mathbb{K}_{\nu} \left(a\|\mathbf{s} - \mathbf{u}\|\right). \quad (4.2)$$

Here, $\mathbb{K}_\nu(\cdot)$ is the Bessel function of the second kind of order $\nu$; $a > 0$ is a spatial scale parameter that sets the speed of decay in correlation of two points with distance; $\nu > 0$ is the smoothness parameter, with a larger $\nu$ resulting a smoother process.

In our simulation study, we set $\sigma^2_{11} = 1$, $\sigma^2_{22} = 4$, $\sigma^2_{12} = \sigma^2_{21} = 1.25^2$, $a = 1$, $\nu_{11} = 1.5$, $\nu_{22} = 1$, and $\nu_{12} = \nu_{21} = 1.25$ to simulate the bivariate spatial process $\mathbf{Y}(\cdot)$. Then we add noise to generate data: $Z_i(\cdot) = Y_i(\cdot) + \epsilon_i(\cdot)$ as in (2.1). We set $\sigma^2_{\epsilon,i} = 0.05\sigma^2_{ii}$. That is, the variance of the measurement error on variable $i$ is 5% of the variance of the $i$th variable, $Y_i(\cdot)$. Therefore, we have $\sigma^2_{\epsilon,1} = 0.05$ and $\sigma^2_{\epsilon,2} = 0.2$. We randomly sample 200 grid cells and refer them as missing-at-random locations. We hold out data for both $Z_1(\cdot)$ and $Z_2(\cdot)$ at these 200 locations. In addition, we define two block regions within $\mathcal{D}$. We then assume data are missing for $Z_1(\cdot)$ in one of these two blocks while data are missing for $Z_2(\cdot)$ in both blocks, as shown in the second row of Figure 1. The block where data are missing for both $Z_1(\cdot)$ and $Z_2(\cdot)$ is called "Block 1" while the other block is called "Block 2" in this simulation scenario.

We implement MFGP, IFGP, SFGP, and MFRK. They all share the same low-rank components, i.e., the same basis functions. Among these four methods, MFRK is different from the other three as it is based on FRK and doesn't include the GGM component. IFGP fits the two variables with the FGP model, independently. Although SFGP jointly fits bivariate data, as does MFGP, it assumes a separable covariance matrix for the GGM component, and thus we expect it to be less flexible than MFGP. To compare these four methods' predictive performance, we calculate the mean squared prediction error (MSPE) for the $i$th variable, $i = 1, 2$, for each of the four methods, IFGP, SFGP, MFRK or MFGP:

$$\text{MSPE}^i_{\mathcal{S}^P} = \frac{1}{|\mathcal{S}^P|} \sum_{\mathbf{s}^P \in \mathcal{S}^P} \left[ Y_i(\mathbf{s}^P) - \widehat{Y}_i(\mathbf{s}^P) \right]^2. \tag{4.3}$$

Here, $\mathcal{S}^P$ represent the set of locations where predictions are made for $Y_i(\cdot)$.

Later, we present MSPE in situations where $\mathcal{S}^P$ varies. For instance, $\mathcal{S}^P$ may be Block 1 (denoted by $b1$), Block 2 (denoted by $b2$), the set of missing-at-random locations (denoted by *points*), or all missing data locations (denoted by *all*). We also report the continuous-rank-probability score (CRPS; [15]), $\text{CRPS}^i_{\mathcal{S}^P}$, for $i = 1, 2$ for all the four methods. Note that for both MSPE and CRPS, smaller values indicate better predictive performance.

Our simulation consists of 100 runs. Simulated data and prediction results from MFGP for a randomly selected run are plotted in Figure 1. Table 1 summarizes the mean and standard error of MSPE and CRPS from the 100 runs for predicting $Y_1(\cdot)$ over Block 1, missing-at-random locations, and all missing locations. Table 2 presents corresponding results for $Y_2(\cdot)$. Among the four methods, MFRK doesn't perform as well as the others for either $Y_1(\cdot)$ or $Y_2(\cdot)$. The three FGP-based methods, IFGP, SFGP, and MFGP, perform similarly in predicting $Y_1(\cdot)$. However, with $Y_2(\cdot)$, we see more substantial differences among them: MFGP performs better than the other two when predicting $Y_2(\cdot)$ at the locations with data missing in both $Z_1(\cdot)$ and $Z_2(\cdot)$, and gives much smaller MSPE over Block 2, the region where data are missing only in $Z_2(\cdot)$. MFRK and IFGP have the largest and second largest means of $\text{MSPE}_{all}$ for $Y_2(\cdot)$, which are around two or three times those of SFGP and MFGP. The standard errors of $\text{MSPE}_{all}$ from MFRK and IFGP are two-to-three times as large as those from SFGP and MFGP. The difference is mainly due to $\text{MSPE}_{b2}$. Similar conclusions can be drawn with CRPS as well.

We use boxplots to display the performance of these four methods by recording MSPE for each of 100 replications, and computing the ratio of MSPEs obtained from IFGP, SFGP and MFRK that of MFGP. If MFGP performs better than the other three methods, we would expect MFGP to have smaller MSPE, and its ratio would be one. Figure 2 presents boxplots of these ratios over all 100 runs. The fact that the majority of ratios are above one indicates that MFGP outperforms the other three methods. The

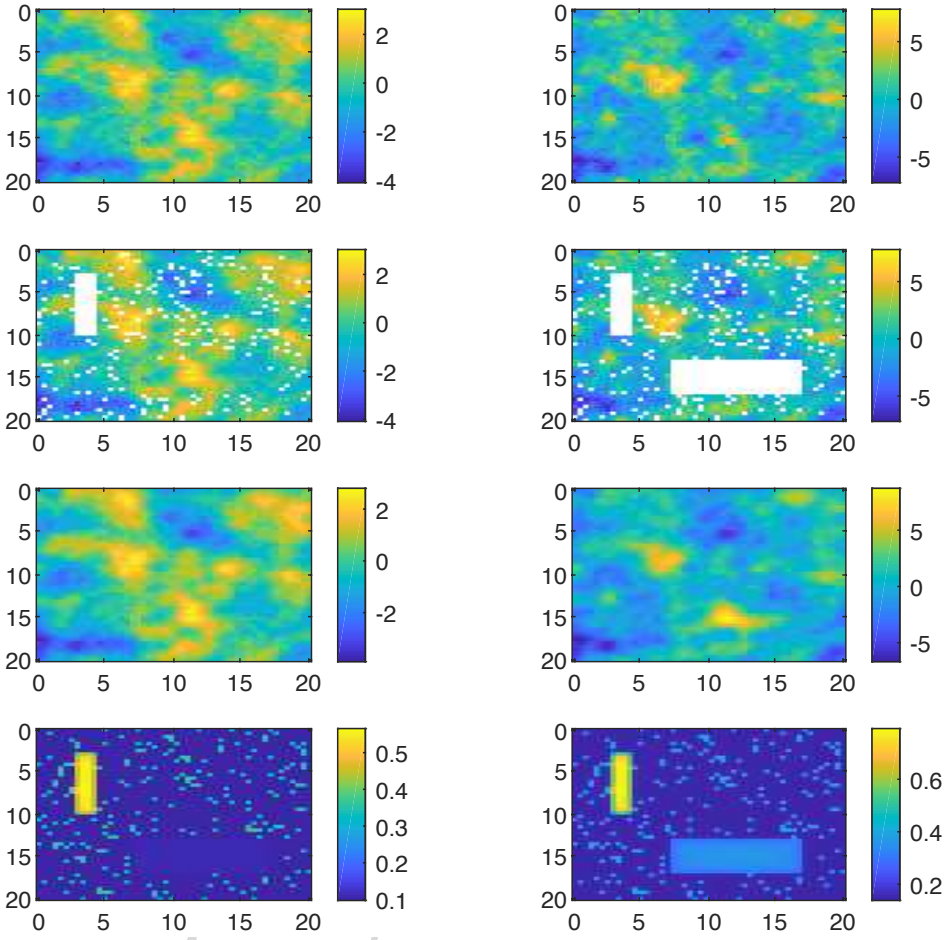


Figure 1: A simulated data set and prediction results from MFGP in Scenario 1. The first row shows the underlying true spatial fields $Y_1(\cdot)$ (left) and $Y_2(\cdot)$ (right). The second row plots the data in which locations with data missing are colored white. The third and fourth rows show the maps of predictions and associated standard errors from MFGP for $Y_1(\cdot)$ (left) and $Y_2(\cdot)$ (right), respectively.

Supplementary Materials presents additional numerical results and similar findings including simulation results from a tri-variate spatial process with the stationary Matérn cross-covariance function in (4.1).

Table 1: Summaries of MSPE and CRPS for predicting $Y_1(\cdot)$ in Scenario 1. The means and standard errors (se) of $\text{MSPE}_{\mathcal{S}^P}$ and $\text{CRPS}_{\mathcal{S}^P}$ are calculated from all 100 runs with $\mathcal{S}^P$ set to Block 1 ($b1$), missing-at-random locations (*points*), and all missing-data locations (*all*). The lowest mean values are highlighted in bold.

| | $\text{MSPE}_{all}$ | | $\text{MSPE}_{b1}$ | | $\text{MSPE}_{points}$ | |
|---|---|---|---|---|---|---|
| Method | mean | se | mean | se | mean | se |
| IFGP | 0.1426 | (0.0314) | 0.2835 | (0.1105) | 0.0919 | (0.0101) |
| SFGP | **0.1388** | (0.0308) | 0.2787 | (0.1091) | **0.0884** | (0.0097) |
| MFGP | 0.1399 | (0.0307) | **0.2782** | (0.1096) | 0.0901 | (0.0104) |
| MFRK | 0.2537 | (0.0520) | 0.4103 | (0.1719) | 0.1973 | (0.0254) |

| | $\text{CRPS}_{all}$ | | $\text{CRPS}_{b1}$ | | $\text{CRPS}_{points}$ | |
|---|---|---|---|---|---|---|
| | mean | se | mean | se | mean | se |
| IFGP | 0.2596 | (0.0378) | 0.3605 | (0.1215) | 0.2269 | (0.0186) |
| SFGP | **0.2557** | (0.0377) | **0.3559** | (0.1226) | **0.2227** | (0.0181) |
| MFGP | 0.2580 | (0.0382) | 0.3565 | (0.1274) | 0.2250 | (0.0184) |
| MFRK | 0.3505 | (0.0500) | 0.4008 | (0.1482) | 0.3265 | (0.0322) |



## 4.2 Scenario 2: Performance under a Nonstationary and Asymmetric Cross-Covariance Model

In Scenario 2, We simulate a bivariate spatial process from a nonstationary and asymmetric cross-covariance model $\text{Cov}(Y_1(\mathbf{s}), Y_2(\mathbf{u})) \neq \text{Cov}(Y_1(\mathbf{u}), Y_2(\mathbf{s}))$. We follow the conditional approach in [9] to simulate $Y_1(\cdot)$ and $Y_2(\cdot)|Y_1(\cdot)$ sequentially. Specifically, we first simulate $Y_1(\cdot)$ from a Gaussian process with the Matérn covariance function with $a = \sigma_{11}^2 = 1$ and $\nu = 1$. To simulate $\{Y_2(\mathbf{s})|Y_1(\cdot) : \mathbf{s} \in \mathcal{D}\}$, we assume:

$$\mathbb{E}\{Y_2(\mathbf{s}) \mid Y_1(\cdot)\} = \int_{\mathcal{D}} b(\mathbf{s}, \mathbf{v}) Y_1(\mathbf{v}) d\mathbf{v}, \tag{4.4}$$

$$\text{Cov}\{Y_2(\mathbf{s}), Y_2(\mathbf{u}) \mid Y_1(\cdot)\} = \sigma_{2|1}^2 \mathcal{M}(\mathbf{s}, \mathbf{u} \mid \nu_{2|1}, a_{2|1}), \tag{4.5}$$

Table 2: Summaries of MSPE and CRPS for predicting $Y_2(\cdot)$ in Scenario 1. The means and standard errors (se) of $\text{MSPE}_{\mathcal{S}^P}$ and $\text{CRPS}_{\mathcal{S}^P}$ are calculated from all 100 runs over $\mathcal{S}^P$ set to Block 1 ($b1$), Block 2 ($b2$), missing-at-random locations (*points*), and all missing-data locations (*all*). The lowest mean values are highlighted in bold.

| | $\text{MSPE}_{all}$ | | $\text{MSPE}_{b1}$ | | $\text{MSPE}_{b2}$ | | $\text{MSPE}_{points}$ | |
|---|---|---|---|---|---|---|---|---|
| | mean | se | mean | se | mean | se | mean | se |
| IFGP | 4.8224 | (2.6171) | 1.8549 | (0.7157) | 9.1649 | (5.5688) | 0.6797 | (0.0645) |
| SFGP | 2.7172 | (1.4022) | 1.8690 | (0.7115) | 4.6555 | (2.9466) | 0.6965 | (0.0665) |
| MFGP | **2.1857** | (1.1275) | **1.6608** | (0.6329) | **3.6616** | (2.3669) | **0.6035** | (0.0547) |
| MFRK | 6.6324 | (3.5282) | 2.5461 | (0.9867) | 12.3105 | (7.5024) | 1.2897 | (0.1517) |

| | $\text{CRPS}_{all}$ | | $\text{CRPS}_{b1}$ | | $\text{CRPS}_{b2}$ | | $\text{CRPS}_{points}$ | |
|---|---|---|---|---|---|---|---|---|
| | mean | se | mean | se | mean | se | mean | se |
| IFGP | 1.0598 | (0.3992) | 0.8878 | (0.2470) | 1.5029 | (0.7291) | 0.6059 | (0.0519) |
| SFGP | 0.9057 | (0.2406) | 0.8896 | (0.2413) | 1.1635 | (0.4385) | 0.6129 | (0.0523) |
| MFGP | **0.8383** | (0.1912) | **0.8371** | (0.2279) | **1.0824** | (0.3670) | **0.5706** | (0.0497) |
| MFRK | 1.3022 | (0.5199) | 0.9777 | (0.3385) | 1.7946 | (0.9667) | 0.8036 | (0.0718) |

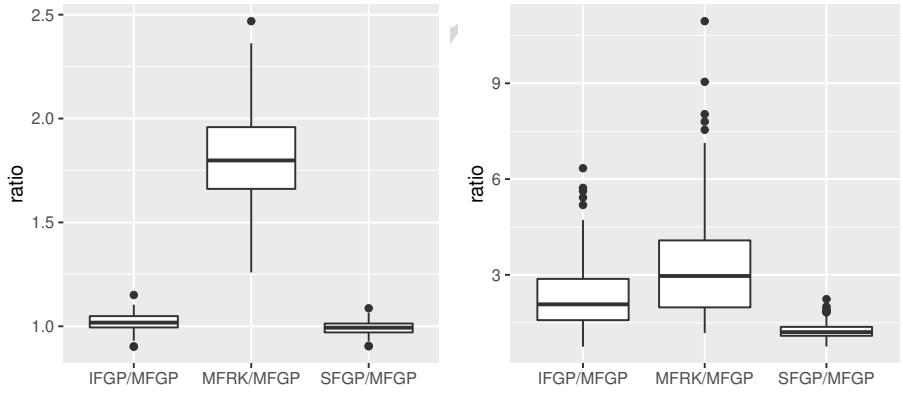


Figure 2: Ratios of MSPEs for predicting $Y_1(\cdot)$ (left) and $Y_2(\cdot)$ (right) in Scenario 1.

where for $\mathbf{u}, \mathbf{s}, \mathbf{v} \in \mathcal{D}$, $b(\cdot, \cdot)$ is called the interaction function. We specify it as: $b(\mathbf{s}, \mathbf{v}) = A\left\{1 - [\ (\|\mathbf{s}\| - \|\mathbf{v} - \mathbf{s}\|)/\gamma\ ]^2\right\}^2 \mathbb{1}_{\|\mathbf{v}-\mathbf{s}\| \leq \gamma/8}$, where $\|\cdot\|$ denotes Euclidean distance and $\mathbb{1}_{\|\mathbf{v}-\mathbf{s}\| \leq \gamma/8}$ is the indicator function whose value is 1 if $\|\mathbf{v} - \mathbf{s}\| \leq \gamma/8$, and 0 otherwise. To simulate $Y_2(\cdot)$, we set $\gamma = 50$, $A = 0.1$, $\sigma^2_{2|1} = 0.5$, $a_{2|1} = 3$, and $\nu_{2|1} = 0.5$. Lastly, we simulate $Z_i(\cdot)$ with $Z_i(\cdot) = Y_i(\cdot) + \epsilon_i(\cdot)$, where $\epsilon_i(\cdot) \overset{iid}{\sim} N(0, \sigma^2_{\epsilon,i})$ for $i = 1, 2$ with $\sigma^2_{\epsilon,1} = 0.05$ and $\sigma^2_{\epsilon,2} = 0.04$.

We carry out 100 runs and implement the four methods as described for Scenario 1. Tables 3 and 4 summarize the mean and standard error of MSPE and CRPS from the 100 runs for predicting $Y_1(\cdot)$ and $Y_2(\cdot)$ over different regions. These results again show that MFGP outperforms the other three methods, giving smaller MSPE and smaller or comparable CRPS overall. In particular, the predictions of $Y_2(\cdot)$ from MFGP are substantially better than those from the other methods over Block 2 where data are missing from $Z_2(\cdot)$ but available for $Z_1(\cdot)$. This demonstrates that MFGP is able to better fit the spatial dependence across the two variables, and thus produce improved predictions compared to the other three methods.

# 5 Applications with Multivariate Remote Sensing Data

In this section, we illustrate MFGP for inference with large multivariate spatial data obtained from remote sensing. [6] present an uncertainty quantification study for the ECOsystem Spaceborne Thermal Radiometer Experiment (ECOSTRESS) which is installed on the International Space Station. In their study, simulation experiments are carried out to quantify the sensitivity of the remote sensing estimation algorithm for deriving evapotranspiration (ET; a quantitative measure of plant water use) from observed radiance spectra. To quantify uncertainties [6] simulate an ensemble of multivariate input spatial fields using SFGP fit to a multivariate

Table 3: Summaries of MSPE and CRPS for predicting $Y_1(\cdot)$ in Scenario 2. The means and standard errors (se) of $\text{MSPE}_{\mathcal{S}^P}$ and $\text{CRPS}_{\mathcal{S}^P}$ are calculated from all 100 runs over $\mathcal{S}^P$ set to Block 1 ($b1$), missing-at-random locations (*points*), and all missing-data locations (*all*). The lowest mean values are highlighted in bold.

| | $\text{MSPE}_{all}$ | | $\text{MSPE}_{b1}$ | | $\text{MSPE}_{points}$ | |
|---|---|---|---|---|---|---|
| | mean | se | mean | se | mean | se |
| IFGP | 0.0496 | (0.0048) | 0.0593 | (0.0141) | 0.0461 | (0.0041) |
| SFGP | 0.0493 | (0.0048) | **0.0556** | (0.0133) | 0.0471 | (0.0045) |
| MFGP | **0.0478** | (0.0044) | 0.0569 | (0.0135) | **0.0445** | (0.0037) |
| MFRK | 0.0502 | (0.0059) | 0.0609 | (0.0156) | 0.0463 | (0.0051) |

| | $\text{CRPS}_{all}$ | | $\text{CRPS}_{b1}$ | | $\text{CRPS}_{points}$ | |
|---|---|---|---|---|---|---|
| | mean | se | mean | se | mean | se |
| IFGP | 0.1682 | (0.0135) | 0.1855 | (0.0425) | 0.1629 | (0.0117) |
| SFGP | 0.1682 | (0.0137) | **0.1780** | (0.0385) | 0.1638 | (0.0139) |
| MFGP | 0.1680 | (0.0138) | 0.1802 | (0.0434) | 0.1630 | (0.0134) |
| MFRK | **0.1672** | (0.0178) | 0.1867 | (0.0457) | **0.1612** | (0.0156) |

Table 4: Summaries of MSPE and CRPS for predicting $Y_2(\cdot)$ in Scenario 2. The means and standard errors (se) of $\text{MSPE}_{\mathcal{S}^P}$ and $\text{CRPS}_{\mathcal{S}^P}$ are calculated from all 100 runs over $\mathcal{S}^P$ set to Block 1 ($b1$), Block 2 ($b2$), missing-at-random locations (*points*), and all missing-data locations (*all*). The lowest mean values are highlighted in bold.

| | $\text{MSPE}_{all}$ | | $\text{MSPE}_{b1}$ | | $\text{MSPE}_{b2}$ | | $\text{MSPE}_{points}$ | |
|---|---|---|---|---|---|---|---|---|
| | mean | se | mean | se | mean | se | mean | se |
| IFGP | 0.6524 | (0.3711) | 0.2580 | (0.1027) | 1.2376 | (0.7895) | 0.0922 | (0.0090) |
| SFGP | 0.6703 | (0.3601) | 0.2774 | (0.1106) | 1.2657 | (0.7673) | 0.0973 | (0.0119) |
| MFGP | **0.5035** | (0.2655) | **0.2457** | (0.1030) | **0.9246** | (0.5668) | **0.0909** | (0.0099) |
| MFRK | 0.8581 | (0.4754) | 0.3887 | (0.1516) | 1.5569 | (1.0122) | 0.1885 | (0.0235) |

| | $\text{CRPS}_{all}$ | | $\text{CRPS}_{b1}$ | | $\text{CRPS}_{b2}$ | | $\text{CRPS}_{points}$ | |
|---|---|---|---|---|---|---|---|---|
| | mean | se | mean | se | mean | se | mean | se |
| IFGP | 0.3840 | (0.1347) | 0.3261 | (0.1165) | 0.5352 | (0.2469) | 0.2286 | (0.0186) |
| SFGP | 0.3930 | (0.1373) | 0.3333 | (0.1178) | 0.5428 | (0.2494) | 0.2346 | (0.0192) |
| MFGP | **0.3576** | (0.1157) | **0.3116** | (0.0995) | **0.4765** | (0.2132) | **0.2279** | (0.0176) |
| MFRK | 0.4693 | (0.1474) | 0.3855 | (0.1349) | 0.6086 | (0.2778) | 0.3193 | (0.0301) |

"truth" data set. Then, the remote sensing estimation algorithm is applied to all locations in each ensemble member. Distributions of estimated ET at locations in the scene, and their relationships to corresponding "truth" data, provide the desired characterization of uncertainty.

[6] use SFGP to fit and then jointly simulate leaf area index (LAI), land surface temperature (LST), and normalized difference vegetation index (NDVI). Here, we use data for these three variables over a $400 \times 400$ grid from two of their scenes, called Scene 1 and Scene 2 in this paper. The data are shown in the top rows of Figure 3 and Figure 4, respectively. More detailed descriptions of these two data sets and background behind them are given in the Supplementary Materials. Note that the sizes of these two tri-variate spatial data sets are about $3 \times 400^2 = 480,000$, which is too large for classical geostatistical methods such as cokriging.

We implement all four methods, MFGP, SFGP, IFGP, and MFRK. Tables 5, 6 and 7 display the summaries of MSPE and CRPS based on predictions of the three variables, NDVI, LAI, and LST. Results are consistent with those from simulation examples: We find that MFGP gives the best predictive performance among all the four methods, in particular in blocks where data are available from some but not all variables.

Improved predictive performance of MFGP comes with a more complicated model specification and thus more computing time: MFRK can be executed in less than 2 minutes; IFGP takes about 1.5 hours; it takes SFGP more than 11 hours; MFGP takes the most time: around 50 hours. However, MFGP reduces MSPE for $Y_2(\cdot)$ by more than 25% compared to IFGP in Scene 1 and about 30% in Scene 2; the reduction of MSPE is even more when we compare MFGP with SFGP and MRFK. The most computationally intensive step in implementing MFGP is the EM algorithm, in which we need to perform numerical optimization with more than 9 parameters, at each iteration. Discussion on possible ways to simplify the parameterization in MFGP, and thus speed up the computation further

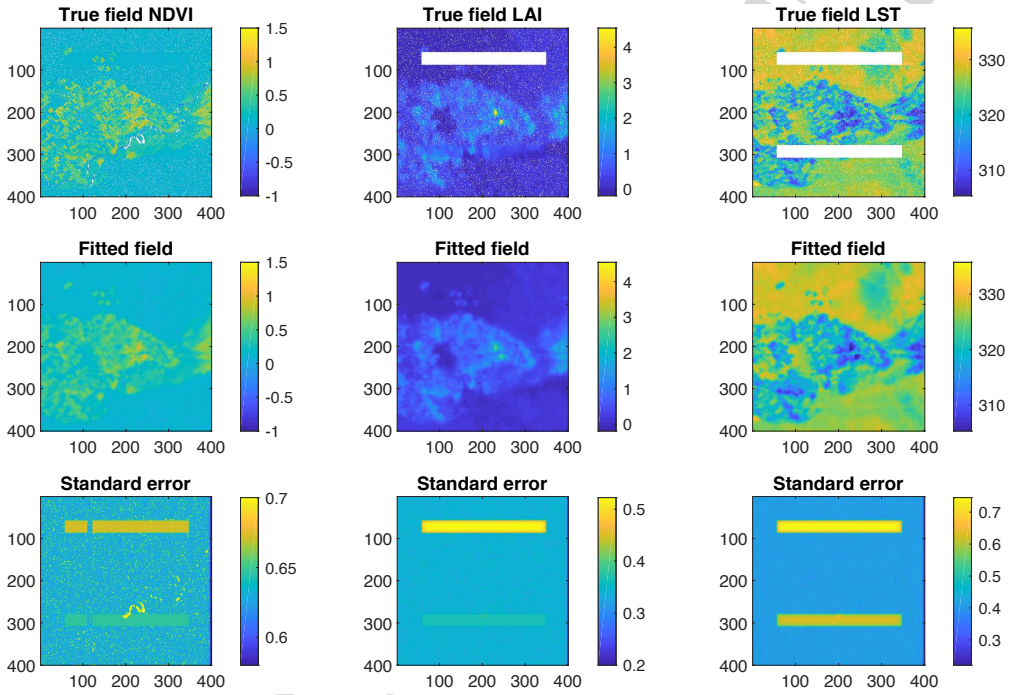


Figure 3: Plots of Data and predictions from MFGP for Scene 1. Data for NDVI (left column), LAI (middle column), and LST (right column) are shown in the top row. MFGP predictions are shown in the middle row, and associate prediction standard errors are shown in the bottom row.

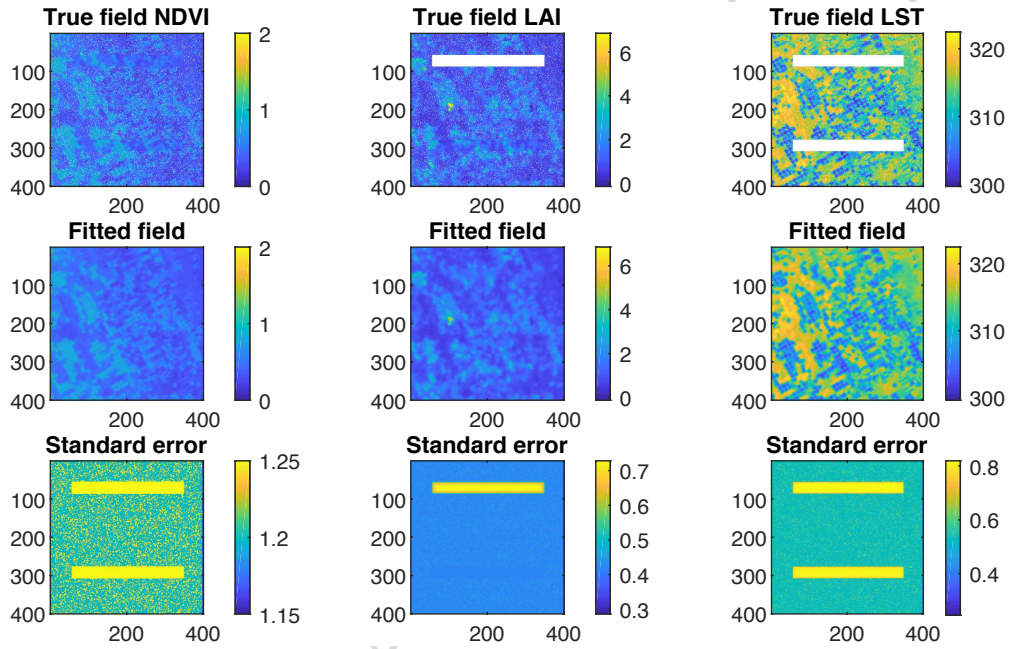


Figure 4: Plots of Data and predictions from MFGP for Scene 2. Data of NDVI (left column), LAI (middle column), and LST (right column) are shown in the top row. MFGP predictions are shown in the middle row, and associated prediction standard errors are shown in the bottom row.

can be found in Section 6.

Finally, note that in order to use the conditional approach we need to specify the order of conditioning. We thus investigate how MFGP performs under all the six possible orders of conditioning, and find that the predictive performance of MFGP is not sensitive to this choice. Details of this sensitivity study are given in the Supplementary Materials.

Table 5: MSPE and CRPS for predicting NDVI ($Y_1(\cdot)$) calculated over missing-data locations from data in both Scene 1 and Scene 2. The lowest value in each column is highlighted in bold.

| | Scene 1 | | Scene 2 | |
|---|---|---|---|---|
| | $\text{MSPE}_{points}$ | $\text{CRPS}_{points}$ | $\text{MSPE}_{points}$ | $\text{CRPS}_{points}$ |
| IFGP | **0.0090** | **0.0518** | **0.0035** | **0.0453** |
| SFGP | 0.0097 | 0.0539 | 0.0045 | 0.0524 |
| MFGP | 0.0096 | 0.0526 | 0.0042 | 0.0497 |
| MFRK | 0.0211 | 0.0878 | 0.0163 | 0.1042 |

Table 6: Summaries of MSPE and CRPS for predicting LAI ($Y_2(\cdot)$) from both Scene 1 and Scene 2. The means and standard errors of $\text{MSPE}_{\mathcal{S}^P}$ and $\text{CRPS}_{\mathcal{S}^P}$ are calculated from locations in $\mathcal{S}^P$ set to be Block 1 ($b1$), missing-at-random locations (*points*), and all missing-data locations (*all*). The lowest value in each column is highlighted in bold.

| | | $\text{MSPE}_{all}$ | $\text{MSPE}_{b1}$ | $\text{MSPE}_{points}$ | $\text{CRPS}_{all}$ | $\text{CRPS}_{b1}$ | $\text{CRPS}_{points}$ |
|---|---|---|---|---|---|---|---|
| Scene 1 | IFGP | 0.0097 | 0.0109 | 0.0082 | 0.0550 | 0.0537 | 0.0561 |
| | SFGP | 0.0113 | 0.0145 | **0.0074** | **0.0502** | **0.0470** | **0.0529** |
| | MFGP | **0.0072** | **0.0047** | 0.0102 | 0.0567 | 0.0487 | 0.0641 |
| | MFRK | 0.0269 | 0.0157 | 0.0410 | 0.1047 | 0.0644 | 0.1373 |
| Scene 2 | IFGP | 0.1289 | 0.2146 | 0.0223 | 0.2313 | 0.3377 | 0.0991 |
| | SFGP | 0.1710 | 0.2963 | **0.0153** | 0.2565 | 0.3977 | **0.0810** |
| | MFGP | **0.0872** | **0.1379** | 0.0243 | **0.1991** | **0.2757** | 0.1039 |
| | MFRK | 0.2618 | 0.3073 | 0.2053 | 0.3765 | 0.4119 | 0.3326 |

Table 7: Summaries of MSPE and CRPS for predicting LST ($Y_3(\cdot)$) from both Scene 1 and Scene 2. The means and standard errors of $\text{MSPE}_{\mathcal{S}^P}$ and $\text{CRPS}_{\mathcal{S}^P}$ are calculated from locations in $\mathcal{S}^P$ set to be Block 1 ($b1$), Block 2 ($b2$), missing-at-random locations ($points$), and all missing-data locations ($all$). The lowest value is highlighted in bold in each column.

| | | $\text{MSPE}_{all}$ | $\text{MSPE}_{b1}$ | $\text{MSPE}_{b2}$ | $\text{MSPE}_{points}$ |
|---|---|---|---|---|---|
| Scene 1 | IFGP | 8.2930 | 4.6564 | 16.7823 | 2.2593 |
| | SFGP | 6.5379 | 3.9618 | 12.6498 | **2.1416** |
| | MFGP | **5.4560** | **2.8397** | **10.7216** | 2.1619 |
| | MFRK | 11.2948 | 6.4759 | 17.2347 | 9.9011 |
| | | $\text{CRPS}_{all}$ | $\text{CRPS}_{b1}$ | $\text{CRPS}_{b2}$ | $\text{CRPS}_{points}$ |
| | IFGP | 1.8255 | 1.5153 | 2.7878 | 1.0145 |
| | SFGP | 1.6717 | 1.4313 | 2.4608 | **0.9896** |
| | MFGP | **1.5788** | **1.2915** | **2.3333** | 0.9978 |
| | MFRK | 2.3474 | 1.8191 | 2.9819 | 2.2154 |
| | | $\text{MSPE}_{all}$ | $\text{MSPE}_{b1}$ | $\text{MSPE}_{b2}$ | $\text{MSPE}_{points}$ |
| Scene 2 | IFGP | 11.4903 | 13.2583 | 17.2495 | 2.1350 |
| | SFGP | 7.2649 | 9.3429 | 9.1974 | 2.2804 |
| | MFGP | **4.1336** | **4.8233** | **5.3685** | **1.7417** |
| | MFRK | 17.0773 | 16.5009 | 20.0083 | 14.1509 |
| | | $\text{CRPS}_{all}$ | $\text{CRPS}_{b1}$ | $\text{CRPS}_{b2}$ | $\text{CRPS}_{points}$ |
| | IFGP | 2.4940 | 2.6885 | 3.4115 | 1.1121 |
| | SFGP | 1.9798 | 2.2382 | 2.3882 | 1.1511 |
| | MFGP | **1.5124** | **1.6929** | **1.7346** | **1.0120** |
| | MFRK | 3.2648 | 3.0131 | 3.7749 | 2.9437 |

## 6 Conclusions and Discussion

In this paper we propose the multivariate fused Gaussian process (MFGP) model which can be used to flexibly model multivariate spatial processes with large data. We demonstrate that MFGP gives superior predictive performance in various simulation scenarios, and in an application to remote sensing data analysis. One advantage MFGP possesses is its flexibility to handle data at different spatial resolutions. MFGP inherits the additive, multiresolution structure of FRK and FGP; the basis functions in the model are completely prespecified and known. This makes it possible to handle data sets with different spatial resolutions, since change-of-support is easily accomplished by aggregating off-line when data are at coarser resolutions than the BAUs.

When then number of variables $q$ is large, the number of MFGP parameters increases substantially. This can make MFGP less than desirable for highly-multivariate spatial processes. The dimension of the matrix $\mathbf{K}$ becomes $\sum_{i=1}^{q} r_i$. For $\{\boldsymbol{\xi}_i\}_{i=1}^{q}$, we need to estimate parameters $\{\tau_i\}_{i=1}^{q}$, $\{\gamma_i\}_{i=1}^{q}$, and $\{\alpha_{i,j,0}, \alpha_{i,j,1}\}_{1\leq j<i\leq q}$. One way to alleviate this difficulty is to introduce additional assumptions such as the Markov property of order one for the GGM component: $\boldsymbol{\xi}_q \perp \boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_{q-2}$, given $\boldsymbol{\xi}_{q-1}$. The directed acyclic graph (DAG) structure used in the meshed GP [33] can also be considered for the GGM component. "Stitching" [12] may also be used to form a sparse graphical model for $\{\boldsymbol{\xi}_i\}_{i=1}^{q}$. This can potentially reduce the number of parameters for $\{\boldsymbol{\xi}_i\}_{i=1}^{q}$. And additional difficulty is that when $q$ is large, $\mathbf{K}$ may no longer be "low-rank" as $\sum_{i=1}^{q} r_i$ increases. [25] recently propose to incorporate regularization terms in model fitting to enforce sparsity in $\mathbf{K}$ and to achieve efficient computation.

One natural extension of MFGP is to relax the assumption of Gaussian distributions and to generalize it for multivariate non-Gaussian spatial data. This can be achieved by embedding the MFGP model in the framework of

the spatial generalized linear models [38]. By assuming a state-space model structure for the low-rank and the multivariate CAR components [10, 28], we may also extend the MFGP model to the space-time setting. These directions may be pursued in future research.

# 7 Appendix

Proof of Proposition 2.1: For $\boldsymbol{\xi}^{(k)} \equiv (\boldsymbol{\xi}_1', \ldots, \boldsymbol{\xi}_k')'$ and $k = 1, \ldots, q$, let $\mathbf{M}_k$ denote the precision matrix of $\boldsymbol{\xi}^{(k)}$. Thus, $\mathbf{M} = \mathbf{M}_q$. We will prove that $\mathbf{M} = |\mathbf{Q}_1| \cdot \cdots \cdot |\mathbf{Q}_q|$ using induction as follows.

When $q = 1$, $\mathbf{M} = \mathbf{M}_1 = \mathbf{Q}_1$. Thus, $|\mathbf{M}| = |\mathbf{Q}_1|$. Therefore, Proposition 2.1 holds when $q = 1$.

When $q = 2$, it is straightforward to show that $\mathbf{M} = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{P}_1'\mathbf{Q}_2 \\ \mathbf{0} & -\mathbf{Q}_2 \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{P}_1 & -\mathbf{I} \end{pmatrix} := \mathbf{B}_{2,1}\mathbf{B}_{2,2}'$, where $\mathbf{B}_{2,1}$ and $\mathbf{B}_{2,2}$ are two upper triangular block matrices. Hence, $|\mathbf{M}| = |\mathbf{Q}_1| \cdot |\mathbf{Q}_2|$. The result in Proposition 2.1 holds when $q = 2$.

Assume that for $q = k \geq 2$, we have $|\mathbf{M}_k| = |\mathbf{Q}_1| \cdot \cdots \cdot |\mathbf{Q}_k|$, and $\mathbf{M}_k = \mathbf{B}_{k,1}\mathbf{B}_{k,2}'$, where $\mathbf{B}_{k,1}$ and $\mathbf{B}_{k,2}$ are two upper triangular block matrices. Then, for $q = k + 1$, in the MFGP model we have $p(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_k, \boldsymbol{\xi}_{k+1}) = p(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_k) \times p(\boldsymbol{\xi}_{k+1} | \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_k)$, where $p(\cdot)$ denotes the probability density function (pdf).

It is straightforward to show that:

$$\begin{aligned} & p(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_k, \boldsymbol{\xi}_{k+1}) \propto exp\left\{-1/2 \boldsymbol{\xi}^{(k)\prime} \mathbf{M}_k \boldsymbol{\xi}^{(k)}\right\} \\ & \times exp\left\{-1/2 \left[\boldsymbol{\xi}_{k+1} - \textstyle\sum_{j=1}^{k} \mathbf{P}_{k+1,j} \boldsymbol{\xi}_j\right]' \mathbf{Q}_{k+1} \left[\boldsymbol{\xi}_{k+1} - \textstyle\sum_{j=1}^{k} \mathbf{P}_{k+1,j} \boldsymbol{\xi}_j\right]\right\} \\ = \; & exp\left\{-1/2 \boldsymbol{\xi}^{(k+1)\prime} \mathbf{M}_{k+1} \boldsymbol{\xi}^{(k+1)}\right\}. \end{aligned}$$

Thus, we have:

$$\mathbf{M}_{k+1} = \begin{pmatrix} \mathbf{B}_{k,1} & \bar{\mathbb{R}}_k' \mathbf{Q}_{k+1} \\ \mathbf{0} & -\mathbf{Q}_{k+1} \end{pmatrix} \begin{pmatrix} \mathbf{B}_{k,2}' & \mathbf{0} \\ \bar{\mathbb{R}}_k & -\mathbf{I} \end{pmatrix} := \mathbf{B}_{k+1,1}\mathbf{B}_{k+1,2}',$$

where $\bar{\mathbb{R}}_k$ is an $N \times (kN)$ matrix given by

$$\bar{\mathbb{R}}_k = (\mathbf{P}_{k+1,1} \quad \mathbf{P}_{k+1,2} \quad \cdots \quad \mathbf{P}_{k+1,k}) .$$

We thus have

$$|\mathbf{M}_{k+1}| = |\mathbf{B}_{k,1}| \times |\mathbf{B}_{k,2}| \times |\mathbf{Q}_{k+1}| = |\mathbf{M}_k| \times |\mathbf{Q}_{k+1}| = |\mathbf{Q}_1| \times |\mathbf{Q}_2| \times \cdots \times |\mathbf{Q}_k| \times |\mathbf{Q}_{k+1}|,$$

which means that the result in Proposition 2.1 holds when $q = k + 1$. This completes the proof of this proposition.



## Acknowledgments

Part of the research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. ©2021 California Institute of Technology. Government sponsorship acknowledged. K. Cawse-Nicholson acknowledges support from ECOSTRESS and the NASA Earth Venture Instruments Program. This work was supported in part by an allocation of computing time from the Ohio Supercomputer Center and by research cyberinfrastructure resources and services provided by the Advanced Research Computing (ARC) center at the University of Cincinnati. This research was part of Li's Ph.D. dissertation supported by the Taft Research Center at the University of Cincinnati. Kang was partially supported by the National Science Foundation (NSF) under award DMS–2053668, NASA-ROSES grant NNH18ZDA001N-SLSCVC, Simons Foundation's Collaboration Award (#317298 and #712755), and the Taft Research Center at the University of Cincinnati.

## References

[1] **Akritas, A. G., Akritas, E. K. and Malaschonok, G. I.** (1996). Various proofs of Sylvester's (determinant) identity, *Mathematics and Computers in Simulation*, **42(4-6)**, 585-593.

[2] **Apanasovich, T. V., Genton, M. G. and Ying Sun, Y.** (2012). A valid matérn class of cross-covariance functions for multivariate random fields with any number of components, *Journal of the American Statistical Association*, **107(497)**, 180-193.

[3] **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(4)**, 825-848.

[4] **Bradley, J. R., Cressie, N. and Shi, T.** (2014). Rejoinder on: Comparing and selecting spatial predictors using local criteria, *TEST*, **24(1)**, 54-60.

[5] **Banerjee, S., Carlin, B. P., Gelfand, A. E., Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., D. Heckerman, D., Smith, A. F. M. and West, M.** (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data (with discussion), *Bayesian Statistics*, **7**, 45-63.

[6] **Cawse-Nicholson, K., Braverman, A., Kang, E. L., Li, M., Johnson, M., Halverson, G., Anderson, M., Hain, C., Gunson, M. and Hook, S.** (2020). Sensitivity and uncertainty quantification for the ECOSTRESS evapotranspiration algorithm - DisALEXI, *International Journal of Applied Earth Observation and Geoinformation*, **89**, 102088.

[7] **Chu, T., Wang, H. and Zhu, J.** (2014). On semiparametric inference of geostatistical models via local Karhunen–Loève expansion, *Journal of the Royal Statistical Society: Series B*, **76(4)**, 817-832.

[8] **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(1)**, 209-226.

[9] **Cressie, N. and Zammit-Mangion, A.** (2016). Multivariate spatial covariance models: a conditional approach, *Biometrika*, **103(4)**, 915-935.

[10] **Cressie, N., Shi, T. and Kang, E. L.** (2010). Fixed rank filtering for spatio-temporal data, *Journal of Computational and Graphical Statistics*, **19(3)**, 724-745.

[11] **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(514)**, 800-812.

[12] **Dey, D., Datta, A. and Banerjee, S.** (2021). Graphical Gaussian process models for highly multivariate spatial data, *Biometrika*, `https://doi.org/10.1093/biomet/asab061`.

[13] **Gelfand, A. E. and Vounatsou, P.** (2003). Proper multivariate conditional autoregressive models for spatial data analysis, *Biostatistics*, **4(1)**, 11–15.

[14] **Genton, M. G. and William Kleiber, W.** (2015). Cross-covariance functions for multivariate geostatistics, *Statistical Science*, **30(2)**, 147-163.

[15] **Gneiting, T. and Raftery, A. E.** (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association*, **102(477)**, 359-378.

[16] **Gneiting, T., Kleiber, W. and Schlather, M.** (2010). Matérn cross-covariance functions for multivariate random fields, *Journal of the American Statistical Association*, **105(491)**, 1167-1177.

[17] **Guinness, J.** (2022). Nonparametric spectral methods for multivariate spatial and spatial-temporal data, *Journal of Multivariate Analysis*, **187**, 104823.

[18] **Jin, X., Carlin, B. P. and Banerjee, S.** (2005). Generalized hierarchical multivariate CAR models for areal data, *Biometrics*, **61(4)**, 950-961.

[19] **Kang, E. L., Cressie, N. and Shi, T.** (2010). Using temporal variability to improve spatial mapping with application to satellite data, *Canadian Journal of Statistics*, **38(2)**, 271-289.

[20] **Katzfuss, M.** (2017). A multi-resolution approximation for massive spatial datasets, *Journal of the American Statistical Association*, **112(517)**, 201-214.

[21] **Katzfuss, M. and Cressie, N.** (2011). Spatio-temporal smoothing and em estimation for massive remote-sensing data sets, *Journal of Time Series Analysis*, **32(4)**, 430-446.

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

[23] **Kleiber, W.** (2017). Coherence for multivariate random fields, *Statistica Sinica*, **27(4)**, 1675-1697.

[24] **Kleiber, W., Nychka, D. and Bandyopadhyay, S.** (2019). A model for large multivariate spatial data sets, *Statistica Sinica*, **29(3)**, 1085-1104.

[25] **Krock, M., Kleiber, W., Hammerling, D. and Stephen Becker, S.** (2021). Modeling massive highly-multivariate nonstationary spatial data with the basis graphical lasso, *arXiv preprint arXiv:2101.02404*.

[26] **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(4)**,423-498.

[27] **Ma, P. and Kang, E. L.** (2020). A fused Gaussian process model for very large spatial data, *Journal of Computational and Graphical Statistics*, **29(3)**, 479-489.

[28] **Ma, P. and Kang, E. L.** (2020). Spatio-temporal data fusion for massive sea surface temperature data from MODIS and AMSR-E instruments, *Environmetrics*, **31(2)**: e2594.

[29] **Mardia., K. V.** (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing, *Journal of Multivariate Analysis*, **24(2)**, 265-284.

[30] **Nguyen, H., Cressie, N. and Braverman, A.** (2012). Spatial statistical data fusion for remote sensing applications, *Journal of the American Statistical Association*, **107(499)**, 1004-1018.

[31] **Nguyen, H., Katzfuss, M., Cressie, N. and Braverman, A.** (2014). Spatio-temporal data fusion for very large remote sensing datasets, *Technometrics*, **56(2)**, 174-185.

[32] **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(2)**, 579-599.

[33] **Peruzzi, M., Banerjee, S. and Finley, A. O.** (2020). Highly scalable Bayesian geostatistical modeling via meshed Gaussian processes on partitioned domains, *Journal of the American Statistical Association*, DOI:10.1080/01621459.2020.1833889.

[34] **Rue, H. and Held, L.** (2005). *Gaussian Markov Random Fields: Theory and Applications*, Chapman and Hall/CRC.

[35] **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(1)**, 111-132.

[36] **Sang, H., Jun, M. and Huang, J.Z.** (2011). Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors, *The Annals of Applied Statistics*, **5(4)**, 2519-2548.

[37] **Schlather, M., Malinowski, A., Menck, P. J., Oesting, M., Strokorb, K.** (2015). Analysis, simulation and prediction of multivariate random fields with package randomfields, *Journal of Statistical Software*, **63(8)**, 1-25.

[38] **Shi, H. and Kang, E. L.** (2017). Spatial data fusion for large non-gaussian remote sensing datasets, *Stat*, **6(1)**, 390-404.

[39] **Shi, T. and Cressie, N.** (2007). Global statistical analysis of MISR aerosol data: a massive data product from NASA's Terra satellite, *Environmetrics*, **18(7)**, 665-680.

[40] **Tzeng, S. and Hsin-Cheng Huang, H. C.** (2018). Resolution adaptive fixed rank kriging, *Technometrics*, **60(2)**, 198-208.

[41] **Zammit-Mangion, A. and Cressie, N.** (2021). FRK: An R package for spatial and spatio-temporal prediction with large datasets, *Journal of Statistical Software*, **98(1)**, 1-48.

[42] **Zhang, L., Banerjee, S. and Andrew O. Finley. A. O.** (2021). High-dimensional multivariate geostatistics: A bayesian matrix-normal approach, *Environmetrics*, **32(4)**: e2675.

**Emily L. Kang**
4199 French Hall, University of Cincinnati, 2815 Commons Way
Cincinnati, OH 45221-0025
E-mail: kangel@ucmail.uc.edu

**Miaoqi Li**
Wells Fargo
E-mail: miaoqili9986@gmail.com

**Kerry Cawse-Nicholson**
4800 Oak Grove Drive M/S 183-518
Pasadena, CA 91109
E-mail: kerry-anne.cawse-nicholson@jpl.nasa.gov

**Amy Braverman**
4800 Oak Grove Drive M/S 158-242
Pasadena, CA 91109
E-mail: amy.braverman@jpl.nasa.gov