BIOMETRIC METHODOLOGY

Biometrics WILEY

Spatial factor modeling: A Bayesian matrix-normal approach for misaligned data

Lu Zhang[1] 🄙 | **Sudipto Banerjee**[2] 🄙

[1] Department of Statistics, Columbia University, New York

[2] Department of Biostatistics, University of California, Los Angeles, California

Correspondence
Sudipto Banerjee, Department of Biostatistics, University of California, Los Angeles, CA.
Email: sudipto@ucla.edu

Abstract

Multivariate spatially oriented data sets are prevalent in the environmental and physical sciences. Scientists seek to jointly model multiple variables, each indexed by a spatial location, to capture any underlying spatial association for each variable and associations among the different dependent variables. Multivariate latent spatial process models have proved effective in driving statistical inference and rendering better predictive inference at arbitrary locations for the spatial process. High-dimensional multivariate spatial data, which are the theme of this article, refer to data sets where the number of spatial locations and the number of spatially dependent variables is very large. The field has witnessed substantial developments in scalable models for univariate spatial processes, but such methods for multivariate spatial processes, especially when the number of outcomes are moderately large, are limited in comparison. Here, we extend scalable modeling strategies for a single process to multivariate processes. We pursue Bayesian inference, which is attractive for full uncertainty quantification of the latent spatial process. Our approach exploits distribution theory for the matrix-normal distribution, which we use to construct scalable versions of a hierarchical linear model of coregionalization (LMC) and spatial factor models that deliver inference over a high-dimensional parameter space including the latent spatial process. We illustrate the computational and inferential benefits of our algorithms over competing methods using simulation studies and an analysis of a massive vegetation index data set.

KEYWORDS
Bayesian inference, factor models, linear models of coregionalization, matrix-normal distribution, multivariate spatial processes, scalable spatial modeling

1 | INTRODUCTION

Statistical modeling for multiple spatially oriented data is required to capture underlying spatial associations in each variable and accounting for inherent associations among the different variables. As an example, to which we return later, consider a set of spatially indexed spectral variables for vegetation activity on the land. Such variables exhibit strong spatial dependence as customarily exhibited through plots of spatial variograms and other exploratory maps. In addition, the variables are assumed to be associated with each other because of shared physical processes that manifest through the observations.

Modeling each variable separately captures the spatial distribution of that variable independent of other variables. Such an analysis ignores associations among

the variables and can impair prediction or interpolation (see, e.g., Wackernagel, 2003; Chiles and Delfiner, 2009; Gelfand and Banerjee, 2010; Cressie and Wikle, 2015). Each of the aforementioned works provides ample evidence, theoretical and empirical, in favor of joint modeling of multiple spatially indexed variables. Joint modeling, or multivariate spatial analysis, is especially pertinent in the presence of spatial misalignment, where not all variables have been observed over the same set of locations. For example, suppose $Y(\mathbf{s})$ is a normalized difference vegetation index (NDVI) and $X(\mathbf{s})$ is red reflectance. If location $\mathbf{s}_0$ has yielded a measurement for $X(\mathbf{s}_0)$ but not for $Y(\mathbf{s}_0)$, then optimal imputation of $Y(\mathbf{s}_0)$ should proceed from $p(Y(\mathbf{s}_0) \mid \mathbf{Y}, \mathbf{X})$, where $\mathbf{Y}$ and $\mathbf{X}$ comprise all measurements on $Y(\mathbf{s})$ and $X(\mathbf{s})$. If the processes $Y()$ and $X()$ are modeled as independent, then the predictive distribution $p(Y(\mathbf{s}_0) \mid \mathbf{Y}, \mathbf{X}) = p(Y(\mathbf{s}_0) \mid \mathbf{Y})$ and will not exploit the possible predictive information present in $X(\mathbf{s}_0)$ for $Y(\mathbf{s}_0)$. This specific issue has also been discussed, with examples, in Banerjee and Gelfand (2002).

Joint modeling is driven by vector-valued latent spatial stochastic processes, such as a multivariate Gaussian process. These are specified with matrix-valued cross-covariance functions (see, e.g., Le and Zidek, 2006; Genton and Kleiber, 2015; Salvaña and Genton, 2020, and references therein) that model pairwise associations at distinct locations. Theoretical properties of cross-covariances are well established, but practical modeling implications and computational efficiency require specific considerations depending upon the application (see, e.g., Le *et al.*, 1997; Sun *et al.*, 1998; Le *et al.*, 2001; Schmidt and Gelfand, 2003; Gamerman and Moreira, 2004; Banerjee *et al.*, 2014).

High-dimensional multivariate spatial models will deal with a large number of dependent variables over a massive number of locations. While analyzing massive spatial and spatial-temporal databases has received attention (see, e.g., Sun *et al.*, 2011; Banerjee, 2017; Heaton *et al.*, 2019; Zhang *et al.*, 2020), the bulk of methods has focused on one or very few (two or three) spatially dependent variables and often have to rely upon restrictive assumptions that preclude full inference on the latent process. With larger numbers of dependent variables, modeling the cross-covariance becomes challenging. Even for stationary cross-covariance functions, where we assume that the associations among the variables do not change over space and the spatial association for each variable depends only on the difference of two positions, matters become computationally challenging.

This article builds upon the popular linear models of coregionalization (Bourgault and Marcotte, 1991; Goulard and Voltz, 1992; Wackernagel, 2003; Gelfand *et al.*, 2004; Chiles and Delfiner, 2009; Genton and Kleiber, 2015). Our contributions include: (i) developing a hierarchical

model with a matrix-normal distribution as a prior for an unknown linear transformation on latent spatial processes, (ii) extending classes of spatial factor models for spatially misaligned data, and (iii) accounting for multiple outcomes over very large number of locations. Spatial factor models have been explored by Wang and Wall (2003), Lopes *et al.* (2008), Ren and Banerjee (2013), and Taylor-Rodriguez *et al.* (2019). Lopes *et al.* (2008) provide an extensive discussion on how hierarchical models emerged from dynamic factor models. Ren and Banerjee (2013) proposed low-rank specifications for spatially varying factors to achieve dimension reduction, but such low-rank specifications tend to oversmooth the latent process from massive data sets containing millions of locations. More recently, Taylor-Rodriguez *et al.* (2019) consider nearest-neighbor Gaussian process (Datta *et al.*, 2016) for spatial factors with the usual constrained loading matrices in nonspatial factor models. These are more restrictive than needed for identifying spatially correlated factors (see, e.g., Ren and Banerjee, 2013).

We develop our modeling framework in Section 2. Section 3 presents some theoretical results about posterior consistency for the proposed models. Simulation studies for exploring the performance of proposed models are summarized in Section 4. Section 5 presents an application to remote-sensed vegetation analysis on land surfaces. We conclude with some discussion in Section 6.

2 | MULTIVARIATE SPATIAL PROCESSES

Let $\mathbf{z}(\mathbf{s}) = (z_1(\mathbf{s}), \dots, z_q(\mathbf{s}))^{\top}$ be a $q \times 1$ stochastic process, where each $z_i(\mathbf{s})$ is a real-valued random variable at location $\mathbf{s} \in \mathcal{D} \subseteq \mathfrak{R}^d$. The process is specified by its mean $\mathrm{E}[z_i(\mathbf{s})] = \mu_i(\mathbf{s})$ and, customarily, second-order stationary covariances $C_{ij}(\mathbf{h}) = \mathrm{Cov}\{z_i(\mathbf{s}), z_j(\mathbf{s} + \mathbf{h})\}$ for $i, j = 1, 2, \dots, q$. These covariances define the matrix-valued $q \times q$ cross-covariance function $\mathbf{C}(\mathbf{h}) = \{C_{ij}(\mathbf{h})\}$ with (i, j)-th entry $C_{ij}(\mathbf{h})$. While there is no loss of generality in assuming the process mean to be zero by absorbing the mean into a separate regression component in the model, as we will do here, modeling the cross-covariance function requires care. From its definition, $\mathbf{C}(\mathbf{h})$ need not be symmetric, but must satisfy $\mathbf{C}(\mathbf{h})^{\top} = \mathbf{C}(-\mathbf{h})$. Also, since $\mathrm{var}\{\sum_i^n \mathbf{a}_i^{\top} \mathbf{z}(\mathbf{s}_i)\} \geq 0$ for any set of finite locations $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n \in \mathcal{D}$ and any set of constant vectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n \in \mathfrak{R}^q$, we have $\sum_{i,j=1}^{n} \mathbf{a}_i^{\top} \mathbf{C}(\mathbf{s}_i - \mathbf{s}_j) \mathbf{a}_i \geq 0$. Genton and Kleiber (2015) provide a comprehensive review of cross-covariance functions.

Perhaps the most widely used approach for constructing multivariate random fields is the linear model of

coregionalization (LMC). This hinges on invertible linear maps of independent spatial processes yielding valid spatial processes. If $\mathbf{f}(\mathbf{s}) = (f_1(\mathbf{s}), f_2(\mathbf{s}), \ldots, f_K(\mathbf{s}))^\top$ is a $K \times 1$ vector of independent spatial processes so that $\mathrm{cov}\{f_i(\mathbf{s}), f_j(\mathbf{s}')\} = 0$ for all $i \neq j$ and any two locations $\mathbf{s}$ and $\mathbf{s}'$ (same or distinct), then LMC (Bourgault and Marcotte, 1991) specifies

$$\mathbf{z}(\mathbf{s}) = \sum_{k=1}^{K} \lambda_k f_k(\mathbf{s}) = \mathbf{\Lambda}^\top \mathbf{f}(\mathbf{s}) , \qquad (1)$$

where $\mathbf{z}(\mathbf{s})$ is $q \times 1$, $\mathbf{\Lambda}$ is $K \times q$, $\lambda_k^\top$ is the k-th row of $\mathbf{\Lambda}$, and each $f_k(\mathbf{s})$ is an independent Gaussian process with correlation function $\rho_{\psi_k}(\cdot, \cdot)$ with parameters ψ_k. The cross-covariance for $\mathbf{z}(\mathbf{s})$ yields nondegenerate process-realizations whenever $K \geq q$ and $\mathbf{\Lambda}$ is nonsingular. To achieve dimension reduction in the number of variables, we restrict $K < q$ so we have nondegenerate realizations in a K-dimensional subspace.

Schmidt and Gelfand (2003) propose multivariate spatial processes through a hierarchical spatial conditional model, whereupon $\mathbf{\Lambda}^\top$ in (1) is a $q \times q$ lower triangular matrix. Other variants of LMC (e.g., Goulard and Voltz, 1992) can also be recast as (1) using linear algebra. The flexibility offered in modeling $\mathbf{\Lambda}$ is appealing and, in particular, can accrue computational benefits in high-dimensional settings. Hence, we build upon (1).

2.1 | A Bayesian LMC (BLMC) factor model

Let $\mathbf{y}(\mathbf{s}) = (y_1(\mathbf{s}), \ldots, y_q(\mathbf{s}))^\top \in \mathbb{R}^q$ denote the $q \times 1$ vector of dependent outcomes in location $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d$, $\mathbf{x}(\mathbf{s}) = (x_1(\mathbf{s}), \ldots, x_p(\mathbf{s}))^\top \in \mathbb{R}^p$ be the corresponding explanatory variables, and $\boldsymbol{\beta}$ be a $p \times q$ regression coefficient matrix in the multivariate spatial model

$$\mathbf{y}(\mathbf{s}) = \boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s}) + \mathbf{\Lambda}^\top \mathbf{f}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}), \ \mathbf{s} \in \mathcal{D}, \qquad (2)$$

where the latent process $\mathbf{\Lambda}^\top \mathbf{f}(\mathbf{s})$ is an LMC as described above. Elements in $\mathbf{f}(\mathbf{s})$ are as described in (1), while the noise process $\boldsymbol{\epsilon}(\mathbf{s}) \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$ with covariance matrix $\mathbf{\Sigma}$. We model $\{\boldsymbol{\beta}, \mathbf{\Lambda}, \mathbf{\Sigma}\}$ using a Matrix-Normal-Inverse-Wishart family. To be precise,

$$\boldsymbol{\beta} \mid \mathbf{\Sigma} \sim \mathrm{MN}(\boldsymbol{\mu}_\beta, \mathbf{V}_\beta, \mathbf{\Sigma}) \, ; \ \mathbf{\Lambda} \mid \mathbf{\Sigma} \sim \mathrm{MN}(\boldsymbol{\mu}_\Lambda, \mathbf{V}_\Lambda, \mathbf{\Sigma});$$
$$\mathbf{\Sigma} \sim \mathrm{IW}(\mathbf{\Psi}, \nu), \qquad (3)$$

where $\boldsymbol{\mu}_\Lambda$ a $q \times K$ matrix and $\mathbf{V}_\Lambda$ a $K \times K$ positive definite matrix. A random matrix $\mathbf{Z}_{n \times p} \sim \mathrm{MN}_{n,p}(\mathbf{M}, \mathbf{U}, \mathbf{V})$ has the probability density function (Dawid, 1981)

$$p(\mathbf{Z} \mid \mathbf{M}, \mathbf{U}, \mathbf{V})$$
$$= \frac{\exp\left[-\frac{1}{2} \mathrm{tr}\{\mathbf{V}^{-1}(\mathbf{Z} - \mathbf{M})^T \mathbf{U}^{-1}(\mathbf{Z} - \mathbf{M})\}\right]}{(2\pi)^{np/2} |\mathbf{V}|^{n/2} |\mathbf{U}|^{p/2}}, \qquad (4)$$

where $\mathrm{tr}(\cdot)$ is the trace function, $\mathbf{M}$ is the mean matrix, $\mathbf{U}$ is the first scale matrix with dimension $n \times n$, and $\mathbf{V}$ is the second scale matrix with dimension $p \times p$. This distribution is equivalent to $\mathrm{vec}(\mathbf{Z}) \sim \mathrm{N}_{np}(\mathrm{vec}(\mathbf{M}), \mathbf{V} \otimes \mathbf{U})$, where $\otimes$ is the Kronecker product and $\mathrm{vec}(\mathbf{Z}) = [\mathbf{z}_1^\top, \ldots, \mathbf{z}_p^\top]^\top$ is the vectorized $n \times p$ random matrix $\mathbf{Z} = [\mathbf{z}_1 : \cdots : \mathbf{z}_p]$. We refer to the model specified through (2)–(3) as the BLMC factor model.

Without misalignment, the observation model in (2) can be cast as

$$\mathbf{Y}_{n \times q} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times q} + \mathbf{F}_{n \times K} \mathbf{\Lambda}_{K \times q} + \boldsymbol{\epsilon}_{n \times q} , \qquad (5)$$

where $\mathbf{Y} = \mathbf{y}(S) = [\mathbf{y}(\mathbf{s}_1) : \cdots : \mathbf{y}(\mathbf{s}_n)]^\top$ is the $n \times q$ response matrix, $\mathbf{X} = \mathbf{x}(S) = [\mathbf{x}(\mathbf{s}_1) : \cdots : \mathbf{x}(\mathbf{s}_n)]^\top$ is the corresponding design matrix with full rank ($n > p$), and $\mathbf{F}$ is the $n \times K$ matrix with j-th column being the $n \times 1$ vector comprising $f_j(\mathbf{s}_i)$'s for $i = 1, 2, \ldots, n$.

The parameters $\mathbf{\Lambda}$ and $\mathbf{F}$ are not jointly identified in factor models and some constraints are required to ensure identifiability (Ren and Banerjee, 2013; Lopes and West, 2004). These constraints are not without problems. For example, a lower trapezoidal (triangular for $K = q$) specification for $\mathbf{\Lambda}$ imposes possibly unjustifiable conditional independence on the spatial processes. Alternatively, ordering the spatial range parameters can ensure identifiability but creates difficulties in computation and interpretation. We avoid such constraints and transform $\boldsymbol{\omega} = \mathbf{F}\mathbf{\Lambda} = [\boldsymbol{\omega}(\mathbf{s}_1) : \cdots : \boldsymbol{\omega}(\mathbf{s}_n)]^\top$ to obtain inference for the latent process. This parameterization yields conditional conjugate distributions and, therefore, efficient posterior sampling. We elucidate below in the context of misaligned data.

2.2 | Inference for spatially misaligned data

Let $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ be the set of locations that have recorded at least one of the observed outcomes and let S_i be the subset of locations that have recorded the i-th response. Then $\cup_{i=1}^{q} S_i = S$ and let $n_i = |S_i|$. Let $\mathcal{M}_i = S \setminus S_i$ denote the set of locations where at least one response, but not the i-th response, is recorded so that $\mathcal{M} = \cup_{i=1}^{q} \mathcal{M}_i$ is the set of all locations with incomplete data. We derive the conditional distribution of $\mathbf{F}$ and of the unobserved responses $\{y_i(\mathcal{M}_i)\}_{i=1}^{q}$ conditional on

$\{\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \{\psi_k\}_{k=1}^K\}$. Let $\mathbf{P}$ be the $(\sum_{i=1}^q n_i) \times nq$ matrix such that $\mathbf{P}\text{vec}(\mathbf{Y}) = [\mathbf{y}(\mathbf{s}_1)_{os_1}^\top, \mathbf{y}(\mathbf{s}_2)_{os_2}^\top, \ldots, \mathbf{y}(\mathbf{s}_n)_{os_n}^\top]^\top$, where the suffix os_i are indexes of the observed responses at $\mathbf{s}_i \in \mathcal{S}$. Thus, $\mathbf{P}$ extracts the observed responses from $\text{vec}(\mathbf{Y})$ in each of the locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. The joint distribution of $\text{vec}(\mathbf{F})$ and $\{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^n$, given by $\{\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \{\psi_k\}_{k=1}^K\}$, can be represented through the augmented linear system,

$$\begin{bmatrix} \{(\mathbf{y}(\mathbf{s}_i) - \mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\beta})_{os_i}\}_{i=1}^n \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{P}(\boldsymbol{\Lambda}^\top \otimes \mathbf{I}_n) \\ \mathbf{I}_K \otimes \mathbf{I}_n \end{bmatrix} \text{vec}(\mathbf{F}) + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{bmatrix}, \tag{6}$$

where $\boldsymbol{\epsilon}_1 \sim N(\mathbf{0}, \oplus_{i=1}^n \{\boldsymbol{\Sigma}_{os_i}\})$, $\boldsymbol{\epsilon}_2 \sim N(\mathbf{0}, \oplus_{k=1}^K \{\boldsymbol{\rho}_{\psi_k}(S, S)\})$, $\boldsymbol{\rho}_{\psi_k}(S, S)$ is the $n \times n$ spatial correlation matrix corresponding to $\mathbf{f}_k = (f_k(\mathbf{s}_1), f_k(\mathbf{s}_2), \ldots, f_k(\mathbf{s}_n))^\top$, and $\oplus_{i=1}^n$ represents the block diagonal operator stacking matrices along the diagonal. Letting $\mathbf{D}_{\Sigma_o}^{-\frac{1}{2}} = \oplus_{i=1}^n \{\boldsymbol{\Sigma}_{os_i}^{-\frac{1}{2}}\}$ and $\mathbf{V}_{\mathbf{F}} = \oplus_{k=1}^K \{\mathbf{V}_k\}$, where $\boldsymbol{\rho}_{\psi_k}^{-1}(S, S) = \mathbf{V}_k^\top \mathbf{V}_k$, we obtain

$$\underbrace{\begin{bmatrix} \mathbf{D}_{\Sigma_o}^{-\frac{1}{2}}\{(\mathbf{y}(\mathbf{s}_i) - \mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\beta})_{os_i}\}_{i=1}^n \\ \mathbf{0} \end{bmatrix}}_{\tilde{\mathbf{Y}}} = \underbrace{\begin{bmatrix} \mathbf{D}_{\Sigma_o}^{-\frac{1}{2}} \mathbf{P}\boldsymbol{\Lambda}^\top \otimes \mathbf{I}_n \\ \mathbf{V}_{\mathbf{F}} \end{bmatrix}}_{\tilde{\mathbf{X}}} \text{vec}(\mathbf{F}) + \underbrace{\begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix}}_{\tilde{\boldsymbol{\eta}}}. \tag{7}$$

The elements of $\tilde{\boldsymbol{\eta}}$ are independent error terms, each with unit variance. The full conditional distribution $\text{vec}(\mathbf{F}) \mid \{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^n, \boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \{\psi_k\}_{k=1}^K$ for the LMC model in (2) then follows

$$\text{vec}(\mathbf{F}) \mid \{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^n, \boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \{\psi_k\}_{k=1}^K$$
$$\sim N((\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{Y}}, (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}). \tag{8}$$

For misaligned data, we will perform Bayesian updating of the outcomes missing at a location $\mathbf{s} \in \mathcal{M}$. Let ms be the suffix that indexes outcomes that are missing at $\mathbf{s} \in \mathcal{M}$. The conditional distribution of $\mathbf{y}(\mathbf{s})_{ms}$ given the parameters $\{\mathbf{F}, \{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^n, \boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}\}$ is

$$N([\boldsymbol{\mu}_{\mathbf{s}}]_{ms} + \boldsymbol{\Sigma}_{[ms,os]} \boldsymbol{\Sigma}_{[os,os]}^{-1}(\mathbf{y}(\mathbf{s})_{os} - [\boldsymbol{\mu}_{\mathbf{s}}]_{os}), \boldsymbol{\Sigma}_{[ms,ms]}$$
$$- \boldsymbol{\Sigma}_{[ms,os]} \boldsymbol{\Sigma}_{[os,os]}^{-1} \boldsymbol{\Sigma}_{[os,ms]}), \tag{9}$$

where $\boldsymbol{\mu}_{\mathbf{s}} = \boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s}) + \boldsymbol{\Lambda}^\top \mathbf{f}(\mathbf{s})$, $\boldsymbol{\Sigma}_{[ms,os]}$ is the submatrix of $\boldsymbol{\Sigma}$ extracted with row and column indices ms and os, respectively. With the priors given in (3), we let $\mathbf{V}_\Lambda = \mathbf{L}_\Lambda \mathbf{L}_\Lambda^\top$ and define $\boldsymbol{\gamma} = [\boldsymbol{\beta}^\top, \boldsymbol{\Lambda}^\top]^\top$. The conditional posterior distribution $\boldsymbol{\gamma} \mid \boldsymbol{\Sigma}, \mathbf{F}, \mathbf{Y}$ can be found from

$$\underbrace{\begin{bmatrix} \mathbf{Y} \\ \mathbf{L}_\beta^{-1} \boldsymbol{\mu}_\beta \\ \mathbf{L}_\Lambda^{-1} \boldsymbol{\mu}_\Lambda \end{bmatrix}}_{\mathbf{Y}^*} = \underbrace{\begin{bmatrix} \mathbf{X} & \mathbf{F} \\ \mathbf{L}_\beta^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_\Lambda^{-1} \end{bmatrix}}_{\mathbf{X}^*} \underbrace{\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\Lambda} \end{bmatrix}}_{\boldsymbol{\gamma}} + \underbrace{\begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{bmatrix}}_{\boldsymbol{\eta}^*}, \tag{10}$$

where $\boldsymbol{\eta}^* \sim MN(\mathbf{0}_{(n+p+K)\times q}, \mathbf{I}_{n+p+K}, \boldsymbol{\Sigma})$. Using standard distribution theory, we can show that $\boldsymbol{\gamma}, \boldsymbol{\Sigma} \mid \mathbf{F}, \mathbf{Y}$ follows $MNIW(\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Psi}^*, \nu^*)$, where

$$\mathbf{V}^* = [\mathbf{X}^{*\top} \mathbf{X}^*]^{-1}, \quad \boldsymbol{\mu}^* = \mathbf{V}^*[\mathbf{X}^{*\top} \mathbf{Y}^*], \quad \boldsymbol{\Psi}^* = \boldsymbol{\Psi} + \mathbf{S}^*,$$
$$\text{and } \nu^* = \nu + n, \tag{11}$$

with $\mathbf{S}^* = (\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*)^\top (\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*)$. In particular, if $\boldsymbol{\Sigma} = \oplus_{i=1}^q \{\sigma_i^2\}$ and each $\sigma_i^2 \sim IG(a, b_i)$ for $i = 1, \ldots q$, then the conditional distribution of σ_i^2 given $\mathbf{Y}, \mathbf{F}$ follows $IG(a^*, b_i^*)$, where

$$a^* = a + \frac{n}{2}, \quad b_i^* = b_i + \frac{1}{2}(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*)_i^\top (\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*)_i,$$
$$i = 1, \ldots, q, \tag{12}$$

and $(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*)_i$ is the i-th column of $\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*$. From (10), $\boldsymbol{\gamma} \mid \boldsymbol{\Sigma}, \mathbf{F}, \mathbf{Y} \sim MN(\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Sigma})$.

The parameters ψ_k, $k = 1, 2, \ldots, K$, by themselves, are not consistently estimable under in-fill asymptotics. Therefore, irrespective of the sample size (within a fixed domain), inference on ψ_k will be sensitive to the choice of the prior. Furthermore, without placing restrictions on the loading matrix or ordering these parameters (Ren and Banerjee, 2013), these parameters are identifiable primarily through the prior. We treat these as unknown and model them using priors based upon customary spatial domain considerations. The full conditional distributions for $\{\psi_k\}_{k=1}^K$ are not available in closed form. However, since $\{\psi_k\}_{k=1}^K$ and $\mathbf{Y}$ are conditionally independent given $\{\mathbf{F}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}\}$, and $\mathbf{f}_k$ are independent for $k = 1, 2, \ldots, K$, we obtain $p(\psi_k \mid \mathbf{F}, \mathbf{Y}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \{\psi_j\}_{j \neq k})$ up to a proportionality constant as

$$p(\mathbf{Y} \mid \mathbf{F}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) \times p(\boldsymbol{\gamma}, \boldsymbol{\Sigma}) \times \prod_{k=1}^K p(\mathbf{f}_k \mid \psi_k)$$
$$\times p(\psi_k) \propto p(\mathbf{f}_k \mid \psi_k) \times p(\psi_k), \tag{13}$$

for each $k = 1, \ldots, K$, where $p(\psi_k)$ is the prior for ψ_k.

Turning to predictions, if $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_{n'}\}$ is a set of new locations, then $\mathbf{Y}_{\mathcal{U}} = \mathbf{y}(\mathcal{U})$ is independent of $\{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^n$ given $\{\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}\}$ and $\mathbf{F}_{\mathcal{U}} = [\mathbf{f}_1(\mathcal{U}) : \ldots : \mathbf{f}_K(\mathcal{U})]^\top$. Then,

$$\mathbf{f}_k(\mathcal{U}) \mid \mathbf{f}_k, \psi_k \sim N(\boldsymbol{\rho}_{\psi_k}(\mathcal{U}, S)\boldsymbol{\rho}_{\psi_k}^{-1}(S, S)\mathbf{f}_k,$$
$$\boldsymbol{\rho}_{\psi_k}(\mathcal{U}, S)\boldsymbol{\rho}_{\psi_k}^{-1}(S, S)\boldsymbol{\rho}_{\psi_k}(S, \mathcal{U})), \tag{14}$$

for each $k = 1, 2, \ldots, K$. It follows that $p(\mathbf{Y}_{\mathcal{U}}, \mathbf{F}_{\mathcal{U}} \mid \{\mathbf{y}(\mathbf{s}_i)_{oi}\}_{i=1}^n)$ is proportional to

$$p(\mathbf{Y}_{\mathcal{U}} \mid \mathbf{F}_{\mathcal{U}}, \boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) \times p(\mathbf{F}_{\mathcal{U}} \mid \mathbf{F}, \{\psi_k\}_{k=1}^K)$$
$$\times p(\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \mathbf{F}, \{\psi_k\}_{k=1}^K \mid \{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^n), \tag{15}$$

ZHANG AND BANERJEE **Biometrics** WILEY | **5**

where we have used the independence between $\mathbf{F}_{\mathcal{U}}$ and $\{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^n$ given $\mathbf{F}$ and $\{\psi_k\}_{k=1}^K$. The distributions in (14) and (15) help in sampling from the posterior predictive distribution over $\mathcal{U}$ using the posterior samples of $\{\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \mathbf{F}, \{\psi_k\}_{k=1}^K\}$. We elaborate below.

2.3 | The block update Markov chain Monte Carlo (MCMC) algorithm

We formulate an efficient MCMC algorithm for obtaining full Bayesian inference as follows. From the l-th iteration with $\{\boldsymbol{\beta}^{(l)}, \boldsymbol{\Lambda}^{(l)}, \boldsymbol{\Sigma}^{(l)}, \{\psi_k^{(l)}\}_{k=1}^K\}$, we generate $\mathbf{F}^{(l+1)}$ from (8). Next, we draw $\{\mathbf{y}(\mathbf{s}_i)_{mi}^{(l+1)}\}_{\mathbf{s}_i \in \mathcal{M}}$ on $\mathcal{M}$ using (9) and then update $\{\boldsymbol{\beta}^{(l+1)}, \boldsymbol{\Lambda}^{(l+1)}, \boldsymbol{\Sigma}^{(l+1)}\}$ using (11). We complete the $(l+1)$-th iteration by drawing $\{\psi_k^{(l+1)}\}_{k=1}^K$ through a Metropolis random walk using (13). Upon convergence, these iterations will generate samples from the desired joint posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\Lambda}, \mathbf{F}, \boldsymbol{\Sigma}, \{\psi_k\}_{k=1}^K \mid \mathbf{Y})$.

For inference on $\mathcal{U}$, we sample $\mathbf{F}_{\mathcal{U}}$ from (14), given the posterior samples of $\mathbf{F}$ and $\{\psi_k\}_{k=1}^K$, then generate posterior predictions of $\mathbf{Y}_{\mathcal{U}}$ given the posterior samples of $\{\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \mathbf{F}_{\mathcal{U}}\}$. Applying the single component adaptive Metropolis (SCAM) algorithm introduced in Haario *et al.* (2005), one can avoid tuning parameters in Metropolis algorithm by warming up each MCMC chain of $\{\psi_k\}_{k=1}^K$ with an adaptive proposal distribution. In our implementation, we use the proposal distribution defined by eq. (2.1) in Roberts and Rosenthal (2009).

We sample $\mathbf{F}$ as a single block through a linear transformation of the $n \times K$ independent parameters from the model in (7). Sampling $\{\boldsymbol{\beta}, \boldsymbol{\Lambda}\}$ follows analogously. We significantly improve convergence by reducing the posterior dependence among the parameters in this Gibbs with Metropolis algorithm (Gelman *et al.*, 2013). Since $\mathbf{F}$ is sensitive to the value of the intercept, we recommend using an intercept-centered latent process to obtain inference for the latent spatial pattern.

2.4 | Scalable modeling

We use a conjugate gradient method (Nishimura and Suchard, 2018) to facilitate sampling of $\mathbf{F}$ when $\rho_{\psi_k}^{-1}(\mathcal{S}, \mathcal{S})$ is sparse for $k = 1, \dots, K$. Here, we develop a scalable BLMC model with each element of $\mathbf{f}(\mathbf{s})$ modeled as a Nearest-Neighbor Gaussian Process (NNGP).

Let each $f_k(\mathbf{s}), \mathbf{s} \in \mathcal{D}$ be an NNGP$(0, \rho_{\psi_k}(\cdot, \cdot))$, which implies that $\mathbf{f}_k \sim \mathrm{N}(\mathbf{0}, \tilde{\rho}_k)$ for each $k = 1, 2, \dots, K$, where $\tilde{\rho}_k = (\mathbf{I} - \mathbf{A}_{\rho_k})^{-1}\mathbf{D}_{\rho_k}(\mathbf{I} - \mathbf{A}_{\rho_k})^{-\top}$, $\mathbf{A}_{\rho_k}$ is a sparse lower

triangular matrix with no more than a specified small number, m, of nonzero entries in each row and $\mathbf{D}_{\rho_k}$ is a diagonal matrix. The diagonal entries of $\mathbf{D}_{\rho_k}$ and the nonzero entries of $\mathbf{A}_{\rho_k}$ are obtained from the conditional variance and conditional expectations for a Gaussian process with covariance function $\rho_{\psi_k}(\mathbf{s}, \mathbf{s}')$. We consider a fixed order of locations in $\mathcal{S}$ and let $N_m(\mathbf{s}_i)$ be the set of at most m neighbors of $\mathbf{s}_i$ among locations $\mathbf{s}_j \in \mathcal{S}$ such that $j < i$. The (i, j)-th entry of $\mathbf{A}_{\rho_k}$ is 0 whenever $\mathbf{s}_j \notin N_m(\mathbf{s}_i)$. If $j_1 < j_2 < \cdots < j_m$ are the m column indices for the nonzero entries in the i-th row of $\mathbf{A}_{\rho_k}$, then the (i, j_k)-th element of $\mathbf{A}_{\rho_k}$ is the k-th element of the $1 \times m$ vector $\mathbf{a}_i^\top = \rho_{\psi_k}(\mathbf{s}_i, N_m(\mathbf{s}_i))\rho_{\psi_k}(N_m(\mathbf{s}_i), N_m(\mathbf{s}_i))^{-1}$. The (i, i)-th diagonal element of $\mathbf{D}_{\rho_k}$ is given by $\rho_{\psi_k}(\mathbf{s}_i, \mathbf{s}_i) - \mathbf{a}_i^\top \rho_{\psi_k}(N_m(\mathbf{s}_i), \mathbf{s}_i)$. Repeating these calculations for each row completes the construction of $\mathbf{A}_{\rho_k}$ and $\mathbf{D}_{\rho_k}$ and yields a sparse $\tilde{\rho}_k^{-1}$. This construction is performed in parallel and requires storage or computation of at most $m \times m$ matrices, where $m << n$, costing $\mathcal{O}(n)$ flops and storage. See Supplementary Appendix B for details.

Sampling $\mathbf{F}$ is computationally expensive, but is expedited by solving $(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \mathbf{v}$ efficiently for any vector $\mathbf{v}$. If $\rho_{\psi_k}(\mathcal{S}, \mathcal{S}) = \mathbf{L}_k \mathbf{L}_k^\top$ has a sparse Cholesky factor $\mathbf{L}_k$, then calculating $\tilde{\mathbf{X}}^\top \mathbf{v}$ is efficient. To be precise, the Woodbury matrix identity yields

$$(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} = (\mathbf{F}\mathbf{D}_{\Sigma_0}^{-1}\mathbf{F}^\top + \oplus_{k=1}^K \{\rho_k^{-1}\})^{-1}$$
$$= \oplus_{k=1}^K \{\rho_k\} - \oplus_{k=1}^K \{\rho_k\}\mathbf{F}\mathbf{G}^{-1}\mathbf{F}^\top \oplus_{k=1}^K \{\rho_k\}, \quad (16)$$

where $\mathbf{F} = (\boldsymbol{\Lambda} \otimes \mathbf{I}_n)\mathbf{P}^\top$ is sparse, $\mathbf{G} = \mathbf{D}_{\Sigma_0} + \mathbf{P}\{\sum_{k=1}^K \lambda_{ik}\lambda_{jk}\rho_k\}_{i,j=1}^p \mathbf{P}^\top$ with $\rho_k = \rho_{\psi_k}(\mathcal{S}, \mathcal{S})$. If all the ρ_k's have similar structures, then permuting $\{\sum_{k=1}^K \lambda_{ik}\lambda_{jk}\rho_k\}_{i,j=1}^p$ with $\mathbf{P}$ in rows and columns often renders structures in ρ_k's that can be exploited by BLMC for very large spatial data sets . For example, if ρ_k's are banded matrices with bandwidth b, then $\mathbf{P}\{\sum_{k=1}^K \lambda_{ik}\lambda_{jk}\rho_k\}_{i,j=1}^p \mathbf{P}^\top$ is also banded with bandwidth bq. Moreover, $\mathbf{D}_{\Sigma_0}$ is a banded matrix with bandwidth $\leq q$. Hence, adding $\mathbf{D}_{\Sigma_0}$ hardly increases the computational burden in the Cholesky decomposition of $\mathbf{G}$ when q is small. Assembling all features of ρ_k, $\mathbf{F}$, and $\mathbf{G}$, the calculation of $(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}\mathbf{u}$ for any $\mathbf{u} = \tilde{\mathbf{X}}^\top \mathbf{v}$ is scalable when multiplying $\mathbf{u}$ with (16).

We conclude this section with a remark on the BLMC model with diagonal $\boldsymbol{\Sigma}$. This specification is desirable for data sets with a massive number of responses q. A diagonal $\boldsymbol{\Sigma}$ avoids the quadratic growth of the number of parameters in $\boldsymbol{\Sigma}$ as q increases. We illustrate an NNGP-based BLMC with diagonal $\boldsymbol{\Sigma}$ in Section 4.2.

3 | ON POSTERIOR CONSISTENCY: LARGE-SAMPLE PROPERTIES OF POSTERIOR ESTIMATES

We present some theoretical results for the models constructed in the previous section. Specifically, we investigate the behavior of the posterior distribution as the sample size increases and establish its convergence to an oracle distribution. Here, for establishing the results, we will assume conjugate Matrix-Normal-Inverse-Wishart (MNIW) models with no misalignment. First, we assume that $\mathbf{y}(\mathbf{s})$ itself is modeled as a spatial process without explicitly introducing a latent process. Let

$$\mathbf{y}(\mathbf{s}) \sim \mathrm{GP}(\boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s}), \mathbf{C}(\cdot,\cdot)), \; \mathbf{C}(\mathbf{s},\mathbf{s}') = \{\rho_\psi(\mathbf{s},\mathbf{s}')$$
$$+(\alpha^{-1}-1)\delta_{\mathbf{s}=\mathbf{s}'}\}\boldsymbol{\Sigma}, \tag{17}$$

where $\rho_\psi(\cdot,\cdot)$ is a spatial correlation function defined through hyperparameter ψ, δ denotes Dirac's delta function, and $\alpha^{-1}\boldsymbol{\Sigma}$ is the nonspatial covariance matrix of $\mathbf{y}(\mathbf{s})$. The fixed scalar α represents the proportion of total variability allocated to the spatial process. This implies that $\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathrm{MN}_{n,q}(\mathbf{X}\boldsymbol{\beta}, \mathcal{K}, \boldsymbol{\Sigma})$, where $\mathcal{K} = \rho_\psi(\mathcal{S},\mathcal{S}) + (\alpha^{-1}-1)\mathbf{I}_n$. We model $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ using the conjugate MNIW prior

$$\boldsymbol{\beta} \mid \boldsymbol{\Sigma} \sim \mathrm{MN}_{p,q}(\boldsymbol{\mu}_\beta, \mathbf{V}_r, \boldsymbol{\Sigma}), \; \boldsymbol{\Sigma} \sim \mathrm{IW}(\boldsymbol{\Psi}, \nu), \tag{18}$$

with prefixed $\{\boldsymbol{\mu}_\beta, \mathbf{V}_r, \boldsymbol{\Psi}, \nu\}$. Closely following the developments in Gamerman and Moreira (2004), we obtain the posterior distribution of $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ as $\mathrm{MNIW}(\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Psi}^*, \nu^*)$, where

$$\mathbf{V}^* = (\mathbf{X}^\top \mathcal{K}^{-1}\mathbf{X} + \mathbf{V}_r^{-1})^{-1}, \; \boldsymbol{\mu}^* = \mathbf{V}^*(\mathbf{X}^\top \mathcal{K}^{-1}\mathbf{Y} + \mathbf{V}_r^{-1}\boldsymbol{\mu}_\beta),$$

$$\boldsymbol{\Psi}^* = \boldsymbol{\Psi} + \mathbf{Y}^\top \mathcal{K}^{-1}\mathbf{Y} + \boldsymbol{\mu}_\beta^\top \mathbf{V}_r^{-1}\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^{*\top}\mathbf{V}^{*-1}\boldsymbol{\mu}^*,$$

$$\text{and } \nu^* = \nu + n. \tag{19}$$

We refer to the above model as the "response" model.

Next, we consider the spatial regression model with the latent process,

$$\mathbf{y}(\mathbf{s}) = \boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s}) + \omega(\mathbf{s}) + \epsilon(\mathbf{s}), \; \mathbf{s} \in \mathcal{D}, \tag{20}$$

where $\omega(\mathbf{s}) \sim \mathrm{GP}(\mathbf{0}_{q\times 1}, \rho_\psi(\cdot,\cdot)\boldsymbol{\Sigma})$ is a latent process and $\epsilon(\mathbf{s}) \sim \mathrm{N}(\mathbf{0}_{q\times 1}, (\alpha^{-1}-1)\boldsymbol{\Sigma})$ is measurement error. Define $\omega = \omega(\mathcal{S}) = [\omega(\mathbf{s}_1) : \cdots : \omega(\mathbf{s}_n)]^\top$. For theoretical tractability, we restrict posterior inference on $\{\boldsymbol{\beta}, \omega, \boldsymbol{\Sigma}\}$, assuming that the scalar α is fixed. Assuming that the joint distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are given in (3) and that $\omega \mid \boldsymbol{\Sigma} \sim \mathrm{MN}_{n\times q}(\mathbf{0}, \rho_\psi(\mathcal{S},\mathcal{S}), \boldsymbol{\Sigma})$, the posterior distribution of $\boldsymbol{\gamma}^\top = [\boldsymbol{\beta}^\top, \omega^\top]$ is $p(\boldsymbol{\gamma}, \boldsymbol{\Sigma} \mid \mathbf{Y}) = \mathrm{MNIW}(\boldsymbol{\mu}_\gamma^*, \mathbf{V}^*, \boldsymbol{\Psi}^*, \nu^*)$,

where

$$\mathbf{V}^* = \begin{bmatrix} \frac{\alpha}{1-\alpha}\mathbf{X}^\top \mathbf{X} + \mathbf{V}_r^{-1} & \frac{\alpha}{1-\alpha}\mathbf{X}^\top \\ \frac{\alpha}{1-\alpha}\mathbf{X} & \rho_\psi^{-1}(\mathcal{S},\mathcal{S}) + \frac{\alpha}{1-\alpha}\mathbf{I}_n \end{bmatrix}^{-1},$$

$$\boldsymbol{\mu}_\gamma^* = \mathbf{V}^* \begin{bmatrix} \frac{\alpha}{1-\alpha}\mathbf{X}^\top \mathbf{Y} + \mathbf{V}_r^{-1}\boldsymbol{\mu}_\beta \\ \frac{\alpha}{1-\alpha}\mathbf{Y} \end{bmatrix}, \tag{21}$$

$$\boldsymbol{\Psi}^* = \boldsymbol{\Psi} + \frac{\alpha}{1-\alpha}\mathbf{Y}^\top \mathbf{Y} + \boldsymbol{\mu}_\beta^\top \mathbf{V}_r^{-1}\boldsymbol{\mu}_\beta - \boldsymbol{\mu}_\gamma^{*\top}\mathbf{V}^{*-1}\boldsymbol{\mu}_\gamma^*$$

$$\text{and } \nu^* = \nu + n.$$

We refer to the above model as the "latent" model.

We establish the posterior consistency of $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ for the response model (17) and the latent model (20). For distinguishing the variables based on the number of observations, we make the dependence upon n explicit. Denote $\mathbf{X}(n)_{n\times p} = [\mathbf{x}(\mathbf{s}_1) : \cdots : \mathbf{x}(\mathbf{s}_n)]^\top$, $\mathbf{Y}(n)_{n\times q} = [\mathbf{y}(\mathbf{s}_1) : \cdots : \mathbf{y}(\mathbf{s}_n)]^\top$, $S(n) = \{\mathbf{s}_1, ..., \mathbf{s}_n\}$, $\mathcal{K}(n) = \mathbf{C}(S(n), S(n)) + (\alpha^{-1}-1)\mathbf{I}_n$, and $\mathbf{J}(n) = \mathbf{X}(n)^\top \mathcal{K}(n)^{-1}\mathbf{X}(n)$. Proofs and technical details are available in Supplementary Appendix C.

Theorem 1. *[Theorem S.1, Theorem S.2] Parameter set $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ is posterior consistent for both conjugate response and latent models if and only if $\lim_{n\to\infty}\lambda_{\min}\{\mathbf{J}(n)\} = \infty$, where $\lambda_{\min}\{\mathbf{J}(n)\}$ is the smallest eigenvalue of $\mathbf{J}(n)$.*

When the explanatory variables share the same spatial correlation with the responses, the necessary and sufficient conditions for Theorem 1 hold (see Remark S.2). When the explanatory variables are themselves regarded as independent observations, the necessary and sufficient conditions in Theorem 1 hold (see Remark S.3).

4 | SIMULATION

We present two simulation examples. The first compares BLMC model with other multivariate Bayesian spatial models. The second assesses our BLMC model when K is not excessively large. BLMC models were implemented in Julia 1.2.0 (Bezanson *et al.*, 2017). We modeled the univariate processes in the proposed BLMC by NNGP. We took the BLMC model proposed by Schmidt and Gelfand (2003) as a benchmark in the first simulation example. The benchmark model was implemented in R 3.4.4 through function *spMisalignLM* in the R package *spBayes* (Finley *et al.*, 2007). The posterior inference for each model was based on MCMC chains with 5000 iterations after a burn-in of 5000 iterations. All models were run on a single 8 Intel Core i7-7700K CPU @ 4.20GHz processor with

32 GB of random-access memory running Ubuntu 18.04.2 LTS. Convergence diagnostics and other posterior summaries were implemented within the Julia statistical environment. Model comparisons were based on parameter estimates (posterior mean and 95% credible interval), root mean squared prediction error (RMSPE), mean squared error of intercept-centered latent processes (MSEL), prediction interval coverage (CVG; the percent of intervals containing the true value), interval coverage for intercept-centered latent process of observed response (CVGL), average continuous rank probability score (CRPS; see Gneiting and Raftery, 2007) for responses, and the average interval score (INT; see Gneiting and Raftery, 2007) for responses and run time. We assessed convergence of MCMC chains by visually monitoring autocorrelations and checking the accuracy of parameter estimates using effective sample size (ESS) (Gelman *et al.*, 2013, section 10.5) and Monte Carlo standard errors (MCSE) with batch size 50 (Flegal *et al.*, 2008). To calculate the CRPS and INT, we assumed that the associated predictive distribution was well approximated by a Gaussian distribution with mean centered at the predicted value and standard deviation equal to the predictive standard error. All NNGP models were specified with at most $m = 10$ nearest neighbors.

4.1 | Simulation Example 1

We simulated the response $\mathbf{y}(\mathbf{s})$ from the LMC model in (2) with $q = 2, p = 2, K = 2$ over 1200 randomly generated locations over a unit square. The size of the data set was kept moderate to enable comparisons with the expensive full GP-based LMC models for experiments conducted on the computing setup described earlier. The explanatory variable $\mathbf{x}(\mathbf{s})$ consists of an intercept and a single predictor generated from a standard normal distribution. An exponential correlation function was used to model $\{\rho_{\psi_k}(\cdot, \cdot)\}_{k=1}^{K}$, that is, $\rho_{\psi_k}(\mathbf{s}, \mathbf{s}') = \exp(-\phi_k \|\mathbf{s} - \mathbf{s}'\|)$, for $\mathbf{s}, \mathbf{s}' \in \mathcal{D}$, where $\|\mathbf{s} - \mathbf{s}'\|$ is the Euclidean distance between $\mathbf{s}$ and $\mathbf{s}'$, and $\psi_k = \phi_k$ is the decay for each k. We randomly picked 200 locations for predicting each response to examine the predictive performance. Supplementary Appendix D presents the fixed parameters generating the data and the subsequent posterior estimates.

For NNGP-based BLMC model, we assigned a flat prior for $\boldsymbol{\beta}$, which makes $\mathbf{L}_{\boldsymbol{\beta}}^{-1}$ in (10) a zero matrix. The prior for $\boldsymbol{\Lambda}$ followed (3) with $\boldsymbol{\mu}_{\boldsymbol{\Lambda}}$ a zero matrix and $\mathbf{V}_{\boldsymbol{\Lambda}}$ a diagonal matrix whose diagonal elements are 25. The prior for $\boldsymbol{\Sigma}$ was set to follow $IW(\boldsymbol{\Psi}, \nu)$ with $\boldsymbol{\Psi} = \text{diag}([1.0, 1.0])$ and $\nu = 3$. For the benchmark LMC, we assigned a flat prior for $\boldsymbol{\beta}$, $IW(\boldsymbol{\Psi}, \nu)$ with $\boldsymbol{\Psi} = \text{diag}([1.0, 1.0])$ and $\nu = 3$ for the cross-covariance matrix $\boldsymbol{\Lambda}^{\top}\boldsymbol{\Lambda}$, and $IG(2, 0.5)$ for each

diagonal element of $\boldsymbol{\Sigma}$. We assigned $\text{unif}(2.12, 212)$ as priors of decays for both models. This implies that the "effective spatial range," which is the distance where spatial correlation drops below 0.05, will be bounded above by $\sqrt{2}$ (the maximum intersite distance within a unit square) and bounded below by 1/100th of that to ensure a wide range.

Table 1 presents posterior estimates of parameters and performance metrics for all candidate models. Both models provided similar posterior inferences for $\{\boldsymbol{\beta}_{21}, \boldsymbol{\beta}_{21}\}$. The 95% credible intervals of $\{\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{12}\}$ all include the true value used to generate the data. The NNGP-based BLMC model and the benchmark LMC model cost 2.38 min and around 18.25 h, respectively. Despite the shorter running time, we observed superior performance of the NNGP-based BLMC than the benchmark LMC for inferring on the latent process using CVGL, MSEL, CRPSL, and INTL. Moreover, the interpolated map of the recovered intercept-centered latent processes (Figure 1) by BLMC and benchmark LMC are almost indistinguishable from each other. BLMC and benchmark LMC produce very similar RMSPEs, CRPSs, and INTs. The differences in estimates between the two models is likely emerging from the different prior settings and sampling schemes. Benchmark LMC restricts the loading matrix $\boldsymbol{\Lambda}$ to be upper triangular, while BLMC does not, resulting in greater flexibility in fitting latent process. On the other hand, the unidentifiable parameter setting of BLMC causes somewhat less stable inference for the hyperparameters $\{\phi_1, \phi_2\}$. The inferences for $\{\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{12}\}$ are also less stable due to the sensitivity of intercept to latent process. For all other parameters including the intercept-centered latent process on 1200 locations, the median ESS is 4111.5. All MCSEs were consistently less than 0.02. These diagnostics suggest adequate convergence of the MCMC algorithm.

4.2 | Simulation Example 2

We generated 100 different data sets using (2) with $\{q = 10, p = 3, K = 50\}$ and a diagonal $\boldsymbol{\Sigma}$ (i.e., independent measurement errors across outcomes). Supplementary Appendix D presents the parameter values used to generate the data sets. We fixed a set of 1200 irregularly situated locations inside a unit square. The explanatory variable $\mathbf{x}(\mathbf{s})$ comprised an intercept and two predictors generated independently from a standard normal distribution. The same set of locations and explanatory variables were used for the 100 data sets. Each $f_k(\mathbf{s})$ was generated using an exponential covariance function, $\{\rho_{\psi_k}(\cdot, \cdot)\}_{k=1}^{K}$, where $\psi_k = \phi_k$ was the decay for $k = 1, \dots, K$. We held out 200 locations for assessing predictive performances.

(a) $\boldsymbol{\omega}_1 + \boldsymbol{\beta}_{11}$ true

(b) $\boldsymbol{\omega}_1 + \boldsymbol{\beta}_{11}$ BLMC

(c) $\boldsymbol{\omega}_1 + \boldsymbol{\beta}_{11}$ benchmark LMC

(d) $\boldsymbol{\omega}_2 + \boldsymbol{\beta}_{12}$ true

(e) $\boldsymbol{\omega}_2 + \boldsymbol{\beta}_{12}$ BLMC

(f) $\boldsymbol{\omega}_2 + \boldsymbol{\beta}_{12}$ benchmark LMC

(g) fitted correlation with $K = 2$

(h) fitted correlation with $K = 4$

(i) fitted correlation with $K = 6$

(j) fitted correlation with $K = 8$

(k) fitted correlation $K = 10$

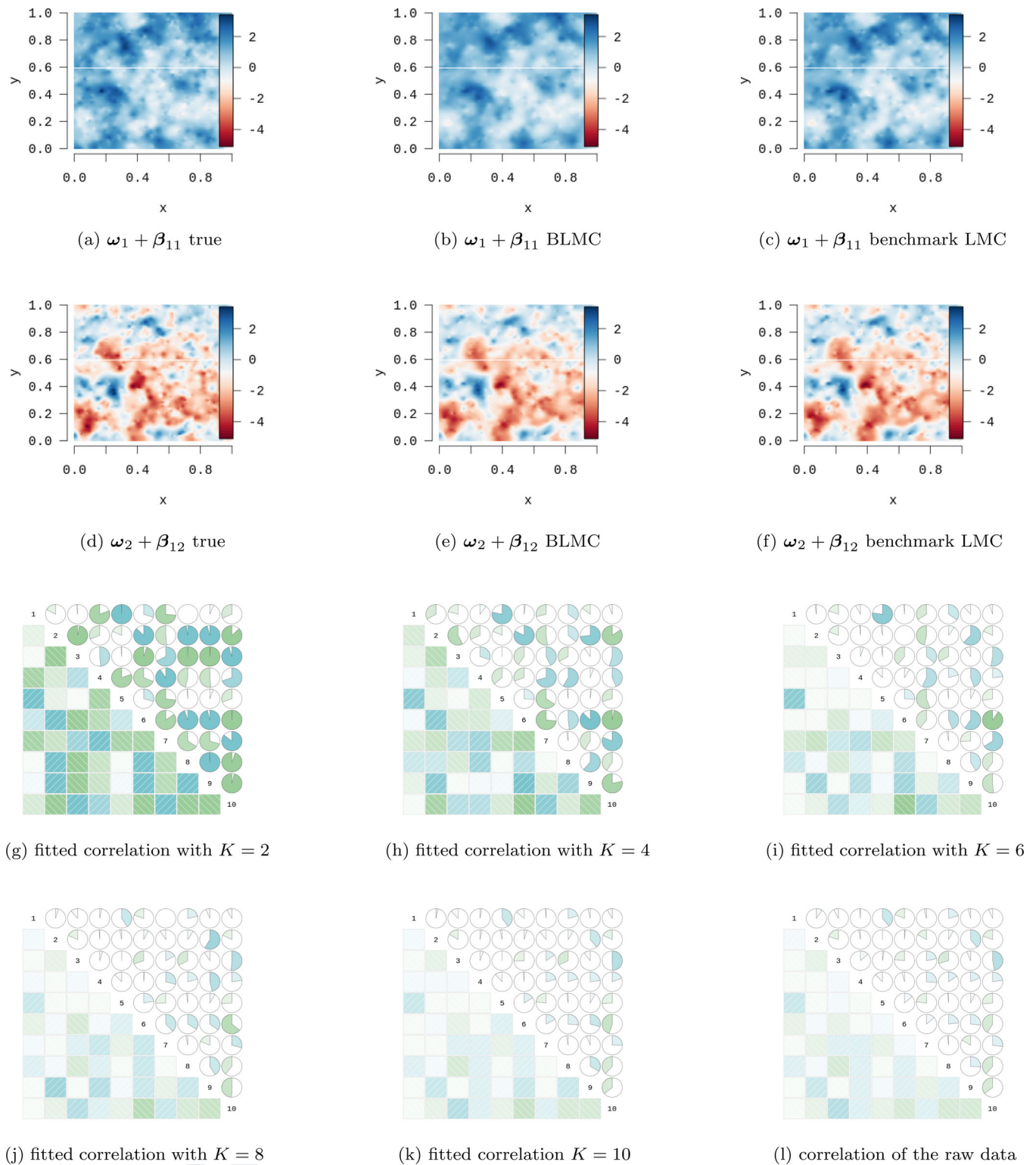(l) correlation of the raw data

FIGURE 1 Interpolated maps of (a) & (d) the true generated intercept-centered latent processes, the posterior means of the intercept-centered latent process $\boldsymbol{\omega}$ from the (b) & (e) NNGP-based BLMC model and the (c) & (f) benchmark LMC model. Heat-maps of the (l) actual finite sample correlation among latent processes and (g)–(k) posterior mean of finite sample correlation among latent processes based on the posterior samples of $\boldsymbol{\Omega}$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

TABLE 1 Simulation study summary table: posterior mean (2.5%, 97.5%) percentiles

		BLMC		Benchmark LMC	
	True	Inference	MCSE	Inference	MCSE
β_{11}	1.0	0.705 (0.145, 1.233)	0.034	0.806 (0.502, 1.131)	0.002
β_{12}	−1.0	−1.24 (−1.998, −0.529)	0.045	−1.1 (−1.533, −0.646)	0.001
β_{21}	−5.0	−4.945 (−5.107, −4.778)	0.002	−4.949 (−5.113, −4.787)	0.004
β_{22}	2.0	1.979 (1.78, 2.166)	0.004	1.974 (1.785, 2.167)	0.002
Σ_{11}	0.4	0.346 (0.283, 0.409)	0.002	0.306 (0.248, 0.364)	0.003
Σ_{12}	0.15	0.133 (0.072, 0.194)	0.003	0.0	–
Σ_{22}	0.3	0.29 (0.198, 0.386)	0.004	0.233 (0.159, 0.334)	0.005
ϕ_1	6.0	8.723 (4.292, 14.065)	0.343	12.839 (8.805, 17.471)	0.23
ϕ_2	18.0	22.63 (15.901, 29.555)	0.416	18.075 (12.99, 23.741)	0.301
RMSPE[a]	–	[0.728, 0.756, 0.742]		[0.725, 0.762, 0.744]	
MSEL[b]	–	[0.136, 0.168, 0.152]		[0.147, 0.192, 0.169]	
CRPS[a]	–	[−0.412, −0.423, −0.418]		[−0.41, −0.427, −0.418]	
CRPSL[b]	–	[−0.035, −0.038, −0.036]		[−0.216, −0.248, −0.232]	
CVG[a]	–	[0.915, 0.955, 0.935]		[0.925, 0.96, 0.9425]	
CVGL[b]	–	[0.946, 0.962, 0.954]		[0.756, 0.773, 0.765]	
INT[a]	–	[3.378, 3.756, 3.567]		[3.347, 3.823, 3.585]	
INTL[a]	–	[0.282, 0.329, 0.305]		[1.875, 2.023, 1.949]	
Time (s)		143		[42047, 23664][c]	

a[response 1, response 2, all responses].
bintercept + latent process on 1000 observed locations for [response 1, response 2, all responses].
c[time for MCMC sampling, time for recovering predictions].

TABLE 2 Simulation study summary table 2: posterior mean (2.5%, 97.5%) percentiles

$K =$	1	2	3	4
CVGL	0.28 (0.14, 0.93)	0.39 (0.17, 0.94)	0.49 (0.2, 0.95)	0.58 (0.25, 0.96)
CVG	0.95 (0.92, 0.98)	0.95 (0.92, 0.98)	0.95 (0.92, 0.98)	0.95 (0.92, 0.98)
RMSPE	2.07 (1.94, 2.18)	1.96 (1.86, 2.05)	1.87 (1.78, 1.97)	1.78 (1.7, 1.85)
MCSE	0.004 (0.002, 0.008)	0.005 (0.002, 0.01)	0.005 (0.003, 0.01)	0.004 (0.003, 0.01)
$K =$	5	6	7	8
CVGL	0.67 (0.29, 0.96)	0.74 (0.33, 0.96)	0.81 (0.38, 0.96)	0.86 (0.44, 0.96)
CVG	0.95 (0.92, 0.98)	0.95 (0.92, 0.98)	0.95 (0.92, 0.98)	0.95 (0.91, 0.98)
RMSPE	1.7 (1.62, 1.77)	1.63 (1.55, 1.69)	1.56 (1.5, 1.63)	1.51 (1.45, 1.57)
MCSE	0.005 (0.003, 0.01)	0.005 (0.003, 0.013)	0.005 (0.003, 0.011)	0.005 (0.003, 0.011)
$K =$	9	10		
CVGL	0.91 (0.59, 0.96)	0.95 (0.92, 0.96)		
CVG	0.95 (0.91, 0.98)	0.95 (0.91, 0.98)		
RMSPE	1.46 (1.41, 1.51)	1.43 (1.38, 1.48)		
MCSE	0.005 (0.003, 0.01)	0.005 (0.003, 0.01)		

For each simulated data set, we fitted the BLMC model specifying a diagonal Σ with K from 1 to 10. Each ϕ_k has a Gamma prior with shape and scale equaling 2 and 4.24, respectively, so that the expected effective spatial range is half of the maximum inter-site distance. We assigned flat prior for β, a vague prior for Λ which follows the prior of Λ in the preceding example and IG(2, 1.0) priors for the diagonal elements of Σ.

The posterior mean and the 95% credible interval of CVGL, CVG, RMSPE, and diagnostics metric MCSE for regression slopes and Σ for 100 simulation studies are summarized by K in Table 2. Inference for CVG and MCSE

were robust to the choice of K. All of the 95% credible intervals for CVG and MCSE were within [0.9, 0.99] and [0.0, 0.02], respectively. As shown in Table 2, the performance metrics were quickly improved as K increased from 1 to 10. On average, RMSPE decreased by about 30.9% and CVGL increased from 28% to 95%. Given that our data come from an LMC model with $K = 50$, we can conclude that BLMC with diagonal Σ is efficient in obtaining inference for the latent processes even when K is not adequately large. We also create heat-maps of the posterior mean of our finite sample correlation matrix among the latent processes based on posterior samples of Ω, where $\Omega = \frac{1}{n}\sum_{i=1}^{n}(\omega(\mathbf{s}_i) - \bar{\omega})(\omega(\mathbf{s}_i) - \bar{\omega})^{\top}$ with $\bar{\omega}$ the vector of column means of ω. Figures 1 (g)–1 (k) depict such heat maps from one of the 100 simulated data sets. As K increases from 2 to 10, the estimated correlation matrix approaches the true correlation matrix. The plots also reveal that the performance of BLMC is sensitive to the choice of K. We recommend choosing K based on scientific considerations for the problem at hand and exploratory data analyses, or checking the RMSPE value for different K and picking K by an elbow rule (Thorndike, 1953).

5 | REMOTE-SENSED VEGETATION DATA ANALYSIS

We apply our proposed models to analyze NDVI and enhanced vegetation indices (EVIs) measuring vegetation activity on the land surface, which can help us understand the global distribution of vegetation types as well as their biophysical and structural properties and spatial variations. Apart from vegetation indices, we consider gross primary productivity (GPP) data, global terrestrial evapotranspiration (ET) product, and land cover data (see Ramon Solano *et al.*, 2010; Mu *et al.*, 2013; Sulla-Menashe and Friedl, 2018, for further details). The geographic coordinates of our variables were mapped on a Sinusoidal (SIN) projection grid. We focus on zone *h08v05*, which covers 11 119 505 to 10 007 555 m south of the prime meridian and 3 335 852 to 4 447 802 m north of the equator. The land is situated in the western United States. Our explanatory variables included an intercept and a binary indicator for no vegetation or urban area through the 2016 land cover data. All other variables were measured through the MODIS satellite over a 16-day period from 2016.04.06 to 2016.04.21. Some variables were rescaled and transformed in exploratory data analysis for the sake of better model fitting. The data sets were downloaded using the R package *MODIS,* and the code for the exploratory data analysis is provided as supplementary material to this article.

Our data comprise 1 020 000 observed locations to illustrate the proposed model. Our spatially dependent outcomes were the transformed NDVI (log(NDVI + 1) labeled as NDVI) and red reflectance (red refl). A Bayesian multivariate regression model, defined by (2) excluding $\Lambda^{\top}\mathbf{f}(\mathbf{s})$, was also fitted for comparisons. All NNGP-based models used $m = 10$ nearest neighbors. We randomly held out 10% of each response and then held all responses over the region 10 400 000 to 10 300 000 m south of the prime meridian and 3 800 000 to 3 900 000 m north of the equator to evaluate the models' predictive performance over a missing region (white square) and randomly missing locations. Figure 2 (a) illustrates the map of the transformed NDVI data.

We fit both models with 5000 iterations after 5000 iterations as burn-in. The priors for all parameters except decays followed those in the Section 4. We assigned Gamma(200, 0.02) and Gamma(200, 0.04) for ϕ_1 and ϕ_2 for BLMC based on fitted variograms to the raw data. All the codes were run with single thread. No other processes were simultaneously run so as to provide an accurate measure of computing time.

Table 3 presents results on the BLMC. The regression coefficients of the index of no vegetation or urban area show relatively low biomass (low NDVI) and high red reflectance over no vegetation or urban area. Estimates of Σ and the finite sample process covariance matrix Ω, as defined in Section 4.2, show a negative association between the residuals and latent processes of transformed NDVI and red reflectance, which satisfies the underlying relationship between two responses. BLMC captured a high negative correlation (≈ -0.87) between the latent processes of two responses, indicating that the spatial pattern of the latent processes of NDVI and red reflectance are almost the reverse of each other. The maps of the latent processes recovered by BLMC, presented in Figure 2, also support this relationship.

We provide RMSPE, CVG, CRPS, INT, MCSE, and run time in Table 3. Apparently BLMC substantially improved predictive accuracy. BLMC's RMSPEs were over 50% less than the Bayesian linear model. CVG is similar between two models, while INT and CRPS also favored BLMC over the Bayesian linear model. Figure 2 presents the estimated latent processes from BLMC. Notably, the BLMC smooths out the predictions in the held-out region. The model's run time was around 38.6 h, which is still impressive given the full model-based analysis it offers for such a massive multivariate spatial data set.

We also fitted a BLMC with diagonal Σ to explore the underlying latent processes of 10 (transformed) responses: (i) NDVI, (ii) EVI, (iii) GPP, (iv) net photosynthesis (PsnNet), (v) red reflectance (red refl), (vi) blue reflectance (blue refl), (vii) average daily global ET, (viii) latent heat flux (LE), (ix) potential ET (PET), and (x) potential
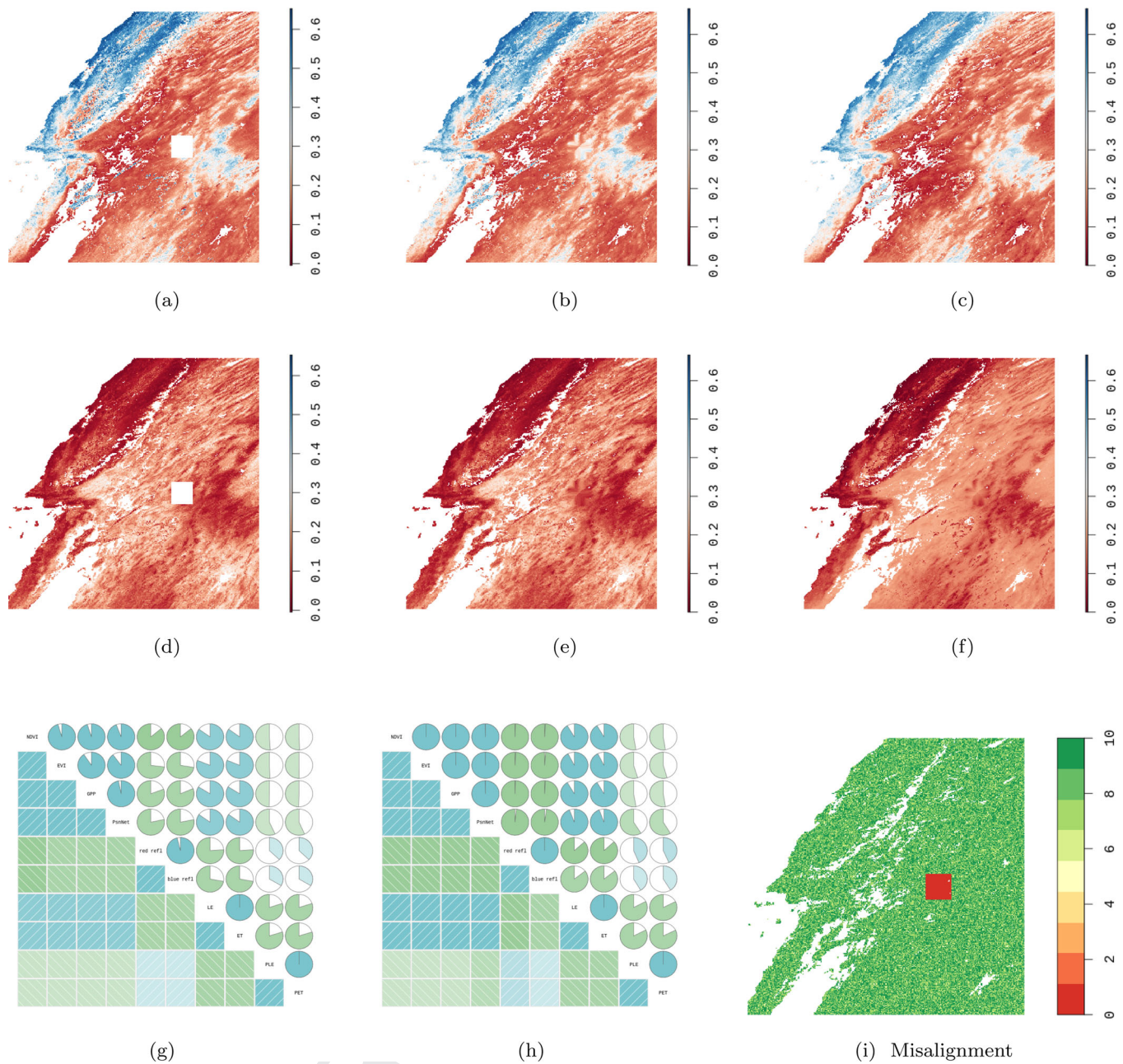
FIGURE 2 Colored NDVI and red reflectance images of western United States (zone h08v05). Maps of raw data (a) & (d) and the posterior mean of the intercept-centered latent process recovered from (b) & (e) BLMC and (c) & (f) BLMC with diagonal Σ. Correlation of responses (g) and posterior mean of finite sample correlation among latent processes from the BLMC model with diagonal Σ (h). Heat-map (i) of counts of observed response. Each observed location is labeled with a color dot whose color represents the count of observed responses on the location. The greener the color, the higher the count. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

LE (PLE). There are, in total, 12 057 locations with no responses and 656 366 observed locations with misaligned data (at least one but not all responses), which cover 65.12% of observed locations. We provide a heat-map (Figure 2 (i)) to present the status of misalignment over the study domain.

Based on the exploratory analysis, we observed two groups of responses that have high within-group correla-

tions but relatively low between-group correlations (see Figure 2 (g)). Hence we picked $K = 2$. Estimates from the BLMC model are presented in Table 4. No vegetation or urban area exhibits lower vegetation indexes (lower NDVI and EVI) and lower production of chemical energy in organic compounds by living organisms (lower GPP and PsnNet). We observe a trend of higher blue reflectance, red reflectance, ET (higher ET LE), and lower potential

TABLE 3 Vegetation data analysis summary Table 1: posterior mean (2.5%, 97.5%) percentiles

	Bayesian linear model inference	BLMC Inference	MCSE
Intercept_1	0.2515 (0.2512, 0.2517)	0.1433 (0.1418, 0.1449)	1.145e-4
Intercept_2	0.1395 (0.1394, 0.1396)	0.1599 (0.159, 0.1608)	6.17e-5
$\text{No vege or urban area}_1$	−0.1337 (−0.1346, −0.1328)	−1.385e-2 (−1.430e-2, −1.342e-2)	1.69e-5
$\text{No vege or urban area}_2$	6.035e-2 (5.992e-2, 6.075e-2)	7.831e-3 (7.584e-3, 8.097e-3)	8.24e-6
Σ_{11}	1.599e-2 (1.594e-2, 1.603e-2)	3.514e-4 (3.477e-4, 3.553e-4)	1.93e-7
Σ_{12}	−6.491e-3 (−6.512e-3, −6.471e-3)	−1.084e-4 (−1.100e-4, −1.067e-4)	8.19e-8
Σ_{22}	3.656e-3 (3.646e-3, 3.667e-3)	1.074e-4 (1.063e-4, 1.084e-4)	4.79e-8
Ω_{11}	–	1.675e-2 (1.674e-2, 1.676e-2)	4.17e-7
Ω_{12}	–	−6.873e-3 (−6.879e-3, −6.867e-3)	1.77e-7
Ω_{22}	–	3.764e-3 (3.760e-3, 3.768e-3)	9.06e-8
ϕ_1	–	3.995 (3.887, 4.075)	7.535e-3
ϕ_2	–	12.376 (11.512, 13.320)	7.60e-3
RMSPE[a]	[0.074, 0.0359, 0.0581]	[0.0326, 0.0171, 0.0260]	
CRPS[a]	[−0.04135, −0.01988, −0.03061]	[−0.01561, −0.00879, −0.0122]	
CVG[a]	[0.956, 0.958, 0.957]	[0.954, 0.947, 0.950]	
INT[a]	[0.3468, 0.1711, 0.2589]	[0.1965, 0.0995, 0.1480]	
Time (min)	11	2318	

a[1st response transformed NDVI, 2nd response red reflectance, all responses].

TABLE 4 Vegetation data analysis summary Table 2: posterior mean (2.5%, 97.5%)

Response	Slope	MCSE	Nugget (Σ_{ii})	MCSE
NDVI	−0.0120 (−0.0124, −0.0116)	1.37e-5	7.46e-4 (7.42e-4, 7.49e-4)	7.60e-8
EVI	−4.38e-3 (−4.68e-3, −4.08e-3)	6.86e-6	8.68e-4 (8.65e-4, 8.7e-4)	3.07e-8
GPP	−0.197 (−0.199, −0.194)	8.31e-5	0.0244 (0.0243, 0.0245)	2.34e-6
PsnNet	−4.48e-3 (−5.39e-3, −3.50e-3)	3.42e-5	5.34e-3 (5.32e-3, 5.36e-3)	3.50e-7
red refl	4.49e-3 (4.20e-3, 4.77e-3)	5.11e-6	9.84e-4 (9.81e-4, 9.87e-4)	3.13e-8
blue refl	0.0123 (0.0121, 0.0124)	2.74e-6	2.60e-4 (2.59e-4, 2.61e-4)	8.81e-9
LE	0.0908 (0.0884, 0.0932)	1.36e-4	0.0531 (0.0529, 0.0533)	2.37e-6
ET	0.0919 (0.0895, 0.0944)	1.49e-4	0.0531 (0.053, 0.0533)	2.27e-6
PLE	−3.64e-3 (−3.98e-3, −3.36e-3)	5.50e-5	2.095e-5 (2.086e-5, 2.104e-5)	1.63e-9
PET	−4.88e-3 (−5.99e-3, −3.96e-3)	1.81e-4	6.50e-5 (6.44e-5, 6.57e-5)	2.20e-8

ET (lower PET PLE) in urban area and area with no vegetation. We provide maps of posterior predictions for all 10 variables in Supplementary Appendix F. The latent processes corresponding to transformed NDVI and red reflectance fitted in two analyses in Figure 2 share a similar pattern. Finally, the heat map of the posterior mean of the finite sample correlation among the latent processes (elements of Ω as defined in Section 4.2) based on BLMC with diagonal Σ, presented in Figure 2 (h), reveals a high underlying correlation among NDVI, EVI, GPP, PsnNet, red and blue reflectance, and that LE and ET are slightly more correlated with NDVI and EVI than PLE and PET. The total run time for BLMC with diagonal Σ was around 60.7 h (3642.25 min).

6 | SUMMARY AND DISCUSSION

We have proposed scalable models for analyzing massive and possibly misaligned multivariate spatial data sets. Our framework offers flexible covariance structures and scalability by modeling the loading matrix of spatial factors using matrix-normal distributions and the factors themselves as NNGPs. This process-based formulation allows us to resolve spatial misalignment by fully model-based imputation. Through a set of simulation examples and an analysis of a massive misaligned data set comprising remote-sensed variables, we demonstrated the inferential and computational benefits accrued from our proposed framework.

This work can be expanded further in at least two important directions. The first is to extend the current methods to spatiotemporal data sets, where multiple variables are indexed by spatial coordinates, as considered here, as well as by temporal indices. Associations are likely to be exhibited across space and time as well as among the variables within a location and time-point. In addition, these variables are likely to be misaligned across time and space. Regarding the scalability of the spatiotemporal process, we can build a dynamic nearest-neighbor Gaussian process (DNNGP) (Datta *et al.*, 2016) to model spatiotemporal factors and one can also envisage temporal dependence on the loading matrix.

A second direction will consider spatially varying coefficient models. We model the regression coefficients β using a spatial (or spatiotemporal) random field to capture spatial (or spatiotemporal) patterns in how some of the predictors impact the outcome. We can assign the prior of the regression coefficients β using a multivariate Gaussian random field with a proportional cross-covariance function. Then the prior of β over observed locations follows a matrix-normal distribution, which is the prior we designed for β in all of the proposed models in this article. While the modification seems to be easy, the actual implementation requires a more detailed exploration, and we leave these topics for further explorations.

From a computational perspective, we clearly need to further explore high-performance computing and high-dimensional spatial models amenable to such platforms. The programs provided in this work are for illustration and have limited usage in graphical processing units (GPUs) computing and parallelized CPU computing. A parallel CPU computing algorithm for the BLMC model can simultaneously sample multiple MCMC chains, improving the performance of the actual implementations. Implementations with modeling methods such as MRA (Katzfuss, 2017) also require dedicated programming with GPU. Other scalable modeling methods that build graphical Gaussian models on space, time, and the number of variables can lead to sparse models for high-dimensional multivariate data and scale not only up to millions of locations and time points, but also to hundreds or even thousands of spatially or spatiotemporally oriented variables. The idea here will be to extend current developments in Vecchia-type models to graphs building dependence among a large number of variables so that the precision matrices across space, time, and variables are sparse. Research on scalable statistical models and high-performance computing algorithms for such models will be of substantial interest to statisticians and environmental scientists.

DATA AVAILABILITY STATEMENT

The simulated and remote-sensed vegetation index data that support the findings in this article are openly available at GitHub at https://github.com/LuZhangstat/Multi_NNGP, Zhang (2021).

ORCID

Lu Zhang https://orcid.org/0000-0001-6829-703X
Sudipto Banerjee https://orcid.org/0000-0002-2239-208X

REFERENCES

Banerjee, S. (2017) High-dimensional Bayesian geostatistics. *Bayesian Analysis*, 12, 583–614.

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: CRC Press.

Banerjee, S. and Gelfand, A. (2002) Prediction, interpolation and regression for spatially misaligned data. *Sankhya: The Indian Journal of Statistics, Series A*, 64, 227–245.

Bezanson, J., Edelman, A., Karpinski, S. and Shah, V.B. (2017) Julia: a fresh approach to numerical computing. *SIAM Review*, 59, 65–98.

Bourgault, G. and Marcotte, D. (1991) Multivariable variogram and its application to the linear model of coregionalization. *Mathematical Geology*, 23, 899–928.

Chiles, J.-P. and Delfiner, P. (2009). *Geostatistics: Modeling Spatial Uncertainty*, 2nd edition. New York: John Wiley & Sons.

Cressie, N. and Wikle, C.K. (2015). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: John Wiley & Sons.

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.

Datta, A., Banerjee, S., Finley, A.O., Hamm, N. A.S. and Schaap, M. (2016) Non-separable dynamic nearest-neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *Annals of Applied Statistics*, 10, 1286–1316.

Dawid, A.P. (1981) Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68, 265–274.

Finley, A.O., Banerjee, S. and Carlin, B.P. (2007) spbayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19, 1.

Flegal, J.M., Haran, M. and Jones, G.L. (2008) Markov chain Monte Carlo: can we trust the third significant figure? *Statistical Science*, 23(2), 250–260.

Gamerman, D. and Moreira, A.R. (2004) Multivariate spatial regression models. *Journal of Multivariate Analysis*, 91, 262–281.

Gelfand, A., Schmidt, A., Banerjee, S. and Sirmans, C. (2004) Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 13, 263–312.

Gelfand, A.E. and Banerjee, S. (2010) Multivariate spatial process models. In: Gelfand, A., Diggle, P., Fuentes, M. and Guttorp, P. (Eds.) *Handbook of Spatial Statistics*. Boca Raton, FL: CRC Press, pp. 495–516.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis, Texts in Statistical Science*. 3rd edition. Boca Raton, FL: Chapman & Hall/CRC.

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

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

Goulard, M. and Voltz, M. (1992) Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, 24, 269–286.

Haario, H., Saksman, E. and Tamminen, J. (2005) Componentwise adaptation for high dimensional MCMC. *Computational Statistics*, 20, 265–273.

Heaton, M.J., Datta, A., Finley, A.O., Furrer, R., Guinness, J., Guhaniyogi, R., et al. (2019) A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24, 398–425.

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

Le, N., Sun, L. and Zidek, J.V. (2001) Spatial prediction and temporal backcasting for environmental fields having monotone data patterns. *Canadian Journal of Statistics*, 29, 529–554.

Le, N.D., Sun, W. and Zidek, J.V. (1997) Bayesian multivariate spatial interpolation with data missing by design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59, 501–510.

Le, N.D. and Zidek, J.V. (2006). *Statistical Analysis of Environmental Space-Time Processes*. New York: Springer Science & Business Media.

Lopes, H.F., Salazar, E. and Gamerman, D. (2008) Spatial dynamic factor analysis. *Bayesian Analysis*, 3(4), 759–792.

Lopes, H.F. and West, M. (2004) Bayesian model assessment in factor analysis. *Statistica Sinica*, 14, 41–67.

Mu, Q., Zhao, M. and Running, S.W. (2013) Modis global terrestrial evapotranspiration (ET) product (nasa mod16a2/a3). Algorithm Theoretical Basis Document, Collection, 5.

Nishimura, A. and Suchard, M.A. (2018) Prior-preconditioned conjugate gradient method for accelerated gibbs sampling in "large n & large p" sparse Bayesian regression. arXiv preprint arXiv:1810.12437 .

Ramon Solano, R., Didan, K., Jacobson, A. and Huete, A. (2010) *Modis Vegetation Index User's Guide*. Tucson, AZ: The University of Arizona.

Ren, Q. and Banerjee, S. (2013) Hierarchical factor models for large spatially misaligned datasets: a low-rank predictive process approach. *Biometrics*, 69, 19–30.

Roberts, G.O. and Rosenthal, J.S. (2009) Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18, 349–367.

Salvaña, M.L.O. and Genton, M.G. (2020) Nonstationary cross-covariance functions for multivariate spatio-temporal random fields. *Spatial Statistics*, 37, 100411.

Schmidt, A.M. and Gelfand, A.E. (2003) A Bayesian coregionalization approach for multivariate pollutant data. *Journal of Geophysical Research: Atmospheres*, 108, 8783.

Sulla-Menashe, D. and Friedl, M.A. (2018) *User Guide to Collection 6 Modis Land Cover (mcd12q1 and mcd12c1) Product*. Reston, VA: USGS, pp. 1–18.

Sun, W., Le, N.D., Zidek, J.V. and Burnett, R. (1998) Assessment of a Bayesian multivariate interpolation approach for health impact studies. *Environmetrics: The Official Journal of the International Environmetrics Society*, 9, 565–586.

Sun, Y., Li, B. and Genton, M. (2011) Geostatistics for large datasets. In: Montero, J., Porcu, E. and Schlather, M. (Eds.), *Advances and Challenges in Space-Time Modelling of Natural Events*. Berlin Heidelberg: Springer-Verlag, pp. 55–77.

Taylor-Rodriguez, D., Finley, A.O., Datta, A., Babcock, C., Andersen, H.E., Cook, B.D., et al. (2019) Spatial factor models for high-dimensional and large spatial data: an application in forest variable mapping. *Statistica Sinica*, 29, 1155–1180.

Thorndike, R.L. (1953) Who belongs in the family. *Psychometrika*, 18, 267–276.

Wackernagel, H. (2003). *Multivariate Geostatistics*, 3 Edition. Berlin: Springer-Verlag.

Wang, F. and Wall, M.M. (2003) Generalized common spatial factor model. *Biostatistics*, 4, 569–582.

Zhang, L. (2021). *Remote-Sensed Vegetation data*. https://github.com/LuZhangstat/Multi_NNGP/tree/master/RDA/data/rawdata.

Zhang, L., Banerjee, S. and Finley, A.O. (2020) High-dimensional multivariate geostatistics: a Bayesian matrix-normal approach. arXiv preprint arXiv:2003.10051.

SUPPORTING INFORMATION

The Web Appendices referenced in Sections 2.4, 3, 4 and 5, the MODIS vegetation indices data analyzed in Section 5, and the Julia code implementing our models are available with this paper at the Biometrics website on Wiley Online Library. The data and code are also available at https://github.com/LuZhangstat/Multi_NNGP.