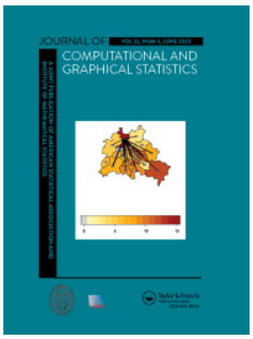

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Modeling Massive Highly Multivariate Nonstationary Spatial Data with the Basis Graphical Lasso

Mitchell L. Krock, William Kleiber, Dorit Hammerling & Stephen Becker

To cite this article: Mitchell L. Krock, William Kleiber, Dorit Hammerling & Stephen Becker (2023): Modeling Massive Highly Multivariate Nonstationary Spatial Data with the Basis Graphical Lasso, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2023.2174126

To link to this article: https://doi.org/10.1080/10618600.2023.2174126

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2023, VOL. 00, NO. 0, 1–16
https://doi.org/10.1080/10618600.2023.2174126

Modeling Massive Highly Multivariate Nonstationary Spatial Data with the Basis Graphical Lasso

Mitchell L. Krock[a], William Kleiber[b], Dorit Hammerling[c], and Stephen Becker[b]

[a]Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL; [b]Department of Applied Mathematics, University of Colorado Boulder, Boulder, CO; [c]Department of Applied Mathematics and Statistics, Colorado School of Mines, Golden, CO

ABSTRACT

We propose a new modeling framework for highly multivariate spatial processes that synthesizes ideas from recent multiscale and spectral approaches with graphical models. The basis graphical lasso writes a univariate Gaussian process as a linear combination of basis functions weighted with entries of a Gaussian graphical vector whose graph is estimated from optimizing an ℓ_1 penalized likelihood. This article extends the setting to a multivariate Gaussian process where the basis functions are weighted with Gaussian graphical vectors. We motivate a model where the basis functions represent different levels of resolution and the graphical vectors for each level are assumed to be independent. Using an orthogonal basis grants linear complexity and memory usage in the number of spatial locations, the number of basis functions, and the number of realizations. An additional fusion penalty encourages a parsimonious conditional independence structure in the multilevel graphical model. We illustrate our method on a large climate ensemble from the National Center for Atmospheric Research's Community Atmosphere Model that involves 40 spatial processes. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2021
Accepted January 2023

KEYWORDS
Climate ensemble; Graphical model; Multivariate Gaussian process; Nonstationary; Spatial basis function

1. Introduction

The past 20 years have seen a surge of interest in developing models for multivariate spatial processes. The major obstacle lies in defining valid cross-covariance functions that can characterize complex interactions between multiple processes; the primary difficulty is that the cross-covariance and marginal covariance functions must work together to provide a nonnegative definite matrix function. Most research has focused on exploring new models or new approaches for defining cross-covariances that are valid for a handful of processes. Many applied problems, such as those in statistical climatology, involve highly-multivariate datasets with dozens to hundreds of variables, and existing approaches fail as strategies to model and understand relevant dependencies between variables. With growing availability of ensemble-based geophysical model output (Eyring et al. 2016), nonstationary spatial data with many realizations (i.e., replicates) of a large spatial field are increasingly common. We propose a method which is able to model multiple variables of a large climate ensemble. Previous techniques are unable to deal with a dataset of this magnitude and to capture nonstationarity in space and between variables.

To introduce the basic ideas, let us fix some notation. The p-variate observational Gaussian process model under consideration is

$$\begin{pmatrix} Y_1(\mathbf{s}) \\ \vdots \\ Y_p(\mathbf{s}) \end{pmatrix} = \begin{pmatrix} \mu_1(\mathbf{s}) \\ \vdots \\ \mu_p(\mathbf{s}) \end{pmatrix} + \begin{pmatrix} Z_1(\mathbf{s}) \\ \vdots \\ Z_p(\mathbf{s}) \end{pmatrix} + \begin{pmatrix} \varepsilon_1(\mathbf{s}) \\ \vdots \\ \varepsilon_p(\mathbf{s}) \end{pmatrix} \tag{1}$$

or in vector form, $\mathbf{Y}(\boldsymbol{s}) = \boldsymbol{\mu}(\boldsymbol{s}) + \mathbf{Z}(\boldsymbol{s}) + \boldsymbol{\varepsilon}(\boldsymbol{s})$. Here, $\mathbf{Y}(\boldsymbol{s}) = (Y_1(\mathbf{s}), \ldots, Y_p(\mathbf{s}))^{\mathrm{T}}$ is the observed process at location $\mathbf{s} \in \mathbb{R}^d$ with mean $\boldsymbol{\mu}(\mathbf{s}) = (\mu_1(\mathbf{s}), \ldots, \mu_p(\mathbf{s}))^{\mathrm{T}}$ and spatially correlated stochastic variation $\mathbf{Z}(\mathbf{s}) = (Z_1(\mathbf{s}), \ldots, Z_p(\mathbf{s}))^{\mathrm{T}}$, which we assume to be a multivariate Gaussian process. The observations are subject to noise $\boldsymbol{\varepsilon}(\mathbf{s}) = (\varepsilon_1(\mathbf{s}), \ldots, \varepsilon_p(\mathbf{s}))^{\mathrm{T}}$, a mean zero multivariate white noise process with covariance matrix $\mathrm{cov}(\boldsymbol{\varepsilon}(\mathbf{s}), \boldsymbol{\varepsilon}(\mathbf{s})) = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2)$.

The paramount issue in working with multivariate processes is specifying the matrix-valued covariance, $\mathbf{C}(\mathbf{s}_1, \mathbf{s}_2) = (C_{ij}(\mathbf{s}_1, \mathbf{s}_2))_{i,j=1}^{p}$, where $C_{ij}(\mathbf{s}_1, \mathbf{s}_2) = \mathrm{cov}(Z_i(\mathbf{s}_1), Z_j(\mathbf{s}_2))$ are the direct and cross-covariance functions. This matrix function must be carefully constrained in order to be a nonnegative definite matrix function. In particular, for arbitrary locations $\{\mathbf{s}_i\}_{i=1}^{n}$, the block matrix $\boldsymbol{\Sigma}$ with (i,j)th block $\mathbf{C}(\mathbf{s}_i, \mathbf{s}_j)$ must be nonnegative definite. Genton and Kleiber (2015) give an overview of cross-covariance functions for multivariate geostatistics which is still relatively up-to-date. Salvaña and Genton (2020) provide a more recent overview with a focus on multivariate spatio-temporal cross-covariance functions. The remainder of this article considers Gaussian process models for continuous spatial processes with no temporal component. Spatial data with locations grouped by region (e.g., by county or state) is known as areal or lattice data. We refer readers interested in high-dimensional multivariate spatio-temporal areal data to Bradley et al. (2015, 2018).

The main extant approaches to generating valid multivariate spatial models are reviewed by Genton and Kleiber (2015)

CONTACT Mitchell Krock mkrock@anl.gov Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL, USA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

and include covariance convolution and kernel convolution (Ver Hoef and Barry 1998; Gaspari and Cohn 1999; Majumdar and Gelfand 2007; Majumdar et al. 2010), the linear model of coregionalization (LMC) (Goulard and Voltz 1992; Schmidt and Gelfand 2003; Wackernagel 2003; Gelfand et al. 2004), latent dimensions (Apanasovich and Genton 2010), dynamical models (Shaddick and Wakefield 2002; Calder 2007; Ippoliti et al. 2012), conditional Bayesian hierarchical structures (Le and Zidek 2006; Pollice and Jona Lasinio 2010), and direct specification as in the multivariate Matèrn (Gneiting et al. 2010; Apanasovich et al. 2012) and nonstationary extensions (Kleiber and Nychka 2012; Kleiber and Porcu 2014). Cressie and Zammit-Mangion (2016) develop a conditional model for multivariate spatial data. Qadir and Sun (2020) and Qadir et al. (2021) work with the coherence function (Kleiber 2017) to provide more flexible behavior than the multivariate Matérn. Vu et al. (2020) construct nonstationary and asymmetric covariances by embedding familiar covariances (e.g., multivariate Matèrn) in a warped domain with warping functions obtained from deep learning. All of these models are designed to handle a few variables, typically fewer than five, and none are designed for high-dimensional multivariate spatial data.

To expand upon the previous statement, multivariate spatial data can be high-dimensional in the sense of a large number of observation locations ($n \gg 0$) and/or a large number of output variables ($p \gg 0$). Computational difficulties in the high-dimensional setting are unavoidable and particularly troublesome when both n and p are large. Even with a valid matrix-valued covariance function specified, that is already an issue if there are more than a handful of variables, likelihood computations and memory requirements for p-variate Gaussian processes at n spatial locations scale as $\mathcal{O}(p^3n^3)$ and $\mathcal{O}(p^2n^2)$, respectively. Salvaña et al. (2021) explore high-performance computing techniques to alleviate these costs. Modeling high-dimensional multivariate spatial processes requires specialized methodology, and most current techniques struggle when both n and p are large.

First, we describe some models for multivariate processes that focus on dealing with a large number of spatial locations ($n \gg 0$). Kleiber et al. (2019) generalize LatticeKrig (Nychka et al. 2015) to the multivariate case, relying on compactly supported basis functions and spatial autoregressive Gaussian Markov random field models for stochastic coefficients to handle massive spatial data. Guinness (2022) is able to avoid maximum likelihood estimation and Bayesian inference by successively imputing data to an expanded lattice domain under a periodic model and efficiently estimating the cross-spectral density using fast Fourier transform techniques. Zhang et al. (2021) perform Bayesian inference on massive multivariate spatial datasets by combining a matrix-normal distribution with Nearest Neighbor Gaussian Processes (NNGP) (Datta et al. 2016). Their model setup and choice of conjugate priors leads to closed form posterior distributions, and scalability is demonstrated on a dataset with over three million spatial locations. Guinness (2022) and Zhang et al. (2021) consider a stationary bivariate setting in their data analyses, although these methods seem in principle able to scale to more than two variables.

Next, we turn attention to the "highly multivariate" problem, where the number of variables p is large. In one of the earliest efforts to tackle this problem, Furrer and Genton (2011) introduce "aggregation cokriging" for the prediction (but not the modeling) of highly multivariate spatial processes. Dey et al. (2021) propose a "stitching" of univariate Gaussian processes that preserves the marginal behavior of each univariate process and also conditional independencies between variables implied by a Gaussian graphical model. Stitching significantly reduces computational costs associated with highly multivariate data, especially when used in conjunction with a "decomposable" sparse graphical model, and it is also amenable to parallel computing.

Finally, we note that several adaptations of LMC are capable of handling data where the number of variables and observation locations are both large. Taylor-Rodriguez et al. (2019) use NNGP to model the latent process of LMC in a two-stage Bayesian hierarchical model, which they use to connect LiDAR maps with forest measurements. Bruinsma et al. (2020) combine a sufficient statistic of the data and an orthogonal basis/loading matrix for substantial computational gains. Zhang and Banerjee (2021) also develop a Bayesian LMC factor model with NNGP in the latent process, but they propose matrix-normal prior distributions and are able to avoid some constraints on the loading matrices used by Taylor-Rodriguez et al. (2019). Meng et al. (2021a) build upon Meng et al. (2021b), where the loading matrix is stochastic with a Gaussian process prior on its elements, and employ variational inference (Titsias and Lawrence 2010) for efficient computations. Liu et al. (2021) also use variational inference for scalable computations and introduce an embedding of the latent Gaussian space in LMC into a higher-dimensional space for more flexibility. We discuss LMC in more detail in Section 2.1, as our formulation can be viewed as the opposite of LMC.

Despite these advances, no one has truly addressed the problem of modeling complicated dependencies in space and across variables when the number of variables and observed spatial locations are both large. The approaches in the previous paragraphs are all designed for modeling stationary processes, except for Meng et al. (2021a, 2021b), who do not attempt to model spatial data. Many datasets (e.g., our example from statistical climatology) exhibit strong nonstationarity both within and between processes, but models for nonstationary multivariate processes are not typically built to handle high-dimensional response vectors. Another issue is the ability to model multivariate processes on the globe, a task for which nonstationary covariance and cross-covariance functions are necessary (Jun 2011). This article fills a major gap in the current literature by introducing methodology for highly multivariate processes observed at a large number of spatial locations with nonstationary spatial and inter-variable dependencies.

We present a model for highly multivariate and nonstationary spatial data that also accommodates estimation and simulation strategies for large networks of observation locations. The essential ideas rely on representing the vector-valued process in a basis expansion with sparsity-inducing Gaussian graphical modeling of the stochastic coefficients. We propose a penalized likelihood framework for estimation and associated optimization

algorithms. In addition to nonstationarity granted by judicious choice of basis functions and stochastic coefficient structure, employing *orthogonal* basis functions allows for rapid computation with nongridded high-dimensional multivariate spatial data on an arbitrary domain. The method is illustrated on a challenging climate data science problem involving $p = 40$ spatial processes from an atmospheric model at thousands of locations over the globe that exhibit strong nonstationarities and cross-process dependencies. Our model provides straightforward interpretations of cross-process dependencies, which, for the climate data example, identify scientifically meaningful and justifiable relationships.

2. Methodology

Our approach relies on a basis expansion of the spatially-correlated components of $\mathbf{Z}(\mathbf{s})$:

$$Z_i(\mathbf{s}) = \sum_{\ell=1}^{L} W_{i\ell}\phi_\ell(\mathbf{s}) \quad (i = 1, \ldots, p) \tag{2}$$

for some classes of basis functions $\{\phi_\ell\}$ and stochastic coefficients $\{W_{i\ell}\}$. For fixed i, this representation subsumes many popular approaches in the spatial statistical literature that have been primarily explored in the univariate setting, including discretized spectral methods, low-rank approaches, and empirical orthogonal functions, among others, despite potential limitations depending upon the choice of basis function (Stein 2014). Few extensions to the multivariate case (2) have been made, with Kleiber et al. (2019) being a notable case for spatial data. Some similar ideas were explored in the context of functional data analysis (Qiao et al. 2019; Fontanella et al. 2020; Zapata et al. 2021). Dey et al. (2022) study the connection between stitching (Dey et al. 2021) and the functional Gaussian graphical model presented in Zapata et al. (2021) and this manuscript. In Section 2.1, we show how the univariate basis graphical lasso directly extends to the multivariate setting (2), and in Section 2.2, we discuss an alternative that is more appropriate for massive multivariate spatial data.

2.1. Basis Graphical Lasso

Our modeling and optimization strategy follows from extensions to our prior work, which we discuss and connect to the multivariate problem in this section. In Krock et al. (2021), we introduced the basis graphical lasso (BGL) to model (2) for the univariate case $p = 1$. The goal of the BGL is to obtain a sparse nonparametric estimate for the precision matrix Q of the mean zero Gaussian graphical vector $\mathbf{W} = (W_1, \ldots, W_L)^{\mathrm{T}}$ when $Z(\mathbf{s}) = \sum_{\ell=1}^{L} W_\ell\phi_\ell(\mathbf{s})$. In other words, we build a Gaussian process by fitting a Gaussian graphical model to the random coefficients of fixed basis functions. Recall that the sparsity pattern of the precision matrix encodes conditional independencies between random variables, with $Q_{ij} = 0$ if and only if W_i and W_j are conditionally independent given all other entries of $\mathbf{W}$. This information is commonly visualized with an undirected graph known as a Gaussian graphical model, where vertices symbolize variables and the lack of an edge between two vertices indicates such a conditional independence (Rue and Held 2005). Inspired by the graphical lasso (Friedman et al. 2008), we use an ℓ_1 penalized likelihood to estimate the graphical model, but in the standard graphical lasso setting $\mathbf{W}$ is observed directly, whereas our model includes basis functions and noise. Our BGL method is viable for a very large spatial sample size (n), and multiple realizations (m) are preferred but not required. We claim that the BGL generalizes to the multivariate setting and also that the optimization routine enjoys a similar computational framework.

With $\mathbf{W}_\ell = (W_{1\ell}, \ldots, W_{p\ell})^{\mathrm{T}} \sim N(\mathbf{0}, \mathbf{Q}_\ell^{-1})$, the multivariate basis model (2) alternatively can be written

$$\mathbf{Z}(\mathbf{s}) = \sum_{\ell=1}^{L} \phi_\ell(\mathbf{s})\mathbf{W}_\ell \tag{3}$$

so that each basis function is weighted with a p-variate random vector. Assuming the weight vectors are independent means that this model amounts to characterizing the inverse covariance matrices $\{\mathbf{Q}_1, \ldots, \mathbf{Q}_L\}$. In contrast, the standard LMC considers (3) where $\{\phi_1(\mathbf{s}), \ldots, \phi_L(\mathbf{s})\}$ are independent Gaussian processes with deterministic weights $\{\mathbf{W}_1, \ldots, \mathbf{W}_L\}$. We note that recent variants of LMC (Taylor-Rodriguez et al. 2019; Bruinsma et al. 2020; Zhang and Banerjee 2021; Meng et al. 2021a; Liu et al. 2021) can handle a large number of variables, and in fact orthogonal basis functions are also exploited for computational gains in Bruinsma et al. (2020), but with a large number of spatial locations, such techniques ultimately amount to efficiently modeling spatially dense univariate Gaussian processes (e.g., Titsias 2009; Datta et al. 2016). The semiparametric latent factor model (Teh et al. 2005) shares a setup similar to LMC where a multi-output Gaussian process is represented as a linear mixture of independent univariate Gaussian processes. With these models, covariance kernel parameters of the independent latent Gaussian processes and the weight vectors (i.e., columns of the loading matrix) must be estimated. Typically, the covariance kernels are stationary and the loading matrix is deterministic. Some nonstationary versions of LMC (Gelfand et al. 2004; Meng et al. 2021a) consider a stochastic loading matrix that depends on the input domain, which entails a prior distribution and a challenging Bayesian framework. To our knowledge, such models have not yet been tested with high-dimensional nonstationary spatial data. On the other hand, our approach associates each basis function with a p-variate Gaussian graphical model, which is much faster from a computational point of view and automatically produces nonstationary covariance and cross-covariance functions with straightforward interpretations.

Let us assume $\boldsymbol{\mu} \equiv 0$ in (1) for simplicity of exposition. In a typical mean function regression context, we can use generalized least squares and profiled likelihoods to estimate the regression coefficients. Given data at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, form the observation vector $\mathbf{Y} = (\mathbf{Y}(\mathbf{s}_1)^{\mathrm{T}}, \ldots, \mathbf{Y}(\mathbf{s}_n)^{\mathrm{T}})^{\mathrm{T}}$. Suppose we have multiple independent realizations $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ of $\mathbf{Y}$. Note that all methodology developed in this paper can work with $m = 1$ realization, but the task of learning the precision matrix of the basis weight vector is better suited for a setting with multiple realizations. Krock et al. (2021) investigated the sensitivity of the basis graphical lasso to the number of realizations. Up to multiplicative and additive constants not depending on the $np \times$

np variance-covariance matrix $\mathbf{\Sigma} = \text{var}(\mathbf{Y})$, the negative log-likelihood is

$$\log\det\mathbf{\Sigma} + \frac{1}{m}\sum_{i=1}^{m}\mathbf{Y}_i^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{Y}_i.$$

Define the sample covariance $\mathbf{S} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}$. Using the cyclic property of trace, we rewrite the negative log-likelihood as

$$\log\det\mathbf{\Sigma} + \text{tr}(\mathbf{S}\mathbf{\Sigma}^{-1})$$

to align with more prevalent notation in graphical lasso literature. Note that $\mathbf{Y}$ is simply a linear combination of the random coefficient vector $\mathbf{W} = (\mathbf{W}_1^{\mathrm{T}},\ldots,\mathbf{W}_L^{\mathrm{T}})^{\mathrm{T}}$ and basis functions, plus noise. That is, $\mathbf{Y} = \mathbf{\Phi}\mathbf{W} + \boldsymbol{\varepsilon}$ for a basis matrix $\mathbf{\Phi}$ and noise vector $\boldsymbol{\varepsilon}$, so $\mathbf{\Sigma} = \mathbf{\Phi}\mathbf{Q}^{-1}\mathbf{\Phi}^{\mathrm{T}} + \mathbf{D}$, where $\text{var}(\mathbf{W}) = \mathbf{Q}^{-1}$ and $\text{var}(\boldsymbol{\varepsilon}) = \mathbf{D}$. The matrices $\mathbf{Q}, \mathbf{\Phi}$, and $\mathbf{D}$ are defined explicitly in Section 3, but it is important to realize here that the matrix algebra produces the same optimization problem as in the univariate case presented in Krock et al. (2021).

The original BGL solves the ℓ_1-penalized maximum likelihood equation

$$\hat{\mathbf{Q}} \in \underset{\mathbf{Q}\succeq 0}{\arg\min}\left(\log\det(\mathbf{\Phi}\mathbf{Q}^{-1}\mathbf{\Phi}^{\mathrm{T}} + \mathbf{D}) + \text{tr}(\mathbf{S}(\mathbf{\Phi}\mathbf{Q}^{-1}\mathbf{\Phi}^{\mathrm{T}} + \mathbf{D})^{-1}) + \|\Lambda \circ \mathbf{Q}\|_1\right). \quad (4)$$

Here, we retain the boldface matrix font to emphasize that the multivariate case entails the same optimization problem, even though the original BGL was proposed for the univariate setting $p = 1$. The notation $\mathbf{Q} \succeq 0$ indicates that $\mathbf{Q}$ is positive semidefinite, and $\|\Lambda \circ \mathbf{Q}\|_1 = \sum_{i,j}\Lambda_{ij}|\mathbf{Q}_{ij}|$, where Λ_{ij} are nonnegative penalty parameters that encourage sparsity in the estimate. Evaluating (4) requires an expensive $\mathcal{O}(p^3n^3)$ Cholesky decomposition. After applying the matrix determinant lemma, the Sherman-Morrison-Woodbury formula, and the cyclic property of trace, we can equivalently minimize

$$\log\det\left(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{\Phi}\right) - \log\det\mathbf{Q} - \text{tr}\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}\mathbf{\Phi}(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{\Phi})^{-1}\right) + \|\Lambda \circ \mathbf{Q}\|_1. \quad (5)$$

Once the matrices $\mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{\Phi}$ and $\mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}\mathbf{\Phi}$ are computed, evaluating (5) only requires Cholesky decompositions in the dimension pL, so we can reduce likelihood evaluations to $\mathcal{O}(p^3L^3)$. However, (5) is nonsmooth and nonconvex with respect to $\mathbf{Q}$, so the minimization problem is nontrivial.

Studying the convexity/concavity structure[1] of (5) suggests a difference-of-convex (DC) algorithm where the next guess for $\mathbf{Q}$ is obtained by solving a convex optimization problem with the concave part linearized at the previous guess. Such an algorithm reads

$$\mathbf{Q}^{(j+1)} = \underset{\mathbf{Q}\succeq 0}{\arg\min}\left(-\log\det\mathbf{Q} + \text{tr}\left(\mathbf{\Psi}^{(j)}\mathbf{Q}\right) + \|\Lambda \circ \mathbf{Q}\|_1\right), \quad (6)$$

where the linearization matrix

$$\mathbf{\Psi}^{(j)} = (\mathbf{Q}^{(j)} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{\Phi})^{-1} + (\mathbf{Q}^{(j)} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}\mathbf{\Phi}(\mathbf{Q}^{(j)} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{\Phi})^{-1} \quad (7)$$

is a function of the previous guess $\mathbf{Q}^{(j)}$ and the aforementioned precomputed matrices. Since a DC algorithm such as (6) is a majorization-minimization algorithm, we are guaranteed that the guesses for $\mathbf{Q}$ create a nonincreasing sequence in the objective function (5). Moreover, (6) is a well-studied problem known as the graphical lasso. Typically, the graphical lasso uses the sample covariance matrix of directly observed, nonnoisy variables to produce a sparse inverse covariance matrix and accordingly a graphical model for the variables. Here we iteratively call the graphical lasso algorithm to estimate a graph for latent basis weights (with additive noise in the observational model), and the linearization matrix (7) acts as the sample covariance in the algorithm. We solve the graphical lasso with the second-order method `QUIC` (Hsieh et al. 2014b).

2.2. Multivariate Basis Graphical Lasso

Although the previous section shows that the BGL can be readily extended to the multivariate setting, the generalization is not well-motivated by a connection to standard multivariate spatial models, and moreover it will require burdensome matrix calculations in the dimension pL. In particular, the BGL must compute the linearization matrix (7) and substitute it into the graphical lasso at each step of the DC algorithm. With our climate data example we use $L = 2000$ basis functions and $p = 40$ variables; an 80,000 dimensional precision matrix is too large for this procedure. Advances in graphical modeling (Fattahi et al. 2019) may allow for estimation of graphs of this magnitude, but storing the dense linearization matrix poses an issue to further scalability. We conclude the article with more commentary about this direct generalization (see Section 5), but here we propose a similar DC algorithm that is more feasible in a highly multivariate setting.

Our basic model still follows a penalized likelihood-based framework, minimizing

$$\hat{\mathbf{Q}} \in \underset{\mathbf{Q}\succeq 0}{\arg\min}\left(\log\det(\mathbf{\Phi}\mathbf{Q}^{-1}\mathbf{\Phi}^{\mathrm{T}} + \mathbf{D}) + \text{tr}(\mathbf{S}(\mathbf{\Phi}\mathbf{Q}^{-1}\mathbf{\Phi}^{\mathrm{T}} + \mathbf{D})^{-1}) + P(\mathbf{Q})\right)$$

for some convex penalty P. However, an ℓ_1 graphical lasso-type penalty by itself does not impose any regularity on the *structure* of coefficient graphs. To motivate our proposal, we recall some recent insights into multivariate modeling that will suggest an appropriate form for P.

The multivariate spectral representation theorem states

$$\mathbf{Z}(\mathbf{s}) = \int \exp(i\boldsymbol{\omega}^{\mathrm{T}}\mathbf{s})\mathbf{W}(\mathrm{d}\boldsymbol{\omega}) \quad (8)$$

for a mean zero stationary process $\mathbf{Z}(\mathbf{s})$, where $\mathbf{W}(\cdot)$ is a complex-valued mean zero random measure vector (Stein 1999). Taking a discretization of the integral, we approximate

$$\int \exp(i\boldsymbol{\omega}^{\mathrm{T}}\mathbf{s})\mathbf{W}(\mathrm{d}\boldsymbol{\omega}) \approx \sum_{\ell}\cos(\boldsymbol{\omega}_\ell^{\mathrm{T}}\mathbf{s} + \theta_\ell)\mathbf{W}_\ell \quad (9)$$

[1] See the appendix of Krock et al. (2021) for classification of convexity/concavity for the terms in (5).

to motivate writing (3) as a linear combination of *independent* random vectors $\mathbf{W}_\ell$, where the coefficient of $\mathbf{W}_\ell$ depends on the phase shift θ_ℓ. As an aside, another justification for modeling the coefficient vectors $\mathbf{W}_\ell$ as independent across ℓ is the Karhunen-Loéve expansion, in which the basis functions ϕ_ℓ are eigenfunctions and the random coefficients are theoretically independent. Indeed, in our climate modeling example below, we use a discrete approximation to the Karhunen-Loéve expansion from which independence of coefficient vectors is expected.

The spectral representation theorem (8) is intimately linked to the spectral density matrix $\mathbf{f}(\cdot)$ of $\mathbf{Z}(\mathbf{s})$ where we identify $\text{var}(\mathbf{W}_\ell)$ with $\mathbf{f}(\boldsymbol{\omega}_\ell)$. One connection to our model is that $\mathbf{Q}_\ell^{-1}$ can be viewed as the spectral density matrix at frequency $\boldsymbol{\omega}_\ell$, but our approach estimates the *inverse* spectral density matrix in a Gaussian graphical framework, and moreover, we consider replacing the harmonic basis functions with other globally supported multiresolution basis functions. Indeed, Kleiber (2017) provides interpretation and exploration of spectral coherence that will additionally motivate our penalized likelihood implementation. Before moving on, it is important to note that although we use the spectral representation theorem (8) to motivate the ensuing approach, our method is general and extends beyond harmonic basis expansions but with similar coherence-like interpretations of coefficient dependence. Guinness (2022) proposes a multivariate space-time model with flexible coherence structures that uses LMC in the spectral domain, but the covariance structure is stationary and the method is viable only for gridded data. We remind the reader that LMC also relies on an independence assumption where $\mathbf{Z}(\mathbf{s})$ is a linear combination of independent univariate Gaussian processes.

The multivariate basis graphical lasso model can be motivated with the same penalized likelihood context as in Section 2.1. Recall the model setup: we work under the additive model (1) with $\mathbf{Z}(\mathbf{s})$ specified as in the basis representation (3), and $\mathbf{W}_i$ and $\mathbf{W}_j$ are independent Gaussian graphical vectors for $i \neq j$. In particular, we assume that $\mathbf{Q}_\ell = \text{var}(\mathbf{W}_\ell)^{-1}$ is a sparse matrix defining a graphical structure at level ℓ. If we consider each $\mathbf{Q}_\ell$ to correspond to an arbitrary sparse graphical model, then we propose P as a graphical lasso regularization for each level:

$$P(\mathbf{Q}_1, \ldots, \mathbf{Q}_L) = \lambda \sum_{\ell=1}^{L} \sum_{i \neq j} \left|(\mathbf{Q}_\ell)_{ij}\right|. \tag{10}$$

This penalty enforces sparsity for each precision matrix but not necessarily any similarity between levels of resolution. Recent development in spectral coherence (Kleiber 2017) suggests that we should expect the coherence of processes arising in practice to vary smoothly across levels. In addition to the ℓ_1 sparsity penalty, we include an ℓ_1 sequentially fused penalty to encourage similarity of the conditional independence structure across adjacent levels of resolution:

$$P(\mathbf{Q}_1, \ldots, \mathbf{Q}_L) = \lambda \sum_{\ell=1}^{L} \sum_{i \neq j} \left|(\mathbf{Q}_\ell)_{ij}\right| + \rho \sum_{\ell=1}^{L-1} \sum_{i \neq j} \left|(\mathbf{Q}_\ell)_{ij} - (\mathbf{Q}_{\ell+1})_{ij}\right|. \tag{11}$$

The new fusion penalty with tuning parameter ρ penalizes precision matrices at adjacent levels if their off-diagonals do not have the same value. This formulation suggests a smoothly-varying graph structure and produces a parsimonious conditional independence structure of the random weights over all levels of resolution. As $\rho \to \infty$, the second penalty effectively shrinks toward a stationary process. Although the final model will be nonstationary, we believe shrinking toward a stationary process is a natural approach to regularization. This type of penalty is most appropriate when basis functions can be ordered meaningfully to represent different levels of resolution, as is the case with an EOF basis.

Assuming that $\mathbf{Q} = \text{diag}(\mathbf{Q}_1, \ldots, \mathbf{Q}_L)$ does not change any reasoning leading to the BGL formulation from Section 2.1 but allows us to reduce computations on a matrix of size $pL \times pL$ to computations on L $p \times p$ matrices. At each step of the DC algorithm (6), we simplify the minimization problem to

$$\underset{\mathbf{Q}_\ell \succeq 0,\ \ell=1,\ldots,L}{\arg\min} \left(\left(\sum_{\ell=1}^{L} -\log\det \mathbf{Q}_\ell + \text{tr}\left(\boldsymbol{\Psi}_\ell \mathbf{Q}_\ell\right) \right) + P(\mathbf{Q}_1, \ldots, \mathbf{Q}_L) \right), \tag{12}$$

where $\boldsymbol{\Psi}_\ell$ is the ℓth block diagonal of the linearization matrix $\boldsymbol{\Psi}$, which depends upon the previous graph guesses. We emphasize that in (6), all matrices are of dimension $pL \times pL$, which can be challenging in three ways: the $\mathcal{O}((pL)^2)$ memory cost to store matrices of this size, the $\mathcal{O}((pL)^3)$ time cost to compute matrix solves for the linearization (7), and finally solving the convex matrix optimization problem. Moreover, the latter two steps are performed at each iteration of the DC algorithm. On the other hand, (12) involves L matrices of dimension $p \times p$, which provides a major improvement over (6) in both memory and computational costs, assuming that calculation of $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_L$ is not prohibitive. An efficient way to compute $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_L$ for orthogonal bases is presented in Section 3.

Using only the sparsity penalty (10) means (12) separates into L independent graphical lasso problems with "sample covariance" matrices $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_L$.[2] Using the fusion penalty (11) gives no such separation of (12) into L independent optimization problems at each DC iteration. Instead, we treat $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_L$ as an array of sample covariance matrices and substitute them into the fused multiple graphical lasso (FMGL) (Yang et al. 2015) for modeling multiple similar graphical models across multiple datasets. A related idea is the joint graphical lasso (Danaher et al. 2014), but their similarity-inducing regularization term penalizes all pairs of graphs rather than just adjacent graphs as in (11), suggesting behavior similar to white noise processes—an unreasonable assumption for most real spatial data applications. The associated algorithm uses a slower optimization approach that calculates eigendecompositions in R. The FMGL algorithm, written in MATLAB, uses the same second-order approximation as `QUIC` (Hsieh et al. 2014b) and also exhibits local quadratic convergence. A general framework for second-order optimization in high-dimensional statistical modeling with regularization is developed in Hsieh et al. (2014a).

[2] See Section 3.1 for a description of this estimation procedure.

3. Implementation Strategy

We highlight some important aspects of implementing the multivariate BGL model. In practice, we independently estimate an error variance for each variable and define the diagonal matrix $\mathbf{D} = I_n \otimes \text{diag}(\tau_1^2, \ldots, \tau_p^2)$ accordingly.[3] Matrices $\mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{\Phi}$ and $\mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}\mathbf{\Phi}$ must be computed effectively—ignoring $\mathbf{D}^{-1}$, the naive matrix multiplications cost $\mathcal{O}(p^3nL^2)$ and $\mathcal{O}(p^3n^2L)$, respectively. Moreover, the block diagonals of the linearization (7) must be retrieved without expending $\mathcal{O}(p^3L^3)$ flops for the matrix inverse $(\mathbf{Q} + \mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{\Phi})^{-1}$. In Krock et al. (2021) with $p = 1$ this linearization step was not an issue—the computational bottleneck was iteratively solving the graphical lasso in the dimension of the basis functions L. Here we are modeling L graphs of dimension p, so the graphical modeling step may no longer be the bottleneck but rather the linearization (7). However, assuming that $\mathbf{Q}$ is block diagonal and basis functions are orthogonal makes this linearization step trivial.

Let Φ be the $n \times L$ basis matrix with (i, j) entry $\phi_j(\mathbf{s}_i)$. Observe that

$$\mathbf{Z}(\mathbf{s}_i) = \sum_{\ell=1}^{L} \phi_\ell(\mathbf{s}_i)\mathbf{W}_\ell = \mathbf{M}\Phi_i^\text{T} \qquad (i = 1, \ldots, n),$$

where $\mathbf{M}$ has columns $\mathbf{W}_1, \ldots, \mathbf{W}_L$ and Φ_i^T is the ith column of Φ^T. Introducing the vec$(\cdot)$ operator, which stacks the columns of a matrix one-by-one into a vector, we write the process observation vector as

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}(\mathbf{s}_1) \\ \vdots \\ \mathbf{Z}(\mathbf{s}_n) \end{pmatrix} = \begin{pmatrix} \mathbf{M}\Phi_1^\text{T} \\ \vdots \\ \mathbf{M}\Phi_n^\text{T} \end{pmatrix} = \text{vec}(\mathbf{M}\Phi^\text{T})$$

where the last equation follows from (16.2.7) of Harville (1997). The vec operator cooperates with the Kronecker product $\otimes$ in the following way: $\text{vec}(ABC) = (C^\text{T} \otimes A)\text{vec}(B)$ whenever ABC is well-defined. For us, this implies

$$\mathbf{Z} = (\Phi \otimes I_p)\mathbf{W}$$

since $\mathbf{W} = \text{vec}(\mathbf{M})$ by construction. Thus, we identify $\mathbf{\Phi} = \Phi \otimes I_p$. Also note that $\mathbf{D}^{-1} = I_n \otimes \text{diag}(\tau_1^{-2}, \ldots, \tau_p^{-2})$ is a Kronecker product. Since $(A \otimes C)(B \otimes D) = AB \otimes CD$ whenever AB and CD are well-defined, we have the Kronecker product representation $\mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{\Phi} = \Phi^\text{T}\Phi \otimes \text{diag}(\tau_1^{-2}, \ldots, \tau_p^{-2})$, which will be sparse for any choice of basis functions. In general, solving systems with $\mathbf{Q} + \mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{\Phi}$ does not have an exploitable structure since $\mathbf{Q}$ is block diagonal yet $\mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{\Phi}$ is a Kronecker product. However, using orthogonal basis functions (i.e., $\Phi^\text{T}\Phi = I_L$) means that $\mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{\Phi}$ is diagonal, so the first term in (7) is block diagonal and can be easily inverted in $\mathcal{O}(Lp^3)$.

The linearization also involves $\mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}\mathbf{\Phi} = (\mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{Y}_\text{data})(\mathbf{Y}_\text{data}^\text{T}\mathbf{D}^{-1}\mathbf{\Phi})/m$, where $\mathbf{Y}_\text{data}$ is the $np \times m$ data matrix with columns of realizations $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$. Using $\text{vec}(ABC) = (C^\text{T} \otimes A)\text{vec}(B)$ again,

$$\mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{Y}_i = \text{vec}(\text{diag}(\tau_1^{-2}, \ldots, \tau_p^{-2})\text{mat}(\mathbf{Y}_i)\Phi)$$

where $\text{mat}(\mathbf{Y}_i)$ is the $p \times n$ matrix with $\text{vec}(\text{mat}(\mathbf{Y}_i)) = \mathbf{Y}_i$. So $\mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{Y}_\text{data}$ can be calculated in $\mathcal{O}(mnpL)$ and has low storage cost $\mathcal{O}(mpL)$. Once $\mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{Y}_\text{data}$ and the block diagonals of $(\mathbf{Q} + \mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{\Phi})^{-1}$ are computed, the block diagonals of the second term in (7) can be computed in $\mathcal{O}(Lp^2m)$.

To summarize, for an orthogonal basis, only $\{\mathbf{Q}_1, \ldots, \mathbf{Q}_L\}$, $\{\tau_1^{-2}, \ldots, \tau_p^{-2}\}$, and $\mathbf{\Phi}^\text{T}\mathbf{D}^{-1}\mathbf{Y}_\text{data}$ must be stored in memory, and we can compute the block diagonals of (7) in $\mathcal{O}(Lp^3 + Lp^2m)$. These L $p \times p$ block matrices are then sent into the FMGL as an array of L "sample covariance" matrices, and the solution to the FMGL problem is the next guess for $\{\mathbf{Q}_1, \ldots, \mathbf{Q}_L\}$, and the entire procedure is repeated until $\|\mathbf{Q}^{(j+1)} - \mathbf{Q}^{(j)}\|_\text{F} / \|\mathbf{Q}^{(j)}\|_\text{F} < \epsilon$, which we set as $\epsilon = 0.05$.

3.1. Initial Guess (Unfused Estimate)

When considering the fusion penalty, a natural initial guess is the corresponding unfused estimate. With $\rho = 0$, the main optimization (12) amounts to solving the graphical lasso independently by level:

$$\underset{\mathbf{Q}_\ell \succeq 0,\ \ell=1,\ldots,L}{\arg\min} \left(\sum_{\ell=1}^{L} -\log\det \mathbf{Q}_\ell + \text{tr}\,(\mathbf{\Psi}_\ell \mathbf{Q}_\ell) + \lambda \sum_{i \neq j} \left|(\mathbf{Q}_\ell)_{ij}\right| \right).$$

The entire algorithm has linear complexity and storage in L in this case.

3.2. Maximum Likelihood Estimate

It is also easy to obtain the unpenalized maximum likelihood estimates for $\mathbf{Q}_1, \ldots, \mathbf{Q}_L$. Recall that $S^{-1} = \arg\min_{Q \succeq 0} -\log\det Q + \text{tr}(SQ)$, assuming S is nonsingular. That is, the maximum likelihood estimate of the precision matrix of a multivariate Gaussian random variable is the inverse of the sample covariance matrix, assuming it is nonsingular. This means that our DC algorithm would simply invert the linearization matrix (or each block diagonal of the linearization matrix in the multivariate case) rather than substitute it into the graphical lasso. Again, $\mathbf{Q}_1, \ldots, \mathbf{Q}_L$ are independent, which implies the same storage and complexity as the unfused estimate in Section 3.1. This algorithm takes many more DC iterations to converge than either of the regularized estimates.

3.3. Estimating the Error Variances

Given a single variable with sample covariance S, we minimize the following function jointly over τ^2 and a few parameters describing a diagonal matrix Q:

$$\log\det\left(Q + \tau^{-2}\Phi^\text{T}\Phi\right) - \log\det Q \\ - \text{tr}\left(\tau^{-4}\Phi^\text{T}S\Phi\left(Q + \tau^{-2}\Phi^\text{T}\Phi\right)^{-1}\right) + n\log\tau^2 + \tau^{-2}\text{tr}(S).$$

Non-boldface font is used here since the estimation is performed separately for each variable. Note that this expression can be simplified with orthogonal bases, and the trace terms can be quickly computed as squared Frobenius norms. The diagonal parameterization of Q and parameter estimation is further discussed in the supplementary material. We record values for $\tau_1^2, \ldots, \tau_p^2$ in Table 1 in the supplementary material that are used throughout the rest of the article. A similar approach used in

[3] See Section 3.3 for a description of this estimation procedure.

Krock et al. (2021) was found to be successful in recovering error variances even with a misspecified spatial covariance structure.

3.4. Cross-Validation

First we describe the cross-validation procedure for a pair of penalty parameters (λ, ρ). Suppose we use k folds and consider t arbitrary pairs of penalties represented by $\{\Lambda_1, \ldots, \Lambda_t\}$. Let $\hat{\mathbf{Q}}_{\Lambda_j}(\mathbf{S})$ be the estimate we get from applying our algorithm with empirical covariance $\mathbf{S} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}$ and penalty pair Λ_j. For $A \subseteq \{1, \ldots, m\}$, let $\mathbf{S}_A = |A|^{-1}\sum_{i\in A} \mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}$. We seek Λ so that $\alpha(\Lambda) = \ell(\hat{\mathbf{Q}}_\Lambda(\mathbf{S}), \mathbf{S})$ is small, where

$$\ell(\mathbf{Q}, \mathbf{S}) = \log\det\left(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{\Phi}\right) - \log\det\mathbf{Q} - \mathrm{tr}\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}\mathbf{\Phi}(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{\Phi})^{-1}\right)$$

is the unpenalized likelihood function in (5). The cross-validation approach is to partition $\{1, \ldots, m\}$ into disjoint sets $\{A_1, \ldots, A_k\}$ and select $\hat{\Lambda} = \arg\min_{\Lambda\in\{\Lambda_1,\ldots,\Lambda_t\}} \hat{\alpha}(\Lambda)$, where $\hat{\alpha}(\Lambda) = k^{-1}\sum_{i=1}^{k} \ell(\hat{\mathbf{Q}}_\Lambda(\mathbf{S}_{A_i^c}), \mathbf{S}_{A_i})$.

Jointly searching over λ and ρ can quickly become unwieldy even when considering a small combination of sparsity and fusion penalties, as noted by Danaher et al. (2014), who instead suggest a dense search for λ with $\rho = 0$ fixed and then a search for ρ with that sparsity value fixed. The individual cross-validation for either λ or ρ follows the same idea: whichever penalty parameter has the lowest negative log-likelihood average across folds is selected.

3.5. Overview of Simulation Study

We conducted an extensive simulation study that examined the proposed estimator of the precision matrices $\mathbf{Q}_1, \ldots, \mathbf{Q}_L$. Here, we provide a high-level overview of the simulation study; full details are available in the supplementary material. Both the basis functions $\phi_1, \ldots, \phi_L$ and the error variances $\tau_1^2, \ldots, \tau_p^2$ are the same as in Section 4.1. We considered two different parameterizations for the precision matrices, and in both cases the graphical structure is identical over all levels. In the simpler case where $\mathbf{Q}_1 = \mathbf{Q}_2 = \cdots = \mathbf{Q}_L$, the exact graphical structure was recovered by the MLE with parameters λ, ρ selected from cross-validation. In a second more complicated case, where the values of the precision matrices grow exponentially over level, selecting parameters with cross-validation was not as helpful. Ignoring the difficulties with cross-validation, the estimator is still able to discern meaningful connections between variables. In both cases, the method correctly identifies patterns in the marginal precision values $(\mathbf{Q}_\ell)_{ii}$ for $i = 1, \ldots, p$ and $\ell = 1, \ldots, L$. Specifically, the estimator accurately recovered the behavior of marginal precision values for different variables that were either constant or exponentially growing over level ℓ.

4. Data Analysis

This section is broken into two main, but related, application and validation efforts. Both are done with the lens of the climate data problem: we begin with some exploratory analyses of the climate dataset, deriving reasonable basis functions and providing discussion to guide intuition for the ensuing model application. In the supplementary material, we detail a simulation study that tests our ability, under a setup similar to the climate example, to recover meaningful and relevant graphs for coefficients at different levels of basis functions under realistic assumptions on possible graph structures. Results from this simulation study are promising and lead us to expect reasonable results with the real climate data as well. The final section in the body of this article provides the full analysis of our model on the Community Atmosphere Model (CAM) data along with scientific interpretations of recovered graphical structures and some implied covariance and cross-covariance patterns. Code that outlines the data analysis procedure is available at *github.com/mlkrock/MultivariateBasisGraphicalLasso*.

4.1. Data Description and Exploratory Analyses

We apply our method to a large climatological dataset from an ensemble study conducted at the National Center for Atmospheric Research (NCAR). Climate variability is typically assessed by examining a collection of numerical climate model simulations, that are computationally and economically expensive to produce. Relationships between variables at different spatial scales are crucial for scientific investigations; hence, a scalable statistical model which can simulate multivariate processes could be a powerful tool for climatologists. Our method allows for efficient emulation and straightforward interpretation of complex geophysical model variable relationships, potentially filling this niche.

A climate model ensemble is typically a collection of climate simulations from the same numerical model using various initial conditions; ours is an extended version of the ensemble described by Baker et al. (2015) with $m = 343$ members. Data are recorded at $n = 48{,}602$ spatial locations over the globe. There are a total of 164 variables available, the majority of which are three-dimensional, meaning they have a third dimension corresponding to 30 vertical atmospheric levels. For our study we only consider the two-dimensional surface variables and a subset thereof. First, the variables are on different scales, so they are standardized with a pixelwise empirical mean and pixelwise empirical standard deviation. Histograms and Q-Q plots were consulted to remove strongly nonnormal variables. Note that our data are yearly averaged quantities, so a Gaussian assumption is generally reasonable. Potential variables were also removed if they were very strongly correlated, suggesting redundant information (e.g., when vectorized across space and realizations, the absolute correlation between two processes was above 0.9). We settled upon the $p = 40$ variables listed in Table 1, which are grouped into five categories: aerosol variables, cloud variables, flux variables, precipitation variables, and transport/state variables. Throughout the rest of the document, aerosol variables are colored red, cloud variables are colored blue, flux variables are colored green, precipitation variables are colored purple, and transport/state variables are colored gray.

With the $p = 40$ variables in hand, our approach relies on first specifying a set of spatial basis functions. We construct such functions as empirical orthogonal functions (EOFs) (Wikle

Table 1. Variable descriptions.

Variable	Description	Units	Category
AODVIS	Aerosol optical depth	550 nm	Aerosol
BURDEN1	Aerosol burden mode 1	kg/m^2	Aerosol
BURDEN2	Aerosol burden mode 2	kg/m^2	Aerosol
BURDEN3	Aerosol burden mode 3	kg/m^2	Aerosol
BURDENBC	Black carbon aerosol burden	kg/m^2	Aerosol
BURDENPOM	POM aerosol burden	kg/m^2	Aerosol
BURDENSEASALT	Seasalt aerosol burden	kg/m^2	Aerosol
BURDENSO4	Sulfate aerosol burden	kg/m^2	Aerosol
BURDENSOA	SOA aerosol burden	kg/m^2	Aerosol
CDNUMC	Vertically-integrated droplet concentration	$1/m^2$	Cloud
CLDHGH	Vertically-integrated high cloud	fraction	Cloud
CLDMED	Vertically-integrated mid-level cloud	fraction	Cloud
CLDTOT	Vertically-integrated total cloud	fraction	Cloud
FLDS	Downwelling longwave flux at surface	W/m^2	Flux
FLNS	Net longwave flux at surface	W/m^2	Flux
FLNSC	Clearsky net longwave flux at surface	W/m^2	Flux
FLNT	Net longwave flux at top of model	W/m^2	Flux
FLNTC	Clearsky net longwave flux at top of model	W/m^2	Flux
FSDS	Downwelling solar flux at surface	W/m^2	Flux
FSDSC	Clearsky downwelling solar flux at surface	W/m^2	Flux
FSNS	Net solar flux at surface	W/m^2	Flux
FSNSC	Clearsky net solar flux at surface	W/m^2	Flux
FSNTC	Clearsky net solar flux at top of model	W/m^2	Flux
FSNTOA	Net solar flux at top of atmosphere	W/m^2	Flux
LHFLX	Surface latent heat flux	W/m^2	Flux
LWCF	Longwave cloud forcing	W/m^2	Cloud
PBLH	PBL height	W/m^2	Transport/State
PS	Surface pressure	Pa	Transport/State
QREFHT	Reference height humidity	kg/kg	Precipitation
SHFLX	Surface sensible heat flux	W/m^2	Flux
SWCF	Shortwave cloud forcing	W/m^2	Cloud
TAUX	Zonal surface stress	N/m^2	Transport/State
TAUY	Meridional surface stress	N/m^2	Transport/State
TGCLDCWP	Total grid-box cloud water path (liquid and ice)	kg/m^2	Cloud
TGCLDIWP	Total grid-box cloud ice water path	kg/m^2	Cloud
TGCLDLWP	Total grid-box cloud liquid water path	kg/m^2	Cloud
TMQ	Total vertically integrated precipitable water	kg/m^2	Precipitation
TREFHT	Surface air temperature at reference height	K	Transport/State
U10	10m wind speed	m/s	Transport/State
PRECT	PRECL Large-scale (stable) precipitation rate (liq + ice) plus PRECC Convective precipitation rate (liq + ice)	m/s	Precipitation

Note that analysis happens on standardized, unitless data.

2010), which are widely used in the atmospheric and climate sciences. Typically, EOFs are used in a temporal context with a single variable. Let's consider a single spatiotemporal variable and suppose we have a matrix B of data with rows indexing n spatial locations and columns indexing t time points. If $B = UDV^{\mathrm{T}}$ is the (economy) SVD of the data matrix, the columns of the orthogonal matrix U, referred to as EOFs, represent the normalized eigenvectors of the process empirical covariance matrix BB^{T}.

We compute the (economy) SVD of the $n \times pm = 48{,}602 \times 13{,}720$ data matrix where a row corresponds to a spatial location and contains the p standardized variables ordered sequentially by realization. Such an approach can be thought of as generating *pooled* EOFs that describe common structure seen among all variables. Using the same notation UDV^{T} for the SVD, we follow common practice and take the first L columns of U to form our $n \times L$ basis matrix Φ. Exploratory analysis suggests truncating after $L = 2000$ EOFs is a reasonable trade-off between using a relatively small number of basis functions and explaining sufficient variance (97.2%). The first two pooled EOFs are displayed in Figure 1. We emphasize that pooling variables together to create EOFs is nontraditional and explains why the first two EOFs account for such little variability. Other orthogonal bases besides EOFs are viable choices but may require model selection criterion like AIC or BIC to select the number of basis functions.

With this formulation, it is important to clarify the role of the additive error term. Here, an interpretation as a traditional spatial statistical nugget effect is not well-motivated; instead, we think of $\boldsymbol{\varepsilon}$ as a fine-scale process that is at smaller scale than the EOFs and absorbs the remaining variability unexplained by the pooling of variables. As noted by Wikle (2010), if enough eigenvectors are used to explain sufficient variation, then it is reasonable to assume that the EOF residuals are uncorrelated in space. This motivates the white noise assumption on $\boldsymbol{\varepsilon}$, which in turn suggests independently estimating $\tau_1^2, \ldots, \tau_p^2$ using the procedure described in Section 3.3. Estimated values for $\tau_1^2, \ldots, \tau_p^2$ and additional interpretations are shown in the supplementary material.

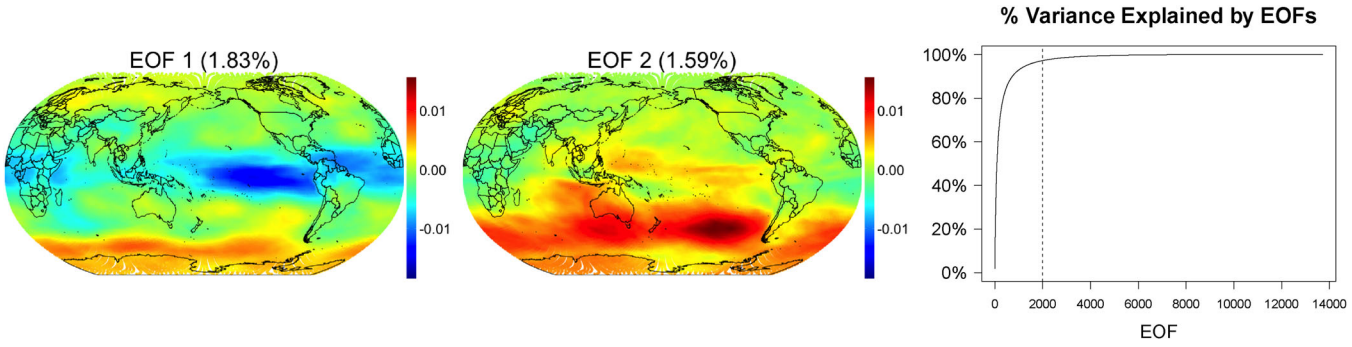

Figure 1. First two pooled EOFs of the standardized CAM data. The two EOFs account for 1.83% and 1.59% of the total variability of all 40 variables, respectively. Rightmost plot shows the cumulative percentage of variability explained by the EOFs, with $L = 2000$ EOFs capturing 97.2% of the total variance.

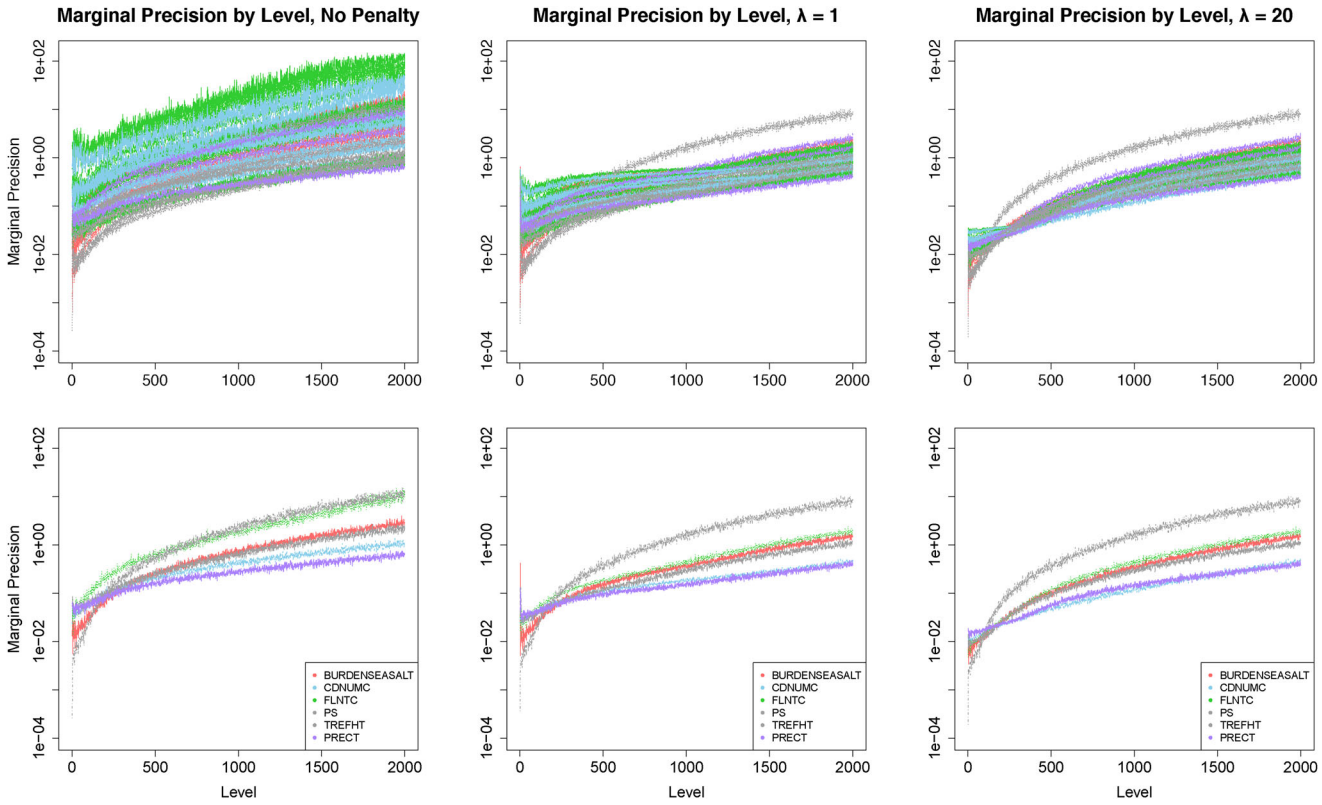

Figure 2. Marginal precision estimates (i.e., diagonals of $\hat{\Omega}_1, \ldots, \hat{\Omega}_{2000}$) shown by level for various penalty choices. Same log scale on y-axis. Bottom row shows a subset of six variables from the top row. The grey pressure variable is noticeably smoother than the rest after regularization is added.

4.2. CAM Data Analysis

We proceed to the data analysis using our estimates for $\tau_1^2, \ldots, \tau_p^2$ and the basis setup from Section 4.1. The remaining question is what penalty parameters to use. With a cross-validation attempt (see supplementary material) and similar difficulties with penalties encountered by Danaher et al. (2014) in mind, we proceed by fixing several penalty pairs and examining the resulting modeling implications. Ideally, we would select a model with a sensible, interpretable graphical neighbor structure over levels. Differences between models with different graph structures may be minor as different graphs can give approximately the same correlation structure. In the supplementary material and remainder of this document, we display several results for $\lambda = 20$ since this looked like an inflection point in a plot of λ versus the total graph sparsity percentage (see Figure 6(a) in supplementary material). We also will occasionally compare $\lambda = 20$ results with those from $\lambda = 1$ to give an idea of how the implied graph structure changes with different penalties.

For the first set of model diagnostics, we examine the estimated precision matrices. Plots of estimated marginal precisions by level are shown in Figure 2. Clearly, the two parameterizations considered in the simulation study (see Section 3.5) are not well-suited to our data, as the estimated marginal precisions are neither constant nor exponentially growing over level. Just as in the simulation study, adding a sparsity penalty causes the pressure variable (colored gray) to have the highest marginal precisions at high levels of resolution, as we would expect from the smoothest variable. Overall, adding a sparsity penalty brings all marginal precisions down by an order of magnitude, with particularly strong shrinkage at lower levels of resolution. Given the focus on regularization, these results may not seem surprising, but

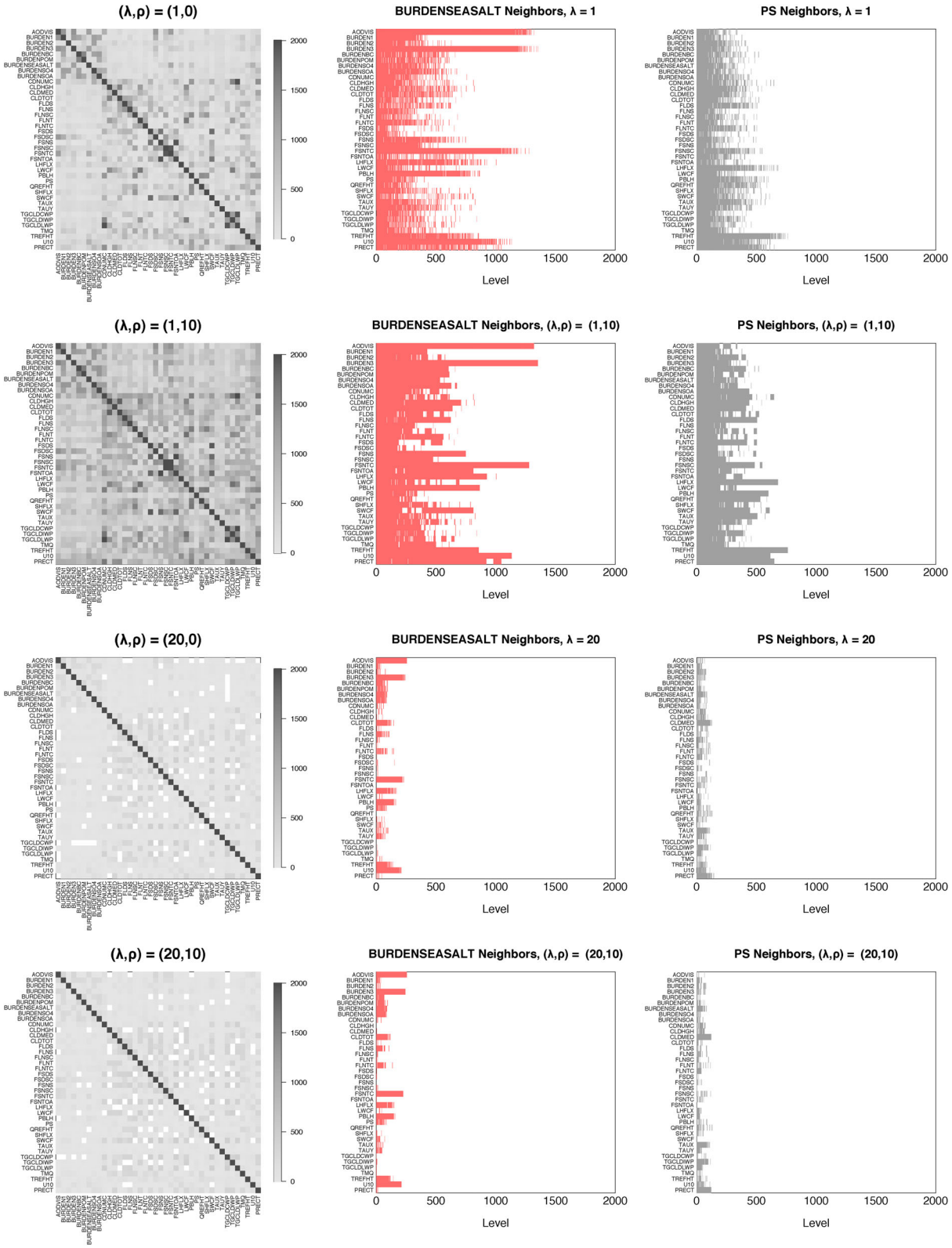

Figure 3. Illustration of how graphical model neighborhoods behave for various penalty choices. Left column counts the nonzeros of $\hat{\mathbf{Q}}_1, \ldots, \hat{\mathbf{Q}}_{2000}$ by level. Center and right columns show how the neighbors of BURDENSEASALT and PS change over level. Variable names along the axes follow the same order as Table 1.

the diagonals of the precision matrix are not penalized in any formulation we have considered. This shrinkage of marginal precisions can be attributed to the larger number of neighbors in the low-penalized graph structures, which means that the (conditional) precision will be higher than in estimates from higher penalties.

Our penalized maximum likelihood procedure produces interpretable graphical models that imply conditional independencies among the variables at varying levels of resolution. Here we study some properties of the estimated graphical models. Figure 3 gives an idea of how the conditional independence structure changes with respect to level for different penalties.

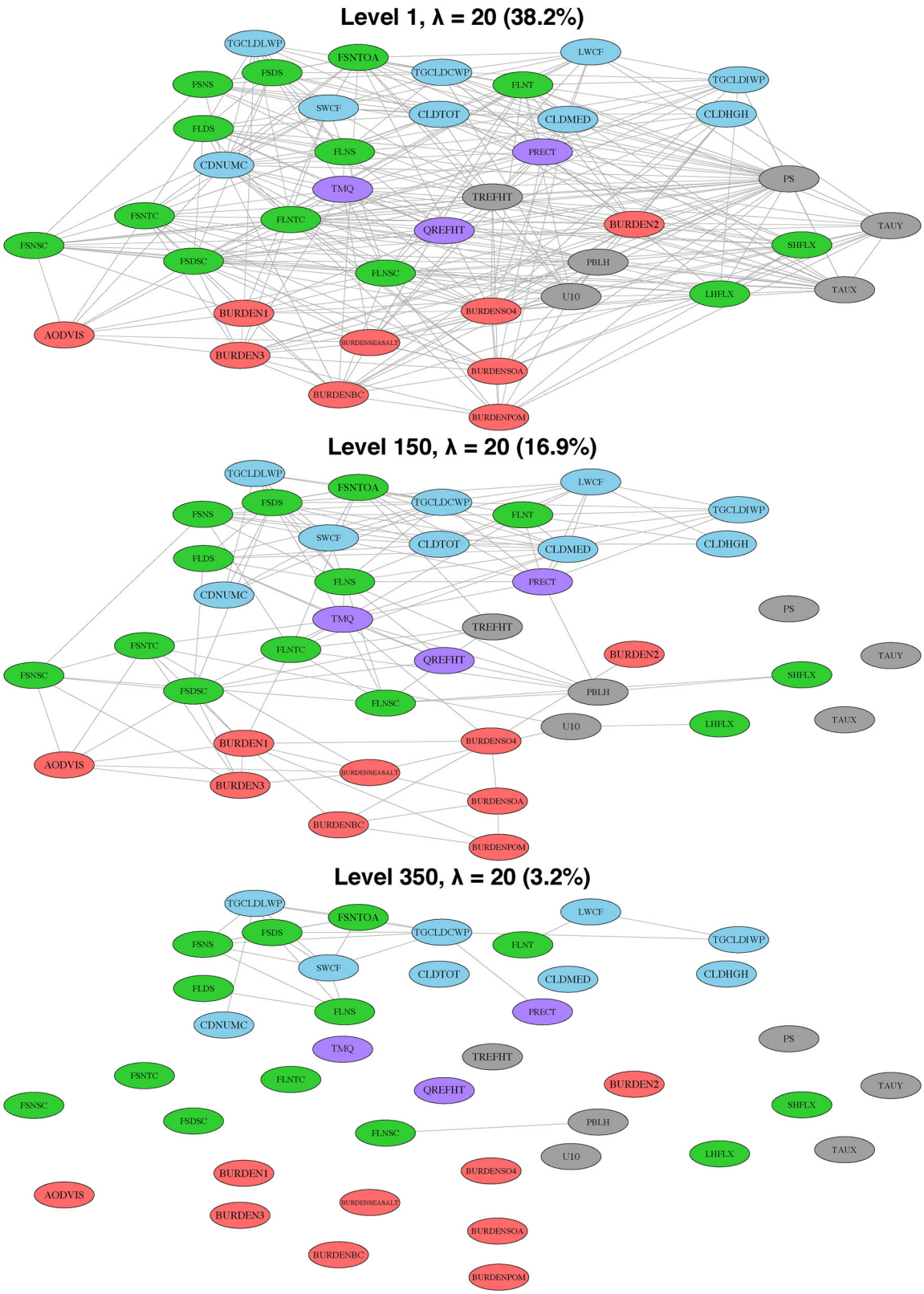

Figure 4. Example estimated graphical models for $\lambda = 20$. Graphs at lower levels (i.e., graph structure of $\mathbf{Q}_\ell$ at smaller values of ℓ) are noisy, with the Level 1 graph containing 38.2% of all possible connections. Higher-level graphs show reasonable variable clusters until eventually no graph edges exist.

The left column counts the nonzeros of the estimated precision matrices $\hat{\mathbf{Q}}_1, \ldots, \hat{\mathbf{Q}}_{2000}$ over level. With $\lambda = 20$ we see several conditional independencies between variables that persist over all levels. The middle and right columns show how the conditional dependence neighborhood structure of two variables (BURDENSEASALT and PS) changes over level, and the impact of the fusion penalty is most apparent in these columns. Broadly speaking, the fusion penalty smooths out the neighbor pattern across levels. Note that the fusion penalty can fuse adjacent nonzeros rather than adjacent zeros and can cause a neighbor

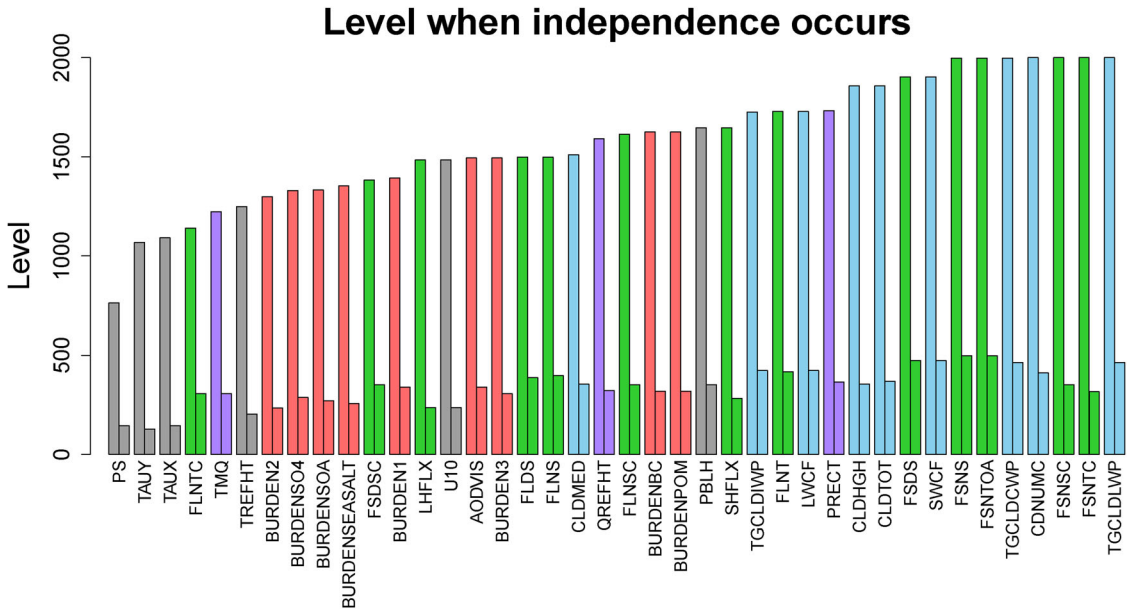

Figure 5. The first level at which a variable becomes (and remains) independent. Here, independence refers to a graphical model where a variable (node) has no edges connecting it to any of the other nodes. Variables are ordered in increasing fashion according to the implied independence level for $\lambda = 1$. Results for $\lambda = 20$ accompany the taller bars and show a similar story for variable groups but with earlier levels of independence.

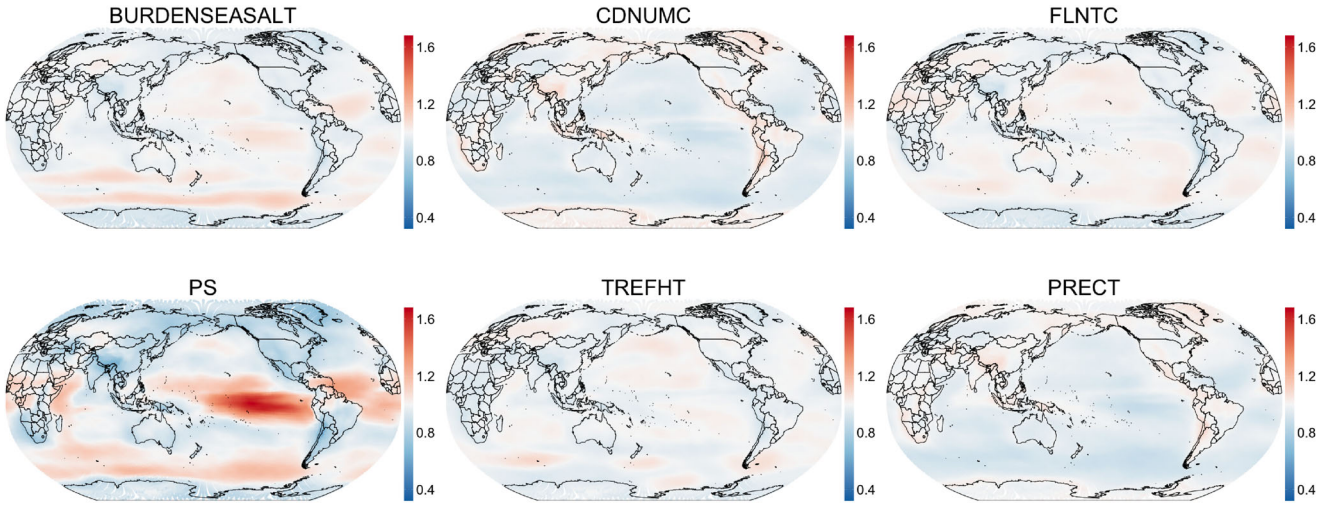

Figure 6. Estimated local standard deviations for six variables. Standardization of the variables is reflected in the color scale, where unit values are colored white. Values above one suggest the empirical standardization does not explain enough variability.

to fuse across all levels if ρ is large enough. Figure 4 displays estimated graphical models for different levels of resolution. At early levels corresponding to large-scale spatial patterns, the graphs are quite dense. Higher-frequency EOFs display interesting and reasonable patterns. For example, near the top of the $\ell = 150$ graph we see groupings among many cloud and precipitation variables, and these connections are more evident at higher levels and even with different penalty parameters (not shown). For higher-level EOFs, past around $\ell = 500$ with $\lambda = 20$, the graphs suggest variable independence. This idea of complete independence is explored in Figure 5 where we display the first level at which each variable is independent and remains independent of all other variables. Note again a natural grouping of variable types, with transport and pressure variables achieving independence much earlier than finer-scale precipitation and cloud variables. Further, as could also be observed in the visualization of the conditional dependence structure for the two selected variables in Figure 3, the level at which independence occurs varies quite dramatically with the sparsity penalty, leading to roughly a four-fold increase in the number of connected levels going from $\lambda = 20$ to $\lambda = 1$.

Now, we examine some spatial properties of our estimates. Figure 6 shows the estimated local standard deviations for a subset of six variables. Recall that the variables were standardized to have an empirical unit standard deviation, so these plots should be interpreted as potential bias corrections where the standardization fails to accurately describe the variability. Most striking is the El Niño effect apparent in the plot for the pressure variable PS. The pattern's presence is unsurprising since El Niño/La Niña are strongly tied to changes in pressure over the Pacific Ocean, and their relative infrequency likely requires more modeling care than just an empirical standardization. Finer-scale variables

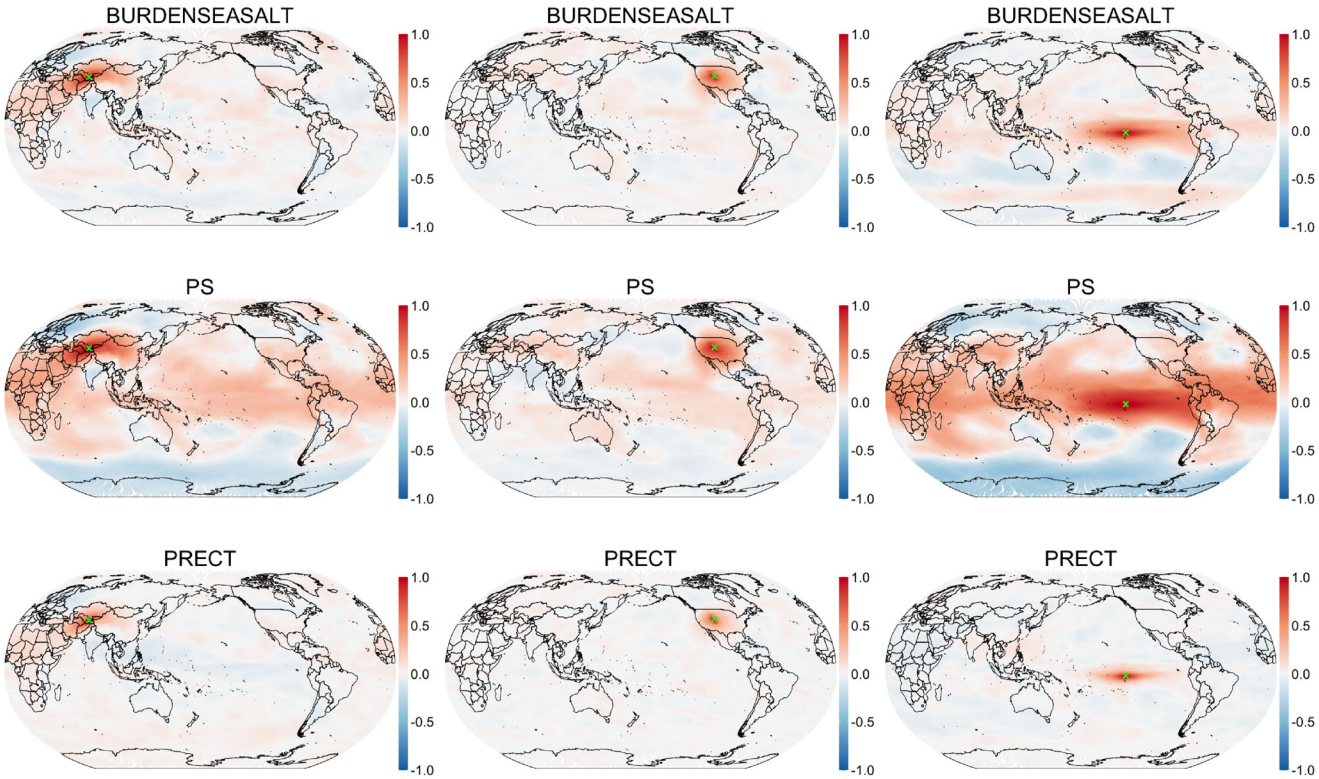

Figure 7. Estimated spatial correlation functions for BURDENSEASALT (top row), PS (middle row), and PRECT (bottom row). Correlation function is centered over Tajikistan in the left column, U.S. in the middle column, and the Pacific Ocean in the right column.

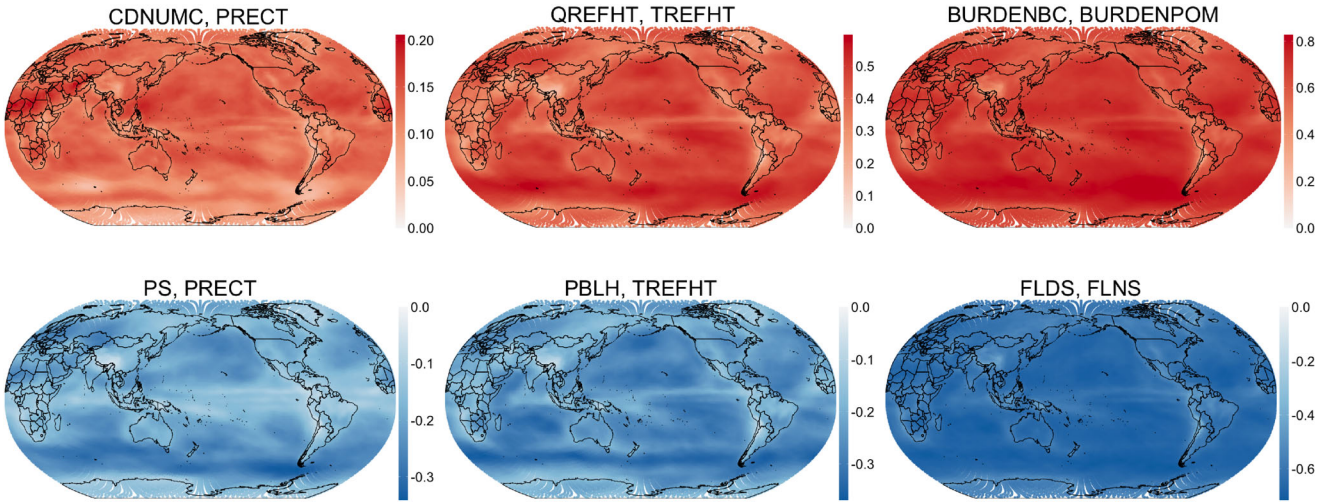

Figure 8. Estimated local cross-correlations for several pairs of variables. Color scale changes between rows to permit easier comparisons between positively correlated variables (top row) and between negatively correlated variables (bottom row).

CDNUMC and PRECT are able to capture distinct behavior in mountainous regions (e.g., Rocky Mountains and Himalayas).

In our last collection of figures, we examine correlation over space as well as across variables. Figure 7 shows variable correlations as a function of space. To be precise, each image shows the correlation between the location marked in green and all other locations. Nonstationarity is evident from the difference in behaviors between land and ocean, and long-range negative correlation is seen to be a possible byproduct of this modeling scheme. In Figure 8, we display estimated local cross-correlations between a few of our selected variables. Expected negative and positive correlations between pairs of variables are correctly captured (e.g., between pressure and precipitation and between cloud droplet concentration and precipitation). Finally, in Figure 9, we show variable cross-correlations as a function of space. To be precise, each image shows the cross-correlation between the first variable at the location marked in green and the second variable at all other locations. The flexibility of the model is again apparent in its nonstationary behavior and various positive and negative cross-correlations. Such behavior is difficult to accommodate using extant models but readily available in our approach without any additional effort (Kleiber and Genton 2013). Both the correlation and cross-correlation functions centered where El Niño/La Niña occur (see right column in

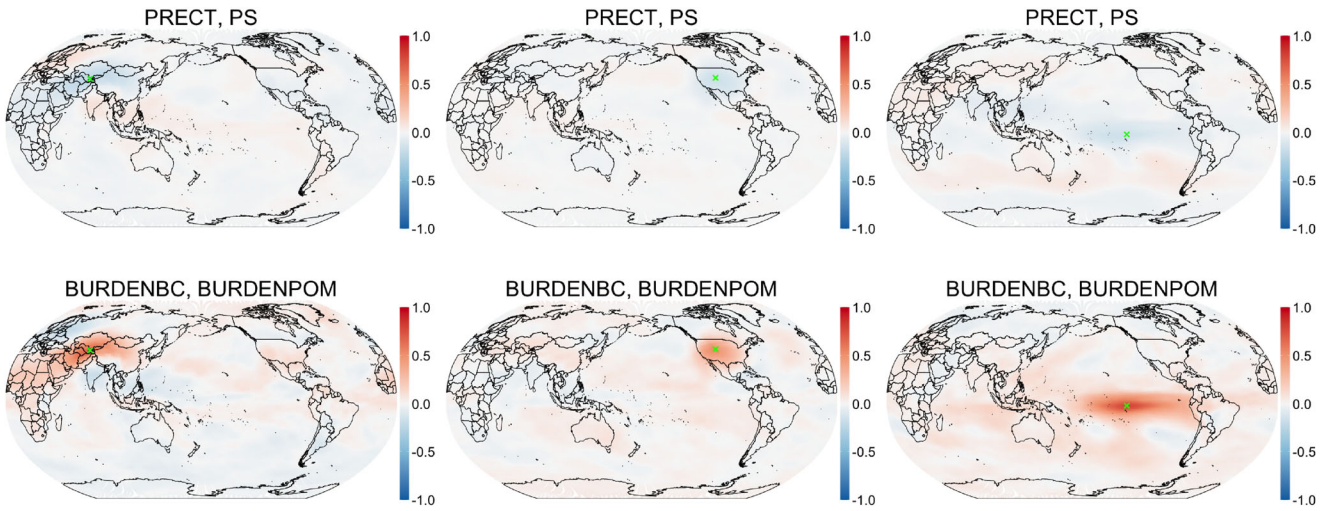

Figure 9. Estimated spatial cross-correlation functions, again centered over Tajikistan in the left column, U.S. in the middle column, and the Pacific Ocean in the right column. Top row shows PRECT and PS, which exhibit negative local cross-correlations, while bottom row shows BURDENBC and BURDENPOM, which exhibit positive local cross-correlations.

Figures 7 and 9) exhibit long-range dependence through the equator across the Pacific Ocean.

We conclude the data analysis with a brief commentary about the timing results, assuming the orthogonal basis has been constructed and $\tau_1^2, \ldots, \tau_p^2$ are estimated. Note that the latter step is fairly quick using the technique in Section 3.3 with an orthogonal basis. The choice of penalty dictates the runtime of the DC algorithm. For the maximum likelihood estimate with no penalty, the tolerance $\epsilon = 0.05$ is reached in 15 DC iterations in 5 seconds. With sparsity penalty $\lambda = 20$, the algorithm converges in two DC iterations in 45 seconds. With sparsity penalty $\lambda = 1$, the algorithm again requires two DC iterations but takes 13 min. With fusion penalty $\rho = 10$, both estimates converged in one DC iteration using the $\rho = 0$ solution as the initial guess, but $(\lambda, \rho) = (20, 10)$ took 7 minutes while $(\lambda, \rho) = (1, 10)$ took 1 min. All experiments were performed in MATLAB on a MacBook Pro with a 6 core 2.6 GHz Intel Core i7 processor and 32 GB of RAM.

5. Discussion and Conclusions

Models for multivariate spatial processes struggle to accommodate nonstationary data with a large number of variables and observation locations. We have presented a multivariate Gaussian process model that will be applicable in a variety of future endeavors. There are many benefits under this framework, including nonstationary covariance and cross-covariance functions, exact likelihood calculations, and cheap computations and storage when $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are independent and basis functions are orthogonal.

Future experiments could relax the assumption of independent weights and estimate the entire sparse $pL \times pL$ precision matrix $\mathbf{Q}$ as in Section 2.1 with the DC algorithm (6). Even if the basis functions are not orthogonal, $\mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{\Phi} = \Phi^{\mathrm{T}}\Phi \otimes \mathrm{diag}(\tau_1^{-2}, \ldots, \tau_p^{-2})$ is sparse due to the Kronecker product with a diagonal matrix. In this formulation, it is crucial to rewrite (7) and calculate this linearization term by solving linear systems with the sparse matrix $\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{\Phi}$ or its sparse Cholesky decomposition. Note that the dense $pL \times pL$ linearization matrix must then be stored in memory, and advanced graphical lasso algorithms (Fattahi et al. 2019) must be explored for the subsequent graph estimation. Modeling dependence across basis functions would allow for more flexibility in the cross-covariances since, with the independence assumption, $\mathrm{cov}(Z_i(\mathbf{s}), Z_j(\mathbf{s}')) = \sum_{\ell=1}^{L} \phi_\ell(\mathbf{s})((\mathbf{Q}_\ell)^{-1})_{ij}\phi_\ell(\mathbf{s}') = \mathrm{cov}(Z_i(\mathbf{s}'), Z_j(\mathbf{s}))$ is symmetric.

We demonstrated that our model can easily fit a large climate ensemble and produce reasonable and interpretable results. It has already been successfully applied in the context of statistical downscaling (Ekanayaka et al. 2022). Our method also easily scales to computer experiments with high-dimensional inputs (e.g., $\mathbf{s} \in \mathbb{R}^d$ with $d \gg 0$). Extending the basis graphical lasso to handle space-time data remains an important task that could be particularly impactful in the highly multivariate setting. Constructing multivariate space-time covariance functions is an active area of research (Alegria et al. 2019; Salvaña and Genton 2020; Porcu et al. 2020; Chen et al. 2021). Indeed, the issues of scalability and flexibility mentioned in the opening sentence of this conclusion are further exacerbated when modeling temporal dependence. A multivariate space-time basis graphical lasso could prove to be a valuable tool that addresses these challenges.

Supplementary Materials

Supplementary material contains: (1) a simulation study (2) additional details regarding CAM data analysis, including estimation of the additive error variance, selection of penalty parameters, and simulations from our multivariate spatial model. Further supplementary materials are available at *https://gdex.ucar.edu/dataset/371_abaker.html*.

Funding

This research was funded by grants NSF DMS-1821074 and NSF DMS-1923062.

ORCID

Mitchell L. Krock http://orcid.org/0000-0002-1628-1842
William Kleiber http://orcid.org/0000-0003-0411-9108
Dorit Hammerling https://orcid.org/0000-0003-3583-3611
Stephen Becker http://orcid.org/0000-0002-1932-8159

References

Alegria, A., Porcu, E., Furrer, R., and Mateu, J. (2019), "Covariance Functions for Multivariate Gaussian Fields Evolving Temporally over Planet Earth," *Stochastic Environmental Research and Risk Assessment*, 33, 1593–1608. [14]

Apanasovich, T., and Genton, M. G. (2010), "Cross-Covariance Functions for Multivariate Random Fields based on Latent Dimensions," *Biometrika*, 97, 15–30. DOI:10.1093/biomet/asp078. [2]

Apanasovich, T. V., Genton, M. G., and 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, 180–193. DOI:10.1080/01621459.2011.643197. [2]

Baker, A. H., Hammerling, D. M., Levy, M. N., Xu, H., Dennis, J. M., Eaton, B. E., Edwards, J., Hannay, C., Mickelson, S. A., Neale, R. B., Nychka, D., Shollenberger, J., Tribbia, J., Vertenstein, M., and Williamson, D. (2015), "A New Ensemble-based Consistency Test for the Community Earth System Model (pyCECT v1.0)," *Geoscientific Model Development*, 8, 2829–2840. *https://gmd.copernicus.org/articles/8/2829/2015/*. [7]

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2015), "Multivariate Spatio-Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics," *The Annals of Applied Statistics*, 9, 1761–1791. DOI:10.1214/15-AOAS862. [1]

——— (2018), "Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data," (with Discussion), *Bayesian Analysis*, 13, 253–310. DOI:10.1214/17-BA1069. Includes comments and discussions by ten discussants and a rejoinder by the authors. [1]

Bruinsma, W. P., Perim, E., Tebbutt, W., Hosking, J. S., Solin, A., and Turner, R. E. (2020), "Scalable Exact Inference in Multi-Output Gaussian Processes," in *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org. [2,3]

Calder, C. A. (2007), "Dynamic Factor Process Convolution Models for Multivariate Space-Time Data with Application to Air Quality Assessment," *Environmental and Ecological Statistics*, 14, 229–247. DOI:10.1007/s10651-007-0019-y. [2]

Chen, W., Genton, M. G., and Sun, Y. (2021), "Space-Time Covariance Structures and Models," *Annual Review of Statistics and Its Application*, 8, 191–215. DOI:10.1146/annurev-statistics-042720-115603. [14]

Cressie, N., and Zammit-Mangion, A. (2016), "Multivariate Spatial Covariance Models: A Conditional Approach," *Biometrika*, 103, 915–935. DOI:10.1093/biomet/asw045. [2]

Danaher, P., Wang, P., and Witten, D. M. (2014), "The Joint Graphical Lasso for Inverse Covariance Estimation Across Multiple Classes," *Journal of the Royal Statistical Society*, Series B, 76, 373–397. DOI:10.1111/rssb.12033. [5,7,9]

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. DOI:10.1080/01621459.2015.1044091. [2,3]

Dey, D., Datta, A., and Banerjee, S. (2021), "Graphical Gaussian Process Models for Highly Multivariate Spatial Data," *Biometrika*, 109, 993–1014. DOI:10.1093/biomet/asab061. [2,3]

——— (2022), "On the Relationship between Graphical Gaussian Processes and Functional Gaussian Graphical Models," *https://arxiv.org/abs/2209.06294*. [3]

Ekanayaka, A., Kang, E., Kalmus, P., and Braverman, A. (2022), "Statistical Downscaling of Model Projections with Multivariate Basis Graphical Lasso," *https://arxiv.org/abs/2201.13111*. [14]

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016), "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization," *Geoscientific Model Development*, 9, 1937–1958. *https://gmd.copernicus.org/articles/9/1937/2016/*. [1]

Fattahi, S., Zhang, R. Y., and Sojoudi, S. (2019), "Linear-Time Algorithm for Learning Large-Scale Sparse Graphical Models," *IEEE Access*, 7, 12658–12672. [4,14]

Fontanella, L., Fontanella, S., Ignaccolo, R., Ippoliti, L., and Valentini, P. (2020), "G-Lasso Network Analysis for Functional Data," in *Functional and High-Dimensional Statistics and Related Fields*, eds. G. Aneiros, I. Horová, M. Hušková, and P. Vieu), pp. 91–98, Cham: Springer. [3]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 9, 432–441. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3019769/*. [3]

Furrer, R., and Genton, M. G. (2011), "Aggregation-Cokriging for Highly Multivariate Spatial Data," *Biometrika*, 98, 615–631. *https://doi.org/10.1093/biomet/asr029*. [2]

Gaspari, G., and Cohn, S. E. (1999), "Construction of Correlation Functions in Two and Three Dimensions," *Quarterly Journal of the Royal Meteorological Society*, 125, 723–757. *https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712555417*. [2]

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004), "Nonstationary Multivariate Process Modeling through Spatially Varying Coregionalization," Test, 13, 263–312. DOI:10.1007/BF02595775. [2,3]

Genton, M. G., and Kleiber, W. (2015), "Cross-Covariance Functions for Multivariate Geostatistics," *Statistical Science*, 30, 147–163. DOI:10.1214/14-STS487. [1]

Gneiting, T., Kleiber, W., and Schlather, M. (2010), "Matérn Cross-Covariance Functions for Multivariate Random Fields," *Journal of the American Statistical Association*, 105, 1167–1177. DOI:10.1198/jasa.2010.tm09420. [2]

Goulard, M., and Voltz, M. (1992), "Linear Coregionalization Model: Tools for Estimation and Choice of Cross-Variogram Matrix," *Mathematical Geology*, 24, 269–286. [2]

Guinness, J. (2022), "Nonparametric Spectral Methods for Multivariate Spatial and Spatial–Temporal Data," *Journal of Multivariate Analysis*, 187, 104823. *https://www.sciencedirect.com/science/article/pii/S0047259X21001019*. [2,5]

Harville, D. A. (1997), *Matrix Algebra from a Statistician's Perspective*, New York: Springer-Verlag. DOI:10.1007/b98818. [6]

Hsieh, C.-J., Dhillon, I. S., Ravikumar, P. K., Becker, S., and Olsen, P. A. (2014a), "QUIC & DIRTY: A Quadratic Approximation Approach for Dirty Statistical Models," in *Advances in Neural Information Processing Systems 27*, pp. 2006–2014. [5]

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. (2014b), "QUIC: Quadratic Approximation for Sparse Inverse Covariance Estimation," *Journal of Machine Learning Research*, 15, 2911–2947. [4,5]

Ippoliti, L., Valentini, P., and Gamerman, D. (2012), "Space-Time Modelling of Coupled Spatiotemporal Environmental Variables," *Journal of the Royal Statistical Society*, Series C, 61, 175–200. DOI:10.1111/j.1467-9876.2011.01011.x. [2]

Jun, M. (2011), "Non-Stationary Cross-Covariance Models for Multivariate Processes on a Globe," *Scandinavian Journal of Statistics*, 38, 726–747. DOI:10.1111/j.1467-9469.2011.00751.x. [2]

Kleiber, W. (2017), "Coherence for Multivariate Random Fields," *Statistica Sinica*, 27, 1675–1697. [2,5]

Kleiber, W., and Genton, M. G. (2013), "Spatially Varying Cross-Correlation Coefficients in the Presence of Nugget Effects," *Biometrika*, 100, 213–220. [13]

Kleiber, W., and Nychka, D. (2012), "Nonstationary Modeling for Multivariate Spatial Processes," *Journal of Multivariate Analysis*, 112, 76–91. DOI:10.1016/j.jmva.2012.05.011. [2]

Kleiber, W., Nychka, D., and Bandyopadhyay, S. (2019), "A Model for Large Multivariate Spatial Data Sets," *Statistica Sinica*, 29, 1085–1104. [2,3]

Kleiber, W., and Porcu, E. (2014), "Nonstationary Matrix Covariances: Compact Support, Long Range Dependence and Quasi-Arithmetic Constructions," *Stochastic Environmental Research and Risk Assessment*, 29, 193–204. [2]

Krock, M., Kleiber, W., and Becker, S. (2021), "Nonstationary Modeling with Sparsity for Spatial Data via the Basis Graphical Lasso,"

Journal of Computational and Graphical Statistics, 30, 375–389. DOI:10.1080/10618600.2020.1811103. [3,4,6,7]

Le, N. D., and Zidek, J. V. (2006), *Statistical Analysis of Environmental Space-Time Processes*, Springer Series in Statistics, New York: Springer. [2]

Liu, H., Ding, J., Xie, X., Jiang, X., Zhao, Y., and Wang, X. (2021), "Scalable Multi-Task Gaussian Processes with Neural Embedding of Coregionalization," *https://arxiv.org/abs/2109.09261*. [2,3]

Majumdar, A., and Gelfand, A. E. (2007), "Multivariate Spatial Modeling for Geostatistical Data Using Convolved Covariance Functions," *Mathematical Geosciences*, 39, 225–245. DOI:10.1007/s11004-006-9072-6. [2]

Majumdar, A., Paul, D., and Bautista, D. (2010), "A Generalized Convolution Model for Multivariate Nonstationary Spatial Processes," *Statistica Sinica*, 20, 675–695. [2]

Meng, R., Lee, H., and Bouchard, K. (2021a), "Stochastic Collapsed Variational Inference for Structured Gaussian Process Regression Network," arXiv: Machine Learning. [2,3]

Meng, R., Soper, B., Lee, H. K., Liu, V. X., Greene, J. D., and Ray, P. (2021b), "Nonstationary Multivariate Gaussian Processes for Electronic Health Records," *Journal of Biomedical Informatics*, 117, 103698. *https://www.sciencedirect.com/science/article/pii/S1532046421000277*. [2]

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015), "A Multiresolution Gaussian Process Model for the Analysis of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 24, 579–599. DOI:10.1080/10618600.2014.914946. [2]

Pollice, A., and Jona Lasinio, G. (2010), "A Multivariate Approach to the Analysis of Air Quality in a High Environmental Risk Area," *Environmetrics*, 21, 741–754. DOI:10.1002/env.1059. [2]

Porcu, E., Furrer, R., and Nychka, D. (2020), "30 Years of Space-Time Covariance Functions," *Wiley Interdisciplinary Reviews: Computational Statistics*, 13, e1512. [14]

Qadir, G. A., Euán, C., and Sun, Y. (2021), "Flexible Modeling of Variable Asymmetries in Cross-Covariance Functions for Multivariate Random Fields," *Journal of Agricultural, Biological and Environmental Statistics*, 26, 1–22. DOI:10.1007/s13253-020-00414-2. [2]

Qadir, G. A., and Sun, Y. (2020), "Semiparametric Estimation of Cross-Covariance Functions for Multivariate Random Fields," *Biometrics*, 77, 547–560. *https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13323*. [2]

Qiao, X., Guo, S., and James, G. M. (2019), "Functional Graphical Models," *Journal of the American Statistical Association*, 114, 211–222. DOI:10.1080/01621459.2017.1390466. [3]

Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications* (Monographs on Statistics and Applied Probability), New York: Chapman & Hall/CRC. [3]

Salvaña, M., Abdulah, S., Huang, H., Ltaief, H., Sun, Y., Genton, M. M., and Keyes, D. (2021), "High Performance Multivariate Geospatial Statistics on Manycore Systems," *IEEE Transactions on Parallel and Distributed Systems*, 32, 2719–2733. [2]

Salvaña, M. L. O., and Genton, M. G. (2020), "Nonstationary Cross-Covariance Functions for Multivariate Spatio-Temporal Random Fields," *Spatial Statistics*, 37, 100411. *https://www.sciencedirect.com/science/article/pii/S2211675320300051*. Frontiers in Spatial and Spatio-temporal Research. [1,14]

Schmidt, A. M., and Gelfand, A. E. (2003), "A Bayesian Coregionalization Approach for Multivariate Pollutant Data," *Journal of Geophysical Research: Atmospheres*, 108. *https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD002905*. [2]

Shaddick, G., and Wakefield, J. (2002), "Modelling Daily Multivariate Pollutant Data at Multiple Sites," *Journal of the Royal Statistical Society*, Series C, 51, 351–372. DOI:10.1111/1467-9876.00273. [2]

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer-Verlag. [4]

——— (2014), "Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data," *Spatial Statistics*, 8, 1–19. DOI:10.1016/j.spasta.2013.06.003. [3]

Taylor-Rodriguez, D., Finley, A., Datta, A., Babcock, C., Andersen, H., Cook, B., Morton, D., and Banerjee, S. (2019), "Spatial Factor Models for High-Dimensional and Large Spatial Data: An Application in Forest Variable Mapping," *Statistica Sinica*, 29, 1155–1180. [2,3]

Teh, Y., Seeger, M. W., and Jordan, M. I. (2005), "Semiparametric Latent Factor Models," in *AISTATS*. [3]

Titsias, M. (2009), "Variational Learning of Inducing Variables in Sparse Gaussian Processes," in *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, Vol. 5 of Proceedings of Machine Learning Research, pp. 567–574. PMLR. *http://proceedings.mlr.press/v5/titsias09a.html*. [3]

Titsias, M., and Lawrence, N. D. (2010), "Bayesian Gaussian Process Latent Variable Model," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, eds. Y. W. Teh and M. Titterington, Vol. 9 of Proceedings of Machine Learning Research, pp. 844–851. Chia Laguna Resort, Sardinia, Italy: PMLR. *http://proceedings.mlr.press/v9/titsias10a.html*. [2]

Ver Hoef, J. M., and Barry, R. P. (1998), "Constructing and Fitting Models for Cokriging and Multivariable Spatial Prediction," *Journal of Statistical Planning and Inference*, 69, 275–294. DOI:10.1016/S0378-3758(97)00162-6. [2]

Vu, Q., Zammit-Mangion, A., and Cressie, N. (2020), "Modeling Nonstationary and Asymmetric Multivariate Spatial Covariances via Deformations," arXiv: Statistics. [2]

Wackernagel, H. (2003), *Multivariate Geostatistics: An Introduction with Applications*, Berlin Heidelberg: Springer. *https://books.google.com/books?id=Rhr7bgLWxx4C*. [2]

Wikle, C. K. (2010), "Low Rank Representations for Spatial Processes," in *Handbook of Spatial Statistics*, pp. 107–118, Boca Raton: Chapman & Hall / CRC. [8]

Yang, S., Lu, Z., Shen, X., Wonka, P., and Ye, J. (2015), "Fused Multiple Graphical Lasso," *SIAM Journal on Optimization*, 25, 916–943. DOI:10.1137/130936397. [5]

Zapata, J., Oh, S. Y., and Petersen, A. (2021), "Partial Separability and Functional Graphical Models for Multivariate Gaussian Processes," *Biometrika*, 109, 665–681. DOI:10.1093/biomet/asab046. [3]

Zhang, L., and Banerjee, S. (2021), "Spatial Factor Modeling: A Bayesian Matrix-Normal Approach for Misaligned Data," *Biometrics*, 109, 665–681. *https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13452*. [2,3]

Zhang, L., Banerjee, S., and Finley, A. O. (2021), "High-Dimensional Multivariate Geostatistics: A Bayesian Matrix-Normal Approach," *Environmetrics*, 32, e2675. *https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2675*. [2]