Title

A flexible class of priors for orthonormal matrices with basis function-specific structure

Permalink

Authors

North, Joshua S
Risser, Mark D
Breidt, F Jay

Publication Date

2024-12-01

DOI

10.1016/j.spasta.2024.100866

Copyright Information

A flexible class of priors for orthonormal matrices with basis function-specific structure

Joshua S. North[1,*], Mark D. Risser[1], and F. Jay Breidt[2]

[1] *Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory*

[2] *Department of Statistics and Data Science, NORC at the University of Chicago*

[*] *Corresponding author: jsnorth@lbl.gov*

Abstract

Statistical modeling of high-dimensional matrix-valued data motivates the use of a low-rank representation that simultaneously summarizes key characteristics of the data and enables dimension reduction. Low-rank representations commonly factor the original data into the product of orthonormal basis functions and weights, where each basis function represents an independent feature of the data. However, the basis functions in these factorizations are typically computed using algorithmic methods that cannot quantify uncertainty or account for basis function correlation structure *a priori*. While there exist Bayesian methods that allow for a common correlation structure across basis functions, empirical examples motivate the need for basis function-specific dependence structure. We propose a prior distribution for orthonormal matrices that can explicitly model basis function-specific structure. The prior is used within a general probabilistic model for singular value decomposition to conduct posterior inference on the basis functions while accounting for measurement error and fixed effects. We discuss how the prior specification can be used for various scenarios and demonstrate favorable model properties through synthetic data examples. Finally, we apply our method to two-meter air temperature data from the Pacific Northwest, enhancing our understanding of the Earth system's internal variability.

Key Words: Bayesian Singular Value Decomposition, Probabilistic Low-Rank Representation, Probabilistic Basis Functions, Stiefel Manifold, Spatio-Temporal Random Effect

1 Introduction

1.1 Orthonormal matrices in statistical modeling

Within the field of statistics, orthonormal matrices are the cornerstone of many modeling approaches, including exploratory data analysis, factor analysis (Harman and Harman, 1976; Mulaik, 2009), principal component analysis (PCA; Hotelling, 1933; Jolliffe, 2002), singular value decomposition (SVD; Stewart, 1993), and proper orthogonal decomposition (POD; Berkooz, 1993). Each of these techniques uses orthonormal matrices to decompose matrix-valued data with the goal of summarizing its key characteristics as well as dimension reduction (Kambhatla and Leen, 1997) and data compression (Chen et al., 2022). Across many areas of science, technology, and medicine, orthonormal matrix factorizations of data are highly useful because the measurements of interest in these fields often arise from lower-dimensional processes with physically interpretable structures. Examples include factor analysis in physiological studies (Fabrigar et al., 1999), PCA in geography (Roden et al., 2015) and ecology (Jackson, 1993; Peres-Neto et al., 2003), and SVD and PCA for medical imaging (Smith et al., 2014).

For mean-zero data $\mathbf{Y} \in \mathbb{R}^{n\times m}$, SVD decomposes $\mathbf{Y} = \mathbf{UDV}'$, where $\mathbf{U} \in \mathbb{R}^{n\times l}$ is an orthonormal matrix, $\mathbf{D} \in \mathbb{R}^{l\times l}$ is a diagonal matrix, $\mathbf{V} \in \mathbb{R}^{m\times l}$ is an orthonormal matrix, and $l = \min\{n, m\}$. Alternatively, PCA decomposes $\mathbf{YY}' = \mathbf{ABA}'$, where now $\mathbf{A} \in \mathbb{R}^{n\times l}$ is an orthonormal matrix whose columns are the eigenvectors of $\mathbf{YY}'$, $\mathbf{B} \in \mathbb{R}^{l\times l}$ is a diagonal matrix whose elements are the eigenvalues of $\mathbf{YY}'$, and $l = \min\{n, m\}$. Note that the equivalence between SVD and PCA comes from $\mathbf{YY}' = (\mathbf{UDV}')(\mathbf{VD}'\mathbf{U}') = \mathbf{UDD}'\mathbf{U}' = \mathbf{ABA}'$, where the diagonal elements of $\mathbf{D}$ are the square root of the eigenvalues of $\mathbf{YY}'$, the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{YY}'$, and the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{Y}'\mathbf{Y}$.

In the climate sciences where data are spatially- and temporally-oriented, the columns of orthonormal matrices define empirical orthogonal functions (EOFs; Lorenz, 1956; North et al., 1982; Hannachi et al., 2007), which are analogous to PCA. EOFs are used to summarize modes of climate variability (see, e.g., Thompson and Wallace, 2000; Mantua and Hare, 2002), identify the drivers of extreme weather events (Grotjahn et al., 2016), and quantify human-induced changes to the global climate system (O'Brien and Deser, 2023). Additionally, spatial modeling of climate data often uses EOFs to incorporate spatial and temporal information via spatially-indexed basis functions and spatial random effects (Stroud et al., 2001; Nychka et al., 2002; Cressie and Johannesson, 2006, 2008).

1.2 Inference and challenges

The basis functions contained in the orthonormal matrices $\mathbf{U}$ and $\mathbf{V}$ and the elements of $\mathbf{D}$ are traditionally computed via iterative methods (Golub and Kahan, 1965; Demmel and Kahan, 1990), which we refer to as classical SVD (C-SVD or C-PCA) henceforth. However, these classical procedures have several important limitations. First, when n is large with respect to m, the basis functions contained in the orthonormal matrices estimated from C-

SVD can be noisy and therefore lose their physical interpretation (Wang and Huang, 2017). C-SVD and C-PCA are not able to distinguish between measurement and signal variation, which means that estimates of the basis functions are heavily influenced by the presence of measurement noise (Bailey, 2012; Epps and Krivitzky, 2019). Furthermore, since their algorithms are deterministic, C-PCA and C-SVD do not provide measures of uncertainty in either the basis functions or their weights. Finally, the estimated basis functions only exhibit dependence or structure implicitly via data correlations since C-PCA and C-SVD cannot take advantage of explicit structure that may be present in the data generating mechanisms.

A variety of approaches have been developed to address limitations associated with C-SVD and C-PCA. Regarding the issue of noise, large n with small m, and structure in the basis functions, a regularized PCA approach can be adopted (Shen and Huang, 2008; Zou et al., 2006; Jolliffe et al., 2002). Wang and Huang (2017) extend the regularization approach by incorporating smoothness and local features into their penalization using smoothing splines and an ℓ_1 penalty, producing spatially explicit orthogonal basis functions. To further account for uncertainty quantification in the basis function, one possibility is to take a Bayesian approach and specify a prior distribution for the orthonormal matrix. The set of orthonormal matrices $\mathcal{V}_{k,n} = \{\mathbf{X} \in \mathbb{R}^{n \times k} : \mathbf{X}'\mathbf{X} = \mathbf{I}_k\}$, where $\mathbf{I}_k$ is the $k \times k$ identity matrix, is called the Stiefel manifold (Chikuse, 2003). Considerable effort has been put into understanding theoretical properties associated with distributions on the Stiefel manifold and optimal methods for computation and sampling (Mardia and Jupp, 1999; Chikuse, 2003; Hoff, 2007, 2009; Byrne and Girolami, 2013; Wang and Gelfand, 2013, 2014; Hernandez-Stumpfhauser et al., 2017; Pourzanjani et al., 2021; Jauch et al., 2021). Hoff (2007) developed a uniform prior for orthonormal basis functions (the invariant or uniform measure on the Stiefel manifold) that enables the specification of a Bayesian SVD model, and showed how to sample from the full conditional distributions of the model. However, the approach in Hoff (2007) requires sampling from the von Mises-Fisher (or Bingam-von Mises-Fisher) distributions, which can be difficult, and does not allow for the basis functions to be structured. Additionally, support for these distributions in probabilistic programming languages such as Stan is limited (Carpenter et al., 2017), providing yet another barrier for implementation. Hoff (2009) and Byrne and Girolami (2013) propose tractable methods for sampling from von Mises-Fisher distributions, but these require the underlying statistical model to abide by specific conditions and forms which limits the application areas. Recent work by Pourzanjani et al. (2021) and Jauch et al. (2021) addresses both sampling and flexibility of distributions on the Stiefel manifold by simulating unconstrained random vectors (i.e., not orthogonal and not unit-length) and then transforming these draws to be orthonormal via an appropriate Jacobian to obtain samples on the Stiefel manifold. Importantly,

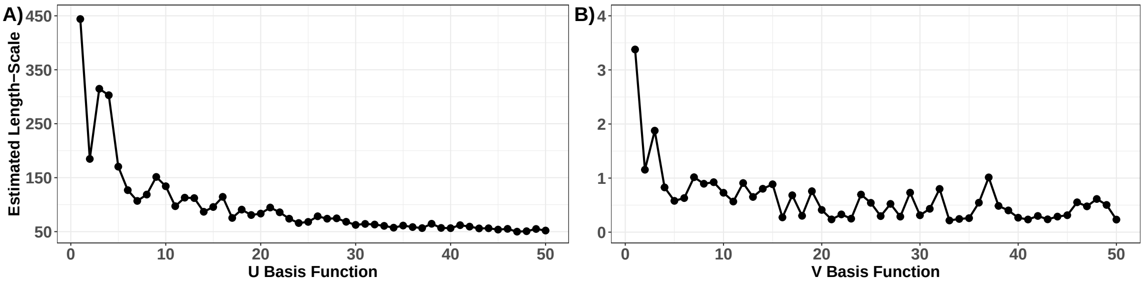


Figure 1: Estimated length-scale from a fitted Gaussian variogram for each spatial and temporal basis function computed from the singular value decomposition of the two-meter air surface temperature data described in Section 5.

these methods are computationally efficient, can be incorporated into probabilistic programming languages, and allow for the basis functions to be modeled dependently. However, the dependence structure is limited in that it is shared across the basis functions and is unable to accommodate the basis function-specific structures that are present in real-world data sets.

Particularly in the climate sciences, the physical structures summarized by orthonormal matrices have different scales (e.g., spatial or temporal), wherein the leading modes or basis functions reflect larger-scale variability while the later modes reflect finer-scale variability. To illustrate this, we calculated the SVD of monthly maximum two-meter air temperature from a $0.25° \times 0.25°$ longitude-latitude grid over the United States Pacific Northwest from 1979 through 2022 (see Section 5 for details on the data) using standard statistical software. We then estimate the length-scale of a Gaussian variogram for each spatial and temporal basis function, the columns of the left- and right- singular matrices, respectively. Figure 1 shows empirical estimates of the length-scale for each basis function for $\mathbf{U}$ and $\mathbf{V}$ in panels a) and b), respectively. From this figure it is clear the length-scale of the leading modes for both the left- and right- singular matrices is at least one order of magnitude larger than that of the later modes, following a quasi-exponentially decreasing trend. This suggests estimating a common spatial or temporal structure for all of the basis functions will miss important features of the data, resulting in oversmoothing and underfitting for the leading modes and undersmoothing and overfitting for the later modes.

1.3 Contributions

Here, we develop a prior distribution for orthonormal matrices that enables basis function specific structure and construct a probabilistic model for SVD. The resulting full conditional distributions for the basis functions are available in closed form, yielding an analytically straightforward posterior for sampling orthonormal matrices. Furthermore, we discuss how the prior can be used for a variety of modeling purposes. Our prior is in general not uniformly distributed on $\mathcal{V}_{k,n}$ (although the uniform distribution is a special case) and we are able to impart information into the prior through our specification of a correlation matrix. We show how the correlation matrix can be specified to either impart smoothing onto the basis functions (producing results similar to Wang and Huang, 2017) or recover the prior developed by Hoff (2007), and also demonstrate how the mean of the full conditional distributions for the basis functions of our probabilistic SVD model coincides with the classical approach (C-SVD) under certain conditions. Our resulting prior, along with the proposed Bayesian hierarchical model, is quite general and allows each basis function to have a unique dependence structure that is learned from the data, which has not been previously possible.

The remainder of the manuscript is organized as follows. Section 2 develops the prior distribution for orthonormal matrices. Section 3 proposes a general probabilistic model for matrix factorizations with a specific focus on SVD and then expands other possible modeling choices. Three simulation studies are conducted in section 4, where we show the importance of basis function-specific structure, the model rank and signal-to-noise ratios, and the impact a linear trend has on basis function recovery. In Section 5, we apply our probabilistic model for SVD to decompose monthly maximum two-meter air temperature into its major modes of variability and provide uncertainty bounds for these modes, allowing better understanding of the spatial relationships in the data and illustrating the importance of basis function-specific structure. Section 6 concludes the paper.

2 A prior distribution for orthonormal matrices with basis function-specific structure

We construct a prior distribution for matrices on the Stiefel manifold $\mathcal{V}_{k,n}$ that is conjugate with a normal likelihood model. The prior is constructed from the projected normal distribution that has been augmented with a latent length (see Wang and Gelfand 2013, 2014 and Hernandez-Stumpfhauser et al. 2017 and the references therein for details on the projected normal).

2.1 Generating mechanism

One method of drawing an orthonormal matrix from the uniform distribution on $\mathcal{V}_{k,n}$ is outlined in the appendix of Hoff (2007). As part of the construction, the underlying normal distribution from which the orthonormal matrix is generated specifies the identity matrix as the covariance, and the resulting distribution is uniform on $\mathcal{V}_{k,n}$. Here, we extend this generating mechanism to allow for structure in its covariance, specific to each column, such that the prior implied by Hoff (2007) is a special case. By construction, the resulting distribution is not necessarily uniform on $\mathcal{V}_{k,n}$.

For fixed k, let $\mathbf{z}_i$ independent $\mathrm{N}_n(\mathbf{0}, \mathbf{\Omega}_i)$ and $\mathbf{\Omega}_i \sim \pi_\Omega$, for $i = 1, 2, \ldots, k$, where π_Ω is a valid distribution for symmetric positive definite matrices. Define $\mathbf{P}_0 = \mathbf{I}_n$, $\mathbf{x}_1 = \mathbf{P}_0\mathbf{z}_1$, and

$$\mathbf{X}_i = [\mathbf{x}_1, \ldots, \mathbf{x}_i], \quad \mathbf{P}_i = \mathbf{I}_n - \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i', \quad \mathbf{x}_{i+1} = \mathbf{P}_i\mathbf{z}_{i+1}$$

for $i = 1, 2, \ldots, k-1$. Then $\mathbf{x}_i|\mathbf{X}_{i-1} \sim \mathrm{N}_n(\mathbf{0}, \mathbf{P}_{i-1}\mathbf{\Omega}_i\mathbf{P}_{i-1}')$ and $\mathbf{x}_i'\mathbf{x}_j = 0$ for $i \neq j$. Further, define

$$\mathbf{w}_i = \frac{\mathbf{x}_i}{(\mathbf{x}_i'\mathbf{x}_i)^{1/2}}, \quad \mathbf{W}_i = [\mathbf{w}_1, \ldots, \mathbf{w}_i] \tag{1}$$

for $i = 1, 2, \ldots, k$. By construction, $\mathbf{W}_k \in \mathcal{V}_{k,n}$ is an orthonormal matrix. The conditional distributions of each column given the preceding columns are $\mathbf{w}_i|\mathbf{W}_{i-1} \sim \mathrm{PN}_n(\mathbf{0}, \mathbf{P}_{i-1}\mathbf{\Omega}_i\mathbf{P}_{i-1}')$, where $\mathrm{PN}_n(\cdot, \cdot)$ denotes the n-dimensional projected normal distribution (Wang and Gelfand, 2013, 2014; Hernandez-Stumpfhauser et al., 2017).

Let $\overset{d}{=}$ denote equality in distribution. We now provide two key properties associated with the distribution of $\mathbf{W} \equiv \mathbf{W}_k$ based on the constructed matrix $\mathbf{X} \equiv \mathbf{X}_k$, with proofs deferred to appendix A.

Proposition 1. *The columns of* $\mathbf{W} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}$ *are exchangeable. That is, for any permutation* π *of the set* $\{1, \ldots, k\}$, $p([\mathbf{w}_1, \ldots, \mathbf{w}_k]) \overset{d}{=} p([\mathbf{w}_{\pi(1)}, \ldots, \mathbf{w}_{\pi(k)}])$.

Proposition 2. $\mathbf{w}_i|\mathbf{W}_{i-1} \overset{d}{=} \mathbf{N}_{i-1}\widetilde{\mathbf{w}}_i|\mathbf{W}_{i-1}$ *where the columns of* $\mathbf{N}_{i-1}$ *form an orthonormal basis for the null space of* $\mathbf{W}_{i-1}$ *and* $\widetilde{\mathbf{w}}_i$*, the projected weight function, satisfies* $\widetilde{\mathbf{w}}_i|\mathbf{W}_{i-1} \sim \mathrm{PN}_{n-i+1}(\mathbf{0}, \mathbf{N}_{i-1}'\mathbf{\Omega}_i\mathbf{N}_{i-1})$.

Proposition 1 implies the columns of $\mathbf{W}$ are exchangeable, and therefore the conditional distribution $\mathbf{w}_i|\mathbf{W}_Q$ is invariant to the choice of subset of columns $Q \subset \{1, \ldots, k\}$. When proposition 1 is taken with proposition 2, the conditional distribution of $\mathbf{w}_i|\mathbf{W}_Q$ given any subset of columns Q is equal in distribution to $\mathbf{N}_Q\widetilde{\mathbf{w}}_i$, where $\mathbf{N}_Q$ is an orthonormal basis for the null space of $\mathbf{W}_Q$ and $\widetilde{\mathbf{w}}_i|\mathbf{W}_Q \sim \mathrm{PN}_{n-|Q|+1}(\mathbf{0}, \mathbf{N}_Q'\mathbf{\Omega}_i\mathbf{N}_Q)$. Therefore, we now focus on a prior distribution for $\widetilde{\mathbf{w}}_i$, the *projected weight function*, where $\widetilde{\mathbf{w}}_i|\mathbf{W}_{-i} \sim \mathrm{PN}_{n-k+1}(\mathbf{0}, \mathbf{N}_i'\mathbf{\Omega}_i\mathbf{N}_i)$ (i.e., $Q = \{1, \ldots, i-1, i+1, \ldots, k\}$) and the columns of $\mathbf{N}_i$ span the null space of $\mathbf{W}_{-i}$.

2.2 Projected normal prior distribution

From the construction in Section 2.1, we have $\widetilde{\mathbf{w}}_i|\mathbf{W}_{-i} \sim \mathrm{PN}_{n-k+1}(\mathbf{0}, \mathbf{N}_i'\mathbf{\Omega}_i\mathbf{N}_i)$. However, sampling from a high-dimensional projected normal distribution is difficult because of the form of the density function. To make sampling from the projected normal tractable, we augment the distribution $\widetilde{\mathbf{w}}_i|\mathbf{W}_{-i}$ using a latent length variable r_i. The joint distribution of $(r_i, \widetilde{\mathbf{w}}_i)|\mathbf{W}_{-i}$ can be derived by transforming the random variable $\mathbf{x}_i$ to spherical coordinates (see supplement S.4), where the density function is

$$p(r_i, \widetilde{\mathbf{w}}_i|\mathbf{W}_{-i}) = (2\pi)^{-n^*/2}|\mathbf{N}_i'\mathbf{\Omega}_i\mathbf{N}_i|^{-1/2}\exp\left\{-\frac{1}{2}(r_i\widetilde{\mathbf{w}}_i)'(\mathbf{N}_i'\mathbf{\Omega}_i\mathbf{N}_i)^{-1}(r_i\widetilde{\mathbf{w}}_i)\right\}r_i^{n^*-1}\mathbb{I}(\widetilde{\mathbf{w}}_i \in \mathcal{V}_{1,n^*}), \tag{2}$$

which we denote as $p(r_i, \widetilde{\mathbf{w}}_i) \sim \mathrm{N}_{n^*}(\mathbf{0}, \mathbf{N}_i'\mathbf{\Omega}_i\mathbf{N}_i)r_i^{n^*-1}$ with $n^* = n - k + 1$. The indicator function $\mathbb{I}(\widetilde{\mathbf{w}}_i \in \mathcal{V}_{1,n^*})$ is an integrating constant that is independent of the angle of $\widetilde{\mathbf{w}}_i$ and dependent only on its length. Note for $k = 1$, the Stiefel manifold $\mathcal{V}_{1,n}$ is the $n-1$-dimensional unit sphere and $\mathcal{V}_{1,n} \equiv \mathbb{S}^{n-1}$. The length variable r_i can be sampled using either a slice sampler (Hernandez-Stumpfhauser et al., 2017) or a Metropolis-Hastings algorithm. However, we have found the slice sampler has numerical issues when n is large, and use a Metropolis-Hastings within Gibbs algorithm (see supplement S.1) for all examples presented herein.

The PN prior is convenient because if the data distribution is normal, the resulting full conditional distribution is proportional to a normal, which is easy to sample from (see Section 3.1 and supplement S.1 for more detail).

2.3 Incorporating explicit structure into the prior

From our formulation of the prior, we have the ability to specify or estimate the correlation structure for the projected basis functions. The non-informative choice is $\mathbf{\Omega}_i \propto \mathbf{I}_n$, implying there is no dependence between the elements of the basis functions. As discussed in the supplement (S.2), when $\mathbf{\Omega}_i \equiv \mathbf{I}$ the generating mechanism is equivalent to that proposed by Hoff (2007), resulting in $\widetilde{\mathbf{w}}_i$ being distributed uniformly on the $(n-k+1)$-dimensional sphere and the prior being equivalent to Hoff (2007).

A more general choice is to model $\mathbf{\Omega}_i = \sigma_i^2\mathbf{C}_i$, where $\mathbf{C}_i$ is a positive-definite correlation matrix that specifies structure among the elements in the ith basis function and σ_i^2 is a common variance parameter for those elements. (While σ_i^2 does not impact the distribution of $\widetilde{\mathbf{w}}_i$ or $\mathbf{w}_i$ directly because they are of unit length, it does affect the joint distribution (2) of $(r_i, \widetilde{\mathbf{w}}_i)$.) In most cases, $\mathbf{C}_i \equiv \mathbf{C}(\boldsymbol{\theta}_i)$ will depend on hyperparameters $\boldsymbol{\theta}_i$ that can either be specified or learned within the hierarchical model. Across many areas of science, including spatial statistics, machine learning, and emulation of complex physical models, the elements of $\mathbf{C}_i$ are modeled via kernel functions $C_\theta(\cdot,\cdot)$ that are positive definite on the domain specified by the input space $\mathcal{S}$. For example, when $\mathcal{S} \subset \mathbb{R}^d$, a popular choice is the

Matérn kernel

$$C_{\nu,\rho}(\mathbf{s},\mathbf{s}') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(2\nu\frac{||\mathbf{s}-\mathbf{s}'||}{\rho}\right)^{\nu} J_{\nu}\left(2\nu\frac{||\mathbf{s}-\mathbf{s}'||}{\rho}\right), \tag{3}$$

defined for $\mathbf{s},\mathbf{s}' \in \mathcal{S}$, where Γ is the gamma function, J_ν is the Bessel function of the second kind, and $\boldsymbol{\theta} = \{\nu,\rho\}$ are hyperparameters that describe the differentiability and length-scale of the implied stochastic process, respectively. Special cases of the Matérn kernel are for $\nu = 0.5$, in which (3) simplifies to the exponential kernel $C_{0.5,\rho}(\mathbf{s},\mathbf{s}') = \exp\{-||\mathbf{s}-\mathbf{s}'||/\rho\}$, and the limit as $\nu \to \infty$, in which (3) reduces to the squared exponential or Gaussian kernel $C_{\infty,\rho}(\mathbf{s},\mathbf{s}') = \exp\{-||\mathbf{s}-\mathbf{s}'||^2/\rho\}$. Kernel functions like the Matérn are useful for modeling generic dependence because they are highly flexible, depend on only a few hyperparameters (each of which is interpretable), yield data-driven smoothing that can characterize nonlinear structures in the underlying data, and require minimal *a priori* or subjective specification. Furthermore, such kernel functions do not require offline tuning of bandwidth or regularization parameters (as is needed in, e.g., smoothing splines; see Wang and Huang, 2017) since these aspects of the kernel can be inferred from the data within the Bayesian hierarchical model.

3 General probabilistic model

Define $\mathbf{Z} \in \mathbb{R}^{n\times m}$ to be the observed data which is modeled as

$$\mathbf{Z} = \mathbf{M} + \mathbf{Y} + \mathbf{A}\boldsymbol{\Xi}\mathbf{B}, \tag{4}$$

where $\mathbf{M} \in \mathbb{R}^{n\times m}$, $\mathbf{Y} \in \mathbb{R}^{n\times m}$, $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}' \in \mathbb{R}^{n\times n}$, $\boldsymbol{\Phi} = \mathbf{B}\mathbf{B}' \in \mathbb{R}^{m\times m}$, and $\boldsymbol{\Xi} = [\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_m]$ with $\boldsymbol{\xi}_i$ independent $N_n(0,\mathbf{I}_m)$ for $i = 1,\ldots,m$. Then $\mathbf{Z}|\mathbf{M},\mathbf{Y},\boldsymbol{\Sigma},\boldsymbol{\Phi} \sim \mathrm{MN}_{n\times m}(\mathbf{M} + \mathbf{Y},\boldsymbol{\Sigma},\boldsymbol{\Phi})$ where MN is the matrix normal distribution, $\mathbf{M}+\mathbf{Y}$ is the mean of $\mathbf{Z}$, $\boldsymbol{\Sigma}$ is the covariance matrix for the rows of $\mathbf{Z}$, $\boldsymbol{\Phi}$ is the covariance matrix for the columns of $\mathbf{Z}$, and the density function is

$$p(\mathbf{Z}|\mathbf{M},\mathbf{Y},\boldsymbol{\Sigma},\boldsymbol{\Phi}) = \frac{1}{(2\pi)^{nm/2}|\boldsymbol{\Phi}|^{n/2}|\boldsymbol{\Sigma}|^{m/2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Phi}^{-1}(\mathbf{Z}-\mathbf{M}-\mathbf{Y})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z}-\mathbf{M}-\mathbf{Y})\right]\right\}. \tag{5}$$

Equation (4) is a mixed-effects model, where $\mathbf{M}$ is a fixed-effect mean structure that is dependent on observed covariates, which we discuss in Section 3.2.2, and $\mathbf{Y}$ is a "smooth" random effect that we will represent using basis functions and weights. Generally, we assume $\mathbf{Y}$ is a mean zero random effect and explains any discrepancy in $\mathbf{Z}$ not explained by $\mathbf{M}$.

For example, if $\mathbf{Z}$ is oriented such that the rows index spatial locations and the columns index temporal observations (or replications), $\mathbf{Y}$ would be considered spatial random effects. We now specify a non-parametric model for the random effects $\mathbf{Y}$ using singular value decomposition and the prior distribution proposed in Section 2.2.

3.1 A probabilistic model for singular value decomposition

For now, we assume the mean of $\mathbf{Z}$ is zero (i.e., $\mathbf{M} = \mathbf{0}$) and focus on a model for $\mathbf{Y}$. In models such as (4), the process $\mathbf{Y}$ can be represented as a reduced-rank process. One example of a reduced-rank model is the singular value decomposition (SVD) $\mathbf{Y} = \mathbf{UDV}'$, where $\mathbf{U} \in \mathbb{R}^{n\times l}$ is an orthonormal matrix, $\mathbf{D} \in \mathbb{R}^{l\times l}$ is a diagonally structured matrix, $\mathbf{V} \in \mathbb{R}^{m\times l}$ is an orthonormal matrix, and $l = \min\{n, m\}$. To reduce the dimension of the process, we set $k < l$ (typically $k \ll l$) where k is some pre-specified value. This results in $\mathbf{Y} \approx \mathbf{UDV}'$, where now $\mathbf{U} \in \mathbb{R}^{n\times k}$, $\mathbf{D} \in \mathbb{R}^{k\times k}$, and $\mathbf{V} \in \mathbb{R}^{m\times k}$ are of reduced dimension. In traditional SVD (similarly in PCA), the amount of variance explained by each component can be used to inform the value of k. For now, we will assume k is fixed and refer the reader to Section 3.1.3 for further discussion.

In (4), $\mathbf{\Phi} = \mathbf{BB}'$ represents the covariance between replicate observations (columns). We make the simplifying assumption $\mathbf{\Phi} = \mathbf{I}_m$ (i.e., independence between replicates) and model all variation in the data through $\mathbf{\Sigma}$, which represents the covariance within observations (rows). The resulting probability model is $\mathbf{Z} \sim \mathrm{MN}_{n\times m}(\mathbf{UDV}', \mathbf{\Sigma}, \mathbf{I}_m)$, where $\mathbf{\Sigma}$ now accounts for the approximation of choosing $k \ll l$ and the density function is

$$p(\mathbf{Z}|\mathbf{U}, \mathbf{D}, \mathbf{V}, \mathbf{\Sigma}, \mathbf{I}_m) = \frac{1}{(2\pi)^{nm/2}|\mathbf{\Sigma}|^{m/2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[(\mathbf{Z} - \mathbf{UDV}')'\mathbf{\Sigma}^{-1}(\mathbf{Z} - \mathbf{UDV}')\right]\right\}. \tag{6}$$

3.1.1 Model priors

To complete our model specification, we assign priors to $\mathbf{U}, \mathbf{D}, \mathbf{V}$, and $\mathbf{\Sigma}$, and estimate the model parameters using Bayesian techniques. Define $\mathbf{U}_{-i} \equiv [\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_k]$, $\mathbf{V}_{-i} \equiv [\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_k]$, $\mathbf{D}_{-i} \equiv \mathrm{diag}(d_1, \ldots, d_{i-1}, d_{i+1}, \ldots, d_k)$, and $\mathbf{E}_{-i} \equiv \mathbf{Z} - \mathbf{U}_{-i}\mathbf{D}_{-i}\mathbf{V}'_{-i}$, so that $\mathbf{Z} - \mathbf{UDV}' = \mathbf{E}_{-i} - d_i\mathbf{u}_i\mathbf{v}'_i$. Factoring the trace of the exponent of (6),

$$\begin{aligned}
\mathrm{tr}[(\mathbf{Z} - \mathbf{UDV}')'\mathbf{\Sigma}^{-1}(\mathbf{Z} - \mathbf{UDV}')] &= \mathrm{tr}[(\mathbf{E}_{-i} - d_i\mathbf{u}_i\mathbf{v}'_i)'\mathbf{\Sigma}^{-1}(\mathbf{E}_{-i} - d_i\mathbf{u}_i\mathbf{v}'_i)] \\
&= \mathrm{tr}[\mathbf{E}'_{-i}\mathbf{\Sigma}^{-1}\mathbf{E}_{-i} - 2d_i\mathbf{v}_i\mathbf{u}'_i\mathbf{\Sigma}^{-1}\mathbf{E}_{-i} + d_i^2\mathbf{v}_i\mathbf{u}'_i\mathbf{\Sigma}^{-1}\mathbf{u}_i\mathbf{v}'_i] \\
&= \mathrm{tr}[\mathbf{E}'_{-i}\mathbf{\Sigma}^{-1}\mathbf{E}_{-i} - 2d_i\mathbf{u}'_i\mathbf{\Sigma}^{-1}\mathbf{E}_{-i}\mathbf{v}_i + d_i^2\mathbf{v}'_i\mathbf{v}_i\mathbf{u}'_i\mathbf{\Sigma}^{-1}\mathbf{u}_i] \\
&= \mathrm{tr}[\mathbf{E}'_{-i}\mathbf{\Sigma}^{-1}\mathbf{E}_{-i} - 2d_i\mathbf{u}'_i\mathbf{\Sigma}^{-1}\mathbf{E}_{-i}\mathbf{v}_i + d_i^2\mathbf{u}'_i\mathbf{\Sigma}^{-1}\mathbf{u}_i].
\end{aligned}$$

The distribution $\mathbf{Z} \sim \mathrm{MN}_{n\times m}(\mathbf{UDV}', \boldsymbol{\Sigma}, \mathbf{I}_m)$ can then be written

$$p(\mathbf{Z}|\mathbf{u}_i, \mathbf{v}_i, d_i, \mathbf{U}_{-i}, \mathbf{D}_{-i}, \mathbf{V}_{-i}, \boldsymbol{\Sigma}) = \tag{7}$$
$$\frac{1}{(2\pi)^{nm/2}|\boldsymbol{\Sigma}|^{m/2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{E}'_{-i}\boldsymbol{\Sigma}^{-1}\mathbf{E}_{-i} - 2d_i\mathbf{u}'_i\boldsymbol{\Sigma}^{-1}\mathbf{E}_{-i}\mathbf{v}_i + d_i^2\mathbf{u}'_i\boldsymbol{\Sigma}^{-1}\mathbf{u}_i\right]\right\},$$

which enables inference on the columns of $\mathbf{U}$ and $\mathbf{V}$ and the elements of $\mathbf{D}$ individually (e.g., inference on $\mathbf{u}_i$ and $\mathbf{v}_i$). Recall from Section 2.2 that $\mathbf{u}_i|\mathbf{U}_{-i} \overset{d}{=} \mathbf{N}_i^u\widetilde{\mathbf{u}}_i|\mathbf{U}_{-i}$ and $\mathbf{v}_i|\mathbf{V}_{-i} \overset{d}{=} \mathbf{N}_i^v\widetilde{\mathbf{v}}_i|\mathbf{V}_{-i}$ where the columns of $\mathbf{N}_i^u$ and $\mathbf{N}_i^v$ span the null space of $\mathbf{U}_{-i}$ and $\mathbf{V}_{-i}$, respectively. We specify the prior distributions

$$\begin{aligned} d_i\widetilde{\mathbf{u}}_i|\mathbf{U}_{-i} &\sim \mathrm{N}_{n-k+1}(\mathbf{0}, \mathbf{N}_i^{u\prime}\boldsymbol{\Omega}_i^u\mathbf{N}_i^u)d_i^{n-k}\mathbb{I}(\widetilde{\mathbf{u}}_i \in \mathcal{V}_{1,n-k+1}) \\ d_i\widetilde{\mathbf{v}}_i|\mathbf{V}_{-i} &\sim \mathrm{N}_{m-k+1}(\mathbf{0}, \mathbf{N}_i^{v\prime}\boldsymbol{\Omega}_i^v\mathbf{N}_i^v)d_i^{m-k}\mathbb{I}(\widetilde{\mathbf{v}}_i \in \mathcal{V}_{1,m-k+1}) \\ d_i &\sim \mathrm{Unif}(0,\infty). \end{aligned} \tag{8}$$

For simplicity, we assume $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$, but this simplification can be relaxed if desired, e.g., by allowing $\boldsymbol{\Sigma}$ to be a structured non-diagonal covariance matrix. Last, we specify $\boldsymbol{\Omega}_i^u = \sigma_{u,i}^2\mathbf{C}_u(\boldsymbol{\theta}_{u,i})$ and $\boldsymbol{\Omega}_i^v = \sigma_{v,i}^2\mathbf{C}_v(\boldsymbol{\theta}_{v,i})$ where $\mathbf{C}_u(\boldsymbol{\theta}_{u,i})$ and $\mathbf{C}_v(\boldsymbol{\theta}_{v,i})$ are valid correlation matrices (i.e., the matrices are positive definite; see Section 2.3) and $\sigma_{u,i}^2$ and $\sigma_{v,i}^2$ are variance parameters. For $\sigma^2, \sigma_{u,i}^2$ and $\sigma_{v,i}^2$ we assign the non-informative half-t prior on the standard deviation as proposed by Huang and Wand (2013); specifically $\sigma \sim \textit{Half-t}(1, A)$, $\sigma_{u,i} \sim \textit{Half-t}(1, A_{u,i})$ and $\sigma_{v,i} \sim \textit{Half-t}(1, A_{v,i})$.

One major benefit of our proposed prior is now realized: the full conditional distribution of $\widetilde{\mathbf{u}}_i$ and $\widetilde{\mathbf{v}}_i$ is proportional to a normal distribution (see supplement S.1). This results in a Gibbs update step for both $\widetilde{\mathbf{u}}_i$ and $\widetilde{\mathbf{v}}_i$ within the larger Markov chain Monte Carlo (MCMC) sampling scheme (shown in supplement S.1), with computational benefits coming from known tricks for sampling from the normal distribution (e.g., the Cholesky decomposition). Additionally, we have the ability to specify, or learn, unique correlation matrices $\mathbf{C}_u(\boldsymbol{\theta}_{u,i})$ and $\mathbf{C}_v(\boldsymbol{\theta}_{v,i})$ for each basis function which, to the best of our knowledge, has not been previously considered.

3.1.2 Special cases

As discussed in Section 2.3, when $\boldsymbol{\Omega}_i \equiv \mathbf{I}$ our specified probabilistic model for SVD is equivalent to the fixed-rank SVD model proposed by Hoff (2007). Another interesting property is the relationship to the classic algorithmic approach, C-SVD. As discussed and shown empirically through simulation in the supplement (S.2.1), when $\boldsymbol{\Omega}_i = \mathbf{I}$ the mean of the full conditional distribution for the basis functions is equivalent to the estimates obtained by

C-SVD.

3.1.3 Model implementation

The SVD model (6) has several parameters that need to be specified: the number of basis functions k, the correlation matrices $\mathbf{C}_u(\boldsymbol{\theta}_{u,i})$ and $\mathbf{C}_v(\boldsymbol{\theta}_{v,i})$, and any hyperparameters associated with the correlation matrices $\boldsymbol{\theta}_{u,i}$ and $\boldsymbol{\theta}_{v,i}$. While in principle the value k can be estimated either informally, e.g., scree plots (Cattell, 1966), or formally, e.g., cross-validation (Wold, 1978) or the variable-rank model proposed by Hoff (2007), that is not the focus of this work. Through empirical testing, we have found that if the true k^* is less than the specified k, then the last $k - k^*$ basis functions of both $\mathbf{U}$ and $\mathbf{V}$ will have posterior credible intervals that cover zero at all, or nearly all, observations implying the basis function is not significant. Conversely, if the true k^* is greater than the specified k, there is little to no impact on the first k basis functions (i.e., the kth basis function is not biased to account for the lost information by not estimating the remaining $k^* - k$ basis functions). In choosing k for the proposed model, an empirical Bayes approach could also be taken. Specifically, one could compute the C-SVD, compute the cumulative amount of variance explained by the basis functions, and inform the value of k based on this "traditional" approach.

Regarding the correlation matrices $\mathbf{C}_{u,i}$ and $\mathbf{C}_{v,i}$, as previously mentioned the hyperparameters $\boldsymbol{\theta}_{u,i}$ and $\boldsymbol{\theta}_{v,i}$ can either be specified directly or learned within the broader hierarchical model. The latter choice would involve specifying a prior $p(\boldsymbol{\theta}_{u,i}, \boldsymbol{\theta}_{v,i})$ for these quantities and subsequently updating them within the MCMC algorithm. In the case of using the Matérn kernel to specify $\mathbf{C}_u(\boldsymbol{\theta}_{u,i})$ and $\mathbf{C}_v(\boldsymbol{\theta}_{v,i})$, recall that $\boldsymbol{\theta}_{u,i} = \{\nu_{u,i}, \rho_{u,i}\}$ and $\boldsymbol{\theta}_{v,i} = \{\nu_{v,i}, \rho_{v,i}\}$, where $\nu_{(\cdot)}$ describes the differentiability of the implied stochastic process and $\rho_{(\cdot)}$ describes the length-scale of the basis functions. We generally recommend setting $\nu_{(\cdot)} = 3.5$ so the basis functions are third-order continuous but not over- or under- smoothed (e.g., infinitely differentiable with $\nu = \infty$ or non-differentiable with $\nu = 0.5$, respectively). If the length-scale parameters are not estimated within the MCMC algorithm, they could be estimated offline via geostatistical techniques, e.g., estimating a semivariogram separately across both the rows and columns. In the simulations presented in Section 4 and for the application in Section 5 we opt to estimate the length-scale parameters within the MCMC algorithm.

3.2 Other modeling choices

Section 3.1 proposes a general model for observed data using a low-rank approach. However, there are other model specifications and corresponding matrix factorizations that can be seen as special cases of the SVD model. We discuss a few of these choices.

3.2.1 Principal components

As discussed in the introduction, PCA and SVD can be shown to produce an equivalent matrix factorization. To this end, we can analogously represent the process $\mathbf{Y} = \mathbf{U}\mathbf{A}$ where $\mathbf{U}$ is an orthonormal matrix of the eigenvectors of $\mathbf{Y}\mathbf{Y}'$, also known as the *principal components*, $\mathbf{A} = \mathbf{D}\mathbf{V}' = [\mathbf{a}_1, \ldots, \mathbf{a}_k]$ where $\mathbf{a}_i \sim N(0, \lambda_i \mathbf{I}_m)$, and $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_k)$ are the eigenvalues of $\mathbf{Y}\mathbf{Y}'$, also known as the *principal loadings*. To estimate $\mathbf{U}, \mathbf{A}$, and $\boldsymbol{\Lambda}$ under this parameterization, there are two choices: (1) factor $\mathbf{E}_{-i} = \mathbf{Z} - \mathbf{U}_{-i}\mathbf{A}_{-i}$ in (7) and we assign the prior $\lambda_i \widetilde{\mathbf{u}}_i | \mathbf{U}_i \sim \mathrm{N}_{n-k+1}(\mathbf{0}, \mathbf{N}_i^{u\prime} \boldsymbol{\Omega}_i^u \mathbf{N}_i^u) \lambda_i^{n-1}$, or (2) estimate the parameters from the SVD model and compute $\mathbf{A}$ as the posterior product of $\mathbf{D}$ and $\mathbf{V}'$. For choice (1), only the columns of $\mathbf{U}$ are dependent where the elements of $\mathbf{A}$ are independent, resulting in only the principal components having dependence. If choice (2) is taken, then the columns of $\mathbf{U}$ and rows $\mathbf{A}$ can be modeled dependently, where $\mathbf{A}$ is dependent through the specification of $\mathbf{V}$. For PCA parameterization, we advocate for choice (2) as there is more control over the model than choice (1).

3.2.2 Including covariates

The general model (4) allows for more complex model structure, such as including covariates. Traditionally, data are centered, or de-trended, prior to computing the SVD/PCA decomposition. However, within (4) a mean term can be accommodated by modeling $\mathbf{M}$. We first consider a linear model for $\mathbf{M}$, $\mathrm{vec}(\mathbf{M}) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X} \in \mathbb{R}^{nm \times p}$ is a matrix of observed covariates and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown parameters. To estimate $\mathbf{U}, \mathbf{D}, \mathbf{V}$ under this parameterization, $\mathbf{E}_{-i} = \mathbf{Z} - [\mathbf{X}\boldsymbol{\beta}] - \mathbf{U}_{-i}\mathbf{D}_{-i}\mathbf{V}'_{-i}$ in (7), where $[\mathbf{X}\boldsymbol{\beta}]$ denotes the reconstructed matrix of size $n \times m$. To estimate $\boldsymbol{\beta}$, we vectorize the model to get $\mathrm{vec}(\mathbf{Z}) \sim MVN_{nm}(\mathbf{X}\boldsymbol{\beta} + \mathrm{vec}(\mathbf{U}\mathbf{D}\mathbf{V}'), \mathbf{I}_m \otimes \boldsymbol{\Sigma})$, assign the diffuse normal prior $\boldsymbol{\beta} \sim MVN_p(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_p)$, with σ_β^2 large, and get a standard normal-normal conjugate update for $\boldsymbol{\beta}$.

This idea can be extended to a nonlinear function, say $\mathrm{vec}(\mathbf{M}) = f(\mathbf{X}, \boldsymbol{\beta})$, where $f()$ is a nonlinear function. For example, generalized additive models (Hastie and Tibshirani, 2017) or differential equations (Berliner, 1996; Wikle, 2003) could be used to model the nonlinear function. However, care will likely need to be taken for the nonlinear case such that the nonlinear function is not too flexible, thereby conflicting with the random effect (e.g., see 4.4).

4 Synthetic data examples

We conduct three simulation studies to illustrate various aspects of the prior. The first simulation provides justification for basis function-specific structure as opposed to a shared structure for all the basis functions. The second illustrates how measurement error and model rank impact basis function recovery. The last simulation investigates the ability to recover covariates when there may be confounding between the fixed and random effects.

4.1 Data generation

For all simulations, the target "true" basis functions $\mathbf{U}$ and $\mathbf{V}$ are simulated according to the generating mechanism described in section 2.1 (e.g., to produce the orthonormal matrix in (1)) with $\mathbf{\Omega}_i^u = \mathbf{C}_u(\boldsymbol{\theta}_{u,i})$ and $\mathbf{\Omega}_i^v = \mathbf{C}_v(\boldsymbol{\theta}_{v,i})$ where the elements of $\mathbf{C}_u(\boldsymbol{\theta}_{u,i})$ and $\mathbf{C}_v(\boldsymbol{\theta}_{v,i})$ are defined by the Matérn correlation function with $\boldsymbol{\theta}_{u,i} = (\nu_{u,i}, \rho_{u,i})$ and $\boldsymbol{\theta}_{v,i} = (\nu_{v,i}, \rho_{v,i})$. Data is simulated according to $Z(x,t) \sim N(M(x,t) + Y(x,t), \sigma^2)$ with $x = x_1, \ldots, x_n$ equally spaced in $\mathcal{X} = [-5, 5]$, $t = t_1, \ldots, t_m$ equally spaced in $\mathcal{T} = [0, 10]$, $n = 100$, $m = 100$, and $Y(x,t)$ being the (x,t) element of the matrix $\mathbf{Y} = \mathbf{UDV}'$. The specification of $\mathbf{M} = [M(x,t)]_{(x,t)\in\mathcal{X}\times\mathcal{T}}$ is described in each of the following subsections. The value of σ^2 is chosen to match a target signal-to-noise ratio (SNR): let $\boldsymbol{\eta}$ be a random $n \times m$ matrix of iid standard normal random variables, then, $\sigma = \sqrt{\frac{var(\mathbf{M}+\mathbf{Y})}{\text{SNR}*var(\boldsymbol{\eta})}}$. Ultimately, the simulated data is $\mathbf{Z} = \mathbf{M}+\mathbf{Y}+\sigma\boldsymbol{\eta}$ (see the supplement Figure S.5) with $var(\mathbf{M}+\mathbf{Y})/var(\mathbf{Z}-\mathbf{M}-\mathbf{Y}) = \text{SNR}$.

4.2 Synthetic example #1: basis function-specific length scales

The first simulation study assesses how our model recovers the underlying basis functions when the true basis functions have differing length-scales. We compare our "variable model," in which we allow each basis function to have unique structure that is estimated from the data, to a "grouped model," in which all basis functions have a shared structure that is also estimated from the data. A distinguishing feature of our methodology is that we can model basis function-specific structure, in comparison to other recent work (Pourzanjani et al., 2021; Jauch et al., 2021) wherein all basis functions have the same length-scales. Both models are described in Section 3.1: in the variable model, $\rho_{\cdot,i}$ and $\rho_{\cdot,j}$ need not be equal, while in the grouped model, we impose the restriction that $\rho_{\cdot,i} = \rho_{\cdot,j}$, for $i, j = 1, \ldots, k$. The grouped model is a special case of the variable model, illustrating the enhanced flexibility of our methodology relative to existing approaches.

To explore the effect of basis function-specific structure, we generate data where the length-scale for each basis function varies from larger to smaller in an exponentially decreasing trend similar to what is shown in Figure 1. To determine the effect of measurement

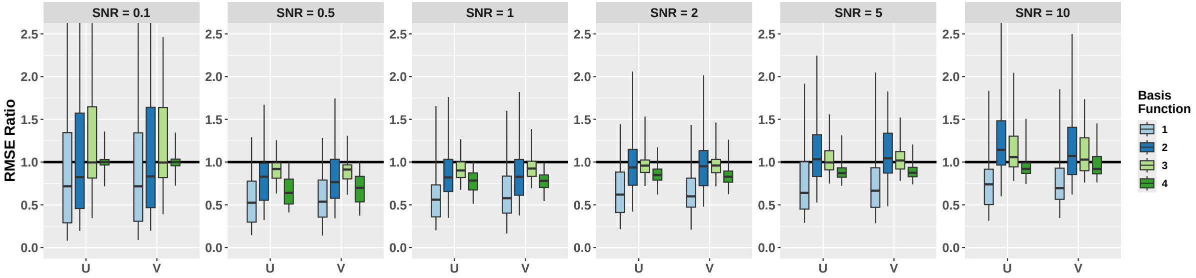


Figure 2: Box plots of the ratio of the RMSE for the variable model divided by the grouped model for **U** (left) and **V** (right) stratified by SNR (sub-panel) and basis function (color) with a horizontal line at 1. For each box, the lower and upper hinge are the 25th and 75th percentiles, respectively, the line within the box is the median, and the lower and upper whiskers are 2.5 and 97.5 percentiles. Note, we have limited the y-axis to ease visual comparison between panels and only the first panel, with a SNR = 0.1, has values outside of the range.

error in conjunction with varying basis function length-scale, we generate data sets with σ^2 chosen such that SNR $= [10, 5, 2, 1, 0.5, 0.1]$. For this simulation study, we do not consider the effect of $\mathbf{M}$, and all data is simulated with $\mathbf{M} \equiv 0$. For all data generation, we specify the true number of basis functions $k = 4$ with covariance parameters $\nu_{(\cdot),i} = 3.5$ for $i = 1, \dots, k$ and $\boldsymbol{\rho}_{(\cdot)} = (3.5, 1, 0.5, 0.25)$ for both $\mathbf{U}$ and $\mathbf{V}$, and diagonal matrix $\mathbf{D} = \text{diag}(40, 30, 20, 10)$. For each SNR, we obtain 10000 posterior samples of the model parameters and discard the first 5000 as burn-in for both the variable and grouped model. The process is repeated 100 times for each SNR to help understand the variability in the results.

For each simulation and model, we calculate the element-wise average root mean squared error (RMSE) of the posterior mean for each basis function in **U** and **V** compared to their corresponding true value. To compare the RMSE estimates of the variable to grouped model, Figure 2 shows the ratio of the RMSE estimate for the variable model over the group model for **U** (left) and **V** (right) stratified by the SNR (sub-panels) and by the basis function (color) along with a horizontal reference line at one.

RMSE ratios less than one favor the variable model. From the figure, we see basis functions 2 and 3 for both **U** and **V** have ratios closest to 1 for all values of SNR. In contrast, basis functions 1 and 4 for both **U** and **V** have ratios that are systematically less than 1 for all values of SNR except 0.1. The reason the variable model has improved RMSE performance for 1 and 4 is because the estimate for ρ for the grouped model is pulled toward the average length-scale value, which is close to the true length scale for basis functions 2 and 3. This bias results in the grouped model over-fitting basis function 1 (since the pooled estimate of the length scale is less than the true length scale) and under-fitting basis function 4 (since the pooled estimate of the length scale is larger than the true length scale); see estimates in Figure S.3 for a visual example of the over- and under-fitting.

In summary, our first synthetic example verifies that when the data have differing structures in the underlying basis functions, failing to account for those different structures results in systematically larger errors in the basis function estimates. The true structures can only be appropriately captured when the underlying statistical model directly accounts for basis function-specific structure.

4.3 Synthetic example #2: model rank

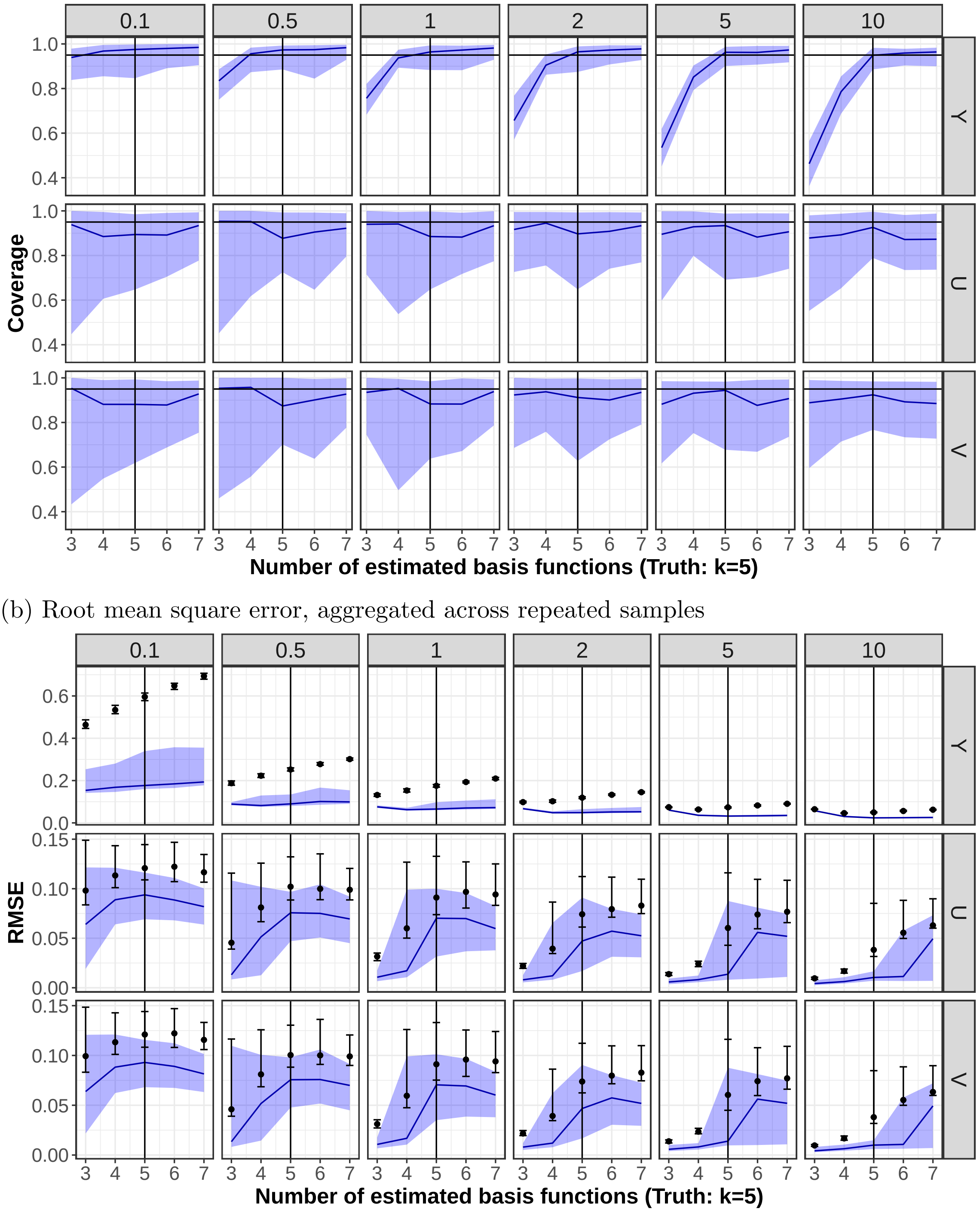
We now conduct a simulation study to illustrate the impact of SNR and model rank k on basis function recovery. To determine the effect of measurement error, we again generate data sets with σ^2 chosen such that SNR $= [10, 5, 2, 1, 0.5, 0.1]$. As with the previous simulation study, all data is simulated with $\mathbf{M} \equiv 0$. For all data generation, we set the true number of basis functions $k^* = 5$ with covariance parameters $(\nu_{(\cdot),i}, \rho_{(\cdot),i}) = (3.5, 3)$ for both $\mathbf{U}$ and $\mathbf{V}$ and for all $i = 1, \ldots, k^*$, and diagonal matrix $\mathbf{D} = \text{diag}(40, 30, 20, 10, 5)$. One realization of the simulated data with SNR $= 1$ and the $\mathbf{U}$ and $\mathbf{V}$ basis functions are shown in Figure S.4 in the supplement.

As discussed in Section 3.1.3, using this model only requires specification of k, the number of basis functions used in $\mathbf{U}$ and $\mathbf{V}$, and kernels for $\mathbf{C}_u(\boldsymbol{\theta})$ and $\mathbf{C}_v(\boldsymbol{\theta})$. To investigate how possible mis-specification of the number of basis functions impacts model recovery, we estimate the model with $k = [3, 4, 5, 6, 7]$ for each level of SNR. Additionally, we specify a Matérn kernel with smoothness parameter $\rho = 3$ for the correlation structure for all basis functions. For each SNR and k combination, we obtain 10000 posterior samples of the model parameters, discarding the first 5000 as burn-in. We repeat this process 100 times.

For each posterior simulation, we calculate the 95% coverage rate (CR) and RMSE for $\mathbf{U}$, $\mathbf{V}$, and the "true" surface $\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}'$. If the true k^* is greater than the specified k, the empirical CR and RMSE are computed only for the first k basis functions and then averaged over the k estimates (e.g., we do not consider the last $k^* - k$ basis functions when computing CR and RMSE). If the true k^* is less than the specified k, the empirical CR and RMSE for the "extra" $k - k^*$ basis functions are compared to the zero line and the reported CR and RMSE values are obtained by averaging over the k estimates. Additionally, for each simulation we computed the C-SVD using the base linear algebra library, *LinearAlgebra.jl*, in Julia (Bezanson et al., 2017) and computed the RMSE of the calculated $\mathbf{U}, \mathbf{V}$, and reconstructed surface $\mathbf{Y}$ assuming the same truncation value k. The coverage rates and the RMSE are shown in Figure 3. The results of one simulation are shown in Figure 4 based on the data shown in Figure S.4 in the supplement.

From Figure 3(a), we see our median coverage rate for the $\mathbf{U}$ (middle row) and $\mathbf{V}$ (bottom row) basis functions (blue line) is near the nominal level (horizontal black line) and the 95%

(a) Coverage rate, aggregated across repeated samples



(b) Root mean square error, aggregated across repeated samples

Figure 3: Validation results from the synthetic data example, showing coverage rate (top) and root mean square error (bottom). In each panel, the solid blue line is the median Monte Carlo coverage rate and shaded regions are the 95% Monte Carlo uncertainty bounds for the coverage rate over synthetic replicates. Results are shown for varying levels of SNR and values of k for the recovered surface $\mathbf{Y}$ (top), $\mathbf{U}$ basis functions (middle), and $\mathbf{V}$ basis functions (bottom). The SNR values range from 0.1 (left) to 10 (right). The black vertical line indicates the true value $k^* = 5$ and the horizontal black line for (a) is at 95% (the nominal coverage rate). In panel (b), the black point and error bars show the median and 95% bootstrapped confidence interval for the RMSE using the algorithmic C-SVD method.

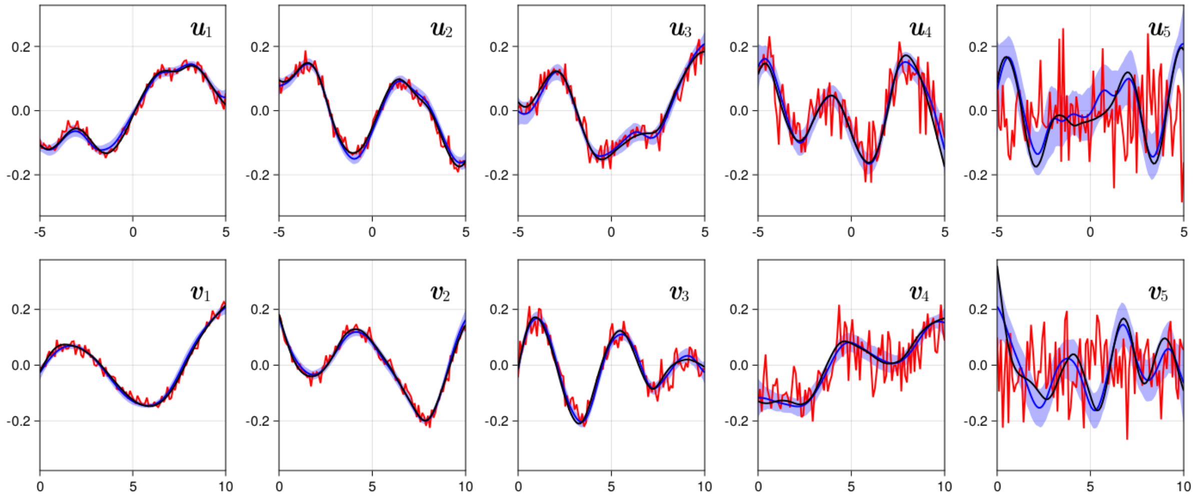


Figure 4: Posterior mean (blue line), 95% credible intervals (shaded blue region), truth (black line), and C-SVD estimate (red line) for the $\mathbf{U}$ and $\mathbf{V}$ basis functions from a random simulation. The data associated with this random simulation is shown in Figure S.4.

Monte Carlo uncertainty bounds (MCUB) for the coverage rate (blue shaded region) covers the nominal level for all SNR levels and regardless of the specification of k. This implies that posterior uncertainties are well calibrated and robust to mis-specifications of the number of estimated basis functions, regardless of the magnitude of the noise. For the recovered data (top row), we see the 95% MCUB cover the nominal level for all SNR levels with k greater than 5. However, for k less than 5, achieving the nominal coverage depends on SNR: in low signal cases (e.g., SNR $= 0.1$), the uncertainties are well calibrated, while posterior uncertainties are too small (i.e., coverage of the truth is much less than the nominal level) when the signal is stronger (SNR > 0.5). This counterintuitive result is due to the impact of unaccounted signal for higher-order basis functions ($i = 4$ and/or $i = 5$) on the signal: for large SNR, individual basis functions both (a) contribute more to the overall uncertainty in the data and also (b) have narrower posterior distributions, such that ignoring one or more true basis functions causes the model to underestimate data uncertainties (e.g., see Figure 4). Conversely, for smaller SNR, there is more uncertainty in each basis function estimate and the impact of higher-order basis functions on the estimated surface is reduced, to the extent that the model can recover the nominal coverage of the data.

For the RMSE, shown in Figure 3(b), the most notable result is that the median RMSE for our approach (blue line) is systematically lower than the corresponding RMSE from the algorithmic C-SVD approach for both data (top row) and basis functions (middle and bottom rows), across SNR levels and specification of k. In other words, estimates of the basis functions in both $\mathbf{U}$ and $\mathbf{V}$ and the recovered data have systematically lower errors than what one can obtain from the algorithmic approach. Regarding RMSE for estimates of the recovered surface, the median error (blue line) decreases as a function of SNR, as expected, and interestingly the data RMSE is relatively insensitive to specification of k. For the $\mathbf{U}$ and $\mathbf{V}$ basis functions (middle and bottom rows, respectively, of Figure 3), we see that trajectories of RMSE estimates for our proposed approach and the C-SVD mirror each other, with our estimates being systematically, but not significantly, lower. However, across SNR levels, the RMSE actually increases as one moves from $k = 3$ to $k = 7$ (even though the true $k^* = 5$). For SNR equal to 5 and 10, we see a dramatic spike in the RMSE estimate and uncertainty for the $\mathbf{U}$ and $\mathbf{V}$ basis functions for $k = 6$ and 7. This is because we are comparing against the zero line for these cases: while the uncertainty bounds for these basis function covers the zero line (as seen in the coverage results in Figure 3a.), there is a lot of variability in these estimates (with relatively lower uncertainty due to larger signal), leading to inflated RMSE values.

In conclusion, this synthetic data example shows the proposed method has well calibrated uncertainty and significantly reduces the impact of measurement noise on the basis function estimates. However, there is a significant trade-off in choosing k to be too small or large based on the magnitude of the SNR. Based on our simulation, there will be significant bias in the recovered surface but *not* in the estimated basis functions if k is too small and the SNR is low. Additionally, there will *not* be significant bias in the recovered surface or in the estimated basis functions if either k is too small and the SNR is large or k is too large. The only trade-off for k too large is inflated RMSE's for the extraneous basis functions, which could lead to underestimated RMSE's in the recovered surface. Therefore, we suggest erring on the side of choosing k to be too large.

4.4 Synthetic example #3: covariates

To illustrate how covariates impact the estimation of the basis functions, we now include the fixed effect $\mathbf{M}$ when simulating data and specify the SNR to be 2. We consider three different cases of the model for $\mathbf{M}$: (M1) independent fixed and random effects, (M2) strongly confounded spatial and temporal fixed and random effects, and (M3) weakly confounded spatial and temporal fixed and random effects. For all three models, we specify $\text{vec}(\mathbf{M}) = \mathbf{X}\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_4) = (-2, 0.6, 1.2, -0.9)$ and $\mathbf{X}$ is a nm by 4 matrix. For each model, the covariates are generated as:

M1 - Each element of $\mathbf{X}$ is i.i.d. $N(0, 0.2^2)$.

M2 - Let $\widetilde{\mathbf{x}}_{1,s}, \widetilde{\mathbf{x}}_{2,s} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_s)$, $\widetilde{\mathbf{x}}_t \sim N_m(\mathbf{0}, \boldsymbol{\Sigma}_t)$, and $\mathbf{x}_{st} \sim \mathrm{N}_{nm}(\mathbf{0}, \boldsymbol{\Sigma}_{st})$ where $\boldsymbol{\Sigma}_s, \boldsymbol{\Sigma}_t$, and $\boldsymbol{\Sigma}_{st}$ are correlation matrices specified using the Matérn kernel with smoothness parameter $\nu = 3.5$ and length-scale parameter $\rho = 3, 3$ and 1, respectively, which is equal to the length-scale of the spatial and temporal random effect, respectively. Then,

model		β_1	β_2	β_3	β_4
	true	-2	0.6	1.2	-0.9
	mean	-2.032	0.632	1.204	-0.873
M1	lower CI	-2.082	0.583	1.156	-0.921
	upper CI	-1.983	0.680	1.252	-0.824
	mean	-1.995	0.636	1.071	-0.875
M2	lower CI	-2.036	0.531	1.014	-0.913
	upper CI	-1.943	0.747	1.118	-0.807
	mean	-2.005	0.601	1.194	-0.908
M3	lower CI	-2.016	0.589	1.179	-0.938
	upper CI	-1.994	0.614	1.209	-0.882

Table 1: Posterior mean (top row), lower 95% credible interval (middle row), and upper 95% credible interval (bottom row) for the regression coefficients of models M1–M3 (top-bottom).

$\mathbf{X} = [\mathbf{x}_{1,s}, \mathbf{x}_{2,s}, \mathbf{x}_t, \mathbf{x}_{st}]$ is a $nm \times 4$ matrix where $\mathbf{x}_{1,s} = \mathbf{I}_m \otimes \widetilde{\mathbf{x}}_{1,s}, \mathbf{x}_{2,s} = \mathbf{I}_m \otimes \widetilde{\mathbf{x}}_{2,s}$, and $\mathbf{x}_t = \widetilde{\mathbf{x}}_t \otimes \mathbf{I}_n$.

M3 - The covariate matrix is created in the same manner as in M2 except the length-scale of $\mathbf{\Sigma}_s, \mathbf{\Sigma}_t$, and $\mathbf{\Sigma}_{st}$ are $\rho = 0.3, 0.3$ and 1, respectively.

For each covariate specification M1–M3, we implement our methodology with $k = 5$, a Matérn kernel with smoothness parameter $\nu = 3$ for the correlation structure for all basis functions, and a diffuse normal prior, $\mathrm{N}(0, 10^2)$, on each element of $\boldsymbol{\beta}$. We obtain 10000 posterior samples of the model parameters, discarding the first 5000 as burn-in. Posterior summaries of the regression coefficients are shown for each model in Table 1. From the table, we see only β_3 from M2 has a credible interval that does not cover the true value, indicating the model is able to reasonably recover the fixed effects under all three scenarios. To determine the model's ability to correctly recover the random effect, we computed the point-wise 95% posterior coverage rate for the random effect $\mathbf{Y} = \mathbf{UDV}'$ for each M1, M2, and M3, which are 0.965, 0.276, and 0.984, respectively. Therefore, when the fixed and random effect are independent or they have different spatial and temporal frequencies (weakly confounded), the model is able to correctly identify both model components. When the fixed and random effects have similar, or in this example equal, spatial and temporal frequencies, the model is unable to properly capture the random effect but can still capture the fixed effect.

Based on previous work by Paciorek (2010) discussing the issue of scale with spatial mixed-effects models, our results are not surprising. Specifically, if the fixed and random effects operate on different scales (either spatially or temporally), Paciorek (2010) rigorously argues the fixed and random effects are identifiable. If they operate on similar (or equivalent) scales, they are not identifiable. If interpretation of the random effect is not important, the random effect can restricted to be orthogonal to the fixed effect, thereby making the random effect identifiable on the space orthogonal to the fixed effect (Reich et al., 2006; Hodges and Reich, 2010; Hanks et al., 2015). However, there has been debate as to the validity of modeling the random effect on the restricted space (Zimmerman and Ver Hoef, 2022). Because this is not the main goal of the paper, for now we simply recommend being cognizant of these issues.

5 Surface air temperature

As discussed in the introduction, empirical orthogonal functions, or EOFs, are commonly used in climate sciences to summarize modes of variability in atmospheric systems. Typically, external factors that could be driving the system are referred to as *climate forcings* and modeled as fixed effects, while “unforced” year-to-year variability is modeled as a spatial, temporal, or spatio-temporal random effect and referred to as *internal variability.* Importantly, when EOF analysis is applied to climate data where the long term trends have been removed, this can be considered a method for characterizing the internal variability of the system. Particularly for extreme temperature events, EOFs are an important tool for understanding how internal variability combines with long-term trends to produce short-term events that have a large impact on human systems (Grotjahn et al., 2016). Historically, estimates of the internal variability are derived from ensembles of climate models and rarely computed from observational data products. Here, we explore our ability to estimate the internal variability of monthly maximum two-meter air temperature in the Pacific Northwest, where it is important to account for spatial and temporal structures in the extreme measurements (again see, e.g., Grotjahn et al., 2016). Such estimates are important for understanding the statistics of monthly maximum temperatures in this region, particularly in light of the recent devastating heatwave that impacted this region in the summer of 2021 (Bercos-Hickey et al., 2022).

We use gridded monthly maximum two-meter air temperature data (tXx) by extracting the largest daily maximum two-meter air temperature each month from the ERA5 reanalysis dataset (Hersbach et al., 2020) at 0.25° horizontal resolution from January 1979 to December 2021. The data are centered by subtracting off the global mean. We focus on the subset of data from 44°- 53°N and 116°- 128°W, for a total of 1813 spatial locations across 516 time points. While it is possible to include relevant covariates for this analysis (e.g., greenhouse gas emissions, the El Niño/Southern Oscillation, urbanization, and drought conditions) using a model for $\mathbf{M}$ (e.g. Section 3.2.2), this would have resulted in a substantial number of parameters to estimate and is not the main focus of this work. Therefore, we opt instead to focus on the model for the random effect and simply centered the data *a priori* to parameter estimation.

As discussed in the introduction, Figure 1 shows empirical evidence that the basis functions resulting from a SVD of tXx may have different structure. We proceed with this assumption. Therefore, we parameterize the covariance matrix for the prior of the spatial basis functions using the Matérn kernel with smoothness $\nu = 3.5$ and the covariance matrix for the prior of the temporal basis functions using the Gaussian kernel. The effective range for both the spatial and temporal basis functions are estimated along with other model parameters. We specify $k = 10$ based on the first 10 basis functions explaining approximately 99% of the variance as determined from the C-SVD decomposition. We obtain 10000 samples from the posterior, discarding the first 5000 as burn-in, where convergence is assessed graphically with no issues detected.

Posterior summaries of three spatial basis functions (2, 5, and 7), three temporal basis functions (2, 5, and 7), and all length-scale estimates are shown in Figure 5 A), B), and C), respectively. We highlight basis function 2 because it has little to no significant difference between C-SVD estimate, and 5 and 7 because they contain many spatial and temporal locations with significant differences. Panel a) depicts summaries of three spatial basis functions $\mathbf{u}_2$ (top), $\mathbf{u}_5$ (middle), and $\mathbf{u}_7$ (bottom), where the left column are the estimates from C-SVD, the middle column are the posterior means from our proposed model, and the right column are the posterior difference between the posterior mean and the algorithmic estimate where locations whose 95% credible interval does not cover zero are denoted with an 'x'. Panel b) contain estimates of three temporal basis functions $\mathbf{u}_2$ (top), $\mathbf{u}_5$ (middle), and $\mathbf{u}_7$ (bottom), where the black line is the C-SVD estimates, blue line is the posterior mean from our proposed model, and blue shaded region are the 95% CIs where a vertical line denotes the 95% CI does not cover the C-SVD estimate. The last panel, c), are posterior mean estimates of the length-scale parameter (dot) and 95% credible intervals (error bars) of the correlation kernel for each spatial (left) and temporal (right) basis functions, where blue estimates correspond to the selected basis functions for panels a) and b). Posterior summaries of all 10 spatial and temporal basis functions are included in the supplement.

Comparing the spatial plots of the posterior mean to the deterministic counterpart (Figure 5A), the posterior estimates are much smoother spatially and for the fifth and seventh basis functions, the estimates are significantly different over much of the spatial region. The estimates, both deterministic and probabilistic alike, have an interpretation that makes sense physically. The second basis function (top row) has a clear land-sea contrast and distinguishes between the plains (purple) and mountains (green). The fifth basis function captures the influence of the low-lying coastal region and foothills of Canadian Rockies (pur-

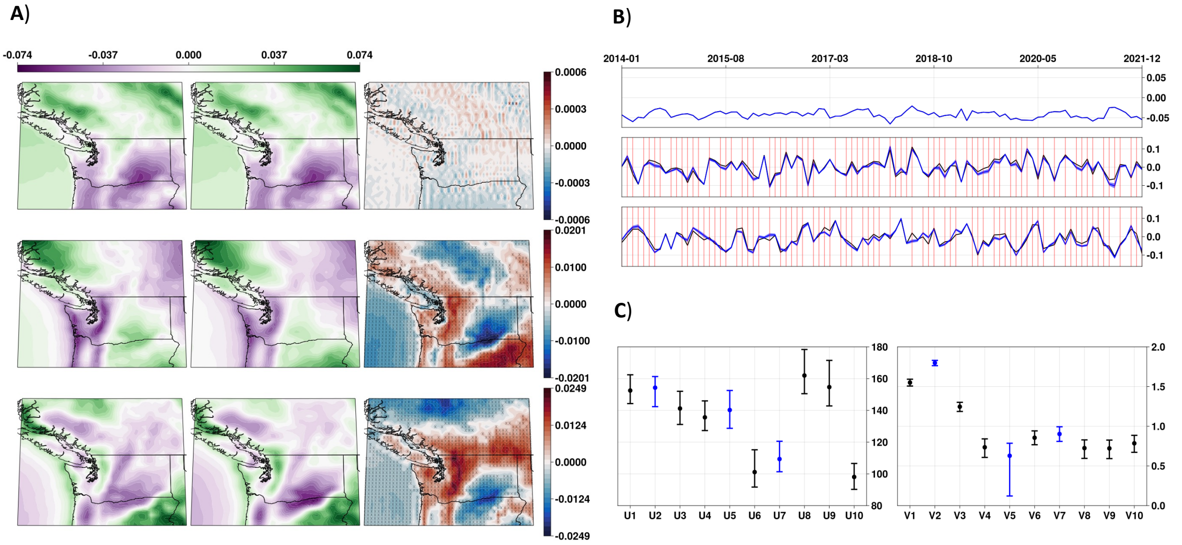


Figure 5: **a)** Estimated spatial basis functions $\mathbf{u}_2$ (top), $\mathbf{u}_5$ (middle), and $\mathbf{u}_7$ (bottom). The left column are the estimates from C-SVD, the middle column are the posterior means, and the right column are the posterior difference between the posterior mean and the algorithmic estimate where locations whose 95% credible interval does not cover zero are denoted with an 'x'. **b)** Estimated temporal basis functions $\mathbf{v}_2$ (top), $\mathbf{v}_5$ (middle), and $\mathbf{v}_7$ (bottom). For each panel, the black line are the estimates from C-SVD, blue line are the posterior means, and blue shaded region are the 95% CIs where a vertical line denotes the 95% CI does not cover the C-SVD estimate. **c)** Posterior mean estimate of the length-scale parameter (dot) and 95% credible intervals (error bars) of the correlation matrix for each spatial (left) and temporal (right) basis functions. Estimates in blue correspond to the selected basis functions for panels **a)** and **b)**.

ple) in contrast to the wet/dry regimes in Canada and Oregon/Washington (green). The seventh combines multiple physical features and aligns well with geographical features such as topography and appears to capture steep gradient contrasts.

Regarding the temporal estimates (Figure 5B), the second basis function (top) does not have any time points with significantly different estimates than the C-SVD counterpart. However, both the fifth (middle) and seventh (bottom) do have significant differences (denoted with the vertical lines), and we see the posterior means produce smoother estimates than the C-SVD counterparts.

Additionally, the basis functions all have different posterior mean length-scale estimates. For the spatial basis functions, $\mathbf{u}_6, \mathbf{u}_7$, and $\mathbf{u}_{10}$ have significantly smaller values than the other six, as determined by the range of the 95% CI (Figure 5C, left), and for the temporal, the first three have significantly larger values than the other seven, as determined by the range of the 95% CI (Figure 5C, right). This shows we are able to capture the spatial and temporal relationship within each basis function and that the spatial and temporal relationship is different across basis functions.

Importantly for climate science, we are able to provide estimates of the internal variability of a system from observational data, in this example monthly maximum two-meter air temperature of daily maxima, by reconstructing the internal variability using posterior estimates of our structured basis functions. The estimates account for measurement uncertainty, spatial and temporal dependence, and have quantifiable uncertainty. These estimates can then be used to account for the internal variability of a system and help isolate the extent to which external factors are driving changes to the system. Additionally, producing ensembles of weather variables like extreme temperature using climate models can be computationally intensive. However, we can now sample directly from the posterior distribution of the internal variability of extreme monthly temperatures, accounting for the spatial structures innate to the underlying data. These posterior samples are analogous to ensembles of the climate system and computationally much cheaper to compute than ensembles of climate model runs.

6 Discussion

We proposed a novel prior distribution for structured orthonormal matrices that is an extension of Hoff (2007), where the individual basis functions can be modeled dependently. The prior is based on the projected normal distribution which we augment with a latent length parameter. When our prior is combined with a normal data model, the resulting full conditional distributions for the basis functions are conjugate, resulting in analytically straightforward MCMC sampling. We describe how the prior can be used to conduct posterior inference on a general class of probabilistic SVD models and how to extend the proposed model to various other applications. We discussed various mathematical properties of our probabilistic SVD model (supplement S.2) and illustrated its capability through multiple simulation studies. The model is then used to draw inference on the internal variability of extreme two-meter air temperature, allowing us to quantify space-time structures in a complex climate process.

The synthetic data examples and application presented in Sections 4 and 5, respectively, all highlight the model's efficacy on gridded, i.e., uniformly spaced, data. However, the model is equally well suited for non-uniformly spaced data so long as the spacing is consistent within space and within time. If the data are not spaced consistently within space and within time, this would constitute a missing data problem, which we plan to explore in future work. In addition, our model assumes normally distributed errors. This assumption can be relaxed by, for example, assuming a hierarchical structure and modeling the mixed-effects as a latent process.

Another area for possible extension could explore the concept of regularized basis functions through the posterior mode of the basis functions. Similar to the Bayesian Lasso (Park and Casella, 2008) or Bayesian Group Lasso spatial data (Hefley et al., 2017), an ℓ_1 penalty could be imposed by representing a Laplace distribution as a scale mixture of normal distributions. The addition of a penalty term, especially a penalty that forces values to zero, could produce sparse dependently structured basis functions whose importance within the spatial context is explored by Wang and Huang (2017).

The choice of the number of basis functions, k, is the only major subjective choice in our proposed probabilistic SVD model. While we show the mis-specification of k does not have a negative impact when erring on the side of k being too large, a more flexible model estimating k is attractive. To estimate k, Hoff (2007) proposed a variable-rank model utilizing the so-called spike-and-slab variable selection prior (Mitchell and Beauchamp, 1988). However, because of the difference in our prior compared to the prior proposed by Hoff (2007), incorporating the spike-and-slab prior into our proposed model would require extra theoretical work. Work focused on estimating the rank k with our framework would produce a very flexible approach for modeling spatio-temporal random effects.

Finally, our proposed prior does have the disadvantage of relying on a column-wise sampling strategy. Specifically, within each MCMC iteration, there is a required $\mathcal{O}(n^3)$ cost of computing the orthonormal basis for the null-space $\mathbf{N}_i^u$ and $\mathbf{N}_i^v$ (see the supplement for more discussion). The additional flexibility our approach offers comes at the cost of the computational gains from the methods by Pourzanjani et al. (2021) and Jauch et al. (2021), which propose solutions to this column-wise strategy.

7 Competing interests

No competing interest is declared.

8 Author contributions statement

J.S.N., M.D.R., and F.J.B. formulated the research question. J.S.N. wrote the code, conducted synthetic examples, and performed the main analysis. J.S.N, M.D.R, and F.J.B wrote and reviewed the manuscript.

9 Acknowledgments

All code is written in Julia (Bezanson et al., 2017) and is available publicly on GitHub at https://github.com/jsnowynorth/BayesianSVD.jl. Additionally, the Bayesian SVD model has been developed into a Julia package *BayesianSVD.jl*, which can also be downloaded from GitHub at https://github.com/jsnowynorth/BayesianSVD.jl for easy use. All data are publicly available at the Copernicus Climate Data Store https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview.

This research was supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and by the Regional and Global Model Analysis Program area within the Earth and Environmental Systems Modeling Program. This research was also supported by the National Science Foundation [OAC-1931363, ACI-1553685] and the National Institute of Food & Agriculture [COL0-FACT-2019]. The research used resources of the National Energy Research Scientific Computing Center (NERSC), also supported by the Office of Science of the U.S. Department of Energy, under Contract No. DE-AC02-05CH11231.

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

References

Bailey, S. (2012). Principal component analysis with noisy and/or missing data. *Publications of the Astronomical Society of the Pacific*, 124(919):1015–1023.

Bercos-Hickey, E., O'Brien, T. A., Wehner, M. F., Zhang, L., Patricola, C. M., Huang, H., and Risser, M. D. (2022). Anthropogenic contributions to the 2021 Pacific Northwest heatwave. *Geophysical Research Letters*, 49(23).

Berkooz, G. (1993). The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25(1):539–575.

Berliner, L. M. (1996). Hierarchical Bayesian time series models. In *Maximum Entropy and Bayesian Methods*, pages 15–22. Springer Netherlands, Dordrecht.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98.

Buhmann, M. D. (2000). A new class of radial basis functions with compact support. *Mathematics of Computation*, 70(233):307–318.

Byrne, S. and Girolami, M. (2013). Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276.

Chen, Q., Cao, J., and Xia, Y. (2022). Physics-enhanced PCA for data compression in edge devices. *IEEE Transactions on Green Communications and Networking*, 6(3):1624–1634.

Chikuse, Y. (2003). *Statistics on Special Manifolds*, volume 174 of *Lecture Notes in Statistics*. Springer New York, New York, NY.

Cressie, N. and Johannesson, G. (2006). Spatial prediction for massive datasets. *Australian Academy of Science*, 1247:1–11.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

Demmel, J. and Kahan, W. (1990). Accurate singular values of bidiagonal matrices. *SIAM Journal on Scientific and Statistical Computing*, 11(5):873–912.

Epps, B. P. and Krivitzky, E. M. (2019). Singular value decomposition of noisy data: Noise filtering. *Experiments in Fluids*, 60(8):126.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3):272–299.

Genton, M. G. (2000). Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312.

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis*, 83(2):493–508.

Golub, G. and Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, 2(2):205–224.

Grotjahn, R., Black, R., Leung, R., Wehner, M. F., Barlow, M., Bosilovich, M., et al. (2016). North American extreme temperature events and related large scale meteorological patterns: A review of statistical methods, dynamics, modeling, and trends. *Climate Dynamics*, 46(3-4):1151–1184.

Hanks, E. M., Schliep, E. M., Hooten, M. B., and Hoeting, J. A. (2015). Restricted spatial regression in practice: Geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, 26(4):243–254.

Hannachi, A., Jolliffe, I. T., and Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology*, 27(9):1119–1152.

Harman, H. H. and Harman, H. H. (1976). *Modern factor analysis.* University of Chicago Press.

Hastie, T. and Tibshirani, R. (2017). *Generalized Additive Models.* Routledge.

Hefley, T. J., Hooten, M. B., Hanks, E. M., Russell, R. E., and Walsh, D. P. (2017). The Bayesian group lasso for confounded spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 22(1):42–59.

Hernandez-Stumpfhauser, D., Breidt, F. J., and van der Woerd, M. J. (2017). The general projected normal distribution of arbitrary dimension: Modeling and Bayesian inference. *Bayesian Analysis*, 12(1):113–133.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334.

Hoff, P. D. (2007). Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association*, 102(478):674–685.

Hoff, P. D. (2009). Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.

Huang, A. and Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8(2):439–452.

Jackson, D. A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology*, 74(8):2204–2214.

Jauch, M., Hoff, P. D., and Dunson, D. B. (2021). Monte Carlo simulation on the Stiefel manifold via polar expansion. *Journal of Computational and Graphical Statistics*, 30(3):622–631.

Jolliffe, I., Uddin, M., and Vines, S. (2002). Simplified EOFs-three alternatives to rotation. *Climate Research*, 20(3):271–279.

Jolliffe, I. T. (2002). *Principal Component Analysis.* Springer Series in Statistics. Springer-Verlag, New York.

Kambhatla, N. and Leen, T. K. (1997). Dimension reduction by local principal component analysis. *Neural Computation*, 9(7):1493–1516.

Lorenz, E. N. (1956). *Empirical orthogonal functions and statistical weather prediction*, volume 1. Massachusetts Institute of Technology, Department of Meteorology Cambridge.

Mantua, N. J. and Hare, S. R. (2002). The Pacific decadal oscillation.

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics.* Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Melkumyan, A. and Ramos, F. (2009). A sparse covariance function for exact gaussian process inference in large datasets. *IJCAI International Joint Conference on Artificial Intelligence*, pages 1936–1942.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023.

Mulaik, S. A. (2009). *Foundations of Factor Analysis*. CRC press.

North, G. R., Bell, T. L., Cahalan, R. F., and Moeng, F. J. (1982). Sampling errors in the estimation of empirical orthogonal functions. *Monthly Weather Review*, 110(7):699–706.

Nychka, D., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2(4):315–331.

O'Brien, J. P. and Deser, C. (2023). Quantifying and understanding forced changes to unforced modes of atmospheric circulation variability over the north Pacific in a coupled model large ensemble. *Journal of Climate*, 36(1):19–37.

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science*, 25(1):107–125.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Peres-Neto, P. R., Jackson, D. A., and Somers, K. M. (2003). Giving meaningful interpretation to ordination axes: Assessing loading significance in principal component analysis. *Ecology*, 84(9):2347–2363.

Pourzanjani, A. A., Jiang, R. M., Mitchell, B., Atzberger, P. J., and Petzold, L. R. (2021). Bayesian inference over the Stiefel manifold via the Givens representation. *Bayesian Analysis*, 16(2):639–666.

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4):1197–1206.

Roden, R., Smith, T., and Sacrey, D. (2015). Geologic pattern recognition from seismic attributes: Principal component analysis and self-organizing maps. *Interpretation*, 3(4):SAE59–SAE83.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034.

Smith, S. M., Hyvärinen, A., Varoquaux, G., Miller, K. L., and Beckmann, C. F. (2014). Group-PCA for very large fMRI datasets. *NeuroImage*, 101:738–749.

Stewart, G. W. (1993). On the early history of the singular value decomposition. *SIAM Review*, 35(4):551–566.

Stroud, J. R., Müller, P., and Sansö, B. (2001). Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 63(4):673–689.

Thompson, D. W. J. and Wallace, J. M. (2000). Annular modes in the extratropical circulation. Part I: Month-to-month variability*. *Journal of Climate*, 13(5):1000–1016.

Wang, F. and Gelfand, A. E. (2013). Directional data analysis under the general projected normal distribution. *Statistical Methodology*, 10(1):113–127.

Wang, F. and Gelfand, A. E. (2014). Modeling space and space-time directional data using projected Gaussian processes. *Journal of the American Statistical Association*, 109(508):1565–1580.

Wang, W.-T. and Huang, H.-C. (2017). Regularized principal component analysis for spatial data. *Journal of Computational and Graphical Statistics*, 26(1):14–25.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, 4(1):389–396.

Wikle, C. K. (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, 84(6):1382–1394.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20(4):397–405.

Zimmerman, D. L. and Ver Hoef, J. M. (2022). On deconfounding spatial confounding in linear models. *The American Statistician*, 76(2):159–167.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

A Appendix

A Proofs of propositions

We now prove the propositions describing the properties of the orthogonal matrix constructed in section 2.1.

Lemma 1. *The generating random variables $\mathbf{z}_j$ and $\mathbf{\Omega}_j$ are exchangeable.*

Proof. The generating random variables $\mathbf{z}_i$ are exchangeable because they all independent and have the same marginal distribution. Specifically, because $\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_k \sim \pi_\Omega$ all have the same distribution, if we marginalize $\mathbf{z}_j$, we get $p(\mathbf{z}_j) = \int_\Omega p(\mathbf{z}_j|\mathbf{\Omega}_j)p(\mathbf{\Omega}_j)d\mathbf{\Omega}$ is the same for all $j = 1, \ldots, k$. □

Lemma 2. *For any permutation π of the columns of the $n \times k$ matrix $\mathbf{X}$, denoted $\mathbf{X}_\pi$, the matrix $\mathbf{P}_\pi \equiv \mathbf{I} - \mathbf{X}_\pi(\mathbf{X}_\pi'\mathbf{X}_\pi)^{-1}\mathbf{X}_\pi'$ is the unique projection onto the orthogonal complement of column space of $\mathbf{X}$. That is, $\mathbf{P}_\pi = \mathbf{P}$.*

Proof. Since $\mathbf{X}$ and $\mathbf{X}_\pi$ share the same column space, the result is immediate by the projection theorem. □

Proposition 1. *The columns of $\mathbf{W} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}$ are exchangeable. That is, for any permutation π of the set $\{1, \ldots, k\}$, $p([\mathbf{w}_1, \ldots, \mathbf{w}_k]) \overset{d}{=} p([\mathbf{w}_{\pi(1)}, \ldots, \mathbf{w}_{\pi(k)}])$.*

Proof. We first show the columns of the matrix $\mathbf{X}$ are exchangeable. That is, for any permutation π of the set $\{1, \ldots, k\}$, $p([\mathbf{x}_1, \ldots, \mathbf{x}_k]) \overset{d}{=} p([\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(k)}])$. Then, we use the exchangeability of $\mathbf{X}$ to show exchangeability of $\mathbf{W}$.

Define $\mathbf{X}_{\pi_j} = [\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(j)}]$ and $\mathbf{P}_{\pi(j)} = \mathbf{I} - \mathbf{X}_{\pi_j}(\mathbf{X}_{\pi_j}'\mathbf{X}_{\pi_j})^{-1}\mathbf{X}_{\pi_j}' = \mathbf{P}_j$. To show exchangeablility, we show the characteristic function of $\mathbf{X}$ is equivalent to the characteristic function of $\mathbf{X}_{\pi_j}$. For a $n \times k$ random matrix $\mathbf{X}$, the characteristic function is defined as $\varphi(\mathbf{X}) = E[\exp\{i\text{tr}(\mathbf{T}'\mathbf{X})\}] = E[\exp\{i\sum_{\ell=1}^k \mathbf{t}_\ell'\mathbf{x}_\ell\}]$, where $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_k]$ is a $n \times k$ matrix, i is the imaginary unit, and $\text{tr}(\cdot)$ is the trace operator. We show the proposition using proof by induction:

1. For $k = 1$, we have $\mathbf{X}_{\pi_1} = \mathbf{x}_{\pi(1)} = \mathbf{P}_0\mathbf{z}_{\pi(1)} \overset{d}{=} \mathbf{P}_0\mathbf{z}_1 = \mathbf{x}_1 = \mathbf{X}_1$, where $\mathbf{z}_{\pi(1)} \overset{d}{=} \mathbf{z}_1$ by lemma 1. Therefore, $\mathbf{X}_1 \overset{d}{=} \mathbf{X}_{\pi_1}$.
2. Assume for $k = j$, $\mathbf{X}_j \overset{d}{=} \mathbf{X}_{\pi_j}$.
3. By the characteristic function of $\mathbf{X}_{\pi_{j+1}}$,

$$\begin{aligned}
\varphi(\mathbf{X}_{\pi_{j+1}}) &= E\left[\exp\left\{i\sum_{\ell=1}^{j+1}\mathbf{t}_\ell'\mathbf{x}_{\pi(\ell)}\right\}\right] \\
&= E\left[\exp\left\{i\sum_{\ell=1}^{j}\mathbf{t}_\ell'\mathbf{x}_{\pi(\ell)}\right\}\exp\left\{i\mathbf{t}_{j+1}'\mathbf{x}_{\pi(j+1)}\right\}\right] \\
&= E\left[\exp\left\{i\sum_{\ell=1}^{j}\mathbf{t}_\ell'\mathbf{x}_{\pi(\ell)}\right\}E[\exp\left\{i\mathbf{t}_{j+1}'\mathbf{x}_{\pi(j+1)}\right\}|\mathbf{X}_{\pi_j}, \mathbf{\Omega}_{\pi(j+1)}]\right] \quad \text{(iterative expectation)} \\
&= E\left[\exp\left\{i\sum_{\ell=1}^{j}\mathbf{t}_\ell'\mathbf{x}_{\pi(\ell)}\right\}E[\exp\left\{i\mathbf{t}_{j+1}'\mathbf{P}_{\pi(j)}\mathbf{z}_{\pi(j+1)}\right\}|\mathbf{X}_{\pi_j}, \mathbf{\Omega}_{\pi(j+1)}]\right] \\
&= E\left[\exp\left\{i\sum_{\ell=1}^{j}\mathbf{t}_\ell'\mathbf{x}_{\pi(\ell)}\right\}\exp\left\{\mathbf{t}_{j+1}'\mathbf{P}_{\pi(j)}\mathbf{\Omega}_{\pi(j+1)}\mathbf{P}_{\pi(j)}'\mathbf{t}_{j+1}\right\}\right].
\end{aligned}$$

The induction hypothesis implies $\{\mathbf{X}_j, \mathbf{P}_j\} \overset{d}{=} \{\mathbf{X}_{\pi_j}, \mathbf{P}_{\pi(j)}\}$. Also, $\mathbf{\Omega}_{\pi(j+1)}$ is independent of $\mathbf{X}_j$ and $\mathbf{P}_j$. Therefore, $\{\mathbf{X}_j, \mathbf{P}_j, \mathbf{\Omega}_{j+1}\} \overset{d}{=} \{\mathbf{X}_{\pi_j}, \mathbf{P}_{\pi(j)}, \mathbf{\Omega}_{\pi(j+1)}\}$ because $\mathbf{X}_{\pi_j} \Rightarrow \mathbf{X}_j$ by induction hypothesis, $\mathbf{P}_{\pi(j)} \equiv \mathbf{P}_j$ by lemma 2, and $\mathbf{\Omega}_{\pi(j+1)} \Rightarrow \mathbf{\Omega}_{j+1}$ because it is independent of $\mathbf{X}_j$ and $\mathbf{P}_j$ and it is exchangeable. Thus,

$$\varphi(\mathbf{X}_{\pi_{j+1}}) = E\left[\exp\left\{i\sum_{\ell=1}^{j}\mathbf{t}'_\ell\mathbf{x}_\ell\right\}\exp\left\{\mathbf{t}'_{j+1}\mathbf{P}_j\mathbf{\Omega}_{j+1}\mathbf{P}'_j\mathbf{t}_{j+1}\right\}\right] = \varphi(\mathbf{X}_{j+1}),$$

and $\mathbf{X}_{j+1} \overset{d}{=} \mathbf{X}_{\pi_{j+1}}$.

The exchangeability of $\mathbf{W}$ follows from the exchangeability of $\mathbf{X}$. Specifically, because the diagonal matrix $\mathbf{R} \equiv (\mathbf{X}'\mathbf{X})^{-1/2} = \text{diag}[(\mathbf{x}'_1\mathbf{x}_1)^{-1/2}, \ldots, (\mathbf{x}'_k\mathbf{x}_k)^{-1/2}] \equiv \text{diag}[r_1, \ldots, r_k]$ where the elements $r_1, \ldots, r_k$ are the norm of the random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$, respectively, is simply a rescaling of the columns of $\mathbf{X}$, and the permutation of the scaling is preserved, $\mathbf{W}_\pi \overset{d}{=} \mathbf{W}$.

□

Proposition 2. $\mathbf{w}_i|\mathbf{W}_{i-1} \overset{d}{=} \mathbf{N}_{i-1}\widetilde{\mathbf{w}}_i|\mathbf{W}_{i-1}$ *where the columns of* $\mathbf{N}_{i-1}$ *form an orthonormal basis for the null space of* $\mathbf{W}_{i-1}$ *and* $\widetilde{\mathbf{w}}_i|\mathbf{W}_{i-1} \sim PN_{n-i+1}(\mathbf{0}, \mathbf{N}'_{i-1}\mathbf{\Omega}_i\mathbf{N}_{i-1})$ *is the projected weight function.*

Proof. The following argument is similar to Hoff (2007), except now we account for dependence structure and the resulting distribution is different. By construction, $\mathbf{w}_i = \mathbf{P}_{i-1}\mathbf{z}_i/(\mathbf{z}'_i\mathbf{P}'_{i-1}\mathbf{P}_{i-1}\mathbf{z}_i)^{1/2}$ where $\mathbf{P}_{i-1}$ has $n-i+1$ eigenvalues equal to 1 and the rest being 0. Let the eigenvalue decomposition be $\mathbf{P}_{i-1} = \mathbf{N}_{i-1}\mathbf{N}'_{i-1}$ where $\mathbf{N}_{i-1}$ is an $n \times (n-i+1)$ matrix whose columns span the null space of $\mathbf{W}_i$. Making the substitution $\mathbf{P}_{i-1} = \mathbf{N}_{i-1}\mathbf{N}'_{i-1}$,

$$\begin{aligned}\mathbf{w}_i &= \frac{\mathbf{P}_{i-1}\mathbf{z}_i}{(\mathbf{z}'_i\mathbf{P}'_{i-1}\mathbf{P}_{i-1}\mathbf{z}_i)^{1/2}} \\ &= \frac{\mathbf{N}_{i-1}\mathbf{N}'_{i-1}\mathbf{z}_i}{(\mathbf{z}'_i\mathbf{N}'_{i-1}\mathbf{N}_{i-1}\mathbf{N}_{i-1}\mathbf{N}'_{i-1}\mathbf{z}_i)^{1/2}} \\ &= \mathbf{N}_{i-1}\frac{\mathbf{N}'_{i-1}\mathbf{z}_i}{(\mathbf{z}'_i\mathbf{N}_{i-1}\mathbf{N}'_{i-1}\mathbf{z}_i)^{1/2}}.\end{aligned}$$

Note that $\mathbf{P}_{i-1} = \mathbf{I} - \mathbf{W}_{i-1}\mathbf{W}'_{i-1}$, so $\mathbf{w}_i|\mathbf{W}_{i-1} \overset{d}{=} \mathbf{N}_{i-1}\frac{\mathbf{N}'_{i-1}\mathbf{z}_i}{(\mathbf{z}'_i\mathbf{N}_{i-1}\mathbf{N}'_{i-1}\mathbf{z}_i)^{1/2}}$. Because $\mathbf{z}_i \sim N_n(\mathbf{0}, \mathbf{\Omega}_i)$, we have $\mathbf{N}'_{i-1}\mathbf{z}_i|\mathbf{W}_{i-1} \sim N_n(\mathbf{0}, \mathbf{N}'_{i-1}\mathbf{\Omega}_i\mathbf{N}_{i-1})$ and $\frac{\mathbf{N}'_{i-1}\mathbf{z}_i}{(\mathbf{z}'_i\mathbf{N}_{i-1}\mathbf{N}'_{i-1}\mathbf{z}_i)^{1/2}} \equiv \widetilde{\mathbf{w}}_i|\mathbf{W}_{i-1} \sim PN(\mathbf{0}, \mathbf{N}'_{i-1}\mathbf{\Omega}_i\mathbf{N}_{i-1})$. □

S Supplemental Material

S.1 Full conditional distributions

The diagonal matrix $\mathbf{D}$ and the length-scale parameters $\rho_{u,i}$ and $\rho_{v,i}$ do not have conjugate updates and so we use a Metropolis-within-Gibbs step to estimate these parameters. For all Metropolis steps, we use a truncated normal for the proposal distribution with the mean set to the most recently accepted value. For $\mathbf{D}$ the upper truncation bound is set to infinity and for $\rho_{u,i}$ and $\rho_{v,i}$ the upper truncation bound is set to the max distance for $\mathbf{U}$ (e.g., greatest distance between spatial locations) and $\mathbf{V}$ (e.g., greatest span between time points) divided by 2, respectively. Because the variance of the proposal can influence the acceptance rate, we automatically tune the proposal variance for each parameter individually such that the acceptance rate is between 25% and 45%.

Sampling Algorithm

For each iteration of the MCMC algorithm, do:

1. Update $\mathbf{D}$ using a Metropolis step
2. For $i \in \{1, \dots, k\}$ update $\mathbf{u}_i|\mathbf{U}_{-i} \overset{d}{=} \mathbf{N}_i^u \widetilde{\mathbf{u}}_i$ where $\widetilde{\mathbf{u}}_i$
$$\begin{aligned}[\widetilde{\mathbf{u}}_i|\cdot] &\sim N(\mathbf{S}_u^{-1}\mathbf{m}_u, \mathbf{S}_u^{-1})\mathbb{I}(\widetilde{\mathbf{u}}_i \in \mathcal{V}_{1,n})\\ \mathbf{m}_u &= d_i \mathbf{N}_i^u{}' \mathbf{\Sigma}^{-1}\mathbf{E}_i\mathbf{v}_i\\ \mathbf{S}_u &= d_i^2(\mathbf{N}_i^u{}'\mathbf{\Omega}_i^u\mathbf{N}_i^u)^{-1} + d_i^2\mathbf{N}_i^u{}'\mathbf{\Sigma}^{-1}\mathbf{N}_i^u.\end{aligned}$$
3. For $i \in \{1, \dots, k\}$ update $\mathbf{v}_i|\mathbf{V}_{-i} \overset{d}{=} \mathbf{N}_i^v \widetilde{\mathbf{v}}_i$ where $\widetilde{\mathbf{v}}_i$
$$\begin{aligned}[\widetilde{\mathbf{v}}_i|\cdot] &\sim N(\mathbf{S}_v^{-1}\mathbf{m}_v, \mathbf{S}_v^{-1})\mathbb{I}(\widetilde{\mathbf{v}}_i \in \mathcal{V}_{1,m})\\ \mathbf{m}_v &= d_i \mathbf{N}_i^v{}' \mathbf{E}_i'\mathbf{\Sigma}^{-1}\mathbf{u}_i\\ \mathbf{S}_v &= d_i^2(\mathbf{N}_i^v{}'\mathbf{\Omega}_i^v\mathbf{N}_i^v)^{-1} + d_i^2\mathbf{u}_i'\mathbf{\Sigma}^{-1}\mathbf{u}_i\mathbf{I}_m.\end{aligned}$$
4. Recall, we parameterize $\mathbf{\Sigma} = \sigma^2\mathbf{I}_n$. The full conditional distribution for σ^2 is
$$\begin{aligned}[a|\cdot] &\sim IG((\xi+1)/2, (1/A^2) + \xi/\sigma)\\ [\sigma^2|\cdot] &\sim IG((nm+\xi)/2, \xi/a + \text{vec}(\mathbf{Z} - \mathbf{UDV}')'\text{vec}(\mathbf{Z} - \mathbf{UDV}')/2).\end{aligned}$$
We specify $\xi = 1$ and $A = 10^5$ which corresponds to the prior $\sigma \sim \text{Half-}t(\xi, A) \equiv \text{Half-cauchy}(A)$.
5. Recall, we parameterize $\mathbf{\Omega}_u(\theta_{u,i}) = \sigma^2_{u,i}\mathbf{C}_u(\theta_{u,i})$. For $i \in \{1, \dots, k\}$ update $\sigma^2_{u,i}$ from
$$\begin{aligned}[a_{u,i}|\cdot] &\sim IG((\xi+1)/2, (1/A^2) + \xi/\sigma_{u,i})\\ [\sigma^2_{u,i}|\cdot] &\sim IG((n-k+1+\xi)/2, \xi/a_{u,i} + (d_i^2\widetilde{\mathbf{u}}_i'(\mathbf{N}_i^u{}'\mathbf{\Omega}_i^u\mathbf{N}_i^u)^{-1}\widetilde{\mathbf{u}}_i)/2),\end{aligned}$$
with $\xi = 1$ and $A = 10^5$.
6. Recall, we parameterize $\mathbf{\Omega}_v(\theta_{v,i}) = \sigma^2_{v,i}\mathbf{C}_v(\theta_{v,i})$. For $i \in \{1, \dots, k\}$ update $\sigma^2_{v,i}$ from
$$\begin{aligned}[a_{v,i}|\cdot] &\sim IG((\xi+1)/2, (1/A^2) + \xi/\sigma_{v,i})\\ [\sigma^2_{v,i}|\cdot] &\sim IG((m-k+1+\xi)/2, \xi/a_{v,i} + (d_i^2\widetilde{\mathbf{v}}_i'(\mathbf{N}_i^v{}'\mathbf{\Omega}_i^v\mathbf{N}_i^v)^{-1}\widetilde{\mathbf{v}}_i)/2),\end{aligned}$$
with $\xi = 1$ and $A = 10^5$.
7. For $i \in \{1, \dots, k\}$ update $\rho^2_{u,i}$ using a Metropolis step
8. For $i \in \{1, \dots, k\}$ update $\rho^2_{v,i}$ using a Metropolis step

S.2 Identity correlation

When $\mathbf{\Omega}_i = \mathbf{I}_n$, $\widetilde{\mathbf{w}}_k$ in proposition 2 is uniformly distributed on the Stiefel manifold. To see this, note that for $\mathbf{z}_k \sim N(0, \mathbf{I})$, $\mathbf{N}_{k-1}'\mathbf{z}_k \sim \text{N}_{n-k+1}(0, \mathbf{I})$, and $\frac{\mathbf{N}_{k-1}'\mathbf{z}_k}{(\mathbf{z}_i'\mathbf{N}_{i-1}\mathbf{N}_{i-1}'\mathbf{z}_i)^{1/2}}$ is uniformly distributed on the $n-k+1$ sphere. Also, we see proposition 1 is now equivalent to Hoff (2007), and $\mathbf{W}$ is the uniform probability measure on $\mathcal{V}_{k,n}$.

The resulting full conditional distributions for $\widetilde{\mathbf{u}}_i$ and $\widetilde{\mathbf{v}}_i$ when $\mathbf{\Omega}_i^u = \mathbf{I}_n$ and $\mathbf{\Omega}_i^v = \mathbf{I}_m$ for the SVD model in S.1 become the von-Mises Fisher distribution, which is equivalent to the full conditionals of Hoff (2007). To see this, note the mean of $[\widetilde{\mathbf{u}}_i|\cdot]$ is $\mathbf{S}_u^{-1}\mathbf{m}_u = \frac{1}{\sigma^2+1}\frac{1}{d_i}\mathbf{N}_i^{u}{'}\mathbf{E}_i\mathbf{v}_i$ and the covariance is $\mathbf{S}_u^{-1} = \left(d_i^2 + \frac{d_i^2}{\sigma^2}\right)^{-1}$. The full conditional distribution

$$
\begin{aligned}
[\widetilde{\mathbf{u}}_i|\cdot] &\propto \exp\left\{-\frac{1}{2}tr\left[-2\widetilde{\mathbf{u}}_i' d_i \mathbf{N}_i^{u}{'}\mathbf{E}_i\mathbf{v}_i\left(\frac{1}{\sigma^2+1}+\frac{1}{\sigma^4+\sigma^2}\right)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}tr\left[-2\widetilde{\mathbf{u}}_i' d_i \mathbf{N}_i^{u}{'}\mathbf{E}_i\mathbf{v}_i\left(\frac{1}{\sigma^2}\right)\right]\right\},
\end{aligned}
$$

which is the kernel of the von-Mises Fisher distribution. The same result holds for $\widetilde{\mathbf{v}}_i$.

S.2.1 Relationship to algorithmic SVD

Computing the SVD of $\mathbf{Y}$,

$$
\begin{aligned}
\mathbf{Y} &= \mathbf{U}\mathbf{D}\mathbf{V}' \\
\mathbf{Y} &= \mathbf{U}_{-i}\mathbf{D}_{-i}\mathbf{V}_{-i}' + d_i\mathbf{u}_i\mathbf{v}_i' \\
\mathbf{Y} - \mathbf{U}_{-i}\mathbf{D}_{-i}\mathbf{V}_{-i}' = \mathbf{E}_{-i} &= d_i\mathbf{u}_i\mathbf{v}_i' \\
\frac{1}{d_i}\mathbf{E}_{-i}\mathbf{v}_i &= \mathbf{u}_i,
\end{aligned}
$$

so the ith basis function can be expressed as a function of the data and other basis functions. The mean of the full conditional distribution $[\widetilde{\mathbf{u}}_i|\cdot]$ is $\mathbf{S}_u^{-1}\mathbf{m}_u = \frac{1}{\sigma^2+1}\frac{1}{d_i}\mathbf{N}_i^{u}{'}\mathbf{E}_i\mathbf{v}_i$, and $E[\widetilde{\mathbf{u}}_i|\cdot] \to \frac{1}{d_i}\mathbf{N}_i^{u}{'}\mathbf{E}_i\mathbf{v}_i$ as $\sigma^2 \to 0$. Mapping to the original space, $\mathbf{N}_i^u E[\widetilde{\mathbf{u}}_i|\cdot] = \frac{1}{d_i}\mathbf{N}_i^u\mathbf{N}_i^{u}{'}\mathbf{E}_i\mathbf{v}_i = \frac{1}{d_i}\mathbf{E}_i\mathbf{v}_i$. While not shown here, the same argument applies for $\mathbf{V}$. Therefore, when the covariance is taken to be the identity matrix, the posterior mean of the basis functions is equivalent to the C-SVD basis functions.

To see the relationship, we repeat one of the simulation conducted in Section 4.3 with $SNR = 5$, $k = 5$, and set the correlation matrices $\mathbf{C}_u$ and $\mathbf{C}_v$ to be the identity. Here, we still estimate the basis function specific variance $\sigma_{u,i}^2$ and $\sigma_{v,i}^2$. We obtain 10000 samples from the posterior, discarding the first 5000 as burnin. The resulting estimates for the $\mathbf{U}$ and $\mathbf{V}$ basis functions are shown in Figure S.1, where the posterior mean of the basis functions (blue) is nearly identical to the C-SVD estimates (red). In all cases, we see the 95% intervals (blue shaded region) cover the C-SVD estimates but has ≈95% coverage of the true line (black).

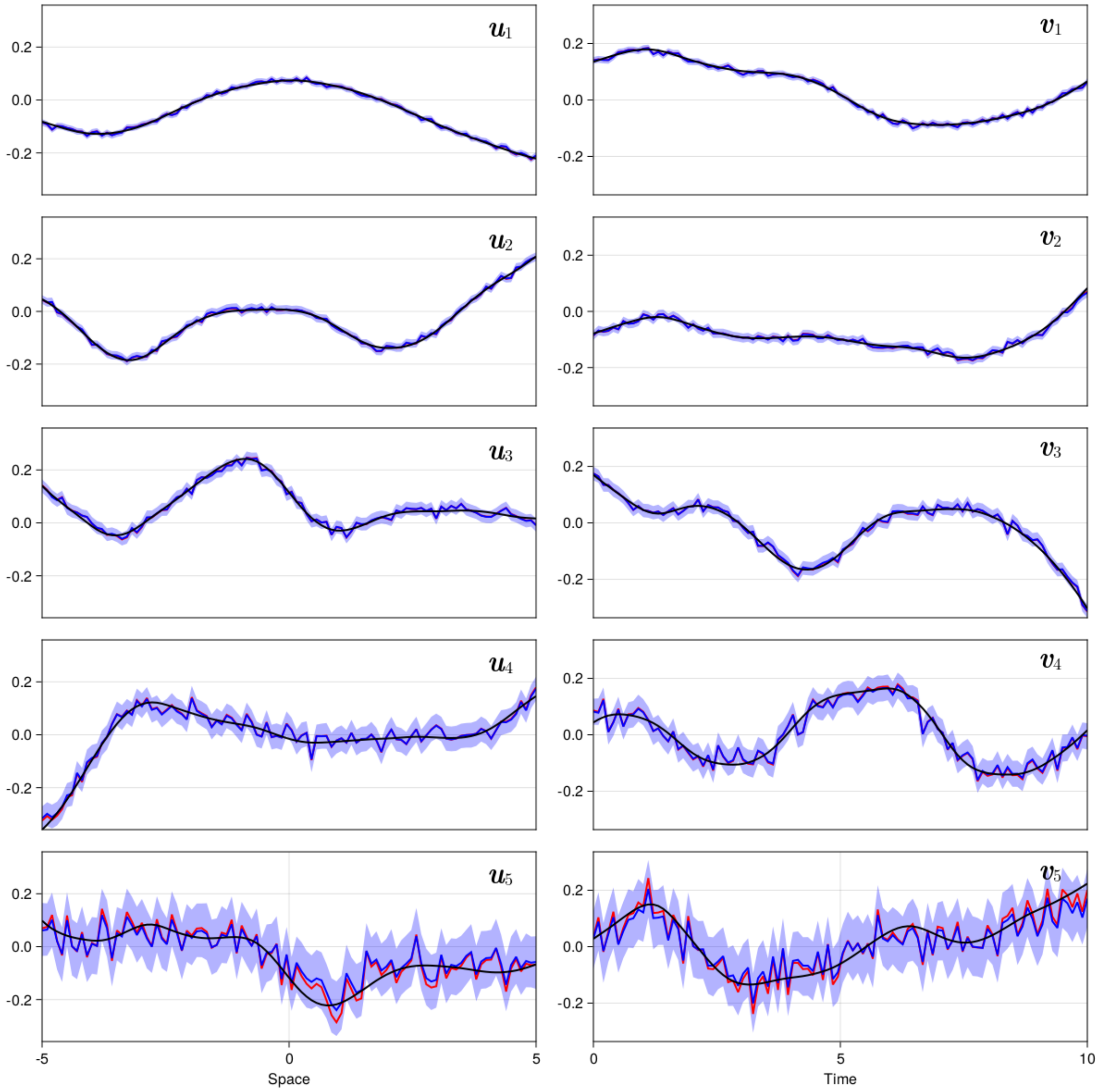


Figure S.1: Posterior mean (blue line), 95% credible intervals (shaded blue region), truth (black line), and C-SVD estimate (red line) for the **U** and **V** basis function.

S.3 Computation and scalability

While the proposed prior is relatively simple to specify and implement, there are some computational aspects to consider. On one hand, the fact that the prior is conjugate with a normal data distribution means that MCMC updates for the columns of $\mathbf{U}$ and $\mathbf{V}$ can be obtained in a straightforward manner. On the other hand, calculating the full conditional distributions (from which the Gibbs draws are sampled) is computationally intensive for large n. From the formulation in Section 3.1, the full conditional distributions for the columns of $\mathbf{U}$ and $\mathbf{V}$ involve matrix inverses $(\mathbf{N}_i^u{}'\mathbf{\Omega}_i^u\mathbf{N}_i^u)^{-1}$ and $(\mathbf{N}_i^v{}'\mathbf{\Omega}_i^v\mathbf{N}_i^v)^{-1}$, respectively (see supplement S.1), each of which are dense $(n-k+1)\times(n-k+1)$ and $(m-k+1)\times(m-k+1)$ matrices, respectively. Therefore, in order to update $\mathbf{U}$ and $\mathbf{V}$ once in an MCMC iteration, we need to calculate $2k$ matrix inverses (one for each of the k columns of $\mathbf{U}$ and $\mathbf{V}$), which is computationally challenging for large n or m. Furthermore, updating the hyperparameters of the kernel (e.g., the length-scale parameters $\rho_{u,i}$ and $\rho_{v,i}$) requires Metropolis-Hastings steps. In this case, the likelihood involves a multivariate Normal density: when the covariance of the multivariate Normal is non-diagonal and dense (as is the case here), the number of flops associated with evaluating the determinant and solving quadratic forms scales with $\mathcal{O}(n^3)$. Again, each iteration of the MCMC requires $2k$ of these calculations. As such, without significant computing resources, the required computation for the model as-is proves difficult for data where either n or m is greater than 1000. More specifically, Figure S.2 shows an estimate of the amount of time needed to update all parameters in a single iteration of the MCMC for the special case of $k = 5$ across different sample sizes n and m on a personal laptop.

In spite of these apparent limitations, there are a variety of approaches one could take to reduce the associated computational burdens of this model. The simplest approach would be to parameterize the covariance matrix as $\mathbf{\Omega}_i = \sigma_i^2\mathbf{C}_i(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is specified and not estimated: this would remove the Metropolis-Hastings steps required to estimate $\rho_{u,i}$ and $\rho_{v,i}$. Alternatively, one could use a compactly-supported kernel function (see, e.g., Wendland, 1995; Buhmann, 2000; Genton, 2000; Gneiting, 2002; Melkumyan and Ramos, 2009) and leverage sparse matrix techniques. These approaches are targeted at reducing the cost of estimating $\mathbf{C}_i(\boldsymbol{\theta}_i)$ and the associated parameters. In either case, however, our proposed prior does have the disadvantage of relying on a column-wise sampling strategy and the corresponding matrix calculations needed to sample each column. Specifically, within each MCMC iteration, there is a required $\mathcal{O}(n^3)$ cost of computing the orthonormal basis for the null-space $\mathbf{N}_i^u$ and $\mathbf{N}_i^v$ and the ensuing inverses. In other words, we must calculate the inverse of $\mathbf{N}_i^{(\cdot)}{}'\mathbf{\Omega}_i^{(\cdot)}\mathbf{N}_i^{(\cdot)}$ which is dense irrespective of the sparsity of $\mathbf{\Omega}_i^{(\cdot)}$. For this reason, implementing sparse matrix techniques for the $\mathbf{\Omega}_i$ will not solve this challenge.

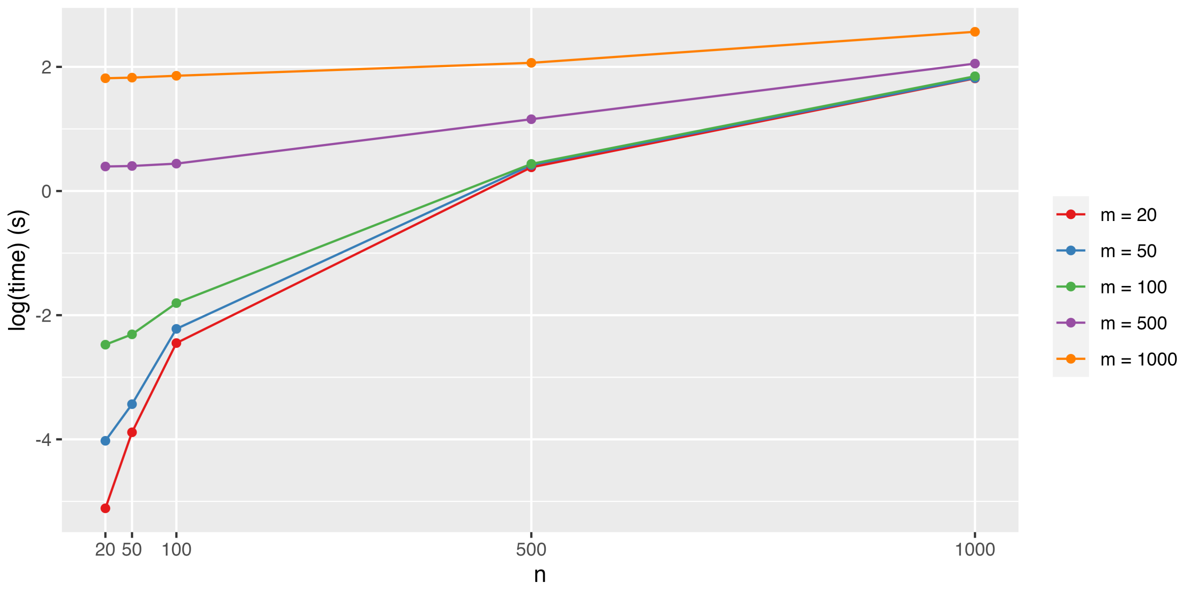


Figure S.2: Median log computation time for a single MCMC iteration as a function of sample size n (x-axis) and m (line color).

S.4 Projected normal distribution

Let $\mathbf{z}_j \sim N_n(\mathbf{0}, \boldsymbol{\Omega}), j = 1, \ldots, K$ and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_K]$. Then $\mathbf{W} = \mathbf{Z}/\|\mathbf{Z}\| \sim PN(\mathbf{0}, \boldsymbol{\Omega})$ where $\mathbf{w}_j$ is a length-n directional vector with $n-1$ angles $\boldsymbol{\theta}_j = [\theta_{1,j}, \theta_{2,j}, \ldots, \theta_{n-1,j}]$. Using spherical coordinates, $r_j = \|\mathbf{z}_j\| = \sqrt{z_{1,j}^2 + \cdots + z_{n,j}^2}$,

$$\begin{aligned} w_{1,j} &= \cos(\theta_{1,j}) \\ w_{2,j} &= \sin(\theta_{1,j})\cos(\theta_{2,j}) \\ &\vdots \\ w_{n-1,j} &= \sin(\theta_{1,j})\cdots\sin(\theta_{n-2,j})\cos(\theta_{n-1,j}) \\ w_{n,j} &= \sin(\theta_{1,j})\cdots\sin(\theta_{n-2,j})\sin(\theta_{n-1,j}) \end{aligned}$$

and

$$\begin{aligned} z_{1,j} &= r_j\cos(\theta_{1,j}) \\ z_{2,j} &= r_j\sin(\theta_{1,j})\cos(\theta_{2,j}) \\ &\vdots \\ z_{n-1,j} &= r_j\sin(\theta_{1,j})\cdots\sin(\theta_{n-2,j})\cos(\theta_{n-1,j}) \\ z_{n,j} &= r_j\sin(\theta_{1,j})\cdots\sin(\theta_{n-2,j})\sin(\theta_{n-1,j}) \end{aligned}$$

with $r_j \geq 0$, $\theta_{1,j}, \theta_{2,j}, \ldots, \theta_{n-2,j} \in [0, \pi]$, and $\theta_{n-1,j} \in [0, 2\pi]$. Augmenting the distribution with its latent length r_j, we get the joint density of $(r_j, \mathbf{w}_j)$ is

$$p(r_j, \mathbf{w}_j) = (2\pi)^{-n/2}|\boldsymbol{\Omega}|^{-1/2}\exp\left\{-\frac{1}{2}(r_j\mathbf{w}_j)'\boldsymbol{\Omega}^{-1}(r_j\mathbf{w}_j)\right\} r_i^{n-1}\mathbb{I}(\mathbf{w}_i \in \mathcal{V}_{1,n}),$$

where the area element on the unit sphere is $r_j^{n-1} sin^{n-2}(\theta_{1,j}) sin^{n-3}(\theta_{2,j}) \ldots sin(\theta_{n-2,j}) dr_j d\theta_{1,j} d\theta_{2,j} \ldots d\theta_{n-1,j}$. For properties of this distribution, see Hernandez-Stumpfhauser et al. (2017).

S.5 Additional simulation figures

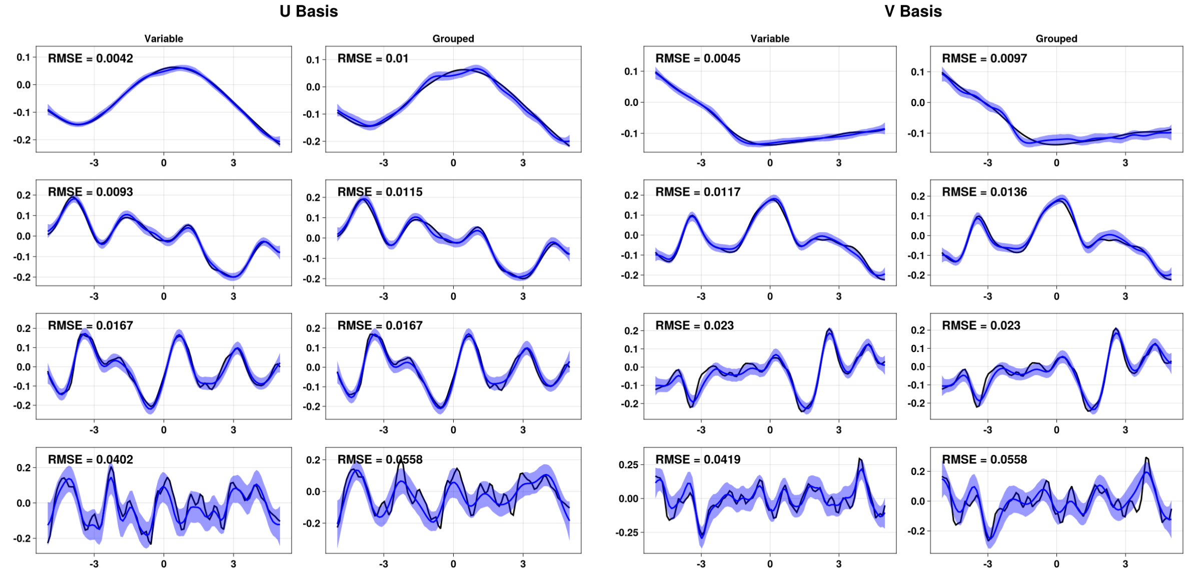


Figure S.3: Randomly chosen results from one simulation described in Section 4.2 with a SNR of 1. The $\mathbf{U}$ ($\mathbf{V}$) basis functions are shown on the left (right) and within the sub-plot the results from the variable (grouped) model are shown on the left (right) where the top row corresponds to the first basis function (e.g., $\mathbf{u}_1$ or $\mathbf{v}_1$) and the bottom row corresponds to the fourth basis function (e.g., $\mathbf{u}_4$ or $\mathbf{v}_4$). In each panel, the black line is the true basis function, blue line is the posterior mean, and blue shaded region denotes the 95% credible interval (CI).

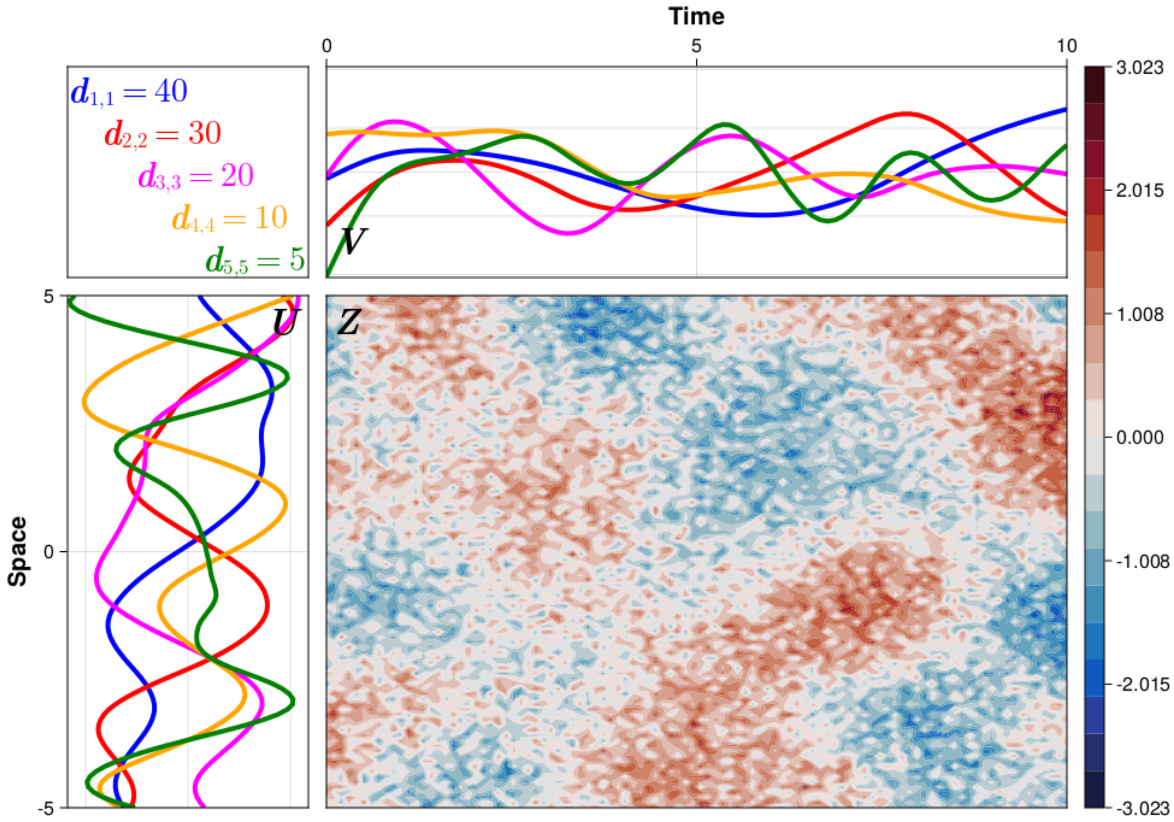


Figure S.4: Randomly simulated data $\mathbf{Z}$ (main plot) with a signal-to-noise ratio of 1 and the randomly simulated $\mathbf{U}$ (left) and $\mathbf{V}$ (top) basis functions.

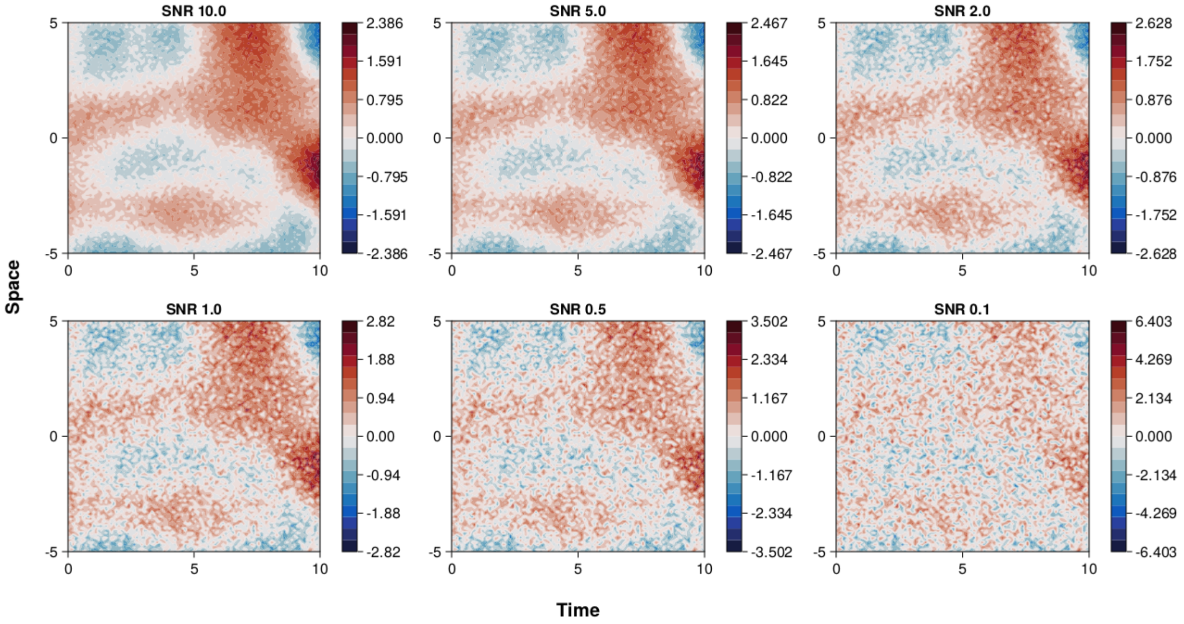


Figure S.5: Example of simulated data with varying levels of signal-to-noise ratio.

S.6 Surface air temperature

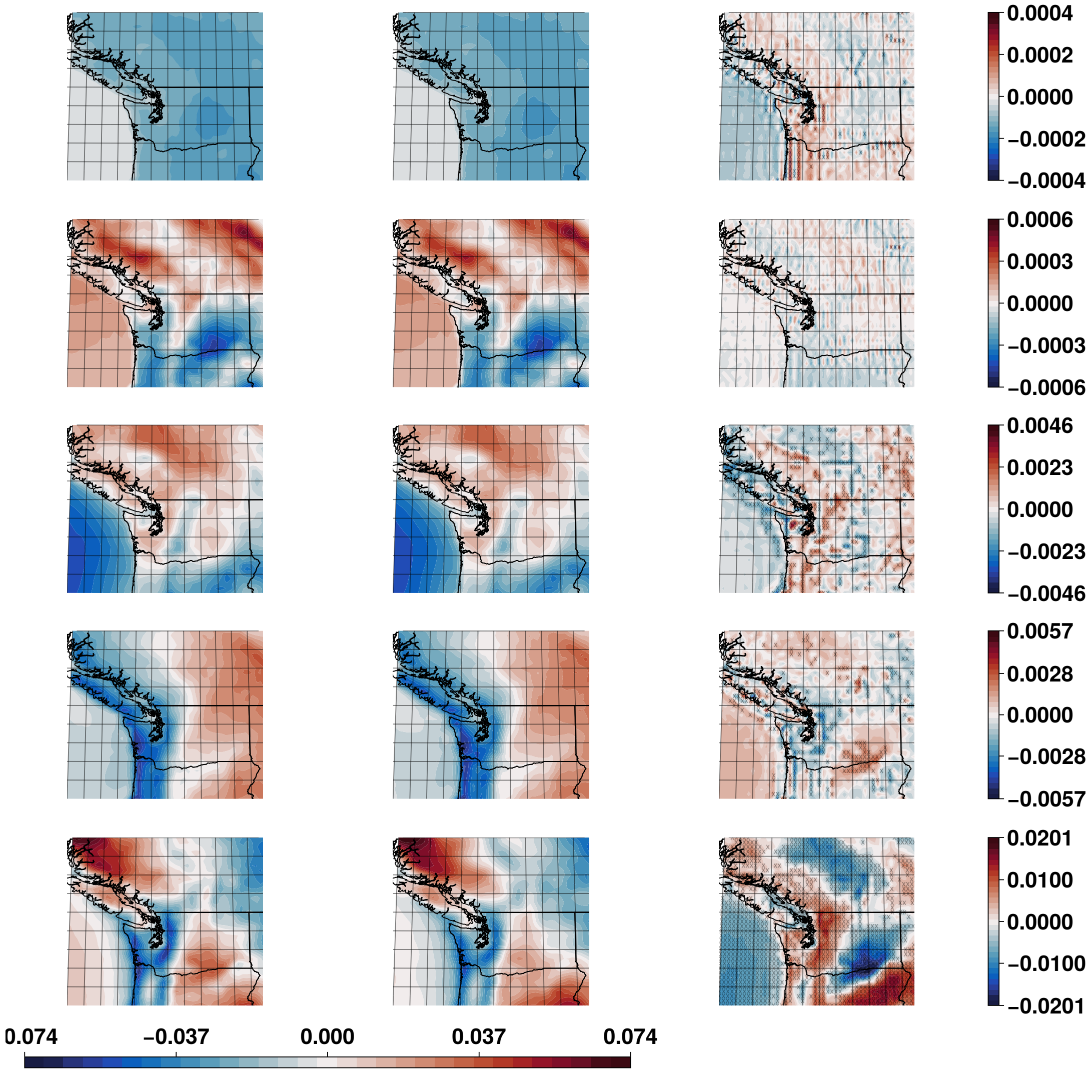


Figure S.6: Estimated spatial basis functions $\mathbf{u}_1$ (top) through $\mathbf{u}_5$ (bottom). The left column are the estimates from the deterministic SVD, the middle column are the posterior means, and the right column are the posterior difference between the posterior mean and the algorithmic estimate where locations whose 95% credible interval does not cover zero are denoted with an ‘x’.

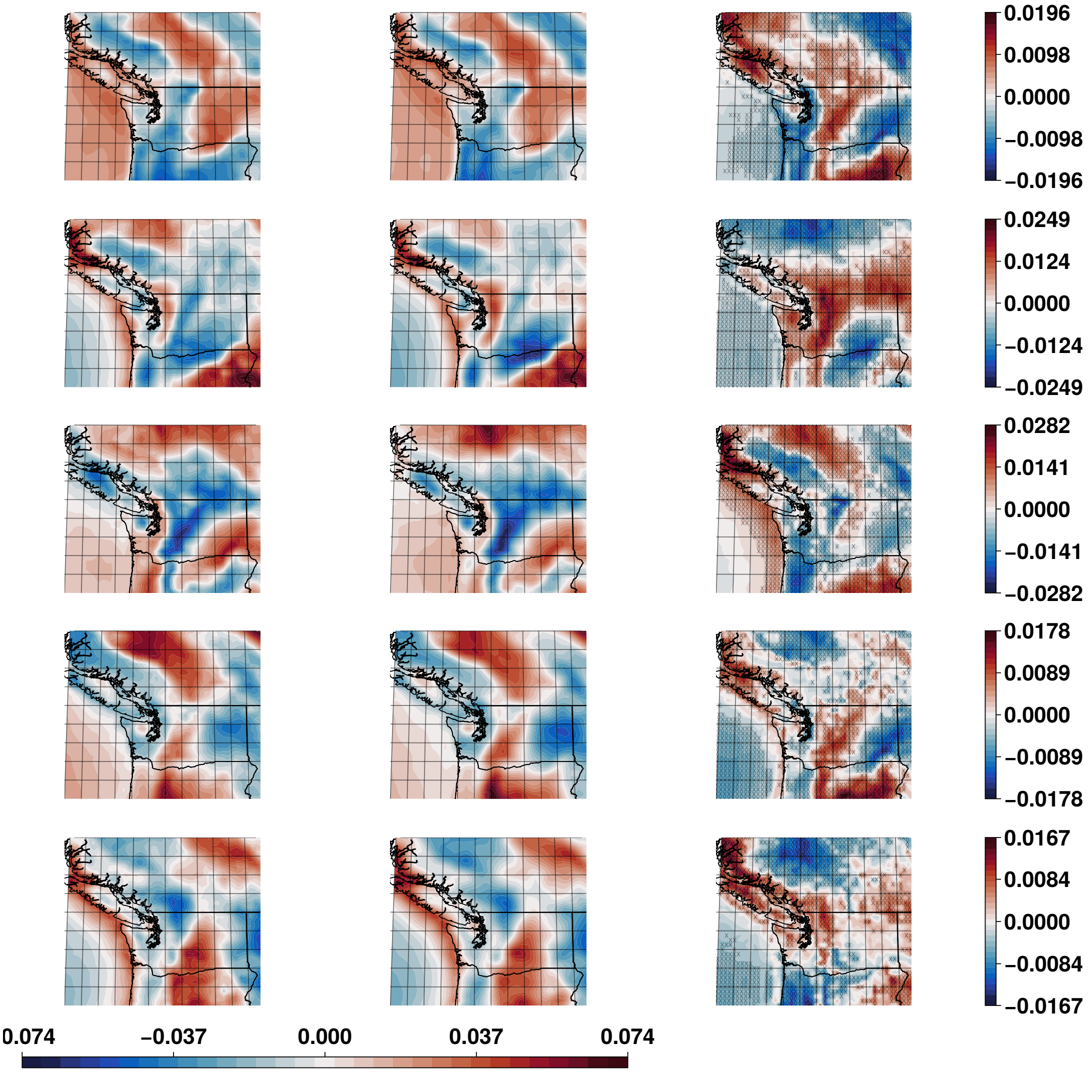


Figure S.7: Estimated spatial basis functions $\mathbf{u}_6$ (top) through $\mathbf{u}_{10}$ (bottom). The left column are the estimates from the deterministic SVD, the middle column are the posterior means, and the right column are the posterior difference between the posterior mean and the algorithmic estimate where locations whose 95% credible interval does not cover zero are denoted with an 'x'.

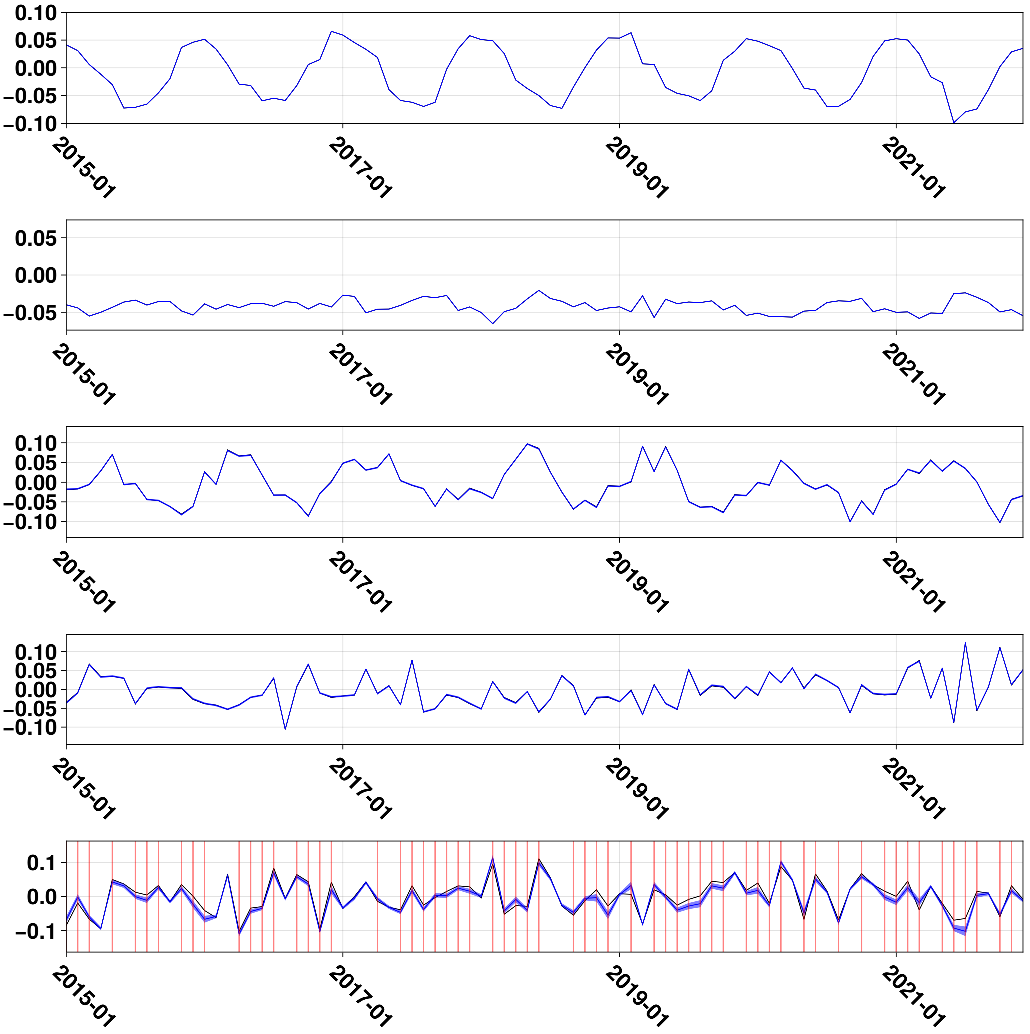


Figure S.8: Estimated temporal basis functions $\mathbf{v}_1$ (top) through $\mathbf{v}_5$ (bottom) from January 2010 to December 2021. The black line is the algorithmic estimate, the solid blue line is the posterior mean, and the blue shaded regions are the 95% credible intervals. Because it is difficult to see the differences, time points where the 95% credible interval does not cover the deterministic estimate are marked with a vertical line.

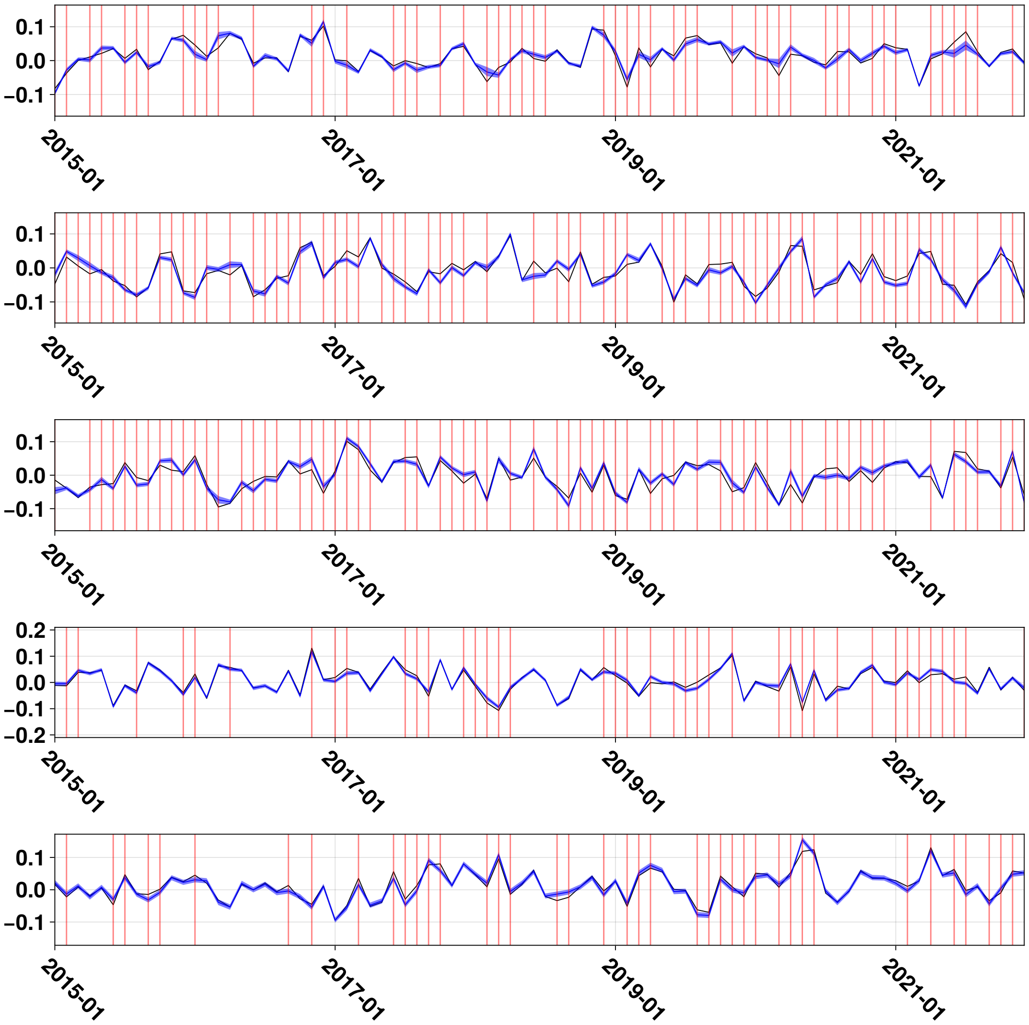


Figure S.9: Estimated temporal basis functions $\mathbf{v}_6$ (top) through $\mathbf{v}_{10}$ (bottom) from January 2010 to December 2021. The black line is the algorithmic estimate, the solid blue line is the posterior mean, and the blue shaded regions are the 95% credible intervals. Because it is difficult to see the differences, time points where the 95% credible interval does not cover the deterministic estimate are marked with a vertical line.