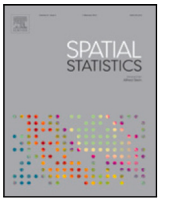
Spatial Statistics 74 (2026) 100988



Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta





# Computationally efficient approximations to conditional autoregressive prior for spatial and spatiotemporal datasets over large lattices

Md Kamrul Hasan Khan [a], Avishek Chakraborty [a,*], Giovanni Petris [a], Ghadeer J.M. Mahdi [b], Barry T. Wilson [c]

[a] Department of Mathematical Sciences, University of Arkansas, United States of America
[b] Mathematics Department, University of Baghdad, Iraq
[c] Northern Research Station, USDA Forest Service, United States of America

ARTICLE INFO

Dataset link: https://github.com/KamrulHKhan/Spectral-models-for-large-spatial-and-spatiotemporal-datasets, https://doi.org/10.7910/DVN/8FG2EV

*Keywords:*
Adjacency matrix
Areal data
Graph Laplacian
Lanczos algorithm
Low-rank approximations
Variance adjustment

ABSTRACT

Models for adjacency-based association represent an important branch of spatial statistics relevant for measurements associated to areal units or geographical networks. Within this domain, conditional autoregressive (CAR) prior represents a well-known tool for modeling dependence through random effects. However, the usual Markov chain Monte Carlo scheme for this framework requires sequential update of these effects from posterior full-conditional distributions resulting in poor convergence and mixing properties. In this article, we propose two alternative methods utilizing spectral properties of the Laplacian of the adjacency matrix. Our first proposal introduces a reduced-rank approximation to the CAR prior and the second proposal improves the quality of approximation by correcting for underestimation of variance parameter. These methods can be extended to any spatiotemporal setting as we have illustrated using a dynamic sequence of spatial random effects. All computational steps necessary to efficiently implement these methods, including determination of the eigen-structure of high-dimensional Laplacian matrices and derivation of posterior distributions, have been thoroughly described. Multiple simulation studies and two real-world datasets, representing diverse kinds of spatial and spatiotemporal response functions as well as adjacency structures on large number of units, are included to illustrate the utility of propose methods. Our findings adequately establish their benefits with respect to predictive performance and computational attributes.

## 1. Introduction

Spatial and spatiotemporal datasets, containing observations indexed by location or both location and time, are commonplace in real-world applications belonging to a diverse range of disciplines (Latimer et al., 2006; Paiva et al., 2014; Tran et al., 2015; Truong et al., 2016; Momenyan and Poorolajal, 2020). Such datasets can often be of massive size when measurements are available at a very fine spatial and/or temporal scale. Depending on how observations are indexed by locations, spatial association is modeled based on distance between locations for point-level datasets or adjacency between units for areal-level datasets. The former kind is most well-studied in literature with Gaussian process (GP) being the ubiquitous modeling tool. Building computationally efficient models

* Corresponding author.
*E-mail addresses:* mkk006@uark.edu (M.K.H. Khan), ac032@uark.edu (A. Chakraborty), gpetris@uark.edu (G. Petris), ghadeer.j@sc.uobaghdad.edu.iq (G.J.M. Mahdi), barry.wilson@usda.gov (B.T. Wilson).

https://doi.org/10.1016/j.spasta.2026.100988
Received 14 October 2025; Received in revised form 1 April 2026; Accepted 2 April 2026
Available online 25 May 2026
2211-6753/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

for large point-level spatial datasets has become an area of active research for about past two decades, see Datta et al. (2016) and the references therein. In this article, we focus on the latter kind which represents an equally important domain of spatial model development and presents computational challenges of different nature, as discussed below.

Within the Bayesian framework, one of the most common approaches for modeling adjacency-based spatial association is to use the family of Markov Random Field-type distributions (Besag, 1974). Here, the focus lies on the precision matrix of the joint distribution which characterizes conditional (in)dependence between observations from different units. One of the most well-known members of this family is the conditional autoregressive (CAR) distribution (Besag and Kooperberg, 1995) which offers the advantage of univariate easy-to-sample posterior full conditional distributions for individual random effects in linear and generalized linear models. So, even with large datasets, one does not have to handle the computational challenges associated with high-dimensional covariance matrices, a feature critical to point-level GP-based spatial models. On the other hand, estimating these large number of random effects one-by-one from full conditional distributions becomes time-consuming and, more importantly, may result in strong autocorrelation and poor mixing performance (Rue and Held, 2005) in a Markov chain Monte Carlo (MCMC) scheme. In the well-known INLA framework (Rue et al., 2009), these challenges are addressed by replacing MCMC-based inference with an approximate Bayesian method that utilizes sparsity of the precision matrix for fast inference.

The original contribution of the current article is to develop low-rank alternatives to the CAR prior utilizing the spectral structure of the Laplacian of adjacency matrix. Laplacian matrix has been used in the context of dimension reduction (McInnes et al., 2020; Zhu and Koniusz, 2022) and Bayesian variable selection (Liu et al., 2014). Its appeal in adjacency-based spatial framework arises from the fact that its eigenvectors can be ordered with respect to their spatial smoothness and offer a natural choice for basis functions over the lattice. In the literature on large point-referenced datasets, use of basis functions is well-studied (Higdon, 2002; Cressie and Johannesson, 2008). Nychka et al. (2002) mentioned the possibility of using eigenvectors of spatial covariance matrix at observed locations as basis functions. However, to the best of our knowledge, Mahdi et al. (2019) represents the only peer-reviewed research to date that explores use of Laplacian matrix to develop computationally efficient models for areal spatial datasets. The current article, founded in their approach, proposes new models making several improvements and extensions as listed below. First, we show that if prior distributions assigned to the basis coefficients are scaled with associated eigenvalues, the induced joint prior on areal-level random effects becomes a low-dimensional approximation to the CAR prior. We present simulation studies to establish that the original approach suffers from overestimation of the variance parameter for spatial random effect that can be corrected by the above-mentioned modification in the prior. Our second idea is the areal-level analogue of how Finley et al. (2009) introduced variance correction to the original predictive process approximation (Banerjee et al., 2008) for large point-level spatial datasets. We show that, by introduction of spatially heteroscedastic yet uncorrelated random noise, one can control the overestimation error for the nugget variance term. In implementing this approach, we had to overcome one additional computational challenge. The precision matrix for CAR prior joint distribution is available in closed form but finding the variances requires inversion of that high-dimensional precision matrix. The Lanczos algorithm (Calvetti et al., 1994) that we use for efficiently computing eigenvectors associated with smallest few eigenvalues, is also utilized to approximate the generalized inverse (Bozzo and Franceschet, 2012) of the precision matrix and its diagonal entries are subsequently extracted to apply variance correction. Finally, we extend the two proposed models — with and without this correction, to settings where spatiotemporal extensions of CAR priors are used. The exact hierarchical specification and MCMC scheme depend on how spatial and temporal dependences interact. For illustrative purpose, we choose a model consisting of a variant of first-order time-autoregressive CAR structure (Rushworth et al., 2014). We analyze several simulated as well as real-world spatial and spatiotemporal datasets to show how our proposed methods perform relative to the original CAR prior. For comparative diagnostics, we present multiple tables and diagrams based on predictive performance in out-of-sample cross validation and computational characteristics. It is important to note that, since the proposed methods apply to any dataset with adjacency-based dependence structure, their domain of potential applications extends well beyond areal datasets — to any setting where measurements are connected through a graphical network such as subsurface flow and transport process in geology (Bhark et al., 2011).

This article is organized as follows. In Section 2, we review the CAR-based hierarchical model and describe how Mahdi et al. (2019) built a reduced-rank representation for areal-level random effects. Section 3 introduces the proposed approximation to the CAR prior and subsequently, improves the approximation through variance adjustment. In Section 4, we illustrate how to extend the proposed methods to a spatiotemporal setting. Section 5 discusses details of spectral computation and metrics used for performance comparison. Sections 6 and 7 present the findings from multiple simulation studies and real data analyses, respectively, that we implemented using R (R Core Team, 2025). We conclude the article in Section 8 reviewing key features of our approach and outlining possible extensions. Throughout this article, the notations $\mathcal{N}, \mathcal{N}_n$ and IG will be used for univariate normal, $n$-variate normal and inverse-gamma probability distributions, respectively. As usual, $I_n$ will denote the $n \times n$ identity matrix.

## 2. Review of CAR prior and a reduced-rank alternative

For any areal data, the region of interest $\mathcal{A}$ is partitioned into a collection of regular or irregularly shaped disjoint areal units $\{A_1, A_2, \ldots, A_S\}$ and each observation in the dataset typically corresponds to one of these areal units. Suppose at areal unit $A_s$, $y_s$ and $X_{s*}$ are recorded as the summary response and the $(p+1)$-dimensional vector of covariates (including intercept), respectively. Now, a simple spatial linear model can be specified for $y_s$ as:

$$y_s = X_{s*}^T \beta + \varphi_s + \epsilon_s, \quad s = 1, 2, \ldots, S \tag{2.1}$$

where $\beta$ is the vector of $(p+1)$-many regression coefficients, $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_S)^T$ is the vector of spatial random effects and, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_S)^T$ is the vector of i.i.d. pure error terms with variance $\sigma^2$. Specifying a prior distribution for $\varphi_s$ requires defining a $S \times S$ adjacency matrix $W$. One can specify entries of $W$ such that $w(s,s') > 0$ if $A_s$ and $A_{s'}$ are defined as neighbors and 0, otherwise. For our analysis, we use binary and symmetric $W$ matrices, the most commonly used choice in literature. The CAR prior, developed by Besag et al. (1991), uses entries from $W$ to specify full conditional distribution for $\varphi_s$ as:

$$\varphi_s | \{\varphi_{s'} : s' \neq s\} \sim \mathcal{N}\left( \frac{\sum_{s'} w(s,s')\varphi_{s'}}{w_{s+}}, \frac{\tau^2}{w_{s+}} \right), \tag{2.2}$$

where $\tau$ is the scale parameter of the conditional distribution and $w_{s+} = \sum_{s'} w(s,s')$ denotes the number of neighbors of areal unit $A_s$. The resulting joint distribution for $\varphi$ becomes:

$$\pi(\varphi) \propto \exp\{-\frac{1}{2\tau^2}\varphi^T(D_w - W)\varphi\}, \tag{2.3}$$

where $D_w$ is a $S \times S$ diagonal matrix with $D_w(s,s) = w_{s+}$. This prior distribution, denoted as $CAR\left(W, \tau^2\right)$, is a common choice in Bayesian areal spatial modeling literature because its univariate full conditional distributions offer intuitive interpretation and, more importantly, convenient Gibbs update within an MCMC scheme. However, it comes with some important limitations. First, since each component of $\varphi$ has to be sequentially sampled at every iteration of the MCMC, computation time increases significantly for a dataset with very large number of areal units. Second, more importantly, this sequential updating of random effects from respective posterior full conditional distributions introduces strong autocorrelation across the successive MCMC draws lowering the effective sample size (ESS), so that one needs a larger thinning interval to obtain approximately independent posterior draws. Third, the prior joint distribution in Eq. (2.3) is improper (due to invariance to translation) and it leads to an improper posterior for $\varphi$ in Eq. (2.1), see Banerjee et al. (2026, Chapter 4). Two solutions discussed there include (i) the $\rho$-CAR approach where the parameter $\rho < 1$ induces propriety with $\rho = 1$ taking us back to usual CAR prior discussed above and, (ii) the ad-hoc "center on the fly" approach where each draw of the $\varphi$ vector within MCMC is mean-centered by subtracting its mean from each component. However, the authors pointed out that, in solution (i), a reasonably strong spatial correlation would need $\rho$ to be fixed at a value very close to 1. All implementations of CAR prior in current article use the specification from Eq. (2.2) in combination with solution (ii).

Mahdi et al. (2019) proposed utilizing the spectral properties of $L = D_w - W$, the Laplacian of $W$ and a singular non-negative definite matrix, to come up with a computationally efficient alternative model for $\varphi$. The authors therein observed that when the Laplacian eigenpairs are labeled as $\{(\lambda_{(i)}, v_i) : i = 1, 2, \ldots, S\}$ according to increasing eigenvalues $\lambda_{(1)} \leq \lambda_{(2)} \leq \cdots \leq \lambda_{(S)}$, the eigenvectors $v_1, v_2, \ldots, v_S$ represent functions on the areal units with decreasing strength of spatial association. Based on this characteristics, they proposed using a certain number ($k$) of eigenvectors associated with small eigenvalues as basis functions and defined the areal-level random effects as a linear combination:

$$\tilde{\varphi} = \theta_1 v_2 + \theta_2 v_3 + \cdots + \theta_k v_{k+1}, \;\; \theta_i \overset{iid}{\sim} \mathcal{N}(0, \tau^2) \text{ for } i = 1, 2, \ldots, k, \tag{2.4}$$

where the eigenvector $v_1 = 1_{S\times 1}$ associated with the smallest eigenvalue (= 0) is excluded to maintain identifiability due to presence of intercept in the fixed effect of Eq. (2.1). In this article, we assume that the adjacency matrix $W$ is connected. The coefficient vector $\theta = \left(\theta_1, \theta_2, \ldots, \theta_k\right)^T$ represents the weights of these basis functions in the areal random effects.

To efficiently compute the eigenvectors associated with $k$ smallest Laplacian eigenvalues, Mahdi et al. (2019) utilized a modified version of Implicit Restarted Lanczos (IRL) method (Calvetti et al., 1994, Algorithm 4.4). Under certain assumptions, this method has a computational complexity to $\mathcal{O}(Sk^2)$ compared to the $\mathcal{O}(S^3)$ complexity of the conventional QR method leading to significant improvement in computational efficiency in terms of reduced time and memory requirements.

With respect to parameter estimation, computational advantage of the specification in Eq. (2.4) is evident. Here, we just need to estimate $k$ coefficients $\theta_1, \theta_2, \ldots, \theta_k$ instead of $S$ random effects $\varphi_1, \varphi_2, \ldots, \varphi_S$ in Eq. (2.1). Moreover, the linear regression-like structure $\tilde{\varphi}$ ensures that these coefficients can be sampled simultaneously from their joint posterior instead of the sequential full conditional updating in CAR leading to significant improvement in autocorrelation and mixing within an MCMC scheme. This is a reduced-rank method as $S$ spatial effects are represented as a linear combination of a much smaller number ($k$) of eigenvectors. The omitted eigenvectors, corresponding to $(S-k-1)$ largest eigenvalues, behave more like random noise with weak spatial association and the pure error term $\epsilon$ approximates their collective contribution. Presence of $\epsilon$ also ensures that, with a reduced-rank model for $\tilde{\varphi}$, we still have a full-rank dispersion at the data-level in Eq. (2.1). In the remainder of this article, we refer to this approach as the spectral method (SM).

## 3. Proposed approximations to CAR prior

We start by exploring the connection between the prior distributions in CAR and SM. The Laplacian $L$ features in both of them. Eq. (2.3) implies that the joint prior for $\varphi$ in CAR is multivariate normal with zero mean and precision matrix $L/\tau^2$. Due to the singularity of $L$, we use, as in Lee and Lee (2013), its Moore–Penrose pseudo inverse (Ben-Israel and Greville, 2003) $L^+$ to define the dispersion matrix for $\varphi$ in CAR as:

$$D(\varphi) = \tau^2 L^+ = \tau^2 \sum_{i=2}^{S} \frac{1}{\lambda_{(i)}} v_i v_i^T \;\; = \;\; \tau^2 \sum_{i=2}^{S} \frac{v_i}{\sqrt{\lambda_{(i)}}} \left( \frac{v_i}{\sqrt{\lambda_{(i)}}} \right)^T. \tag{3.1}$$

On the other hand, marginalizing the coefficients out of the prior distribution of $\tilde{\varphi}$ in Eq. (2.4) for SM, we get a multivariate normal distribution with zero mean and dispersion matrix

$$D(\tilde{\varphi}) = \tau^2 \sum_{i=2}^{k+1} v_i v_i^T \tag{3.2}$$

Comparison between expressions from Eqs. (3.1) and (3.2) suggests that, if one rescales the eigenvector coefficients in Eq. (2.4) using the square-root of the corresponding eigenvalues, the marginal prior distributions of $\tilde{\varphi}$ and $\varphi$ will both be zero-mean multivariate normal distributions with the dispersion matrix of the former being a truncated version (consisting of the first $k$ out of $(S-1)$ terms) of the same for the latter. We use this notion to modify Eq. (2.4) as:

$$\tilde{\varphi} = \theta_1 v_2 + \theta_2 v_3 + \cdots + \theta_k v_{k+1}, \;\; \theta_i \overset{ind}{\sim} \mathcal{N}\Big(0, \frac{\tau^2}{\lambda_{(i+1)}}\Big) \text{ for } i = 1, 2, \ldots, k. \tag{3.3}$$

Redefining the prior distributions for the eigenvector coefficients as above yields an intuitive interpretation. Using how Mahdi et al. (2019) defined total variability of a function depending on the adjacency structure, it is easy to see that prior variance of the weight of any eigenvector is inversely proportional to its total variability. That means, more smoothly an eigenvector changes on the lattice, we assign a more diffused prior distribution to its weight parameter. With decreasing spatial smoothness, the prior distribution for the weight gets more tightly condensed around the mean zero. Hence, Eq. (3.3) effectively (i) asserts that, a priori, eigenvectors representing stronger spatial association are more likely to have significant effects and, (ii) assigns stronger prior shrinkage to effects of eigenvectors exhibiting weaker spatial pattern.

We briefly note that our idea of using the Laplacian eigenvectors reweighed by corresponding eigenvalues has an analogue for point-level spatial datasets. If a point-level zero-mean spatial process $\{w(s) : s \in D\}$ is assigned a GP prior with positive definite covariance function $C(s, s')$, using Karhunen–Loéve expansion for the GP prior we can express $w(s) = \sum_{i=1}^{\infty} H_i h_i(s)$ where $H_i \overset{ind}{\sim} \mathcal{N}(0, \alpha_i)$ with $\alpha_i = i$-th largest eigenvalue of $C(s, s')$ and $h_i(\cdot)$ the corresponding eigenfunction. Restricting this sum to a finite number of terms (as $i \uparrow$, $\text{var}(H_i) \downarrow$) will lead to a reduced-rank representation of the $w(\cdot)$ process, see Sang and Huang (2012) and Banerjee et al. (2026, Chapter 13.3.2) for related discussion. We can view our model for $\varphi$ as a discrete, lattice-version of this approach with $L^+$ representing the (unscaled) dispersion matrix and, in Eq. (3.3), $\big(\theta_i, v_{i+1}, 1/\lambda_{(i+1)}\big)$ playing the role of $\big(H_i, h_i(\cdot), \alpha_i\big)$.

We refer to the model for $\tilde{\varphi}$ in Eq. (3.3) as the modified spectral method (MSM) hereafter. Appendix A.1 quantifies the Kullback–Leibler (KL) divergence between the marginal distributions induced on the response vector $y$ by the CAR and MSM prior on random effects and establishes its upper bound depending on the $(S-k-1)$ discarded Laplacian eigenvalues as well as the ratio of variance parameters $\tau^2$ and $\sigma^2$.

Conceptually, the difference between CAR and MSM can be viewed in terms of local *vs.* global model for areal random effects. In CAR, each one of the $S$ areal units has its own random effect that is influenced by random effects from units in the vicinity. So, there is scope for posterior learning from the data about localized patterns. On the other hand, MSM combines a relatively small number of basis functions (i.e., the eigenvectors) with known spatial patterns according to a set of global weights (i.e. the components of $\theta$) to produce the random effects and posterior learning occurs for these weight parameters only. Hence, the dimension-reduction achieved in moving from CAR to MSM comes with the assumption that a linear combination of some known functions can explain the spatial variation throughout the region reasonably well. In practice, performance of MSM relative to CAR will depend on the extent to which this assumption is supported by the data.

### 3.1. Correction in variance estimation

Since MSM uses $k$ many free parameters $\theta_1, \theta_2, \ldots, \theta_k$ to construct $S$-many spatial random effects with $k << S$, it leads to underestimation of the variances of those random effects compared to CAR. It is easy to see that

$$\begin{aligned} var(\varphi_s) - var(\tilde{\varphi}_s) &= \left(\tau^2 \sum_{i=2}^{S} \frac{1}{\lambda_{(i)}} v_i v_i^T\right)_{ss} - \left(\tau^2 \sum_{i=2}^{k+1} \frac{1}{\lambda_{(i)}} v_i v_i^T\right)_{ss} \\ &= \tau^2 \sum_{i=2}^{S} \frac{1}{\lambda_{(i)}} v_{si}^2 - \tau^2 \sum_{i=2}^{k+1} \frac{1}{\lambda_{(i)}} v_{si}^2 = \tau^2 \sum_{i=k+2}^{S} \frac{1}{\lambda_{(i)}} v_{si}^2 \;\; > 0, \end{aligned} \tag{3.4}$$

where $v_{si}$ is the $s$th entry of the eigenvector $v_i$. Underestimation of uncertainty is systemic to reduced-rank methods. Finley et al. (2009) observed the same in the context of the predictive process approximation for large point-level spatial datasets. Consequently, in practical implementations, the underestimation of spatial uncertainty is compensated by overestimation of nugget variance parameter $\sigma^2$. To address such bias in estimation of variance parameters, we adopt a technique similar to Finley et al. (2009) by introducing a heteroscedastic error term in MSM to account for the excess variance in Eq. (3.4). Thus, we have a new model for spatial random effects given by

$$\ddot{\varphi}_s = \tilde{\varphi}_s + \tilde{\epsilon}_s, \;\; \tilde{\epsilon}_s \overset{ind}{\sim} \mathcal{N}\Big(0, \tau^2 d_s\Big), \; d_s = L^+_{ss} - \sum_{i=2}^{k+1} \frac{1}{\lambda_{(i)}} v_{si}^2 \;\; \text{for } s = 1, 2, \ldots, S. \tag{3.5}$$

It follows that $\text{var}(\ddot{\varphi}_s) = \text{var}(\varphi_s)$, as desired. We refer to this approach as variance adjusted modified spectral method (VA-MSM).

## 4. Extension to spatiotemporal datasets

Now that the proposed approximation methods are developed in a spatial-only setting, we show how they can be applied to analyze spatiotemporal datasets over large number of areal units at discrete time-points. Let $y_{st}$ be the summary response and $X_{st*}$ be a $p\times 1$ vector of predictors (including intercept) at areal unit $A_s$ at time $t$. The spatiotemporal regression equation can be written as:

$$y_{st} = X_{st*}^T\beta + \varphi_{st} + \epsilon_{st};\ s = 1,2,\dots,S \text{ and } t = 1,2,\dots,T, \tag{4.1}$$

where $\varphi_{st}$ is the spatiotemporal random effect, and $\epsilon_{st} \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$ represents the pure error uncorrelated in space and time. There exists a rich family of literature on ways to model the spatiotemporal random effect such as separable space–time covariance structure (Hamaamin et al., 2016), spatially varying linear time trend (Bernardinelli et al., 1995), additive space–time effects (Knorr-Held, 2000) etc. See Lee et al. (2018) for an overview of some of the available model choices. The steps to implement the MSM and VA-MSM methods depend on the specifics of how spatial association is being modeled in combination with temporal effect. Here, we choose to illustrate with an example where members of $\{\varphi_{st}\}$ are modeled as a dynamic sequence of spatial random effect vectors interconnected through a first-order autoregressive structure. We refer to this as autoregressive CAR (ARCAR) prior and specify its form below:

$$\varphi_{*t} = \eta\varphi_{*(t-1)} + \zeta_{*t};\ t = 1,2,\dots,T, \tag{4.2}$$

where $\zeta_{*t} = \{\zeta_{st} : s = 1,2,\dots,S\} \sim CAR(W,\tau^2)$ and $\varphi_{*0} = \{\varphi_{s0} : s = 1,2,\dots,S\} \sim CAR\big(W, \frac{\tau^2}{1-\eta^2}\big)$ with $|\eta| < 1$. This form is similar to the one used by Rushworth et al. (2014) with some minor differences. The reason behind this choice to illustrate use of MSM and VA-MSM is that, since it includes a time series of spatial random effects with CAR prior, the computational advantage of the proposed methods is expected to be many-fold compared to settings with a single set of spatial random effects. We provide details for extension of both methods below.

### 4.1. Spatiotemporal extension of MSM

As before, we assume a connected adjacency structure so that $L$ has only one zero eigenvalue with (unnormalized) eigenvector $1_{S\times 1}$. There are $(T+1)$ many mutually independent vectors of spatial random effects $\tilde{\varphi}_{*0}, \zeta_{*1}, \zeta_{*2}, \dots, \zeta_{*T}$ in Eq. (4.2) and we use a common value of $k$ to choose the number of eigenpairs for each one of them. Let, $V = [v_2, v_3, \dots, v_{k+1}]$ where $v_i$ be the eigenvector corresponding to the $i$th smallest eigenvalue, $\lambda_{(i)}$, of the Laplacian matrix $L$. Then, using the low-rank representation of the MSM, we can write $\tilde{\zeta}_{*t} = V\theta_{*t}$ and $\tilde{\varphi}_{*0} = V\theta_{*0}$ where $\theta_{*t}$ and $\theta_{*0}$ are vectors of size $k$ with $\theta_{it} \overset{ind}{\sim} \mathcal{N}(0,\tau^2/\lambda_{(i+1)})$ and $\theta_{i0} \overset{ind}{\sim} \mathcal{N}\Big(0, \frac{\tau^2}{\lambda_{(i+1)}(1-\eta^2)}\Big)$ for $i = 1,2,\dots,k$ and $t = 1,2,\dots,T$. The reduced-rank representation of $\varphi_{st}$ can be written as:

$$\begin{aligned}
\tilde{\varphi}_{*0} &= \tilde{\zeta}_{*0} = V\theta_{*0};\\
\tilde{\varphi}_{*1} &= \eta\tilde{\varphi}_{*0} + \tilde{\zeta}_{*1} = \eta V\theta_{*0} + V\theta_{*1} = V(\eta\theta_{*0} + \theta_{*1});\\
\tilde{\varphi}_{*2} &= \eta\tilde{\varphi}_{*1} + \tilde{\zeta}_{*2} = V(\eta^2\theta_{*0} + \eta\theta_{*1} + \theta_{*2});\\
&\vdots\\
\tilde{\varphi}_{*T} &= \eta\tilde{\varphi}_{*(T-1)} + \tilde{\zeta}_{*T} = V\Big(\sum_{t=0}^{T}\eta^{T-t}\theta_{*t}\Big).
\end{aligned}$$

Whereas one can certainly use different values of $k$ at different time points, the advantage of working with a single $V$ matrix, as seen in above system of equations, will no longer apply in that case. We reparametrize the linear combinations of $\theta_{*t}$ for $t = 0,1,2,\dots,T$ as follows

$$\begin{aligned}
\psi_{*0} &= \theta_{*0};\\
\psi_{*1} &= \eta\theta_{*0} + \theta_{*1} = \eta\psi_{*0} + \theta_{*1};\\
\psi_{*2} &= \eta^2\theta_{*0} + \eta\theta_{*1} + \theta_{*2} = \eta\psi_{*1} + \theta_{*2};\\
&\vdots\\
\psi_{*T} &= \sum_{t=0}^{T}\eta^{T-t}\theta_{*t} = \eta\psi_{*(T-1)} + \theta_{*T};
\end{aligned}$$

where $\psi_{*0} \sim \mathcal{N}_k\Big(0, \frac{\tau^2}{1-\eta^2}\,\Lambda^{-1}\Big)$, $\psi_{*t}|\psi_{*(t-1)} \sim \mathcal{N}_k(\eta\,\psi_{*(t-1)}, \tau^2\Lambda^{-1})$ for $t = 1,2,\dots,T$, and $\Lambda$ is a $k\times k$ diagonal matrix with $i$th diagonal entry $\lambda_{(i+1)}$. Hence, our proposed first-order autoregressive MSM (ARMSM) becomes:

$$\tilde{\varphi}_{*t} = V\psi_{*t},\ \text{for } t = 0,1,\dots,T, \tag{4.3}$$

where $\psi_{*0} \sim \mathcal{N}_k\Big(0, \frac{\tau^2}{1-\eta^2}\,\Lambda^{-1}\Big)$ and $\psi_{*t}|\psi_{*(t-1)} \sim \mathcal{N}_k(\eta\,\psi_{*(t-1)}, \tau^2\Lambda^{-1})$ for $t = 1,2,\dots,T$.

### 4.2. Spatiotemporal extension of VA-MSM

Similar to MSM, ARMSM underestimates the spatial variance of ARCAR model at every areal unit $A_s$ and at every time point $t$ since

$$\text{var}(\varphi_{s0}) - \text{var}(\tilde{\varphi}_{s0}) = \frac{\tau^2}{1-\eta^2}\left(\sum_{i=2}^{S}\frac{1}{\lambda_{(i)}}v_{si}^2 - \sum_{i=2}^{k+1}\frac{1}{\lambda_{(i)}}v_{si}^2\right)$$

$$\text{var}(\varphi_{st}|\varphi_{s(t-1)}) - \text{var}(\tilde{\varphi}_{st}|\tilde{\varphi}_{s(t-1)}) = \tau^2\left(\sum_{i=2}^{S}\frac{1}{\lambda_{(i)}}v_{si}^2 - \sum_{i=2}^{k+1}\frac{1}{\lambda_{(i)}}v_{si}^2\right),\ 1 \le t \le T$$

To rectify this variance underestimation, we use the technique of adding spatially heteroscedastic noise term from Section 3.1, separately at each time point to define $\ddot{\zeta}_{*t} = \tilde{\zeta}_{*t} + \tilde{\epsilon}_{*t}$ and $\ddot{\varphi}_{*0} = \tilde{\varphi}_{*0} + \tilde{\epsilon}_{*0}$ where $\tilde{\epsilon}_{*t}$ and $\tilde{\epsilon}_{*0}$ are vectors of size $S$, $\tilde{\epsilon}_{st} \overset{ind}{\sim} \mathcal{N}(0, \tau^2 d_s)$, $\tilde{\epsilon}_{s0} \overset{ind}{\sim} \mathcal{N}\left(0, \frac{\tau^2}{(1-\eta^2)}d_s\right)$ and $d_s = L_{ss}^{+} - \sum_{i=2}^{k+1}\frac{1}{\lambda_{(i)}}v_{si}^2$ for $s = 1, 2, \ldots, S$ and $t = 1, 2, \ldots, T$. Using this representation of VA-MSM, $\varphi_{st}$ can be written as:

$$\begin{aligned}
\ddot{\varphi}_{*0} &= \ddot{\zeta}_{*0} = \tilde{\varphi}_{*0} + \tilde{\epsilon}_{*0} = V\theta_{*0} + \tilde{\epsilon}_{*0};\\
\ddot{\varphi}_{*1} &= \eta\ddot{\varphi}_{*0} + \ddot{\zeta}_{*1} = \eta V\theta_{*0} + \eta\tilde{\epsilon}_{*0} + V\theta_{*1} + \tilde{\epsilon}_{*1} = V(\eta\theta_{*0} + \theta_{*1}) + \eta\tilde{\epsilon}_{*0} + \tilde{\epsilon}_{*1};\\
\ddot{\varphi}_{*2} &= \eta\ddot{\varphi}_{*1} + \ddot{\zeta}_{*2} = V(\eta^2\theta_{*0} + \eta\theta_{*1} + \theta_{*2}) + \eta^2\tilde{\epsilon}_{*0} + \eta\tilde{\epsilon}_{*1} + \tilde{\epsilon}_{*2};\\
&\vdots\\
\ddot{\varphi}_{*T} &= \eta\ddot{\varphi}_{*(T-1)} + \ddot{\zeta}_{*T} = V\left(\sum_{t=0}^{T}\eta^{T-t}\theta_{*t}\right) + \sum_{t=0}^{T}\eta^{T-t}\tilde{\epsilon}_{*t}.
\end{aligned}$$

We separately reparametrize the two sets of linear combinations of $\theta_{*t}$ and $\tilde{\epsilon}_{*t}$ for $t = 0, 1, 2, \ldots, T$, as:

| | |
|---|---|
| $\psi_{*0} = \theta_{*0};$ | $\ddot{\epsilon}_{*0} = \tilde{\epsilon}_{*0};$ |
| $\psi_{*1} = \eta\theta_{*0} + \theta_{*1} = \eta\psi_{*0} + \theta_{*1};$ | $\ddot{\epsilon}_{*1} = \eta\tilde{\epsilon}_{*0} + \tilde{\epsilon}_{*1} = \eta\ddot{\epsilon}_{*0} + \tilde{\epsilon}_{*1};$ |
| $\psi_{*2} = \eta^2\theta_{*0} + \eta\theta_{*1} + \theta_{*2} = \eta\psi_{*1} + \theta_{*2};$ | $\ddot{\epsilon}_{*2} = \eta^2\tilde{\epsilon}_{*0} + \eta\tilde{\epsilon}_{*1} + \tilde{\epsilon}_{*2} = \eta\ddot{\epsilon}_{*1} + \tilde{\epsilon}_{*2};$ |
| $\vdots$ | $\vdots$ |
| $\psi_{*T} = \sum_{t=0}^{T}\eta^{T-t}\theta_{*t} = \eta\psi_{*(T-1)} + \theta_{*T};$ | $\ddot{\epsilon}_{*T} = \sum_{t=0}^{T}\eta^{T-t}\tilde{\epsilon}_{*t} = \eta\ddot{\epsilon}_{*(T-1)} + \tilde{\epsilon}_{*T};$ |

It follows from the above reparametrization that

- $\psi_{*0} \sim \mathcal{N}_k\left(0, \frac{\tau^2}{1-\eta^2}\,\Lambda^{-1}\right)$, $\psi_{*t}|\psi_{*(t-1)} \sim \mathcal{N}_k(\eta\,\psi_{*(t-1)}, \tau^2\Lambda^{-1})$ for $t = 1, 2, \ldots, T$, and $\Lambda$ is a $k \times k$ diagonal matrix with $i$th diagonal entry $\lambda_{(i+1)}$.
- $\ddot{\epsilon}_{*0} \sim \mathcal{N}_S\left(0, \frac{\tau^2}{1-\eta^2}\,\Delta\right)$, $\ddot{\epsilon}_{*t}|\ddot{\epsilon}_{*(t-1)} \sim \mathcal{N}_S(\eta\,\ddot{\epsilon}_{*(t-1)}, \tau^2\Delta)$ and $\Delta$ is a $S \times S$ diagonal matrix with $s$th diagonal entry $d_s$.

Hence, our proposed first-order autoregressive VA-MSM (ARVA-MSM) becomes:

$$\ddot{\varphi}_{*t} = V\psi_{*t} + \ddot{\epsilon}_{*t} = \tilde{\varphi}_{*t} + \ddot{\epsilon}_{*t},\ \text{for } t = 0, 1, 2, \ldots, T, \tag{4.4}$$

where $\psi_{*0} \sim \mathcal{N}_k\left(0, \frac{\tau^2}{1-\eta^2}\,\Lambda^{-1}\right)$, $\psi_{*t}|\psi_{*(t-1)} \sim \mathcal{N}_k(\eta\,\psi_{*(t-1)}, \tau^2\Lambda^{-1})$, $\ddot{\epsilon}_{*0} \sim \mathcal{N}_S\left(0, \frac{\tau^2}{1-\eta^2}\,\Delta\right)$, and $\ddot{\epsilon}_{*t}|\ddot{\epsilon}_{*(t-1)} \sim \mathcal{N}_S(\eta\,\ddot{\epsilon}_{*(t-1)}, \tau^2\Delta)$ for $t = 1, 2, \ldots, T$.

## 5. Computational details

Now that the proposed methods are developed in both spatial and spatiotemporal settings, it is important to discuss the key computational steps to implement them. In Section 5.1, we discuss how the Moore–Penrose pseudo inverse of the Laplacian is computed. Section 5.2 discusses the diagnostic tools we can use to assess the performance of these methods relative to CAR. The posterior distributions for parameters in the spatial (CAR, SM, MSM and VA-MSM) and spatiotemporal (ARCAR, ARMSM and ARVA-MSM) models are described in Appendices A.2, A.3 and A.4.

### 5.1. G-inverse computation

VA-MSM requires the diagonal entries of the dispersion matrix for spatial random effects in CAR prior. In Eq. (3.1), we specify that matrix using generalized inverse $L^{+}$ in terms of the eigenvalues and eigenvectors of the Laplacian. Direct computation of this expression is computationally challenging for large values of $S$ since it requires determining all $(S-1)$ eigenpairs. Hence, we utilize the stretch approximation method (Bozzo and Franceschet, 2012) to approximate $L^{+}$ using a relatively small number (that we denote here using $m_0$) of Laplacian eigenpairs. We briefly describe the method below.

We begin by computing Laplacian eigenpairs $\{(\lambda_{(i)}, \upsilon_i) : i = 2, 3, \ldots, m_0\}$ corresponding to $(m_0 - 1)$ smallest eigenvalues, exactly the kind of computation already done in MSM and VA-MSM, using IRL method, to determine the eigenvectors representing the spatial random effects $\tilde{\varphi}$ and $\ddot{\varphi}$, respectively. An approximate spectral representation of the generalized inverse is defined by replacing remaining $(S - m_0)$ eigenvalues with a suitable value $c$ as:

$$\tilde{L}^{+(m_0)} := \sum_{i=2}^{m_0} \frac{1}{\lambda_{(i)}} \upsilon_i \upsilon_i^T + \sum_{i=m_0+1}^{S} \frac{1}{c} \upsilon_i \upsilon_i^T = \sum_{i=2}^{m_0} \frac{1}{\lambda_{(i)}} \upsilon_i \upsilon_i^T - \sum_{i=1}^{m_0} \frac{1}{c} \upsilon_i \upsilon_i^T + \sum_{i=1}^{S} \frac{1}{c} \upsilon_i \upsilon_i^T$$
$$= \frac{1}{c} I_S - \frac{1}{c} \upsilon_1 \upsilon_1^T + \sum_{i=2}^{m_0} \Big( \frac{1}{\lambda_{(i)}} - \frac{1}{c} \Big) \upsilon_i \upsilon_i^T. \tag{5.1}$$

where $I_S$ denotes the $S \times S$ identity matrix. Bozzo and Franceschet (2012) computed the approximation accuracy using relative error, which is the ratio of the 2-norm (largest singular value, denoted by $\|\cdot\|_2$) for the difference matrix between generalized inverse and its approximation, and the 2-norm of the generalized inverse itself. They have determined that the value of $c$ that minimizes this relative error satisfies $\frac{1}{c} = \frac{1}{2}\Big( \frac{1}{\lambda_{(m_0+1)}} + \frac{1}{\lambda_{(S)}} \Big)$ and, with this choice, the relative error of stretch approximation is bounded as

$$\frac{\|L^+ - \tilde{L}^{+(m_0)}\|_2}{\|L^+\|_2} \leq \frac{\lambda_{(2)}}{2} \Big( \frac{1}{\lambda_{(m_0+1)}} - \frac{1}{\lambda_{(S)}} \Big), \tag{5.2}$$

In our analysis, we first determine the lowest $m_0 \in \mathbb{N}$ that makes the upper bound in Eq. (5.2) smaller than a pre-specified error threshold and plug-in that value in Eq. (5.1) to approximate $L^+$ with $\tilde{L}^{+(m_0)}$.

### 5.2. Criteria for performance evaluation

Both MSM and VA-MSM require the user to choose the number of eigenvectors $k$ (more on that in Section 8) to use in the expression for areal random effects. Since the strength of spatial association varies between different datasets, the required number of eigenpairs should be determined by the data. Therefore, as the first step in every data analysis included in this article, we perform out-of-sample cross-validation (CV) to find the optimal value of $k$.

In statistical literature, CV is typically carried out by randomly allocating the observations into training and test pools. Here, we take a different approach. Since spatial models predict at any areal unit (with missing response) by borrowing information from its neighboring cells, the prediction is expected to be better if test cells are scattered randomly over the study region and are mostly surrounded by training cells. On the other hand, when data from all units in a large contiguous block are missing, the predictive performance is expected to depend on nature of spatial association in the data and the model's ability to capture that accurately. The prediction is expected to be better on the boundary units of the block due to the availability of training data from nearby units. However, as we move to units in the interior, prediction uncertainty accumulates since it is based on neighboring values that are also predicted by the model. So, it will be interesting to see how predictions based on a CAR prior that uses unit-wise local random effects compare against those from MSM and VA-MSM that use global combinations of eigenvectors. In real-world instances of geographically-indexed datasets over large regions, such contiguous patches of missingness can occur for reasons such as difficulty in data collection from inaccessible terrains or unavailability of remote sensing imagery due to widespread weather events. In this article, we always perform CV by removing a large block from the data as a test set and fitting the models on the remaining training dataset. This approach is referred to as *single block spatial CV* hereafter. The optimal $k$ is selected based on two criteria: (i) mean square error (MSE), computed as the average of the squared differences between the hidden test observations and their corresponding MCMC draws after burning and thinning; and (ii) MCMC time, i.e., run time for MCMC iterations. The model that has the lowest MSE with a reasonably smaller computation time is deemed to be the best, and the number of eigenpairs used in that model is chosen as the optimal value of $k$.

After implementing MSM and VA-MSM with $k$ chosen as above, we compare their predictive and computational performance relative to CAR via single block spatial CV. Predictive performance is evaluated using three measures: (i) absolute bias: computed as the absolute difference between the test observations and their corresponding posterior medians; (ii) uncertainty, measured as the width of the 90% highest posterior density (HPD) interval and (iii) empirical coverage, calculated as the proportion of the test observations that fall within the corresponding 90% HPD interval. The model with minimum absolute bias and uncertainty with desired coverage is considered the best. Computational performance is assessed using three measures. The first two are: (i) MCMC time and (ii) effective sample size (ESS) proportion, the ratio of the ESS calculated from posterior samples of log likelihood and the number of MCMC draws (both calculated after burn-in and thinning). A small value of ESS indicates high autocorrelation and poor mixing among the MCMC draws. To calculate ESS, we use coda (Plummer et al., 2006) R package. Then we compute the effective computation time by taking the ratio of ESS proportion and MCMC time, and report (iii) the relative time efficiency calculated as the ratio of effective computation time for the proposed methods and the corresponding value for CAR.

We also explore the effect of contiguity in the test block on the predictive performance. For this, in addition to single block spatial CV, another spatial CV is performed with a similar size of the test set. However, the test units in this case come from three smaller disjoint blocks instead of one large block. We refer to this as *three blocks spatial CV*. The predictive performance is examined using the same three measures described above.

In case of spatial datasets, smoothness represents another basis for performance comparison. The metric should measure to what extent the predictions from the candidate models capture the true smoothness of the test blocks. See Appendix A.5 on definition of such metric and its application using the simulation examples from Section 6.2.2.

## 6. Simulation studies

Now, we present multiple simulation studies focussing on different aspects of the proposed methods. In Section 6.1, we explore the computational requirements associated with implementation of the Lanczos algorithm. Sections 6.2 and 6.3 evaluate performance of the proposed spatial and spatiotemporal models, respectively, on simulated datasets.

### 6.1. Lanczos method implementation

Spatial and spatiotemporal models based on MSM and VA-MSM require computation of a selected number of eigenpairs of the $L$ matrix for (i) reduced-rank representation of the random effects and (ii) calculation of $L^+$ only in the latter. We do this computation using Implicit Restarted Lanczos with Leja Shifts with Free Left Endpoint (IRL-LSFLE) method from RSpectra (Qiu et al., 2016) R package. Mahdi et al. (2019) describes that, under certain assumption on sparsity of the neighborhood structure, it has $O(Sk^2)$ complexity for computing eigenpairs corresponding to $k$ smallest or largest eigenvalues of a $S \times S$ real symmetric sparse matrix. The algorithm converges fast when computing eigenpairs associated with eigenvalues having largest magnitudes. Consequently, instead of computing the eigenpairs for the $k$ smallest eigenvalues of the $L$, we opt to calculate the eigenpairs corresponding to the $k$ largest eigenvalues of the $(L - aI_S)^{-1}$ using the IRL-LSFLE with shift-and-invert mode where $I_S$ denotes $S \times S$ identity matrix. Here $a$ is a fixed real number and we choose $a = -0.001$ because $L$ has all non-negative eigenvalues. Since $L$ is singular, $a$ cannot be zero. Whereas RSpectra package uses $k$ Lanczos basis vectors by default, in this study, we use $3k$ basis vectors that will result in faster convergence although with higher memory usage. See Mahdi et al. (2019) for simulation studies that show how changes in grid size, neighborhood structure and value of $k$ influence the computation time and peak memory requirements.

Now, we discuss computation of $L^+$ for VA-MSM and its temporal extension. For the data analyses included in Sections 6.2.1, 6.3.1 and 7.2, we can compute the Moore–Penrose inverse exactly without using any approximation. For all other data analyses, the gridsizes are significantly larger and we replace $L^+$ using its stretch approximation from Section 5.1. As stated therein, we fix an upper bound $E_0$ of the approximation error and find the smallest integer $m_0$ such that $\lambda_{(m_0+1)} \geq \left(\frac{1}{\lambda_{(S)}} + 2\,\frac{E_0}{\lambda_{(2)}}\right)^{-1}$. Appendix A.6 describes how $m_0$ varies depending on the grid size, sparsity level in neighborhood and choice of $E_0$, as well as, how it can be computed. In this article, we set $E_0 = 10^{-4}$, used a first-order neighborhood structure (cells sharing an edge or even a corner point will be considered as neighbors), started by computing 4000 eigenpairs of $L$ at a time, and have always found $\lambda_{(m_0+1)}$ to satisfy the aforementioned inequality.

### 6.2. Simulations for spatial data

In this section, we conduct several simulation studies for spatial datasets. In Section 6.2.1, we focus on accuracy of parameter estimation from SM, MSM, and VA-MSM. Then, in Section 6.2.2, we showcase three synthetic data examples to assess and compare the predictive and computational performance of the MSM and VA-MSM against CAR model. All outputs presented in Sections 6 and 7 are based on running MCMC algorithms for 60,000 iterations, burning-in the initial 50,000 draws and thinning the rest at every 10-th iteration. While comparing the MCMC runtimes for different methods, it is essential to know about their respective computing environments, see Appendix A.7 for related information.

#### 6.2.1. Parameter estimation

For this study, we first generate 40,000 areal units by dividing the unit-square into identical gridcells and assign a first-order neighborhood structure to them. Now, we choose our simulation model for the response variable to have an intercept-only fixed effect part ($\beta_0 = 1$) added to the spatial random effects drawn from an improper CAR prior (utilizing the spectral structure of $L^+$) with $\tau^2 = 1$ and a pure error term with $\sigma^2 = 1$. We run this simulation model 30 times to generate as many datasets. Then, we fit each dataset with SM, MSM, and VA-MSM using $k \in \{10, 25, 50, 100, 200, 400\}$ eigenpairs and compared the 90% HPD intervals of three model parameters $(\beta_0, \tau^2, \sigma^2)$ from these runs. We find that $\beta_0$ has always been estimated with comparable accuracy but interesting differences emerge for inference related to values of $\sigma^2$ and $\tau^2$, as shown in Fig. 1.

For estimation of spatial variance parameter $\tau^2$, both MSM and VA-MSM exhibit satisfactory performance whereas SM shows significant overestimation. The reason for this behavior becomes evident by comparing the expression of spatial variance from Eqs. (3.1) and (3.2). The former expression representing the dispersion structure used for these simulations depends on all Laplacian eigenpairs. Since the latter expression (representing SM) only includes the first $k$ of those eigenvectors *without* the eigenvalues, $\tau^2$ has to get significantly larger to offset the absence of scaling effect of those small eigenvalues in the denominator. Increase in value of $k$ significantly reduces the overestimation bias for $\tau^2$ in SM. Patterns observed in estimation of nugget variance $\sigma^2$ follow the explanations provided in Section 3.1. The underestimation of spatial variance in SM and MSM is compensated by overestimation of nugget variance. In fact, the extent of similarity between the credible intervals for $\sigma^2$ produced by these two methods makes it difficult to separately identify their line plots throughout the top row of Fig. 1. Since VA-MSM is developed specifically to address this estimation bias, it produces HPD intervals covering the true parameter value. In all panels of Fig. 1, the HPD intervals from all methods get significantly narrower as we use more and more Laplacian eigenpairs. This is expected as well since the simulation model, i.e. the CAR prior, is based on all 40,000 Laplacian eigenpairs. As $k$ keeps increasing, the contribution of omitted eigenpairs becomes smaller and smaller. In summary, this simulation exercise very clearly illustrates the nature of improvement achieved by moving from SM to MSM and VA-MSM as we seek to use them to substitute the CAR prior. For all subsequent analysis in this article, we are going to use only MSM and VA-MSM (and their spatiotemporal extensions) for comparison against CAR-based approaches.

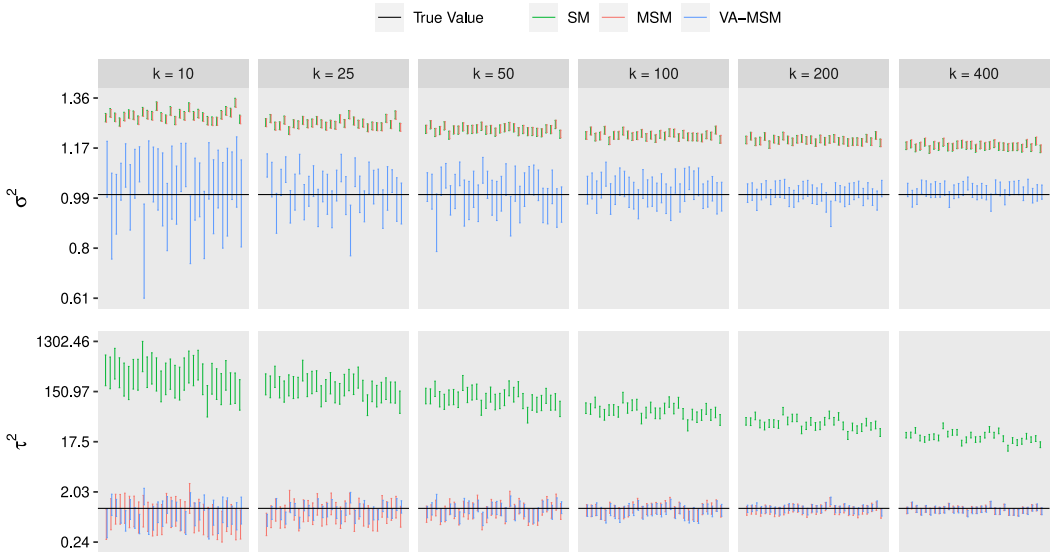



**Fig. 1.** Line plots of 90% HPD credible intervals for (top) $\sigma^2$ and (bottom) $\tau^2$, obtained from SM, MSM and VA-MSM for varying number of eigenpairs ($k$). The horizontal lines represent the corresponding true parameter values used in simulation.

*6.2.2. Performance comparison in general simulation models*

To explore robustness of proposed methods for different kinds of spatial datasets, we consider three simulation studies in this section. The first one utilizes a Gaussian linear spatial model but its correlation structure does *not* originate from a CAR prior. The second example considers a non-Gaussian highly nonlinear spatial function on an irregular grid. The third one uses the same function, although on a regular grid of significantly larger size.

For the first example, we consider 40,000 areal units by dividing $[0,5] \times [0,5]$ into identical gridcells. Similar to Section 6.2.1, we construct the response vector on this grid combining an intercept-only fixed effect part, a spatial random effect and a pure error term, with the latter two having unit variance each. However, the vector of spatial random effects now comes from a nonstationary GP based on spatially-varying mixture component kernels and a Matérn correlation function (smoothness = 2.5), see Paciorek and Schervish (2006) and the documentation for R package 'convoSPAT' (Risser and Calder, 2017) for relevant details on the nonstationary correlation model and its implementation.

For the second example, we move beyond the Gaussian framework and define the following nonlinear function over $(h_1, h_2) \in [0,1] \times [0,1]$:

$$\textbf{Function} - \textbf{1} : f_1(h_1, h_2) = 10 + 15\,\frac{\ln(1+h_1)}{1+h_2^2} + 11 \sin 2\pi h_1 \cos 2\pi h_2 + 5h_1^2(1-h_2^2),$$

As our choice of irregular grid, we consider the US census tracts, extracted from R package 'tigris' (Walker, 2025) for the default year 2024. We define the neighborhood structure by identifying census tracts with adjacent boundaries. Extracting the largest connected component consisting of 83,280 tracts, we standardize their centroids within unit-square before evaluating Function-1 at those points. Fig. 2 presents the two spatial maps generated from aforementioned simulations.

As mentioned in Section 5.2, we begin by computing the optimal number of eigenpairs ($k$) for the MSM and VA-MSM using single block spatial CV for each of these two datasets. For the nonstationary GP dataset, a block of dimensions $60 \times 60$($\sim 9\%$ of all gridcells) is considered as the test data, and both MSM and VA-MSM are fitted on the remaining training data for $k \in \{10, 25, 50, 100, 200, 400\}$. We repeat the process for Function-1 on irregular grid as well, albeit with one change. As contiguous US has a longer east–west stretch, we consider rectangular test blocks (east–west span of the block almost twice its north–south span) containing $\sim$ 8%–9% of the total number of tracts. Both MSM and VA-MSM are fitted on the remaining training data points for $k \in \{50, 100, 200, 400, 600, 800\}$. After replicating this process with 30 different choices for the test block in each dataset, we compute average MSE and MCMC runtime. Fig. 3 represents these findings for all candidate values of $k$ using MSM and VA-MSM for both datasets.

In all these plots, computation time increases with the increase in the value of $k$, as expected. For nonstationary GP dataset, use of $k = 200$ eigenpairs leads to the lowest MSE for VA-MSM. For MSM, the decrease in MSE for increasing $k$ to 400 from 200 is nearly negligible to justify the additional runtime. Hence, we choose $k = 200$ for both methods. Shifting to the plots for Function-1 on irregular grid, $k = 600$ and $k = 200$ are chosen for MSM and VA-MSM, respectively, as suggested by the respective MSE trends.

Using these selected values of $k$, we now examine predictive performances of MSM and VA-MSM via cross-validation on the above-mentioned test blocks, based on the measures specified in Section 5.2. For comparison purpose, we also include similar performance metrics obtained from the CAR prior as well as INLA (Martins et al., 2013; Lindgren and Rue, 2015, https://www.r-inla.org/). Table 1 summarizes these findings for the single block and the three blocks spatial CV. For the latter, each of the three

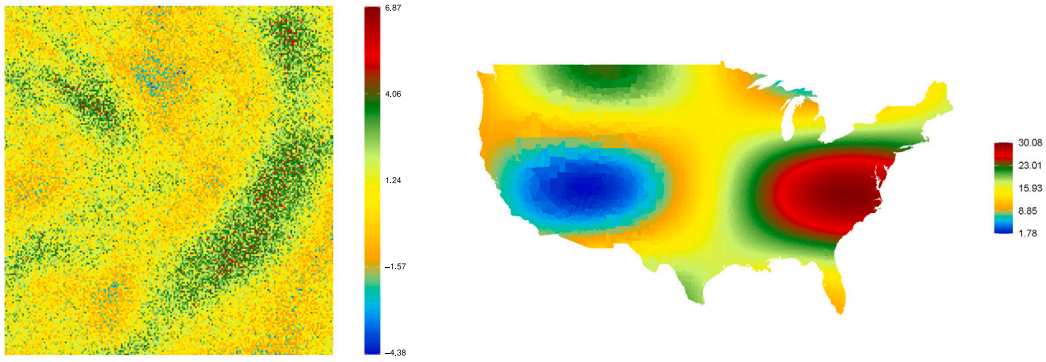



**Fig. 2.** Maps of spatial data simulated using (left) nonstationary GP on regular grid and (right) Function-1 on irregular grid.

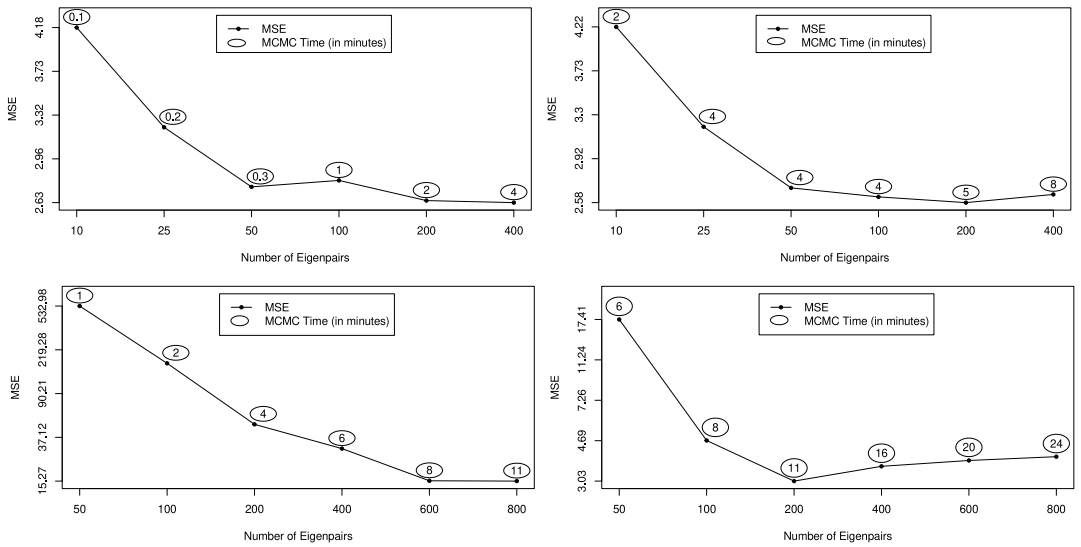



**Fig. 3.** MSE and MCMC runtimes using (left column) MSM and (right column) VA-MSM for (top row) nonstationary GP and (bottom row) Function-1 on irregular grid.

**Table 1**
Cross validation-based predictive comparison for simulated spatial datasets.

| Models | Single block spatial CV | | | Three blocks spatial CV | | |
|---|---|---|---|---|---|---|
| | Absolute bias | Uncertainty | Empirical coverage | Absolute bias | Uncertainty | Empirical coverage |
| Nonstationary GP on Regular Grid | | | | | | |
| CAR | 0.960 | 3.443 | 0.845 | 0.863 | 3.411 | 0.882 |
| INLA | 0.965 | 3.448 | 0.843 | 0.864 | 3.414 | 0.882 |
| MSM | 0.960 | 3.522 | 0.855 | 0.865 | 3.369 | 0.878 |
| VA-MSM | 0.951 | 3.465 | 0.853 | 0.861 | 3.347 | 0.877 |
| Function-1 on Irregular Grid | | | | | | |
| CAR | 1.807 | 0.563 | 0.204 | 0.561 | 0.494 | 0.485 |
| INLA | 8.754[a] | 0.564 | 0.301 | 8.328[a] | 0.525 | 0.528 |
| MSM | 2.377 | 1.086 | 0.304 | 1.285 | 0.797 | 0.454 |
| VA-MSM | 1.025 | 1.202 | 0.525 | 0.609 | 1.187 | 0.729 |

[a] See related explanation in text.

non-contiguous blocks within a test set consists of 35 × 35 gridcells (for nonstationary GP dataset) and ∼ 2.5% of total number of tracts (for Function-1 on irregular grid) and selection of the test set is replicated 30 times.

**Table 2**
Computational assessment for simulated spatial datasets.

| Models | Single block spatial CV | | | | | |
|---|---|---|---|---|---|---|
| | Nonstationary GP | | | Function-1 on irregular grid | | |
| | $S = 40,000$ | | | $S = 83,280$ | | |
| | MCMC runtime | ESS | Relative Efficiency | MCMC runtime | ESS | Relative Efficiency |
| CAR | 14 | 0.268 | – | 27 | 0.412 | – |
| MSM | 2 | 1 | 26.119 | 8 | 0.950 | 7.782 |
| VA-MSM | 5 | 0.017 | 0.178 | 11 | 0.085 | 0.506 |

INLA does not use MCMC. It took ∼1 min to converge for each dataset.

**Table 3**
Comparison of VA-MSM performance with and without Marginalization.

| VA-MSM marginalization | Single block spatial CV | | | | | | Three blocks spatial CV | | |
|---|---|---|---|---|---|---|---|---|---|
| | Predictive performance | | | Computational assessment | | | Predictive performance | | |
| | Absolute bias | Uncertainty | Empirical coverage | MCMC time (in minutes) | ESS proportion | Rel. Time efficiency | Absolute bias | Uncertainty | Empirical coverage |
| Nonstationary GP | | | | | | | | | |
| No | 0.951 | 3.465 | 0.853 | 5 | 0.017 | 0.178 | 0.861 | 3.347 | 0.877 |
| Yes | 0.953 | 3.461 | 0.851 | 45 | 1 | 1.166 | 0.860 | 3.350 | 0.878 |
| Function-1 on Irregular Grid | | | | | | | | | |
| No | 1.025 | 1.202 | 0.525 | 11 | 0.085 | 0.506 | 0.609 | 1.187 | 0.729 |
| Yes | 1.082 | 1.242 | 0.550 | 84 | 0.934 | 0.729 | 0.611 | 1.207 | 0.735 |

Notably, the performance of all four methods are nearly identical for nonstationary GP dataset. This is due to the fact that its spatial random effects still come from a multivariate normal joint distribution, although with a different correlation structure. Hence, fitting this dataset with CAR and related prior distributions amounts to a relatively modest misspecification causing the empirical coverage rates to fall marginally below the target level of 90%. As expected, with smaller test blocks in three blocks spatial CV, the performance improves for all candidate models in terms of lower bias and uncertainty and higher coverage rates.

Shifting our focus to Function-1 on irregular grid, we see a considerable worsening of the predictive performance across all models and none of them could attain desired level of 90% coverage. This is anticipated due to highly non-linear, non-Gaussian nature of the spatial pattern in the simulated surface. In the single block segment, CAR and INLA exhibit significantly lower uncertainty compared to the our proposed methods. However, both MSM and VA-MSM show marked improvement in the coverage rate compared to that of CAR implying wider predictive intervals have actually been beneficial. Specifically, VA-MSM achieves significantly better coverage than INLA. In the three block segment, all models expectedly improve their coverage rates with VA-MSM again outperforming the other three by a large margin with over 70% out-of-sample coverage. We found that unusually high values of absolute bias from INLA in the single- and three-blocks spatial CV are artifacts of its poor performing in *only one* of the test sets in each case. Excluding those sets results in triplets of $(1.193, 0.566, 0.311)$ and $(0.907, 0.527, 0.547)$ in the single- and three-blocks spatial CV, respectively. Comparison of performance metric between the upper and lower blocks of Table 1 reveals that the predictive performance of INLA falls significantly below that of VA-MSM when we switch to a highly non-linear, non-Gaussian spatial functions (Since the former uses approximate posterior distributions to fast inference, it may not work efficiently with datasets of that kind.).

We move to computational assessment of proposed methods for these two datasets. In Table 2, MSM turns out to be highly efficient compared to CAR due to much shorter runtime and higher ESS. VA-MSM runs significantly faster than CAR but exhibits very low ESS due to presence of large number of parameters in $\ddot{\varphi}$ resulting in diminished relative efficiency. INLA does not use MCMC and, as expected, has the shortest runtime.

To address the low ESS in VA-MSM, we explore an alternative estimation scheme by marginalizing $\ddot{\varphi}$ out. In Table 3, we compare the original and marginalized sampling methods in terms of predictive and computational measures. Although marginalization drastically improves ESS as anticipated, its considerably longer runtime offsets its utility to some extent. This is due to the fact that margination brings a dispersion structure (see Appendix A.2 on the sampling distributions for marginalized VA-MSM) that necessitates several computationally challenging calculations at every iteration which could be avoided (or precomputed before MCMC) in the original VA-MSM estimation scheme. Since the predictive performance of both variants of VA-MSM match each other closely for both datasets, we decide to use the former (not marginalized) version for all subsequent analysis in this article owing to its significantly shorter runtime.

For the final simulation study in this section, we construct a very large regular $450 \times 450$ grid ($S = 202,500$) within the unit square and reuse Function-1 on this grid. The left panel of Fig. 4 presents the spatial map for this dataset.

As with previous simulation studies, we begin by computing the optimal number of eigenpairs ($k$) for the MSM and VA-MSM using single block spatial CV. In this cross-validation, a block of dimensions $130 \times 130$ ($\sim 8.3\%$ of all gridcells) is considered as the

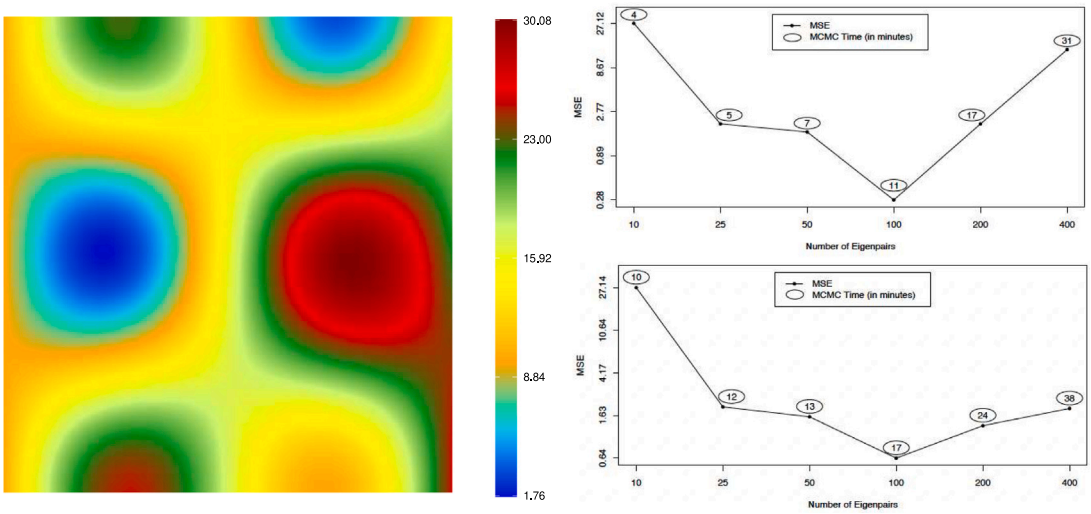



**Fig. 4.** (left) Map of spatial data simulated using Function-1 on a large, regular grid. (right) MSE and MCMC runtimes using (top) MSM and (bottom) VA-MSM.

**Table 4**
Cross validation-based model comparison for Function-1 on Large, Regular Grid.

| Models | Single block spatial CV | | | | | | Three blocks spatial CV | | |
|---|---|---|---|---|---|---|---|---|---|
| | Predictive performance | | | Computational assessment | | | Predictive performance | | |
| | Absolute bias | Uncertainty | Empirical coverage | MCMC time (in minutes) | ESS proportion | Rel. Time efficiency | Absolute bias | Uncertainty | Empirical coverage |
| CAR | 1.879 | 0.412 | 0.142 | 59 | 0.358 | – | 0.611 | 0.356 | 0.287 |
| MSM | 0.376 | 0.622 | 0.535 | 11 | 1.000 | 15.000 | 0.231 | 0.616 | 0.752 |
| VA-MSM | 0.532 | 0.859 | 0.565 | 17 | 0.074 | 0.718 | 0.222 | 0.860 | 0.854 |

test data, and both MSM and VA-MSM are fitted on the remaining training data for $k \in \{10, 25, 50, 100, 200, 400\}$. After replicating this process with 30 different choices for the test block, we compute average MSE and MCMC runtime. The right panel of Fig. 4 represents these findings for all candidate values of $k$ using MSM and VA-MSM. It can be seen that MSE becomes the smallest for $k = 100$ with reasonable computation time in both MSM and VA-MSM. Therefore, we select $k = 100$ for both MSM and VA-MSM to further analyze this dataset. Notably, these chosen values for $k$ are different from the ones we selected based on Fig. 3 where we used Function-1 on an irregular grid. This shows that, even for the same function, the difference in adjacency structure (regular vs irregular) may change the spectral properties of the Laplacian matrix yielding different outcomes for optimal number of eigenpairs.

Using the chosen values of $k$, the predictive and computational performances of CAR, MSM and VA-MSM are now examined via cross-validation, based on the measures specified in Section 5.2. Table 4 summarizes these findings for the single- and three-block spatial CV. For the latter, each of the three non-contiguous blocks within a test set consists of $80 \times 80$ gridcells and selection of the test set is replicated 30 times.

CAR exhibits the lowest uncertainty with the largest absolute bias and, more importantly, very low coverage rate. Both of the proposed methods show marked improvement in coverage rates (approximately 3- to 4-fold rise) along with smaller bias, implying wider predictive intervals have actually been beneficial. However, none of the three models could attain desired level of 90% coverage which is anticipated due to highly nonlinear, nonGaussian nature of the function. Computation-wise, MSM turns out to be significantly more efficient compared to CAR due to much smaller runtime and higher ESS. VA-MSM, employed in its non-marginalized version, yields very low ESS as expected. All methods exhibit reduced bias and improved coverage rate when we move to three-block spatial CV owing to smaller size of individual blocks. It should also be noted that, compared to the simulation study of Function-1 on irregular grid (see lower block of Table 1), the empirical coverage rates of MSM show very strong improvement (margin of improvement relatively modest for VA-MSM) in the regular grid version.

We complete this study presenting a visual comparison based on spatial pattern of the predicted surfaces for four selected test blocks from the single block spatial CV. In Fig. 5, we present the true test surfaces vs their posterior median estimates from each of CAR, MSM, and VA-MSM for Function-1 on large, regular grid. In this plot matrix, the columns correspond to different test blocks and the locations of these blocks inside the entire region are shown in the topmost square.

The first and third blocks in Fig. 5 show that the CAR model makes poor predictions by over and/or underestimating throughout these blocks and generates spurious spatial patterns. On the other hand, spatial patterns in the true data is captured well by the MSM and VA-MSM, although the former performs better in predicting the shape of the blue patch in the first block and red region

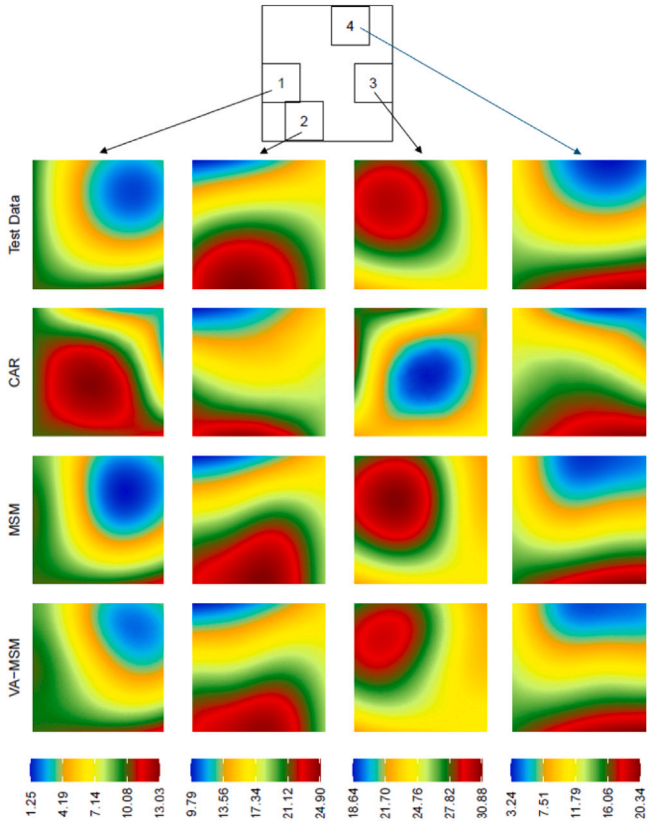



**Fig. 5.** Spatial map of true data vs posterior median using CAR, MSM and VA-MSM for test blocks taken from single block spatial CV for Function-1 on large, regular grid.

of the third block. However, none of these models correctly predicts the shape of (i) the red region at the southern part in the second block and (ii) the blue region at the north-eastern part in the fourth block. Such small discrepancies are observed in other test blocks as well. Overall, these spatial maps are coherent with the very low coverage rate exhibited by CAR in Table 4 relative to its competitors.

### 6.3. Simulations for spatiotemporal data

We expand the simulation studies to spatiotemporal setting. In Section 6.3.1, we examine the accuracy of parameter estimation. In Section 6.3.2, we compare ARCAR, ARMSM, and ARVA-MSM models with respect to the predictive performance and computational efficiency.

#### 6.3.1. Parameter estimation

We compare ARMSM and ARVA-MSM with respect to accuracy of parameter estimation from spatiotemporal data simulated using ARCAR prior. For this purpose, we divide unit-square into identical grid cells of dimensions $200 \times 200$ and consider 50 time points. We simulate 30 data sets from Eq. (4.1) using (i) intercept-only fixed part with $\beta_0 = 1$, (ii) the spatiotemporal random effect following an ARCAR prior with first-order neighborhood structure, $\eta = 0.6$ and $\tau^2 = 1$ and, (iii) a pure error term with $\sigma^2 = 1$. We fit the ARMSM and ARVA-MSM on each of the 30 data sets and construct the 90% HPD intervals of these parameters using $k \in \{10, 25, 50, 100, 200, 400\}$-many Laplacian eigenpairs. Fig. 6 presents the line plots of the 90% HPD intervals with an increasing order of the number of eigenpairs.

Similar to the spatial models, we do not show the output for $\beta_0$ because both models estimate $\beta_0$ correctly. The line plots show that the ARVA-MSM estimates all these parameters correctly. On the other hand, the ARMSM model estimates $\tau^2$ and $\eta$ correctly but overestimates the nugget variance as expected. Also, with the increase in $k$, the intervals become narrower. As $k$ gets larger, the ARMSM marginally underestimates the value of $\tau^2$ potentially affecting estimation of $\eta$ as well. It is no surprise that performance of ARVA-MSM is most satisfactory for parameter estimation as it was specifically developed to bridge the difference between the dispersion structures of ARCAR and ARMSM prior, the former being utilized in this simulation model.

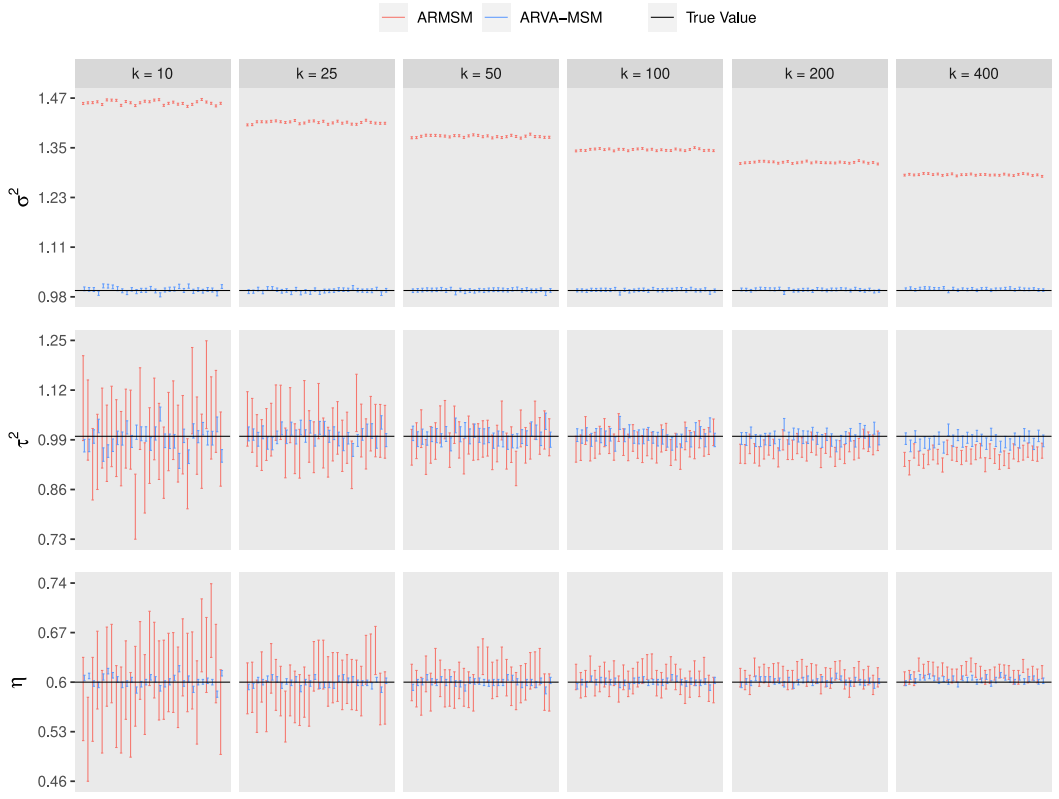



**Fig. 6.** Line plots of 90% HPD credible intervals for (top) $\sigma^2$, (middle) $\tau^2$ and, (bottom) $\eta$, obtained from ARMSM and ARVA-MSM for varying number of eigenpairs (k). The horizontal lines represent the corresponding true parameter values used in simulation.

### 6.3.2. *Performance comparison*

Now, we compare the predictive performance and computational efficiency between ARCAR, ARMSM, and ARVA-MSM. To do this, we consider (i) $T = 25$ time points and (ii) a $300 \times 300$ grid within the unit square and simulate data from the following function:

$$\textbf{Function} - \textbf{2}: \quad f_2(h_1, h_2, t) = a_t \frac{\ln(1+h_1)}{1+h_2^2} + b_t \sin 2\pi h_1 \cos 2\pi h_2 + c_t\, h_1^2(1-h_2^2),$$

$$a_t = 15\frac{\sqrt{1+t/T}}{\ln(2+t/T)} + \mathcal{N}(0,1), \quad b_t = 5\,|\cos 2\pi(1+(1-t)/T)| + \mathcal{N}(0,1),$$

$$c_t = 5\,\ln(2+(1-t)/T) + \mathcal{N}(0,1)$$

where $(h_1, h_2) \in [0,1] \times [0,1]$ and $t \in \{1, 2, \ldots, T\}$. Fig. 7 displays this data as temporal snapshots of spatial surfaces.

As in Section 5.2, we start with single block spatial CV to determine the optimal number of eigenpairs, $k$, for ARMSM and ARVA-MSM. We select a block of dimensions $65 \times 65$ and consider the observations from that block across all time points as the test data ($\sim 4.7\%$ of the total number of observations) for cross-validation. We replicate this procedure 30 times. We fit the ARMSM and ARVA-MSM on each training dataset for $k \in \{5, 10, 15, 20, 30\}$, compute average MSE and MCMC runtime. These outputs are displayed in Fig. 8.

In both ARMSM and ARVA-MSM, MSE attains its minimum at $k = 15$. Within each model, the runtimes do not vary significantly for different choices of $k$ since the candidate values are relatively close to each other to make a notable difference in runtimes. We select $k = 15$ for ARMSM and ARVA-MSM in the subsequent analysis.

Now we examine the computational and predictive performance of the ARCAR, ARMSM and ARVA-MSM through cross-validation using the measures specified in Section 5.2. We use the same test blocks used to select the optimal value of $k$ for ARMSM and ARVA-MSM. Table 5 summarizes these outputs within the single block spatial CV segment.

As seen with earlier simulation examples, CAR-based model exhibits the lowest uncertainty with significantly larger bias and, more importantly, very low coverage rate. Compared to that, both of the proposed methods show very strong improvement in coverage rates along with smaller bias. However, none of the three models could attain desired level of 90% coverage. Computation-wise, ARMSM exhibits very strong relative efficiency owing to the best ESS proportion as well as runtimes that are more than ten

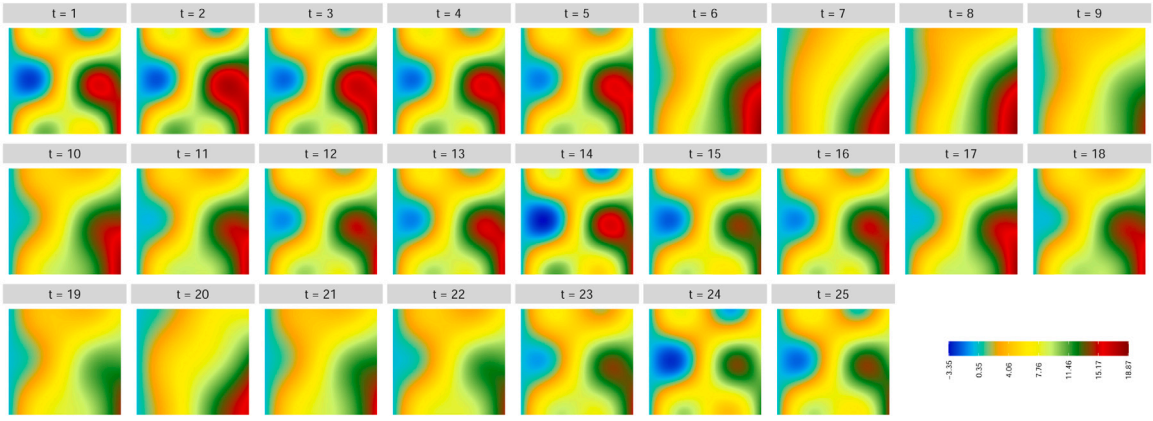



**Fig. 7.** Maps of spatiotemporal data simulated using Function-2.

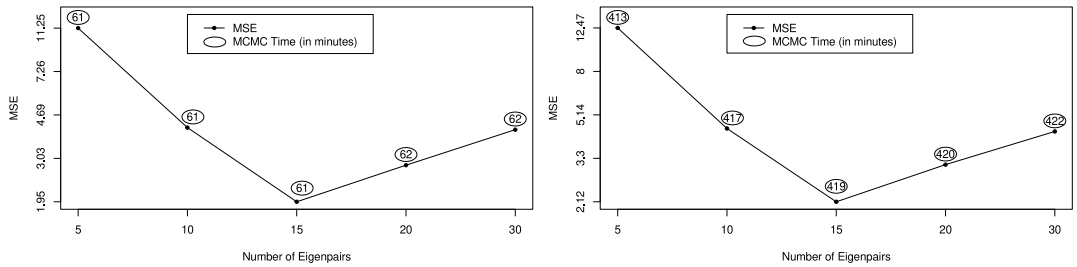



**Fig. 8.** MSE and MCMC runtimes for (left) ARMSM and (right) ARVA-MSM for simulated dataset from Function-2.

**Table 5**
Cross validation-based model comparison for simulated spatiotemporal dataset.

| Models | Single block spatial CV | | | | | | Three blocks spatial CV | | |
|---|---|---|---|---|---|---|---|---|---|
| | Predictive performance | | | Computational assessment | | | Predictive performance | | |
| | Absolute bias | Uncertainty | Empirical coverage | Duration (in minutes) | ESS proportion | Rel. Time efficiency | Absolute bias | Uncertainty | Empirical coverage |
| ARCAR | 2.914 | 0.168 | 0.017 | 718 | 0.342 | – | 2.410 | 0.173 | 0.022 |
| ARMSM | 0.976 | 1.991 | 0.619 | 61 | 1.000 | 34.462 | 0.592 | 2.023 | 0.864 |
| ARVA-MSM | 1.010 | 2.091 | 0.621 | 419 | 0.305 | 1.531 | 0.605 | 2.124 | 0.871 |

times shorter than the same for ARCAR, as expected. ARVA-MSM, both runtime- and ESS-wise, falls short of ARMSM while still outperforming ARCAR in terms of runtime and relative efficiency.

We also perform three blocks spatial CV. Here, in each of the 30 replications, we consider three disjoint small blocks each with dimensions $40 \times 40$ and treat all observations from these blocks across all time points as the test data. The summary outputs of the same three predictive performance comparison measures are shown in three blocks spatial CV segment of Table 5. Still ARCAR performs very inefficiently with a large bias and very low coverage. Both of the proposed methods exhibit coverage rates very close to desired level of 90% along with smaller absolute bias, implying wider predictive intervals have actually been beneficial.

We complete this study by presenting (Fig. 9) a visual comparison of the spatial patterns in the posterior median predicted surfaces for four selected test blocks, each at a different time point, extracted from the single block spatial CV output. The topmost square therein displays the locations of the four test blocks within the entire region and the maps for each block-time combination occupy a specific column in the plot matrix.

All columns generally show that the maps predicted by proposed methods resemble, in terms of spatial patterns, the true maps much better compared to those from ARCAR. However, a closer look reveals existence of specific regions or patterns in these blocks, where models mis-predict to different extent. For the first block, the blue region gets overestimated by all models whereas the red region is underestimated by ARMSM and ARVA-MSM to some extent. In the second block, ARMSM and ARVA-MSM capture the patterns of the data although the former makes the red and blue regions thicker and thinner, respectively, to some extent. On the other hand, ARCAR only captures some part of the blue region and it mostly over- or under-predicts in the remaining part of the block. In the third and the fourth blocks, ARCAR creates spurious spatial patterns completely absent in the true maps. ARMSM and ARVA-MSM capture the spatial patterns well in these two blocks although, in the fourth block, the red and blue patches in the northern section are predicted to be a bit wider than the actual by the both of these methods.

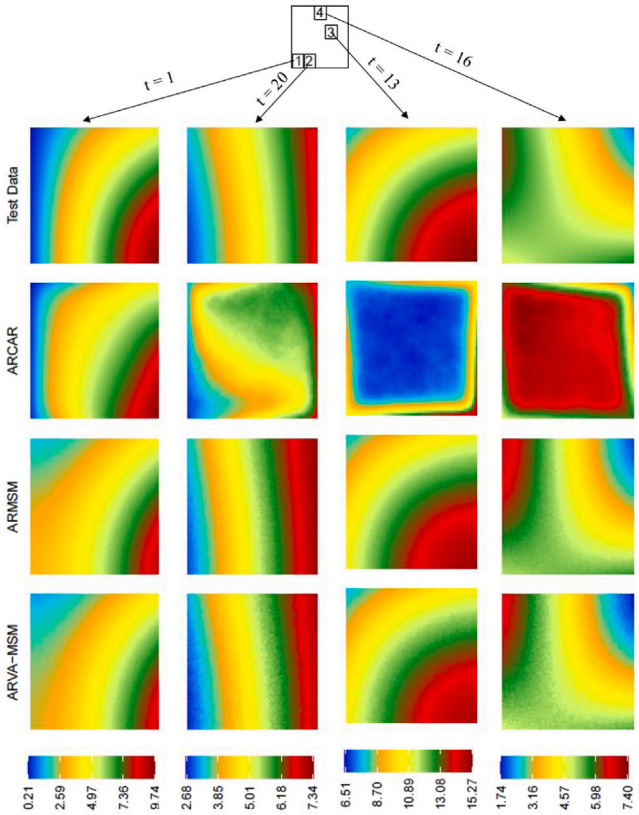



**Fig. 9.** Maps of true data vs posterior median using ARCAR, ARMSM and ARVA-MSM for four test block-time combinations, taken out of single block spatial CV for Function-2.

## 7. Real data analysis

We turn to real-world datasets to illustrate the use of proposed spatial and spatiotemporal models. First, we assess them in spatial setting analyzing median home values across US census tracts in Section 7.1. Next, in Section 7.2, we conduct a comparative analysis in spatiotemporal setting using Landsat imagery data.

### *7.1. Spatial analysis of US median home value data*

As an illustrative example, we choose American Community Survey (ACS) 5-year median home value by census tract data that can be accessed using tidycensus (Walker et al., 2023) R package. Our spatial domain is the contiguous USA (excluding Alaska and Hawaii), and the 5-year period is chosen to be 2013–2017. Walker (2022) previously analyzed 5-year median home value data for Dallas-Fort Worth metropolitan area. Following their work, we model median home value in log scale and start with the same set of tract-specific predictors: median number of rooms for homes, total population, median age of the population, median income of households, percentage of the population age 25 and up with a four-year college degree, percentage of the population born outside the United States, percentage of the population that identifies as non-Hispanic white, percentage of owner-occupied housing units, population density (measured as the ratio of total population and the area of the corresponding tract) and, median age of housing structures (computed by subtracting median year built of housing structures from the mid-point of the 5-year ACS period). We remove 2800 tracts where at least one of the predictors are missing. Out of the remaining 69738 census tracts, 69625 tracts are connected (determined using igraph (Csárdi et al., 2024) R package) and we only consider this set in the following analysis. Fig. 10 displays the map of log median home value data.

Before applying the spatial models, we perform exploratory analysis. We first standardized all covariates within $[0, 1]$ range and check for multicollinearity based on the variance inflation factor (VIF) computed using R package car (Fox et al., 2012). We had to remove only one covariate, the total population, based on a VIF value above 5. Then, we run a linear regression to identify and remove insignificant predictors based on 5% significance level. In our case, the only insignificant covariate to be excluded was population density.

Now, we determine the optimal number of eigenpairs, $k$, for MSM and VA-MSM based on cross-validation. As contiguous US has a longer east–west stretch, we consider rectangular test block (east–west span of the block almost twice its north–south span)

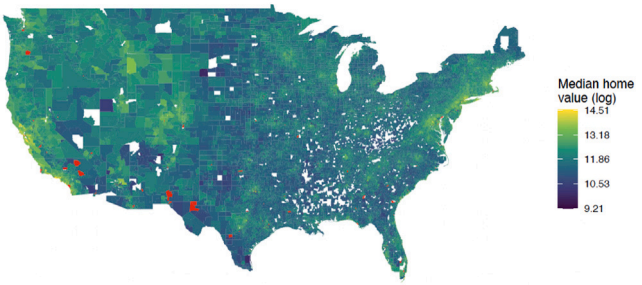



**Fig. 10.** Map of 5-year median home value (in log scale) by US census tract during 2013–2017. The red patches represent 907 tracts with missing response. The white patches represent 2800 tracts with one or more missing covariates that we discard from analysis.

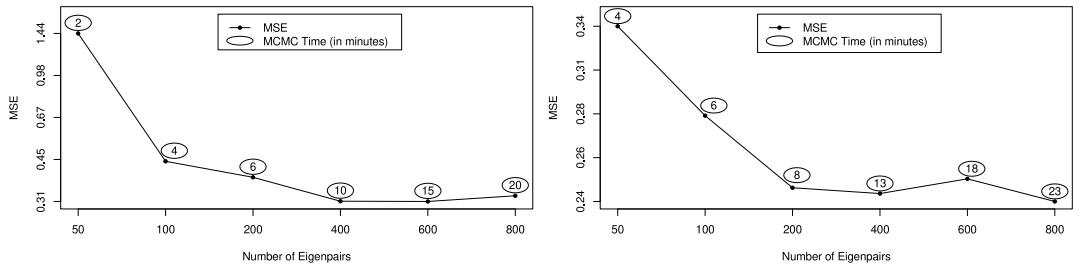



**Fig. 11.** MSE and MCMC runtimes using (left) MSM and (right) VA-MSM for log median home value by US census tract.

**Table 6**
Spatial model comparison for (log) median home value by US Census Tract.

| Models | Single block spatial CV | | | | | | Three blocks spatial CV | | |
|---|---|---|---|---|---|---|---|---|---|
| | Predictive performance | | | Computational assessment | | | Predictive performance | | |
| | Absolute bias | Uncertainty | Empirical coverage | MCMC time (in minutes) | ESS proportion | Rel. Time efficiency | Absolute bias | Uncertainty | Empirical coverage |
| CAR | 0.305 | 1.035 | 0.824 | 24 | 0.288 | – | 0.253 | 1.003 | 0.880 |
| MSM | 0.332 | 1.102 | 0.821 | 10 | 1.275 | 10.637 | 0.299 | 1.021 | 0.828 |
| VA-MSM | 0.294 | 0.990 | 0.825 | 8 | 0.022 | 0.233 | 0.265 | 0.974 | 0.859 |

containing around 8–9% of the total number of tracts. This process was replicated 30 times and, for each replication, MSM and VA-MSM are implemented on the corresponding training dataset with $k \in \{50, 100, 200, 400, 600, 800\}$ eigenpairs. Since we model the median home value in log scale, all of the following outputs are presented in that scale as well. Fig. 11 shows the average MSE across the replications and MCMC runtime for each choice of $k$.

MSM shows a gradual decline in the MSE as $k$ increases to 400. After that, MSE marginally increases and the computational cost increases significantly. For VA-MSM, MSE declines sharply up to $k = 200$. After that, the decrease in the MSE becomes too small to justify the additional computational cost. Therefore, we consider $k = 400$ and $k = 200$ in MSM and VA-MSM, respectively, for the following analysis.

The predictive and computational performances of CAR, MSM and VA-MSM are now examined via cross-validation using the above-mentioned test blocks for both functions. Measures specified in Section 5.2 are computed and averaged across test blocks for a comparative evaluation. Table 6 summarizes these diagnostics within the single block spatial CV segment.

CAR and VA-MSM show comparable predictive performance with VA-MSM achieving the smallest uncertainty. Although the performance of MSM is not on par with others, the values are not significantly different from the corresponding values in CAR and VA-MSM. The models achieved comparable coverage levels. Computation-wise, MSM outperformed the other methods due to low runtime and high ESS. VA-MSM, although about three times faster than CAR, again suffered from a very low ESS value.

Next, we perform three blocks spatial CV. For this purpose, we consider three disjoint small rectangular blocks, each with $\sim$ 2.5–3% of the total number of tracts as the test data in the cross-validation. We replicate this process 30 times. The predictive performance is compared based on the same three measures used in single block spatial CV. These outputs are presented in the three block spatial CV segment of Table 6. For all methods, decrease in bias is observed compared to the corresponding values in the single block version along with modest increase in coverage rate as well.

We complete this study presenting a visual comparison based on spatial pattern of the predicted surfaces for four selected test blocks. In Fig. 12, the topmost map displays the locations of the four test blocks within contiguous US. We extracted one of these blocks from single block spatial CV and other three from different replications of three blocks spatial CV. In the plot matrix, each

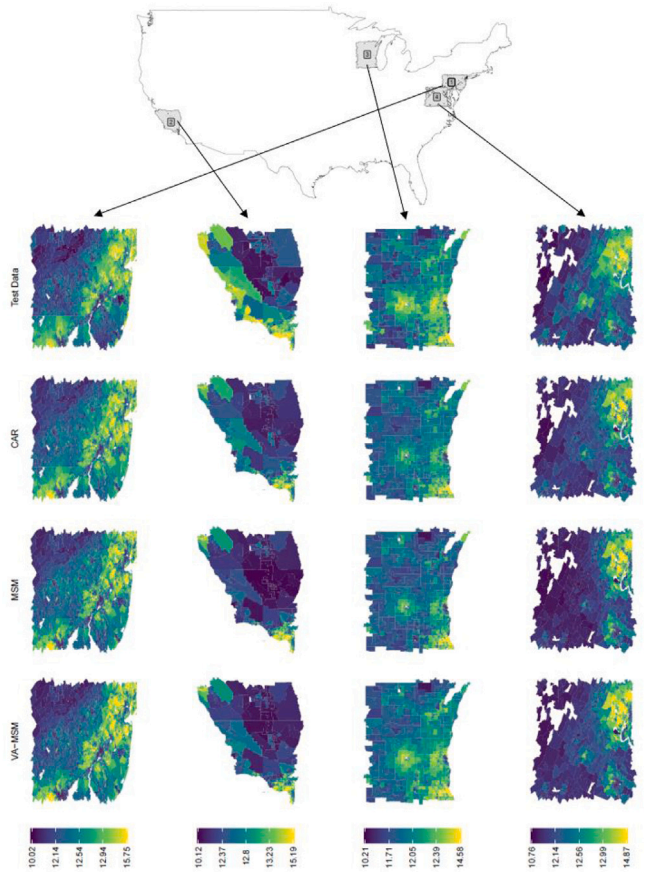



**Fig. 12.** Spatial map of true (log) median home value data vs posterior median using CAR, MSM and VA-MSM for test blocks chosen from (column 1) single block spatial CV and (columns 2-4) three blocks spatial CV.

column represents one block's true test surface and the corresponding posterior median estimate from each of CAR, MSM, and VA-MSM. The predicted surfaces from the candidate methods look very comparable to each other and they broadly retain the spatial patterns exhibited by the true test data. However, a closer inspection reveals certain areas of under-performance such as the western side of the second test block which has been underestimated by all models.

### *7.2. Spatiotemporal analysis of Landsat imagery data*

Tasseled Cap (TC) features, computed using reflectance detected from Earth's surface by satellites, are known to be useful for mapping land cover, forest characteristics and biomass (Zheng et al., 2014; Karlson et al., 2015). One of the TC features is brightness that we choose to analyze as an illustrative application for the proposed methods in a spatiotemporal setting. Our region of interest spans a square-shaped area of 230,400 hectares located near Langlade, Shawano, and Menominee counties in north-central Wisconsin, USA. The dataset on brightness values for this region comes from Landsat 7 ETM+ imagery, collected during the decade 2003–2012 as part of the Web-Enabled Landsat Data (WELD) project for the contiguous United States and Alaska. WELD imagery is mosaicked into pixels at the native 30-meter resolution. Our study region, a part of WELD tile H20V05 (Wilson, 2015), consists of $1600 \times 1600$ such pixels. The left panel of Fig. 13 shows the study region within the WELD tile and the state of Wisconsin. Temporally, the data are available as monthly composite values over the 10-year span. Khan et al. (2021) utilized the TC features dataset from this region for forest inventory analysis aggregating every $16 \times 16$ adjacent pixels into a single gridcell. In our study, we adopt a finer $8 \times 8$ aggregation to obtain a spatiotemporal dataset with 40,000 gridcells and 120 time points.

This dataset has a significant proportion of missing values ($\sim 26\%$) with 100% missing values at six months (April 2004, June 2003, November 2008, and December of 2006, 2011, 2012) due either to missing records in the data repository or to pixels that were masked out because of the presence of clouds, snow or artifacts of sensor failure. The temporal and spatial patterns of missingness are presented in the center and right panels, respectively, of Fig. 13.

To fill in the missing positions, Khan et al. (2021) evaluated the performance of a group of spatiotemporal models for logarithm of brightness. Their chosen model includes a fixed effect, a first-order autoregressive temporal effect and, a temporally independent

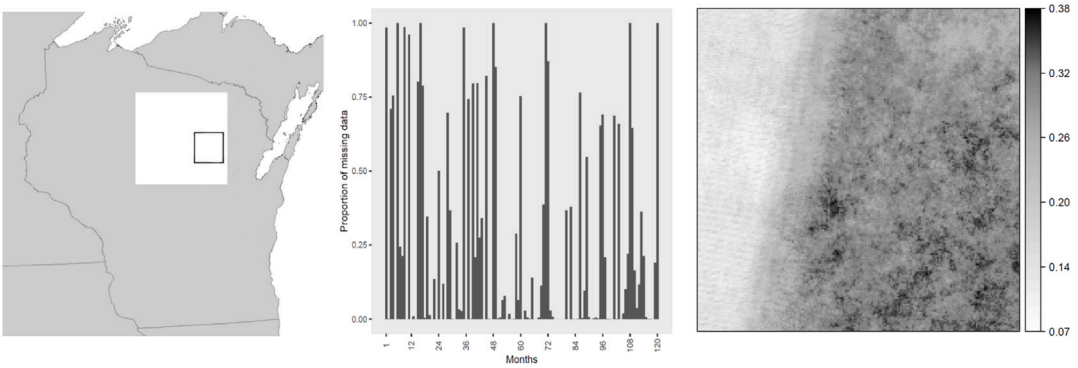



**Fig. 13.** (left) The study region (smaller square) is shown inside the WELD H20V05 tile (larger square) within the map of Wisconsin. Proportion of missing brightness data in the study area during 2003–2012 is displayed: (center) temporally, for each month and (right) spatially, for each gridcell.

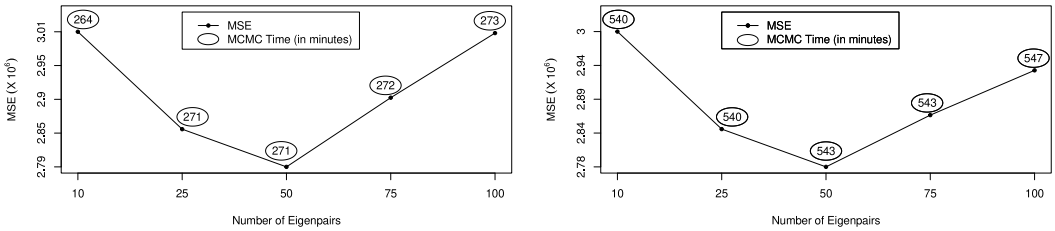



**Fig. 14.** MSE and MCMC runtimes for (left) ARMSM and (right) ARVA-MSM for spatiotemporal data on brightness with $\eta = 0$. Although brightness is modeled in logarithmic scale, MSE outputs are shown in the original scale.

**Table 7**
Cross validation-based model comparison for spatiotemporal brightness data.

| Models | | Single block spatial CV | | | | | | Three blocks spatial CV | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Predictive performance | | | Computational assessment | | | Predictive performance | | |
| | | Absolute bias | Uncertainty | Empirical coverage | MCMC time (in minutes) | ESS proportion | Rel. Time efficiency | Absolute bias | Uncertainty | Empirical coverage |
| ARCAR | $\eta = 0$ | 797.016 | 2471.436 | 0.821 | 1038 | 0.016 | – | 742.630 | 2425.593 | 0.839 |
| ARMSM | | 793.283 | 3057.888 | 0.881 | 271 | 0.855 | 199.283 | 765.675 | 3032.621 | 0.888 |
| ARVA-MSM | | 793.283 | 3043.705 | 0.880 | 543 | 0.012 | 1.425 | 765.975 | 3019.892 | 0.886 |

sequence of spatial processes with CAR prior. The third component fits within the ARCAR framework that we discussed in Section 4 if we fix the AR coefficient $\eta = 0$. We retain their chosen model and explore replacing the ARCAR prior for the third component with ARMSM and ARVA-MSM. For each of them, we implement single block spatial CV to determine the optimal number of eigenpairs, $k$. We select a block of dimensions $40 \times 40$ and consider all available measurements from that block across all time points as the test data (between 2.7%–3.5% of the total number of available observations) for cross-validation. This procedure is replicated 30 times. We fit the ARMSM and ARVA-MSM on each training dataset for $k \in \{10, 25, 50, 75, 100\}$, compute average MSE across replications as well as MCMC runtime. These outputs are displayed in Fig. 14.

MSE attains its lowest values at $k = 50$ for both ARMSM and ARVA-MSM. Within each model, the runtimes do not vary significantly for different choices of $k$. We choose $k = 50$ for both ARMSM and ARVA-MSM for subsequent analysis of brightness data.

Now, we examine the computational and predictive performance of the ARCAR, ARMSM and ARVA-MSM through cross-validation using the measures specified in Section 5.2. Table 7 displays these outputs within the single block spatial CV segment.

Matching the pattern broadly observed in simulation studies for single block spatial cv, ARCAR exhibits lowest uncertainty with the largest bias and, more importantly, coverage rate falling short of the desired level of 90%. On the other hand, both of the proposed methods nearly reach the target coverage level along with smaller bias. Computation-wise, ARMSM is most efficient, by a wide margin as expected, due to the shortest runtimes and largest ESS values. Although ARCAR and ARVA-MSM have very low ESS values, the latter outperforms the former with $> 45\%$ shorter runtime and $\sim 42\%$ improvement in relative time efficiency.

We also perform three blocks spatial CV. Here, in each of the 30 replications, we consider three disjoint small blocks each with dimensions $25 \times 25$ and treat all available observations from these blocks across all time points as the test data. The summary

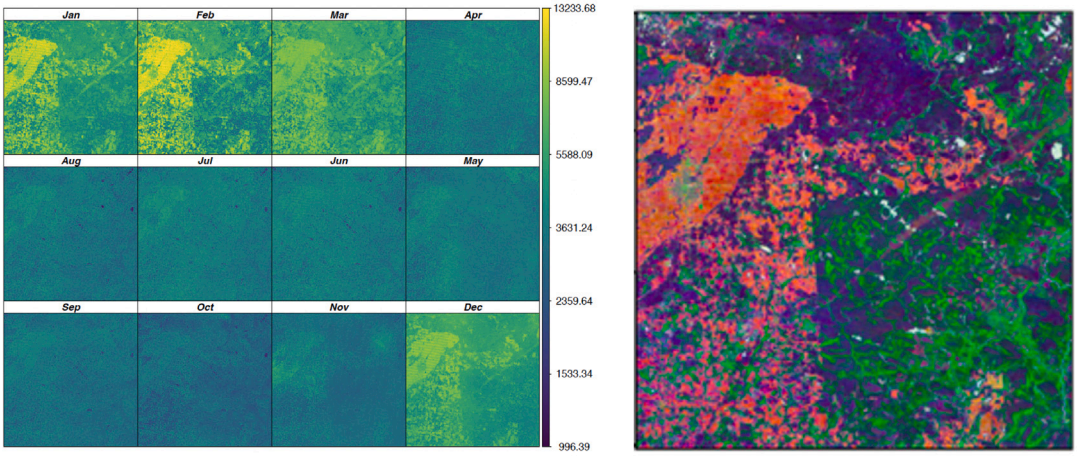



**Fig. 15.** (left) Monthly Map of 10-year average brightness using ARMSM with fixed $\eta = 0$. Although brightness is modeled in logarithmic scale, values are shown in the original scale. (right) Spatial variability of the land cover within the study area, depicted as an RGB composite image of mean TC brightness, greenness, and wetness from the WELD imagery.

outputs of the same three predictive performance comparison measures are shown in three blocks spatial CV segment of Table 7. The predictive performance improved significantly in ARCAR (lower bias and improved coverage) compared to the single-block version although the coverage rate still trails the other two methods. The proposed methods have shown modest reduction in bias as well.

Predicted brightness values at missing location-time combinations based on any of the three candidate models can be used to obtain a complete 10-year spatiotemporal imagery over the study region. Since the general pattern of these maps remains broadly similar between the models, we only display the output from ARMSM in the left panel of Fig. 15 as month-wise maps of 10-year averaged brightness. Brightness is related to overall surface albedo which varies according to the type of land cover as well as seasonality in vegetation and weather. Hence, to explain the spatial and month-to-month variations in these maps, we display various kinds of land covers and uses within the study region in the right panel of Fig. 15. It shows agricultural fields (orange pixels), deciduous (purple) and evergreen (green) forests, developed land (gray), as well as scattered water bodies (white) and wetlands (pink). The eastern half of the study area is predominantly forested whereas the agricultural land is mostly located in the northwestern part near the town of Antigo. Since snow-covered fields have larger albedo values than forests, the winter maps (December-March) exhibit significant contrast in brightness values between the eastern and western portions that dissipates during rest of the year due to the absence of snow. A closer look at the winter maps also reveals a narrow, almost linear trajectory in the northeastern part of the region having higher brightness values relative to the neighboring forest area. It represents the path of a tornado that traversed through this area on June 07, 2007 (around the middle of our study period) resulting in near-complete removal of trees along its swath (Wilson, 2015). Whereas gradual recovery occurred in subsequent years, the overall 10-year averaged brightness along this trajectory remains higher than that from the neighboring dense forest in winter months.

## 8. Discussion

In this article, we developed two computationally efficient alternative methods to CAR for modeling spatial and spatiotemporal datasets on a lattice. Our methods use the spectral structure of the Laplacian to construct reduced-rank basis function representations of areal random effects, an idea first mentioned in Mahdi et al. (2019). Our first proposal MSM shrinks, a priori, the weights of these functions in inverse proportion of their spatial smoothness. Simulation studies establish that this modification rectifies significant overestimation of the spatial variance parameter that could potentially occur in the original approach. Our second proposal VA-MSM further refines the model by improving the estimation of the nugget variance parameter using another set of uncorrelated and spatially heteroscedastic random effects. Finally, we have also developed spatiotemporal extension to each of these methods illustrating with a model involving a time series of spatial random effects. Whereas all examples in this article use the Gaussian linear model at the data-level, these methods can easily be extended to other types of response (binary or count) through a generalized linear model, in the same way CAR is applied in those settings.

Expressing spatial random effects on a graph as linear combination of eigenvectors has been explored in Hughes and Haran (2013), albeit with a different objective of dimension reduction to resolve confounding between fixed and random effects. Hence, the hierarchical model proposed therein, in spite of its apparent similarity with MSM, has major structural differences with the latter. In Appendix A.8, we describe the contrasts between two approaches.

Since VA-MSM represents a more complete approximation of the CAR prior over MSM, it has resulted in more accurate variance estimation in the two simulation studies where CAR prior is used for data generation. Notably, for the simulation models involving general spatial and spatiotemporal functions, CAR has shown very low coverage rates and both of the proposed methods have been

able to improve that by a large margin. We found VA-MSM to be generally more robust to model misspecification. In the lower block of Table 1, where we used Function-1 to simulate on an irregular grid, VA-MSM was able to outperform CAR, INLA and MSM by a large margin in terms of out-of-sample empirical coverage. INLA, in spite of its ability to produce fast inference for large datasets, suffers from low predictive accuracy relative to VA-MSM for non-Gaussian spatial functions. Comparative assessment of Tables 1 and 4 also reveals that predictive performance of VA-MSM, relative to MSM, was more robust to switching between regular and irregular adjacency structures for the same spatial function. For the two real datasets analyzed in this article, proposed methods have again achieved either comparable or better coverage rates than CAR in single-block spatial CV. Use of smaller contiguous test blocks, as expected, leads to performance improvement for all methods.

MSM exhibits significant computational advantage CAR and VA-MSM, as all simulation studies and real data analyses clearly indicate, which is a consequence of reducing the parameter size from a large number of unit-specific random effects to weights of a (relatively) small number of basis functions. In spatial-only settings, VA-MSM exhibits runtimes consistently in between those for CAR and MSM (when same value of $k$ is chosen). However, its low ESS values weaken or reverse its relative time efficiency over CAR. This occurs due to the presence of a large number of parameters in $\ddot{\varphi}$. As we have shown in two of the simulation studies, an alternative sampling scheme marginalizing $\ddot{\varphi}$ leads to marked improvement in ESS, albeit at the expense of significantly longer runtime and without any notable difference in predictive performance. Hence, we made limited use of this approach in this article and aim to further explore its potential, or alternative ideas, in our future research. Notably, when we move from spatial to spatiotemporal setting, the runtime differences between the proposed methods and CAR become much more pronounced, indicating their usefulness in analyzing large spatiotemporal datasets. In those cases, in spite of its lower ESS values, VA-MSM was able to achieve a relative computational efficiency $> 1$ over CAR, as shown in Tables 5 and 7. There is scope of further exploration of the feasibility and benefit of marginalized sampling scheme for VA-MSM within spatiotemporal models.

One important step in implementing the proposed methods is to choose an appropriate number ($k$) of Laplacian eigenpairs. In this article, we made this choice by comparing several candidate values of $k$ based on out-of-sample MSE and runtime. Based on the datasets analyzed in Sections 6 and 7, these MSE values were broadly found to either (i) reach the minimum before starting to increase, or (ii) keep declining initially and nearly plateauing after a certain point to justify the extra computational cost. Our simulation studies have shown that, even for the same spatial pattern, change in the adjacency structure (such as regular vs. irregular grid) can lead to significantly different choice for $k$. Additionally, the optimal $k$ values for fitting MSM and VA-MSM (or their spatiotemporal variants) to the same dataset do not necessarily match and should be determined independently. We also explored use of more automated alternatives to computationally intensive CV, such as DIC and WAIC, for $k$ selection. In our simulation study, when the actual spatial pattern was not generated from a Gaussian random effect model (e.g., Function-1 in Section 6.2.2), values of these measures for MSM kept decreasing significantly with increase in choice of $k$, so we could not use them to identify optimal $k$.

One alternative approach for $k$ selection would be to use the upper bound for KL divergence (derived in Appendix A.1) based on Laplacian eigenvalues and identify the value of $k$ at which improvement in KL distance becomes minimal or starts plateauing. Since the bound involves reciprocal of the discarded Laplacian eigenvalues, we can use the stretch approximation from Section 5.1 to approximately compute $L^+$. However, as our simulation studies in Sections 6.2.2 and 6.3.2 show, when the response comes from a general, non-Gaussian spatial function, MSM and VA-MSM may achieve significantly better predictive performance compared to CAR. In those cases, selecting a value of $k$ to optimize the KL divergence with respect to CAR may not be a reasonable strategy.

Instead of focussing on $k$ selection, one can also treat it as a model parameter similar to the work of Mahdi et al. (2019), where two algorithms, based on reversible-jump MCMC (Richardson and Green, 1997) and spike and slab prior (Ghosh and Clyde, 2011), respectively, were used for this purpose. These algorithms use a large number of basis functions and allow a subset of varying size to contribute to the model, at a given iteration. We explored use of the spike and slab prior to parametrize $k$ in MSM since the orthogonality of the Laplacian eigenvectors enables parallel updating the binary indicators for their coefficients. Appendix A.2 includes posterior sampling steps for the variable-$k$ MSM and, in Appendix A.9, a simulation study is presented comparing its predictive performance against the fixed-$k$ version. As demonstrated therein, adding the extra sampling step to update $k$ has not resulted in any predictive improvement or reduction in computational effort. Based on these findings, we decided to stick to CV-based choice of fixed $k$ for all analyses presented in this article. However, exploring efficient options to select or to parametrize $k$ still represents one direction for future research. In case of the latter, one should also look at (i) how the variance of the heteroscedastic random effect term in VA-MSM will change iteration-wise if $k$ is allowed to vary and, (ii) the computational complexities associated with such change.

Extending both of MSM and VA-MSM to multivariate areal data models will significantly broaden their utility. It will be similar to how we extended these methods to spatiotemporal datasets if we use pairwise correlation between variables to replace correlation across time points. The exact details of hierarchical structure in such an extension will depend on how spatial random effects interact with correlation between variables. For example, if one chooses the linear models of coregionalization (Jin et al., 2007) that involve a sequence of variable-specific spatial random effects with CAR prior, they bear structural resemblance to the ARCAR model that we used in this article, so development of the corresponding extensions of MSM and VA-MSM can benefit from the computations presented in Section 4.

Note that, in VA-MSM, $\tilde{\epsilon}_s$ is assumed to be independent across locations implying that the difference between a priori spatial covariances of these two processes $\text{cov}(\varphi_s, \varphi_{s'}) - \text{cov}(\ddot{\varphi}_s, \ddot{\varphi}_{s'}) = \tau^2 \sum_{i=k+2}^{S} \frac{1}{\lambda_{(i)}} v_{si} v_{s'i}$ is not compensated. Since $v_{k+2}, v_{k+3}, \ldots, v_S$ are associated with $(S-k-1)$ largest eigenvalues of $L$ representing highly varying, noise-like, functions on the lattice, we expect this residual covariance to be weak. This allows us to retain a low-rank covariance structure in spite of variance adjustment. It is relevant

to mention that, in the literature for point level spatial model, covariance tapering has been used to allow such residual processes to have non-diagonal, yet sparse, dispersion structure for short-range correlation modeling, see Sang and Huang (2012).

If one uses $\rho$-CAR in place of improper CAR with $\rho$ fixed at 0.99 (to ensure strong spatial association, a priori), the eigenpairs can be extracted from its precision matrix $(D_w - \rho W)$ and subsequent steps can be carried out with minor modifications. In that case, one does not need to exclude any eigenvector (unlike what we had to do in this article) and, due to nonsingularity of the precision matrix, the generalized inverse in VA-MSM would become the regular inverse. It should be noted that $\rho \approx 1$ implies near-singularity and determining approximate inverse Laplacian based on Lanczos algorithm remains computationally challenging. As far as the choice of adjacency matrix is concerned, any symmetric adjacency matrix can be used in place of binary adjacency matrix chosen in this article. A more challenging question is how one should apply the proposed methods in any hierarchical model where the adjacency matrix is parametrized such as local CAR prior (Lee and Mitchell, 2012), boundary detection models (Ma et al., 2010) and adaptive spatiotemporal smoothing (Rushworth et al., 2017). If entries of $W$ matrix change from one iteration to the next, that will change the eigenvalues, their ordering and eigenvectors. So, we no longer have the advantage of applying IRL just once before the MCMC and new computations will be necessary at every iteration. How to design an estimation scheme for this setting while optimizing the additional computational load is an interesting topic for future research.

We conclude this article by noting that the proposed methods open up a range of interesting theoretical questions for future research. Clearly, spectral properties of the Laplacian, i.e., how quickly its eigenvalues decay depending on the grid size and/or adjacency structure are the key factors controlling performance of these methods. Our simulation studies in Section 6.2.2 have shown that, even for the same spatial response, switching between regular and irregular grids may change the value for optimal number of eigenpairs and influence the predictive performance. A theoretical justification for the same will be immensely useful. In Appendix A.1, we have provided analytic expression and upper bounds for the KL divergence between response marginal distributions induced by CAR and MSM prior. Extending that analysis to VA-MSM prior will be a natural extension. Exploring connections between the prior and posterior supports in MSM and VA-MSM and those in CAR, restricted to functions on the lattice with a minimum threshold on spatial smoothness, represents another potential direction. One benefit of this kind of studies will be that, in addition to offering important insight in the model performance, they may also provide theory-driven guidance on practical aspects such as choice of $k$ that one can utilize in place of or in combination with the CV-based exercises employed in this article. See Kirichenko and Van Zanten (2017) for theoretical developments related to function estimation on a graph based on its Laplacian. Also of interest will be to verify if the posterior consistency holds for estimation of nugget variance in VA-MSM.

## Acknowledgments

The authors sincerely thank the anonymous reviewers for their comments and suggestions which enhanced the quality of content and presentation in this article. Md Kamrul Hasan Khan gratefully acknowledges utilization of computational resources from the Arkansas High Performance Computing Center which is funded through multiple National Science Foundation, United States grants and the Arkansas Economic Development Commission.

## Appendix

### A.1. Kullback–Leibler divergence analysis

Let us use $p$ and $\tilde{p}$ to denote the induced multivariate normal distributions on the response vector $y$ when the spatial random effects are assigned CAR and MSM prior distributions, respectively. These distributions share the same mean vector $X\beta$ with different dispersion matrices $\Sigma_p = \tau^2 L^+ + \sigma^2 I_S$ and $\Sigma_{\tilde{p}} = \tau^2 L_k^+ + \sigma^2 I_S$ where $L_k^+ = \sum_{i=2}^{k+1} \frac{1}{\lambda_{(i)}} v_i v_i^T$. (Note that $\lambda_{(i)} \uparrow$ as $i \uparrow$.) We first state the following well-known result on KL divergence between two $S$-variate normal distributions $f_1 = \mathcal{N}_S(\mu_1, \Sigma_1)$ and $f_2 = \mathcal{N}_S(\mu_2, \Sigma_2)$:

$$2 * KL(f_1 \parallel f_2) = \ln|\Sigma_2| - \ln|\Sigma_1| + trace(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1) - S.$$

Applying this result to our setting with $f_1 = p$ and $f_2 = \tilde{p}$, we get

$$2 * KL(p \parallel \tilde{p}) = \ln|\Sigma_{\tilde{p}}| - \ln|\Sigma_p| + trace(\Sigma_{\tilde{p}}^{-1}\Sigma_p) + 0 - S.$$

Note that the eigenvalues of $\Sigma_p$ are $\sigma^2 + \tau^2/\lambda_{(2)}, \sigma^2 + \tau^2/\lambda_{(3)}, \ldots, \sigma^2 + \tau^2/\lambda_{(S)}, \sigma^2$ and the eigenvalues of $\Sigma_{\tilde{p}}$ are $\sigma^2 + \tau^2/\lambda_{(2)}, \sigma^2 + \tau^2/\lambda_{(3)}, \ldots, \sigma^2 + \tau^2/\lambda_{(k+1)}, \underbrace{\sigma^2, \sigma^2, \ldots, \sigma^2}_{S-k \text{ many}}$. Their eigenvectors are identical. Hence,

$$\ln|\Sigma_{\tilde{p}}| - \ln|\Sigma_p| = (S-k-1)\ln\sigma^2 - \sum_{i=k+2}^{S} \ln(\sigma^2 + \tau^2/\lambda_{(i)}) = -\sum_{i=k+2}^{S} \ln\left(1 + \frac{\tau^2}{\sigma^2}\frac{1}{\lambda_{(i)}}\right)$$

Using orthonormality of eigenvectors and the property of trace function, we can write

$$trace(\Sigma_{\tilde{p}}^{-1}\Sigma_p) = \sum_{i=2}^{k+1} \frac{1}{\sigma^2 + \tau^2/\lambda_{(i)}}\left(\sigma^2 + \tau^2/\lambda_{(i)}\right) + \sum_{i=k+2}^{S} \frac{1}{\sigma^2}\left(\sigma^2 + \tau^2/\lambda_{(i)}\right) + \frac{1}{\sigma^2}\sigma^2$$

$$= S + \frac{\tau^2}{\sigma^2} \sum_{i=k+2}^{S} \frac{1}{\lambda_{(i)}}$$

Combining above expressions, we get

$$2 * KL(p \parallel \tilde{p}) = \sum_{i=k+2}^{S} \left( \frac{\tau^2}{\sigma^2} \frac{1}{\lambda_{(i)}} - \ln\left(1 + \frac{\tau^2}{\sigma^2} \frac{1}{\lambda_{(i)}}\right)\right) \leq \frac{\tau^2}{\sigma^2} \sum_{i=k+2}^{S} \frac{1}{\lambda_{(i)}}$$

Hence, as $k \uparrow$, $KL(p \parallel \tilde{p}) \downarrow$ bounded by the tail-sum of the sequence of reciprocal of Laplacian eigenvalues. Since $\lambda_{(k+2)}, \lambda_{(k+3)}, \ldots, \lambda_{(S)}$ represent largest $(S - k - 1)$ Laplacian eigenvalues, another useful upper bound may be obtained based on the inequality $a^2 > a - \ln(1 + a)$ for $a > 0$:

$$2 * KL(p \parallel \tilde{p}) \leq \left(\frac{\tau^2}{\sigma^2}\right)^2 \sum_{i=k+2}^{S} \frac{1}{\lambda_{(i)}^2}$$

For any specific $k$, $KL(p \parallel \tilde{p})$ becomes smaller if (i) discarded eigenvalues of $L$ have larger magnitudes or, (ii) the ratio $\tau^2/\sigma^2$ is smaller (i.e., the nonspatial source of heterogeneity becomes relatively more dominant).

### *A.2. Posterior distributions in spatial models*

Let, $y = \{y_s : s = 1, 2, \ldots, S\}$, $\varphi = \{\varphi_s : s = 1, 2, \ldots, S\}$ and $\ddot{\varphi} = \{\ddot{\varphi}_s : s = 1, 2, \ldots, S\}$ be $S \times 1$ vectors, $X$ is a $S \times (p+1)$ matrix with rows from $\{X_{s*}^T : s = 1, 2, \ldots, S\}$, $V = [v_2, v_3, \ldots, v_{k+1}]$ is a $S \times k$ matrix with its $s$th row represented as $V_{s*}^T$, $\theta = \{\theta_i : i = 1, 2, \ldots, k\}$ is a $k \times 1$ vector, $\Lambda$ is a $k \times k$ diagonal matrix with $i$th diagonal entry $\lambda_{(i+1)}$ and $\Delta$ is a $S \times S$ diagonal matrix with $d_s$ as the $s$th diagonal entry.

**CAR:**

$$\beta|- \sim \mathcal{N}_{p+1}(\Sigma_\beta\, \mu_\beta, \Sigma_\beta),\ \sigma^2|- \sim \text{IG}(a_1, b_1),\ \varphi_s|- \sim \mathcal{N}(\sigma^2_{\varphi_s}\, \mu_{\varphi_s}, \sigma^2_{\varphi_s}),\ \tau^2|- \sim \text{IG}(a_2, b_2),$$

where

$$\mu_\beta = \frac{1}{\sigma^2} X^T (y - \varphi) + \alpha_0/\nu_\beta^2 \text{ and } \Sigma_\beta^{-1} = \frac{1}{\sigma^2} X^T X + I_{p+1}/\nu_\beta^2;$$

$$a_1 = a_0 + S/2 \text{ and } b_1 = b_0 + \sum_{s=1}^{S} (y_s - X_{s*}^T \beta - \varphi_s)^2/2;$$

$$\mu_{\varphi_s} = \frac{1}{\sigma^2}(y_s - X_{s*}^T\, \beta) + \frac{1}{\tau^2} \sum_{s'} w(s, s')\varphi_{s'} \text{ and } (\sigma^2_{\varphi_s})^{-1} = \frac{1}{\sigma^2} + \frac{w_{s+}}{\tau^2};$$

$$a_2 = a_0 + (S-1)/2 \text{ and } b_1 = b_0 + \varphi^T L\, \varphi/2.$$

**SM:**

$$\beta|- \sim \mathcal{N}_{p+1}(\Sigma_\beta\, \mu_\beta, \Sigma_\beta),\ \sigma^2|- \sim \text{IG}(a_1, b_1),\ \theta|- \sim \mathcal{N}_k(\Sigma_\theta\, \mu_\theta, \Sigma_\theta),\ \tau^2|- \sim \text{IG}(a_2, b_2),$$

where

$$\mu_\beta = \frac{1}{\sigma^2} X^T (y - V\theta) + \alpha_0/\nu_\beta^2 \text{ and } \Sigma_\beta^{-1} = \frac{1}{\sigma^2} X^T X + I_{p+1}/\nu_\beta^2;$$

$$a_1 = a_0 + S/2 \text{ and } b_1 = b_0 + \sum_{s=1}^{S} (y_s - X_{s*}^T \beta - V_{s*}^T \theta)^2/2;$$

$$\mu_\theta = \frac{1}{\sigma^2} V^T (y - X\, \beta) \text{ and } \Sigma_\theta^{-1} = \frac{1}{\sigma^2} V^T V + \frac{1}{\tau^2} I_k;$$

$$a_2 = a_0 + k/2 \text{ and } b_1 = b_0 + \theta^T \theta/2.$$

**MSM:**

$$\beta|- \sim \mathcal{N}_{p+1}(\Sigma_\beta\, \mu_\beta, \Sigma_\beta),\ \sigma^2|- \sim \text{IG}(a_1, b_1),\ \theta|- \sim \mathcal{N}_k(\Sigma_\theta\, \mu_\theta, \Sigma_\theta),\ \tau^2|- \sim \text{IG}(a_2, b_2),$$

where

$$\mu_\beta = \frac{1}{\sigma^2} X^T (y - V\theta) + \alpha_0/\nu_\beta^2 \text{ and } \Sigma_\beta^{-1} = \frac{1}{\sigma^2} X^T X + I_{p+1}/\nu_\beta^2;$$

$$a_1 = a_0 + S/2 \text{ and } b_1 = b_0 + \sum_{s=1}^{S} (y_s - X_{s*}^T \beta - V_{s*}^T \theta)^2/2;$$

$$\mu_\theta = \frac{1}{\sigma^2} V^T (y - X\, \beta) \text{ and } \Sigma_\theta^{-1} = \frac{1}{\sigma^2} V^T V + \frac{1}{\tau^2} \Lambda;$$

$$a_2 = a_0 + k/2 \text{ and } b_1 = b_0 + \theta^T \Lambda\, \theta/2.$$

**Variable-$k$ MSM:** Here, the number of eigenpairs contributing to the model in any iteration is allowed to vary. So, an indicator vector $\gamma_{k\times 1}$ is introduced to represent zero/nonzero status of the members of $\theta$ coefficient vector with $k_\gamma = \sum_{j=1}^{k} \gamma_j$ denoting number of nonzero coefficients within $\theta$. Also, in this approach, $\tau^2$ is reparametrized as $\sigma^2\tau^2$ to facilitate marginalization of $\theta$ coefficients.

Wherever $\gamma$ is used in the subscript for any vector or matrix, it indicates restriction to the components corresponding to positions $\tilde{\gamma} = \{j : \gamma_j = 1\}$. Hence, (i) $\theta_\gamma$ denotes the subvector of $\theta$ corresponding to indices in $\tilde{\gamma}$, (ii) $V_\gamma, V_{o\gamma}$ and $V_{m\gamma}$ denote the submatrices formed by selecting the columns corresponding to $\tilde{\gamma}$, and (iii) $\Lambda_\gamma$ denotes the submatrix of $\Lambda$ obtained by selecting both rows and columns based on $\tilde{\gamma}$. Additionally, $V_{*j}$ denotes the $j$th column of $V$.

$$\beta|- \sim \mathcal{N}_{p+1}(\Sigma_\beta\ \mu_\beta, \Sigma_\beta),\ \sigma^2|- \sim \text{IG}(a_1, b_1),\ \theta_\gamma|- \sim \mathcal{N}_{k_\gamma}(\Sigma_\theta\ \mu_\theta, \sigma^2\Sigma_\theta),\ \tau^2|- \sim \text{IG}(a_2, b_2),$$

Posterior odds of $(\gamma_j = 1) = \frac{\pi}{1-\pi}\left(\frac{\tau^2+\Lambda_{jj}}{\Lambda_{jj}}\right)^{-1/2} \exp\left\{\frac{1}{2\sigma^2}\left(\frac{\tau^2+\Lambda_{jj}}{\tau^2}\right)^{-1}\left(V_{*j}^T(y - X\beta)\right)^2\right\}$ for $j = 1, 2, \ldots, J$.

where

$$\mu_\beta = \frac{1}{\sigma^2}X^T(y - V_\gamma\ \theta_\gamma) + \alpha_0/\nu_\beta^2 \text{ and } \Sigma_\beta^{-1} = \frac{1}{\sigma^2}X^TX + I_{p+1}/\nu_\beta^2;$$
$$a_1 = a_0 + S/2 + k_\gamma/2 \text{ and } b_1 = b_0 + \frac{1}{2}(y - X\beta - V_\gamma\ \theta_\gamma)^T(y - X\beta - V_\gamma\ \theta_\gamma) + \frac{1}{2\tau^2}\theta_\gamma^T\ \Lambda_\gamma\ \theta_\gamma;$$
$$\mu_\theta = V_\gamma^T(y - X\ \beta) \text{ and } \Sigma_\theta = \left(I_{k_\gamma} + \frac{1}{\tau^2}\Lambda_\gamma\right)^{-1};$$
$$a_2 = a_0 + k_\gamma/2 \text{ and } b_2 = b_0 + \theta_\gamma^T\ \Lambda_\gamma\ \theta_\gamma/2.$$

**VA-MSM:**

$$\beta|- \sim \mathcal{N}_{p+1}(\Sigma_\beta\ \mu_\beta, \Sigma_\beta),\ \sigma^2|- \sim \text{IG}(a_1, b_1),\ \ddot{\varphi}|- \sim \mathcal{N}_S(\Sigma_{\ddot{\varphi}}\ \mu_{\ddot{\varphi}}, \Sigma_{\ddot{\varphi}}),$$
$$\theta|- \sim \mathcal{N}_k(\Sigma_\theta\ \mu_\theta, \Sigma_\theta),\ \tau^2|- \sim \text{IG}(a_2, b_2),$$

where

$$\mu_\beta = \frac{1}{\sigma^2}X^T\ (y - \ddot{\varphi}) + \alpha_0/\nu_\beta^2 \text{ and } \Sigma_\beta^{-1} = \frac{1}{\sigma^2}\ X^TX + I_{p+1}/\nu_\beta^2,$$
$$a_1 = a_0 + S/2 \text{ and } b_1 = b_0 + \sum_{s=1}^{S}(y_s - X_{s*}^T\beta - \ddot{\varphi}_s)^2/2$$
$$\mu_{\ddot{\varphi}} = \frac{1}{\sigma^2}\ (y - X\beta) + \frac{1}{\tau^2}\ \Delta^{-1}V\theta \text{ and } \Sigma_{\ddot{\varphi}}^{-1} = \frac{1}{\sigma^2}\ I_S + \frac{1}{\tau^2}\ \Delta^{-1},$$
$$\mu_\theta = \frac{1}{\tau^2}\ V^T\Delta^{-1}\ddot{\varphi} \text{ and } \Sigma_\theta^{-1} = \frac{1}{\tau^2}\left(V^T\Delta^{-1}V + \Lambda\right),$$
$$a_2 = a_0 + S/2 + k/2 \text{ and } b_2 = b_0 + (\ddot{\varphi} - V\theta)^T\Delta^{-1}(\ddot{\varphi} - V\theta)/2 + \theta^T\Lambda\ \theta/2$$

**VA-MSM Marginalized:** In this approach, reparametrizing $\tau^2$ as $\sigma^2\tau^2$ leads to standard sampling distribution for $\sigma^2$.

$$\beta|- \sim \mathcal{N}_{p+1}(\Sigma_\beta\ \mu_\beta, \Sigma_\beta),\ \sigma^2|- \sim \text{IG}(a_1, b_1), \theta|- \sim \mathcal{N}_k(\Sigma_\theta\ \mu_\theta, \sigma^2\ \Sigma_\theta),$$
$$\tau^2|- \propto \left|(I + \tau^2\Delta)\right|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta - V\theta)^T(I + \tau^2\Delta)^{-1}(y - X\beta - V\theta)\right\}$$
$$(\tau^2)^{-a_0-k/2-1} \exp\left\{-\frac{1}{\tau^2}\left(b_0 + \frac{1}{2\sigma^2}\theta^T\Lambda\theta\right)\right\},$$

where

$$\mu_\beta = \frac{1}{\sigma^2}X^T(I + \tau^2\Delta)^{-1}(y - V\theta) + \alpha_0/\nu_\beta^2 \text{ and } \Sigma_\beta^{-1} = \frac{1}{\sigma^2}\ X^T(I + \tau^2\Delta)^{-1}X + I_{p+1}/\nu_\beta^2,$$
$$a_1 = a_0 + S/2 + k/2 \text{ and } b_1 = b_0 + \frac{1}{2}(y - X\beta - V\theta)^T(I + \tau^2\Delta)^{-1}(y - X\beta - V\theta) + \frac{1}{2\tau^2}\theta^T\Lambda\theta$$
$$\mu_\theta = V^T(I + \tau^2\Delta)^{-1}(y - X\beta) \text{ and } \Sigma_\theta^{-1} = V^T(I + \tau^2\Delta)^{-1}V + \frac{1}{\tau^2}\Lambda$$

$\tau^2$ is updated using a random-walk Metropolis-Hastings step with lognormal proposal distribution.

### *A.3. Posterior distributions in spatiotemporal models*

Define $\psi = (\psi_{1*}^T, \psi_{2*}^T, \ldots, \psi_{k*}^T)^T$ with $\psi_{i*} = \{\psi_{it} : t = 1, 2, \ldots, T\}$, $\ddot{\varepsilon} = (\ddot{\varepsilon}_{1*}^T, \ddot{\varepsilon}_{2*}^T, \ldots, \ddot{\varepsilon}_{S*}^T)^T$, with $\ddot{\varepsilon}_{s*} = \{\ddot{\varepsilon}_{st} : t = 1, 2, \ldots, T\}$, $y_{s*} = \{y_{st} : t = 1, 2, \ldots, T\}$, $X_{s**}$ as a $T\times(p+1)$ matrix with rows from $\{X_{st*}^T : t = 1, 2, \ldots, T\}$, and $\tilde{V}_{s*i} = v_{s(i+1)}I_T$ for $i = 1, 2, \ldots, k$ and $s = 1, 2, \ldots, S$. Now, the prior joint distribution of $\psi_{i*}$ and $\ddot{\varepsilon}_{s*}$ can be written as $\psi_{i*} \overset{iid}{\sim} \mathcal{N}_T(0, \Omega_\psi)$ and $\ddot{\varepsilon}_{s*} \overset{iid}{\sim} \mathcal{N}_T(0, \Omega_{\ddot{\varepsilon}})$, respectively.

**ARCAR:**

$$\beta|- \sim \mathcal{N}_{p+1}(\Sigma_\beta\ \mu_\beta,\ \Sigma_\beta),\ \sigma^2|- \sim \text{IG}(a_1, b_1),\ \tau^2|- \sim \text{IG}(a_2, b_2),\ \varphi_{st}|- \sim \mathcal{N}(\sigma^2_{\varphi_{st}}\mu_{\varphi_{st}}, \sigma^2_{\varphi_{st}}),$$

$$\pi(\eta|-) \propto (1-\eta^2)^{S/2} \exp\Big\{ -\frac{1}{2\tau^2}\Big(\eta^2 \sum_{t=1}^{T-1} \varphi_{*t}^T L \varphi_{*t} - 2\eta \sum_{t=1}^{T} \varphi_{*(t-1)} L \varphi_{*t}\Big)\Big\},$$

where

$$\mu_\beta = \sum_{s=1}^{S}\sum_{t=1}^{T} X_{st*}^T (y_{st} - \varphi_{st})/\sigma^2 + \alpha_0/\nu_\beta^2 \text{ and } \Sigma_\beta^{-1} = \sum_{s=1}^{S}\sum_{t=1}^{T} X_{st*}^T X_{st*}/\sigma^2 + I_{p+1}/\nu_\beta^2,$$

$$a_1 = a_0 + S \times T/2 \text{ and } b_1 = b_0 + \sum_{s=1}^{S}\sum_{t=1}^{T} (y_{st} - X_{st*}^T\beta - \varphi_{st})^2/2,$$

$$a_2 = a_0 + \frac{(S-1)(T+1)}{2} \text{ and}$$

$$b_2 = b_0 + \frac{(1-\eta^2)}{2}\varphi_{*0}^T L \varphi_{*0} + \frac{1}{2}\sum_{t=1}^{T} (\varphi_{*t} - \eta\varphi_{*(t-1)})^T L (\varphi_{*t} - \eta\varphi_{*(t-1)}),$$

$$\mu_{\varphi_{st}} = \begin{cases} \eta\varphi_{s(t+1)} + \frac{1}{w_{s+}}\sum_{s'} w(s,s')(\varphi_{s't} - \eta\varphi_{s'(t+1)}) & \text{for } t=0, \\ \frac{1}{\sigma^2}(y_{st} - X_{st*}^T\beta) + \frac{1}{\tau^2}\Big( w_{s+}\eta(\varphi_{s(t-1)} + \varphi_{s(t+1)}) + (1 \\ +\eta^2)\sum_{s'} w(s,s')\varphi_{s't} - \eta\sum_{s'} w(s,s')(\varphi_{s'(t-1)} + \varphi_{s'(t+1)})\Big) & \text{for } t=1,2,\ldots,(T-1), \\ \frac{1}{\sigma^2}(y_{st} - X_{st*}^T\beta) + \frac{1}{\tau^2}\Big( w_{s+}\eta\varphi_{s(t-1)} \\ +\sum_{s'} w(s,s')(\varphi_{s't} - \eta\varphi_{s'(t-1)})\Big) & \text{for } t=T, \end{cases}$$

$$\sigma^2_{\varphi_{st}} = \begin{cases} \frac{\tau^2}{w_{s+}} & \text{for } t=0, \\ \Big(\frac{1}{\sigma^2} + \frac{w_{s+}(1+\eta^2)}{\tau^2}\Big)^{-1} & \text{for } t=1,2,\ldots,(T-1), \\ \Big(\frac{1}{\sigma^2} + \frac{w_{s+}}{\tau^2}\Big)^{-1} & \text{for } t=T. \end{cases}$$

**ARMSM:**

$$\beta|- \sim \mathcal{N}_{p+1}(\Sigma_\beta\, \mu_\beta,\ \Sigma_\beta),\ \sigma^2|- \sim \text{IG}(a_1,b_1),\ \tau^2|- \sim \text{IG}(a_2,b_2),\ \psi_{*0} \sim \mathcal{N}_k(\eta\,\psi_{*1}, \tau^2\Lambda^{-1}),$$

$$\psi|- \sim \mathcal{N}_{Tk}(\Sigma_\psi^{-1}\mu_\psi, \Sigma_\psi^{-1}),\ \pi(\eta|-) \propto (1-\eta^2)^{k/2}\exp\Big\{ -\frac{1}{2\tau^2}\Big( \eta^2 \sum_{t=1}^{T-1} \psi_{*t}^T \Lambda \psi_{*t}$$

$$-2\eta \sum_{t=1}^{T} \psi_{*(t-1)}\Lambda\psi_{*t}\Big)\Big\},$$

where $\mu_\psi = (\mu_{\psi\,1}^T, \mu_{\psi\,2}^T, \ldots, \mu_{\psi\,k}^T)^T$, $\Sigma_\psi$ is a tridiagonal matrix and

$$\mu_\beta = \sum_{s=1}^{S}\sum_{t=1}^{T} X_{st*}^T (y_{st} - V_{s*}^T\psi_{*t})/\sigma^2 + \alpha_0/\nu_\beta^2 \text{ and } \Sigma_\beta^{-1} = \sum_{s=1}^{S}\sum_{t=1}^{T} X_{st*}^T X_{st*}/\sigma^2 + I_{p+1}/\nu_\beta^2,$$

$$a_1 = a_0 + S \times T/2 \text{ and } b_1 = b_0 + \sum_{s=1}^{S}\sum_{t=1}^{T} (y_{st} - X_{st*}^T\beta - V_{s*}^T\psi_{*t})^2/2,$$

$$a_2 = a_0 + \frac{k(T+1)}{2} \text{ and}$$

$$b_2 = b_0 + \frac{(1-\eta^2)}{2}\psi_{*0}^T\Lambda\psi_{*0} + \frac{1}{2}\sum_{t=1}^{T} (\psi_{*t} - \eta\psi_{*(t-1)})^T \Lambda\, (\psi_{*t} - \eta\psi_{*(t-1)}),$$

$$\mu_{\psi\,i} = \frac{1}{\sigma^2}\sum_{s=1}^{S} \tilde{V}_{s*i}^T (y_{s*} - X_{s*}\beta) \text{ and } \Sigma_{\psi\,ii} = \frac{1}{\sigma^2} I_T + \Omega_\psi^{-1}.$$

**ARVA-MSM:**

$$\beta|- \sim \mathcal{N}_{p+1}(\Sigma_\beta\, \mu_\beta,\ \Sigma_\beta),\ \sigma^2|- \sim \text{IG}(a_1,b_1),\ \tau^2|- \sim \text{IG}(a_2,b_2),\ \psi_{*0} \sim \mathcal{N}_k(\eta\,\psi_{*1}, \tau^2\Lambda^{-1}),$$
$$\psi|- \sim \mathcal{N}_{Tk}(\Sigma_\psi^{-1}\mu_\psi, \Sigma_\psi^{-1}),\ \ddot{\varepsilon}_{*0} \sim \mathcal{N}_S(\eta\,\ddot{\varepsilon}_{*1}, \tau^2\Delta),\ \ddot{\varepsilon}|- \sim \mathcal{N}_{ST}(\Sigma_{\ddot{\varepsilon}}^{-1}\mu_{\ddot{\varepsilon}}, \Sigma_{\ddot{\varepsilon}}^{-1}),$$

$$\pi(\eta|-) \propto (1-\eta^2)^{(S+k)/2}\exp\Big\{ -\frac{1}{2\tau^2}\Big( \eta^2 \Big( \sum_{t=1}^{T-1} \psi_{*t}^T\Lambda\psi_{*t} + \sum_{t=1}^{T-1} \ddot{\varepsilon}_{*t}^T\Delta^{-1}\ddot{\varepsilon}_{*t}\Big)$$

$$-2\eta\Big(\sum_{t=1}^{T} \psi_{*(t-1)}\Lambda\psi_{*t} + \sum_{t=1}^{T} \ddot{\varepsilon}_{*(t-1)}\Lambda\ddot{\varepsilon}_{*t}\Big)\Big)\Big\},$$

where $\mu_\psi = (\mu_{\psi\,1}^T, \mu_{\psi\,2}^T, \ldots, \mu_{\psi\,k}^T)^T$, $\mu_{\ddot{\epsilon}} = (\mu_{\ddot{\epsilon}\,1}^T, \mu_{\ddot{\epsilon}\,2}^T, \ldots, \mu_{\ddot{\epsilon}\,S}^T)^T$, $\Sigma_\psi$ and $\Sigma_{\ddot{\epsilon}}$ are a tridiagonal matrix and

$$\mu_\beta = \sum_{s=1}^{S}\sum_{t=1}^{T} X_{st*}^T(y_{st} - V_{s*}^T\psi_{*t} - \ddot{\epsilon}_{st})/\sigma^2 + \alpha_0/\nu_\beta^2 \text{ and } \Sigma_\beta^{-1} = \sum_{s=1}^{S}\sum_{t=1}^{T} X_{st*}^T X_{st*}/\sigma^2 + I_{p+1}/\nu_\beta^2,$$

$$a_1 = a_0 + S\times T/2 \text{ and } b_1 = b_0 + \sum_{s=1}^{S}\sum_{t=1}^{T}(y_{st} - X_{st*}^T\beta - V_{s*}^T\psi_{*t} - \ddot{\epsilon}_{st})^2/2,$$

$$a_2 = a_0 + \frac{k\,(T+1)}{2} + \frac{S\,(T+1)}{2} \text{ and}$$

$$b_2 = b_0 + \frac{(1-\eta^2)}{2}\,\psi_{*0}^T\Lambda\psi_{*0} + \frac{1}{2}\sum_{t=1}^{T}(\psi_{*t} - \eta\psi_{*(t-1)})^T\,\Lambda\,(\psi_{*t} - \eta\psi_{*(t-1)})$$

$$+ \frac{(1-\eta^2)}{2}\,\ddot{\epsilon}_{*0}\Delta^{-1}\ddot{\epsilon}_{*0} + \frac{1}{2}\sum_{t=1}^{T}(\ddot{\epsilon}_{*t} - \eta\,\ddot{\epsilon}_{*(t-1)})^T\Delta^{-1}\,(\ddot{\epsilon}_{*t} - \eta\,\ddot{\epsilon}_{*(t-1)}),$$

$$\mu_{\psi\,i} = \frac{1}{\sigma^2}\sum_{s=1}^{S}\tilde{V}_{s*i}^T(y_{s*} - X_{s**}\beta - \ddot{\epsilon}_{s*}) \text{ and } \Sigma_{\psi\,ii} = \frac{1}{\sigma^2}I_T + \Omega_\psi^{-1},$$

$$\mu_{\ddot{\epsilon}\,s} = \frac{1}{\sigma^2}\Big(y_{s*} - X_{s**}\beta - \sum_{i=1}^{k}\tilde{V}_{s*i}\psi_{i*}\Big) \text{ and } \Sigma_{\ddot{\epsilon}\,ss} = \frac{1}{\sigma^2}I_T + \Omega_{\ddot{\epsilon}}^{-1}.$$

### *A.4. Derivation of posterior distributions for spatiotemporal random effects*

The posterior distributions for spatiotemporal random effects in each of ARCAR, ARMSM and ARVA-MSM, mentioned in Appendix A.3, require significant computational effort. Hence, for the convenience of readers, we present the details of those derivations below.

**Posterior distribution of $\varphi_{st}$ for ARCAR**: Fig. A.1 presents the dependence structure for spatiotemporal random effects $\{\varphi_{st}\}$ in ARCAR.

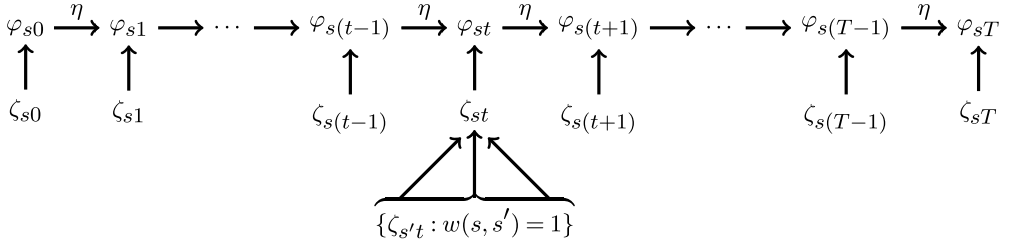



**Fig. A.1.** Hierarchical dependence structure for spatiotemporal random effects $\{\varphi_{st}\}$ in ARCAR.

At $t = 0$, the posterior distribution of $\varphi_{s0}$ includes contribution from two sources: its own spatial prior and the future random effect at the same location $\varphi_{s1}$ which depends on it through $\eta$.

$$\begin{aligned}
\pi(\varphi_{s0}|-) &\propto \pi(\varphi_{s0}|-)\,\pi(\varphi_{s1}|\varphi_{s0},-)\\
&\propto \exp\Big\{-\frac{w_{s+}(1-\eta^2)}{2\tau^2}\Big(\varphi_{s0} - \frac{\sum_{s'} w(s,s')\varphi_{s'0}}{w_{s+}}\Big)^2\Big\} \times \exp\Big\{-\frac{w_{s+}}{2\tau^2}\Big(\varphi_{s1} - \eta\varphi_{s0} - \frac{\sum_{s'} w(s,s')(\varphi_{s'1} - \eta\varphi_{s'0})}{w_{s+}}\Big)^2\Big\}\\
&\propto \exp\Big\{-\frac{w_{s+}}{2\tau^2}\Big(\varphi_{s0}^2 - 2\varphi_{s0}\Big((1-\eta^2)\frac{\sum_{s'} w(s,s')(\varphi_{s'0})}{w_{s+}} + \eta\varphi_{s1} - \eta\frac{\sum_{s'} w(s,s')(\varphi_{s'1} - \eta\varphi_{s'0})}{w_{s+}}\Big)\Big\}\\
&\propto \exp\Big\{-\frac{w_{s+}}{2\tau^2}\Big(\varphi_{s0}^2 - 2\varphi_{s0}\Big(\eta\varphi_{s1} + \frac{\sum_{s'} w(s,s')(\varphi_{s'0} - \eta\varphi_{s'1})}{w_{s+}}\Big)\Big)\Big\}
\end{aligned}$$

At $t = 1, 2, \ldots, (T-1)$, the posterior distribution for $\varphi_{st}$ includes contribution from three sources: the likelihood of response $y_{st}$ as well as immediate past and future random effects at the same location $\varphi_{s(t-1)}$ and $\varphi_{s(t+1)}$, respectively. So, for $t = 1, 2, \ldots, (T-1)$,

$$\begin{aligned}
\pi(\varphi_{st}|-) &\propto \pi(y_{st}|-)\,\pi(\varphi_{st}|\varphi_{s(t-1)},-)\,\pi(\varphi_{s(t+1)}|\varphi_{st},-)\\
&\propto \pi(y_{st}|-)\,\exp\Big\{-\frac{w_{s+}}{2\tau^2}\Big(\varphi_{st} - \eta\varphi_{s(t-1)} - \frac{\sum_{s'} w(s,s')(\varphi_{s't} - \eta\varphi_{s'(t-1)})}{w_{s+}}\Big)^2\Big\}\\
&\quad\times \exp\Big\{-\frac{w_{s+}}{2\tau^2}\Big(\varphi_{s(t+1)} - \eta\varphi_{st} - \frac{\sum_{s'} w(s,s')(\varphi_{s'(t+1)} - \eta\varphi_{s't})}{w_{s+}}\Big)^2\Big\}\\
&\propto \pi(y_{st}|-)\,\exp\Big\{-\frac{w_{s+}}{2\tau^2}\Big((1+\eta^2)\varphi_{st}^2 - 2\varphi_{st}\Big(\eta(\varphi_{s(t-1)} + \varphi_{s(t+1)})\\
&\quad + \frac{\sum_{s'} w(s,s')(\varphi_{s't} - \eta\varphi_{s'(t-1)})}{w_{s+}} - \eta\frac{\sum_{s'} w(s,s')(\varphi_{s'(t+1)} - \eta\varphi_{s't})}{w_{s+}}\Big)\Big)\Big\}
\end{aligned}$$

$$
\begin{aligned}
&\propto \pi(y_{st}|-)\ \exp\Big\{-\frac{w_{s+}}{2\tau^2}\Big((1+\eta^2)\varphi_{st}^2-2\varphi_{st}\Big(\eta(\varphi_{s(t-1)}+\varphi_{s(t+1)})\\
&\quad+(1+\eta^2)\frac{\sum_{s'}w(s,s')\varphi_{s't}}{w_{s+}}-\eta\frac{\sum_{s'}w(s,s')(\varphi_{s'(t-1)}+\varphi_{s'(t+1)})}{w_{s+}}\Big)\Big)\Big\}\\
&\propto \exp\Big\{-\frac{1}{2\sigma^2}(y_{st}-X_{st*}^T\beta-\varphi_{st})^2\Big\}\ \exp\Big\{-\frac{w_{s+}(1+\eta^2)}{2\tau^2}\Big(\varphi_{st}^2\\
&\quad-2\varphi_{st}\frac{1}{w_{s+}(1+\eta^2)}\Big(w_{s+}\eta(\varphi_{s(t-1)}+\varphi_{s(t+1)})+(1+\eta^2)\sum_{s'}w(s,s')\varphi_{s't}-\eta\sum_{s'}w(s,s')(\varphi_{s'(t-1)}+\varphi_{s'(t+1)})\Big)\Big)\Big\}\\
&\propto \exp\Big\{-\frac{1}{2}\Big(\Big(\frac{1}{\sigma^2}+\frac{w_{s+}(1+\eta^2)}{\tau^2}\Big)\varphi_{st}^2-2\varphi_{st}\Big(\frac{1}{\sigma^2}(y_{st}-X_{st*}^T\beta)+\frac{1}{\tau^2}\Big(w_{s+}\eta(\varphi_{s(t-1)}\\
&\quad+\varphi_{s(t+1)})+(1+\eta^2)\sum_{s'}w(s,s')\varphi_{s't}-\eta\sum_{s'}w(s,s')(\varphi_{s'(t-1)}+\varphi_{s'(t+1)})\Big)\Big)\Big)\Big\}
\end{aligned}
$$

And, for $t=T$, the posterior distribution for $\varphi_{sT}$ includes contribution from two sources: the likelihood of response $y_{sT}$ as well as immediate past random effect at the same location $\varphi_{s(T-1)}$.

$$
\begin{aligned}
\pi(\varphi_{sT}|-) &\propto \pi(y_{sT}|-)\,\pi(\varphi_{sT}|\varphi_{s(T-1)},-)\\
&\propto \exp\Big\{-\frac{1}{2\sigma^2}(y_{sT}-X_{sT*}^T\beta-\varphi_{sT})^2\Big\}\times\exp\Big\{-\frac{w_{s+}}{2\tau^2}\Big(\varphi_{sT}-\eta\varphi_{s(T-1)}-\frac{\sum_{s'}w(s,s')(\varphi_{s'T}-\eta\varphi_{s'(T-1)})}{w_{s+}}\Big)^2\Big\}\\
&\propto \exp\Big\{-\frac{1}{2\sigma^2}(\varphi_{sT}^2-2\varphi_{sT}(y_{sT}-X_{sT*}^T\beta))\Big\}\times\exp\Big\{-\frac{w_{s+}}{2\tau^2}\Big(\varphi_{sT}^2-2\varphi_{sT}\Big(\eta\varphi_{s(T-1)}+\frac{\sum_{s'}w(s,s')(\varphi_{s'T}-\eta\varphi_{s'(T-1)})}{w_{s+}}\Big)\Big)\Big\}\\
&\propto \exp\Big\{-\frac{1}{2}\Big(\Big(\frac{1}{\sigma^2}+\frac{w_{s+}}{\tau^2}\Big)\varphi_{st}^2-2\varphi_{st}\Big(\frac{1}{\sigma^2}(y_{sT}-X_{sT*}^T\beta)+\frac{1}{\tau^2}\Big(w_{s+}\eta\varphi_{s(T-1)}\times+\sum_{s'}w(s,s')(\varphi_{s'T}-\eta\varphi_{s'(T-1)})\Big)\Big)\Big\}
\end{aligned}
$$

Hence, the posterior distribution is $\varphi_{st}|-\sim\mathcal{N}(\sigma^2_{\varphi_{st}}\mu_{\varphi_{st}},\sigma^2_{\varphi_{st}})$ where

$$
\mu_{\varphi_{st}}=\begin{cases}
\eta\varphi_{s(t+1)}+\frac{1}{w_{s+}}\sum_{s'}w(s,s')(\varphi_{s't}-\eta\varphi_{s'(t+1)}) & \text{for } t=0,\\
\frac{1}{\sigma^2}(y_{st}-X_{st*}^T\beta)+\frac{1}{\tau^2}\Big(w_{s+}\eta(\varphi_{s(t-1)}+\varphi_{s(t+1)})+(1\\
+\eta^2)\sum_{s'}w(s,s')\varphi_{s't}-\eta\sum_{s'}w(s,s')(\varphi_{s'(t-1)}+\varphi_{s'(t+1)})\Big) & \text{for } t=1,2,\ldots,(T-1),\\
\frac{1}{\sigma^2}(y_{st}-X_{st*}^T\beta)+\frac{1}{\tau^2}\Big(w_{s+}\eta\varphi_{s(t-1)}+\sum_{s'}w(s,s')(\varphi_{s't}-\eta\varphi_{s'(t-1)})\Big) & \text{for } t=T,
\end{cases}
$$

$$
\sigma^2_{\varphi_{st}}=\begin{cases}
\frac{\tau^2}{w_{s+}} & \text{for } t=0,\\
\Big(\frac{1}{\sigma^2}+\frac{w_{s+}(1+\eta^2)}{\tau^2}\Big)^{-1} & \text{for } t=1,2,\ldots,(T-1),\\
\Big(\frac{1}{\sigma^2}+\frac{w_{s+}}{\tau^2}\Big)^{-1} & \text{for } t=T.
\end{cases}
$$

**Posterior distribution of $\psi$ for ARMSM**: Fig. A.2 presents the dependence structure for spatiotemporal random effects $\{\tilde{\varphi}_{st}\}$ in ARMSM.

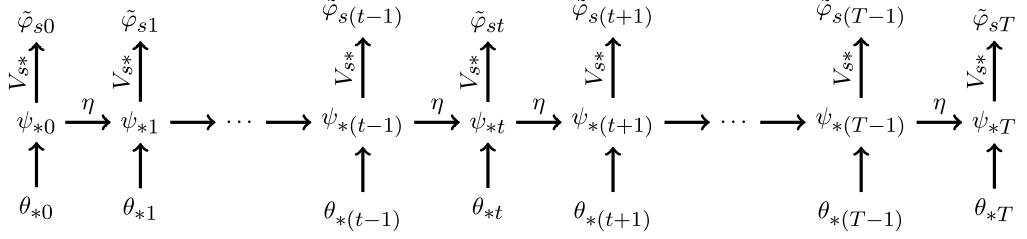



**Fig. A.2.** Hierarchical dependence structure for spatiotemporal random effects $\{\tilde{\varphi}_{st}\}$ in ARMSM.

The posterior distribution of $\psi$ can be written as follows:

$$
\begin{aligned}
\pi(\psi|-) &\propto \prod_{s=1}^{S}\pi(y_{s*}|-)\,\pi(\psi)\\
&\propto \prod_{s=1}^{S}\pi(y_{s*}|-)\prod_{i=1}^{k}\pi(\psi_{i*})\\
&\propto \prod_{s=1}^{S}\exp\Big\{-\frac{1}{2\sigma^2}\Big(y_{s*}-X_{s**}\beta-\sum_{i=1}^{k}\tilde{V}_{s*i}\psi_{i*}\Big)^T\Big(y_{s*}-X_{s**}\beta-\sum_{i=1}^{k}\tilde{V}_{s*i}\psi_{i*}\Big)\Big\}\times\prod_{i=1}^{k}\exp\Big\{-\frac{1}{2}\psi_{i*}\Omega_{\psi}^{-1}\psi_{i*}\Big\}
\end{aligned}
$$

$$
\begin{aligned}
&\propto \prod_{s=1}^{S} \exp\Big\{-\frac{1}{2\sigma^2}\Big(\sum_{i=1}^{k} \psi_{i*}^T \tilde{V}_{s*i}^T \tilde{V}_{s*i} \psi_{i*} - 2\sum_{i=1}^{k} \psi_{i*}^T \tilde{V}_{s*i}^T (y_{s*} - X_{s**}\beta)\Big)\Big\} \times \prod_{i=1}^{k} \exp\{-\frac{1}{2}\psi_{i*}\Omega_\psi^{-1}\psi_{i*}\} \\
&\propto \prod_{i=1}^{k} \exp\Big\{-\frac{1}{2\sigma^2}\Big(\sum_{s=1}^{S} \psi_{i*}^T \tilde{V}_{s*i}^T \tilde{V}_{s*i} \psi_{i*} - 2\sum_{s=1}^{S} \psi_{i*}^T \tilde{V}_{s*i}^T (y_{s*} - X_{s**}\beta)\Big)\Big\} \times \prod_{i=1}^{k} \exp\{-\frac{1}{2}\psi_{i*}\Omega_\psi^{-1}\psi_{i*}\} \\
&\propto \prod_{i=1}^{k} \exp\{-\frac{1}{2\sigma^2}\big(\psi_{i*}^T\psi_{i*} - 2\psi_{i*}^T \sum_{s=1}^{S} \tilde{V}_{s*i}(y_{s*} - X_{s**}\beta)\big)\} \times \prod_{i=1}^{k} \exp\{-\frac{1}{2}\psi_{i*}\Omega_\psi^{-1}\psi_{i*}\} \\
&\propto \prod_{i=1}^{k} \exp\Big\{-\frac{1}{2}\Big(\psi_{i*}^T\Big(\frac{1}{\sigma^2} I_T + \Omega_\psi^{-1}\Big)\psi_{i*} - 2\,\frac{1}{\sigma^2}\psi_{i*}^T \sum_{s=1}^{S} \tilde{V}_{s*i}^T (y_{s*} - X_{s**}\beta)\Big)\Big\}.
\end{aligned}
$$

Hence, the posterior distribution is $\psi|- \sim \mathcal{N}_{Tk}(\Sigma_\psi^{-1}\mu_\psi, \Sigma_\psi^{-1})$ where $\mu_\psi = (\mu_{\psi\,1}^T, \mu_{\psi\,2}^T, \ldots, \mu_{\psi\,k}^T)^T$ and $\Sigma_\psi$ is a tridiagonal matrix with

$$
\mu_{\psi\,i} = \frac{1}{\sigma^2}\sum_{s=1}^{S} \tilde{V}_{s*i}^T (y_{s*} - X_{s**}\beta) \text{ and } \Sigma_{\psi\,ii} = \frac{1}{\sigma^2} I_T + \Omega_\psi^{-1}.
$$

**Posterior distribution of $\psi$ and $\ddot{\epsilon}$ for for ARVA-MSM**: Fig. A.3 presents the dependence structure for spatiotemporal random effects $\{\ddot{\varphi}_{st}\}$ in ARVA-MSM.

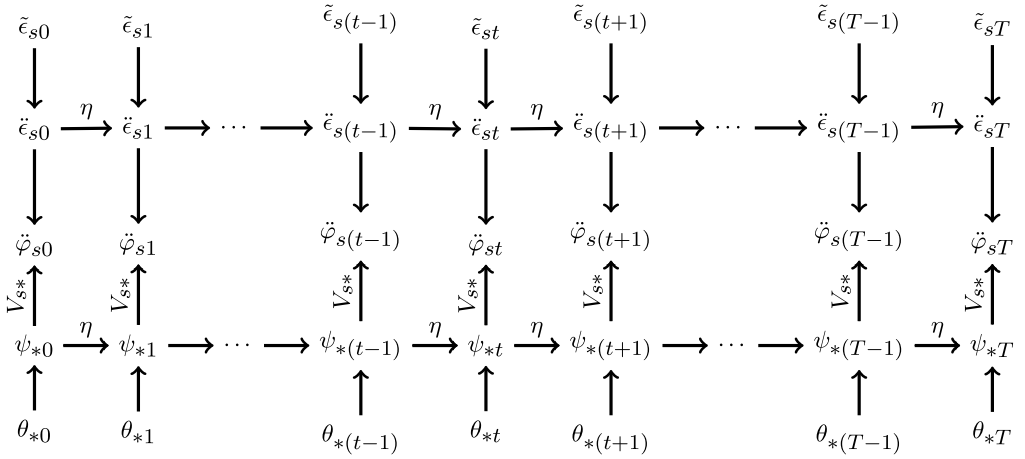



**Fig. A.3.** Hierarchical dependence structure for spatiotemporal random effects $\{\ddot{\varphi}_{st}\}$ in ARVA-MSM. Note that bottom-half the diagram represents dependence explained by $k$ Laplacian eigenvectors in the matrix $V$. The top-half of the diagram represents the heteroscedastic noise terms introduced for variance adjustment.

Temporal dependence exists within latent processes $\psi$ and $\ddot{\epsilon}$. The posterior distribution of $\psi$ is:

$$
\begin{aligned}
\pi(\psi|-) &\propto \prod_{s=1}^{S} \pi(y_{s*}|-)\,\pi(\psi) \\
&\propto \prod_{s=1}^{S} \pi(y_{s*}|-) \prod_{i=1}^{k} \pi(\psi_{i*}) \\
&\propto \prod_{s=1}^{S} \exp\Big\{ -\frac{1}{2\sigma^2}\Big( y_{s*} - X_{s**}\beta - \sum_{i=1}^{k} \tilde{V}_{s*i}\psi_{i*} - \ddot{\epsilon}_{s*} \Big)^T \Big( y_{s*} - X_{s**}\beta - \sum_{i=1}^{k} \tilde{V}_{s*i}\psi_{i*} - \ddot{\epsilon}_{s*} \Big)\Big\} \prod_{i=1}^{k} \exp\Big\{ -\frac{1}{2}\psi_{i*}\Omega_\psi^{-1}\psi_{i*} \Big\} \\
&\propto \prod_{s=1}^{S} \exp\Big\{ -\frac{1}{2\sigma^2}\Big( \sum_{i=1}^{k} \psi_{i*}^T \tilde{V}_{s*i}^T \tilde{V}_{s*i}\psi_{i*} - 2\sum_{i=1}^{k} \psi_{i*}^T \tilde{V}_{s*i}^T (y_{s*} - X_{s**}\beta - \ddot{\epsilon}_{s*}) \Big)\Big\} \times \prod_{i=1}^{k} \exp\Big\{ -\frac{1}{2}\psi_{i*}\Omega_\psi^{-1}\psi_{i*} \Big\} \\
&\propto \prod_{i=1}^{k} \exp\Big\{ -\frac{1}{2\sigma^2}\Big( \sum_{s=1}^{S} \psi_{i*}^T \tilde{V}_{s*i}^T \tilde{V}_{s*i}\psi_{i*} - 2\sum_{s=1}^{S} \psi_{i*}^T \tilde{V}_{s*i}^T (y_{s*} - X_{s**}\beta - \ddot{\epsilon}_{s*}) \Big)\Big\} \times \prod_{i=1}^{k} \exp\Big\{ -\frac{1}{2}\psi_{i*}\Omega_\psi^{-1}\psi_{i*} \Big\} \\
&\propto \prod_{i=1}^{k} \exp\Big\{ -\frac{1}{2\sigma^2}\big(\psi_{i*}^T\psi_{i*} - 2\psi_{i*}^T \sum_{s=1}^{S} \tilde{V}_{s*i}(y_{s*} - X_{s**}\beta - \ddot{\epsilon}_{s*})\big) \Big\} \times \prod_{i=1}^{k} \exp\{-\frac{1}{2}\psi_{i*}\Omega_\psi^{-1}\psi_{i*}\} \\
&\propto \prod_{i=1}^{k} \exp\Big\{ -\frac{1}{2}\Big( \psi_{i*}^T\Big(\frac{1}{\sigma^2} I_T + \Omega_\psi^{-1}\Big)\psi_{i*} - 2\,\frac{1}{\sigma^2}\psi_{i*}^T \sum_{s=1}^{S} \tilde{V}_{s*i}^T (y_{s*} - X_{s**}\beta - \ddot{\epsilon}_{s*}) \Big)\Big\}.
\end{aligned}
$$

Hence, the posterior distribution is $\psi|- \sim \mathcal{N}_{Tk}(\Sigma_\psi^{-1}\mu_\psi, \Sigma_\psi^{-1})$ where $\mu_\psi = (\mu_{\psi\,1}^T, \mu_{\psi\,2}^T, \ldots, \mu_{\psi\,k}^T)^T$ and $\Sigma_\psi$ is a tridiagonal matrix with

$$\mu_{\psi\,i} = \frac{1}{\sigma^2}\sum_{s=1}^{S}\tilde{V}_{s*i}^T(y_{s*} - X_{s**}\beta - \ddot{\epsilon}_{s*}) \text{ and } \Sigma_{\psi\,ii} = \frac{1}{\sigma^2}I_T + \Omega_\psi^{-1}.$$

The posterior distribution of $\ddot{\epsilon}$ is:

$$\begin{aligned}
\pi(\ddot{\epsilon}|-) &\propto \prod_{s=1}^{S}\pi(y_{s*}|-)\,\pi(\ddot{\epsilon}_{s*}) \\
&\propto \prod_{s=1}^{S}\exp\Big\{-\frac{1}{2\sigma^2}\Big(y_{s*} - X_{s**}\beta - \sum_{i=1}^{k}\tilde{V}_{s*i}\psi_{i*} - \ddot{\epsilon}_{s*}\Big)^T\Big(y_{s*} - X_{s**}\beta - \sum_{i=1}^{k}\tilde{V}_{s*i}\psi_{i*} - \ddot{\epsilon}_{s*}\Big)\Big\} \times \exp\Big\{-\frac{1}{2}\ddot{\epsilon}_{s*}^T\Omega_{\ddot{\epsilon}}^{-1}\ddot{\epsilon}_{s*}\Big\} \\
&\propto \prod_{s=1}^{S}\exp\Big\{-\frac{1}{2\sigma^2}\Big(\ddot{\epsilon}_{s*}^T\ddot{\epsilon}_{s*} - 2\ddot{\epsilon}_{s*}^T\Big(y_{s*} - X_{s**}\beta - \sum_{i=1}^{k}\tilde{V}_{s*i}\psi_{i*}\Big)\Big)\Big\} \times \exp\Big\{-\frac{1}{2}\ddot{\epsilon}_{s*}^T\Omega_{\ddot{\epsilon}}^{-1}\ddot{\epsilon}_{s*}\Big\} \\
&\propto \prod_{s=1}^{S}\exp\Big\{-\frac{1}{2}\ddot{\epsilon}_{s*}^T\Big(\frac{1}{\sigma^2}I_T + \Omega_{\ddot{\epsilon}}^{-1}\Big)\ddot{\epsilon}_{s*} - 2\frac{1}{\sigma^2}\ddot{\epsilon}_{s*}^T\Big(y_{s*} - X_{s**}\beta - \sum_{i=1}^{k}\tilde{V}_{s*i}\psi_{i*}\Big)\Big\}.
\end{aligned}$$

Hence, the posterior distribution is $\ddot{\epsilon}|- \sim \mathcal{N}_{ST}(\Sigma_{\ddot{\epsilon}}^{-1}\mu_{\ddot{\epsilon}}, \Sigma_{\ddot{\epsilon}}^{-1})$ where $\mu_{\ddot{\epsilon}} = (\mu_{\ddot{\epsilon}\,1}^T, \mu_{\ddot{\epsilon}\,2}^T, \ldots, \mu_{\ddot{\epsilon}\,S}^T)^T$ and $\Sigma_{\ddot{\epsilon}}$ is a tridiagonal matrix with

$$\mu_{\ddot{\epsilon}\,s} = \frac{1}{\sigma^2}\Big(y_{s*} - X_{s**}\beta - \sum_{i=1}^{k}\tilde{V}_{s*i}\psi_{i*}\Big) \text{ and } \Sigma_{\ddot{\epsilon}\,ss} = \frac{1}{\sigma^2}I_T + \Omega_{\ddot{\epsilon}}^{-1}.$$

### A.5. Analysis of spatial smoothness

In case of any spatial function, a feature of interest is its smoothness. For areal-level measurements, functions are defined on the areal units and their smoothness can be defined using the adjacency structure. We use what Belkin and Niyogi (2002) defined as the "smoothness functional": $f_G^T L f_G$ where $f_G = \{f_1, f_2, \ldots, f_S\}$ represents collection of function values on the areal units, and $L$ is the Laplacian matrix. Smaller the value of smoothness functional is, smoother the function is. Here, we report the relative smoothness which is the ratio of the smoothness functional values for the true surface and predicted surface. Relative smoothness values below and above 1 indicate under- and over-smoothed predicted surface, respectively, relative to the true surface.

We present comparison of spatial smoothness between the two proposed methods and CAR for the third simulation function (Function-1 on large, regular grid) from Section 6.2.2 of the main article. For each replication of the single block spatial CV, we compute the relative smoothness of the test surface, as defined above, from each method. Since we have repeated the Single block spatial CV 30 times, 30 relative smoothness values are obtained method-wise for each function. Fig. A.4 displays the box plots of these values. Whereas the median relative smoothness of CAR is found to be closer to 1 compared to the proposed methods, it exhibits much higher variability and extremes (significantly over- and under-smoothed predictions) across replications. The relative smoothness values from MSM and VA-MSM remain stable across replications, although the former is seen to generate predictions that are marginally smoother and closer to the true smoothness compared to the latter.

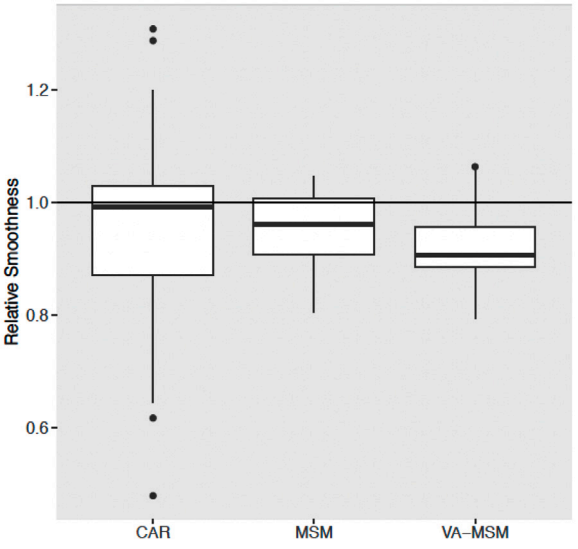



**Fig. A.4.** Box plots comparing the relative smoothness of the predictions over the test blocks generated using CAR, MSM and VA-MSM for Function-1 on a large, regular grid.

### A.6. Approximate generalized inverse for Laplacian

To implement the stretch approximation with given error bound $E_0$, one has to find the required number of eigenpairs $m_0$ such that $\lambda_{(m_0+1)} \geq \left(\frac{1}{\lambda_{(S)}} + 2\,\frac{E_0}{\lambda_{(2)}}\right)^{-1}$. Ways to find $m_0$ include (i) using IRL-LSFLE algorithm in a trial-and-error mode (that we refer to as direct computation) or, (ii) running a batchwise computation. The implementation of the latter proceeds as follows. Let us denote the batch size as $B$. First, evaluate the largest eigenvalue $\lambda_{(S)}$ of $L$. Then compute the eigenpairs associated with $B$ largest magnitude eigenvalues of $\tilde{L} = L - (\lambda_{(S)} - b)I_S$ where $b$ is a positive real number. Let $\tilde{\lambda}_{(i)}$ be the $i$th largest eigenvalue (in terms of magnitude) of $\tilde{L}$ and $\tilde{v}_i$ is the corresponding eigenvector. It follows that, $\tilde{\lambda}_{(i)} + (\lambda_{(S)} - b)$ is the $i$th smallest eigenvalue of $L$ and $\tilde{v}_i$ is the corresponding eigenvector. If $\lambda_{(m_0+1)}$ is not found in the first batch, in the second step, redefine $\tilde{L} = \tilde{L} - \sum_{i=1}^{B} \tilde{\lambda}_{(i)}\tilde{v}_i\tilde{v}_i^T$ and compute its eigenpairs for $B$ largest magnitude eigenvalues. The eigenpairs of $L$ corresponding to next $B$ smallest eigenvalues can be evaluated from the eigenpairs of redefined $\tilde{L}$ using the same relationship as in the first step. This process is repeated until $m_0$ is determined.

For the direct computation, when a step fails to find $m_0$ (i.e. when none of computed eigenvalues exceeds the threshold), all outputs from that step are discarded and a larger number of eigenpairs are computed from scratch. On the other hand, a batchwise method retrieves the eigenpairs gradually (in small batches) so that all of them will be useful after the method stops. For that reason, one may observe lesser computation time under batchwise method in some experiments. However, it is important to note that, this approach loses the sparsity property of $L$ from the second step resulting in significant increase memory requirement.

Before comparing computational performance of these two approaches, let us check how $m_0$ varies based on our chosen error bound $E_0$. For this analysis, we consider two grid sizes by dividing the unit square into $150 \times 150$ and $200 \times 200$ blocks, and assign first and second-order neighborhoods to each of them. We consider three different candidates for $E_0$ at $0.01, 0.001$ and $0.0001$. Using direct computation, we determine the required number of eigenpairs $m_0$ to attain the corresponding error bound, as shown in Table A.1. This table reveals several interesting trends. First, given a grid size and neighborhood structure, any decrease in $E_0$ corresponds to nearly proportional jump in $m_0$. Second, for a given grid size and fixed $E_0$, lower sparsity in adjacency matrix allows a smaller $m_0$. Finally, between two grid sizes with comparable sparsity levels, the larger grid requires more eigenvectors with the difference increasing sharply as $E_0$ keeps getting smaller.

**Table A.1**
Required number of Laplacian Eigenvectors ($m_0$) in stretch approximation.

| Grid size | | 150 × 150 | | 200 × 200 | |
|---|---|---|---|---|---|
| Neighborhood | | First-order | Second-order | First-order | Second-order |
| $E_0$ = | $10^{-2}$ | 49 | 48 | 49 | 49 |
| | $10^{-3}$ | 407 | 389 | 413 | 397 |
| | $10^{-4}$ | 3132 | 2885 | 3444 | 3074 |

Returning to the comparison of aforementioned two approaches for eigenvector computation in stretch approximation, we look at peak memory consumption to determine 4000 Laplacian eigenpairs. We retain our choices for grids and neighborhoods as in Table A.1. In batchwise approach, we consider the batch size as 500 and $b = 0.1$. Table A.2 displays the striking advantage of direct computation in terms of peak memory requirement that is 50 to 100 times lower than that for the batchwise computation. As explained above, sparsity (or lack of it) becomes the key distinguishing factor. These numbers indicate that, even if the direct method has to be run several times to get to the required number of eigenpairs, its peak computational demand is very likely to remain below that of a batchwise algorithm. Therefore, we always use the former for all analysis presented in the main article.

**Table A.2**
Comparison of Peak Memory[a] demand for 4000 Laplacian Eigenpairs.

| Grid size | 150 × 150 | | 200 × 200 | |
|---|---|---|---|---|
| Neighborhood | First-order | Second-order | First-order | Second-order |
| Direct Computation | 687 | 687 | 1221 | 1221 |
| Batchwise[b] Computation | 33 436 | 33 351 | 104 980 | 105 134 |

[a] Rounded to nearest MB.
[b] Batch size = 500.

### A.7. Codes for model implementation

Since all simulation studies and real data analyses included in Sections 6 and 7 of the main article compare our proposed methods against CAR in terms of computational attributes, it is important to clarify how the parameter estimation from competing models were actually performed in software. There are packages available in R (R Core Team, 2025) to implement CAR for spatial data (CARBayes; Lee, 2013) and ARCAR for spatiotemporal data (CARBayesST; Lee et al., 2018). However, instead of using them directly, the first author had to download the source codes to make the necessary modifications as described below:

- The current versions of these packages (CARBayes 6.1.1 and CARBayesST 4.0) do not permit users to input the adjacency matrix either as a sparse matrix or as a list of region id numbers for neighbors. Consequently, with a large number of areal units, generating a neighborhood matrix becomes infeasible due to the memory requirement. To address this, we made modifications at relevant places to enable the adjacency matrix to be entered as a sparse matrix, list, or 'nb' class as specified in the 'spdep' R package (Bivand and Wong, 2018).

- The specification of ARCAR in CARBayesST package is not identical to what the authors used in this work. For the first-order autoregressive structure over time, we assumed a stationary distribution for the random effects at $t = 0$ which resulted in presence of $(1-\eta^2)$ in the distribution of $\psi_0$ in ARMSM (see Equation (4.3) in the main article) and distributions of $\psi_0$ and $\ddot{\epsilon}_{*0}$ in ARVA-MSM (see Equation (4.4) in the main article). The original implementation in the package omits that part.

It is also important to note that these R source codes update the spatial and spatiotemporal random effects (in CAR and ARCAR, respectively) and their associated parameters within the MCMC schemes utilizing C++ via the Rcpp package (Eddelbuettel and Joseph, 2018). Our proposed methods, i.e., (a) MSM, variable-$k$ MSM, VA-MSM with and without marginalization, in spatial setting, and (b) ARMSM and ARVA-MSM, in spatiotemporal setting, are also executed in R with their entire MCMC schemes implemented using C++ via the Rcpp package. Within C++, LAPACK (Anderson et al., 1999) is employed to handle some of the most computationally demanding tasks required to update the spatiotemporal random effects $\psi$ in ARMSM, and $\psi$ and $\ddot{\epsilon}$ in ARVA-MSM. Additionally, R was configured to use the Intel Math Kernel Library (Intel, 2024) for improved matrix algebra performance.

All runtimes reported in the main article are as observed while implementing the codes using a single 3.79 GHz processor in a Windows machine with 128 GB RAM.

### A.8. Contrasts between MSM and Hughes and Haran (2013)

In Hughes and Haran (2013), a reduced-rank model for spatial random effects on a graph is developed. Similar to MSM, their specification also involves (i) use of adjacency matrix $W$ and (ii) linear combination of a relatively small number of eigenvectors. However, the two models have major structural differences that we clarify in Table A.3. Here, $P_X$ denotes the usual projection matrix onto the span of covariate columns, calculated as $X(X^TX)^{-1}X^T$.

**Table A.3**
List of key differences between MSM and Hughes and Haran (2013) model.

| Feature | MSM | Hughes and Haran (2013) |
|---|---|---|
| Objective | Reduced rank representation of CAR prior | Dimension reduction to resolve confounding between fixed and random effects |
| Eigenvectors are extracted from | $L = D_w - W$ | $(I_{p+1} - P_X)\, W\, (I_{p+1} - P_X)$ |
| Eigenvalues can be | Nonnegative only | Any real value[a] |
| Prior distribution on eigenvector coefficients | Zero-mean independent normal, variance depends on corresponding eigenvalue | Zero-mean multivariate normal, dispersion matrix depends on $L$ |

[a] Although eigenvectors corresponding to negative eigenvalues were not considered for inclusion.

### A.9. Simulation study for variable-k MSM

We explore if number of eigenpairs $k$ can be set as a model parameter so that we can entirely skip the procedure of determining the optimal $k$ based on MSE and runtimes, as described in the main article. We use spike and slab prior for the members of $\theta$ vector, i.e., a priori, $\theta_i \overset{ind}{\sim} (1-\pi)\delta_0 + \pi\mathcal{N}(0, \sigma^2\tau^2/\lambda_{(i+1)})$, where $\delta_0$ denotes the degenerate distribution at 0. We set $\pi = 0.5$. Since the columns of $V$ matrix (Laplacian eigenvectors) are orthonormal, binary indicator variables corresponding to members of $\theta$ can easily be sampled in parallel (same feature has been used by Ghosh and Clyde (2011) using data augmentation for general covariate matrices). To assess its performance and compare against that of fixed-$k$ MSM, we choose one of the simulated datasets from Section 6.2.2 of the main article where Function-1 is used on an irregular grid with 83,280 cells. As shown therein (see bottom left panel of Fig. 3 in main article), $k = 600$ was chosen as the optimal value for MSM. Here, we run the variable-$k$ version of MSM at that value as well as at candidate values immediate before (400) and after (800) it. In Table A.4, we present the predictive performance from single block spatial CV with 30 replications (same as what we used in the main article for this dataset) comparing against corresponding values for Fixed-$k$ MSM.

**Table A.4**
Predictive assessment of variable-$k$ MSM.

| # Eigenpairs | MSE | Absolute bias | Uncertainty | Empirical coverage |
|---|---|---|---|---|
| Single block spatial CV using Function-1 on Irregular Grid | | | | |
| 400 | 32.252 (29.533) | 3.119 (2.681) | 0.765 (1.153) | 0.243 (0.346) |
| 600 | 16.581 (15.368) | 2.738 (2.377) | 0.605 (1.086) | 0.189 (0.304) |
| 800 | 18.097 (15.270) | 2.816 (2.374) | 0.573 (1.109) | 0.174 (0.301) |

Within parentheses, corresponding values for Fixed-$k$ MSM are mentioned.

This table informs us about practical utility of Variable-$k$ MSM. First, its out-of-sample predictive performance is sensitive to how many eigenpairs are included in the $V$ matrix. Hence, just as in Fixed-$k$ MSM, one still has to search for the optimal $k$. Second, given the same value of $k$ in both variants, Variable-$k$ approach generates higher out-of-sample MSE. Its predictive intervals have smaller uncertainty and, coupled with higher absolute bias, they yield significantly lower empirical coverage than the Fixed-$k$ version. Overall, this simulation study does not provide any evidence that treating $k$ as a parameter leads to any kind of predictive improvement or reduction in computational effort.

## Codes and datasets availability

All R codes and the step-by-step instructions in README.pdf file, to run the models from this article to reproduce the included tables and figures, are available at https://github.com/KamrulHKhan/Spectral-models-for-large-spatial-and-spatiotemporal-datasets. The ACS 5-Year Median Home Value by Census Tract dataset can be extracted from tidycensus (Walker et al., 2023) R package. The dataset on Brightness (.Rdata file) is derived from the publicly available archive of Landsat imagery and can be downloaded from Dataverse at https://doi.org/10.7910/DVN/8FG2EV.

## References


Anderson, E., Bai, Z., Bischof, C., Blackford, L.S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D., 1999. LAPACK Users' Guide. SIAM.

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2026. Hierarchical Modeling and Analysis for Spatial Data, third ed. CRC Press.

Banerjee, S., Gelfand, A.E., Finley, H., 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70, 825–848.

Belkin, M., Niyogi, P., 2002. Using manifold stucture for partially labeled classification. Adv. Neural Inf. Process. Syst. 15.

Ben-Israel, A., Greville, T.N., 2003. Generalized Inverses: Theory and Applications, vol. 15, Springer Science & Business Media.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., Songini, M., 1995. Bayesian analysis of space–time variation in disease risk. Stat. Med. 14, 2433–2443.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B Stat. Methodol. 192–236.

Besag, J., Kooperberg, C., 1995. On conditional and intrinsic autoregressions. Biometrika 82, 733–746.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Statist. Math. 43, 1–20.

Bhark, E.W., Jafarpour, B., Datta-Gupta, A., 2011. A generalized grid connectivity–based parameterization for subsurface flow model calibration. Water Resour. Res. 47.

Bivand, R., Wong, D.W.S., 2018. Comparing implementations of global and local indicators of spatial association. TEST 27, 716–748. http://dx.doi.org/10.1007/s11749-018-0599-x.

Bozzo, E., Franceschet, M., 2012. Approximations of the generalized inverse of the graph Laplacian matrix. Internet Math. 8, 456–481.

Calvetti, D., Reichel, L., Sorensen, D.C., 1994. An implicitly restarted lanczos method for large symmetric eigenvalue problems. Electron. Trans. Numer. Anal. 2, 1–21.

Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Ser. B 70, 209–226.

Csárdi, G., Nepusz, T., Traag, V., Horvát, S., Zanini, F., Noom, D., Müller, K., 2024. igraph: Network analysis and visualization in R. https://CRAN.R-project.org/package=igraph.

Datta, A., Banerjee, S., Finley, A.E., 2016. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J. Amer. Statist. Assoc. 111, 800–812.

Eddelbuettel, D.B., Joseph, J., 2018. Extending R with C++: A brief introduction to rcpp. Amer. Statist. 72, 28–36. http://dx.doi.org/10.1080/00031305.2017.1375990.

Finley, A.O., Sang, H., Banerjee, S., Gelfand, A.E., 2009. Improving the performance of predictive process modeling for large datasets. Comput. Statist. Data Anal. 53, 2873–2884.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., et al., 2012. Package 'Car'. R Foundation for Statistical Computing, Vienna, p. 16.

Ghosh, J., Clyde, M.A., 2011. Rao-blackwellization for Bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach. J. Amer. Statist. Assoc. 106, 1041–1052.

Hamaamin, Y.A., Nejadhashemi, A.P., Zhang, Z., Giri, S., Woznicki, S.A., 2016. Bayesian regression and neuro-fuzzy methods reliability assessment for estimating streamflow. Water 8, 287.

Higdon, D., 2002. Space and space–time modeling using process convolutions. In: Anderson, C., Barnett, V., El-Shaarawi, P.C.A. (Eds.), Quantitative Methods for Current Environmental Issues. Springer, London, pp. 37–56.

Hughes, J., Haran, M., 2013. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. J. R. Stat. Soc. Ser. B Stat. Methodol. 75, 139–159.

Intel, 2024. Intel oneAPI math kernel library (oneMKL) developer reference.

Jin, X., Banerjee, S., Carlin, B.P., 2007. Order-free co-regionalized areal data models with application to multiple-disease mapping. J. R. Stat. Soc. Ser. B Stat. Methodol. 69, 817–838.

Karlson, M., Ostwald, M., Reese, H., Sanou, J., Tankoano, B., Mattsson, E., 2015. Mapping tree canopy cover and aboveground biomass in Sudano-Sahelian woodlands using landsat 8 and random forest. Remote. Sens. 7, 10017–10041.

Khan, M.K.H., Chakraborty, A., Petris, G., Wilson, B.T., 2021. Constrained functional regression of national forest inventory data over time using remote sensing observations. J. Amer. Statist. Assoc. 116, 1168–1180.

Kirichenko, A., Van Zanten, H., 2017. Estimating a smooth function on a large graph by Bayesian Laplacian regularisation. Electron. J. Stat. 11, 891–915.

Knorr-Held, L., 2000. Bayesian modelling of inseparable space–time variation in disease risk. Stat. Med. 19, 2555–2567.

Latimer, A.M., Wu, S., Gelfand, A.E., Silander, Jr., J.A., 2006. Building statistical models to analyze species distributions. Ecol. Appl. 16, 33–50.

Lee, D., 2013. CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. J. Stat. Softw. 55, 1–24.

Lee, W., Lee, Y., 2013. Random-effect models with singular precision. J. Statist. Plann. Inference 143, 2128–2141.

Lee, D., Mitchell, R., 2012. Boundary detection in disease mapping studies. Biostatistics 13, 415–426.

Lee, D., Rushworth, A., Napier, G., 2018. Spatio-temporal areal unit modeling in R with conditional autoregressive priors using the CARBayesST package. J. Stat. Softw. 84, 1–39. http://dx.doi.org/10.18637/jss.v084.i09.

Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with R-INLA. J. Stat. Softw. 63, 1–25.

Liu, F., Chakraborty, S., Li, F., Liu, Y., Lozano, A.C., 2014. Bayesian regularization via graph Laplacian. Bayesian Anal. 9, 449–474.
Ma, H., Carlin, B.P., Banerjee, S., 2010. Hierarchical and joint site-edge methods for medicare hospice service region boundary analysis. Biometrics 66, 355–364.
Mahdi, G.J., Chakraborty, A., Arnold, M.E., Rebelo, A.G., 2019. Efficient Bayesian modeling of large lattice data using spectral properties of Laplacian matrix. Spat. Stat. 29, 329–350.
Martins, T.G., Simpson, D., Lindgren, F., Rue, H., 2013. Bayesian computing with INLA: new features. Comput. Statist. Data Anal. 67, 68–83.
McInnes, L., Healy, J., Melville, J., 2020. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint 1802.03426v3.
Momenyan, S., Poorolajal, J., 2020. Modeling of spatio-temporally clustered survival HIV/AIDS data in the presence of competing risks setting. Spat. Stat. 38, 100460.
Nychka, D., Wikle, C., Royle, J.A., 2002. Multiresolution models for nonstationary spatial covariance functions. Stat. Model. 2, 315–331.
Paciorek, C.J., Schervish, M.J., 2006. Spatial modelling using a new class of nonstationary covariance functions. Environmetrics: Off. J. Int. Environmetrics Soc. 17, 483–506.
Paiva, T., Chakraborty, A., Reiter, J., Gelfand, A., 2014. Imputation of confidential data sets with spatial locations using disease mapping models. Stat. Med. 33, 1928–1945.
Plummer, M., Best, N., Cowles, K., Vines, K., 2006. CODA: Convergence diagnosis and output analysis for MCMC R. News 6, 7–11.
Qiu, Y., Mei, J., Guennebaud, G., Niesen, J., 2016. Rspectra: solvers for large scale eigenvalue and svd problems. R package version 0.12-0 405.
R Core Team, 2025. R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing. Vienna, Austria.
Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components. J. R. Stat. Soc. Ser. B 59, 731–792.
Risser, M.D., Calder, C.A., 2017. Local likelihood estimation for covariance functions with spatially-varying parameters: the convoSPAT package for R. J. Stat. Softw. 81, 1–32.
Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. Chapman and Hall/CRC.
Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71, 319–392.
Rushworth, A., Lee, D., Mitchell, R., 2014. A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in greater London. Spat. Spatio-Temporal Epidemiology 10, 29–38.
Rushworth, A., Lee, D., Sarran, C., 2017. An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk. J. R. Stat. Soc. Ser. C. Appl. Stat. 66, 141–157.
Sang, H., Huang, J.Z., 2012. A full scale approximation of covariance functions for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 74, 111–132.
Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4489–4497.
Truong, L.T., Kieu, L.-M., Vu, T.A., 2016. Spatiotemporal and random parameter panel data models of traffic crash fatalities in Vietnam. Accid. Anal. Prev. 94, 153–161.
Walker, K., 2022. Analyzing US Census Data: Methods, Maps, and Models in R. Chapman and Hall/CRC.
Walker, K., 2025. tigris: Load census TIGER/Line shapefiles r package version 2.2.1. http://dx.doi.org/10.32614/CRAN.package.tigris.
Walker, K., Herman, M., Eberwein, K., 2023. Tidycensus: Load US census boundary and attribute data as 'tidyverse' and 'sf'-ready data frames. R package version 1.5.
Wilson, B.T., 2015. Harmonic analysis of dense time series of landsat imagery for modeling change in forest conditions. In: Pushing Boundaries: New Directions in Inventory Techniques & Applications; Forest Inventory and Analysis (FIA) Symposium. pp. 200–205.
Zheng, S., Cao, C., Dang, Y., Xiang, H., Zhao, J., Zhang, Y., Wang, X., Guo, H., 2014. Int. J. Remote Sens. 35, 29–43.
Zhu, H., Koniusz, P., 2022. Generalized Laplacian eigenmaps. Adv. Neural Inf. Process. Syst. 35, 30783–30797.