Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Bayesian Geostatistics Using Predictive Stacking

Lu Zhang, Wenpin Tang & Sudipto Banerjee

To cite this article: Lu Zhang, Wenpin Tang & Sudipto Banerjee (2026) Bayesian Geostatistics Using Predictive Stacking, Journal of the American Statistical Association, 121:554, 1549-1561, DOI: 10.1080/01621459.2025.2566449

To link to this article: https://doi.org/10.1080/01621459.2025.2566449

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2026, VOL. 121, NO. 554, 1549–1561: Theory and Methods
https://doi.org/10.1080/01621459.2025.2566449

Bayesian Geostatistics Using Predictive Stacking

Lu Zhang[a], Wenpin Tang[b], and Sudipto Banerjee[c]

[a]Division of Biostatistics, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA; [b]Department of Industrial Engineering and Operations Research, Columbia University, New York, NY; [c]Department of Biostatistics, University of California Los Angeles, Los Angeles, CA

ABSTRACT

We develop Bayesian predictive stacking for geostatistical models, where the primary inferential objective is to provide inference on the latent spatial random field and conduct spatial predictions at arbitrary locations. We exploit analytically tractable posterior distributions for regression coefficients of predictors and the realizations of the spatial process conditional upon process parameters. We subsequently combine such inference by stacking these models across the range of values of the hyper-parameters. We devise stacking of means and posterior densities in a manner that is computationally efficient without resorting to iterative algorithms such as Markov chain Monte Carlo (MCMC) and can exploit the benefits of parallel computations. We offer novel theoretical insights into the resulting inference within an infill asymptotic paradigm and through empirical results showing that stacked inference is comparable to full sampling-based Bayesian inference at a significantly lower computational cost. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received May 2023
Accepted September 2025

KEYWORDS
Bayesian inference; Gaussian processes; Geostatistics; Stacking

1. Introduction

Geostatistics (Cressie 1993; Chilés and Delfiner 1999; Zimmerman and Stein 2010; Banerjee 2019) refers to the study of a spatially distributed variable of interest, which in theory is defined at every point over a bounded study region of interest. Customary geostatistical modeling proceeds from a latent stochastic process over space that specifies the probability law for the measurements on the variable as a partial realization of the process over a finite set of locations. Inference is sought for the underlying spatial process, which is subsequently used for spatial predictions (Stein 1999) to grasp the scientific phenomenon under study. The spatial process is often assumed to be stationary and specified by parameters representing the sill, the nugget, the range and, possibly, the smoothness of the process. We collectively refer to these as process parameters that are often empirically estimated from measurements at sampled locations using the variogram.

Likelihood-based inference for this process is, however, thwarted by the absence of classical consistent estimators of the process parameters in a customarily preferred infill asymptotic paradigm (see, e.g., Stein 1999; Zhang 2004; Zhang and Zimmerman 2005; Kaufman and Shaby 2013; Tang, Zhang, and Banerjee 2021). Bayesian inference for geostatistical data (Handcock and Stein 1993; Berger, Oliveira, and Sansó 2001; Banerjee, Carlin, and Gelfand 2014; Li, Sun, and Zhu 2023), while not relying upon asymptotic inference, is also not entirely straightforward. Specifically, irrespective of how many spatial locations yield measurements, the likelihood fails to mitigate the prior distributions' impact on the inference. This is undesirable since prior elicitation for the process parameters is challenging. Objective priors for spatial process models have also been pursued, but interpreting such information in practice and their implications in scientific contexts are not uncontroversial. The related question of how effectively (or poorly) the realized data identify these process parameters (in an exact sense from finite samples) has also received commentary (Hodges 2013; Bose, Hodges, and Banerjee 2018; De Oliveira and Han 2022).

It is, therefore, not unreasonable to pursue methods that will yield robust inference for the spatial process and for spatial predictions of the outcome at arbitrary points ("kriging") while circumventing inference on the weakly identified parameters. Instead of seeking families of prior distributions for such parameters, recent efforts at computationally efficient algorithms for geostatistical models have proposed multi-fold cross-validation methods (Finley et al. 2019) to fix the values of weakly identified parameters. However, the metrics for ascertaining optimal values of such parameters are somewhat arbitrary and may not offer robust inference.

Our primary contribution here is to develop and explore Bayesian predictive stacking of geostatistical models. Stacking is a model averaging procedure for generating predictions (Wolpert 1992; Breiman 1996; Clyde and Iversen 2013). Stacking methods and algorithms in diverse data analytic applications are rapidly evolving and a comprehensive review is beyond the scope of this manuscript. Significant developments of stacking methodology in Bayesian analysis have been achieved

CONTACT Sudipto Banerjee *sudipto@ucla.edu* Department of Biostatistics, University of California Los Angeles, 650 Charles E. Young Drive South, Los Angeles, CA 90095-1772

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

in recent years (Le and Clarke 2017; Yao et al. 2018, 2021; Yao, Vehtari, and Gelman 2022), but, to the best of our knowledge, developments in the context of spatial data analysis are lacking. Stacking can be regarded as an alternative to Bayesian model averaging (Madigan et al. 1996; Hoeting et al. 1999). Assume that there are G candidate models $\mathbb{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_G\}$ and each model $\mathcal{M}_g$ is indexed according to a set of fixed values of certain spatial covariance parameters so that the exact posterior distribution is analytically tractable (Section 2). We follow a specific formulation of a spatial hierarchical linear model that seamlessly evinces the familiar closed-form posterior distributions of regression coefficients, random effects and variance components.

The inferential properties of conventional stacking are largely available for exchangeable models that do not apply to geostatistics. Therefore, we offer theoretical insights within an infill asymptotic paradigm by exploiting tractability offered by the Matérn covariance kernel and conjugate Bayesian linear regression (Section S2). Section 3 offers implementation details, including two algorithms: (i) stacking of means, which combines posterior predictive means; and (ii) stacking of posterior predictive densities (Yao et al. 2018), which combines posterior predictive densities. These methods are evaluated theoretically and empirically through simulation experiments (Section 4) demonstrating that stacked inference is comparable to full Bayesian inference using MCMC at significantly less computational expense. An illustrative data analysis is presented in Section 5 including comparisons with machine learning interpolation. Section 6 concludes with pointers to future research.

2. Bayesian Spatial Models and Stacking Algorithms

2.1. Overview

Let $y(s)$ be a spatially indexed outcome at location $s \in \mathcal{D} \subset \mathbb{R}^d$ and $x(s)$ is a $p \times 1$ vector of predictors observed at s. A customary geostatistical model is

$$y(s) = x(s)^{\mathrm{T}}\beta + z(s) + \varepsilon(s), \tag{2.1}$$

where β is the $p \times 1$ vector of slopes, $z(s) \sim \mathrm{GP}(0, \sigma^2 R_\Phi(\cdot,\cdot))$ is a zero-centered spatial Gaussian process on $\mathbb{R}^d$ with spatial correlation function $R_\Phi(\cdot,\cdot)$ indexed by parameters Φ, σ^2 is the spatial variance parameter ("partial sill") and $\varepsilon(s) \sim \mathcal{N}(0, \tau^2)$ is a white noise process with variance τ^2 ("nugget") capturing measurement error. Processes $z(\cdot)$ and $\epsilon(\cdot)$ are assumed to be independent. Let $\chi = \{s_1, \ldots, s_n\} \in \mathcal{D}$ be a set of n spatial locations yielding measurements $y = (y(s_1), \ldots, y(s_n))^{\mathrm{T}}$ with known values of predictors at these locations collected in the $n \times p$ full rank matrix $X = (x(s_1), \ldots, x(s_n))^{\mathrm{T}}$. We let $z = (z(s_1), \ldots, z(s_n))^{\mathrm{T}}$ denote the realization of $z(s)$ over χ and let $R_\Phi(\chi) = (R_\Phi(s_i, s_j))_{1 \le i,j \le n}$ be the $n \times n$ spatial correlation matrix constructed from the correlation function.

Bayesian modeling extends (2.1) by assigning proper prior distributions to the parameters $\{\beta, z, \sigma^2, \Phi, \tau^2\}$ and drawing samples from the posterior distribution $p(\beta, \sigma^2, \Phi, \tau^2 \mid y)$. However, evaluating the posterior distribution, particularly when employing iterative Markov chain Monte Carlo algorithms (MCMC, Robert and Casella 1999), is cumbersome due to slow convergence of weakly identified process parameters. Instead, we exploit exact distributions for conjugate Bayesian spatial models by fixing some process parameters and subsequently implement stacked posterior inference over candidate values of the process parameters.

Our approach can be broadly described as follows. We separate our model parameters $\Theta = \{\theta_1, \theta_2\}$ into two sets $\theta_1 = \{\beta, z, \sigma^2\}$ and $\theta_2 = \{\Phi, \delta^2\}$, where $\delta^2 = \tau^2/\sigma^2$. We use conjugate distribution theory so that $p(\theta_1 \mid y, \theta_2)$ is available in closed form. We are primarily concerned with Bayesian inference on θ_1 averaging out the effects of θ_2. Therefore, our desired posterior distribution is $p(\theta_1 \mid y) = \int p(\theta_1 \mid y, \theta_2) p(\theta_2 \mid y) d\theta_2$. The key bottleneck here is that $p(\theta_2 \mid y)$ is intractable and will require either MCMC or iterative quadrature such as Integrated Nested Laplace Approximations (INLA, Rue, Martino, and Chopin 2009) or variational inference (Ren et al. 2011; Blei, Kucukelbir, and McAuliffe 2017). However, such algorithms are thwarted by convergence issues arising especially from weakly identified parameters Φ in the spatial correlation kernel. Therefore, we reformulate the inference problem by writing $p(\theta_1 \mid y) = \int p(\theta_1 \mid y, \theta_2) p(\theta_2 \mid y) d\theta_2 \approx \sum_{g=1}^{G} w_g p(\theta_1 \mid y, \theta_{2g})$, where the collection of weights w_g replace $p(\theta_2 \mid y)$. While this may seem to resemble quadrature, a key distinction is that we find the weights using convex optimization with scoring rules and do not attempt to approximate $p(\theta_2 \mid y)$. Once the optimal weights, say $\hat{w}_g$ are computed, posterior inference for quantities of interest subsequently proceed from the "stacked posterior" $\tilde{p}(\cdot \mid y) = \sum_{g=1}^{G} \hat{w}_g p(\cdot \mid y, \theta_{2g})$. To circumvent iterative algorithms, we employ conjugate Bayesian spatial models with closed form expressions for $p(\theta_1 \mid y, \theta_2)$ and then proceed to obtain the stacked posterior by averaging over candidate values of θ_2. Section 2.2 derives the analytically accessible models that are used in Section 2.3 where we devise the stacking methodology and algorithms.

2.2. Conjugate Bayesian Spatial Model

We extend (2.1) to a conjugate Bayesian hierarchical spatial model,

$$\begin{aligned} &y \mid z, \beta, \sigma^2 \sim \mathcal{N}(X\beta + z, \delta^2\sigma^2 I_n), \quad z \mid \sigma^2 \sim \mathcal{N}(0, \sigma^2 R_\Phi(\chi)), \\ &\beta \mid \sigma^2 \sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta), \quad \sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma), \end{aligned} \tag{2.2}$$

where we fix Φ, the noise-to-spatial variance ratio $\delta^2 := \frac{\tau^2}{\sigma^2}$, and μ_β, V_β, a_σ, and b_σ are fixed hyper-parameters specifying the prior distributions for β and σ^2. This design enables closed-form posterior distributions, as summarized below (also see Kitanidis 1986; Handcock and Stein 1993; Gaudard et al. 1999; Banerjee 2020, for exact Bayesian inference). Let $\gamma = (\beta^{\mathrm{T}}, z^{\mathrm{T}})^{\mathrm{T}}$. The posterior distribution of (γ, σ^2) from (2.2) is

$$p(\gamma, \sigma^2 \mid y, \Phi, \delta^2) = \underbrace{\mathrm{IG}(\sigma^2; a_\sigma^*, b_\sigma^*)}_{p(\sigma^2 \mid y)} \times \underbrace{\mathcal{N}(\gamma; \hat{\gamma}, \sigma^2 M_*)}_{p(\gamma \mid \sigma^2, y)}, \tag{2.3}$$

where $\hat{\gamma} = M_* X_*^{\mathrm{T}} V_*^{-1} y_*$, $y_* = [y, \mu_\beta, 0]^{\mathrm{T}}$, $a_\sigma^* = a_\sigma + n/2$, $b_\sigma^* = b_\sigma + \frac{1}{2}(y_* - X_*\hat{\gamma})^{\mathrm{T}} V_*^{-1}(y_* - X_*\hat{\gamma})$, $M_*^{-1} = X_*^{\mathrm{T}} V_*^{-1} X_*$, $X_*^{\mathrm{T}} = \begin{bmatrix} X^{\mathrm{T}} & I_p & 0 \\ I_n & 0 & I_n \end{bmatrix}$ and $V_* = \begin{bmatrix} \delta^2 I_n & 0 & 0 \\ 0 & V_\beta & 0 \\ 0 & 0 & R_\Phi(\chi) \end{bmatrix}$. The posterior distribution $p(\gamma \mid y, \Phi, \delta^2)$ obtained after integrating out σ^2 is multivariate Student's t (i.e., $t_{2a_\sigma^*}(\gamma; \hat{\gamma}, (b_\sigma^*/a_\sigma^*)M_*)$) with degrees of freedom $2a_\sigma^*$, location $\hat{\gamma}$ and scale matrix $(b_\sigma^*/a_\sigma^*)M_*$. The derivation follows standard results from the Normal-Gamma family of distributions (Kitanidis 1986; Handcock and Stein 1993). The posterior distributions remain well-defined in the limit as $\delta^2 \to 0$, as elaborated in Section S1 in the supplement.

Furthermore, let $\tilde{\chi} = \{\tilde{s}_1, \ldots, \tilde{s}_m\}$ be a set of m unknown points in $\mathcal{D}$, $\tilde{z}$ and $\tilde{y}$ be the $m \times 1$ vectors with elements $z(\tilde{s}_i)$ and $y(\tilde{s}_i)$ for $i = 1, 2, \ldots, m$. Let $\tilde{X} = (x(\tilde{s}_1), \ldots, x(\tilde{s}_m))^{\mathrm{T}}$ be the $m \times p$ matrix that carries the values of predictors at $\tilde{\chi}$ and let $J_\Phi(\chi, \tilde{\chi}) = (R_\Phi(s, s'))_{\{s \in \chi, s' \in \tilde{\chi}\}}$ Then, spatial predictive inference follows from the posterior distribution

$$p(\tilde{z}, \tilde{y} \mid y, \Phi, \delta^2) = \int p(\tilde{y} \mid \tilde{z}, \beta, \sigma^2, \Phi, \delta^2) p(\tilde{z} \mid z, \sigma^2, \Phi, \delta^2) \, p(\gamma, \sigma^2 \mid y, \Phi, \delta^2) \, d\gamma \, d\sigma^2 \,, \tag{2.4}$$

which is again a multivariate t distribution with degrees of freedom $2a_\sigma^*$, location $\tilde{\mu}$ and scale matrix $(b_\sigma^*/a_\sigma^*)\tilde{M}$ where $\tilde{\mu} = W\hat{\gamma}$, $\tilde{M} = WM_*W^{\mathrm{T}} + M_2$, $M_1 = R_\Phi(\tilde{\chi}) - J_\Phi^{\mathrm{T}}(\chi, \tilde{\chi}) R_\Phi^{-1}(\chi) J_\Phi(\chi, \tilde{\chi})$, $W = \begin{bmatrix} 0 & J_\Phi^{\mathrm{T}}(\chi, \tilde{\chi}) R_\Phi^{-1}(\chi) \\ \tilde{X} & J_\Phi^{\mathrm{T}}(\chi, \tilde{\chi}) R_\Phi^{-1}(\chi) \end{bmatrix}$ and $M_2^{-1} = \begin{bmatrix} \frac{1}{\delta^2} I_m + M_1^{-1} & -\frac{1}{\delta^2} I_m \\ -\frac{1}{\delta^2} I_m & \frac{1}{\delta^2} I_m \end{bmatrix}$. The predictive distributions $p(z(s_0) \mid y, \Phi, \delta^2)$ and $p(y(s_0) \mid y, \Phi, \delta^2)$ are also available in analytic form as a univariate t distributions for any single point $s_0 \in \mathcal{D}$. Bayesian inference proceeds from exact posterior samples obtained from (2.3). We first draw values of $\sigma^2 \sim IG(a_\sigma^*, b_\sigma^*)$ followed by a single draw of $\gamma \sim N(\hat{\gamma}, \sigma^2 M_*)$ for each drawn value of σ^2. This yields samples $\{\gamma, \sigma^2\}$ from (2.3). Predictive inference for the latent process $z(s_0)$ and the outcome $y(s_0)$ is obtained by sampling from (2.4) by drawing a value of $\tilde{z} \sim N(\mu_z(\gamma), \sigma^2 M_1)$ with $\mu_z(\gamma) := J_\Phi^{\mathrm{T}}(\chi, \tilde{\chi}) R_\Phi^{-1}(\chi) z$ for each value of $\{\gamma, \sigma^2\}$ drawn above (see sec. 3.4 in Banerjee 2020), then drawing a value of $\tilde{y} \sim N(\tilde{X}\beta + \tilde{z}, \sigma^2 \delta^2 I_m)$ for each drawn value of β (extracted from γ), σ^2 and $\tilde{z}$.

This direct sampling is possible if Φ and δ^2 are fixed. However, these parameters are not consistently estimable (Zhang 2004; Tang, Zhang, and Banerjee 2021), and trying to estimate them from the data impedes the convergence of the MCMC algorithms. Diggle and Ribeiro (2007) proposed inference with discrete priors on these parameters (Ribeiro Jr et al. 2007), which still entails evaluating potentially numerically unstable conditional posterior densities. Alternate approaches that use K-fold cross-validation have been explored with limited success (Finley et al. 2019). Instead, we avoid numerically computing marginal posterior distributions and pursue optimization based on stacking over a set of fixed values of $\{\Phi, \delta^2\}$.

2.3. Stacking Algorithms for Bayesian Spatial Models

Let $\{\mathcal{M}_g, g = 1, \ldots, G\}$ be the set of candidate models. The Bayes predictor for $y(s_0)$ under model $\mathcal{M}_g$, for each $g = 1, \ldots, G$, is $\mathbb{E}_g(y(s_0) \mid y, \mathcal{M}_g)$, where $\mathbb{E}_g(\cdot \mid y, \mathcal{M}_g)$ is the expectation with respect to $p(y(s_0) \mid y, \mathcal{M}_g) = t_{2a_\sigma^*}(y(s_0); h_g^{\mathrm{T}}\hat{\gamma}_g, (b_\sigma^*/a_\sigma^*) h_g^{\mathrm{T}} M_* h_g)$ with $h_g^{\mathrm{T}} = [x^{\mathrm{T}}(s_0), J_{\Phi_g}(s_0, \chi) R_{\Phi_g}^{-1}(\chi)]$ and a_σ^*, b_σ^*, $\hat{\gamma}_g$ and M_* given by (2.3) with $\Phi = \Phi_g$ and $\delta^2 = \delta_g^2$. Stacking will combine the G Bayes predictors as a weighted average,

$$\sum_{g=1}^{G} w_g \mathbb{E}_g(y(s_0) \mid y, \mathcal{M}_g) = \sum_{g=1}^{G} w_g h_g^{\mathrm{T}} \hat{\gamma}_g, \tag{2.5}$$

where $\{w_1, \ldots, w_G\}$ are the stacking weights. We refer to (2.5) as the stacked predictor. Subject to the constraint that stacking weights are nonnegative and their sum equals one, we define the corresponding stacked predictive density as $\sum_{g=1}^{G} w_g p(y(s_0) \mid y, \mathcal{M}_g)$. We consider two stacking algorithms: *stacking of means* and *stacking of predictive densities*.

Stacking of means: This is the most natural stacking algorithm adapted from Breiman (1996). Define the leave-one-out (LOO) Bayes predictor for $y(s_i)$ under model $\mathcal{M}_g$ as $\hat{y}_g(s_i) = \mathbb{E}_g(y(s_i) \mid y_{-i}, \mathcal{M}_g) = h_{g,-i}^{\mathrm{T}} \hat{\gamma}_{g,-i}$, where y_{-i} is the data without the ith observation, $h_{g,-i}^{\mathrm{T}}$ is defined as in (2.5) but with s_i and $\chi_{-i} = \chi \setminus \{s_i\}$ replacing s_0 and χ, respectively, and $\hat{\gamma}_{g,-i}$ is obtained from (2.3) applied to the data without s_i. The expectation $\mathbb{E}_g(\cdot \mid y_{-i}, \mathcal{M}_g)$ is calculated with respect to $p(y(s_i) \mid y_{-i}, \mathcal{M}_g) = t_{2a_{\sigma,-i}^*}\left(y(s_i); h_{g,-i}^{\mathrm{T}} \hat{\gamma}_{g,-i}, \frac{b_{\sigma,-i}^*}{a_{\sigma,-i}^*} h_{g,-i}^{\mathrm{T}} M_{*,-i} h_{g,-i} + \delta^2\right)$, where $a_{\sigma,-i}^*$, $b_{\sigma,-i}^*$ and $M_{*,-i}$ are, again, provided in (2.3) for the data excluding s_i. Stacking of means determines the optimal weights as

$$\arg\min_{w} \sum_{i=1}^{n} \left(y(s_i) - \sum_{g=1}^{G} w_g \hat{y}_g(s_i) \right)^2 . \tag{2.6}$$

Stacking of predictive densities: Following the generalized Bayesian stacking framework in Yao et al. (2018), we devise a second stacking algorithm for spatial analysis, which we refer to as stacking of predictive densities. This algorithm finds the distribution in the convex hull $\mathcal{C} = \{\sum_{g=1}^{G} w_g \times p(\cdot \mid \mathcal{M}_g) : \sum_g w_g = 1, w_g \geq 0\}$ that is optimal according to some proper scoring functions. Here $p(\cdot \mid \mathcal{M}_g)$ refers to the distribution of interest under model $\mathcal{M}_g$. Let $\mathcal{S}_1^G = \{w \in [0,1]^G : \sum_{g=1}^{G} w_g = 1\}$ and $p_t(\cdot \mid y)$ be the true posterior predictive distribution. Using the logarithmic score (corresponding to the KL divergence), we seek w so that

$$\max_{w\in\mathcal{S}_1^G}\frac{1}{n}\sum_{i=1}^{n}\log\left(\sum_{g=1}^{G}w_g\underbrace{t_{2a^*_{\sigma,-i}}\left(y(s_i);h_{g,-i}^{\mathrm{T}}\hat{\gamma}_{g,-i},\frac{b^*_{\sigma,-i}}{a^*_{\sigma,-i}}h_{g,-i}^{\mathrm{T}}M_{*,-i}h_{g,-i}+\delta^2\right)}_{p(y(s_i)\,|\,y_{-i},\mathcal{M}_g)}\right). \tag{2.7}$$

The optimal distribution $\sum_{g=1}^{G}w_g p(y(s_i)\,|\,y_{-i},\mathcal{M}_g)$ provides a "likelihood" of observing $y(s_i)$ on location s_i given other data. Therefore, $\prod_{i=1}^{n}\sum_{g=1}^{G}w_g p(y(s_i)\,|\,y_{-i},\mathcal{M}_g)$ serves as a pseudo-likelihood that measures the performance of prediction based on the weighted average of the LOO predictors for all observed locations.

Stacking using K-fold cross-validation predictors: Solving stacking weights relies upon computing the Bayes predictor and predictive density. Computing the exact LOO Bayes predictor and predictive densities for all observed locations $\{s_1,\ldots,s_n\}$ requires refitting a model n times. For a Gaussian latent variable model with the number of parameters larger than the sample size n, there are limited choices for approximating LOO predictors accurately without the onerous computation (see, e.g., Vehtari et al. 2016). Instead of using the LOO predictors, computing predictors through K-fold cross-validation (CV) is more practical. For each fold in K-fold CV, Cholesky decompositions—a major computational bottleneck—are required for each fold, whereas for LOO-CV, they are required for each observation. As a result, LOO-CV is typically more computationally intensive (see, e.g., Vehtari and Lampinen 2002, and references therein for computational details including greater computational savings and numerical stability achieved by K-fold CV). Using K-fold cross-validation instead of LOO in stacking is explored in Breiman (1996), who demonstrated that 10-fold cross-validation provides more efficient predictors than LOO. If the data is partitioned into K folds and $y[-k]$ denotes the observed outcomes that are not included in the kth fold, then, following (2.7), we have $p(y(s_i)\,|\,y[-k],\mathcal{M}_g) = t_{2a^*_\sigma[-k]}\left(y(s_i);h_g[-k]^{\mathrm{T}}\hat{\gamma}_g[-k],\frac{b^*_\sigma[-k]}{a^*_\sigma[-k]}h_g[-k]^{\mathrm{T}}M_*[-k]\right.$ $\left.h_g[-k]+\delta^2\right)$, where s_i is in kth folder. Here, $a^*_\sigma[-k]$, $b^*_\sigma[-k]$, $h_g[-k]$, and $M_*[-k]$ correspond to $a^*_{\sigma,-i}$, $b^*_{\sigma,-i}$, $h_{g,-i}$, and $M_{*,-i}$ in $p(y(s_i)\,|\,y_{-i},\mathcal{M}_g)$, but are derived using data excluding the kth folder instead of just the ith observation. The K-fold cross-validation Bayes predictor for $y(s_i)$ under model $\mathcal{M}_g$ is $\hat{y}_g(s_i)=\mathbb{E}_g(y(s_i)\,|\,y[-k],\mathcal{M}_g)=h_g[-k]^{\mathrm{T}}\hat{\gamma}_g[-k]$. For stacking of predictive densities, the optimal distribution changes into $\sum_{g=1}^{G}w_g p(y(s_i)\,|\,y[-k],\mathcal{M}_g)$.

Reconstructing stacked posterior distributions: Once the stacking weights are calculated using either stacking of means or stacking of predictive densities, we use them to reconstruct the posterior distributions of interest as

$$p(\cdot\,|\,y)=\sum_{g=1}^{G}\hat{w}_g p(\cdot\,|\,y,\mathcal{M}_g)\,, \tag{2.8}$$

where $\cdot$ represents the inferential quantity of interest. We refer to (2.8) as the *stacked posterior density*. For parameter inference we take $\cdot$ as $\{\beta,z,\sigma^2\}$, while for predictive inference we use $y(s_0)$ at an arbitrary location.

3. Implementation of Stacking Algorithms

Our theoretical development and implementation of the stacking algorithms center on Matérn models. Supplement S2 provides theoretical insights into posterior inference under the Matérn framework and supports the use of stacking in this context. The Matérn model is a special case of (2.1), where $z(s)$ is modeled with the isotropic Matérn correlation function,

$$R_\Phi(s,s'):=\frac{(\phi|s-s'|)^\nu}{\Gamma(\nu)2^{\nu-1}}K_\nu(\phi|s-s'|),\quad \Phi=\{\phi,\nu\}\,. \tag{3.1}$$

Here ϕ is the decay parameter, $\nu>0$ is a fixed smoothness parameter, $\Gamma(\cdot)$ is the Gamma function, and $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν (Abramowitz and Stegun 1965, sec. 10).

We outline algorithms that compute the weights for spatial stacking. We first partition the data into K-folds based on locations. Let $X=[x(s_1):\cdots:x(s_n)]^{\mathrm{T}}$ be the design matrix with $X[k]$, $y[k]$ and $\chi[k]$ denoting the predictors, outcome and observed locations from kth fold, respectively, and $X[-k]$, $y[-k]$ and $\chi[-k]$ denoting respective data not in kth fold. Let n_k be the number of observed locations for the kth fold. The values of the prefixed hyper-parameters $\{\phi,\nu,\delta^2\}$ of the conjugate Bayesian spatial regression model are picked from the grid G_{all}, which is expanded over the grids of candidate values as the Cartesian product $G_\phi\times G_\nu\times G_\delta^2$ for $\{\phi,\nu,\delta^2\}$. We compute the posterior expectation $\mathbb{E}(y[k]\,|\,y[-k],\phi,\nu,\delta^2)$ for $k=1,\ldots,K$ and all candidate $\{\phi,\nu,\delta^2\}$ when using stacking of means to obtain the stacking weights. Algorithm 1 describes the procedure with additional details in Section S3.1. Algorithm 1 structures the computation for different choices of δ^2 to be nested in each folder k. This structure allows the re-use of the correlation matrix for the same $\{\phi,\nu\}$ in the same folder for different δ^2, but it only works for the shared candidate values G_δ^2.

Let $\hat{Y}_{kg}$ be the $n_k\times 1$ vector with elements $\mathbb{E}(y[k]\,|\,y[-k],\phi,\nu,\delta^2)$ for each $\{\phi,\nu,\delta^2\}\in G_{all}$, where $g=1,\ldots,G$ indexes the distinct combination of $\{\phi_g,\nu_g,\delta_g^2\}\in G_{all}$ and $G=|G_{all}|$. We suppress the index g in Algorithm 1 for ease of notation. We construct $\hat{Y}_g=(\hat{Y}_{1g}^{\mathrm{T}},\ldots,\hat{Y}_{Kg}^{\mathrm{T}})^{\mathrm{T}}$ as the $n\times 1$ vector with $n=\sum_{k=1}^{K}n_k$, and $\hat{Y}=[\hat{Y}_1:\ldots:\hat{Y}_G]$ as an $n\times G$ matrix. Algorithm 1 describes the explicit steps for computing $\mathbb{E}(y[k]\,|\,y[-k],\phi,\nu,\delta^2)$. Let $w=(w_1,w_2,\ldots,w_G)^{\mathrm{T}}$ be the stacking weights obtained as $\underset{w}{\mathrm{argmin}}\{(y-\hat{Y}w)^{\mathrm{T}}(y-\hat{Y}w)\}$ under constraints $\sum_{g=1}^{G}w_g=1$ and $w_g\geq 0$ for $g=1,\ldots,G$. We formulate this as a quadratic programming (QP) problem

Algorithm 1 Computing stacking weights using stacking of means

Input:, X, y, χ, prior parameters μ_β, V_β, a_σ, b_σ, G_ϕ, G_ν, G_{δ^2}, and K
Output: $w = \{w_{\phi,\nu,\delta^2}\}_{(\phi,\nu,\delta^2)\in G_{all}}$: Stacking weights
Compute $X^{(k)}_{\text{prod}} = X^{\mathrm{T}}[-k]X[-k]$, $X^{(k)}_y = X^{\mathrm{T}}[-k]y[-k]$ and n_k for $k = 1, \ldots, K$
for $\{\phi, \nu\} \in G_\phi \times G_\nu$ **do**
  for $k = 1$ to K **do**
    Calculate $R^{-1}_{\phi,\nu}(\chi[-k])$ and store $J_{\phi,\nu}(\chi[k], \chi[-k])$
    for δ^2 in G_{δ^2} **do**
      Set $W = \begin{bmatrix} X[k]^{\mathrm{T}} \\ R^{-1}_{\phi,\nu}(\chi[-k])J_{\phi,\nu}(\chi[k], \chi[-k])^{\mathrm{T}} \end{bmatrix}$ and $m_* = \begin{bmatrix} V_\beta^{-1}\mu_\beta + \delta^{-2}X^{(k)}_y \\ \delta^{-2}y[-k] \end{bmatrix}$
      Set $M_*^{-1} = \begin{bmatrix} \delta^{-2}X^{(k)}_{\text{prod}} + V_\beta^{-1} & \delta^{-2}X^{\mathrm{T}}[-k] \\ \delta^{-2}X[-k] & R^{-1}_{\phi,\nu}(\chi[-k]) + \delta^{-2}I_{n-n_k} \end{bmatrix}$
      Set $\mathbb{E}(y[k] \mid y[-k], \phi, \nu, \delta^2) = W^{\mathrm{T}}M_*^{-1}m_*$
    end for
  end for
end for
Construct $\hat{Y}$ using $\mathbb{E}(y[k] \mid y[-k], \phi, \nu, \delta^2)$ for $k = 1, \ldots K$ and $\{\phi, \nu, \delta^2\} \in G_{all}$.
Solve convex optimization problem: $\arg\min_w (y - \hat{Y}w)^{\mathrm{T}}(y - \hat{Y}w)$ under constraints $\sum_{g=1}^{G} w_g = 1$ and $w_g \geq 0$ for $g = 1, \ldots, G$, where $G = |G_{all}|$.

Algorithm 2 Stacking weights calculation using stacking of predictive densities

Input:, X, y, χ, prior parameters μ_β, V_β, a_σ, b_σ, G_ϕ, G_ν, G_{δ^2}, and K
Output: $w = \{w_{\phi,\nu,\delta^2}\}_{(\phi,\nu,\delta^2)\in G_{all}}$: Stacking weights
for $\{\phi, \nu\}$ in grid expanded by G_ϕ and G_μ **do**
  for $k = 1$ to K **do**
    for δ^2 in G^2_δ **do**
      Compute $\mathbb{E}(y[k] \mid y[-k], \phi, \nu, \delta^2)$ (Follow line 1-7 in Algorithm 1)
      for s in $\chi[k]$ **do**
        Compute $lp_{(\phi,\nu,\delta^2)}(s) = \log(p(y(s) \mid y[-k], \phi, \nu, \delta^2))$
      end for
    end for
  end for
end for
Calculate weights by maximizing $\sum_{s\in\chi} \log\left(\sum_{(\phi,\nu,\delta^2)\in G_{all}} \exp\{lp_{(\phi,\nu,\delta^2)}(s)\} * w_{(\phi,\nu,\delta^2)}\right)$ under constrains $\sum_{(\phi,\nu,\delta^2)\in G_{all}} w_{(\phi,\nu,\delta^2)} = 1$ and $w_{(\phi,\nu,\delta^2)} > 0$

and use the `quadprog` package in R and the solver `Mosek` (Andersen and Andersen 2000) in `Julia` to solve for the weights (see Section S3.5).

We obtain the stacking weights for predictive densities by evaluating the log point-wise predictive density, $(lp_{(\phi,\nu,\delta^2)}(s))$, of $y(s)$ for all locations in each fold for all candidate models. The log point-wise predictive density is derived explicitly in Section S3.2, while Section S3.3 devises a Monte Carlo algorithm for stacking of predictive densities. The stacking weights for predictive densities are calculated using *Mosek* in R and *Ipopt* in `Julia`, both employing interior-point methods for optimization problems with logarithmic objectives and linear constraints. Algorithm 2 summarizes the procedure with further details provided in Section S3.1.

While inferential performance of these algorithms is promising (theoretically and asymptotically) when the candidate values of the correlation parameters fall in a reasonable domain, the choice of G_ϕ, G_ν, and G^2_δ still impact the performance of stacking in practical analysis. Here, we offer guidance on specifying $G_\phi, G_\nu, G^2_\delta$. More comprehensive evaluation and discussion are presented in Section 4. Based on the property of Matérn kernel, popular choices for ν include 0.5, 1.0, 1.5, and 1.75 or 2. Matérn kernels with $\nu > 2$ generate overly smoothed processes and cause numerical instabilities. The range of candidate values for ϕ are determined by a lower and upper bound of range along with the choices for ν. In simulations, we choose equally spaced candidate values for G_ϕ. We note that an even grid is not equivalent to a uniform prior unlike, for example, Kazianka and Pilz (2012). G_ϕ serves as a discretized domain for ϕ and, as we described and showed in Section S4.4, one can hardly obtain inference about hyper-parameters through stacking. The choice of candidate values for δ^2 is more subtle. In our implementation, we use quantiles of beta distribution $\text{Beta}(a_1, a_2)$ to select candidate values of $\delta^2/(1+\delta^2)$, which falls between 0 and 1. This is based on the fact that when $\sigma^2 \sim \text{IG}(a_1, b)$, $\tau^2 \sim \text{IG}(a_2, b)$ and the two parameters are independent, $\delta^2/(1+\delta^2) \sim \text{Beta}(a_1, a_2)$. The two

shape parameters a_1, a_2 are determined from values of nugget and partial sill estimated from an empirical semivariogram. We choose b to be the larger value of the estimated nugget and partial sill in our implementation. Since the posteriors of σ^2 and τ^2 are not independent, other choices for G_δ^2 are preferred when additional information about δ^2 is available.

4. Simulation

4.1. Simulation Settings

We present four simulation experiments to evaluate predictive performance using our stacking algorithms. The data for these experiments are generated using (2.1) on locations sampled uniformly over a unit square $[0,1]^2$ with R_Φ being the Matérn covariogram in (3.1). The sample size n of the simulated datasets ranges from 200 to 900, and we randomly pick $n_h = 100$ observations for checking predictive performance. The vector $x(s)$ consists of an intercept and a predictor generated from a standard normal distribution. We use the parameter values $\beta = (1,2)^\top$, $\sigma^2 = 1$, $\tau^2 = 1$, $\nu = 1$, and $\phi = 7$ and 2 to generate data for the first and second simulation, respectively. For the third and fourth simulation, we set $\phi = 20$ and 2, respectively, with $\sigma^2 = 1$, $\tau^2 = 0.3$ and $\nu = 0.5$.

We analyze our data using the $K = 10$-fold stacking Algorithms 1 and 2 with candidate values $\nu \in G_\nu = \{0.5, 1, 1.5, 1.75\}$. The candidate values for ϕ are selected so that the "effective spatial range", which refers to the distance where spatial correlation drops below 0.05, covers 0.1 and 0.6 times $\sqrt{2}$ (the maximum inter-site distance within a unit square) for all candidate values of ν. Here we set $G_\phi = \{3, 14, 25, 36\}$. It is worth noting that the actual value of ϕ in the second and fourth simulation are 2, which are smaller than the lowest candidate value in G_ϕ. The values of ϕ were deliberately chosen to be large and small to investigate the behavior of the proposed algorithms. Finally, we specify G_{δ^2} to comprise the 0.05, 0.35, 0.65, and 0.95th quantiles of a beta distribution with expectations of σ^2 and τ^2 equal to their data generating values. We assign an $\text{IG}(a_\sigma, b_\sigma)$ prior with $a_\sigma = b_\sigma = 2$ for σ^2. The prior of β is $\text{N}(\mu_\beta, V_\beta)$ where $\mu_\beta = 0$ and $V_\beta = 4 \cdot I$. For each simulated dataset, we implement stacking of means and of predictive densities to obtain the expected outcome $\hat{y}(s)$ based on the held out observed locations. The predictive accuracy is evaluated by the root mean squared prediction error over a set of n_h hold-out locations in set $\mathcal{S}_h$ ($\text{RMSPE} = \sqrt{\sum_{s\in\mathcal{S}_h}((\hat{y}(s) - y(s))^2)/n_h}$). We also compute the posterior expected values of the latent process $\hat{z}(s)$ for $z(s)$ on all of the n sampled locations in $\mathcal{S}$ and evaluate the root mean squared error for $z(s)$ ($\text{RMSEZ} = \sqrt{\sum_{s\in\mathcal{S}}(\hat{z}(s) - z(s))^2/n}$). To further evaluate the distribution of predicted values, we compute the mean log point-wise predictive density for the n_h held out locations ($\text{MLPD} = \sum_{s\in\mathcal{S}_h}\{\log(\sum_{g=1}^G w_g p(y(s) \mid y, \mathcal{M}_g))\}/n_h$).

Apart from stacking, we also implement a fully Bayesian model with priors on the hyperparameters using Markov chain Monte Carlo (MCMC) sampling for comparison. In addition, we carry out exact Bayesian inference using the conjugate model in Section 2.2 with hyper-parameters fixed at the exact value (denoted as $\mathcal{M}_0$). We use the same priors for σ^2 and β as those in stacking implementations. For the rest of the priors needed in full MCMC sampling, we assign uniform priors U(3, 36) for ϕ and U(0.25, 2) for ν, and an IG(2, 2) prior for τ^2. Sampling is fitted through the *spLM* function in the *spBayes* package in R. The diagnostic metrics are computed based on 1000 posterior samples retained after convergence was diagnosed over a burn-in period of 10,000 initial iterations. The algorithm for recovering the expected z and the log point-wise predictive density based on the output of `spLM` is presented in Section S3.4 of the supplement. We monitor all diagnostic metrics for prediction for all competing algorithms. To measure uncertainty of the diagnostic metrics, we generate 60 datasets for each sample size in each simulation, fit each dataset with the four competing methods and record the diagnostic metrics of each model fitting.

4.2. Predictive Performances

The aforementioned methods exhibit different behaviors in the four simulation studies. Figure 1 compares predictive performance for the first and the third study. The results for Simulation 2 and 4 closely mirror those of the first simulation and are included in Section S4.1 of the supplement for brevity. There are no pronounced distinctions in predictive performance among the competing models in the first simulation. In the third simulation, however, stacking of means seems to deliver better estimates for the latent process at observed and unobserved locations than stacking of predictive densities (based on RMSEZ), while stacking of predictive densities outperforms stacking of means in terms of the log point-wise predictive density (based on MLPD). This is unsurprising as we optimize prediction error in the stacking of means and we maximize the log predictive densities in the stacking of predictive densities. Treating the fully Bayesian model with priors on all hyperparameters (fitted using MCMC) as a benchmark, we find that stacking of predictive densities is very competitive in terms of MLPD. The performance of latent process estimation for the full Bayesian model falls between stacking of means and stacking of predictive densities. All competing algorithms deliver very similar prediction accuracy for the outcome at unobserved locations, while stacking of means slightly outperforms the full Bayesian model with regard to the medians of the RMSPEs for all fittings; both are slightly better than stacking of predictive densities, The conjugate Bayesian model M_0 provides the best point estimates for the outcome and latent processes based on RMSPE and RMSEZ, but is less impressive in terms of MLPD. These results indicate that stacking of means excels in point estimation, while stacking of predictive densities is preferable for interval estimation.

We check the counts of the nonzero weights in stacking and find that stacking of means tends to produce a slightly smaller number of nonzero weights than stacking of predictive densities. The number of nonzero weights is small for both stacking algorithms. This sparsity is not an artifact of our methodology but rather a known characteristic of stacking. As first reported by (Section 9 Breiman 1996), the author observed that stacking combines a "surprisingly few" number of models. Our findings are consistent with this observation: on average, there are around 3.6 and 4.3 out of 64 weights that are greater than 0.001 in the simulation studies for stacking of means and stacking of predic-

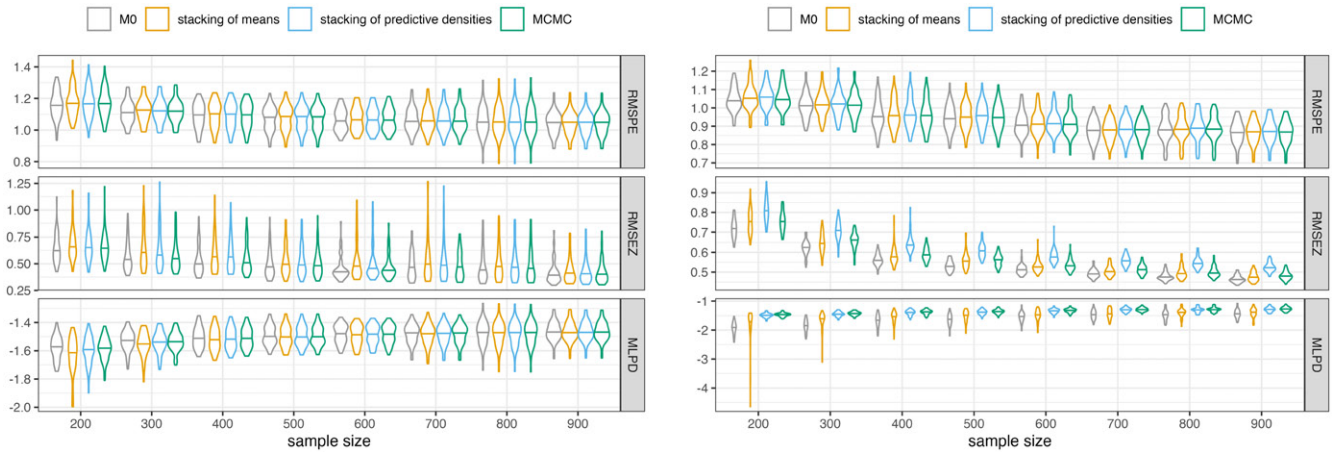

Figure 1. Distributions of the diagnostic metrics for prediction performance for the first simulation (left) and the third simulation (right). Each distribution is depicted through a violin plot. The horizontal line in each violin plot indicates the median.

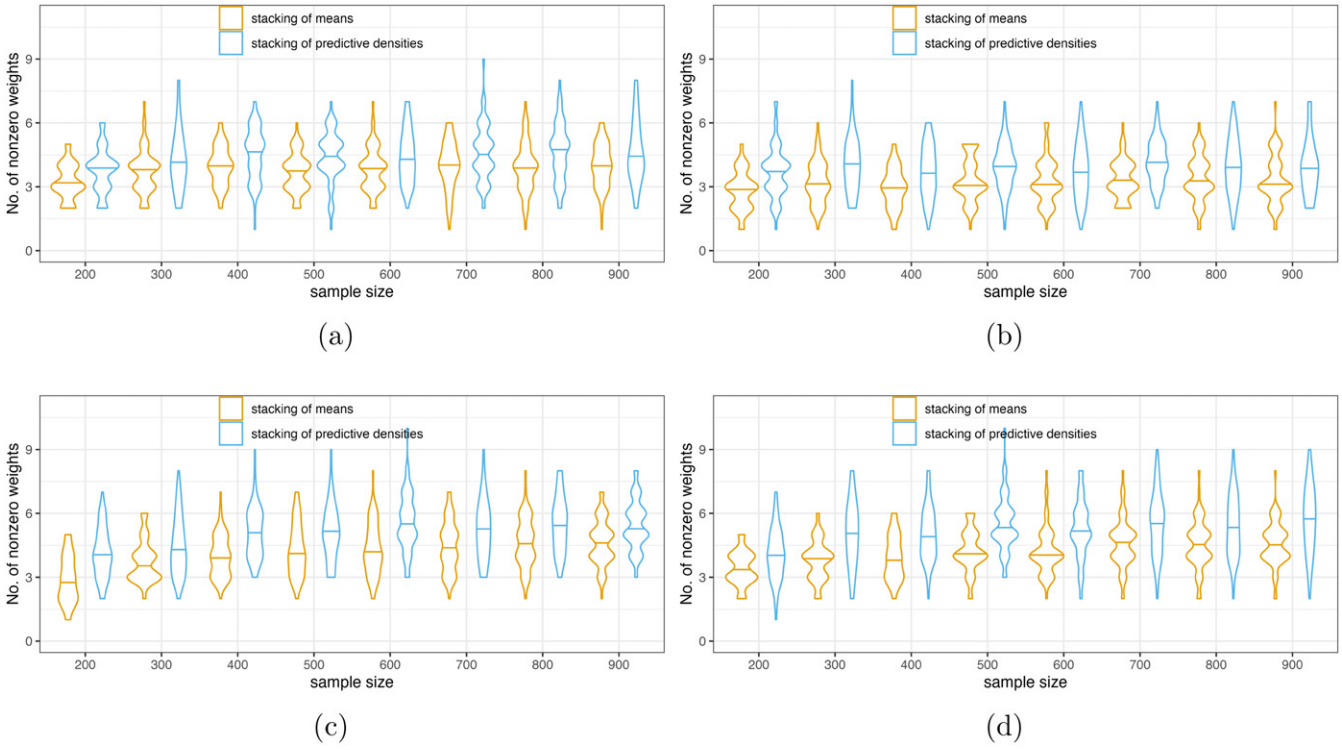

Figure 2. Distributions of the counts of nonzero weights in the first (a), second (b), third (c), and fourth (d) simulation. The distribution of the counts are described through violin plots whose horizontal lines indicate the medians.

tive densities, respectively. This number is relatively consistent when the sample sizes increase. Plots for the distributions of nonzero weights counts are provided in Figure 2. To further explore this, we visualize the distributions of the nonzero weight values in Supplement S4.2. These plots reveal a strongly right-skewed distribution, confirming that stacking not only selects a small set of models but further concentrates the predictive influence on just a few top performers. This inherent parsimony highlights stacking's ability to perform implicit model selection.

Building on the preceding broader analysis, we now attend to more specific scrutiny of inferential performance. We examine a case from Simulation 1 featuring 800 observations, another from Simulation 2 with 600 observations, a third from Simulation 3 containing 400 observations, and a fourth from simulation 4 containing 200 observations to enhance our evaluation of stacking's predictive performance. These four examples are intentionally selected to represent typical inferential behavior across varying parameter settings and sample sizes. The results from these selected examples are consistent with our empirical findings across 60 replicates and align well with the diagnostic metrics for predictive performance reported for the full set of simulations. Figure 3 directly compares 95% credible intervals (CIs) and point estimates obtained through stacking and MCMC methods for the first and the third cases. Results for the second and fourth cases, which closely resemble those of the first case, are provided in Section S4.1 of the supplement for brevity. Here, the 95% CIs for stacking are based on 900 draws from the stacked posterior distribution. Specifically, we generate 900

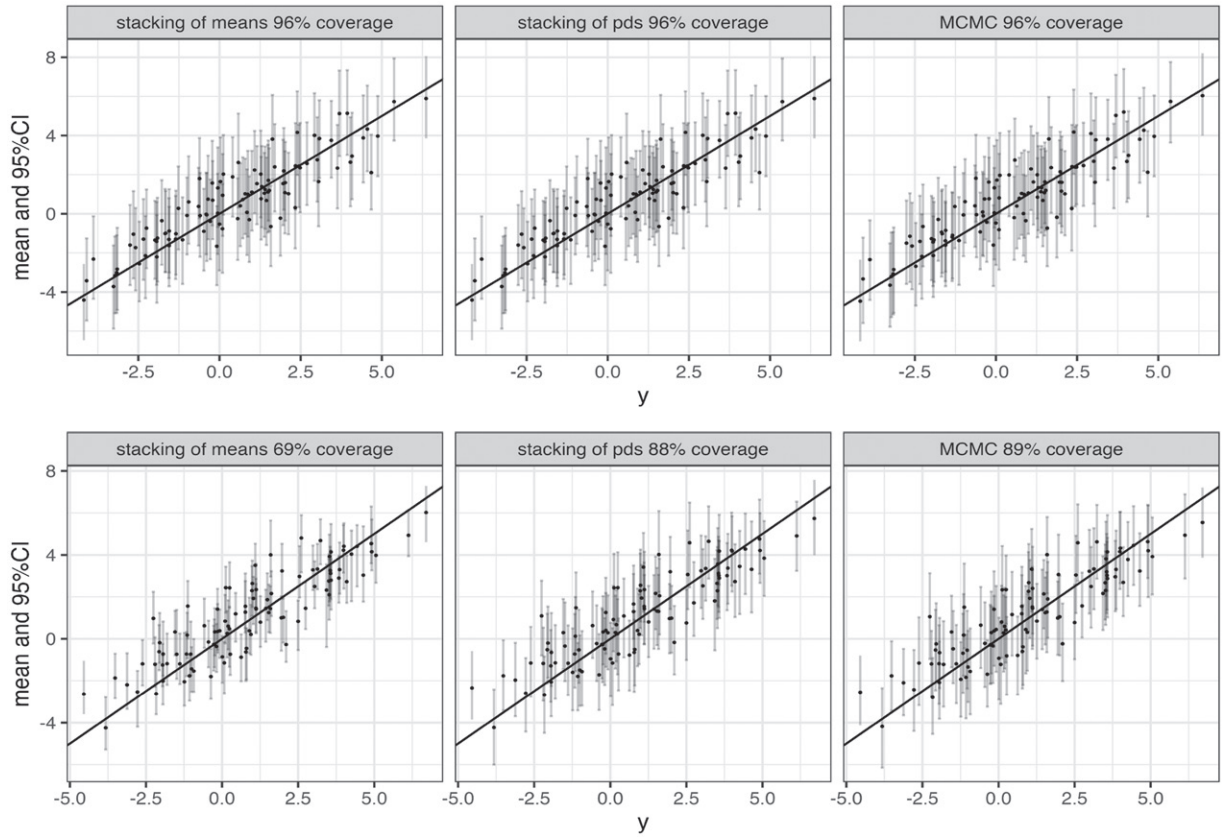

Figure 3. Scatterplots for predicted versus actual outcomes at 100 unobserved locations with 95% credible intervals from Simulation 1 with 800 observations (upper row) and another example from Simulation 3 with 400 observations (lower row). Each plot includes a solid black line representing the 45-degree reference line, with titles indicating 95% CI coverage. "pds" denotes predictive densities.

posterior samples from each candidate model with nonzero stacking weights and then randomly sample 900 draws from this pool according to the stacking weights. Stacking of predictive density appears to closely align with inference from MCMC, while stacking of means tends to marginally underestimate the CI widths, particularly in the setting with a smaller range. In addition, following a reviewer's suggestion, we include a comparison with the Bayesian method implemented in the `geoR` package (Ribeiro Jr et al. 2007) in Section S5 of the supplement, as an additional reference for interested readers.

Section S4.3 illustrates the interpolated maps of the predicted outcome at held out locations and the expected latent processes over all locations generated by different fitting algorithms. The posterior predictive means, $\mathbb{E}(y(s) \mid y)$, and $\mathbb{E}(z(s) \mid y)$, share similar patterns with the raw data. The mean of $z(s)$ estimated by stacking of predictive densities appears smoother than those estimated from $\mathcal{M}_0$, full Bayes and stacking of means, while the predicted mean of $y(s)$ at unobserved locations are indistinguishable across all methods.

4.3. Running Time Comparisons

Computer programs for reproducing the simulation studies are hosted in the GitHub repository *https://github.com/LuZhangstat/spatial_stacking*. Comparisons in predictive performances presented above are conducted in R. For the running time comparisons reported here, the stacking algorithms are implemented using `Julia-1.11.1`. The MCMC sampling algorithms are executed through the package `spBayes` in R-4.3.2, which relies on underlying functions written in C++ for computational efficiency. We report the time for obtaining weights for stacking, and we consider the sampling time for $\{\phi, \nu, \sigma^2, \tau^2\}$ using MCMC (no sampling of $\{\beta, z\}$ and no predictions). The timing comparisons are based upon experiments on a Linux system equipped with 64 AMD EPYC 7513 32-Core Processors. Parallel computing is performed using 8 threads. Figure 4 summarizes the running time for the three competing algorithms. On average, the stacking of means is 496 times faster than MCMC, while stacking of predictive densities is only slightly slower being around 483 times faster than MCMC sampling. These experiments clearly establish that predictive stacking algorithms are efficient alternatives to MCMC for estimating latent spatial processes and predicting spatial outcomes.

4.4. Impact of Hyperparameter Selection

We further explore improving inferential performance of stacking by judicious choices of $\{\phi, \nu, \delta^2\}$. We explore the inferential impact of selecting candidate values for the prefixed hyperparameters from their posterior distributions. These posterior distributions are evaluated from MCMC samples of the full Bayesian model. The method for choosing candidate values as outlined in previous subsections is now referred to as the

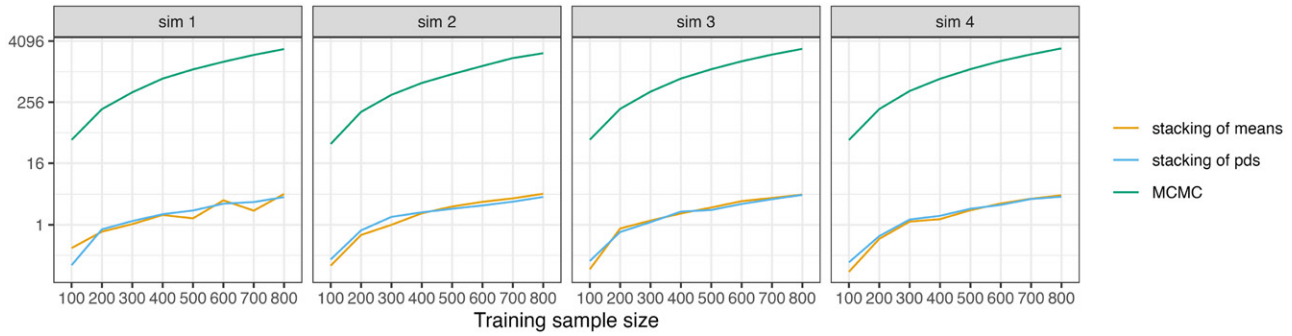

Figure 4. Running time comparison for stacking and MCMC sampling.

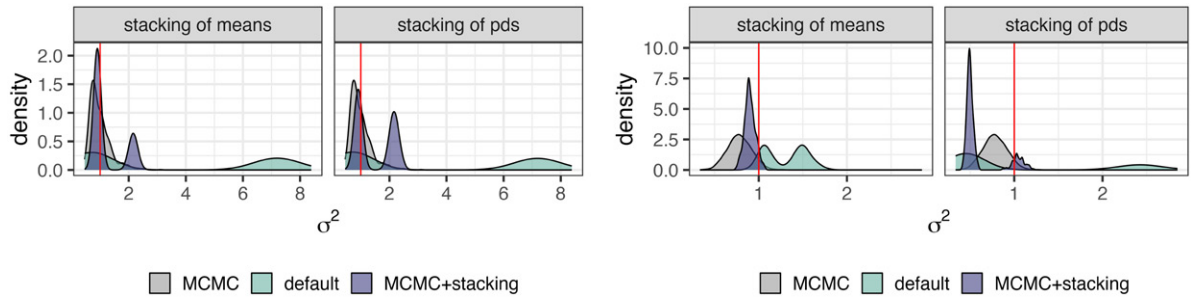

Figure 5. Densities of σ^2 for the example with 800 observations from simulation 1 (left) and the example with 400 observations from simulation 2 (right). Vertical red lines indicate the actual σ^2 values. Grey densities represent MCMC-recovered posterior distributions of σ^2. "Default" and "MCMC+Stacking" show stacking results using two methods for selecting ϕ, ν, δ^2 candidates. Left panel: stacking of means. Right panel: stacking of predictive densities.

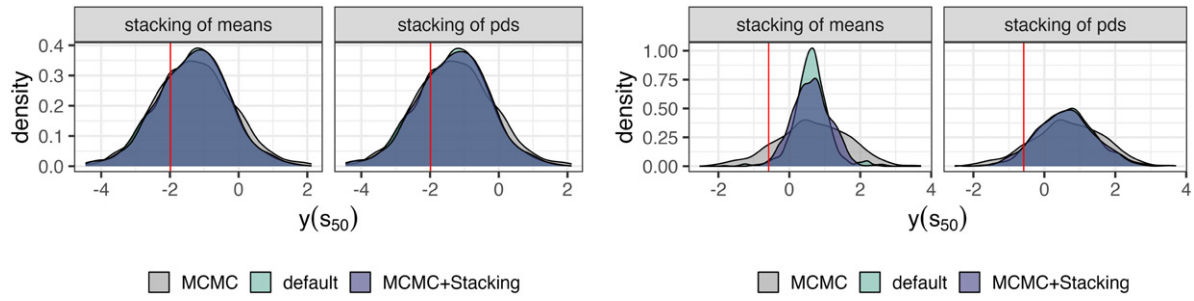

Figure 6. Predictive densities of the outcome at 50th point in the example with 800 observations from simulation 1 (left) and the example with 400 observations from simulation 3 (right). Vertical red lines indicate the actual values. Grey densities represent MCMC-recovered posterior distributions. "Default" and "MCMC+Stacking" show stacking results using two methods for selecting ϕ, ν, δ^2 candidates. Left panel: stacking of means. Right panel: stacking of predictive densities

"default" method. Given that the default algorithm in the simulation studies involves 64 candidate models, we randomly select 64 samples of $\{\phi, \nu, \delta^2\}$ as candidate values. Although obtaining marginal posterior samples is impractical when implementing the stacking algorithm, this approach serves as our benchmark, assuming full knowledge of the marginal posterior. We also present the posterior distributions recovered by MCMC as the gold standard for comparisons.

Figure S16 depicts a comparison of diagnostic metrics and reveals that there is no significant improvement in predictive performance by selecting the prefixed correlation parameters through posterior distributions. To further check the impact, we revisit the four selected examples in Section 4.2, featuring 800, 600, 400, and 200 observations from simulations 1, 2, 3, and 4, respectively. Representative results from the first and third examples are included in the main paper, with all simulation results detailed in the supplement for completeness. Figure 5 compares the posterior distributions for σ^2. Regardless of the method used for selecting candidate values for the prefixed hyper-parameters, the marginal distributions recovered through stacking invariably exhibits multimodal behavior. This reinforces the claim that stacking does not provide effective inference for covariance parameters, including σ^2. Furthermore, we observe that the distribution of σ^2 recovered by stacking is highly dependent on the choice of candidate values for the correlation parameter. Similarly, for τ^2, we observe multimodality in the third example, as shown in Figure S18. Intriguingly, the variations in the distribution of σ^2 resulting from stacking do not impact the predictive distribution of outcomes. We examine the predictive distribution of the outcome at several unobserved points and present some typical examples in Figure 6 and S20 . The predictive distributions obtained from stacking are largely consistent regardless of hyperparameter selection methods. Additionally, the predictive distributions recovered by stacking of means tend to concentrate around the mode compared to those recovered by stacking of predictive

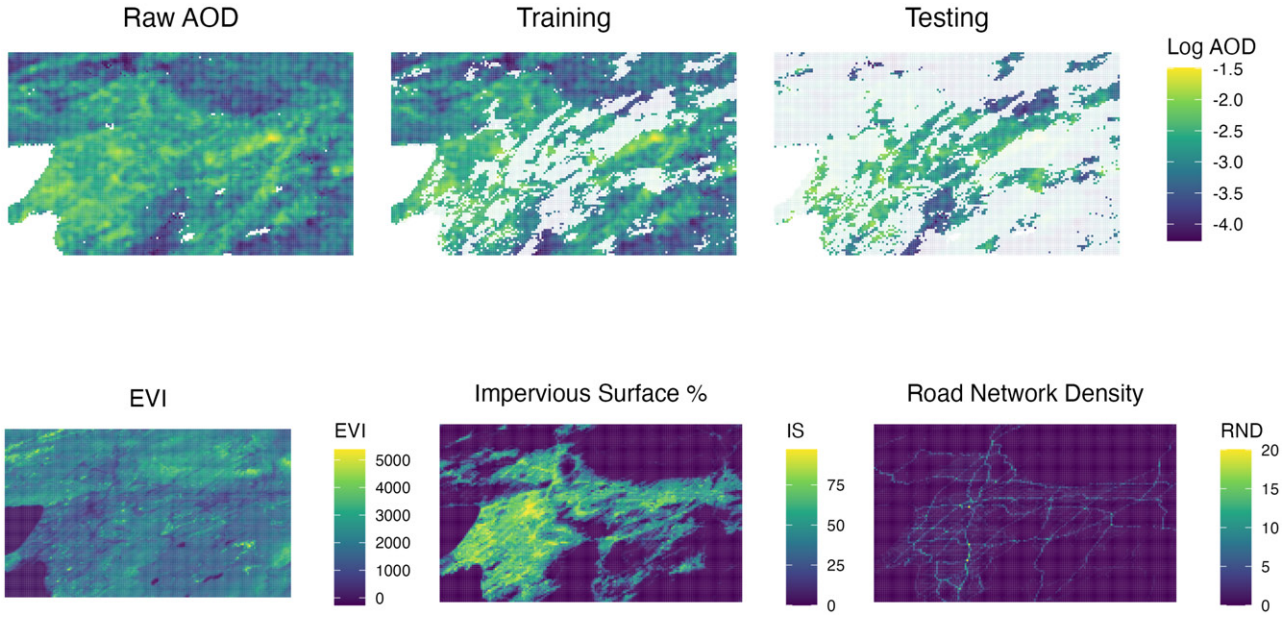

Figure 7. Upper: MODIS Aerosol Optical Depth (AOD) Visualization. From left to right: Log-transformed AOD values, training data (pixels not covered by clouds), and testing data (cloud-covered pixels), as of September 13th, 2018. Lower: Visualization of Regression Model Predictors. From left to right: Enhanced Vegetation Index (EVI) from MODIS MOD13A2 products, impervious surface percentage from USGS NLCD 2018, and weighted road network density from OpenStreetMap, all resampled to 1-km resolution.

densities. The distributions for the intercept recovered by stacking may have larger variance according to Figure S19 , while those for regression coefficients β_2 closely align with inference from MCMC. We conclude that the improvement of selecting candidate values for $\{\phi, \delta^2, \nu\}$ from the posterior is limited based on these results.

5. AOD Prediction

We use $K = 10$-fold Bayesian stacking to analyze Aerosol Optical Depth (AOD) observations from satellite technologies in global aerosol research. This is an increasingly important field across multiple disciplines such as environmental health, climatology, atmospheric science, and remote sensing (Voiland 2010). Unlike ground-based monitoring that is limited by regional coverage and budget, satellite-derived AOD data provides a more expansive picture of aerosol distributions on a global scale. However, cloud screening and conditions of high surface reflectance can result in a significant proportion of missing data in AOD satellite observations (Li et al. 2009). Prevailing AOD interpolation algorithms often rely upon random forests and neural networks (Fan and Sun 2023; Aguilera et al. 2023). Nonetheless, Gaussian process models present a competitive alternative by accommodating data-driven processes and providing essential uncertainty quantification. We apply our Bayesian spatial regression model using stacking to analyze 1-km MODIS AOD products (MCD19A2) over the Greater Los Angeles area (Lyapustin et al. 2018) and assess their predictive performances in the context of large-scale AOD retrieval and uncertainty quantification.

Figure 7 is a base image with near-complete AOD coverage on September 13th, 2018, encompassing 16,003 pixels that do not encroach over water bodies. We use the cloud pattern from August 24th, 2018, to partition the data. Pixels not obscured by clouds comprise the training set (totaling 11,857 pixels), while the remaining 4146 pixels form the testing set. We use log-transformed AOD as the outcome and five predictors (resampled to a 1-km resolution): the x-y coordinates, the Enhanced Vegetation Index (EVI) from the 16-day MODIS MOD13A2 products, the impervious surface percentage from the USGS National Land Cover Database (NLCD) 2018, and the weighted road network density from OpenStreetMap (Figure 7).

Following the prototype in simulation studies, we set $G_\nu = (0.5, 1.0, 1.5, 1.75)$. The candidate values for ϕ were set at $(0.1, 0.4, 0.7, 1.0)$ so that the practical range of the process spans from 5km to 25km. This range was conservatively estimated based on the empirical semivariogram of the residual of a linear regression model (Figure S22). We used the estimated values for σ^2 and τ^2 from the empirical semivariogram to determine G_{δ^2}. We assigned $\sigma^2 \sim \text{IG}(2, 0.05)$ and the remaining settings following those used in our simulation studies. In addition to stacking, we expanded the analysis by including five competitive algorithms: Deep Learning, Random Forest (RF) (Breiman 2001), Gradient Boosting Machine (GBM) (Friedman 2001), an Ensemble Model integrating the aforementioned three algorithms, and Bayesian Linear Regression (BLR) without spatial effects. We implemented the first four algorithms on H2O (Cook 2016), a popular open-source platform designed for big data analytics. Specifically, we used H2O's default settings for the Deep Learning algorithm (Candel et al. 2016) and similarly for the GBM. For the RF model, we determined the optimal configuration, setting the number of trees to 120 and the maximum tree depth to 60-via a Cartesian grid search. We validated these learner models with 10-fold cross-validation. H2O's Ensemble method is another stacking algorithm which finds the optimal combination of predictions from the fitted DL, RF, and GBM models. The BLR model was fitted through R package `brms`. We assigned $N(0, 2^2)$ to our regression coefficients and a half-Cauchy prior half-Cauchy$(0, 0.5)$ to the standard deviation of the error.

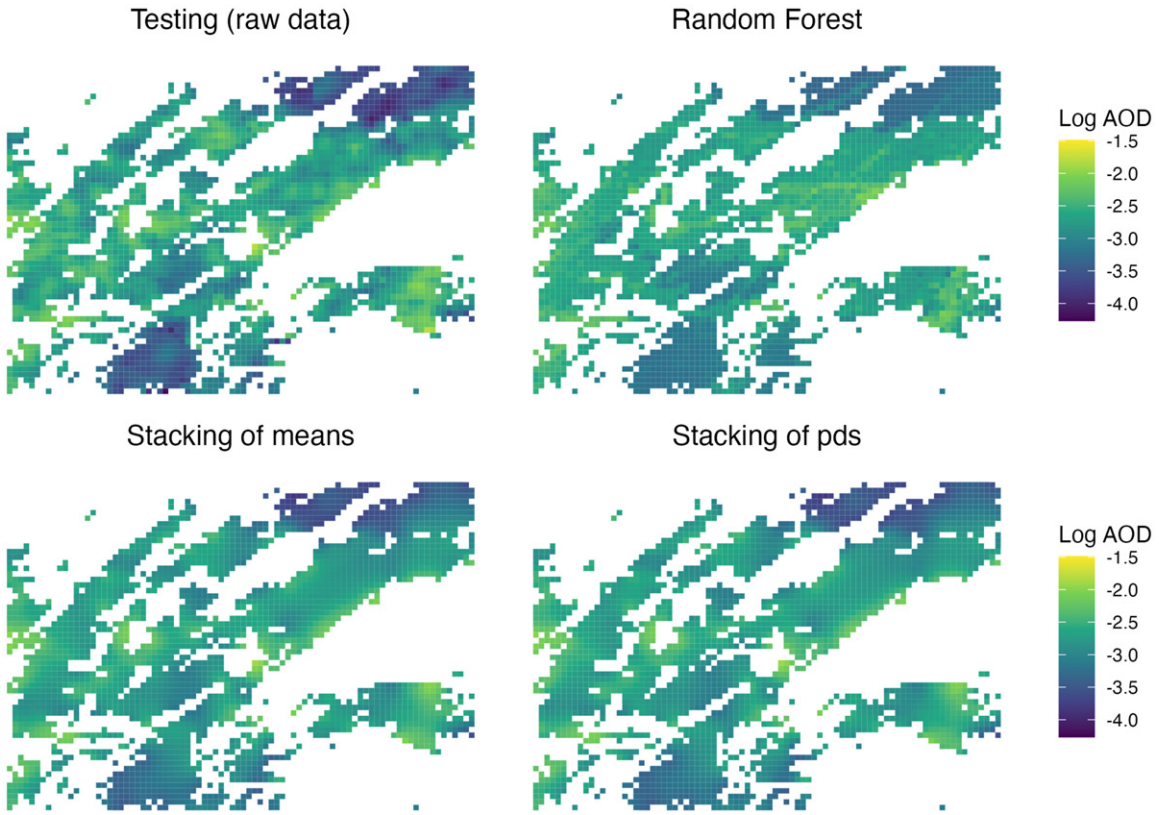

Figure 8. Interpolated and testing data AOD in the selected region.

Table 1. Comparative performance metrics for AOD interpolation.

Method	RMSPE	MAE	R	95% CIC	99% CIC
Bayesian linear regression	0.019	0.0147	0.693	**94.9%**	**99.4%**
Deep Learning (H2O)	0.0166	0.0127	0.792	NA	NA
Gradient Boosting (H2O)	0.0161	0.0122	0.797	NA	NA
Random Forest (H2O)	0.0153	0.0115	0.821	NA	NA
Ensemble model (H2O)	0.0154	0.0116	0.816	NA	NA
Stacking of means	**0.010**	**0.007**	**0.927**	75.9%	87.0%
Stacking of posterior density	**0.010**	**0.007**	0.926	82.1%	93.9%

NOTE: Best results are in bold.

For stacking and BLR, the AOD interpolation was retrieved using posterior mean. We calculated the root mean squared prediction error (RMSPE), mean absolute error (MAE) and Pearson correlation coefficient (R) by comparing the interpolated AOD with the testing data AOD, and we have summarized these results in Table 1. The two stacking algorithms significantly outperformed the other competitive algorithms in terms of Pearson correlation coefficients, RMSPE and MAE. Figure 8 showcases predictions at a selection of central testing locations, where we note that the stacking algorithms produce smoother interpolations. We compare the 95% CI for the testing AOD in Figure S23 and provide the CI coverage (CIC) for 95% CI and 99% CI in Table 1. Although the BLR model achieved excellent CI coverage, the CIs it produced are remarkably wider than those obtained through stacking. This analysis demonstrates the effectiveness of our stacking algorithms in AOD interpolation, notably outperforming prevailing models in accuracy and uncertainty quantification with limited predictors, showing great potential for future remote sensing data interpolation.

6. Conclusion and Future Work

We devise Bayesian inference for geostatistical data using predictive stacking. The underlying idea is to optimally average a finite set of analytically tractable conditional posterior distributions by fixing the values of certain parameters in the spatial covariance kernel. We explore inferential performance through simulations and analysis of an AOD data. The empirical results reveal that Bayesian predictive stacking delivers predictions comparable to full Bayesian inference obtained using MCMC samples, but at significantly lower computational costs. Our proposed algorithms are implemented in parallel using efficient storage and, hence, comprise an efficient alternative to full Bayesian inference using MCMC samples. We can build upon our current framework to extend Bayesian stacking for multivariate geostatistics using conjugate matrix-variate normal-Wishart families (Zhang, Banerjee, and Finley 2021) and conjugate exponential families for non-Gaussian data (Bradley, Holan, and Wikle 2020). A recent article by Pan et al. (2024) extends the current framework to spatial-temporal generalized linear models and another by Presicce and Banerjee (2025) develops a Bayesian transfer learning framework for massive spatial datasets using predictive stacking. Stacked Bayesian inference for high-dimensional geostatistics (e.g., building on the conjugate frameworks in Zhang, Datta, and Banerjee 2019; Banerjee 2020) is also possible as is the development of stacking methods for pooling inference across subsets of data (e.g., as an alternative to meta-kriging described in Guhaniyogi and Banerjee 2018). Finally, the R package `spStack` available for download from *https://cran.r-project.org/package=spStack* implements Bayesian predictive

stacking for Gaussian and non-Gaussian data based on the methodology developed in this article.

Acknowledgments

The authors thank the Editor, Associate Editor, and three anonymous reviewers for several helpful comments that improved the manuscript.

Supplementary Materials

The online supplement includes additional theoretical results (Sections S1 and S2), computational details (Section S3), additional results from simulation experiments (Sections S4 and S5) designed to investigate the robustness and computational efficiency of predictive stacking for spatial data analysis, and some additional predictive analysis of the AOD dataset (Section S6). A consolidated version including the main article and supplement is also available as an arXiv preprint at *https://arxiv.org/abs/2304.12414v4*.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Lu Zhang was supported by the National Institutes of Health (NIH) under Grant P20HL176204 and by the Department of Health and Human Services (HHS) under grant P30ES007048. Wenpin Tang was supported by the National Science Foundation under grant DMS 2113779. Sudipto Banerjee was supported, in part, by the National Science Foundation (NSF) through grant DMS-2113778 and by the National Institute of Health through NIGMS R01GM148761.

ORCID

Lu Zhang http://orcid.org/0000-0001-6829-703X
Wenpin Tang http://orcid.org/0000-0001-7228-1954
Sudipto Banerjee http://orcid.org/0000-0002-2239-208X

References

Abramowitz, M., and Stegun, A. (1965), *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, Mineola, NY: Dover. [1552]

Aguilera, R., Luo, N., Basu, R., Wu, J., Clemesha, R., Gershunov, A., and Benmarhnia, T. (2023), "A Novel Ensemble-based Statistical Approach to Estimate Daily Wildfire-Specific PM2. 5 in California (2006–2020)," *Environment International*, 171, 107719. [1558]

Andersen, E. D., and Andersen, K. D. (2000), "The MOSEK Interior Point Optimizer for Linear Programming: An Implementation of the Homogeneous Algorithm," in *High Performance Optimization*, eds. H. Frenk, K. Roos, T. Terlaky, S. Zhang, pp. 197–232, New York: Springer. [1553]

Banerjee, S. (2019), "Geostatistics for Environmental Processes," in *Handbook of Environmental and Ecological Statistics*, eds. A. E. Gelfand, M. Fuentes, J. A. Hoeting and R. L. Smith, Boca Raton, FL: CRC Press, pp. 81–96. [1549]

——— (2020), "Modeling Massive Spatial Datasets Using a Conjugate Bayesian Linear Modeling Framework," *Spatial Statistics*, 37, 100417. [1550,1551,1559]

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: CRC Press. [1549]

Berger, J. O., Oliveira, V. D., and Sansó, B. (2001), "Objective Bayesian Analysis of Spatially Correlated Data," *Journal of the American Statistical Association*, 96, 1361–1374. [1549]

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112, 859–877. [1550]

Bose, M., Hodges, J. S., and Banerjee, S. (2018), "Toward a Diagnostic Toolkit for Linear Models with Gaussian-Process Distributed Random Effects," *Biometrics*, 74, 863–873. [1549]

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2020), "Bayesian Hierarchical Models with Conjugate Full-Conditional Distributions for Dependent Data from the Natural Exponential Family," *Journal of the American Statistical Association*, 115, 2037–2052. [1559]

Breiman, L. (1996), "Stacked Regressions," *Machine Learning*, 24, 49–64. [1549,1551,1552,1554]

——— (2001), "Random Forests," *Machine Learning*, 45, 5–32. [1558]

Candel, A., Parmar, V., LeDell, E., and Arora, A. (2016), "Deep Learning with h2o," *H2O. ai Inc*, 1–21. [1558]

Chilés, J., and Delfiner, P. (1999), *Geostatistics: Modeling Spatial Uncertainty*, New York: Wiley. [1549]

Clyde, M., and Iversen, E. S. (2013), "Bayesian Model Averaging in the m-open Framework," *Bayesian Theory and Applications*, 14, 483–498. [1549]

Cook, D. (2016), *Practical Machine Learning with H2O: Powerful, Scalable Techniques for Deep Learning and AI*, Sebastopol, CA: 'O'Reilly Media, Inc." [1558]

Cressie, N. (1993), *Statistics for Spatial Data* (rev. ed.), New York: Wiley-Interscience. [1549]

De Oliveira, V., and Han, Z. (2022), "On Information about Covariance Parameters in Gaussian Matern Random Fields," *Journal of Agricultural, Biological and Environmental Statistics*, 27, 690–712. [1549]

Diggle, P., and Ribeiro, P. (2007), *Model-based Geostatistics*, New York: Springer. [1551]

Fan, Y., and Sun, L. (2023), "Satellite Aerosol Optical Depth Retrieval based on Fully Connected Neural Network (FCNN) and A Combine Algorithm of Simplified Aerosol Retrieval Algorithm and Simplified and Robust Surface Reflectance Estimation (SREMARA)," *IEEE Journal of Selected Topics in Applied Earth Observations*. [1558]

Finley, A. O., Datta, A., Cook, B. C., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019), "Efficient Algorithms for Bayesian Nearest Neighbor Gaussian Processes," *Journal of Computational and Graphical Statistics*, 28, 401–414. [1549,1551]

Friedman, J. H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, 29, 1189–1232. [1558]

Gaudard, M., Karson, M., Linder, E., and Sinha, D. (1999), "Bayesian Spatial Prediction," *Environmental and Ecological Statistics*, 6, 147–171. [1550]

Guhaniyogi, R., and Banerjee, S. (2018), "Meta-kriging: Scalable Bayesian Modeling and Inference for Massive Spatial Datasets," *Technometrics*, 60, 430–444. [1559]

Handcock, M. S., and Stein, M. L. (1993), "A Bayesian Analysis of Kriging," *Technometrics*, 35, 403–410. [1549,1550,1551]

Hodges, J. S. (2013), *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*, Boca Raton, FL: Chapman and Hall/CRC. [1549]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors, *Statistical Science*, 14, 382–417. [1550]

Kaufman, C. G., and Shaby, B. A. (2013), "The Role of the Range Parameter for Estimation and Prediction in Geostatistics," *Biometrika*, 100, 473–484. [1549]

Kazianka, H., and Pilz, J. (2012), "Objective Bayesian Analysis of Spatial Data with Uncertain Nugget and Range Parameters," *Canadian Journal of Statistics*, 40, 304–327. [1553]

Kitanidis, P. K. (1986), "Parameter Uncertainty in Estimation of Spatial Functions: Bayesian Analysis," *Water Resources Research*, 22, 499–507. [1550,1551]

Le, T., and Clarke, B. (2017), "A Bayes Interpretation of Stacking for m-complete and m-open Settings," *Bayesian Analysis*, 12, 807–829. [1550]

Li, C., Sun, S., and Zhu, Y. (2023), "Fixed-Domain Posterior Contraction Rates for Spatial Gaussian Process Model with Nugget," *Journal of the American Statistical Association*, 119, 1336–1347. [1549]

Li, Z., Zhao, X., Kahn, R., Mishchenko, M., Remer, L., Lee, K.-H., Wang, M., Laszlo, I., Nakajima, T. & Maring, H. (2009), "Uncertainties in Satellite Remote Sensing of Aerosols and Impact on Monitoring its Long-Term Trend: A Review and Perspective," *Annales Geophysicae*, 27, 2755–2770. [1558]

Lyapustin, A., Wang, Y., Korkin, S., and Huang, D. (2018), "Modis Collection 6 Maiac Algorithm," *Atmospheric Measurement Techniques*, 11, 5741–5765. [1558]

Madigan, D., Raftery, A. E., Volinsky, C., and Hoeting, J. (1996), "Bayesian Model Averaging," in *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR*. [1550]

Pan, S., Zhang, L., Bradley, J. R., and Banerjee, S. (2024), "Bayesian Inference for Spatial-temporal Non-Gaussian Data Using Predictive Stacking," arXiv preprint arXiv:2406.04655. [1559]

Presicce, L., and Banerjee, S. (2025), "Bayesian Transfer Learning for Artificially Intelligent Geospatial Systems: A Predictive Stacking Approach," arXiv preprint arXiv:2410.09504v3. [1559]

Ren, Q., Banerjee, S., Finley, A. O., and Hodges, J. S. (2011), "Variational Bayesian Methods for Spatial Data Analysis," *Computational Statistics & Data Analysis*, 55, 3197–3217. [1550]

Ribeiro Jr, P. J., Diggle, P. J., Ribeiro Jr, M. P. J., and Suggests, M. (2007), "The geoR package," *R news*, 1, 14–18. [1551,1556]

Robert, C. P., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York, NY: Springer. [1550]

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian models by using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society*, Series B, 71, 319–392. [1550]

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer-Verlag. [1549]

Tang, W., Zhang, L., and Banerjee, S. (2021), "On Identifiability and Consistency of the Nugget in Gaussian Spatial Process Models," *Journal of the Royal Statistical Society*, Series B, 83, 1044–1070. [1549,1551]

Vehtari, A., and Lampinen, J. (2002), "Bayesian Model Assessment and Comparison Using Cross-Validation Predictive Densities," *Neural Computation*, 14, 2439–2468. [1552]

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., and Winther, O. (2016), "Bayesian Leave-One-Out Cross-Validation Approximations for Gaussian Latent Variable Models," *The Journal of Machine Learning Research*, 17, 3581–3618. [1552]

Voiland, A. (2010), "Aerosols: Tiny Particles, Big Impact," *NASA Earth Observatory*, 2. [1558]

Wolpert, D. H. (1992), "Stacked Generalization," *Neural Networks*, 5, 241–259. [1549]

Yao, Y., Pirš, G., Vehtari, A., and Gelman, A. (2021), "Bayesian Hierarchical Stacking: Some Models are (Somewhere) Useful," *Bayesian Analysis*, 1, 1–29. [1550]

Yao, Y., Vehtari, A., and Gelman, A. (2022), "Stacking for Non-Mixing Bayesian Computations: The Curse and Blessing of Multimodal Posteriors," *Journal of Machine Learning Research*, 23, 1–45. [1550]

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018), "Using Stacking to Average Bayesian Predictive Distributions," (with Discussion), *Bayesian Analysis*, 13, 917–1007. [1550,1551]

Zhang, H. (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261. [1549,1551]

Zhang, H., and Zimmerman, D. L. (2005), "Towards Reconciling Two Asymptotic Frameworks in Spatial Statistics," *Biometrika*, 92, 921–936. [1549]

Zhang, L., Banerjee, S., and Finley, A. O. (2021), "High-Dimensional Multivariate Geostatistics: A Bayesian Matrix-Normal Approach," *Environmetrics*, 32, e2675. [1559]

Zhang, L., Datta, A., and Banerjee, S. (2019), "Practical Bayesian Modeling and Inference for Massive Spatial Data Sets on Modest Computing Environments," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12, 197–209. [1559]

Zimmerman, D., and Stein, M. (2010), "Classical Geostatistical Methods," in *Handbook of Spatial Statistics*, eds. A. E. Gelfand, P. Diggle, P. Guttorp and M. Fuentes, pp. 29–44, Boca Raton, FL: CRC Press. [1549]