Generating Independent Replicates Directly from the Posterior Distribution for a Class of Spatial Hierarchical Models

Jonathan R. Bradley[1] and Madelyn Clinch[2]

Abstract

Markov chain Monte Carlo (MCMC) allows one to generate dependent replicates from a posterior distribution for effectively any Bayesian hierarchical model. However, MCMC can produce a significant computational burden. This motivates us to consider finding expressions of the posterior distribution that are computationally straightforward to obtain independent replicates from directly. We focus on a broad class of Bayesian hierarchical models for spatially dependent data, which are often modeled via a latent Gaussian process (LGP). First, we derive a new class of distributions referred to as the generalized conjugate multivariate (GCM) distribution. The GCM distribution's theoretical development follows that of the conjugate multivariate (CM) distribution with two main differences: the GCM allows for latent Gaussian process assumptions, and the GCM explicitly accounts for hyperparameters through marginalization. The development of GCM is needed to obtain independent replicates directly from the exact posterior distribution, which has an efficient regression form. Hence, we refer to our method as Exact Posterior Regression (EPR). Simulation studies with weakly stationary spatial processes and spatial basis function expansions are provided. We provide an analysis of poverty incidence from the U.S. Census Bureau, and an analysis of high-dimensional remote sensing data. Supplementary materials for this article are available online.

Keywords: Bayesian hierarchical model; Big data; Gibbs sampler; Log-Linear Models; Markov chain Monte Carlo; Non-Gaussian.

[1] (to whom correspondence should be addressed) Department of Statistics, Florida State University, 117 N. Woodward Ave., Tallahassee, FL 32306-4330, jrbradley@fsu.edu

[2] Department of Statistics, Florida State University, 117 N. Woodward Ave., Tallahassee, FL 32306-4330

1 Introduction

MCMC has become an invaluable tool in statistics and is covered in standard text books (Robert and Casella, 2004). MCMC is an all-purpose strategy that allows one to obtain dependent samples from a generic posterior distribution. There are several theoretical considerations that one needs to consider when implementing MCMC to obtain samples from the posterior distribution including, ergodicity, irreducibility, and positive recurrence of the MCMC. In addition to theoretical considerations, practical implementation issues arise, including, a potential for high computational costs, assessing convergence (Gelman and Rubin, 1992; Cowles and Carlin, 1996), tuning the MCMC (Roberts and Rosenthal, 2009), and computing the effective sample size of the Markov chain (Vats et al., 2019), among other considerations. One of the current state-of-the-art techniques in MCMC is Hamiltonian Monte Carlo (HMC, Neal, 2011). HMC is a Metropolis−Hastings algorithm, where Hamiltonian dynamic evolution is used to propose a new value. In general, HMC leads to "fast mixing" (i.e., converges relatively quickly to the posterior distribution) because it provides a sample from the joint posterior distribution of all processes and parameters, and moreover, has been optimized efficiently using the software `Stan` (Carpenter et al., 2017).

Of course, MCMC is not needed if one can obtain independent replicates directly from the posterior distribution efficiently. In this article, we revisit the problem of generating independent replicates directly from the posterior distribution for a broad class of hierarchical models. This class of hierarchical models could be applied in many of the settings in which one would use a latent Gaussian process model (LGP, e.g., see Gelfand and Schliep, 2016). Much of the current literature does not consider solving this problem, since obtaining independent replicates directly from the exact posterior distribution for Bayesian spatial LGPs is a difficult problem, and MCMC can easily be adapted to many settings. We consider Bayesian spatial hierarchical models for Gaussian distributed data, Poisson distributed data, and binomial distributed data. The samples from our proposed model are independently drawn, and hence avoid issues with convergence, tuning, and positive autocorrelations in a MCMC. Moreover, our exact replicates have an interpretable projection formulation. This regression-type projection can be computed efficiently using known block matrix inversion formulas (Lu and Shiou, 2002). Thus, we refer to our method as Exact Posterior Regression (EPR), which is the one of the contributions of this article.

Conjugate prior distributions are often restricted to the data type. For example, for binomial, negative binomial, Bernoulli, and multinomial distributed data, the fixed and random effects are conjugate with the multivariate logit-beta distribution (Gao and Bradley, 2019; Bradley et al., 2019), which is the special case of the conjugate multivariate (CM) distribution. Similarly, Poisson and Weibull distributed data are conjugate with the multivariate log-gamma distribution (Bradley et al., 2018; Hu and Bradley, 2018; Xu et al., 2023; H.-C.Yang et al., 2019; Parker et al., 2020, 2021), another special case of the CM distribution. Finally, mixed effects models for Gaussian distributed data regularly make use of Gaussian priors for fixed and random effects (Gelman et al., 2013), which is also a type of CM distribution. Thus, our second major contribution is to extend the conjugate multivariate (CM) distribution (Bradley et al., 2020a) to allow for Gaussian priors. Additionally, conjugate prior distributions and the CM distribution do not allow one to explicitly account for hyperparameters without the use of MCMC or approximate Bayesian techniques. Thus, in our extension of the CM to Gaussian prior specifications we marginalize across hyperparameters. We call this new distribution the generalized CM (GCM) distribution, which allows for standard Gaussian priors (e.g., see Gelfand and Schliep, 2016, for a recent discussion). Furthermore, we develop conditional distributions for GCM distributed random vectors.

A key step in our formulation is the incorporation of what we call a "discrepancy term," which is simply an additive term introduced into a mixed effects model similar to that of Bradley et al. (2020b) and Bradley et al. (2023). This term has been interpreted as a way to incorporate signal-to-noise dependence (Bradley et al., 2020b) in Generalized Linear Mixed Effects Models (GLMM), and has also been interpreted as a type of model averaging (Bradley et al., 2023). Classical spatial hierarchical models set these discrepancy terms equal to zero. When these terms are not set equal to zero and instead given a type of improper prior then we show that the implied posterior distribution for fixed effects, random effects, and discrepancy terms will be of the form of a GCM, which we can directly sample from. Draws of the fixed and random effects will then be used for inference (e.g., regression estimation and spatial prediction) bypassing the need for MCMC.

We emphasize the high potential impact of the contributions of EPR and GCM, since much of the literature places a high consistent emphasis on using MCMC strategies to obtain asymptotically exact correlated samples from the posterior distribution. In the context of generalized linear mixed effects models (GLMM) and LGPs, EPR has the potential to circumvent the use of MCMC or approximate Bayesian strategies in several settings where it is commonly used. For example, at the time of writing this manuscript the following papers use MCMC in a spatial LGP setting: Kang et al. (2023), Konomi et al. (2023), Porter et al. (2023), Vranckx et al. (2023), and Zhang et al. (2023a), among others. All of these analyses can easily be adapted to be implemented using EPR, which completely avoids MCMC.

EPR allows one to efficiently analyze several types of correlated spatial data. In particular, we consider modeling three "types of data," namely, conditionally Gaussian, Poisson, and binomial distributed spatial data. Computationally expensive MCMC techniques have become a standard for modeling spatial data (Robert and Casella, 2011; Gelfand and Schliep, 2016). Also, a common approximate Bayesian technique used frequently in the spatial statistics literature is referred to as integrated nested Laplace approximations (INLA, Lindgren et al., 2022). In this article, we compare MCMC and INLA applied to traditional LGPs to EPR.

To summarize, the contributions of this article can be classified into three groups:

1. The first group of contributions of this article develops the GCM distribution. This includes integral expressions for the GCM distribution and the conditional GCM distribution (i.e., the conditional distribution of one sub-vector of a GCM) up to a proportionality constant. The key literature on conjugate modeling began with Diaconis and Ylvisaker (1979)'s semi-

nal paper which formally developed univariate conjugate models for the exponential family. Then Chen and Ibrahim (2003) developed Diaconis and Ylvisaker (1979)'s work in the context of fixed effects models and Bradley et al. (2020a) developed Diaconis and Ylvisaker (1979)'s work in the context of mixed effects models. However, all of these papers require one to match the form of the prior distribution with that of the likelihood. The development of the GCM allows one to consider Gaussian priors. Moreover, this literature often does not emphasize hyperparameters; however, our development explicitly addresses hyperparameters through marginalization. It should be noted that the theoretical development of the GCM is similar to that of the CM distribution (Bradley et al., 2020a). However, the GCM has an enormous practical advantage over the CM by allowing one to use a more standard class of prior distributions (i.e., Gaussian) for spatial data and avoids MCMC updates of hyperparameters. For example, when using the CM for a Poisson data settings, one uses multivariate log-gamma priors for fixed and random effects and updates shape/rate parameters in an MCMC. When taking a GCM approach one can use Gaussian priors for many (but not all) of the fixed and random effects and avoid sampling hyperparameters in an MCMC.

2. The second group of contributions of this article is that we show that one can completely avoid the use of MCMC in settings in which a Bayesian LGP model could be used. In particular, our Bayesian hierarchical model results in a GCM posterior distribution for discrepancy terms, fixed, and random effects, with independent replicates that one can compute without approximations, which we call EPR. Much of the Bayesian literature is shifting focus on avoiding MCMC through the use of approximate Bayesian methods (e.g., see Wainwright and Jordan, 2008; Rue et al., 2009) or through direct sampling of the posterior distributions in special cases for Gaussian data (Zhang et al., 2021; van Erven and Szabó, 2021; Shirota et al., 2023; Zhang et al., 2023b). Recently, Bradley et al. (2023) developed an exact sampler from the posterior distribution for a particular deep Bayesian statistical model for Gaus-

sian and non-Gaussian spatio-temporal data referred to as the deep hierarchical generalized transformation model. EPR adds to this growing literature by allowing one to independently sample from the posterior from a broad class of spatial hierarchical models. By "broad" we mean that similar versions of our proposed model can be written for many existing LGPs. In this article, we consider settings that incorporate spatial basis functions, weakly stationary spatial processes, and conditional autoregressive models.

3. The third group of contributions is the development of EPR for high-dimensional settings. Specifically, we use standard block matrix algebra techniques along with dimension reduction to aid in the computation of EPR (see Theorems 3.3 and 3.4). We demonstrate the size of data that can be efficiently analyzed with EPR through a benchmark high-dimensional dataset consisting of binary cloud mask data from Bradley et al. (2020a), where we analyze spatial Bernoulli observations on the order of 2.4 million observations.

The remainder of the article proceeds as follows. Before we introduce our proposed hierarchical model, we will first provide derivations of the GCM and conditional GCM distribution (i.e., the conditional distribution of one sub-vector of a GCM) in Section 2. We emphasize that GCM random vectors are derived through how they are simulated. Then, in Section 3 we show that our proposed model's posterior distribution for discrepancy terms, fixed, and random effects is GCM, and we describe how to efficiently sample independent replicates directly from the marginal posterior of the fixed effects, and random effects (which we call EPR). Illustrations are provided in Section 4, which includes several simulations/comparisons (15 in total) including common models used in spatial statistics: weakly stationary spatial processes, spatial basis function expansions, and conditional autoregressive models. The main goal of our illustrations is to compare EPR to several traditional Bayesian spatial LGPs. Proofs and additional details, examples, and simulations are given in the Supplementary Material. A discussion is given in Section 5.

2 Preliminary Derivations: The Generalized Conjugate Multivariate Distribution

We now derive the *generalized conjugate multivariate* (GCM) distribution. This development is similar to the development of the CM distribution from Bradley et al. (2020a). The difference between the GCM and CM is that the GCM drops the assumption of identical classes of Diaconis-Ylvisaker (DY) random variable (Diaconis and Ylvisaker, 1979), and marginalizes across a generic d-dimensional real-valued parameter vector $\boldsymbol{\theta}$. We give a review of both the DY and CM distributions, along with a notation table, in Supplementary Appendix A to provide the reader additional preliminary information. The GCM is needed for our main contribution of EPR in Section 4.

The GCM is defined by the transformation,

$$\mathbf{y}_M = \boldsymbol{\mu}_M + \mathbf{V}_M \mathbf{D}(\boldsymbol{\theta}) \mathbf{w}_M, \tag{1}$$

where the $n \equiv \sum_{k=1}^{K} n_k$-dimensional random vector $\mathbf{w}_M = (\mathbf{w}_1', \ldots, \mathbf{w}_K')'$, the n_k-dimensional random vector $\mathbf{w}_k = (w_{k,1}, \ldots, w_{k,n_k})'$ with (k,i)-th element $w_{k,i} \sim \mathrm{DY}(\alpha_{k,i}, \kappa_{k,i}; \psi_k)$, "DY" is a short hand for the DY distribution, the subscript "M" stands for "Multi-type" (as there are multiple types of DY random variables indexed by k), the $n \times n$ real-valued matrix $\mathbf{V}_M$ is an invertible covariance parameter matrix, $\kappa_{k,i} > 0$, $\alpha_{k,i}/\kappa_{k,i} \in \mathscr{Y}_k$ defines the support for $\alpha_{k,i}$, $\mathscr{Y}_k$ is the support of $w_{k,i}$, and $\boldsymbol{\mu}_M$ is an unknown n-dimensional real-valued location parameter vector. Let $\boldsymbol{\alpha}_k = (\alpha_{k,1}, \ldots, \alpha_{k,n_k})'$, $\boldsymbol{\kappa}_k = (\kappa_{k,1}, \ldots, \kappa_{k,n_k})'$, $\boldsymbol{\alpha}_M = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_K')'$, and $\boldsymbol{\kappa}_M = (\boldsymbol{\kappa}_1', \ldots, \boldsymbol{\kappa}_K')'$. The function ψ_k is referred to as the unit log partition function, and we consider $\psi_1(w) = w^2$, $\psi_2(w) = \exp(w)$, and $\psi_3(w) = \log\{1 + \exp(w)\}$ for real-valued w. It is known that $w_{1,i}$ is normally distributed with mean $\frac{\alpha_{1,i}}{2\kappa_{1,i}}$ and variance $\frac{1}{2\kappa_{1,i}}$, $w_{2,i}$ is the log of a gamma random variable with shape $\alpha_{2,i}$ and rate $\kappa_{2,i}$, and $w_{3,i}$ is the logit of a beta random variable with shape parameters $\alpha_{3,i}$ and rate $\kappa_{3,i} - \alpha_{3,i}$ (Bradley et al., 2020a).

Let $\mathbf{D} : \Omega \rightarrow \mathbb{R}^n \times \mathbb{R}^n$ be a known $n \times n$ matrix valued function, such that $\mathbf{D}(\boldsymbol{\theta})^{-1}$ exists for every d-dimensional $\boldsymbol{\theta} \in \Omega$ for a generic real-valued set Ω. Let $\boldsymbol{\theta}$ be distributed according to the proper density $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is independent of $\boldsymbol{\mu}_M$, $\boldsymbol{\alpha}_M$, $\boldsymbol{\kappa}_M$, and $\mathbf{V}_M$. Sampling from the marginal distribution $\mathbf{y}_M|\boldsymbol{\mu}_M, \mathbf{V}_M, \boldsymbol{\alpha}_M, \boldsymbol{\kappa}_M$ (marginalizing across $\boldsymbol{\theta}$) is straightforward; namely, first sample $\boldsymbol{\theta}$ from $\pi(\boldsymbol{\theta})$ and then compute the transformation in (1) to produce a sample from $f(\mathbf{y}_M|\boldsymbol{\mu}_M, \mathbf{V}_M, \boldsymbol{\alpha}_M, \boldsymbol{\kappa}_M)$. The probability density function (pdf) for $\mathbf{y}_M|\boldsymbol{\mu}_M, \mathbf{V}_M, \boldsymbol{\alpha}_M, \boldsymbol{\kappa}_M$ is stated in Theorem 2.1.

Theorem 2.1. *Let $\mathbf{y}_M$ be defined as in (1). Then the pdf for $\mathbf{y}_M$ is given by,*

$$\begin{aligned} &f(\boldsymbol{y}_M|\boldsymbol{\mu}_M, \boldsymbol{V}_M, \boldsymbol{\alpha}_M, \boldsymbol{\kappa}_M) \\ &= \int_{\Omega} \pi(\boldsymbol{\theta}) \mathscr{N}_M \exp\left[\boldsymbol{\alpha}_M' \boldsymbol{D}(\boldsymbol{\theta})^{-1} \boldsymbol{V}_M^{-1}(\boldsymbol{y}_M - \boldsymbol{\mu}_M) - \boldsymbol{\kappa}_M' \boldsymbol{\psi}_M \left\{\boldsymbol{D}(\boldsymbol{\theta})^{-1} \boldsymbol{V}_M^{-1}(\boldsymbol{y}_M - \boldsymbol{\mu}_M)\right\}\right] d\boldsymbol{\theta}, \qquad (2) \end{aligned}$$

where $\mathscr{N}_M = \frac{\left\{\prod_{k=1}^{K} \prod_{i=1}^{n_k} \mathscr{N}_k(\kappa_{k,i}, \alpha_{k,i})\right\}}{\det\{\boldsymbol{D}(\boldsymbol{\theta})\}\det(\boldsymbol{V}_M)}$, $\boldsymbol{y}_M \in \mathscr{S}$, $\mathscr{S} = \{\boldsymbol{y}_M : \boldsymbol{y}_M = \boldsymbol{\mu}_M + \boldsymbol{V}_M \boldsymbol{D}(\boldsymbol{\theta})\boldsymbol{c}, \boldsymbol{c} = \{c_{k,i}\}, c_{k,i} \in \mathscr{Y}_k, \boldsymbol{\theta} \in \Omega, i = 1, \ldots, n_k, k = 1, \ldots, K\}$, $\alpha_{k,i}/\kappa_{k,i} \in \mathscr{Z}_k$, $\kappa_{k,i} > 0$, $\boldsymbol{\psi}_M\{\boldsymbol{V}_M(\boldsymbol{y}_M - \boldsymbol{\mu}_M)\} = \left(\boldsymbol{\psi}_1\{\boldsymbol{J}_1 \boldsymbol{V}_M(\boldsymbol{y}_M - \boldsymbol{\mu}_M)\}', \ldots, \boldsymbol{\psi}_K\{\boldsymbol{J}_K \boldsymbol{V}_M(\boldsymbol{y}_M - \boldsymbol{\mu}_M)\}')\right)'$, the $n_k \times n$ matrix $\boldsymbol{J}_k = \left(\mathbf{0}_{n_k, \sum_{j=1}^{k-1} n_j}, \boldsymbol{I}_{n_k}, \mathbf{0}_{n_k, \sum_{j=k+1}^{K} n_j}\right)$ for $1 < k < K$, $\boldsymbol{J}_1 = \left(\boldsymbol{I}_{n_1}, \mathbf{0}_{n_1, \sum_{j=2}^{K} n_j}\right)$, $\boldsymbol{J}_K = \left(\mathbf{0}_{n_k, \sum_{j=1}^{K-1} n_j}, \boldsymbol{I}_{n_k}\right)$, $\mathbf{0}_{n,m}$ is an $n \times m$ matrix of zeros, $\boldsymbol{I}_{n_k}$ is an $n_k \times n_k$ identity matrix, the n-dimensional vector $\boldsymbol{\alpha}_M = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_K')'$, and the n-dimensional vector $\boldsymbol{\kappa}_M = (\boldsymbol{\kappa}_1', \ldots, \boldsymbol{\kappa}_K')'$.

Proof: See Supplementary Appendix B.

Note that we use the notation of a bold $\boldsymbol{\psi}_j(\mathbf{h})$ to represent a vector with i-th component (not bolded) $\psi_j(h_i)$ for $\mathbf{h} = (h_1, \ldots, h_n)'$. We use the shorthand GCM$(\boldsymbol{\alpha}_M, \boldsymbol{\kappa}_M, \boldsymbol{\mu}_M, \mathbf{V}_M, \pi, \mathbf{D}; \boldsymbol{\psi}_M)$ for the density in (2).

Sampling directly from a GCM distribution requires two items:

1. One must be able to sample the random vector $\boldsymbol{\theta}$ directly from its prior distribution π.

2. One must be able to the sample independent DY random variables contained in the vector $\mathbf{w}_M$.

In this article, the parameter vector $\boldsymbol{\theta}$ typically consists of variance parameters and spatial range parameters. These parameters will be given independent inverse gamma prior or uniform prior distributions, which one can sample from directly. Additionally, the class of hierarchical models in Section 3 lead to DY random variables that are either independent univariate normal, beta, or gamma random variables, which are straightforward to simulate from directly using standard software. The fact that we can sample independent replicates of a GCM random vector directly is crucial in Section 3, where we show that a certain class of hierarchical models leads to a posterior distribution for discrepancy terms, fixed, and random effects that is GCM (i.e., is of the form in Theorem 2.1), and hence, one can directly sample from it.

A related distribution to the GCM is what we call the conditional GCM. By "conditional GCM," we mean the conditional distribution of $\mathbf{y}^{(1)}$ given $\mathbf{y}^{(2)}$ when $(\mathbf{y}^{(1)},\mathbf{y}^{(2)})$ are jointly GCM. We provide the integral expression for the conditional GCM in Theorem 2.2 up to a proportionality constant.

Theorem 2.2. *Let $\boldsymbol{y}_M = (\boldsymbol{y}^{(1)\prime},\boldsymbol{y}^{(2)\prime})' \sim \mathrm{GCM}(\boldsymbol{\alpha}_M,\boldsymbol{\kappa}_M,\boldsymbol{\mu}_M,\boldsymbol{V}_M,\pi,\boldsymbol{D};\boldsymbol{\psi}_M)$, where $\boldsymbol{y}^{(1)}$ is r-dimensional and $\boldsymbol{y}^{(2)}$ is $(n-r)$-dimensional. Also, let $\boldsymbol{V}_M^{-1} = (\boldsymbol{H},\boldsymbol{Q})$, where $\boldsymbol{H}$ is a $n \times r$ and $\boldsymbol{Q}$ is $n \times (n-r)$. Then, it follows*

$$f(\boldsymbol{y}^{(1)}|\boldsymbol{y}^{(2)},\boldsymbol{\mu}_M,\boldsymbol{V}_M,\boldsymbol{\alpha}_M,\boldsymbol{\kappa}_M)$$
$$\propto \int_\Omega \frac{\pi(\boldsymbol{\theta})}{\det\{\boldsymbol{D}(\boldsymbol{\theta})\}}\exp\left[\boldsymbol{\alpha}_M'\boldsymbol{D}(\boldsymbol{\theta})^{-1}\boldsymbol{H}\boldsymbol{y}^{(1)} - \boldsymbol{\alpha}_M'\boldsymbol{\mu}_M^* - \boldsymbol{\kappa}_M'\boldsymbol{\psi}_M\left\{\boldsymbol{D}(\boldsymbol{\theta})^{-1}\boldsymbol{H}\boldsymbol{y}^{(1)} - \boldsymbol{\mu}_M^*\right\}\right]d\boldsymbol{\theta},$$

where $\boldsymbol{\mu}_M^* = \boldsymbol{D}(\boldsymbol{\theta})^{-1}\boldsymbol{V}_M^{-1}\boldsymbol{\mu}_M - \boldsymbol{D}(\boldsymbol{\theta})^{-1}\boldsymbol{Q}\boldsymbol{y}^{(2)}$.

Proof: See Supplementary Appendix B.

We use the shorthand cGCM$(\boldsymbol{\alpha}_M, \boldsymbol{\kappa}_M, \boldsymbol{\mu}_M^*, \mathbf{H}, \pi, \mathbf{D}; \boldsymbol{\psi}_M)$ for the conditional GCM in Theorem 2.2. It is not known how to simulate directly from a cGCM.

3 Methodology

In this section, we outline how to sample from the posterior distribution for discrepancy terms, fixed, and random effects from a general class of spatial hierarchical models. We define EPR in Section 3.1 for areal spatial data, discuss hyperprior specifications in Section 3.2, define the extension to spatial process models in Section 3.3, and discuss computational issues and implementation in Sections 3.4 and 3.5. The statement of our model in Section 3.1 is given for data distributed according to a generic member of the exponential family. To aid practitioners interested in using this method, we provide more accessible model statements for specific cases (i.e., Gaussian, binomial, and Poisson data) in Supplementary Appendix C.

3.1 Exact Posterior Regression for Regional Data

Suppose we observe data from the exponential family, let the total number of observations be denoted with n, and denote the n-dimensional data vector with $\mathbf{z} = (Z_1, \ldots, Z_n)'$. Let Z_i represent the data at region i (e.g., counties, census tracts, etc.). Then assume Z_i belongs to a member of the exponential family of distributions. In particular, we assume one of the following:

$$Z_i | Y_i, b_{i,k} \sim \mathrm{EF}(Y_i, b_{i,k}, \psi_k); \; i = 1, \ldots, n, \; k = 1, 2, 3, \tag{3}$$

where "EF" is a shorthand for the natural exponential family (see Supplementary Appendix A for more details), and $b_{i,k}\psi_k(Y_i)$ is the log-partition function. For example, when $b_{i,1} = \frac{1}{2\sigma_i^2}$ with $\sigma_i^2 >$

0 and $\psi_1(Y_i) = Y_i^2$ we have that $Z_i|Y_i, b_{i,1}$ is normally distributed with mean Y_i and variance σ_i^2. When $b_{i,2} \equiv 1$ and $\psi_2(Y_i) = \exp(Y_i)$ we have that $Z_i|Y_i, b_{i,2}$ is Poisson distributed with mean $\exp(Y_i)$. Similarly, when $b_{i,3} = m_i$ with integer $m_i \geq 1$ and $\psi_3(Y_i) = \log\{1+\exp(Y_i)\}$ we have that $Z_i|Y_i, b_{i,3}$ is binomial distributed with sample size m_i and probability of success $\exp(Y_i)/\{1+\exp(Y_i)\}$. Let the n-dimensional vector $\mathbf{b}_k \equiv (b_{1,k}, \ldots, b_{n,k})'$. In this article, we consider these three cases (i.e., normal, Poisson and binomial distributed cases), and note that binomial distributed data allows for Bernoulli distributed data as a special case (i.e., $m_i = 1$), and multinomial distributed data when using a stick-breaking representation of the multinomial (e.g., see Bradley et al., 2019, for stick-breaking in the context of CM prior distributions). Organize the latent random variable Y_i into the n-dimensional vector $\mathbf{y} = (Y_1, \ldots, Y_n)'$.

Consider the following linear model assumption for $\mathbf{y}$ (McCullagh and Nelder, 1989):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\eta} + (\boldsymbol{\xi} - \boldsymbol{\delta}_y), \tag{4}$$

where $\mathbf{X}$ is a $n \times p$ matrix of known covariates, and $\boldsymbol{\beta}$ is an unknown p-dimensional vector of regression coefficients. Let $\boldsymbol{\beta}$ have a Gaussian prior with p-dimensional location vector $\mathbf{D}_\beta(\boldsymbol{\theta})\boldsymbol{\delta}_\beta$, and $p \times p$ covariance matrix $\mathbf{D}_\beta(\boldsymbol{\theta})\mathbf{D}_\beta(\boldsymbol{\theta})'$, where $\mathbf{D}_\beta(\boldsymbol{\theta}) : \Omega \to \mathbb{R}^p \times \mathbb{R}^p$. Let $\mathbf{G}$ be a $n \times r$ matrix of coefficients for the r-dimensional random effects $\boldsymbol{\eta}$. Several choices for $\mathbf{G}$ are available including a known pre-specified matrix of basis functions (e.g., splines (Wahba, 1990), wavelets (Novikov et al., 2005), Moran's I basis functions (Hughes and Haran, 2013), etc.), or a matrix square root of a known spatial covariance matrix. We assume $\boldsymbol{\eta}$ is Gaussian with r-dimensional location vector $\mathbf{D}_\eta(\boldsymbol{\theta})\boldsymbol{\delta}_\eta$ and $r \times r$ covariance matrix $\mathbf{D}_\eta(\boldsymbol{\theta})\mathbf{D}_\eta(\boldsymbol{\theta})'$, where $\mathbf{D}_\eta(\boldsymbol{\theta}) : \Omega \to \mathbb{R}^r \times \mathbb{R}^r$. Let $\boldsymbol{\theta}$ be a generic d-dimensional parameter vector with prior distribution $\pi(\boldsymbol{\theta})$. The fourth term $\boldsymbol{\delta}_y$ has recently been introduced to the spatial mixed effects model literature (Bradley et al., 2020b, 2023), and models the error introduced by allowing the mixed effects representation $\widehat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\eta} + \boldsymbol{\xi}$ to be different from the natural parameter $\mathbf{y}$. That is, $\boldsymbol{\delta}_y$ is the error term caused by incorporating non-zero $\widehat{\mathbf{y}} - \mathbf{y} (= \boldsymbol{\delta}_y)$.

Traditionally, the fine-scale variability term $\boldsymbol{\xi}$ is assumed to be Gaussian (Cressie and Wikle, 2011). In our framework, we will be able to draw independent replicates from the exact posterior distribution for discrepancy terms, fixed, and random effects if we specify $\boldsymbol{\xi}$ to be a cGCM that is "close" to a Gaussian distribution. Specifically, let the distribution for $\boldsymbol{\xi}$ be proportional to a cGCM$(\boldsymbol{\alpha}_\xi, \boldsymbol{\kappa}_\xi, \boldsymbol{\delta}_\xi^*, \mathbf{H}_\xi, \pi_\xi, \mathbf{D}_\xi; \boldsymbol{\psi}_\xi)$, where the $2n$-dimensional discrepancy parameter $\boldsymbol{\delta}_\xi^* = (\boldsymbol{\delta}_y' - \boldsymbol{\beta}'\mathbf{X}' - \boldsymbol{\eta}'\mathbf{G}', \boldsymbol{\delta}_\xi')'$, $\boldsymbol{\delta}_y$ and $\boldsymbol{\delta}_\xi$ are n-dimensional real-vectors, and $2n \times n$ matrix-valued precision parameter $\mathbf{H}_\xi = (\sigma_\xi \mathbf{I}_n, \mathbf{I}_n)'$. The $2n$-dimensional shape parameter $\boldsymbol{\alpha}_\xi = \mathbf{0}_{2n,1}$ when the data is assumed Gaussian, and $\boldsymbol{\alpha}_\xi = (\alpha_\xi \mathbf{1}_{1,n}, \mathbf{0}_{1,n})'$ when the data is assumed to be distributed according to the Poisson or binomial distributions, where $\alpha_\xi > 0$ and $\mathbf{1}_{r,n}$ is a $r \times n$ matrix of ones. The $2n$-dimensional shape parameter $\boldsymbol{\kappa}_\xi = (\mathbf{0}_{1,n}, \frac{1}{2}\mathbf{1}_{1,n})'$ when the data is assumed to be either Gaussian or Poisson distributed, and $\boldsymbol{\kappa}_\xi = (2\alpha_\xi \mathbf{1}_{1,n}, \frac{1}{2}\mathbf{1}_{1,n})'$ when the data is assumed to be distributed according to the binomial distribution. Let $\mathbf{D}_\xi \equiv \sigma_\xi \mathbf{I}_{2n}$ with $\sigma_\xi^2 > 0$ and $\pi_\xi(\theta) = I(\theta = \sigma_\xi^2)$ with $I(\cdot)$ defined to be the indicator function. The unit-log partition function $\boldsymbol{\psi}_\xi$ is,

$$\boldsymbol{\psi}_\xi(\mathbf{h}) = (\psi_k(h_1), \ldots, \psi_k(h_n), \psi_1(h_1^*), \ldots, \psi_1(h_n^*))',$$

for any $\mathbf{h} = (h_1, \ldots, h_n, h_1^*, \ldots, h_n^*)' \in \mathbb{R}^{2n}$. It is straightforward to verify that when $\alpha_\xi = 0$ we have that cGCM$(\boldsymbol{\alpha}_\xi, \boldsymbol{\kappa}_\xi, \boldsymbol{\delta}_\xi^*, \mathbf{H}_\xi, \pi_\xi, \mathbf{D}_\xi; \boldsymbol{\psi}_\xi)$ is proportional to a Gaussian distribution with mean $\boldsymbol{\delta}_\xi$ and covariance $\sigma_\xi^2 \mathbf{I}_n$ with $\sigma_\xi \in \boldsymbol{\theta}$. This choice of cGCM with $\alpha_\xi > 0$ will ensure that the implied posterior distribution for discrepancy terms, fixed, and random effects has parameters that do not lie on the boundary of the parameter space. In the case of a Gaussian data model, this cGCM distribution is exactly a Gaussian distribution, where $\boldsymbol{\xi}$ follows a normal distribution with mean $\boldsymbol{\delta}_\xi$ and covariance $\sigma_\xi^2 \mathbf{I}_n$. For Poisson and binomial data, $\boldsymbol{\xi} | \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\delta}_y, \boldsymbol{\delta}_\xi$ is proportional to a conditional GCM distribution that is "close" to a Gaussian distribution as α_ξ approaches zero. By "proportional to a conditional GCM" we mean that we do not include the normalizing constant for

$f(\boldsymbol{\xi}|\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{\delta}_y,\boldsymbol{\delta}_\xi)$ in our expression of the hierarchical model for binomial and Poisson cases, since this normalizing constant is unknown. More technical details on the fine-scale model specification and how this class of cGCM is "close" to a Gaussian is provided in Supplementary Appendix D.

The terms $\mathbf{X}\boldsymbol{\beta}$, $\mathbf{G}\boldsymbol{\eta}$, and $\boldsymbol{\xi}$ are covered in standard textbooks in spatio-temporal statistics (Cressie and Wikle, 2011), and are referred to as large-scale variability, small-scale variability, and fine-scale variability, respectively. In more recent literature a fourth term has been considered (Bradley et al., 2020b, 2023); that is, the $(2n+p+r)$-dimensional vector $\boldsymbol{\delta} = (\boldsymbol{\delta}_y', \boldsymbol{\delta}_\beta', \boldsymbol{\delta}_\eta', \boldsymbol{\delta}_\xi')'$ discrepancy parameter. Previous uses of a discrepancy term have led to MCMC algorithms to sample from $f(\boldsymbol{\zeta}|\mathbf{z})$ that were computationally more efficient, in terms of Central Processing Unit (CPU) times, than MCMC algorithms to sample from $f(\boldsymbol{\zeta}|\mathbf{z},\mathbf{q}=\mathbf{0}_{n,1})$ derived from traditional LGPs (e.g., see Bradley et al., 2020b; Bradley, 2022). In our case, a *particular form* of $\boldsymbol{\delta}$ leads the fixed and random effects to be distributed according to a GCM, which from Section 2, we know how to sample from directly without approximations and without MCMC. Specifically, let $\boldsymbol{\delta} = -\text{blkdiag}(\mathbf{I}_n, \mathbf{D}_\beta(\boldsymbol{\theta})^{-1}, \mathbf{D}_\eta(\boldsymbol{\theta})^{-1}, \frac{1}{\sigma_\xi}\mathbf{I}_n)\mathbf{Q}\mathbf{q}$, where "blkdiag" is the block diagonal operator and $\mathbf{Q}$ are an $(2n+p+r)\times n$ eigenvectors of the orthogonal complement of the $(2n+p+r)\times(n+p+r)$ matrix,

$$\mathbf{H} = \begin{pmatrix} \mathbf{I}_n & \mathbf{X} & \mathbf{G} \\ \mathbf{0}_{p,n} & \mathbf{I}_p & \mathbf{0}_{p,r} \\ \mathbf{0}_{r,n} & \mathbf{0}_{r,p} & \mathbf{I}_r \\ \mathbf{I}_n & \mathbf{0}_{n,p} & \mathbf{0}_{n,r} \end{pmatrix}, \tag{5}$$

so that $\mathbf{Q}\mathbf{Q}' = \mathbf{I}_{2n+p+r} - \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'$ and $\mathbf{H}'\mathbf{Q} = \mathbf{0}_{n+p+r,n}$, where recall that idempotent matrices have eigenvalues equal to zero or one. The free parameter $\mathbf{q}$ is now referred to as the "discrepancy term," which is assumed unknown. Several LGPs in the literature set $\mathbf{q} = \mathbf{0}_{n,1}$. However, if one instead assumes an improper prior on $\mathbf{q}$ then the posterior distribution of $\boldsymbol{\zeta} = (\boldsymbol{\xi}', \boldsymbol{\beta}', \boldsymbol{\eta}')'$ and $\mathbf{q}$ is

GCM, as seen in Theorem 3.1 below.

Theorem 3.1. *Suppose $Z_i|Y_i, b_{i,k}$ are independently distributed according to (3). Assume the model for $\boldsymbol{y}$ in (4), the improper prior $f(\boldsymbol{q}) = 1$, and let the hyperparameters $\boldsymbol{\theta}$ have a proper prior distribution $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \cap \{\sigma_i^2\}^c)\prod_{i=1}\pi(\sigma_i^2)$ where "c" denotes the set complement. Let $\pi(\boldsymbol{\theta})$ be any proper distribution that can be simulated from directly. Then*

$$(\boldsymbol{\zeta}', \boldsymbol{q}')'|z \sim \text{GCM}(\boldsymbol{\alpha}_M, \boldsymbol{\kappa}_M, \mathbf{0}_{2n+p+r,1}, \boldsymbol{V}_M, \pi, \boldsymbol{D}; \boldsymbol{\psi}_M),$$

where $\boldsymbol{D}(\boldsymbol{\theta})^{-1} = \text{blkdiag}(\boldsymbol{I}_n, \boldsymbol{D}_\beta(\boldsymbol{\theta})^{-1}, \boldsymbol{D}_\eta(\boldsymbol{\theta})^{-1}, \frac{1}{\sigma_\xi}\boldsymbol{I}_n)$, $\boldsymbol{V}_M^{-1} = (\boldsymbol{H}, \boldsymbol{Q})$ is defined by (5), $\boldsymbol{D}_\sigma = \text{diag}\left(\frac{1}{\sigma_i^2} : i = 1, \ldots, n\right)$, the $(2n+p+r)$-dimensional unit-log partition function $\boldsymbol{\psi}_M(\boldsymbol{h}) = \left(\psi_k(h_1), \ldots, \psi_k(h_n), \psi_1(h_1^*), \ldots, \psi_1(h_{n+p+r}^*)\right)'$ for $(2n+p+r)$-dimensional real-valued vector $\boldsymbol{h} = (h_1, \ldots, h_n, h_1^*, \ldots, h_{n+p+r}^*)'$, and the $(2n+p+r)$-dimensional location and shape/scale parameter vectors are defined as follows: $\boldsymbol{\alpha}_M = (z'\boldsymbol{D}_\sigma', \mathbf{0}_{1,n+p+r})'$ and $\boldsymbol{\kappa}_M = (\frac{1}{2}\mathbf{1}_{1,n}\boldsymbol{D}_\sigma', \frac{1}{2}\mathbf{1}_{1,n+p+r})'$ when the data is normally distributed; $\boldsymbol{\alpha}_M = (z' + \alpha_\xi\mathbf{1}_{1,n}, \mathbf{0}_{1,n+p+r})'$ and $\boldsymbol{\kappa}_M = (\mathbf{1}_{1,n}, \frac{1}{2}\mathbf{1}_{1,n+p+r})'$ when the data is Poisson distributed; and $\boldsymbol{\alpha}_M = (z' + \alpha_\xi\mathbf{1}_{1,n}, \mathbf{0}_{1,n+p+r})'$ and $\boldsymbol{\kappa}_M = (\boldsymbol{m}' + 2\alpha_\xi\mathbf{1}_{1,n}, \frac{1}{2}\mathbf{1}_{1,n+p+r})'$ when the data is binomial distributed.

Proof: See Supplementary Appendix B.

In Supplementary Appendix B, we also show that the posterior distribution for $\boldsymbol{\zeta}$ and $\mathbf{q}$ in Theorem 3.1 is proper. This is an important consideration, since we specify an improper prior for the discrepancy parameter $\mathbf{q}$.

In Theorem 3.1 the presence of $\alpha_\xi > 0$, arising from our cGCM specification for $\boldsymbol{\xi}$, leads to strictly positive elements in the vectors $\boldsymbol{\alpha}_M$ and $\boldsymbol{\kappa}_M$ when the data vector $\mathbf{z}$ contains zero elements, which can occur in the Poisson or binomial data settings. Hence, the presence of a cGCM (chosen to be close to a Gaussian) fine-scale term allows one to avoid the boundaries of the parameter space, leading to a well-defined GCM that one can sample from directly. Theorem 3.1 allows one to obtain replicates directly from the posterior distribution $f(\boldsymbol{\zeta},\mathbf{q}|\mathbf{z})$ using a familiar projection expression, as seen below in Theorem 3.2.

Theorem 3.2. *Denote a replicate of $\boldsymbol{\zeta}$, $\boldsymbol{q}$, and $\boldsymbol{y}$ using $f(\boldsymbol{\zeta},\boldsymbol{q}|z)$ from Theorem (3.1) with $\boldsymbol{\zeta}_{rep}$, $\boldsymbol{q}_{rep}$, and $\boldsymbol{y}_{rep}$. Then*

$$\boldsymbol{\zeta}_{rep} = (\boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{w} \tag{6}$$

$$\boldsymbol{q}_{rep} = \boldsymbol{Q}'\boldsymbol{w}, \tag{7}$$

$$\boldsymbol{y}_{rep} = (\boldsymbol{I}_n, \mathbf{0}_{n,n+p+r})\boldsymbol{H}\boldsymbol{\zeta}_{rep} + (\boldsymbol{I}_n, \mathbf{0}_{n,n+p+r})\boldsymbol{Q}\boldsymbol{q}_{rep} = (\boldsymbol{I}_n, \mathbf{0}_{n,n+p+r})\boldsymbol{w} \tag{8}$$

where the $(2n+p+r)$-dimensional random vector $\boldsymbol{w}$ is GCM$(\boldsymbol{\alpha}_M, \boldsymbol{\kappa}_M, \mathbf{0}_{2n+p+r,1}, \boldsymbol{I}_{2n+p+r}, \pi, \boldsymbol{D}; \boldsymbol{\psi}_M)$, where $\boldsymbol{\alpha}_M$, $\boldsymbol{\kappa}_M$, π, $\boldsymbol{D}$, and $\boldsymbol{\psi}_M$ are the same as defined in Theorem 3.1.

Proof: See Supplementary Appendix B.

In Theorem 3.2, the vector $(2n+p+r)$-dimensional vector $\mathbf{w} \equiv \mathbf{D}(\boldsymbol{\theta})\mathbf{w}_M = (\mathbf{y}'_{rep}, \mathbf{w}'_{\beta}, \mathbf{w}'_{\eta}, \mathbf{w}'_{\xi})'$, where $\mathbf{y}_{rep}$ is easy to generate since it consists of independent DY random variables, $\mathbf{w}_{\beta} \sim$ GCM$(\mathbf{0}_{p,1}, \frac{1}{2}\mathbf{1}_{p,1}, \mathbf{0}_{p,1}, \mathbf{I}_p, \pi, \mathbf{D}_{\beta}(\boldsymbol{\theta}); \boldsymbol{\psi}_1)$, $\mathbf{w}_{\eta} \sim$ GCM$(\mathbf{0}_{r,1}, \frac{1}{2}\mathbf{1}_{r,1}, \mathbf{0}_{r,1}, \mathbf{I}_r, \pi, \mathbf{D}_{\eta}(\boldsymbol{\theta}); \boldsymbol{\psi}_1)$, and $\mathbf{w}_{\xi}$ is n-dimensional consisting of independent Gaussian random variables with mean zero and variance σ^2_{ξ}. Thus, it is straightforward to compute $\mathbf{w}$ when it is straightforward to sample from the marginal prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ when setting the discrepancy parameter equal to zero. To do this one can, for example, sample from the joint distribution of $\mathbf{w}_{\beta}$ and $\boldsymbol{\theta}$, where first one samples $\boldsymbol{\theta}$ from π then samples $\mathbf{w}_{\beta}$ from a Gaussian distribution with mean zero and covariance matrix $\mathbf{D}_{\beta}(\boldsymbol{\theta})\mathbf{D}_{\beta}(\boldsymbol{\theta})'$. The projection $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}$ can be computed on the order of $n+p^3+r^3$ operations with storage on the order of $n(p+r)+p^2+r^2$, when $\mathbf{G}$ is dense. When $\mathbf{G}$ is identity with $r=n$, $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}$ can be computed on the order of p^3 operations with storage on the order of $np+p^2$. For the details on computing $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}$ see Section 3.4.

The solution in the Gaussian special case is very similar to that in Murphy (2007) and Zhang et al. (2023b). Namely, a different regression arises in (6) from Murphy (2007) and Zhang et al. (2023b) due to our incorporation of a fine-scale variability term. Recall that the presence of fine-scale terms is particularly important for non-Gaussian data, since shape parameters and rate parameters in $\boldsymbol{\alpha}_M$ and $\boldsymbol{\kappa}_M$ in Theorem 3.1 are non-zero when count-valued observations are zero (i.e., the first stack components of $\boldsymbol{\alpha}_M$ and $\boldsymbol{\kappa}_M$) leading to a proper GCM. Thus, one exciting feature of Theorem 3.2 is that we obtain Gaussian like simulations of replicates from the marginal posterior distribution $f(\boldsymbol{\zeta}|\mathbf{z})$ for non-Gaussian data. Equation (6) can also be seen as a parsimonious special case of the sampler in Bradley et al. (2023) with considerably fewer parameters.

The class of distributions that our model's posterior belongs to is GCM, and in Supplementary Appendix E we show that the traditional LGP's posterior distribution (setting $\alpha_\xi = 0$ and $\mathbf{q} = \mathbf{0}_{n,1}$) is a conditional GCM. In Section 4, we empirically investigate the consequences for using a GCM posterior distribution for $\boldsymbol{\zeta}$ and $\mathbf{q}$ instead of the more traditional cGCM posterior distribution for $\boldsymbol{\zeta}$ by generating the data from a traditional LGP model and comparing several metrics.

Theorem 3.2 and Supplementary Appendix E provides the motivation for including the discrepancy parameter $\mathbf{q}$. Namely, this discrepancy parameter leads to easy-to-compute direct simulations from the posterior distribution for $\boldsymbol{\zeta}$ and $\mathbf{q}$, whereas, the cGCM posterior distribution for $\boldsymbol{\zeta}$ can not be sampled from directly. However, the incorporation of $\mathbf{q}$ leads to a model that is clearly overparameterized. Thus, a simple solution is to perform inference on $\boldsymbol{\zeta}$ using exact replicates from (6), which generates values from the marginal distribution $f(\boldsymbol{\zeta}|\mathbf{z})$. Then use the estimator of $\mathbf{q} = \mathbf{0}_{n,1}$. This is the general strategy used in the CM literature (Bradley et al., 2020a) implemented using a type of block Gibbs sampler. Let $\widehat{\mathbf{y}}$ represent the profile of $\mathbf{y}$ using the plug-in estimator $\mathbf{q} = \mathbf{0}_{n,1}$, so that $\widehat{\mathbf{y}}_{rep} = (\mathbf{I}_n, \mathbf{0}_{n,n+p+r})\mathbf{H}\boldsymbol{\zeta}_{rep} = \mathbf{X}\boldsymbol{\beta}_{rep} + \mathbf{G}\boldsymbol{\eta}_{rep} + \boldsymbol{\xi}_{rep}$, where $\boldsymbol{\zeta}_{rep} = (\boldsymbol{\xi}'_{rep}, \boldsymbol{\beta}'_{rep}, \boldsymbol{\eta}'_{rep})'$. Moreover, one might similarly use $\widetilde{\mathbf{y}}_{rep} = \mathbf{X}\boldsymbol{\beta}_{rep} + \mathbf{G}\boldsymbol{\eta}_{rep}$ for inference on $\mathbf{y}$, which would implicitly estimate both $\mathbf{q}$ and $\boldsymbol{\xi}$ to be zero after marginalizing them from the posterior distribution for $\boldsymbol{\zeta}$ and $\mathbf{q}$.

The random vector $\mathbf{y}_{rep}$ has a very important interpretation. If one assumes $Z_i|Y_i$ is distributed according to the natural exponential family in (3), and Y_i is independently distributed according to the DY distribution then we have that the implied posterior distribution for $\{Y_i\}$ is equal in distribution to $\mathbf{y}_{rep}$ in Theorem 3.2. Thus, $\mathbf{y}_{rep}$ represents a replicate from the posterior distribution from a saturated model. Recall in the goodness-of-fit literature that saturated models define a separate parameter for each datum and is meant to overfit the data, and then, measures of deviance from the saturated model are used to select more parsimonious models (e.g., see Bradley, 2022, for a recent paper). This provides additional motivation for using the marginal distribution $f(\boldsymbol{\zeta}|\mathbf{z})$ and $\widehat{\mathbf{y}}_{rep}$ (or $\widetilde{\mathbf{y}}_{rep}$) to perform inference on $\mathbf{y}$, which implies the use of the estimator of $\mathbf{q} = \mathbf{0}_{n,1}$ (and $\boldsymbol{\xi} = \mathbf{0}_{n,1}$). In the recent literature $\mathbf{y}_{rep} - \widehat{\mathbf{y}}_{rep}$ $(= -\boldsymbol{\delta})$ is referred to as "discrepancy error," and hence we refer to $\boldsymbol{\delta}$ as a discrepancy term (Bradley et al., 2020b, 2023).

It is important to clarify how our model could be considered an LGP. By "latent" we mean that a process is not directly observed, but rather the observation Z is observed. In our new model there are three latent vectorized processes, namely $\widetilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\eta}$, $\widehat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\eta} + \boldsymbol{\xi}$, and $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\eta} + \boldsymbol{\xi} - \boldsymbol{\delta}_y$. We assume $\boldsymbol{\beta}$ is given a Gaussian prior and $\boldsymbol{\eta}$ is given a Gaussian process model, and hence, $\widetilde{\mathbf{y}}$ is considered a LGP. However, for Poisson and binomial data $\boldsymbol{\xi}$ is given a conditional GCM prior, and thus, $\widehat{\mathbf{y}}$ is a latent process but is non-Gaussian. Similarly, $\mathbf{y}$ includes $\boldsymbol{\delta}_y$ which is not Gaussian distributed, and hence, $\mathbf{y}$ is a latent process but is non-Gaussian.

3.2 Hyperprior Considerations

In general, $\boldsymbol{\theta}$ consists of variance/covariance parameters for fixed and random effects. Thus, the purpose of $\boldsymbol{\theta}$ and the hyperprior $\pi(\boldsymbol{\theta})$ is to allow our variance/covariance parameters to be unknown. When multiplying the likelihood, prior distributions, and process models the matrix $\mathbf{D}(\boldsymbol{\theta})$ appears in our expression of the posterior distribution for $\boldsymbol{\zeta}$ and $\mathbf{q}$, and represents a matrix square root factorization of covariance parameters. While we account for hyperparameter uncertainty, only the marginal posterior distribution for $\boldsymbol{\zeta}$ is used for inference (i.e., $f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\eta} | \mathbf{z})$), and consequently, the role of the hyperprior specification is diminished. However, EPR is extremely flexible and allows one to modify the dependence structure of hyperparameters easily, as $\pi(\boldsymbol{\theta})$ is allowed to be any non-conjugate hyperprior that is proper. To demonstrate that large modifications to our procedure are not required when changing hyperpriors, we consider several hyperpriors with different dependence structure that are common in spatial statistics, including diagonal covariance matrices (Sections 4.1 and 4.4), covariance matrices based on the exponential covariogram (Section 4.2), and covariance matrices based on the conditional autoregressive model (Section 4.3).

In the context of generalized linear mixed effects models, there are prior specifications for $\boldsymbol{\theta}$ that are either difficult to simulate from or are not possible to simulate from. For example, improper priors such as $f(\boldsymbol{\theta}) = 1$ or Jeffreys prior (Jeffreys, 1946), among others, can not be sampled from directly. Other priors may require MCMC to sample from including constrained/truncated Bayesian models with intractable normalizing constants (e.g., see Dunson and Neelon, 2003, among others), which would then obfuscate our goal of obtaining an "MCMC free" method. The hyperparameter $\boldsymbol{\theta}$ represents the variance/covariance parameters, and hence, its dimension is at most on the order of $r^2 + p^2$, and in dimension reduction settings, these values will be small (i.e., $r, p \ll n$). However, of course, one should expect computational difficulties with simulating from the Gaussian priors/process models (for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$) and hyperpriors as the dimension of $\boldsymbol{\theta}$ grows, as this is a common problem for many spatial statistical models (Bradley et al., 2016).

3.3 Spatial Process Modeling with Exact Posterior Regression

The mixed effects model specification in Section 3.1 may be deceptively simple; however, we emphasize that several modern statistical models can use EPR including process models (e.g., spatial and spatio-temporal statistical models). See Hodges (2013) for a thorough treatment of how spatial and temporal statistical models can be written as a richly parameterized mixed effects model.

Although, of course, process models are different from mixed effects models, implementation of additive process models are similar to that of mixed effects models for a given collection of location/times. For example, consider locations $\mathbf{s} \in D$, where D is a generic spatial domain (e.g., a lattice or subset of $\mathbb{R}^d$). We introduce process into our notation functionally so that, for example, Z_i is written as $Z(\mathbf{s}_i)$, where $\mathbf{s}_1, \ldots, \mathbf{s}_n \in D$. Consider the following multivariate spatial statistical model,

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta} + \mathbf{g}(\mathbf{s})'\boldsymbol{\eta} + (\xi(\mathbf{s}) - \delta(\mathbf{s})); \, \mathbf{s} \in D,$$

where $\mathbf{x}(\mathbf{s})$ is a p-dimensional vector of spatially varying covariates, $\mathbf{g}(\mathbf{s})$ is a r-dimensional vector of spatial basis functions, $\xi(\mathbf{s})$ is a random process, and $\delta(\mathbf{s})$ is an unknown mean function. Suppose we are interested in estimation and prediction at the observed locations $D_O = \{\mathbf{s}_i : i = 1, \ldots, n\}$ and an additional m locations $D_P \in \{\mathbf{u}_1, \ldots, \mathbf{u}_m\} \subset D$. Let $T = n + m$.

Then stacking over locations in $D_O \cup D_P$ yields,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\eta} + (\boldsymbol{\xi} - \boldsymbol{\delta}_y), \tag{9}$$

where "$\cup$" is the set union, n-dimensional vector $\mathbf{y} = (Y(\mathbf{s}) : \mathbf{s} \in D_O)'$, and the $n \times p$ matrix $\mathbf{X} = (\mathbf{x}(\mathbf{s}) : \mathbf{s} \in D_O)'$, where we note that $\mathbf{X}$ can be computed by pre-multiplying the covariates stacked over $D_O \cup D_P$ by a $n \times T$ incidence matrix $\mathbf{E} = (\mathbf{e}(\mathbf{s}) : s \in D_O \cup D_P)'$ with T-dimensional vector $\mathbf{e}(\mathbf{s}) \equiv (I(\mathbf{s} = \mathbf{s}_1), \ldots I(\mathbf{s} = \mathbf{s}_n), I(\mathbf{s} = \mathbf{u}_1), \ldots, I(\mathbf{s} = \mathbf{u}_m))'$ and $I(\cdot)$ denoting the indicator function. That is $\mathbf{X} = \mathbf{E}\mathbf{X}_T$, where the $T \times p$ matrix $\mathbf{X}_T = (\mathbf{x}(\mathbf{s}) : \mathbf{s} \in D_O \cup D_P)'$. In a similar manner let the $n \times T$ matrix $\mathbf{G} = \mathbf{E}\mathbf{G}_T$, where the $T \times T$ matrix $\mathbf{G}_T = (\mathbf{g}(\mathbf{s}) : \mathbf{s} \in D_O \cup D_P)'$. Here, we let $\mathbf{G}_T$ be the matrix square root of a parameterized covariance matrix (e.g., $\mathbf{G}_T$ may be the Cholesky of a $T \times T$ covariance matrix with (i, j)-th element defined by the exponential covariogram). We let $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\delta}_y$ in (9) have the same specifications as in Section 3 with $\mathbf{D}_\eta \equiv \mathbf{I}_r$. Comparing our mixed effects model setup in (4) and the process model specification in (9) we see that process modeling can be implemented with EPR. That is, Theorems 3.1 and 3.2 (i.e., EPR) can be applied to the stacked expression in Equation (9). We illustrate this with spatial basis function expansions, weakly stationary spatial processes, and the conditional autoregressive model (Besag et al., 1991) in Section 4. To predict the process at both observed and prediction locations, posterior summaries of $\widetilde{\boldsymbol{y}}_T = \mathbf{X}_T\boldsymbol{\beta} + \mathbf{G}_T\boldsymbol{\eta}$ will be used.

In the case where inference on $Y(\mathbf{s}_0)$ with $\mathbf{s}_0 \notin D_O \cup D_P$ is of interest then one does not need to modify D_P and re-run EPR provided one can compute $\mathbf{x}(\mathbf{s}_0)$ and $\mathbf{g}(\mathbf{s}_0)$. In this case, one can compute $\widetilde{\mathbf{Y}}^{[b]}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)'\boldsymbol{\beta}^{[b]}_{rep} + \mathbf{g}(\mathbf{s}_0)'\boldsymbol{\eta}^{[b]}_{rep}$, where $\boldsymbol{\beta}^{[b]}_{rep}$ and $\boldsymbol{\eta}^{[b]}_{rep}$ are the b-th independent replicate of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ obtained when sampling with $\mathbf{s}_0 \notin D_O \cup D_P$. Then posterior summaries of $\widetilde{\mathbf{Y}}^{[b]}(\mathbf{s}_0)$ across b can be used for inference on $Y(\mathbf{s}_0)$.

In practice, EPR may not always be scale-able for process modeling with large n and T, since it is not always straightforward to simulate directly from the prior distribution, nor is it always straightforward to compute $\mathbf{G}$. In this article, we consider one example with a reduced rank assumption (Cressie and Johannesson, 2008; Banerjee et al., 2008; Hughes and Haran, 2013) by defining $\mathbf{G}$ to consist of $r < T$ spatially referenced basis functions (e.g., see Section 4.1). Although we consider $r < T$ to achieve scalability, there are options to consider when implementing EPR with $r \geq T$. In particular, one might consider the "data subset model" from (Bradley, 2021) to achieve scale-able inference, or sparse matrix Cholesky decompositions (e.g., see Datta et al., 2016).

Computing $\mathbf{G}$ can be computationally challenging in certain cases. For example, if $\mathbf{G}$ is interpreted as a Cholesky matrix from a known covariance matrix with parameter vector $\boldsymbol{\theta}$ then there are well-known computational considerations here. In particular, sparse Cholesky strategies such as those used in Datta et al. (2016) would be needed in this setting. These difficulties are exacerbated by the fact that as one continually samples $\boldsymbol{\theta}$ from its hyperprior, one needs to recompute $\mathbf{G}$ if $\mathbf{G}$ is parameterized. In high-dimensional settings it would be more amenable to consider fixed classes (i.e., $\mathbf{G}$ not parameterized) of point-referenced (Wikle, 2010) or regional basis functions (Bradley et al., 2015), such as the Obled-Cruetin basis functions (Bradley et al., 2017), which would only need to be computed one time and without Cholesky factorizations.

3.4 Computational Considerations

For large n the EPR formulation may not look practically feasible. However, standard block matrix inversion techniques can be used to reduce the order of operations to inverses of $r \times r$ matrices, $p \times p$ matrices, and $n \times n$ diagonal matrices (Lu and Shiou, 2002).

Theorem 3.3. *The following expression holds,*

$$(\boldsymbol{H}'\boldsymbol{H})^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{A}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \\ -(\boldsymbol{D} - \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{A}^{-1} & (\boldsymbol{D} - \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \end{pmatrix}, \tag{10}$$

where $\boldsymbol{A} = 2\boldsymbol{I}_n$, the $n \times (p+r)$ matrix $\boldsymbol{B} = (\boldsymbol{X}, \boldsymbol{G})$, the $(p+r) \times (p+r)$ matrix

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_p & \boldsymbol{X}'\boldsymbol{G} \\ \boldsymbol{G}'\boldsymbol{X} & \boldsymbol{G}'\boldsymbol{G} + \boldsymbol{I}_r \end{pmatrix}, \tag{11}$$

the $(p+r) \times (p+r)$ matrix

$$(\boldsymbol{D} - \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} = \begin{pmatrix} \boldsymbol{A}^{*-1} + \boldsymbol{A}^{*-1}\boldsymbol{B}^*(\boldsymbol{D}^* - \boldsymbol{C}^*\boldsymbol{A}^{*-1}\boldsymbol{B}^*)^{-1}\boldsymbol{C}^*\boldsymbol{A}^{*-1} & -\boldsymbol{A}^{*-1}\boldsymbol{B}^*(\boldsymbol{D}^* - \boldsymbol{C}^*\boldsymbol{A}^{*-1}\boldsymbol{B}^*)^{-1} \\ -(\boldsymbol{D}^* - \boldsymbol{C}^*\boldsymbol{A}^{*-1}\boldsymbol{B}^*)^{-1}\boldsymbol{C}^*\boldsymbol{A}^{*-1} & (\boldsymbol{D}^* - \boldsymbol{C}^*\boldsymbol{A}^{*-1}\boldsymbol{B}^*)^{-1} \end{pmatrix},$$

the $p \times p$ matrix $\boldsymbol{A}^* = \frac{1}{2}\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_p$, the $p \times r$ matrix $\boldsymbol{B}^* = \frac{1}{2}\boldsymbol{X}'\boldsymbol{G}$, the $r \times p$ matrix $\boldsymbol{C}^* = \frac{1}{2}\boldsymbol{G}'\boldsymbol{X}$, and the $r \times r$ matrix $\boldsymbol{D}^* = \frac{1}{2}\boldsymbol{G}'\boldsymbol{G} + \boldsymbol{I}_r$.

Proof: See Supplementary Appendix B.

Theorem 3.3 allows us to reduce the inverse of the $(n+p+r) \times (n+p+r)$ matrix $\mathbf{H}'\mathbf{H}$ to the inverse of the $p \times p$ matrix $\mathbf{A}^*$, and the $r \times r$ matrix $(\mathbf{D}^* - \mathbf{C}^*\mathbf{A}^{*-1}\mathbf{B}^*)^{-1}$. When p and r are both "small," these inverses are computationally efficient.

Simulation from the marginal posterior distribution of $\boldsymbol{\zeta}$ using EPR does not necessarily require first computing a matrix of the form $(\mathbf{H}'\mathbf{H})^{-1}$, storing this matrix, and then computing a $(n+p+r)$-dimensional vector of the form $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}$. In fact this order of operations may require impractical storage, since the $(n+p+r) \times (n+p+r)$ matrix $(\mathbf{H}'\mathbf{H})^{-1}$ may be high-dimensional. To avoid these issues one can instead compute/store the $(n+p+r)$-dimensional vector of the form $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}$ that avoids storage of high-dimensional matrices.

Theorem 3.4. *Let $\boldsymbol{w} = (\boldsymbol{w}_e', \boldsymbol{w}_\beta', \boldsymbol{w}_\eta', \boldsymbol{w}_\xi{}')'$, $\boldsymbol{w}_e \in \mathbb{R}^n$, $\boldsymbol{w}_\beta \in \mathbb{R}^p$, $\boldsymbol{w}_\eta \in \mathbb{R}^r$, and $\boldsymbol{w}_\xi \in \mathbb{R}^n$. Then the following expression holds,*

$$(\boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{w} = \begin{pmatrix} (\boldsymbol{F} - \boldsymbol{K}\boldsymbol{L}^{-1}\boldsymbol{K}')^{-1}(\boldsymbol{R} - \boldsymbol{K}\boldsymbol{L}^{-1}\boldsymbol{P}) \\ -\boldsymbol{L}^{-1}\boldsymbol{K}'(\boldsymbol{F} - \boldsymbol{K}\boldsymbol{L}^{-1}\boldsymbol{K}')^{-1}(\boldsymbol{R} - \boldsymbol{K}\boldsymbol{L}^{-1}\boldsymbol{P}) + \boldsymbol{L}^{-1}\boldsymbol{P} \end{pmatrix}, \qquad (12)$$

where the $(n+p)$-dimensional vector $\boldsymbol{R} = (\boldsymbol{w}_e' + \boldsymbol{w}_\xi{}', \boldsymbol{w}_e'\boldsymbol{X} + \boldsymbol{w}_\beta')'$, the r-dimensional vector $\boldsymbol{P} = \boldsymbol{G}'\boldsymbol{w}_e + \boldsymbol{w}_\eta$, the $r \times (n+p)$ matrix $\boldsymbol{K}' = (\boldsymbol{G}', \boldsymbol{G}'\boldsymbol{X})$, the $r \times r$ matrix $\boldsymbol{L} = \boldsymbol{G}'\boldsymbol{G} + \boldsymbol{I}_r$, the $(n+p) \times (n+p)$ matrix $\boldsymbol{F} = \begin{pmatrix} 2\boldsymbol{I}_n & \boldsymbol{X} \\ \boldsymbol{X}' & \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_p \end{pmatrix}$, and the $(n+p) \times (n+p)$ matrix

$$\boldsymbol{F} - \boldsymbol{K}\boldsymbol{L}^{-1}\boldsymbol{K}' = \begin{pmatrix} \boldsymbol{F}_1 & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{12}' & \boldsymbol{F}_2 \end{pmatrix}.$$

The $(n+p) \times (n+p)$ matrix,

$$(\boldsymbol{F} - \boldsymbol{K}\boldsymbol{L}^{-1}\boldsymbol{K}')^{-1} = \begin{pmatrix} \boldsymbol{F}_{11} & \boldsymbol{F}_{12} \\ \boldsymbol{F}_{21} & \boldsymbol{F}_{22} \end{pmatrix},$$

where the $n \times n$ matrix $\boldsymbol{F}_1 = 2\boldsymbol{I}_n - \boldsymbol{G}\boldsymbol{L}^{-1}\boldsymbol{G}'$, the $n \times p$ matrix $\boldsymbol{B}_{12} = \boldsymbol{X} - \boldsymbol{G}\boldsymbol{L}^{-1}\boldsymbol{G}'\boldsymbol{X}$, the $p \times p$ matrix $\boldsymbol{F}_2 = \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_p - \boldsymbol{X}'\boldsymbol{G}\boldsymbol{L}^{-1}\boldsymbol{G}'\boldsymbol{X}$, the $n \times n$ matrix $\boldsymbol{F}_{11} = \boldsymbol{F}_1^{-1} + \boldsymbol{F}_1^{-1}\boldsymbol{B}_{12}(\boldsymbol{F}_2 - \boldsymbol{B}_{12}'\boldsymbol{F}_1^{-1}\boldsymbol{B}_{12})^{-1}\boldsymbol{B}_{12}'\boldsymbol{F}_1^{-1}$, the $n \times p$ matrix $\boldsymbol{F}_{12} = -\boldsymbol{F}_1^{-1}\boldsymbol{B}_{12}(\boldsymbol{F}_2 - \boldsymbol{B}_{12}'\boldsymbol{F}_1^{-1}\boldsymbol{B}_{12})^{-1}$, the $p \times n$ matrix $\boldsymbol{F}_{21} = -(\boldsymbol{F}_2 - \boldsymbol{B}_{12}'\boldsymbol{F}_1^{-1}\boldsymbol{B}_{12})^{-1}\boldsymbol{B}_{12}'\boldsymbol{F}_1^{-1}$, the $p \times p$ matrix $\boldsymbol{F}_{22} = (\boldsymbol{F}_2 - \boldsymbol{B}_{12}'\boldsymbol{F}_1^{-1}\boldsymbol{B}_{12})^{-1}$, and the $n \times n$ matrix $\boldsymbol{F}_1^{-1} = \frac{1}{2}\boldsymbol{I}_n + \frac{1}{4}\boldsymbol{G}(\boldsymbol{L} - \frac{1}{2}\boldsymbol{G}'\boldsymbol{G})^{-1}\boldsymbol{G}'$.

Proof: See Supplementary Appendix B.

Careful examination of the order of operations show that Theorem (3.4) allows one to compute the vector $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}$ by storing/computing the $n \times p$ matrix $\mathbf{X}$, the $n \times r$ matrix $\mathbf{G}$ (when $r = n$ we set $\mathbf{G} = \mathbf{I}_n$), the $r \times r$ matrix $\mathbf{L}^{-1}$, the $r \times r$ matrix $(\mathbf{L} - \frac{1}{2}\mathbf{G}'\mathbf{G})^{-1}$, the $p \times p$ matrix $\mathbf{F}_2$, and the $p \times p$ matrix $\mathbf{F}_{22}$. These computations are straightforward when r and p are "small" or when p is small and $\mathbf{G}$ is diagonal.

3.5 Implementation of Exact Posterior Regression

The following gives step-by-step instructions on obtaining efficient independent replicates directly from the posterior distribution of $\boldsymbol{\zeta}$ (i.e., $f(\boldsymbol{\zeta}|\mathbf{z})$) using Theorem 3.1, which we refer to as EPR. We consider Gaussian data with unknown non-constant variance, Poisson data, and binomial data.

1. Store/compute the $n \times p$ matrix $\mathbf{X}$, the $n \times r$ matrix $\mathbf{G}$, the $r \times r$ matrix $\mathbf{L}^{-1}$, the $r \times r$ matrix $(\mathbf{L} - \frac{1}{2}\mathbf{G}'\mathbf{G})^{-1}$, the $p \times p$ matrix $\mathbf{F}_2$, and the $p \times p$ matrix $\mathbf{F}_{22}$. Set b = 1.

2. Simulate $\mathbf{w}$ according to Theorem 3.2.

3. From Theorem 3.2 compute

$$\begin{aligned}
\boldsymbol{\zeta}_{rep}^{[b]} &= (\boldsymbol{\xi}_{rep}^{[b]\prime}, \boldsymbol{\beta}_{rep}^{[b]\prime}, \boldsymbol{\eta}_{rep}^{[b]\prime})' = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w} \\
\mathbf{y}_{rep}^{[b]} &= (\mathbf{I}_n, \mathbf{0}_{n,n+p+r})\mathbf{w} \\
\widehat{\mathbf{y}}_{rep}^{[b]} &= \mathbf{X}\boldsymbol{\beta}_{rep}^{[b]} + \mathbf{G}\boldsymbol{\eta}_{rep}^{[b]} + \boldsymbol{\xi}_{rep}^{[b]} \\
\widetilde{\mathbf{y}}_{rep}^{[b]} &= \mathbf{X}\boldsymbol{\beta}_{rep}^{[b]} + \mathbf{G}\boldsymbol{\eta}_{rep}^{[b]},
\end{aligned}$$

with the value of $\mathbf{w}$ generated in Step 2. Efficient computation of $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}$ can be achieved via Theorem 3.4 and the values stored in Item 1.

4. Set $b = b+1$. Repeat Steps $2-3$ until $b = B$ when $\mathbf{G}$ does not consist of unknown parameters. Repeat Steps $1-3$ until $b = B$ when $\mathbf{G}$ is parameterized.

The goal of this algorithm is to provide B independent replicates of $(\boldsymbol{\zeta}_{rep}, \mathbf{y}_{rep}, \widehat{\mathbf{y}}_{rep}, \widetilde{\mathbf{y}}_{rep})$. Since Theorem 3.2, produces a single draw of $(\boldsymbol{\zeta}_{rep}, \mathbf{y}_{rep}, \widehat{\mathbf{y}}_{rep}, \widetilde{\mathbf{y}}_{rep})$, we repeat the computations in Theorem 3.2 (in Step 3) B times to produce B independent replicates of $(\boldsymbol{\zeta}_{rep}, \mathbf{y}_{rep}, \widehat{\mathbf{y}}_{rep}, \widetilde{\mathbf{y}}_{rep})$. Summaries of the B independent posterior replicates are used for inference (e.g., component-wise means, variances, quantiles, etc.).

To sample from the GCM posterior according to Theorem 3.2, the main computational bottleneck is the computation of $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{w}$. Theorem 3.4 reduces the computational complexity from the naive calculation of an inverse of a $(n+p+r) \times (n+p+r)$ matrix to the computational complexity arising from an inverse of a $n \times n$ diagonal matrix, inverses of $p \times p$ matrices, and inverses of $r \times r$ matrices. Thus, not only are we able to obtain independent replicates drawn directly from the posterior, if the values of p and r are small (i.e., a dimension reduction setting), these replicates are easy to compute.

When using known basis function expansions to define $\mathbf{G}$, repeated matrix operations that one might see in a Gibbs sampler are avoided, since matrix inversions are only required *a single* *time* in Step 1. Additionally, B does not have to be as large as what one requires for an MCMC, since one does not require a burn-in period, thinning, or have concerns of mixing and positive autocorrelations in the MCMC.

4 Illustrations

We provide illustrations of EPR in several settings covered in standard textbooks on spatial and spatio-temporal modeling (e.g., see Cressie, 1993; Rue and Held, 2005; Cressie and Wikle, 2011; Banerjee et al., 2015, among others). In particular, we illustrate EPR in the context of spatial basis function expansions, weakly stationary spatial processes, conditional autoregressive models, and dimension reduction in high-dimensional settings.

In our illustrations, we offer comparisons to standard spatial statistical models (i.e., LGPs) fitted via INLA and MCMC. By standard we mean $\mathbf{q}$ equal to a zero vector and $\alpha_\xi = 0$ to produce an LGP. Moreover, we simulate the data from an LGP so that $\mathbf{q}$ is equal to the zero vector. As a result, the models fitted with INLA and MCMC are correctly specified to not include a discrepancy term, however, our model implemented with EPR is misspecified, since it models $\boldsymbol{\xi}$ with a conditional GCM (for non-Gaussian data) and $\mathbf{q}$ as not necessarily equal to zero. This misspecified setting is of particular interest because it is arguably more common to drop $\mathbf{q}$ from the additive model and generate $\boldsymbol{\xi}$ from a normal distribution (or set equal to zero). Furthermore, these simulation studies will allow us to empirically investigate the implications of including both our conditional GCM model for $\boldsymbol{\xi}$ and the discrepancy term $\mathbf{q}$ on inference. We provide simulation studies for the case where the data is generated with a discrepancy term in Supplementary Appendix F.

We also provide a sensitivity study in Supplementary Appendix G on the choice of B and found that $B = 100$ provides similar results as $B = 500$ and $B = 1,000$. As such, we set $B = 100$ in this section. All prior specifications used for the models fitted via INLA, MCMC and EPR are listed in Supplementary Appendix H. To avoid an unfavorable specification, the models fitted with the computational tools INLA and MCMC are based on their package's default prior specifications or flat prior specifications when no default is available.

Our simulations are fairly low dimensional so that it is straightforward to obtain results over multiple replicates (i.e., $n = 400$). As such, we include a high-dimensional illustration in Section 4.4. This example demonstrates the scale in which EPR can be implemented.

4.1 Spatial Basis Function Expansions

Spatial basis function expansions have become a standard in spatial statistics, with common classes of basis functions including Fourier basis functions, wavelet basis functions (Huang and Cressie, 1999), radial basis functions (Cressie and Johannesson, 2008), and splines (Wahba, 1990), among others. In this section, we compare models that make use of basis function expansions fitted with EPR, INLA, and MCMC. The Bayesian hierarchical models implemented with INLA and MCMC both correctly assume that no discrepancy term is present. Additionally, the models that apply both INLA and MCMC use improper priors for fixed and random effects (i.e., $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$), and the variance parameter for normal data. For ease of exposition, we refer to the model fitted with both INLA and MCMC as the "null-discrepancy model," since the data model is specified without a discrepancy term. Similarly, we call the model fitted with EPR as the "discrepancy model," since it incorrectly includes an additional discrepancy term in the additive model. The discrepancy model assumes an inverse gamma priors on variance parameters. The range parameter is given a uniform zero to 0.5 prior. See Supplementary Appendix H for more details. The Pólya-Gamma augmentation technique (Polson et al., 2013) is a particularly efficient approach to fit latent Gaussian process models using MCMC, and is one of the more computationally competitive techniques in MCMC, which we use for Bernoulli data in this section. In the Poisson MCMC implementation we make use of a new extremely efficient algorithm by D'Angelo and Canale (2022) that extends the Pólya-Gamma augmentation technique for Poisson data. MCMC was implemented with 10,000 replicates with a burn-in of 5,000.

We simulate data from the following models,

$$
\begin{aligned}
Z_1(\mathbf{s})|\eta_1,\ldots,\eta_{30},\xi(\mathbf{s}),\sigma^2(\mathbf{s}) &\sim \text{Normal}\left(-1-x_1(\mathbf{s})-x_2(\mathbf{s})+\sum_{j=1}^{30}g_j(\mathbf{s})\eta_j+\xi(\mathbf{s}),\sigma^2(\mathbf{s})\right)\\
Z_2(\mathbf{s})|\eta_1,\ldots,\eta_{30},\xi(\mathbf{s}) &\sim \text{Poisson}\left\{\exp\left(-1+0.5\,x_1(\mathbf{s})+0.4\,x_2(\mathbf{s})+\sum_{j=1}^{30}g_j(\mathbf{s})\eta_j+\xi(\mathbf{s})\right)\right\}\\
Z_3(\mathbf{s})|\eta_1,\ldots,\eta_{30},\xi(\mathbf{s}) &\sim \text{Bernoulli}\left\{\frac{\exp\left(-2-x_1(\mathbf{s})-2\,x_2(\mathbf{s})+\sum_{j=1}^{30}g_j(\mathbf{s})\eta_j+\xi(\mathbf{s})\right)}{1+\exp\left(-2-x_1(\mathbf{s})-2\,x_2(\mathbf{s})+\sum_{j=1}^{30}g_j(\mathbf{s})\eta_j+\xi(\mathbf{s})\right)}\right\},
\end{aligned}
\tag{13}
$$

where $\{\eta_j\}$ are independently distributed according to a normal distribution with mean zero and variance 0.04, $s \in \{0, 0.002, \ldots, 1\}$, $\sigma^2(\mathbf{s})$ are uniform distributed over 0.15 to 2, and for each $\mathbf{s}$ we sample $\xi(\mathbf{s})$ independently from a normal distribution with mean zero and variance of 0.02 for Bernoulli data, variance of 0.01 for Poisson data, and variance of 0.15 for normal data. We observe $n = 400$ randomly selected locations, $x_1(s)$ is an independent Bernoulli random variable with probability $\exp(s)/(1+\exp(s))$, $x_2(s)$ is an independent Bernoulli random variable with probability $\exp(-0.01s)/(1+\exp(-0.01s))$, $g_j(s) = \exp(-||s-u_j||^2)$, $\{u_j\}$ are equally spaced across the spatial domain, and $||\cdot||$ is the Euclidean distance. All models use the basis functions and covariates that generate the data.

In Figure 1, we see that each method is fairly comparable in terms of predictive performance. Moreover, the discrepancy model implemented with EPR tends to give larger measures of uncertainty than the null-discrepancy model fitted with INLA and MCMC, both of which produce credible intervals that do not contain the true values of the latent process. The discrepancy model implemented with EPR is misspecified in this simulation study, since it includes a conditional GCM model for $\boldsymbol{\xi}$ (for non-Gaussian settings) and the discrepancy term $\mathbf{q}$. In general, as you include additional random effects into a model, one obtains larger credible intervals. As a result,

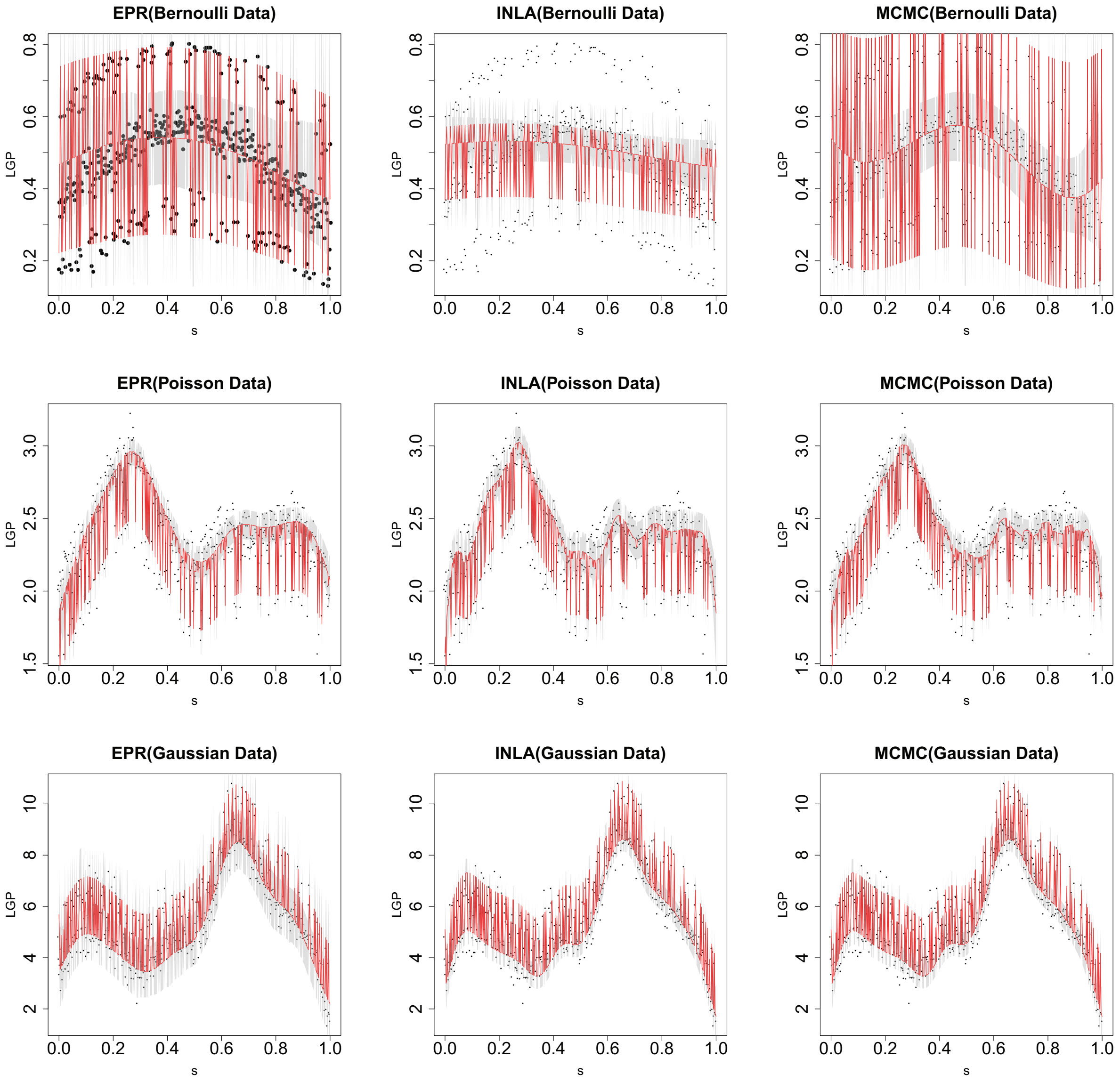

Figure 1: Predictions for models fitted with EPR, INLA, and MCMC computed for the spatial basis function illustration. The first row presents results for binary spatial data, the second row presents results for Poisson spatial data, and the third row presents results for Gaussian spatial data. The black points represent the true value of the latent process, the red line represents the posterior mean of $\widetilde{\mathbf{y}}$, and the shaded region represents pointwise 95% credible intervals, respectively.

EPR's credible intervals tend to be more conservative (i.e., larger), which in turn lead to credible intervals that tend to cover the generative true values more often. The fact that the predictions are similar is notable, since INLA and MCMC are both approximate methods (MCMC is exact in the limit), whereas EPR is an exact method for a different hierarchical model. Moreover, EPR (and INLA) does not require the additional overhead of MCMC diagnostics.

To assess the performance over multiple replicates, we use the central processing unit (CPU) time (seconds), the mean squared error (MSE) between the estimated regression coefficients and $\{\eta_j\}$ and corresponding true values, the mean squared prediction error (MSPE) between the latent process and predicted latent process (using $\widetilde{\mathbf{y}}$), and the continuous rank probability score (CRPS) (Gneiting and Katzfuss, 2014) averaged over missing locations and scaled so that small values are preferable. The CRPS is useful since it is metric that evaluates the entire predictive distribution so that uncertainty in the predictions is considered. In Table 1, we provide the average MSPE, MSE, CRPS, and CPU plus or minus two standard deviations over 50 independent replicates by method and data type.

In Table 1, the confidence intervals (CIs) for MSPE in the normal and Poisson setting are all overlapping, suggesting that the discrepancy model implemented with EPR does the same as the null-discrepancy model implemented with either INLA or MCMC. However, in the logistic regression case the discrepancy model implemented with EPR has preferable MSPE than the null-discrepancy model fitted with either INLA or MCMC. Similarly, the discrepancy model fitted with EPR appears preferable in terms of CRPS to the null-discrepancy model in the logistic (fitted with either INLA or MCMC) and Poisson (fitted with MCMC) regression cases. The CRPS appears indistinguishable in the normal regression case for all methods/computational tools, as the CIs overlap. The MSE is preferable for the discrepancy model fitted with EPR in the logistic and normal regression cases, but appears indistinguishable (confidence intervals overlap) to that of the null-discrepancy model fitted with MCMC in the Poisson setting. EPR is consistently and considerably preferable in terms CPU time in all settings. The performance in CPU time is especially notable, since the null-discrepancy model fitted with INLA and MCMC are both approximate methods (MCMC is exact in the limit), whereas EPR is an exact MCMC free method for a differ-

	Type	MSPE	MSE	CRPS	CPU
EPR	Logistic	0.0038 (0.0033,0.0042)	0.189 (0.166,0.211)	0.167 (0.150,0.184)	0.487 (0.468,0.507)
INLA	Logistic	0.0068 (0.0056,0.0081)	2.098 (1.397,2.799)	0.215 (0.190,0.239)	2.320 (2.241,2.398)
MCMC	Logistic	0.0046 (0.0040,0.0051)	3.626 (2.715,4.537)	0.567 (0.560,0.574)	125.115 (123.291,126.939)
EPR	Poisson	0.011 (0.01095,0.01186)	0.0095 (0.0090,0.01000)	0.060 (0.058,0.062)	0.46 (0.45,0.47)
INLA	Poisson	0.014 (0.013,0.015)	20.552 (17.204,23.899)	0.064 (0.062,0.066)	2.202 (2.135,2.270)
MCMC	Poisson	0.012 (0.0118,0.0131)	0.088 (0.065,0.111)	8.68 (8.62,8.73)	23.737 (23.676,23.798)
EPR	Normal	0.202 (0.197,0.207)	1.863 (1.810,1.916)	0.292 (0.278,0.307)	0.346 (0.327,0.365)
INLA	Normal	0.202 (0.196,0.208)	28.634 (24.378,32.891)	0.283 (0.278,0.290)	2.589 (2.515,2.663)
MCMC	Normal	0.202 (0.196,0.208)	26.730 (22.836,30.624)	0.284 (0.278,0.290)	24.203 (23.369,25.037)

Table 1: Fifty independent replicates data vectors are drawn according to (13), and several methods are applied to each replicated data vector. The first column indicates EPR, INLA, and MCMC, which are applied to the discrepancy model and null-discrepancy model (see Supplementary Appendix H for more detail). The type column indicates logistic regression, Poisson regression, and normal regression. The values represent averages over 50 independent simulated data sets and the parenthetical represent the confidence interval (CI) (i.e., average plus or minus two standard deviations). The MSE, MSPE, CRPS, and CPU (in seconds) are indicated in the column header. The MSPE for Poisson regression is computed on the log-scale so that the values are easier to present, where logistic spatial regression's MSPE was computed on the expit scale. Outliers were removed when computing MSE values.

ent hierarchical model. That is, EPR produces similar-to-better predictions and superior regression estimates in a faster time than that of the state-of-the-art approximate Bayes and MCMC based methods in this study.

4.2 Weakly Stationary Spatial Processes

A classical assumption for spatially referenced data is that the latent spatial process is weakly stationary. In particular, weakly stationary spatial processes have mean zero and the covariance of the process at any two locations is a positive definite function evaluated at the spatial lag, where this covariance function is referred to as a covariogram. In this section, we compare models fitted with EPR, INLA, and MCMC. We fit an LGP implemented with MCMC through the R package spBayes (Finley et al., 2012) using the exponential covariogram. The exponential covariogram is a well-known choice, but there are several other choices available (e.g, see Cressie, 1993, among others). The simulated data are generated as follows,

$$\begin{aligned} Z_1(\mathbf{s})|\nu(\mathbf{s}),\sigma^2(\mathbf{s}) &\sim \text{Normal}\left(-x(\mathbf{s})+\nu(\mathbf{s}),\sigma^2(\mathbf{s})\right) \\ Z_2(\mathbf{s})|\nu(\mathbf{s}) &\sim \text{Poisson}\left\{\exp\left(3+2\,x(\mathbf{s})+\nu(\mathbf{s})\right)\right\} \\ Z_3(\mathbf{s})|\nu(\mathbf{s}) &\sim \text{Bernoulli}\left\{\frac{\exp\left(-x(\mathbf{s})+\nu(\mathbf{s})\right)}{1+\exp\left(-x(\mathbf{s})+\nu(\mathbf{s})\right)}\right\}, \end{aligned} \tag{14}$$

where $x(\cdot)$ are generated from a standard uniform distribution of a 15×15 grid of the unit square, $\sigma^2(\mathbf{s})$ are uniform distributed over 0.15 to 2, and $\nu(\mathbf{s})$ is generated as a weakly stationary spatial process with exponential covariogram with range parameter 0.25, nugget zero for Gaussian and Bernoulli data, nugget 0.64 for Poisson data, and variance 2 on a 15×15 grid of the unit square. The slope, intercept, and nugget for the Poisson example were chosen so that the percent of zero count-valued observations to be small (roughly one percent) to avoid zero inflation. The discrepancy model is fitted with EPR, and specifies inverse gamma priors on all variance parameters with shape 1 and rate parameter given a gamma hyperprior with shape and rate set to 1. The range parameter is given a uniform zero to 0.5 prior. In this example, the discrepancy model specifies **G** via a Cholesky matrix of the covariance matrix formed by the exponential covariogram. The default prior specifications are used when implementing the stochastic partial differential equation (SPDE) approach (Lindgren et al., 2011) using INLA, and spBayes default prior specifications are used for the LGP implementation via MCMC (see Supplementary Appendix H for more details). The data models for SPDE and LGP correctly do not include a discrepancy term.

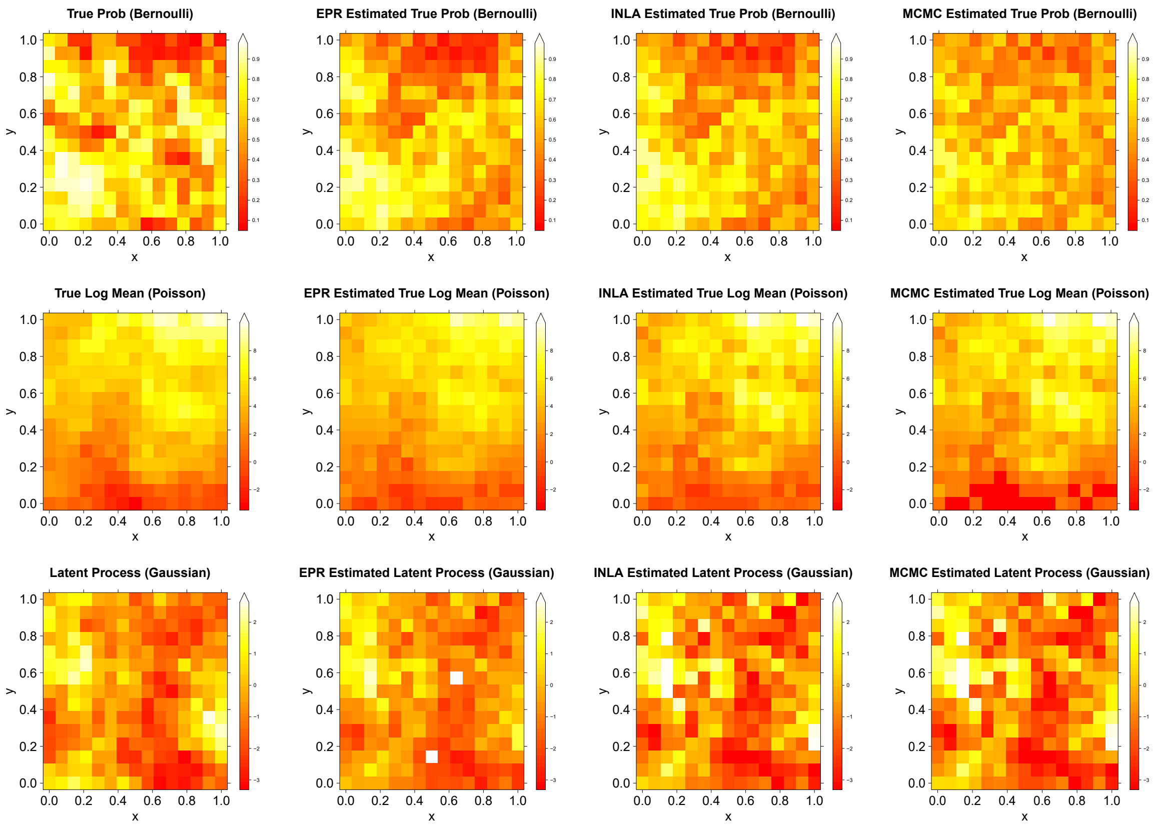

Figure 2: Illustration of predictions from models fitted with EPR, INLA, and MCMC for weakly stationary processes. The first row presents results for binary spatial data, the second row presents results for Poisson spatial data, and the third row presents results for Gaussian spatial data. The left column contains the latent process on the inverse link scale. Second, Third, and Fourth columns display the posterior mean when using EPR, INLA, and MCMC, respectively. Let $\mathbf{s} = (x, y)'$.

In Figure 2, we provide plots of one simulated replicate and fitted means computed using EPR, INLA, and MCMC. The fitted posterior standard deviations for this example are provided in Figure 3. In general, we see that all methods perform similarly for this example, however, the discrepancy model implemented with EPR tends to have larger posterior standard deviation. MCMC can result in larger estimates of variability, as quite often positive dependence is introduced via rejection steps in a Metropolis algorithm. EPR's larger uncertainty quantification here are due to the fact that the model EPR is derived from is misspecified in this simulation study, since it includes $\mathbf{q}$. Similar to Section 4.1, we see that EPR's credible intervals tend to be more conservative

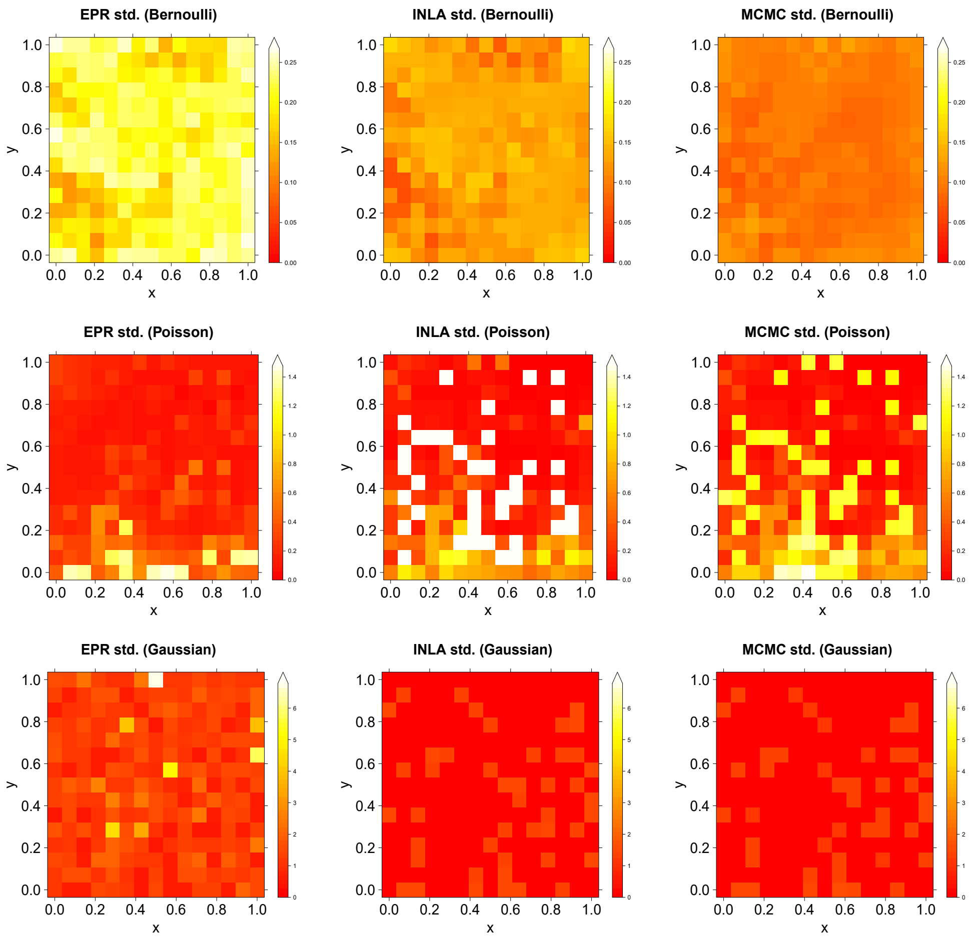

Figure 3: Illustration of posterior standard deviations from models fitted with EPR, INLA, and MCMC for weakly stationary processes for the simulated replicate in the first column of Figure 2. The first row presents results for binary spatial data, the second row presents results for Poisson spatial data, and the third row presents results for Gaussian spatial data. First, second, and third columns display the posterior standard deviations computed using discrepancy model with EPR, SPDE/INLA, and LGP/MCMC, respectively.

(i.e., larger), which in turn lead to credible intervals that tend to cover the generative true values more often. The fact that the predictions are similar is again notable since, INLA and MCMC are both approximate methods (MCMC is exact in the limit), whereas EPR is an exact MCMC free method for a different model.

Instead of writing SPDE implemented with INLA, we write SPDE/INLA, and similarly we

	Type	MSPE	MSE	CRPS	CPU
EPR	Logistic	0.0288 (0.0270, 0.0306)	0.551 (0.322, 0.780)	0.631 (0.610, 0.651)	15.224 (14.801, 15.647)
INLA	Logistic	0.0313 (0.0293, 0.0333)	0.660 (0.303, 1.018)	0.651 (0.608, 0.694)	8.461 (6.150, 10.771)
MCMC	Logistic	0.0383 (0.0354, 0.0412)	0.503 (0.240, 0.756)	1.009 (0.964, 1.054)	363.162 (346.763, 379.560)
EPR	Poisson	0.564 (0.544, 0.583)	4.662 (4.019, 5.307)	0.591 (0.567, 0.614)	16.645 (15.081, 18.208)
INLA	Poisson	0.845 (0.818, 0.873)	1.328 (0.788, 1.868)	0.581 (0.563, 0.599)	5.285 (4.892, 5.678)
MCMC	Poisson	0.815 (0.790, 0.840)	8.771 (3.777, 13.765)	0.510 (0.497, 0.524)	392.913 (363.218, 422.608)
EPR	Normal	0.736 (0.709, 0.763)	0.571 (0.382, 0.759)	0.529 (0.506, 0.551)	14.862 (14.543, 15.181)
INLA	Normal	1.423 (1.370, 1.475)	1.437 (0.905, 1.970)	0.553 (0.538, 0.567)	8.669 (8.129, 9.209)
MCMC	Normal	1.469 (1.413, 1.525)	1.940 (1.416, 2.464)	0.586 (0.568, 0.604)	122.448 (121.527, 123.369)

Table 2: Fifty independent replicates data vectors are drawn according to (14), and several methods are applied to each replicated data vector. The first column indicates EPR, INLA, and MCMC, which are applied to the discrepancy model, SPDE, and an LGP (see Supplementary Appendix H for more detail). The type column indicates logistic regression, Poisson regression, and normal regression. The values represent averages over 50 independent simulated data sets and the parenthetical represent the confidence interval (CI) (i.e., average plus or minus two standard deviations). The MSE, MSPE, CRPS, and CPU (in seconds) are indicated in the column header. The MSPE for Poisson regression is computed on the log-scale so that the values are easier to present, where logistic spatial regression's MSPE was computed on the expit scale.

write LGP/MCMC. In Table 2, when comparing the confidence intervals for MSPE in the Poisson and normal data cases, we see that the discrepancy model fitted with EPR outperforms SPDE/INLA and the LGP/MCMC model. In the logistic regression case, the MSPE of the discrepancy model fitted with EPR is preferable to that of the LGP/MCMC model. The discrepancy model fitted with EPR is preferable in terms of MSE for the normal case, has overlapping CIs with SPDE/INLA and LGP/MCMC in the logistic case, and is less preferable than SPDE/INLA in the Poisson regression case. The discrepancy model fitted with EPR is preferable to LGP/MCMC in terms of CRPS for the normal and logistic settings, is less preferable than LGP/MCMC in the Poisson regression case, and has overlapping CIs with SPDE/INLA in all cases. For all three types of spatial linear mixed models the LGP/MCMC has a considerably larger CPU time than SPDE/INLA and the discrepancy model implemented with EPR, and SPDE/INLA has moderately smaller CPU time than the discrepancy model implemented with EPR. EPR performs marginally slower than it did in Section 4.1, since **G** needs to be computed every step of the sampler, whereas the radial basis function in Section 4.1 only needed to be computed once.

4.3 Intrinsic Conditional Autoregressive Model for American Community Survey Poverty Estimates

The U.S. Census Bureau's ACS provides demographic statistics over several geographies and over 1-year and 5-year time periods (Torrieri, 2007). As such, it has become a very useful tool for poverty estimation (Molina and Rao, 2010). Small area estimation of poverty is a crucial and standard problem in both demography and official statistics (Rao, 2003), since it is a key variable for determining economic disparities, public policy, and monitoring the financial circumstances at various levels of geography (e.g., see Theil, 1996). Considering the wide-applicability of EPR, it is important to investigate its performance in standard settings such as poverty estimation. Consequently, in this section, we compare a Besag, York, and Mollié (BYM, Besag et al., 1991) model fitted with INLA and MCMC to a discrepancy model (implemented with EPR) for poverty estimation over U.S. counties in Florida in 2019 using ACS 1-year period estimates.

Standard demographic related covariates are used; namely, ACS five year period estimates of the median age, the ratio of the population of males to females, and the population (on the log scale) of those who identify as white alone, black or African American alone, and Asian alone. We assume the population of those under the poverty status as binomial distributed with m_i representing the i-th county's population. The discrepancy model assumes $\sigma_\xi^2 = 0.5$ (chosen with

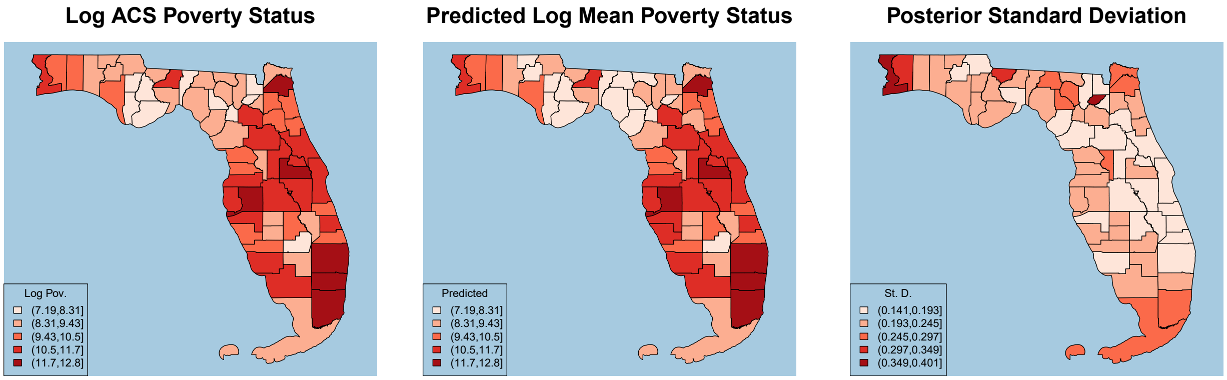

Figure 4: We plot the log ACS estimates (as a reference) along with a plot of the posterior mean and standard deviations computed from 500 independent replicates of EPR.

cross validation), and the default prior specifications are used for INLA and MCMC for the BYM model. Let $\mathbf{W}$ be the row normalized first order binary adjacency matrix. When implementing the discrepancy model with EPR we define $\mathbf{G}$ to be the value such that $\mathbf{G}\mathbf{G}' = (\mathbf{I}_n - \mathbf{W})^{-1}$, which is the matrix square root of the covariance matrix (computed using the spectral decomposition) implied by the intrinsic conditional autoregressive model with unit variance. We apply EPR according to Section 3.5 with $B = 500$ independent replicates from the posterior distribution for $\boldsymbol{\zeta}$ and BYM with MCMC using the R package `CARBayes` with 20,000 replicates with a burn-in of 10,000 (Lee, 2013).

Plots of the predicted mean and standard deviation of $\widetilde{\mathbf{y}}$ versus the log-data using EPR are provided in Figure 4. In general, we see predictions that reflect the pattern of the data with spatial smoothing. Table 3 contains several metrics comparing the predictive performance of the discrepancy model implemented with EPR, a BYM fitted with INLA, and a BYM fitted with MCMC. The leave-one-out cross validation error (Wahba, 1990) is used to assess the predictive performance. Specifically, an observation is left out, and the model is used to predict this value. We compute the relative cross-validation (CV) error and a leave-one-out strictly proper

	CV	CRPS	CPU
EPR	0.1558	0.1699	49.14
INLA	0.1567	0.1756	243.51
MCMC	0.8461	1.2678	661.49

Table 3: Let CV be the relative leave-one-out cross-validation error for poverty computed on the logit scale. That is, let CV $\equiv \underset{i\in\{1,\ldots,n\}}{\text{mean}}\left\{\text{abs}\left(\text{logit}\left(\frac{Z_i}{m_i}\right)-E_{-i}\left[\widetilde{Y}_i\right]\right)/\text{abs}(\text{logit}\left(\frac{Z_i}{m_i}\right))\right\}$, where E_{-i} is the posterior expected value that leaves out Z_i, "abs" is the absolute value operator, and "logit" is the logit operator. In the column CRPS we evaluate the average CRPS evaluated at the leave one out logit-value. We also provide the CPU time (seconds) to compute the leave-out-out cross-validation criterion.

scoring rule (e.g., see Yao et al., 2018, among others). Specifically, the relative CV is defined as CV $\equiv \underset{i\in\{1,\ldots,n\}}{\text{mean}}\left\{\text{abs}\left(\text{logit}\left(\frac{Z_i}{m_i}\right)-E_{-i}\left[\widetilde{Y}_i\right]\right)/\text{abs}(\text{logit}\left(\frac{Z_i}{m_i}\right))\right\}$, where E_{-i} is the posterior expected value that leaves out Z_i, "abs" is the absolute value operator, $\widetilde{Y}_i$ is the i-th component of $\widetilde{\mathbf{y}}$ and "logit" is the logit operator. We use the word "relative" since the absolute error $\text{abs}\left(\text{logit}\left(\frac{Z_i}{m_i}\right)-E_{-i}\left[\widetilde{Y}_i\right]\right)$ is weighted by (or is relative to) $1/\text{abs}(\text{logit}\left(\frac{Z_i}{m_i}\right))$. The leave-one-out CRPS $\equiv \sum_{i=1}^{n}\text{crps}(\text{Normal}\left[E_{-i}(\widetilde{Y}_i), var_{-i}(\widetilde{Y}_i)\right], \text{logit}(Z_i/m_i))$, where var_{-i} is the posterior variance that leaves out Z_i, and Φ is the standard normal cumulative distribution function, ϕ is the standard normal pdf function, and $\text{crps}(\text{Normal}(\mu,\sigma^2),x) \equiv \sigma\left\{\frac{x-\mu}{\sigma}\left[2\Phi\left(\frac{x-\mu}{\sigma}\right)-1\right]+2\phi\left(\frac{x-\mu}{\sigma}\right)-\frac{1}{\sqrt{\pi}}\right\}$ is the continuous rank probability score from Gneiting et al. (2005). The relative CV suggest that leave-one-out predictions are roughly within 15% of the hold-out proportion for the discrepancy model implemented with EPR and BYM/INLA, and BYM/MCMC has a large relative CV at 85%. These results suggest that the discrepancy model implemented with EPR and BYM/INLA are comparable in terms of CV and CRPS. In general, BYM/INLA and the discrepancy model implemented with EPR produce more similar estimates of the regression coefficients (MSE between these two estimated regression coefficients is 0.2523), and BYM/MCMC and BYM/INLA produce dissimilar regression estimates (MSE between these two estimated regression coefficients is

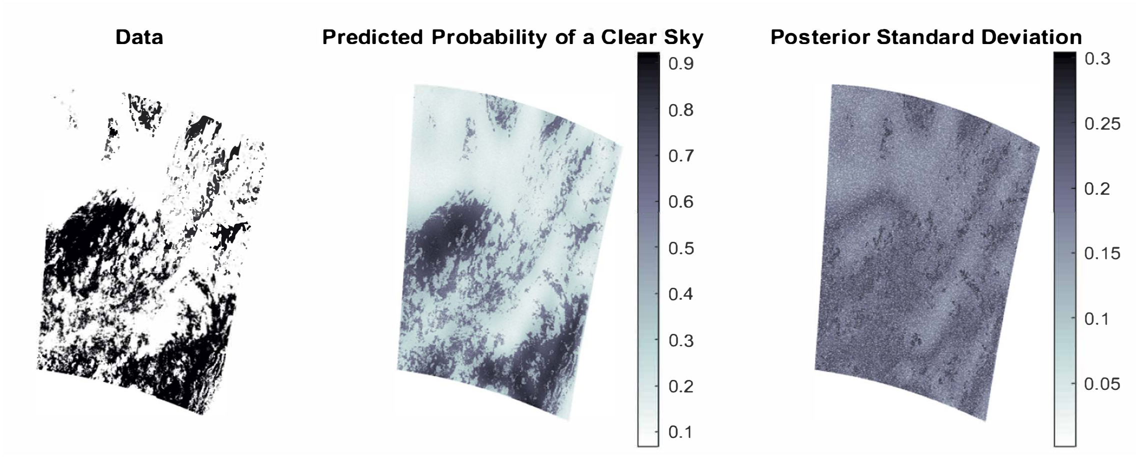
Data
Predicted Probability of a Clear Sky
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Posterior Standard Deviation
0.3
0.25
0.2
0.15
0.1
0.05

Figure 5: In the left panel, we plot MODIS cloud mask data from Bradley et al. (2020a) (white-value indicates a cloud, and black value indicates clear sky), in the middle panel we plot the predicted probability of a clear sky, and in the right panel we plot the posterior standard deviation.

1.4634). Computationally, EPR is preferable in terms of CPU time followed closely by INLA. MCMC took considerably longer to implement the leave-one-out analysis.

4.4 Benchmark Binary MODIS Cloud Mask Data

In this section, we consider the benchmark cloud mask dataset recorded via the Moderate Resolution Imaging Spectroradiometer (MODIS) onboard the National Aeronautics and Space Administration's (NASA) Terra satellite (Sengupta et al., 2016; Bradley et al., 2020a; Lee and Haran, 2022). The MODIS instrument transforms spectral radiance measurements into a level-2 (i.e.,1 km by 1 km grid) cloud mask (i.e., binary values) through cloud-detection algorithms. This is a fairly high-dimensional dataset with $2,748,620$ observations. We define the discrepancy model with the same exact (low-dimensional) basis functions and covariates used in Bradley et al. (2020a) to illustrate dimension reduction for EPR. We also use the same prior specifications as in Section 4.1.

We holdout 5% of the data to estimate a threshold on EPR predicted probability to clas-

Method	False Positive	False Negative	CPU
EPR	0.21	0.27	10.4 hours
MCMC	0.22	0.28	16 hours
SVM	0.11	0.53	3 days

Table 4: The false positive rates and false negative rates are based on the 5% testing data. To classify a predicted clear sky as cloud or no cloud we use another 5% of the data as a validation dataset to determine a threshold on the predicted probability, where the threshold is chosen to minimize the false positive and false negative rate in the validation dataset. The values for MCMC and SVM were computed in Bradley et al. (2020a). The final column presents approximate CPU times.

sify the presence of a cloud, and hold out another 5% testing data for cross-validation so that $n = 2,473,758$. A plot of the data, EPR-based predicted probability of clear sky, and prediction standard deviations are displayed in Figure 5. It is clear from the plot that the predicted probability of clear sky follows the patterns of the observed dataset. That is, when the response is a 1 (indicated with black) the predicted probability is larger and vice versa. In Table 4, we compare false positive and false negative rates of EPR to that computed using the latent CM (or LCM) model (Bradley et al., 2020a) implemented with MCMC and to support vector machines (SVM, Hastie et al., 2009). The discrepancy model implemented with EPR not only produces smaller false positive and false negative rates than LCM/MCMC, the CPU time is faster on the order of hours. SVM has smaller false positive rates, but considerably higher false negative rates, and is considerably more time intensive than both the discrepancy model implemented with EPR and LCM/MCMC.

5 Discussion

This paper describes how to efficiently sample independent replicates directly from the posterior distribution of fixed and random effects using a broad class of spatial latent Gaussian process models. This development required the introduction of the GCM distribution and the conditional GCM distribution. The use of the GCM allows one to consider any class of CM's for their prior distributions on fixed and random effects. Our development explicitly addresses hyperparameters through marginalization. We make use of the GCM in settings where one traditionally would use a LGP to produce what we call EPR, which represents an efficiently generated independent sample from the posterior distribution. We show that the posterior distribution for fixed and random effects for our hierarchical model (referred to as the discrepancy model) are GCM, which we can directly sample from without approximations and without MCMC. Furthermore, we use matrix algebra techniques and dimension reduction to aid in the computation of EPR in high-dimensional settings. The performance of our model is particularly exciting in the misspecified case (i.e., the case when the discrepancy parameter $\mathbf{q} = \mathbf{0}_{n,1}$). The misspecified discrepancy model produces larger more conservative credible intervals, while still being able to produce comparable predictions and regression point estimates to those of the correctly specified traditional LGP model.

The results in this paper solve an important problem for Bayesian analysis that is regularly overlooked (i.e., obtaining efficient independent replicates directly from the posterior distribution in Bayesian spatial hierarchical models). One might also consider empirical Bayesian variations of EPR as our solution also allows one to sample independent replicates directly from the posterior predictive distribution when using point mass specification of π (i.e.,point mass on an estimate), which avoids MCMC in empirical Bayesian settings as well. Specifically, Theorems 3.1 $-$ 3.3 can be used with a plug-in estimator of $\boldsymbol{\theta}$. However, plug-in estimators have unchecked sampling variability (provided that the plug-in estimator is a non-constant function of the data), and the development of the GCM provides a straightforward solution that accounts for all sources of variability.

A key area of future development is the use of model selection criteria for the discrepancy model. For example, the Akaike information criterion (AIC, Akaike, 1974), Bayesian information criterion (BIC, Schwarz, 1978), and the likelihood ratio (e.g., see Casella and Berger, 2002, for a standard reference) all require methodological/computational development for use on the discrepancy model since the expression of the marginal likelihood $f(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\theta})$ contains integrals (across $\mathbf{q}$, $\boldsymbol{\eta}$, and $\boldsymbol{\xi}$) that are not known. Other related criteria such as the conditional Akaike information criterion (Vaida and Blanchard, 2005) similarly requires an expression of the marginal data model (that integrates out $\mathbf{q}$) $f(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\xi})$ that does not have a closed form. The nested relationship between the discrepancy model and the traditional LGP (see Supplementary Appendix E) may aid with deriving these expressions. In our applications, we used other traditional criteria to evaluate the discrepancy model that do not require approximating an integral. In particular, we used the leave-one-out cross-validation and cross-validation to assess predictive performance (Hastie et al., 2009) and the continuous rank probability score (CRPS) to assess both prediction and uncertainty quantification (Gneiting and Raftery, 2007). In our real data analyses these metrics were all similar in value between the discrepancy model and a comparable hierarchical model that did not include a discrepancy parameter, and the discrepancy model was more computationally efficient in terms of CPU time. In general, one should keep in mind statistical accuracy and computation convenience when choosing a model.

While we feel that the results in this manuscript represent a significant advancement in Bayesian modeling of spatial data, it is important to state that MCMC and INLA will always be standard tools. In particular, in this paper, EPR has only been developed in the context of settings where generalized linear mixed models are appropriate, where the data are conditionally Gaussian, binomial, and Poisson. This implies that finite mixture models, Dirichlet process models, zero-inflated models, extreme-value models, and other settings, currently, can not be implemented using EPR. Moreover, inference using EPR is limited to summaries of regression coefficients and the latent process, since $\boldsymbol{\theta}$ is marginalized. Additionally, EPR is not computationally feasible when both n is large and the basis function matrices are dense and full rank. Thus, when n is large we make use of dimension reduction to perform inference and we include the MODIS cloud-mask illustration of this case. This opens up the possibility for further methodological/computational development of EPR in these settings. However, we hope the theory developed in this article leads to further theoretical developments that allows one to sample independent replicates from the posterior distribution in other settings.

Supplemental Materials

Appendices: This document includes the necessary reviews, notation tables, proofs of technical results, example model specifications, notations, additional details on the fine-scale variance parameter, additional simulations, and prior specifications used in the illustrations. (supplement.pdf)

R Code: All R code use in the illustrations. Please read the file "Read Me" found in the zip file for more details. Several of the files are quite large, and for access to these files please see the file "LargeCodeFiles" contained in the zip file for more details. (Code.zip)

Acknowledgments

Jonathan R. Bradley's research was partially supported by the U.S. National Science Foundation (NSF) under NSF grant DMS-2310756. The authors are deeply appreciative to several helpful discussions from Drs. Scott H. Holan and Christopher K. Wikle at the University of Missouri. We are also deeply appreciative of the feedback from the editor, associate editor, and two anonymous referees.

Disclosure Statement

No potential competing interest was reported by the authors.

References

Akaike, H. (1974). “A new look at the statistical model identification.” *IEEE transactions on automatic control*, 19, 6, 716–723.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*. London, UK: Chapman and Hall.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). “Gaussian predictive process models for large spatial data sets.” *Journal of the Royal Statistical Society, Series B*, 70, 825–848.

Besag, J. E., York, J. C., and Molliè, A. (1991). “Bayesian image restoration, with two applications in spatial statistics (with discussion).” *Annals of the Institute of Statistical Mathematics*, 43, 1–59.

Bradley, J., Holan, S., and Wikle, C. (2015). “Multivariate spatio-temporal models for high-dimensional areal data with application to Longitudinal Employer-Household Dynamics.” *The Annals of Applied Statistics*, 9, 1761–1791.

— (2018). “Computationally Efficient Distribution Theory for Bayesian Inference of High-Dimensional Dependent Count-Valued Data.” *Bayesian Analysis*, 13, 253–302.

Bradley, J., Wikle, C., and Holan, S. (2017). “Regionalization of multiscale spatial processes using a criterion for spatial aggregation error.” *Journal of the Royal Statistical Society: Series B*, 79, 815–832.

Bradley, J. R. (2021). “An Approach to Incorporate Subsampling Into a Generic Bayesian Hierarchical Model.” *Journal of Computational and Graphical Statistics*, 30, 4, 889–905.

— (2022). "Joint Bayesian Analysis of Multiple Response-Types Using the Hierarchical Generalized Transformation Model." *Bayesian Analysis*, 17, 127–164.

Bradley, J. R., Cressie, N., and Shi, T. (2016). "A comparison of spatial predictors when datasets could be very large." *Statistics Surveys*, 10, 100–131.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2020a). "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.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2019). "Spatio-temporal models for big multinomial data using the conditional multivariate logit-beta distribution." *Journal of Time Series Analysis*, 40, 3, 363–382.

— (2020b). "Hierarchical Models for Spatial Data with Errors that are Correlated with the Latent Process." *Statistica Sinica*, 30, 80–109.

Bradley, J. R., Zhou, S., and Liu, X. (2023). "Deep hierarchical generalized transformation models for spatio-temporal data with discrepancy errors." *Spatial Statistics*, 100749.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A probabilistic programming language." *Journal of Statistical Software*, 76, 1, 1–32.

Casella, G. and Berger, R. (2002). *Statistical Inference*. Pacific Grove, CA: Duxbury.

Chen, M. H. and Ibrahim, J. G. (2003). "Conjugate priors for generalized linear models." *Statistica Sinica*, 13, 2, 461–476.

Cowles, M. K. and Carlin, B. P. (1996). "Markov chain Monte Carlo convergence diagnostics: a comparative review." *Journal of the American Statistical Association*, 91, 434, 883–904.

Cressie, N. (1993). *Statistics for Spatial Data,* rev. edn. New York, NY: Wiley.

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

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets." *Journal of the American Statistical Association*, 111, 514, 800–812.

Diaconis, P. and Ylvisaker, D. (1979). "Conjugate priors for exponential families." *The Annals of Statistics*, 17, 269–281.

Dunson, D. B. and Neelon, B. (2003). "Bayesian inference on order-constrained parameters in generalized linear models." *Biometrics*, 59, 2, 286–295.

D'Angelo, L. and Canale, A. (2022). "Efficient posterior sampling for Bayesian Poisson regression." *Journal of Computational and Graphical Statistics*, 1–10.

Finley, A. O., Banerjee, S., and Carlin, B. (2012). "Package 'spBayes'." http://cran.r-project.org/web/packages/spBayes/spBayes.pdf. Retrieved January, 2013.

Gao, H. and Bradley, J. R. (2019). "Bayesian analysis of areal data with unknown adjacencies using the stochastic edge mixed effects model." *Spatial Statistics*, 31, 100357.

Gelfand, A. E. and Schliep, E. M. (2016). "Spatial statistics and Gaussian processes: a beautiful marriage." *Spatial Statistics*, 18, 86–104.

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

Gelman, A. and Rubin, D. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7, 473–511.

Gneiting, T. and Katzfuss, M. (2014). "Probabilistic forecasting." *Annual Review of Statistics and Its Application*, 1, 125–151.

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102, 477, 359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. (2005). "Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation." *Monthly Weather Review*, 133, 5, 1098–1118.

H.-C.Yang, Hu, G., and Chen, M.-H. (2019). "Bayesian Variable Selection for Pareto Regression Models with Latent Multivariate Log Gamma Process with Applications to Earthquake Magnitudes." *Geosciences*, 9, 4, 169.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Hodges, J. S. (2013). *Richly parameterized linear models: additive, time series, and spatial models using random effects*. CRC Press.

Hu, G. and Bradley, J. R. (2018). "A Bayesian spatial–temporal model with latent multivariate log-gamma random effects with application to earthquake magnitudes." *Stat*, 7, 1, e179.

Huang, H. and Cressie, N. (1999). "Empirical Bayesian Spatial Prediction Using Wavelets." *Bayesian Inference in Wavelet Based Models, eds P. Mueller and B. Vidakovich.*, , 141.

Hughes, J. and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed model." *Journal of the Royal Statistical Society, Series B*, 75, 139–159.

Jeffreys, H. (1946). "An invariant form for the prior probability in estimation problems." *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186, 1007, 453–461.

Kang, H. B., Jung, Y. J., and Park, J. (2023). "Fast Bayesian Functional Regression for Non-Gaussian Spatial Data." *Bayesian Analysis*, 1, 1, 1–32.

Konomi, B. A., Kang, E. L., Almomani, A., and Hobbs, J. (2023). "Bayesian Latent Variable Co-kriging Model in Remote Sensing for Quality Flagged Observations." *Journal of Agricultural, Biological and Environmental Statistics*, 1–19.

Lee, B. S. and Haran, M. (2022). "PICAR: An efficient extendable approach for fitting hierarchical spatial models." *Technometrics*, 64, 2, 187–198.

Lee, D. (2013). "CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors." *Journal of Statistical Software*, 55, 13, 1–24.

Lindgren, F., Bolin, D., and Rue, H. (2022). "The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running." *Spatial Statistics*, 50, 100599.

Lindgren, F., Rue, H., and Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach." *Journal of the Royal Statistical Society, Series B*, 73, 423–498.

Lu, T.-T. and Shiou, S.-H. (2002). "Inverses of 2× 2 block matrices." *Computers & Mathematics with Applications*, 43, 1-2, 119–129.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. London, UK: Chapman and Hall.

Molina, I. and Rao, J. (2010). "Small area estimation of poverty indicators." *Canadian Journal of statistics*, 38, 3, 369–385.

Murphy, K. P. (2007). "Conjugate Bayesian analysis of the Gaussian distribution." *https://www.cs.ubc.ca/ murphyk/Papers/bayesGauss.pdf*.

Neal, R. M. (2011). "MCMC using Hamiltonian dynamics." *Handbook of Markov chain Monte Carlo*, 2, 11, 2.

Novikov, I. Y., Protasov, V. Y., and Skopina, M. A. (2005). *Wavelet Theory*. US: American Mathematical Society.

Parker, P. A., Holan, S. H., and Janicki, R. (2020). "Conjugate Bayesian unit-level modelling of count data under informative sampling designs." *Stat*, 9, 1, e267.

Parker, P. A., Holan, S. H., and Wills, S. A. (2021). "A general Bayesian model for heteroskedastic data with fully conjugate full-conditional distributions." *Journal of Statistical Computation and Simulation*, 91, 15, 3207–3227.

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya-Gamma latent variables." *Journal of the American Statistical Association*, 108, 1339–1349.

Porter, E. M., Franck, C. T., and Ferreira, M. A. (2023). "Objective Bayesian Model Selection for Spatial Hierarchical Models with Intrinsic Conditional Autoregressive Priors." *Bayesian Analysis*, 1, 1, 1–27.

Rao, J. K. (2003). *Small area estimation*. John Wiley & Sons.

Robert, C. and Casella, G. (2011). "A short history of MCMC: Subjective recollections from incomplete data." *Handbook of markov chain monte carlo*, 49.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York, NY: Springer.

Roberts, G. O. and Rosenthal, J. S. (2009). "Examples of adaptive MCMC." *Journal of computational and graphical statistics*, 18, 2, 349–367.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations." *Journal of the Royal Statistical Society, Series B*, 71, 319–392.

Schwarz, G. (1978). "Estimating the dimension of a model." *The Annals of Statistics*, 461–464.

Sengupta, A., Cressie, N., Kahn, B. H., and Frey, R. (2016). "Predictive inference for big, spatial, non-Gaussian data: MODIS cloud data and its change-of-support." *Australian & New Zealand Journal of Statistics*, 58, 1, 15–45.

Shirota, S., Finley, A. O., Cook, B. D., and Banerjee, S. (2023). "Conjugate sparse plus low rank models for efficient Bayesian interpolation of large spatial data." *Environmetrics*, 34, 1, e2748.

Theil, H. (1996). *Studies in global econometrics*. Springer Science & Business Media.

Torrieri, N. (2007). "America is changing, and so is the census: The American Community Survey." *American Statistician*, 61, 16–21.

Vaida, F. and Blanchard, S. (2005). "Conditional Akaike information for mixed-effects models." *Biometrika*, 92, 351 – 370.

van Erven, T. and Szabó, B. (2021). "Fast exact Bayesian inference for sparse signals in the normal sequence model." *Bayesian Analysis*, 16, 3, 933–960.

Vats, D., Flegal, J. M., and Jones, G. L. (2019). "Multivariate output analysis for Markov chain Monte Carlo." *Biometrika*, 106, 2, 321–337.

Vranckx, M., Faes, C., Molenberghs, G., Hens, N., Beutels, P., Van Damme, P., Aerts, J., Petrof, O., Pepermans, K., and Neyens, T. (2023). “A spatial model to jointly analyze self-reported survey data of COVID-19 symptoms and official COVID-19 incidence data.” *Biometrical Journal*, 65, 1, 2100186.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Wainwright, M. J. and Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

Wikle, C. K. (2010). “Low-rank representations for spatial processes.” In *Handbook of Spatial Statistics*, eds. A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, 107–118. Boca Raton, FL: Chapman & Hall/CRC Press.

Xu, Z., Bradley, J. R., and Sinha, D. (2023). “Latent multivariate log-gamma models for high-dimensional MultiType responses with application to daily fine particulate matter and mortality counts.” *The Annals of Applied Statistics*, 17, 2, 1175–1198.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). “Using stacking to average Bayesian predictive distributions (with discussion).” *Bayesian Analysis*, 13, 3, 917–1007.

Zhang, B., Sang, H., Luo, Z. T., and Huang, H. (2023a). “Bayesian clustering of spatial functional data with application to a human mobility study during COVID-19.” *The Annals of Applied Statistics*, 17, 1, 583–605.

Zhang, L., Banerjee, S., and Finley, A. O. (2021). “High-dimensional multivariate geostatistics: A Bayesian matrix-normal approach.” *Environmetrics*, 32, 4, e2675.

Zhang, L., Tang, W., and Banerjee, S. (2023b). “Exact Bayesian Geostatistics Using Predictive Stacking.” *arXiv preprint arXiv:2304.12414*.