Statistica Sinica Preprint No: SS-2016-0230	
Title	Hierarchical Models for Spatial Data with Errors that are Correlated with the Latent Process
Manuscript ID	SS-2016-0230
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0230
Complete List of Authors	Jonathan R. Bradley Christopher K. Wikle and Scott H. Holan
Corresponding Author	Jonathan R. Bradley
E-mail	bradley@stat.fsu.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Hierarchical Models for Spatial Data with Errors that are Correlated with the Latent Process

Jonathan R. Bradley[1], Christopher K. Wikle[2], Scott H. Holan[23]

Abstract

Prediction of a spatial process using a "big dataset" has become a topical area of research over the last decade. The available solutions often involve placing strong assumptions on the error process associated with the data. Specifically, it has typically been assumed that the data are equal to the spatial process of principal interest plus a mutually independent error process. This is done to avoid modeling confounded cross-covariances between the signal and noise within an additive model. In this article, we consider an alternative latent process modeling schematic where it is assumed that the error process is spatially correlated *and* correlated with the latent process of interest. We show that such error process dependencies allow one to obtain precise predictions, and avoids confounded error covariances within the expression of the marginal distribution of the data. We refer to these covariances as "non-confounded discrepancy error covariances." Additionally, a "process augmentation" technique is developed to aid in computation. Demonstrations are provided through simulated examples and through an application using a large dataset consisting of the U.S. Census Bureau's American Community Survey 5-year period estimates of median household income on census tracts.

Keywords: Bayesian, Low rank, Machine learning, Mixed effects model, Nonresponse, Parsimony.

[1](to whom correspondence should be addressed) Department of Statistics, Florida State University, bradley@stat.fsu.edu

[2]Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211-6100

[3]U.S. Census Bureau, 4600 Silver Hill Road, Washington D. C., 20233-9100

1 Introduction

In this article, we introduce a general class of additive spatial models, where it is assumed that the error process is spatially correlated *and* correlated with the latent process of interest. Adopting these discrepancy error covariances could have an important impact in a variety of applications. For example, in federal/official statistics, assuming independent "survey error" is standard, and is a key component of the ubiquitous Fay-Herriot model (Fay and Herriot, 1979). However, in many settings, it is well-known that these errors are dependent. For example, disseminated estimates are sometimes modified/suppressed based on the value of the latent process due to disclosure limitations, which may induce correlations between the survey error and the latent process (e.g., see Quick et al. (2015), for a review in the spatial setting). The error induced by nonresponse may be due to the value of the latent process, and consequently, we expect cross-correlations between the error process and the latent process (Groves et al., 2001). Also, sampling designs are often motivated by the spatial structure of the latent process, which may result in correlations between the latent process and the survey error (Wikle and Royle, 2005; Holan and Wikle, 2012). Thus, we are partially motivated by investigating whether or not these dependencies exist among the increasingly popular American Community Survey (ACS) period estimates (e.g., see Torrieri, 2007).

Although it is reasonable to expect the presence of such discrepancy error covariances, these covariances are often ignored. For example, standard spatial statistics textbooks focus almost exclusively on the case of a spatially correlated error term that is independent of a mutually independent error term (Cressie, 1993; Cressie and Wikle, 2011; Banerjee et al., 2015). This is partially because there is a problem of confounding between the covariances of the signal and covariances of the noise (Cressie, 1993). Henceforeth, we use the terms "signal" and "latent process" interchangeably. Confounding between *fixed effects* and *spatial random effects* has become an important topic covered in the spatial statistics literature (e.g.,

Clayton et al., 1993; Reich et al., 2006; Hodges and Reich, 2011). More recently, spatial basis functions (i.e., the "Moran's I basis functions") have been developed to account for confounding the spatial setting (Griffith, 2000, 2002, 2004; Hughes and Haran, 2013; Bradley et al., 2015a), which have a connection to the classical Moran's I statistic (Moran, 1950). However, to our knowledge, there has not been any work that addresses confounding between covariances of the signal and covariances of the discrepancy error.

Thus, the goal of this paper is to provide a way to leverage error process dependencies in a manner that is computationally feasible and accounts for confounding in the marginal covariance matrix associated with the data. We achieve this goal by introducing non-negligible discrepancy error covariances that are not present in the marginal distribution of the data. We refer to this class of measurement-error covariances matrices as *non-confounded discrepancy error covariances*. We give a very general form of the non-confounded discrepancy error covariance matrices, and provide several parameterizations to use in practice. In particular, we show that the standard uncorrelated discrepancy error assumption is a special case of the non-confounded discrepancy error covariances. Then, we provide parameterizations that represent slight departures from the standard assumption of uncorrelated discrepancy errors. To aid researchers in assessing the appropriateness of these assumptions, we develop a covariance penalized error (Efron, 2004) as a measure of out-of-sample error.

To date, there are no competing spatial methodologies that capitalize on correlations between the error and the latent process in such a computationally efficient manner. These dependencies lead to improvements in predictions of the latent process. However, the general approach to leverage dependence between an error process and a latent process has been exploited in other settings outside spatial statistics. For example, in time-series there is a methodological approach referred to as "leverage effects" used within stochastic volatility models (e.g., see Black, 1976, for an early reference), where the volatility is assumed to be correlated with the latent process. There are also similar relationships with "feedback

models" (Zeger and Liang, 1991).

To aid in computation we consider using a type of data augmentation approach (e.g., see Tanner and Wong, 1987; Albert and Chib, 1993; Wakefield and Walker, 1999; Wolpert and Ickstadt, 1998, among others). In our implementation we augment the process and not the data, and hence, we refer to this strategy as "process augmentation." The implementation of our process augmentation approach involves two steps. The first step of our procedure involves fitting any well-defined Bayesian spatial model (e.g., see Banerjee et al., 2008; Cressie and Johannesson, 2008; Lindgren et al., 2011; Datta et al., 2014; Nychka et al., 2015; Katzfuss, 2017). The second step involves posterior predictive simulation. Thus, our proposed model can be seen as a diagnostic tool; that is, after one fits a Bayesian statistical model, the second step of our method can be performed to check whether improvements in prediction are made through the incorporation of discrepancy error covariances. Additionally, this two step procedure shows that estimation of the covariances associated with the data, and regression parameters are unaffected by incorporating non-confounded discrepancy error covariances.

The remainder of the paper is organized as follows. In Section 2, we introduce the non-confounded discrepancy error covariances, and describe several special cases. Additionally, we provide the kriging predictor (Cressie, 1993) to aid in the interpretation of the non-confounded discrepancy error covariances. Then, in Section 3 we describe how to address out-of-sample performance and robustness to departures from model assumptions. Next, in Section 4, we describe implementation using a process augmentation approach. In Section 5, we use simulation studies to illustrate the high predictive performance of our method, and we demonstrate our method using a large dataset consisting of ACS estimates defined on census tracts. We end with a discussion in Section 6. For convenience of exposition, proofs of technical results, model selection, and model fitting are provided in a Supplemental Appendix.

2 Methodology

We start our exposition with the motivating difficulty of incorporating discrepancy error covariances; namely, confounded cross-covariances. Then, we introduce the non-confounded discrepancy error covariance matrix (Section 2.1). Next, we provide several special cases and properties of the non-confounded discrepancy error covariance matrix in Sections 2.2 − 2.5. To aid the interpretation of the non-confounded discrepancy error covariance matrix we discuss these special cases in context of kriging in Section 2.6.

2.1 Non-Confounded Discrepancy Error Covariances

Suppose we observed data at a finite number of locations denoted with $\mathbf{s}_1, \ldots, \mathbf{s}_m \in D \subset \mathbb{R}^d$, where D represents the spatial domain of interest. The observed data are then denoted by $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_m)\}$, and in general the data are realized through a spatial process, and $Z(\mathbf{s})$ is defined at unobserved locations $\mathbf{s} \notin \{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$. Suppose we are interested in predicting at the set of locations $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$, which is not necessarily equal to the set $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$. We follow the hierarchical modeling approach common in the spatial statistics literature (e.g., see standard textbooks, Cressie and Wikle, 2011; Banerjee et al., 2015) and assume $Z(\mathbf{s}) = Y(\mathbf{s}) + \delta(\mathbf{s});\ \mathbf{s} \in D$, where $Y(\cdot)$ represents the latent process of interest, and the data process represents a corrupted version of the latent process, where the corruption is represented additively with the error process $\delta(\cdot)$. The spatial Gaussian process modeling literature assumes that both $Y(\mathbf{s})$ and $\delta(\mathbf{s})$ are Gaussian for any $\mathbf{s} \in D$.

Let $\mu(\cdot)$ be the mean function for both $Y(\cdot)$ and $Z(\cdot)$, and organize the vectors $\mathbf{z} = \{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_m)\}^\top$, $\mathbf{y} = \{Y(\mathbf{u}_1), \ldots, Y(\mathbf{u}_n)\}^\top$, $\boldsymbol{\delta} \equiv \{\delta(\mathbf{u}_1), \ldots, \delta(\mathbf{u}_n)\}^\top$, and $\boldsymbol{\mu} \equiv \{\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_m)\}^\top$. Additionally, define the $n \times n$ matrices $\boldsymbol{\Sigma}_Y = \text{cov}(\mathbf{y})$, $\boldsymbol{\Sigma}_\delta = \text{cov}(\boldsymbol{\delta})$, and

$\boldsymbol{\Sigma}_{Y,\delta} = \text{cov}(\mathbf{y}, \boldsymbol{\delta})$. Then the probability density function of $\mathbf{z}$ is,

$$\begin{aligned} &f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y = \mathbf{C}_Y, \boldsymbol{\Sigma}_\delta = \mathbf{C}_\delta, \boldsymbol{\Sigma}_{Y,\delta} = \mathbf{C}_{Y,\delta}) \\ &\propto |\mathbf{C}_Y + \mathbf{C}_\delta + 2\mathbf{C}_{Y,\delta}|^{-1/2} \text{exp}\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^\top (\mathbf{C}_Y + \mathbf{C}_\delta + 2\mathbf{C}_{Y,\delta})^{-1} (\mathbf{z}-\boldsymbol{\mu})\right\}, \end{aligned} \tag{1}$$

where $\mathbf{C}_Y$, $\mathbf{C}_\delta$, and $\mathbf{C}_{Y,\delta}$ are distinct values in the parameter space of $\boldsymbol{\Sigma}_Y$, $\boldsymbol{\Sigma}_\delta$, and $\boldsymbol{\Sigma}_{Y,\delta}$, respectively. Notice that for discussion we have set $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\} = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$, but this is not always true, and the method generalizes easily to this case. The confounding problem is immediately apparent in (1). The commutative property of matrices gives us, for example,

$$\begin{aligned} f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y = \mathbf{C}_Y, \boldsymbol{\Sigma}_\delta = \mathbf{C}_\delta, \boldsymbol{\Sigma}_{Y,\delta} = \mathbf{C}_{Y,\delta}) &= f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y = \mathbf{C}_\delta, \boldsymbol{\Sigma}_\delta = \mathbf{C}_Y, \boldsymbol{\Sigma}_{Y,\delta} = \mathbf{C}_{Y,\delta}) \\ &= f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y = \mathbf{C}_Y, \boldsymbol{\Sigma}_\delta = 2\mathbf{C}_{Y,\delta}, \boldsymbol{\Sigma}_{Y,\delta} = \frac{1}{2}\mathbf{C}_\delta). \end{aligned}$$

To mitigate these confounding issues it is often assumed that $\boldsymbol{\Sigma}_{Y,\delta}$ is an $n \times n$ matrix of zeros (denoted with $\mathbf{0}_{n,n}$) and $\boldsymbol{\Sigma}_\delta = \sigma^2\mathbf{I}_n$, where $\sigma^2 > 0$ and $\mathbf{I}_n$ is an $n \times n$ identity matrix. (e.g., see Banerjee et al., 2008; Cressie and Johannesson, 2008; Finley et al., 2009; Lindgren et al., 2011; Sang and Huang, 2012; Nychka et al., 2015, among others). This assumption gives

$$\begin{aligned} &f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma}_w, \boldsymbol{\Sigma}_\delta = \sigma^2\mathbf{I}_n, \boldsymbol{\Sigma}_{Y,\delta} = \mathbf{0}_{n,n}) \\ &\propto |\boldsymbol{\Sigma}_w + \sigma^2\mathbf{I}_n|^{-1/2} \text{exp}\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^\top \left(\boldsymbol{\Sigma}_w + \sigma^2\mathbf{I}_n\right)^{-1} (\mathbf{z}-\boldsymbol{\mu})\right\}, \end{aligned} \tag{2}$$

where $\boldsymbol{\Sigma}_w$ is a generic positive semi-definite matrix in the parameter space of $\boldsymbol{\Sigma}_Y$. Here, confounded cross-covariances are not present in the likelihood for $\mathbf{z}$; however, we are no longer able to capitalize on the dependence between $Y(\cdot)$ and $\delta(\cdot)$ and covariances between $\delta(\cdot)$ at different locations. In this article, we introduce a non-zero structure to the cross-covariance parameter $\boldsymbol{\Sigma}_{Y,\delta}$ *and* we introduce a compatible (possibly non-diagonal) $\boldsymbol{\Sigma}_\delta$ so that we obtain the likelihood in (2). This leads to what we call the "General Assumption."

General Assumption: Let $\mathbf{\Sigma}_\delta = \mathbf{\Sigma}_Y + \mathbf{\Sigma}_w - 2\mathbf{\Sigma}_{Y,w} + \sigma^2\mathbf{I}_n$ and $\mathbf{\Sigma}_{Y,\delta} = \mathbf{\Sigma}_{Y,w} - \mathbf{\Sigma}_Y$, where $\mathbf{\Sigma}_w$ is a positive semi-definite matrix, $\mathbf{\Sigma}_{Y,w}$ is an $n \times n$ real matrix, and

$$\text{cov}\left\{ \begin{pmatrix} \mathbf{y} \\ \boldsymbol{\delta} \end{pmatrix} | \mathbf{\Sigma}_Y, \mathbf{\Sigma}_w, \mathbf{\Sigma}_{Y,w}, \sigma^2 \right\} = \begin{pmatrix} \mathbf{\Sigma}_Y & \mathbf{\Sigma}_{Y,w} - \mathbf{\Sigma}_Y \\ \mathbf{\Sigma}_{Y,w}^\top - \mathbf{\Sigma}_Y^\top & \mathbf{\Sigma}_Y + \mathbf{\Sigma}_w - 2\mathbf{\Sigma}_{Y,w} + \sigma^2\mathbf{I}_n \end{pmatrix}, \quad (3)$$

is positive semi-definite.

Now, substituting the General Assumption into (1), we obtain

$$\begin{aligned} & f(\mathbf{z}|\boldsymbol{\mu}, \mathbf{\Sigma}_Y, \mathbf{\Sigma}_\delta = \mathbf{\Sigma}_Y + \mathbf{\Sigma}_w - 2\mathbf{\Sigma}_{Y,w} + \sigma^2\mathbf{I}_n, \mathbf{\Sigma}_{Y,\delta} = \mathbf{\Sigma}_{Y,w} - \mathbf{\Sigma}_Y) \\ & \propto |\mathbf{\Sigma}_w + \sigma^2\mathbf{I}_n|^{-1/2} \text{exp}\left\{ -\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^\top \left(\mathbf{\Sigma}_w + \sigma^2\mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right\}, \end{aligned} \quad (4)$$

since,

$$\begin{aligned} & \text{cov}(\mathbf{z}|\mathbf{\Sigma}_Y, \mathbf{\Sigma}_w, \mathbf{\Sigma}_{Y,w}, \sigma^2) \\ & = \text{cov}(\mathbf{y}|\mathbf{\Sigma}_Y, \mathbf{\Sigma}_w, \mathbf{\Sigma}_{Y,w}, \sigma^2) + \text{cov}(\boldsymbol{\delta}|\mathbf{\Sigma}_Y, \mathbf{\Sigma}_w, \mathbf{\Sigma}_{Y,w}, \sigma^2) + 2\text{cov}(\mathbf{y}, \boldsymbol{\delta}|\mathbf{\Sigma}_Y, \mathbf{\Sigma}_w, \mathbf{\Sigma}_{Y,w}, \sigma^2) \\ & = \mathbf{\Sigma}_Y + \mathbf{\Sigma}_Y + \mathbf{\Sigma}_w - 2\mathbf{\Sigma}_{Y,w} + \sigma^2\mathbf{I} + 2\mathbf{\Sigma}_{Y,w} - 2\mathbf{\Sigma}_Y \\ & = \mathbf{\Sigma}_w + \sigma^2\mathbf{I}_n. \end{aligned}$$

The likelihood in (4) is of the same form as the likelihood in Equation (2), which did not have confounded cross-covariances! The difference in our approach is that we *also* have correlations between $Y(\cdot)$ and $\delta(\cdot)$. That is,

$$\text{cov}\left(\mathbf{y}, \boldsymbol{\delta}\right) = \mathbf{\Sigma}_{Y,w} - \mathbf{\Sigma}_Y.$$

Thus, the General Assumption gives us a Gaussian likelihood in (2) with unconfounded cross-covariances, while simultaneously allowing for cross-correlations between the latent process

and the error process.

A special case of the General Assumption is the more conventionally used uncorrelated signal and noise with no cross-spatial dependence among the discrepancy errors. Specifically, let $\mathbf{\Sigma}_w = \mathbf{\Sigma}_{Y,w} = \mathbf{\Sigma}_Y$ so that,

$$\text{cov}\left\{ \begin{pmatrix} \mathbf{y} \\ \boldsymbol{\delta} \end{pmatrix} | \mathbf{\Sigma}_Y, \mathbf{\Sigma}_w, \mathbf{\Sigma}_{Y,w}, \sigma^2 \right\} = \begin{pmatrix} \mathbf{\Sigma}_Y & \mathbf{0}_{n,n} \\ \mathbf{0}_{n,n} & \sigma^2 \mathbf{I}_n \end{pmatrix}, \tag{5}$$

which is positive definite. We shall henceforth refer to $\mathbf{\Sigma}_w = \mathbf{\Sigma}_{Y,w} = \mathbf{\Sigma}_Y$ as the "Standard Assumption." This helps provide interpretations of these new matrix valued parameters $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_{Y,w}$. That is, as the difference between $\mathbf{\Sigma}_{Y,w}$ and $\mathbf{\Sigma}_Y$ increases, we obtain larger cross dependence between the signal $Y(\cdot)$ and the noise $\delta(\cdot)$. The moment properties of this model also provide interpretation of each matrix valued parameter.

Since we are primarily interested in developing non-confounded additive error covariances, for illustration we use standard "off-the-shelf" covariance functions to define $\mathbf{\Sigma}_Y$ and $\mathbf{\Sigma}_w$, respectively. In particular, when $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are continuous we use the Matérn covariogram (Matérn, 1960). When $\mathbf{s}_1, \ldots, \mathbf{s}_n$ defines a lattice, we use a conditional autoregressive model (Besag, 1974). In general, our framework can be implemented using any well-defined covariance function.

2.2 Moment Properties of the General Assumption

This section covers basis moment results, which lead to illuminating interpretations of $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_{Y,w}$ defined in the General Assumption.

Proposition 1 (Moment Properties): Let the data vector $\boldsymbol{z}$ have probability density function according to (1). Suppose the General Assumption from Section 2.1 holds. Then we have the following moment results:

a. $\text{cov}(\boldsymbol{z}|\boldsymbol{\Sigma}_w, \sigma^2) = \boldsymbol{\Sigma}_w + \sigma^2 \boldsymbol{I}_n$;

b. $E(\boldsymbol{z}|\boldsymbol{\Sigma}_w, \sigma^2) = \boldsymbol{\mu}$;

c. $\text{cov}(\boldsymbol{z}, \boldsymbol{y}|\boldsymbol{\Sigma}_w, \sigma^2) = \boldsymbol{\Sigma}_{Y,w}^\top$;

d. $\text{cov}(\boldsymbol{z}|\mathbf{y}, \boldsymbol{\Sigma}_w, \sigma^2) = \boldsymbol{\Sigma}_w + \sigma^2 \boldsymbol{I}_n - \boldsymbol{\Sigma}_{Y,w}^\top \boldsymbol{\Sigma}_Y^{-1} \boldsymbol{\Sigma}_{Y,w}$;

e. $E(\boldsymbol{z}|\mathbf{y}, \boldsymbol{\Sigma}_w, \sigma^2) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_{Y,w}^\top \boldsymbol{\Sigma}_Y^{-1}(\mathbf{y} - \boldsymbol{\mu})$.

Proof: Follows immediately from the General Assumption and rules for conditional and marginal distributions of Gaussian random vectors (Ravishanker and Dey, 2002).

Proposition 1(a,b) are the motivating features discussed in Section 2.1 (i.e., the marginal density of $\mathbf{z}$ does not contain confounded cross-correlations). However, it also shows that the off-diagonals of $\boldsymbol{\Sigma}_w$ represent the cross-spatial correlations of the data, and σ^2 represents extra variability not accounted for in $\boldsymbol{\Sigma}_w$; this has a similar role as the nugget in classical spatial statistics (see Cressie, 1993; Cressie and Wikle, 2011; Banerjee et al., 2015, for standard references). Additionally, Proposition 1(c) shows that $\boldsymbol{\Sigma}_{Y,w}^\top$ represents the cross-covariance between $\mathbf{z}$ and $\mathbf{y}$.

Proposition 1(d,e) implies that the data are *not* conditionally unbiased and *not* conditionally uncorrelated given the latent process $Y(\cdot)$. In the ACS example (Section 5.3), $Z(\cdot)$ represents the *disseminated* (log transform) median income of individuals in a particular census tract, while $Y(\cdot)$ represents the actual (log transform) median income of individuals in the census tract. The difference between $Y(\cdot)$ and $Z(\cdot)$ may be due to the culmination of the sampling design, nonresponse bias, modifications due to disclosure avoidance concerns, and many other sources of error. Thus, it may be reasonable to assume that the disseminated ACS data has some bias and/or unaccounted for covariability among the survey errors.

2.3 Special Case 1

In this section, we consider a slight departure from the Standard Assumption that $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$. Specifically, we assume $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_Y$. In this case, the expression of (3) is given by,

$$\text{cov}\left\{\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\delta}\end{pmatrix}\right\} = \begin{pmatrix}\boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{Y,w} - \boldsymbol{\Sigma}_Y \\ \boldsymbol{\Sigma}_{Y,w}^\top - \boldsymbol{\Sigma}_Y^\top & \boldsymbol{\Sigma}_Y + \boldsymbol{\Sigma}_w - 2\boldsymbol{\Sigma}_{Y,w} + \sigma^2\mathbf{I}_n\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_Y & -\boldsymbol{\Sigma}_1 \\ -\boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_1 + \sigma^2\mathbf{I}_n\end{pmatrix}, \tag{6}$$

where $\boldsymbol{\Sigma}_1 \equiv \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_w$. In Supplemental Appendix A, we show that (6) is positive definite provided that $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_1$ are positive definite. We refer to (6) as "Special Case 1." Here, we see an indirect relationship between the signal-to-noise cross-covariance and the marginal covariance of the discrepancy error. Additionally, Special Case 1 leads to conditionally unbiased data (substitute $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_Y$ into Proposition 1(e)). Some might consider Special Case 1 to be more realistic in the official statistics setting. As discussed in the Introduction, federal agencies go through extensive work to produce highly accurate estimates; however, there is no guarantee that discrepancy error correlations are not present among these estimates.

2.4 Special Case 2

Consider the assumption that $\boldsymbol{\Sigma}_w \neq \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$. In this case, the expression of (3) is given by,

$$\text{cov}\left\{\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\delta}\end{pmatrix}\right\} = \begin{pmatrix}\boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{Y,w} - \boldsymbol{\Sigma}_Y \\ \boldsymbol{\Sigma}_{Y,w}^\top - \boldsymbol{\Sigma}_Y^\top & \boldsymbol{\Sigma}_Y + \boldsymbol{\Sigma}_w - 2\boldsymbol{\Sigma}_{Y,w} + \sigma^2\mathbf{I}_n\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_Y & \mathbf{0}_{n,n} \\ \mathbf{0}_{n,n} & \boldsymbol{\Sigma}_2 + \sigma^2\mathbf{I}_n\end{pmatrix}, \tag{7}$$

where $\boldsymbol{\Sigma}_2 \equiv \boldsymbol{\Sigma}_w - \boldsymbol{\Sigma}_Y$. It is immediate that (7) is positive definite provided that $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_Y$ are positive definite. We refer to the assumption that $\boldsymbol{\Sigma}_w \neq \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$ as "Special Case

2." In (7) we see that within discrepancy error covariances are present, but cross-covariances between the signal and the noise are not present. Additionally, Proposition 1(e) shows that Special Case 2 also implies that the data is conditionally unbiased for the latent process.

In general, any valid covariance function can be used to represent $\boldsymbol{\Sigma}_2$. For illustration, in the case where $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ consists of point-referenced locations we define the (i, j)-th element of $\boldsymbol{\Sigma}_2$ to be formed by the Matérn covariogram with correlation parameter $\tau > 0$ and variance parameter $\sigma_Y^2 > 0$ (Matérn, 1960). In the case where $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ consists of areal locations we can define $\boldsymbol{\Sigma}_2$ to be the covariance from a conditional autoregressive model with correlation parameter $\tau > 0$ and variance parameter $\sigma_Y^2 > 0$ (Besag, 1974). See Supplemental Appendix B for more details.

Special Case 2 is equivalent to the Standard Assumption after transformation. Specifically, let $\mathbf{z}^* = (\frac{1}{\sigma^2}\boldsymbol{\Sigma}_2 + \mathbf{I}_n)^{-1}\mathbf{z}$. In a similar manner define $\mathbf{y}^* = (\frac{1}{\sigma^2}\boldsymbol{\Sigma}_2 + \mathbf{I}_n)^{-1}\mathbf{y}$ and $\boldsymbol{\delta}^* = (\frac{1}{\sigma^2}\boldsymbol{\Sigma}_2 + \mathbf{I}_n)^{-1}\boldsymbol{\delta}$. It follows that,

$$\text{cov}\left\{\begin{pmatrix} \mathbf{y}^* \\ \boldsymbol{\delta}^* \end{pmatrix}\right\} = \begin{pmatrix} (\frac{1}{\sigma^2}\boldsymbol{\Sigma}_2 + \mathbf{I}_n)^{-1}\boldsymbol{\Sigma}_Y(\frac{1}{\sigma^2}\boldsymbol{\Sigma}_2 + \mathbf{I}_n)^{-1} & \mathbf{0}_{n,n} \\ \mathbf{0}_{n,n} & \sigma^2\mathbf{I}_n \end{pmatrix},$$

which has the same form as (5). Consequently, in Section 5 we emphasize Special Case 2 less than other choices because implementing Special Case 2 is identical (after transformation) to implementing a model using the Standard Assumption.

2.5 Special Case 3

In this section, we consider a slight departure from the Standard Assumption that $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$. Specifically, we assume $\boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_Y$. In this case, the expression of (3) is

given by,

$$\text{cov}\left\{\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\delta}\end{pmatrix}\right\} = \begin{pmatrix}\boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{Y,w}-\boldsymbol{\Sigma}_Y\\ \boldsymbol{\Sigma}_{Y,w}^\top-\boldsymbol{\Sigma}_Y^\top & \sigma^2\mathbf{I}_n-2(\boldsymbol{\Sigma}_{Y,w}-\boldsymbol{\Sigma}_Y)\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma}^\top & \sigma^2\mathbf{I}_n-2\boldsymbol{\Sigma}\end{pmatrix}, \quad (8)$$

where $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}_{Y,w} - \boldsymbol{\Sigma}_Y$. In Supplemental Appendix A, we show that (8) is positive semi-definite provided $\boldsymbol{\Sigma}_Y$ is positive semi-definite and $\boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_{Y,w}^\top \boldsymbol{\Sigma}_Y^{-1} \boldsymbol{\Sigma}_{Y,w}$ is positive semi-definite. We refer to the assumption that $\boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_Y$ as "Special Case 3."

2.6 Special Case 4

Assume that $\boldsymbol{\Sigma}_w \approx \boldsymbol{\Sigma}_Y$ and $\boldsymbol{\Sigma}_{Y,w} \approx \boldsymbol{\Sigma}_Y$, where $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_{Y,w}$ are defined to be projections onto a reduced dimensional space. Specifically, let $\boldsymbol{\Psi} \in \mathbb{R}^n \times \mathrm{R}^r$ be a $n \times r$ $(r \leq n)$ matrix consisting of spatial basis functions evaluated at $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. Several examples of $\boldsymbol{\Psi}$ are provided in the Supplemental Appendix C. Then, we assume that the $\boldsymbol{\Sigma}_w$ is the left-and-right-projections of $\boldsymbol{\Sigma}_Y$ onto the column space of $\boldsymbol{\Psi}$, and $\boldsymbol{\Sigma}_{Y,w}$ is the right projection of $\boldsymbol{\Sigma}_Y$ onto the column space of $\boldsymbol{\Psi}$. That is,

$$\boldsymbol{\Sigma}_w = \underset{\mathbf{K}\in\mathbb{R}^r\times\mathbb{R}^r}{\arg\min}||\boldsymbol{\Psi}\mathbf{K}\boldsymbol{\Psi}^\top - \boldsymbol{\Sigma}_Y||_F^2 = \boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_Y\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top \quad (9)$$

$$\boldsymbol{\Sigma}_{Y,w} = \underset{\mathbf{C}\in\mathbb{R}^n\times\mathbb{R}^r}{\arg\min}||\mathbf{C}\boldsymbol{\Psi}^\top - \boldsymbol{\Sigma}_Y||_F^2 = \boldsymbol{\Sigma}_Y\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top, \quad (10)$$

respectively. The operator $||\cdot||_F^2$ is known as the Frobenius norm, and for any square real-valued matrix $\mathbf{M}$ we have that $||\mathbf{M}||_F^2 = \text{trace}(\mathbf{M}^\top\mathbf{M})$. The expressions on the far right-hand side of (9) and (10) are an immediate consequence of a result in Cressie and Johannesson (2008). For notational convenience denote the hat matrix $\mathbf{P} = \boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top$.

In this case, the expression of (3) is given by,

$$\text{cov}\left\{\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\delta}\end{pmatrix}\right\} = \begin{pmatrix}\boldsymbol{\Sigma}_Y & -\boldsymbol{\Sigma}_Y(\mathbf{I}_n-\mathbf{P})\\ -(\mathbf{I}_n-\mathbf{P})\boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_Y(\mathbf{I}_n-\mathbf{P})-(\mathbf{I}_n-\mathbf{P})\boldsymbol{\Sigma}_Y\mathbf{P}+\sigma^2\mathbf{I}_n\end{pmatrix}. \quad (11)$$

In Supplemental Appendix A, we show that (11) is positive semi-definite provided that $\boldsymbol{\Sigma}_Y$ is positive semi-definite. The cross-covariance term between the signal and noise is determined by the covariance of $\mathbf{y}$ and the basis functions evaluated at all locations in $\{\mathbf{s}_1,\ldots,\mathbf{s}_n\}$. Thus, it is especially important that the basis functions are chosen carefully. In Supplemental Appendix C, we describe an algorithm to select the basis functions using an out-of-sample measure of error (see Section 3).

There is an interesting relationship between (11) and the Standard Assumption. Namely, if we set $\boldsymbol{\Sigma}_Y = \mathbf{PHP}$ for some $n \times n$ positive definite matrix $\mathbf{H}$, we obtain,

$$\text{cov}\left\{\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\delta}\end{pmatrix}\right\} = \begin{pmatrix}\mathbf{PHP} & \mathbf{0}_{n,n}\\ \mathbf{0}_{n,n} & \sigma^2\mathbf{I}_n\end{pmatrix}.$$

Thus, to capture discrepancy error dependencies we require $\boldsymbol{\Sigma}_Y$ to have columns that fall in the orthogonal column space of $\mathbf{P}$. In our empirical results we have checked to make sure all eigenvectors of $\boldsymbol{\Sigma}_Y$ are in the orthogonal complement space of $\mathbf{P}$.

2.7 The Kriging Predictor With Non-Confounded discrepancy error Covariances

The traditional kriging predictor (e.g., see Matheron, 1963; Cressie, 1990, among others) is a standard optimal predictor (in terms of minimizing mean squared prediction error) in spatial statistics, and should be discussed under the General Assumption. In Supplemental Appendix A, we show that the kriging predictor, when one assumes $\boldsymbol{\Sigma}_w \neq \boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_Y$, is

given by,

$$E(\mathbf{y}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_w, \boldsymbol{\Sigma}_{Y,w}, \boldsymbol{\Sigma}_Y, \sigma^2) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_{Y,w}^\top \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu}). \tag{12}$$

This kriging predictor has covariance (see Supplemental Appendix A),

$$\text{cov}(\mathbf{y}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_w, \boldsymbol{\Sigma}_{Y,w}, \boldsymbol{\Sigma}_Y, \sigma^2) = \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_{Y,w}^\top \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} \boldsymbol{\Sigma}_{Y,w}. \tag{13}$$

The special cases discussed in Sections 2.1 − 2.5 lead to illuminating special cases of the kriging predictor.

Proposition 2 (kriging Predictors): Let the data vector $\boldsymbol{z}$ *have probability density function according to (1). Suppose the General Assumption from Section 2.1 holds. Then we have the following expressions of the kriging predictor and kriging covariances:*

a. *Standard Assumption (*$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$*):*

$$E(\mathbf{y}|\mathbf{z}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_w \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu});$$

$$\text{cov}(\mathbf{y}|\mathbf{z}) = \boldsymbol{\Sigma}_w - \boldsymbol{\Sigma}_w \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} \boldsymbol{\Sigma}_w;$$

b. *Special Case 1 (*$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_Y$*):*

$$E(\mathbf{y}|\mathbf{z}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_w \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu});$$

$$\text{cov}(\mathbf{y}|\mathbf{z}) = \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_w \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} \boldsymbol{\Sigma}_w;$$

c. *Special Case 2 (*$\boldsymbol{\Sigma}_w \neq \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$*):*

$$E(\mathbf{y}|\mathbf{z}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_Y \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu});$$

$$\text{cov}(\mathbf{y}|\mathbf{z}) = \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_Y \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} \boldsymbol{\Sigma}_Y;$$

d. *Special Case 3 (*$\boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_Y$*):*

$$E(\mathbf{y}|\mathbf{z}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_{Y,w}^\top \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu});$$

$$\text{cov}(\mathbf{y}|\mathbf{z}) = \boldsymbol{\Sigma}_w - \boldsymbol{\Sigma}_{Y,w}^\top \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} \boldsymbol{\Sigma}_{Y,w};$$

e. *Special Case 4 (*$\boldsymbol{\Sigma}_w \approx \boldsymbol{\Sigma}_{Y,w} \approx \boldsymbol{\Sigma}_Y$*):*

$$E(\mathbf{y}|\mathbf{z}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_Y \mathbf{P} \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} (\mathbf{z} - \boldsymbol{\mu});$$
$$\text{cov}(\mathbf{y}|\mathbf{z}) = \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_Y \mathbf{P} \left(\boldsymbol{\Sigma}_w + \sigma^2 \mathbf{I}_n\right)^{-1} \mathbf{P} \boldsymbol{\Sigma}_Y.$$

Proof: Follows immediately from the General Assumption and rules for conditional and marginal distributions of Gaussian random vectors (Ravishanker and Dey, 2002).

In Proposition 2 (a,b), the kriging predictors are the same. However, the kriging variances are larger when assuming Special Case 1 than the kriging variances under the Standard Assumption. In Proposition 2 (c,d,e), both the kriging predictor and kriging covariance differ from the expressions that are developed under the Standard Assumption. In Proposition 2(d), if we set $\boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y \mathbf{P}$ and $\boldsymbol{\Sigma}_w = \mathbf{P}\boldsymbol{\Sigma}_Y \mathbf{P}$ we obtain the same kriging predictor as in Special Case 4. Henceforth, we use $\boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y \mathbf{P}$ when applying Special Case 3.

3 An Empirical Measure of Out-of-Sample Error

The strategy presented in Sections 2.3 − 2.5 are all based on making a small change to the Standard Assumption that $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$. By providing these flexible modeling assumptions, it is incumbent on us to provide an empirical measure to assess the appropriateness of these new assumptions in practice. We propose the following criterion:

$$\begin{aligned} &\sum_{\mathbf{s}} \left\{ Z(\mathbf{s}) - \widehat{Y}(\mathbf{s}) \right\}^2 + 2 \sum_{\mathbf{s}} \text{cov} \left\{ Z(\mathbf{s}) - Y(\mathbf{s}) \right\} \widehat{Y}(\mathbf{s}) \\ &= \sum_{\mathbf{s}} \left\{ Y(\mathbf{s}) - \widehat{Y}(\mathbf{s}) \right\}^2 + \sum_{\mathbf{s}} \left\{ Z(\mathbf{s}) - Y(\mathbf{s}) \right\}^2 + 2 \sum \left\{ Z(\mathbf{s}) - Y(\mathbf{s}) \right\} Y(\mathbf{s}), \end{aligned} \tag{14}$$

where $\widehat{Y}(\mathbf{s})$ is a generic real-valued function of $\mathbf{z}$, which represents a prediction at $\mathbf{s} \in D$. Notice that the left hand side of Equation (14) is the so-called covariance penalized error introduced in Efron (1983), and is a measure of out-of-sample error. This out-of-sample

criterion is preferable to others because it is general enough to capture non-confounded discrepancy error covariances. In practice, we use the posterior expected value of the cross product between $Z(\mathbf{s}) - Y(\mathbf{s})$ and $\widehat{Y}(\mathbf{s})$ to estimate the covariance term on the left-hand side of (14). See Supplemental Appendix B for more details surrounding implementation.

In Special Cases 1 and 3 we use the estimated covariance penalized error to select the parameters that define $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. Recall, we consider Matérn and CAR model specifications of the matrices. Similarly, in Special Case 4, the choice of basis functions evaluated at pre-specified locations partially defines the discrepancy error covariances. Consequently, we provide a stepwise algorithm that uses the covariance penalized error in (14) (see Supplemental Appendix C).

4 Bayesian Implementation: Process Augmentation

4.1 Process Augmentation

It is often useful to introduce an artificial latent random variable, such that upon marginalization, one obtains the original joint probability function (e.g., see Tanner and Wong, 1987; Albert and Chib, 1993; Wakefield and Walker, 1999; Wolpert and Ickstadt, 1998, among others). We extend this strategy to our setting by developing a similar "process augmentation" approach. Specifically, let

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \delta(\mathbf{s}) \tag{15}$$

$$\delta(\mathbf{s}) = w(\mathbf{s}) - Y(\mathbf{s}) + \epsilon(\mathbf{s}); \;\; \mathbf{s} \in D, \tag{16}$$

where $w(\mathbf{s}) - Y(\mathbf{s}) \neq 0$ for at least one location $\mathbf{s} \in D$, $w(\cdot)$ is a Gaussian process with mean function $\mu(\cdot)$, and $\epsilon(\cdot)$ is a mutually independent error term with mean zero and variance σ^2.

Let $\mathbf{z}_n = \{Z(\mathbf{u}_1), \ldots, Z(\mathbf{u}_n)\}^\top$, $\mathbf{w} = (w(\mathbf{u}_1), \ldots, w(\mathbf{u}_n))^\top$, and $\boldsymbol{\epsilon} = (\epsilon(\mathbf{u}_1), \ldots, \epsilon(\mathbf{u}_n))^\top$,

so that (15) becomes

$$\mathbf{z}_n = \mathbf{y} + \boldsymbol{\delta} \tag{17}$$

$$\boldsymbol{\delta} = \mathbf{w} - \mathbf{y} + \boldsymbol{\epsilon}. \tag{18}$$

Let $\text{cov}(\mathbf{y}, \mathbf{w}) \equiv \boldsymbol{\Sigma}_{Y,w}$ and $\text{cov}(\mathbf{w}) \equiv \boldsymbol{\Sigma}_w$. Then from (17) and (18) it follows that

$$\begin{aligned}\boldsymbol{\Sigma}_{Y,\delta} &= \boldsymbol{\Sigma}_{Y,w} - \boldsymbol{\Sigma}_Y \\ \boldsymbol{\Sigma}_\delta &= \boldsymbol{\Sigma}_Y + \boldsymbol{\Sigma}_w - 2\boldsymbol{\Sigma}_{Y,w} + \sigma^2 \mathbf{I}_n,\end{aligned}$$

and hence, we obtain the General Assumption. Many of the special cases can arise through Equations (17) and (18). For example, Special Case 3 occurs when we define $\text{cov}(\mathbf{w}) = \text{cov}(\mathbf{y})$. The remaining special cases are organized into Proposition 3.

Proposition 3 (Process Augmentation, Special Cases): Assume the model in (15) and (16); complete regularity conditions are provided in Supplemental Appendix D. Let $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \in D$ *be a generic collection of points in* D*. Suppose the General Assumption from Section 2.1 holds. Then we have the following*

a. *The Standard Assumption (*$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$*) can be obtained by assuming* $w(\cdot) \equiv Y(\cdot)$.

b. *Special Case 1 (*$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{Y,w} \neq \boldsymbol{\Sigma}_Y$*) can be obtained by assuming* $Y(\cdot) \equiv w(\cdot) + \epsilon_Y(\cdot)$*, where* $\epsilon_Y(\cdot)$ *is independent of* $w(\cdot)$ *and is a Gaussian process.*

c. *Special Case 2 (*$\boldsymbol{\Sigma}_w \neq \boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y$*) can be obtained by assuming* $w(\cdot) \equiv Y(\cdot) + \epsilon_w(\cdot)$*, where* $\epsilon_w(\cdot)$ *is independent of* $Y(\cdot)$ *and is a Gaussian process with mean zero.*

d. *Special Case 3 (*$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_Y \neq \boldsymbol{\Sigma}_{Y,w}$*) can be obtained by assuming* $w(\cdot)$ *and* $Y(\cdot)$ *are*

identically distributed, where $w(\cdot)$ and $Y(\cdot)$ are dependent Gaussian processes.

e. *Special Case 4 ($\boldsymbol{\Sigma}_w \approx \boldsymbol{\Sigma}_{Y,w} \approx \boldsymbol{\Sigma}_Y$) can be obtained by assuming $\boldsymbol{w} \equiv \boldsymbol{\mu} + \boldsymbol{\Psi}\boldsymbol{\eta} + \boldsymbol{\xi}$, where $\boldsymbol{\Psi}$ is defined in Section 2.6, and $\boldsymbol{\eta}$ is a mean-zero Gaussian random vector such that $\text{cov}(\boldsymbol{\eta}) = (\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_Y\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$ and $\text{cov}(\boldsymbol{y}, \boldsymbol{\eta}) = \boldsymbol{\Sigma}_Y\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$, and $\boldsymbol{\xi}$ is an independent n-dimensional Gaussian random vector.*

Proof: Follows immediately from the General Assumption and rules for conditional and marginal distributions of Gaussian random vectors (Ravishanker and Dey, 2002).

In our exposition, the process $w(\cdot)$ is interpreted as an artificial quantity that will aid in implementation (see discussion in Tanner and Wong, 1987; Albert and Chib, 1993; Wakefield and Walker, 1999; Wolpert and Ickstadt, 1998, among others). However, several current methods interpret $w(\cdot)$ as an approximation of $Y(\cdot)$ (Banerjee et al., 2008; Cressie and Johannesson, 2008; Sang and Huang, 2012), but still paradoxically assume the Standard Assumption that $w(\cdot) \equiv Y(\cdot)$. We assert that it is more realistic to assume that $w(\cdot)$ is separate from $Y(\cdot)$ when $w(\cdot)$ represents an approximation of $Y(\cdot)$.

Consider Special Case 4, which from Proposition 3, is equivalent to assuming $w(\mathbf{s}) = \sum_{i=1}^{r} \psi_i(\mathbf{s})\eta_i$, where $\{\eta_i\}$ are random effects and $\{\psi_i\}$ are real-valued spatial basis functions. Consider the Karhunen-Loéve representation of a spatial random process $Y(\cdot)$ (Karhunen, 1947; Loéve, 1978),

$$Y(\mathbf{s}) = \sum_{i=1}^{\infty} \psi_i(\mathbf{s})\alpha_i, \tag{19}$$

where $\{\psi_i(\cdot)\}$ are orthonormal and the random variables $\{\alpha_i\}$ are uncorrelated with mean zero and variance $\{\lambda_i\}$. Let $w(\cdot)$ be modeled using the *truncated* Karhunen-Loéve expansion,

$$w(\mathbf{s}) = \sum_{i=1}^{r} \psi_i(\mathbf{s})\alpha_i,$$

which for "large" r, $w(\cdot)$ approximates $Y(\cdot)$. Now, setting $w(\cdot) \equiv Y(\cdot)$ under the Standard Assumption is the same as claiming,

$$\sum_{i=r+1}^{\infty} \psi_i(\mathbf{s})\alpha_i = 0, \tag{20}$$

$$Y(\mathbf{s}) = \sum_{i=1}^{r} \psi_i(\mathbf{s})\alpha_i \tag{21}$$

$$Z(\mathbf{s}) = \sum_{i=1}^{r} \psi_i(\mathbf{s})\alpha_i + \epsilon(\mathbf{s}), \tag{22}$$

for *every* $\mathbf{s} \in D$. Recently, Stein (2014) has shown inferential problems with the KL-divergence measure when making the assumptions in (20), (21), and (22). Some have tried to adjust for this by using the Standard Assumption and placing an artificial model (i.e., tapered covariance, or white noise) for $\sum_{i=r+1}^{\infty} \psi_i(\mathbf{s})\alpha_i$, typically chosen for computational reasons (Finley et al., 2009; Sang and Huang, 2012). Instead of placing an artificial model on $\sum_{i=r+1}^{\infty} \psi_i(\mathbf{s})\alpha_i$ we choose to model $Y(\cdot)$ in (19) directly through Special Case 4.

For this heuristic, Special Case 4 can be seen as the following assumption:

$$Y(\mathbf{s}) = \sum_{i=1}^{\infty} \psi_i(\mathbf{s})\alpha_i$$

$$Z(\mathbf{s}) = \sum_{i=1}^{r} \psi_i(\mathbf{s})\alpha_i + \epsilon(\mathbf{s}).$$

Thus, if the Standard Assumption is correct then we can use our modeling approach to make the correct assumption on $Y(\cdot)$ and $\sum_{i=r+1}^{\infty} \psi_i(\mathbf{s})\alpha_i$, but use an approximation to model $Z(\cdot)$. Thus, in this case, the discrepancy error variances are induced through a misspecified/approximated model for $Z(\cdot)$.

In Section 5, we compare the predictor under Special Case 4 to several methods that assume an approximated process $w(\cdot)$ is equivalent to the exact process $Y(\cdot)$. Specifically, we compare to the Full-Scale Approximation (FSA), modified predictive processes (MPP),

and Bayesian fixed rank kriging (FRK) (Cressie and Johannesson, 2008; Finley et al., 2009; Sang and Huang, 2012). See Supplemental Appendix E for a review of these methods.

4.2 Posterior Predictive Distributions

The presence of $w(\cdot)$ can be used to obtain computationally efficient predictions. The following technical results demonstrate this for the Bayesian setting by showing a useful conditional independence property of the posterior predictive distribution of $\mathbf{y}$.

Theorem 1: Let $\mathcal{S} \subset D \subset \mathbb{R}^d$ be an open set. For each $k \in \mathbb{N} = \{1, 2, 3, \ldots\}$ and finite collection of locations $\mathbf{s}_1, \ldots, \mathbf{s}_k \in \mathcal{S}$. Define the n-dimensional random vectors $\boldsymbol{w} = \{w(\mathbf{s}_{k+1}), \ldots, w(\mathbf{s}_{k+n})\}^\top$, $\mathbf{z} = \{Z(\mathbf{s}_{k+1}), \ldots, Z(\mathbf{s}_{k+n})\}^\top$, and $\boldsymbol{\epsilon} = \{\epsilon(\mathbf{s}_{k+1}), \ldots, \epsilon(\mathbf{s}_{k+n})\}^\top$, for $\{\mathbf{s}_{k+1}, \ldots, \mathbf{s}_{k+n}\} \in D$. Suppose all marginal and conditional densities are proper (see Regularity Conditions in Supplemental Appendix D). Also, let $f(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_k)|\boldsymbol{w}, \boldsymbol{\theta})$ be Kolmogorov consistent. Then, there exists a probability space (with sample space Ω, sigma-algebra $\mathcal{A}$, and probability measure $\mathbb{P}$) and stochastic process $Y : \mathcal{S} \times \Omega \to \mathbb{R}$, such that

$$\begin{aligned}\mathbb{P}\left\{Y(\mathbf{s}_1) \in A_1, \ldots, Y(\mathbf{s}_k) \in A_k\right\} &= \int_{A_1} \cdots \int_{A_k} f(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_k)|\boldsymbol{w}, \boldsymbol{\theta}, \mathbf{z}) dY(\mathbf{s}_1) \ldots dY(\mathbf{s}_k) \\ &= \int_{A_1} \cdots \int_{A_k} f(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_k)|\boldsymbol{w}, \boldsymbol{\theta}) dY(\mathbf{s}_1) \ldots dY(\mathbf{s}_k),\end{aligned} \tag{23}$$

for all $\mathbf{s}_1, \ldots, \mathbf{s}_k \in \mathcal{S}$, $k \in \mathbb{N}$, and measurable sets $A_i \subset \mathbb{R}$; $i = 1, \ldots, k$.

Proof: See Supplemental Appendix F.

The conditional independence property in Equation (23) has been referred to as "Bayesianly

unidentified" (e.g., see Banerjee et al., 2015, page 157). However, this is different from no Bayesian learning, which would imply $f(Y(\mathbf{s}_1),\ldots,Y(\mathbf{s}_k)|\mathbf{w},\boldsymbol{\theta},\mathbf{z}) = f(Y(\mathbf{s}_1),\ldots,Y(\mathbf{s}_k))$. This independence property is not present in our modeling framework. To clarify the model structure we have included a pictorial representation of the augmented joint probability density function in Figure 1.

In practice we do not observe the entire spatial field D nor do we predict over the (possibly) uncountably infinite spatial domain. Thus, for practical purposes it is important to state the following corollary.

Corollary 1: Define the n-dimensional random vector $\boldsymbol{y} = \{Y(\boldsymbol{u}_1),\ldots,Y(\boldsymbol{u}_n)\}^\top$ *and the m-dimensional random vectors* $\boldsymbol{w} = \{w(\mathbf{s}_1),\ldots,w(\mathbf{s}_m)\}^\top$, $\mathbf{z} = \{Z(\mathbf{s}_1),\ldots,Z(\mathbf{s}_m)\}^\top$, *and* $\boldsymbol{\epsilon} = \{\epsilon(\mathbf{s}_1),\ldots,\mathbf{s}_m\}^\top$, *for* $\{\mathbf{s}_1,\ldots,\mathbf{s}_m\} \in D$ *and* $\{\mathbf{s}_1,\ldots,\mathbf{s}_n\} \subset \mathbb{R}^d \in D$. *Let* $(\Omega,\mathcal{A},\mathbb{P})$ *denote the probability space for* $\boldsymbol{y}$, $\boldsymbol{w}$, $\mathbf{z}$, *and* $\boldsymbol{\epsilon}$. *Assume that for every* $\omega \in \Omega$ *and* $\boldsymbol{h} \in \mathbb{R}^m$ *the set* $\{\omega : \mathbf{z}(\omega) = \boldsymbol{h}\} = \{\omega : \boldsymbol{\epsilon}(\omega) = \boldsymbol{h} - \mathbf{y}_Z(\omega)\}$ *(i.e., the additive model holds almost surely). Also assume that* $\mathbf{y}$ *and* $\boldsymbol{\epsilon}$ *are mutually independent. Let the probability density functions for* $\boldsymbol{y}|\boldsymbol{w},\boldsymbol{\theta},\mathbf{z}$ *and* $\boldsymbol{y}|\boldsymbol{w},\boldsymbol{\theta}$ *exist, and denote them with* $f(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{\theta},\mathbf{z})$ *and* $f(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{\theta})$, *respectively. Let* $\boldsymbol{\theta}$ *be a generic k-dimensional real-valued parameter vector. Then,* $f(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{\theta},\mathbf{z}) = f(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{\theta})$.

Proof: Follows immediately from Theorem 1 by setting $k = n$.

In Corollary 1, the probability density function $f(\mathbf{y}|\mathbf{w},\boldsymbol{\theta},\mathbf{z})$ is the so-called predictive distribution for $\mathbf{y}$ (Berger, 1985). (That is, say $\mathbf{w}^{[0]}$ and $\boldsymbol{\theta}^{[0]}$ are generated from their respective posterior distributions $f(\mathbf{w},\boldsymbol{\theta}|\mathbf{z})$, then a random vector with probability density function $f(\mathbf{y}|\mathbf{z})$ can be simulated from $f(\mathbf{y}|\mathbf{w}^{[0]},\boldsymbol{\theta}^{[0]},\mathbf{z})$.) This leads to a composition sampling approach for model implementation, where one first simulates $\mathbf{w}^{[0]}$ and $\boldsymbol{\theta}^{[0]}$ from $f(\mathbf{w},\boldsymbol{\theta}|\mathbf{z})$, and then simulates from $f(\mathbf{y}|\mathbf{w}^{[0]},\boldsymbol{\theta}^{[0]})$.

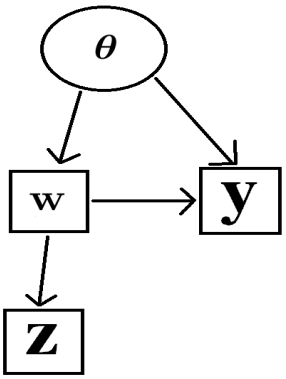

Figure 1: A pictorial representation of process augmentation. Circles represent parameters, squares represent random vectors that stack the data, the augmented process, and the latent process over different spatial locations. These vectors are defined in Corollary 1. The n-dimensional vector $\mathbf{z}$ is the observed data vector, $\mathbf{w}$ is the associated n-dimensional vector of the augmented values, and $\mathbf{y}$ is N-dimensional vector of the values of the latent process.

This implies there are two steps involved with obtaining posterior replicates of $\mathbf{y}$. That is, simulating first from the posterior distribution of $\mathbf{w}$ and $\boldsymbol{\theta}$, and second simulating from $f(\mathbf{y}|\mathbf{w}, \boldsymbol{\theta})$.

4.3 Bayesian Inference of the Latent Process

For each special case we assume $\mathbf{w} = \boldsymbol{\Psi}\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a mean-zero Gaussian random vector such that $\text{cov}(\boldsymbol{\eta}|\boldsymbol{\theta}) = (\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top\mathbf{H}(\boldsymbol{\theta})\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$. For point referenced data the (i,j)-th element of $\mathbf{H}(\boldsymbol{\theta})$ is $C_M(||\mathbf{s}_i - \mathbf{s}_j||; \boldsymbol{\theta})$, and for areal data $\mathbf{H}(\boldsymbol{\theta})$ is the covariance from a CAR model, where $\boldsymbol{\theta} = (\sigma^2_Y, \tau)'$. The assumptions made for implementation are outlined in Table 1.

In each setting, the procedure for posterior inference on $Y(\cdot)$ starts with obtaining B posterior replicates of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$, which we denote with $\boldsymbol{\beta}^{[b]}$, $\boldsymbol{\eta}^{[b]}$, $\boldsymbol{\xi}^{[b]}$, and $\boldsymbol{\theta}^{[b]}$, for $b = 1, \ldots, B$. Let the N-dimensional vector $\mathbf{w}^{[b]} = \left(w^{[b]}(\mathbf{s}_1), \ldots, w^{[b]}(\mathbf{s}_m)\right)^\top$ and $w^{[b]}(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top\boldsymbol{\beta} + \boldsymbol{\psi}(\mathbf{s})^\top\boldsymbol{\eta}^{[b]} + \xi^{[b]}(\mathbf{s})$ for each b and $\mathbf{s}$. In general, the full-conditional distributions for $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$ are well-known (e.g., see Cressie and Wikle, 2011, Ch. 7), and straightforward to

Model	$\boldsymbol{\Sigma}_w$	$\boldsymbol{\Sigma}_Y$	$\boldsymbol{\Sigma}_{Y,w}$	$\text{cov}(\mathbf{y}, \boldsymbol{\eta})$
Standard Assumption	Let $\boldsymbol{\Sigma}_w = \mathbf{PHP}$, where for point referenced data the (i,j)-th element of $\mathbf{H}$ is $C_M(\|\mathbf{s}_i - \mathbf{s}_j\|)$, and for areal data $\mathbf{H}$ is the covariance from a CAR model.	$\boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma}_w$	$\boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_w$	$\text{cov}(\mathbf{y}, \boldsymbol{\eta}) = \mathbf{PH}\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$
Special Case 1	Let $\boldsymbol{\Sigma}_w = \mathbf{PHP}$, where for point referenced data the (i,j)-th element of $\mathbf{H}$ is $C_M(\|\mathbf{s}_i - \mathbf{s}_j\|)$, and for areal data $\mathbf{H}$ is the covariance from a CAR model.	Let $\boldsymbol{\Sigma}_Y = \mathbf{PHP} + \boldsymbol{\Sigma}_1$. For point referenced data the (i,j)-th element of $\boldsymbol{\Sigma}_1$ is $C_M(\|\mathbf{s}_i - \mathbf{s}_j\|)$, and for areal data $\boldsymbol{\Sigma}_1$ is the covariance from a CAR model. Recall from Section 4, the parameters of $\boldsymbol{\Sigma}_1$ are chosen using the covariance penalized error.	$\boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_w$	$\text{cov}(\mathbf{y}, \boldsymbol{\eta}) = \mathbf{PH}\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$
Special Case 3	Let $\boldsymbol{\Sigma}_w = \mathbf{PHP}$, where for point referenced data the (i,j)-th element of $\mathbf{H}$ is $C_M(\|\mathbf{s}_i - \mathbf{s}_j\|)$, and for areal data $\mathbf{H}$ is the covariance from a CAR model.	$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_Y$	Following the discussion at the end of Section 2.7, we set $\boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y\mathbf{P}$.	$\text{cov}(\mathbf{y}, \boldsymbol{\eta}) = \mathbf{H}\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$
Special Case 4	$\boldsymbol{\Sigma}_w = \mathbf{P}\boldsymbol{\Sigma}_Y\mathbf{P}$	For point referenced data the (i,j)-th element of $\boldsymbol{\Sigma}_Y$ is $C_M(\|\mathbf{s}_i - \mathbf{s}_j\|)$ and for areal data $\boldsymbol{\Sigma}_Y$ is the covariance from a CAR model.	$\boldsymbol{\Sigma}_{Y,w} = \boldsymbol{\Sigma}_Y\mathbf{P}$	$\text{cov}(\mathbf{y}, \boldsymbol{\eta}) = \boldsymbol{\Sigma}_Y\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$

Table 1: Assumptions made for Bayesian inference in Section 4.3. The choice of $\boldsymbol{\Sigma}_w = \mathbf{PHP}$ stays the same. Thus, to obtain this $\boldsymbol{\Sigma}_w$ we set $\mathbf{w} = \boldsymbol{\Psi}\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a mean-zero Gaussian random vector such that $\text{cov}(\boldsymbol{\eta}) = (\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top\mathbf{H}\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}$. In the fifth column we give the expression of $\text{cov}(\mathbf{y}, \boldsymbol{\eta})$ that produces $\boldsymbol{\Sigma}_{Y,w}$. We do not include Special Case 2, which can be implemented using the Standard Assumption after transformation.

compute. For ease of exposition, we outline the final statistical model and the corresponding full conditional distributions in Supplemental Appendix B.

We point out that the computationally intensive likelihood associated with $Y(\cdot)$ is not needed to obtain posterior replicates of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$. Instead, we only need to use the probability density functions, $[\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\theta}]$, $[\boldsymbol{\eta}|\boldsymbol{\theta}]$, $[\boldsymbol{\xi}|\boldsymbol{\theta}]$, $[\boldsymbol{\beta}]$, and $[\boldsymbol{\theta}]$. Thus, obtaining $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$ only requires Br^3 computations, generally speaking. Additionally, questions of confounding between $Y(\cdot)$ and $\delta(\cdot)$ are avoided when estimating these parameters, since obtaining $\boldsymbol{\beta}^{[b]}$, $\boldsymbol{\eta}^{[b]}$, $\boldsymbol{\xi}^{[b]}$, and $\boldsymbol{\theta}^{[b]}$ does not require joint modeling of $Y(\cdot)$ and $\delta(\cdot)$.

With MCMC replicates of random effects and parameters in-hand, we can now use the posterior predictive distribution of $\mathbf{y}$ to obtain samples from $[\mathbf{y}|\mathbf{z}]$. Using Corollary 1 and standard results for the Gaussian distribution, the predictive distribution for $Y(\mathbf{s})$ is given

by

$$Y(\mathbf{s})|\boldsymbol{\eta},\boldsymbol{\xi},\boldsymbol{\theta},\mathbf{z} \sim \text{Gau}\left(\mathbf{e}(\mathbf{s})^\top E(\mathbf{y}|\boldsymbol{\eta},\boldsymbol{\xi},\boldsymbol{\theta}), \mathbf{e}(\mathbf{s})^\top \mathbf{K}(\boldsymbol{\theta})\mathbf{e}(\mathbf{s})\right), \tag{24}$$

where $\mathbf{K}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_Y - \text{cov}\left(\mathbf{y},\boldsymbol{\eta}|\boldsymbol{\theta}\right)(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})(\boldsymbol{\Psi}^\top\mathbf{H}(\boldsymbol{\theta})\boldsymbol{\Psi})^{-1}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})\text{cov}\left(\boldsymbol{\eta},\mathbf{y}|\boldsymbol{\theta}\right)$, the elemental vector $\mathbf{e}(\mathbf{s}^*) \equiv (I(\mathbf{s}^* = \mathbf{s}) : \mathbf{s} \in \{\mathbf{u}_1,\ldots,\mathbf{u}_n\})^\top$, and $I(\cdot)$ is the indicator function. The choice of $\mathbf{K}$ depends on which special case we are considering. These choices are outlined in Table 1, where we explicitly provide $\boldsymbol{\Sigma}_Y$ and $\text{cov}(\mathbf{y},\boldsymbol{\eta}|\boldsymbol{\theta})$, which is needed to compute $\mathbf{K}$.

For each b and $\mathbf{s}$,

$$\begin{aligned} Y^{[b]}(\mathbf{s}) &= \mathbf{e}(\mathbf{s})^\top E\left(\mathbf{y}|\boldsymbol{\eta}^{[b]},\boldsymbol{\xi}^{[b]},\boldsymbol{\theta}^{[b]}\right) + \{\mathbf{e}(\mathbf{s})^\top\mathbf{K}(\boldsymbol{\theta}^{[b]})\mathbf{e}(\mathbf{s})\}^{1/2}\phi \\ &= \mathbf{x}(\mathbf{s})^\top\boldsymbol{\beta}^{[b]} + \mathbf{e}(\mathbf{s})^\top\text{cov}\left(\mathbf{y},\boldsymbol{\eta}|\boldsymbol{\theta}^{[b]}\right)\mathbf{K}(\boldsymbol{\theta}^{[b]})^{-1}\boldsymbol{\eta}^{[b]} + \mathbf{e}(\mathbf{s})^\top\boldsymbol{\xi}^{[b]} + \{\mathbf{e}(\mathbf{s})^\top\mathbf{K}(\boldsymbol{\theta}^{[b]})\mathbf{e}(\mathbf{s})\}^{1/2}\phi, \end{aligned}$$

where ϕ is a draw from a standard normal random distribution. Then, posterior predictions and prediction variances of $Y(\cdot)$ can be estimated by the following,

$$\begin{aligned} \widehat{E}(Y(\mathbf{s})|\mathbf{z}) &= \frac{1}{B}\sum_{b=1}^{B} Y^{[b]}(\mathbf{s}) \\ \widehat{\text{var}}(Y(\mathbf{s})|\mathbf{z}) &= \frac{1}{B}\sum_{b=1}^{B}\left\{Y^{[b]}(\mathbf{s}) - \widehat{E}(Y(\mathbf{s})|\mathbf{z})\right\}^2, \end{aligned} \tag{25}$$

where we let $\widehat{\mathbf{y}} \equiv \left\{\widehat{E}(Y(\mathbf{u}_i)|\mathbf{z}) : i = 1,\ldots,n\right\}^\top$. Notice that the order of computations need to obtain predictions of $Y(\cdot)$ is order N (i.e., we simulate from Equation (24) B times). Additionally one does not need to store $\mathbf{K}$, but only needs to store the n values $\{\mathbf{e}(\mathbf{s})^\top\mathbf{K}(\boldsymbol{\theta}^{[b]})\mathbf{e}(\mathbf{s})\}^{1/2}$, and the $n \times r$ matrix $\boldsymbol{\Psi}$; since r is presumed to be small storage is not difficult in this setting. Additionlly, we do not compute and store $\mathbf{H}$, nor do we compute and store $\mathbf{K}$ before we obtain $\{\mathbf{e}(\mathbf{s})^\top\mathbf{K}(\boldsymbol{\theta}^{[b]})\mathbf{e}(\mathbf{s})\}^{1/2}$. The order of matrix multiplications are carefully chosen to avoid storing an $n \times n$ dense matrix. For example, to obtain $\mathbf{K}$ we compute the first column of $\mathbf{H}$ and pre-multiply the column by the n-dimensional vector

$(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top$, which gives an r-dimensional vector. The remaining steps are described in Supplemental Appendix B.

In principle, one could use *any* prior on $\boldsymbol{\theta}$, but not every prior on $\boldsymbol{\theta}$ is computationally feasible. We suggest using a discrete uniform prior for computational reasons. Suppose the discrete uniform prior on $\boldsymbol{\theta}$ takes on M values. Then, one only needs to store the n values in $\{\mathbf{e}(\mathbf{s})^\top\mathbf{K}(\boldsymbol{\theta})\mathbf{e}(\mathbf{s})\}^{1/2}$ for each of the M values of the support of $\boldsymbol{\theta}$. However, if the prior distribution has continuous support then the n values in $\{\mathbf{e}(\mathbf{s})^\top\mathbf{K}(\boldsymbol{\theta})\mathbf{e}(\mathbf{s})\}^{1/2}$ would need to be computed each time $\boldsymbol{\theta}$ is updated within a Gibbs sampler, which is not computationally feasible. Of course, in low dimensions a continuous support for $\boldsymbol{\theta}$ would be straightforward to implement.

To choose the support of the discrete uniform distribution, we suggest considering several discrete fixed intervals for the range parameter of a Matérn or conditional autoregressive model, and using Spiegelhalter et al. (2002)'s deviance information criterion to select among the candidate supports. Our independent simulation studies suggest that the results are robust to this specification provided that the discrete uniform support is not too coarse (i.e., M is small). In Supplemental Appendix B we outline the statistical model in full, and provide the full-conditional distributions for the Gibbs sampler.

5 Empirical Results

In this section we provide several analyses including a simulation studies, and an analysis of median household income using a large dataset consisting of ACS 5-year period estimates at the census tract level. We do not provide empirical results of Special Case 2 because it is closely related to the Standard Assumption (see Section 2.4). Also, following the discussion at the end of Section 2.7, we only investigate predictions using Special Case 4 and not Special Case 3 (our parameterization of Special Case 3 produces the same kriging predictor as Special

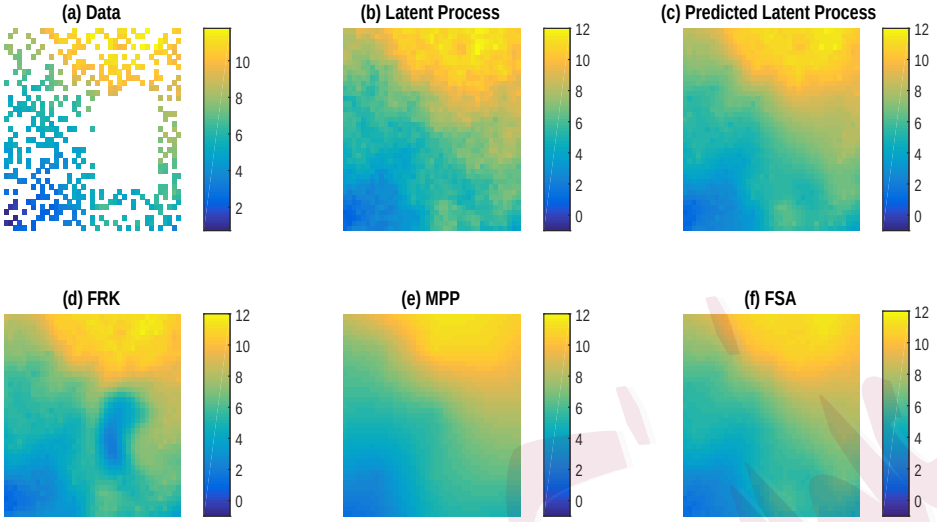

Figure 2: In Panel (a), we display the simulated data $\{Z(\cdot)\}$ over a collection of observed locations $D_O \subset D$, which are generated by randomly selecting points outside of a rectangular region in D. Here, D is a 40×40 grid $D \equiv \{(s_1, s_2)^\top : s_1, s_2 = 0, 0.025, \ldots, 1\}$. White areas indicate a missing observation. Panel (b) represents a simulation of the latent process with a Matérn covariance function as specified in the last row of Table 1. In Panels (c), (d), (e), (f) we present the posterior expected values of $Y(\cdot)$ from MD = SC4, FRK, MPP, and FSA defined in Table 1, respectively.

Case 4). A simulation study of Special Case 1 is provided in Supplemental Appendix G.

5.1 Simulation Study of Special Case 4: Robust to Departures from the Standard Assumption

In this simulation study, we compare Bayesian prediction under Special Case 4 to a modified predictive process (MPP) approach, Bayesian fixed rank kriging (FRK), and the full-scale approximation (FSA) Finley et al. (2009); Kang and Cressie (2011); Sang and Huang (2012). The Gibbs sampler for Bayesian FRK is outlined in Supplemental Appendix B, the spBayes R-package is used to compute the MPP predictor (Finley et al., 2015), and Matlab code to compute an empirical Bayes implementation of FSA Castillo and Tajbakhsh (2015). The same equally spaced knot locations are used for all methods that share the same rank r.

We consider several choices of r for each method. The goal of this analysis is to show that Special Case 4 is robust to departures in the Standard Assumption, by comparing to other reasonable choices for spatial prediction in the literature.

We generate a random process on a 40×40 grid $D \equiv \{(s_1, s_2)^\top : s_1, s_2 = 0, 0.025, \ldots, 1\}$, and let $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ be generated from a Matérn process with unknown mean $\mathbf{x}(\mathbf{s})^\top \boldsymbol{\beta} = 2 + u_1 + 7u_2$ for $\mathbf{s} = (u_1, u_2) \in D$, smoothing parameter 0.5 (i.e., an exponential covariagram), and unit variance. The range parameter of the Matérn process is $\tau = 1/12$, so that spatial range is moderate at $1/4$. Let $\text{var}\{\epsilon(\cdot)\} \equiv 0.5$ so that the signal to noise ratio is large (≈ 10).

To obtain simulated data we add independent error,

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s}); \ \ \mathbf{s} \in D_O,$$

where $\{\epsilon(\mathbf{s}_i) : i = 1, \ldots, m\}$ consists of i.i.d. Gaussian random variables with mean zero and variance 0.5. Notice that $Z(\cdot)$ is not generated according to the General Assumption. In Figure 3(a), we plot Z over our choice of observed data locations, which is missing at random at locations outside of a large square region in D. Notice that our model makes an incorrect assumption about the data process, but correctly specifies the latent process.

The bisquare radial basis function from Cressie and Johannesson (2008) are used. These basis functions depend on a collection of r knots. The location and number of these knots were found using an algorithm that compares estimates of the latent process to estimates of $w(\cdot)$. In Supplemental Appendix C, we outline how we select basis functions. For the predictions in Figure 3 we set $r = 100$.

The predictions in Figure 3 are based on the single realization of $Z(\cdot)$ and are given in Figure 3(a). Let the total mean squared prediction error be defined as

$$MSPE(\text{MD}) \equiv \frac{1}{|D|} \sum \left[Y(\mathbf{s}) - E_{MD} \left\{ Y(\mathbf{s}) | \mathbf{z} \right\} \right]^2; \ \text{MD} = \text{SC4}, \text{FRK}, \text{MPP}, \text{FSA},$$

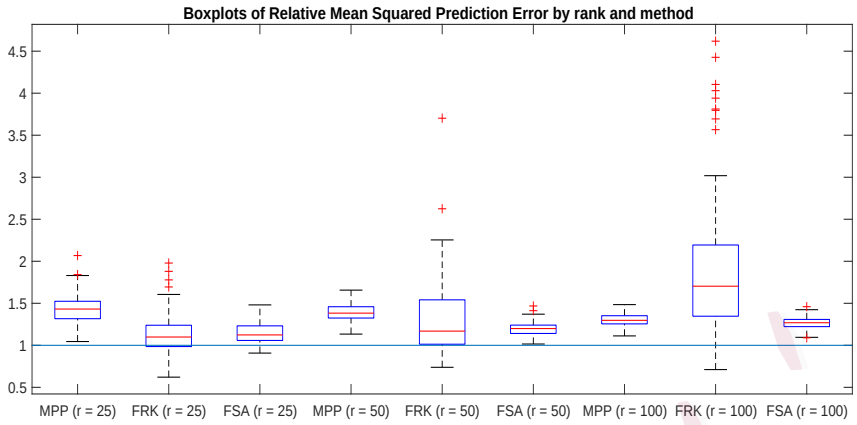

Figure 3: Boxplots of the rMSPE value defined in (26). The x-axis contains labels indicating the method and the choice of the rank r. Recall that the y-axis is a relative MSPE to SC4. When the rMSPE is greater than 1 this indicates that the MSPE for SC4 is smaller than the MSPE of the model labled on the x-axis.

where here "MD" stands for model in which the posterior expected value is taken and "SC4" stands for Special Case 4.

For the example in Figure 3, the total mean squared prediction errors are as follows: 0.08 for SC4 approach, 0.75 for FRK, 0.23 for MPP, and 0.12 for FSA. Thus, it is clear that the mean square prediction error is considerably smaller using the SC4 approach. For Bayesian FRK and MPP we see a strange circular artifact in the large square missing region in D, which is a well-known consequence of low rank modeling (Datta et al., 2014). This deficiency in low rank modeling is not present when using our model, nor is it present for the full-scale approximation of Sang and Huang (2012). These behaviors are consistent over multiple replicates. That is, after generating 100 different sets of $\{Z(\mathbf{s})\}$ and $\{Y(\mathbf{s})\}$ we obtain the results presented in Figure 4. Here, we plot the relative mean squared prediction error (rMSPE),

$$rMSPE(MD) = \frac{MSPE(\text{MD})}{MSPE(\text{SC4})}; \text{ MD} = \text{SC4}, \text{BFRK}, \text{MPP}, \text{FSA}. \tag{26}$$

For each choice of r, we see that we tend to have values of rMSPE over 1, which indicates that our approach (i.e., Model 1) outperforms each of it's competitors. Furthermore, for each method, as we increase the rank, the range of rMSPE-values tends to move further away from one, which indicates increased performances using the SC4 approach.

The data is simulated in a manner such that the optimal kriging predictor should outperform SC4. The mean squared prediction error of the optimal kriging predictor tends (over the replicate simulations) to be around 0.02. Additionally, the rMSPE associated with the traditional kriging predictor tends to be around 0.91. This suggests that the predictions using SC4 produces predictions with mean squared prediction error close (albeit greater than) to that of the optimal predictor.

5.2 Census-Tract Level ACS 5-Year Period Estimates of Median Household Income

The U.S. Census Bureau's American Community Survery (ACS) is a key data source for U.S. demographics. ACS estimates have a unique structure, where the estimates are reported for different time-periods; specifically, the ACS currently produces 1-year and 5-year period estimates of various U.S. demographic variables. We consider a subset of the ACS data and provide a spatial analysis of median household income for the 2009 to 2013 time-period. We consider a fairly large dataset with 72,361 observations, which consists of estimates of median household income over all census tracts in the contiguous U.S. A subset of the data is presented in Figure 4(a).

This particular example is especially interesting from the point of view of spatial analysis, as ACS period estimates of median household income have been modeled using Moran's I (MI) basis functions (e.g., see Bradley et al., 2015b, 2016). The previous literature suggests that the MI basis functions require a large r to obtain a reasonable fit using a low rank

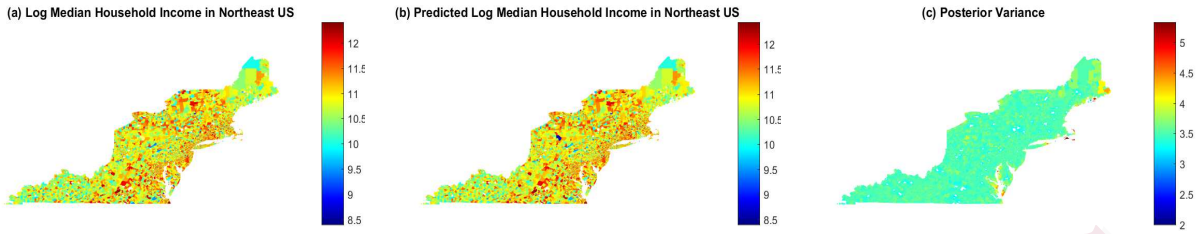

Figure 4: In Panel (a) we plot the log of the ACS 2013 5-year period estimates of median household income over census tracts in the northeast U.S. Panel (b) contains the predicted log median income, and Panel (c) displays the corresponding posterior variances. Data, predictions, and prediction variances are available over the entire U.S., however we only plot the northeast U.S. in Panels (a), (b), and (c) for visualization purposes.

approach; for example, in Bradley et al. (2015b) $r = 4,750$. Hence, one of our goals is to see if our approach allows one to model w with fewer basis functions than what has been common in the literature. Furthermore, federal datasets, like the public-use ACS estimates, are typically modeled using the assumption of independent "survey errors" (i.e., δ) (Fay and Herriot, 1979). Thus, evidence suggesting the survey errors are dependent may have an important implication for modeling federal data.

For this example, let $\{R(\mathbf{s}) : \mathbf{s} \in D\}$ represent the median household income over the census tracts in the contiguous U.S. (denoted by D). Histograms of the logarithm of R appear roughly symmetric indicating that normality of $Z(\cdot) = \log\{R(\cdot)\}$ is reasonable. Thus, for this example, we assume that Z (i.e., the logarithm of the data) is Gaussian. Additionally, the survey variances are converted to the log scale using the delta method (Oehlert, 1992). We use an intercept-only model and set $\mathbf{X}$ equal to a vector of ones. Currently, there are two methods that one might use to analyze this areal dataset, first is a version of Bayesian kriging found in Hughes and Haran (2013, among others). Another method that is common to fit data of this type is a conditional autoregressive model (Besag, 1974). In Supplemental Appendix H, we clearly layout the model assumptions, and how Special Case 1, Bayesian

fixed rank kriging, and the CAR model are implemented.

All Markov chains in this section use a burn-in of 1,000 and generate $B = 10,000$ posterior replications. Convergence was assessed visually through trace plots of the sample chain, with no lack of convergence detected. The rank r is specified according to Supplemental Appendix C. Here, we find that $r = 30$ is reasonable, and upon comparison to historical choices of r when using MI basis functions, we see that we obtain significant dimension reduction. For this dataset $\sigma^2_Y = 9$ and $\tau = 0.91$ minimize the covariance penalized error in (14). In Figure 4(b,c) we display a subset of the posterior mean and variances. Here, we see that the overall pattern of the predictions are similar to, but smoother than, the spatial patterns of the data. Additionally, estimates of Efron (2004)'s covariance penalized error (see (14)) is -0.9874 for our model, -0.1321 for FRK, and 0.9153 for the CAR model. Thus, it appears we are outperforming competing methods for spatial prediction.

6 Discussion

In this article, we have introduced a hierarchical model that leverages spatial dependencies within the error process associated with the data, and leverages cross-dependencies between the error process and the latent process of principal interest. This is done by introducing non-degenerate discrepancy error covariances that are not confounded within the marginal distribution of the data. The "Standard Assumption" of uncorrelated discrepancy errors is an important special case of our general parameterization, which occurs when three matrix valued parameters are equal to each other. We consider four additional special cases motivated by allowing for the matrix valued parameters to differ.

Our parameterization between the error process and the latent process leads to a computationally useful process augmentation approach. The first step in our implementation is to fit a model for the process that does not assume a dependent error process (see Supplemental

Appendix B). Then, the second step in our implementation is to produce predictions of a latent process that is dependent on the error process. The predictions in the second step of the algorithm are based on the output from the first step of the algorithm (see Section 5.2 and Supplemental Appendix B). This feature of our model greatly increases the applicability of approach because any statistical model that is based on the assumption of mutually independent error can be used within the first step of this algorithm.

Our empirical results suggest that our approach is robust to departures of our model assumptions. We illustrated this by implementing Special Case 3, when the data was generated using the full-rank Matérn specification. Low rank statistical models are known to be sensitive to the setting when the data are very sparse over the spatial domain. Hence, the data was generated to have sparsity. In this setting, our predictor outperformed many of the useful current methods special prediction. These result suggest that if the assumption of a dependent error process is not correct, then the non-confounded discrepancy error covariances may still be useful.

The non-confounded discrepancy error covariance can also be scaled to large datasets, and we demonstrated this using a dataset consisting of ACS estimates of median household income defined on census tracts. We would like to emphasize that we have produced precise predictions that have a full-rank specification, and obtained the computational gains of a reduced rank model. In a sense, we have paradoxically enjoyed the best of both worlds (i.e., a full rank specification and low rank specification). Furthermore, this example demonstrated that there appears to be dependent error in a popular survey dataset. This is a setting where it is typically assumed that the error process is independent of the latent process (i.e., the Fay-Herriot model used in federal statistics). This could make an important impact in small area estimation because the Fay-Herriot model is a ubiquitous choice for modeling area-level data found in the official statistics literature.

There are many opportunities for future research. For example, basis function selection

(e.g., see Huang et al., 2006; Bradley et al., 2011) is an important and recurring inferential question. We specified knot locations so that the augmented process was close to the latent process. Similarly, one could use this type of strategy to choose the class of spatial basis functions. In general, there is great potential to use this new modeling paradigm to specify parsimonious models in an informed manner.

Supplementary Materials

Proofs of technical results, a review of current methods in spatial statistics, an additional simulation, and additional details on model specification and implementation are all provided in the Supplemental Appendix.

Acknowledgments

We would like to express our gratitude to the editor, the associate editor, and the referees for their very helpful comments that improved this manuscript. This research was partially supported by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF grant SES-1132031, funded through the NSF-Census Research Network (NCRN) program. This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the NSF or the U.S. Census Bureau.

References

Albert, J. H. and Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88, 669–679.

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.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York, NY: Springer-Verlag.

Besag, J. E. (1974). "Spatial Interaction and the statistical analysis of lattice systems (with discussion)." *Journal of the Royal Statistical Society, Series B*, 36, 192–236.

Black, B. (1976). "Studies of Stock Price Volatility Changes." *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economic Statistics*, 177–181.

Bradley, J. R., Cressie, N., and Shi, T. (2011). "Selection of rank and basis functions in the Spatial Random Effects model." In *Proceedings of the 2011 Joint Statistical Meetings*, 3393–3406. Alexandria, VA: American Statistical Association.

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

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2015b). "Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates." *Stat*, 4, 255–270.

— (2016). "Bayesian spatial change of support for count-valued survey data." *Journal of the American Statistical Association*, 111, 472 – 487.

Castillo, E. and Tajbakhsh, B. M. C. . S. D. (2015). "Geodesic Gaussian Processes for the Reconstruction of a Free-Form Surface." *Technometrics*, 57, 87–99.

Clayton, D., Bernardinelli, L., and Montomoli, C. (1993). "Spatial correlation in ecological analysis." *International Journal of Epidemiology*, 6, 1193–1202.

Cressie, N. (1990). "The origins of kriging." *Mathematical Geology*, 22, 239–252.

— (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. (2014). "Hierarchical Nearest-Nerighbor Gaussian Process Models for Large Geostatistical Datasets." *arXiv preprint: 1406.7343*.

Efron, B. (1983). "Estimating the error rate of a prediction rule: Improvement on cross-validation." *Journal of the Americal Statistical Association*, 78, 316–331.

— (2004). "The estimation of prediction error: Covariance penalties and cross-validation." *Journal of the Americal Statistical Association*, 99, 619–642.

Fay, R. and Herriot, R. (1979). "Estimates of income for small places: an application of James-Stein procedures to census data." *Journal of the American Statistical Association*, 74, 269–277.

Finley, A. O., Banerjee, S., and Carlin, B. (2015). "Package 'spBayes'." http://cran.r-project.org/web/packages/spBayes/spBayes.pdf. Retrieved April, 2015.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). "Improving the performance of predictive process modeling for large datasets." *Computational Statistics and Data Analysis*, 53, 2873–2884.

Griffith, D. (2000). "A linear regression solution to the spatial autocorrelation problem." *Journal of Geographical Systems*, 2, 141–156.

— (2002). "A spatial filtering specification for the auto-Poisson model." *Statistics and Probability Letters*, 58, 245–251.

— (2004). "A spatial filtering specification for the auto-logistic model." *Environment and Planning A*, 36, 1791–1811.

Groves, R., Dillman, D. A., Eltinge, J. L., and Little, R. J. A. (2001). *Survey Nonresponse (Wiley Series in Survey Methodology)*. New York, NY: Wiley-Interscience.

Hodges, J. S. and Reich, B. J. (2011). "Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love." *The American Statistician*, 64, 325–334.

Holan, S. H. and Wikle, C. K. (2012). "Semiparametric Dynamic Design of Monitoring Networks for Non-Gaussian Spatio-Temporal Data." In *Spatio-Temporal Design Advances in Efficient Data Acquisition*, eds. J. Mateu and W. Muller. New York, NY: Wiley.

Huang, H. C., Hsu, N. J., Theobald, D., and Breidt, F. J. (2006). "Spatial LASSO with applications to GIS model selection."

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.

Kang, E. L. and Cressie, N. (2011). "Bayesian inference for the Spatial Random Effects model." *Journal of the American Statistical Association*, 106, 972 – 983.

Karhunen, K. (1947). "Uber lineare Methoden in der Wahrscheinlichkeitsrechnung." *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys*, 37, 1–49.

Katzfuss, M. (2017). "A multi-resolution approximation for massive spatial datasets." *Journal of the American Statistical Association*, 112, 201–214.

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.

Loéve, M. (1978). "Probability theory, vol. ii." *Graduate texts in mathematics*, 46, 0–387.

Matérn, B. (1960). "Spatial Variation." *Meddelanden fran Statens Skogsforskningsinstitut*, 49, 1–144.

Matheron, G. (1963). "Principles of geostatistics." *Economic Geology*, 58, 1246–1266.

Moran, P. A. P. (1950). "Notes on Continuous Stochastic Phenomena." *Biometrika*, 37, 17–23.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). "A multi-resolution Gaussian process model for the analysis of large spatial data sets." *Journal of Computational and Graphical Statistics*, 2, 579–599.

Oehlert, G. (1992). "A note on the delta method." *The American Statistician*, 46, 27 – 29.

Quick, H., Holan, S. H., Wikle, C. K., and Reiter, J. P. (2015). "Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography." *Spatial Statistics*, 14, 439–451.

Ravishanker, N. and Dey, D. K. (2002). *A First Course in Linear Model Theory*. Boca Raton, FL: Chapman and Hall/CRC.

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006). "Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models." *Biometrics*, 62, 1197–1206.

Sang, H. and Huang, J. (2012). "A full-scale approximation of covariance functions for large spatial data sets." *Journal of the Royal Statistical Society: Series B*, 74, 111–132.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society, Series B*, 64, 583–616.

Stein, M. (2014). "Limitations on low rank approximations for covariance matrices of spatial data." *Spatial Statistics*, 8, 1–19.

Tanner, M. A. and Wong, W. H. (1987). "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association*, 82, 528–540.

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

Wakefield, J. and Walker, S. (1999). "Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical Models by Using Auxiliary Variables." *Journal of the Royal Statistical Society*, 82, 331–344.

Wikle, C. K. and Royle, J. A. (2005). "Dynamic design of ecological monitoring networks for nonGaussian spatio-temporal data." *Environmetrics*, 16, 507–522.

Wolpert, R. and Ickstadt, K. (1998). "Poisson/gamma random field models for spatial statistics." *Biometrika*, 85, 251–267.

Zeger, S. L. and Liang, K.-Y. (1991). "Feedback models for discrete and continuous time series." *Statistica Sinica*, 1, 51–64.