Deep Hierarchical Generalized Transformation Models for Spatio-Temporal Data with Discrepancy Errors

Jonathan R. Bradley,[1] Shijie Zhou[2], and Xu Liu[2]

Abstract

"Discrepancy error covariance" refers to the cross-covariance between the signal and the noise terms in an additive model. Traditionally, the signal and noise are assumed independent in additive models to avoid issues with confounding and non-identifiable expressions in the marginal likelihood. This assumption is made even in settings where it is known that discrepancy error covariances exists. Recently, a model has been proposed that allows for discrepancy error covariances that avoids issues with confounding. These models introduce a telescoping sum within the additive model's expression such that the latent process of interest is dependent on other terms of the telescoping sum that are included as part of the noise. However, when evaluating the telescoping sum one obtains signal and noise terms that are independent, which avoids such concerns with confounding. The current model that allows for discrepancy error covariances only includes two terms in this telescoping sum, and consequently, a natural extension is to include more terms within the telescoping sum, which leads to a deep architecture to the statistical model. We refer to this model as the "deep hierarchical generalized transformation" (DHGT) model due to a relationship with the recently introduced hierarchical generalized transformation model. We show that the DHGT is extremely efficient to implement, and can allow for exact Bayesian implementation without the use of MCMC (i.e., we can sample directly from its posterior distribution). We illustrate the DHGT using a simulation and an analysis of the 2017 Haypress wildfire downloaded from the Geospatial Multi-Agency Coordination (GeoMAC) database. These illustrations show that discrepancy errors that arise from common model misspecifications in the spatio-temporal setting can be leveraged to improve prediction.

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

[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

Bayesian models for dependent data often assume that the signal and noise terms in additive models are independent of each other (Cressie, 1993; Gelman et al., 2013; Banerjee et al., 2015; Cressie and Wikle, 2011). However, this assumption is often false. For example, survey statistics estimates are often modified based on disclosure limitations, and can create a type of signal-to-noise dependence (e.g., see Quick et al., 2013). Survey errors such as nonresponse bias (Groves et al., 2001) are often a consequence of the value of the signal. Model misspecification can also induce signal to noise dependence, since the fitted misspecified model is close to the latent process, but the misspecification introduces an unaccounted for error (Bradley et al., 2020). Additionally, several spatial sampling designs naturally lead to discrepancy errors (Wikle and Royle, 2005; Holan and Wikle, 2012).

There are very few models that allow for known discrepancy error covariances in a formal statistical framework. In time-series, there are models that include "leverage effects" in stochastic volatility models (Black, 1976), which assume the volatility to be correlated with the latent process in a particular way. Part of the difficulty with assuming signal-to-noise cross-dependence in a traditional spatio-temporal additive model is that confounding between the marginal signal and noise covariances occur in a hierarchical model (Bradley et al., 2020).

More recently Bradley et al. (2020) developed an approach that makes use of "process augmentation," which is similar to but different from data augmentation (e.g., see Tanner and Wong, 1987; Albert and Chib, 1993; Wakefield and Walker, 1999; Wolpert and Ickstadt, 1998, among others). In particular, a process is introduced into the additive model expression, and cancels out through the use of a telescoping sum to avoid issues with confounding. That is, denote data with Z, two independent random variables with $Y^{(1)}$ and $Y^{(2)}$, and an additive mean-zero independent

error term ε. Then, assume

$$Z = Y^{(2)} + \delta$$
$$\delta = Y^{(1)} - Y^{(2)} + \varepsilon, \tag{1}$$

where notice $Z = Y^{(1)} + \varepsilon$, which is the more traditional independent signal-and-noise additive model, since $Y^{(2)}$ cancels out through the telescoping sum in the above expression. However, $cov(Y^{(2)}, \delta) = cov(Y^{(2)}, Y^{(1)}) - var(Y^{(2)})$, which is not necessarily zero and hence the signal $Y^{(2)}$ is correlated with the noise term δ. Thus, one simultaneously obtains a model for Z that avoids confounding between a signal and a noise term (i.e., between $Y^{(1)}$ and ε), however there is an implied cross covariance between the signal $Y^{(2)}$ and the error δ. Bradley et al. (2020) show that a Bayesian implementation of this process augmentation strategy for discrepancy error modeling is fairly straightforward. First, one samples from the traditional posterior distribution for $Y^{(1)}$, and then samples $Y^{(2)}$ from the distribution for $Y^{(2)}|Y^{(1)}$. This general strategy has been extended to allow for possibly non-Gaussian data in Bradley (2022b) and Nandy et al. (2022). Here, $Y^{(1)}$ and $Y^{(2)}$ can be considered as successive transformations, both of which are aimed at estimating the latent process for the data; hence, this model is referred to as the hierarchical generalized transformation (HGT) model.

The interpretation of the augmented processes $Y^{(1)}$ and $Y^{(2)}$ as successive transformations is similar to the successive transformations used in a neural network, where neural networks refer to the transformations as *activation functions* (Bishop et al., 1995). One main difference between the transformations in an HGT (i.e., $Y^{(1)}$ and $Y^{(2)}$) and the activation functions in a neural network is that HGT treats $Y^{(1)}$ and $Y^{(2)}$ as unknown, whereas activation functions are prespecified and known (e.g., the sigmoid activation function).

Consider Figure 1 for a naive example implementation of an HGT, and let $[Y^{(1)}|Z]$ be the bracket notation for the distribution of $Y^{(1)}$ given Z. In the left panel of Figure 1, we provide the

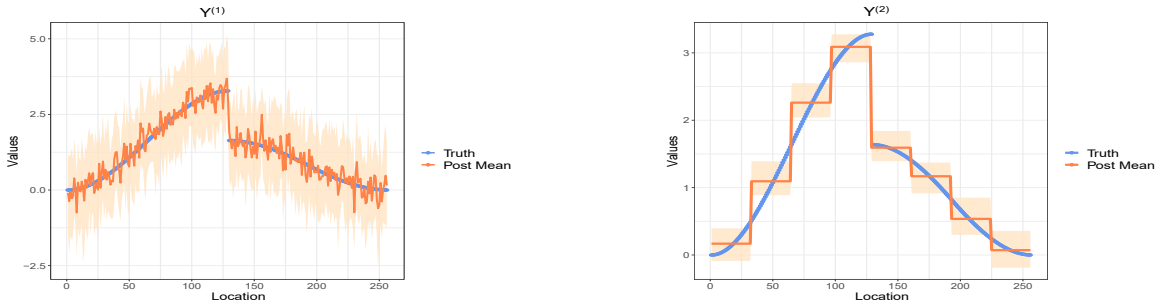

Figure 1: The blue line represents the piecewise polynomial test function, the orange lines indicate the posterior mean, and the transparent region represent 95% credible intervals. The x-axis indicates location s. The left panel displays the posterior mean from $[Y^{(1)}|Z]$, which is chosen to overfit the data. The right panel displays the posterior mean from $[Y^{(2)}|Y^{(1)}]$, which uses Haar wavelets. See Section 4.1 for more details. There are discontinuities in the credible interval in the top right panel due to the use of Haar wavelets.

posterior mean of $Y^{(1)}$ from a posterior distribution $[Y^{(1)}|Z]$ that perfectly interpolates the data. Then in the right panel, we provide the posterior mean of $Y^{(2)}$ from a second layer posterior distribution $[Y^{(2)}|Y^{(1)}]$, which makes use of Haar wavelet basis functions (Novikov et al., 2005). In the HGT, posterior samples from the first layer augmented process $Y^{(1)}$ are used as "new transformed data" in the second layer posterior distribution $[Y^{(2)}|Y^{(1)}]$. Inferences on the latent process are then based on summaries of $Y^{(2)}$ (i.e., $Y^{(1)}$ is marginalized). The implicit assumption here is that samples of $Y^{(1)}$ from $[Y^{(1)}|Z]$ are reasonable to use in place of Z, since samples from the overfitted $[Y^{(1)}|Z]$ are close to Z. That is, we assume that summaries of $Y^{(1)}$ are a reasonable "initial estimate" of the latent process, which motivates its use as new transformed data for inference on $Y^{(2)}$. Then samples of $Y^{(2)}$ are averaged to produce a"second estimate" of the latent process. We say that Figure 1 is a "naive" example, since the posterior mean of $Y^{(2)}$ clearly has problematic discontinuities that arise from using too few Haar wavelets; however, despite this misspecification, the posterior mean of $Y^{(2)}$ more precisely predicts the true mean than the initial overfitted $Y^{(1)}$ in terms of mean squared prediction error.

The first major contribution of this article is to propose a natural extension of the HGT by

increasing the terms in the telescoping sum in the expression of the discrepancy error. This deep representation will lead to J processes $Y^{(1)},\ldots,Y^{(J)}$, that will allow one to incorporate several modeling strategies into a single joint statistical model. For example, one might consider a $Y^{(3)}$ that smooths across the discontinuities in the second panel of Figure 1. Another illustrative example is in the context of modeling wildfire spread, where one could use a spatial basis function expansion model for the logit probability of a fire (e.g.,see Cressie and Johannesson, 2008; Wikle, 2010; Paciorek, 2007, among several others). However, this basis function expansion does not allow for cellular automata (CA) dynamics that are implicit in wildfire spread (Hooten and Wikle, 2010; Quaife and Speer, 2021; Currie et al., 2019; Achtemeier, 2003; Albinet et al., 1986; Duarte, 1997; Wolfram, 1983). By the phrase "CA dynamics," we are referring to the auto-regressive terms in statistical CA models (Hooten and Wikle, 2010) that dictate how the probability of a fire changes over time. Using process augmentation one can incorporate CA dynamics directly into the Bayesian model. This "deep" HGT (DHGT) framework is particularly exciting because it allows one to sequentially consider several models within a single well defined uncertainty quantification Bayesian framework. Moreover, combining these models in this way, as opposed to other existing strategies (e.g., see Bayesian model averaging Raftery et al., 1997), allows one to leverage covariances in the discrepancy error (e.g., $Y^{(1)}-Y^{(2)}$) to improve prediction (e.g., see Bradley et al., 2020, for a result that states if signal-to-noise dependence is present then incorporating discrepancy error covariances leads to smaller mean squared prediction error).

To understand why the DHGT can aid with prediction, it is important to notice that the augmented processes $Y^{(1)},\ldots,Y^{(J-1)}$ are all assumed misspecified, however, $Y^{(j)}$ is specified to be from a "more realistic" model than $Y^{(j-1)}$. By "more realistic" we mean that the assumptions of the j-th Layer Bayesian hierarchical model (BHM) is known to be more realistic in practice than the assumptions of the $(j-1)$-th Layer BHM. Hence, as one sequentially samples from these posterior distributions, one is sequentially sampling from progressively more realistic models. For example, in our application of predicting wildfires, we define the model for $Y^{(1)}$ to assume inde-

pendence and perfectly fit the data in its posterior mean. Since the first layer overfits the data, the original features of the data are not completely smoothed across, and hence, may be reasonable to use as data in the second layer BHM. The model for $Y^{(2)}$ assumes a type of spatio-temporal basis function expansion that ignores temporal dynamics. Although the BHM for $Y^{(2)}$ is misspecified (since temporal dynamics are ignored), spatio-temporal basis function expansion models are arguably more realistic than an overfitted model that assumes the data is independent. Finally the model for $Y^{(3)}$ is specified in a way that incorporates both spatio-temporal basis functions and temporal dynamics, and hence is a more realistic model than $Y^{(2)}$, which ignores temporal dynamics. In this article, we focus on two types of misspecification that arise in spatial and spatio-temporal applications: (1) when too few basis functions are used to model spatial dependencies (Stein, 2014), and (2) ignoring spatio-temporal dynamics (Wikle and Hooten, 2010).

Our second main contribution is that the implied sampler from the Bayesian hierarchical model can be implemented in a dynamic programming fashion. That is, sampling the current layer $Y^{(j)}$ only requires knowledge on the previous layer $Y^{(j-1)}$, but is conditionally independent of $Y^{(k)}$ for $k \leq j-2$. This is a similar property that makes back-propagation in traditional feed-forward neural networks particularly efficient (e.g., see Bishop et al., 1995, for a standard reference). The current deep Bayesian models more often require sophisticated Markov Chain Monte Carlo (MCMC) techniques such as Hamiltonian Monte Carlo (HMC; Neal, 2011) that requires all previous layers $Y^{(1)},\ldots,Y^{(j-1)}$ at each update (i.e., they do not allow for dynamic programming) (e.g., see Papamarkou et al., 2022, for a complete discussion).

Our third main contribution is to introduce how Bayesian implementation can be done exactly without the use of Markov chain Monte Carlo (MCMC) under a certain specification of the DHGT. In particular, MCMC can be avoided when specifying the first augmented process to be a particular conjugate saturated model (as done in Bradley, 2022b) and the remaining augmented processes specified to be a Gaussian mixed effects model with improper priors. This specification leads to an exact sampler through conjugacy (Diaconis and Ylvisaker, 1979). There has been a renewed effort

in the literature to develop general models that allow one to sample directly from the posterior distribution without the use of MCMC. For example, see Gong (2019) who produce an exact procedure to preserve differential privacy, Zhang et al. (2021) for a particular Gaussian model, Bradley (2022a) for an exact procedure using conjugate models, and van Erven and Szabó (2021) for exact Bayesian inference in variable selection. Considering that there has been a boom in approximate Bayesian methodologies that do not require MCMC (e.g., see, Rue et al. (2009), Wainwright et al. (2008), and Katzfuss and Guinness (2021)), exact methods represent an important key future direction of Bayesian analysis.

The remainder of the paper is organized as follows. In Section 2, we introduce the notion of deep discrepancy error covariances and we present the DHGT. Section 3, we describe exact Bayesian inference without the use of MCMC under a particular specification of the DHGT, we refer to as the conjugate DHGT. In Section 4, we illustrate how to leverage discrepancy errors induced by model misspecification. We consider two common misspecifications in spatio-temporal models. In particular, in a simulation study we illustrate the methodology and high predictive performance of our method when discrepancy errors are introduced from a poor specification of basis functions. In an analysis of the 2017 Haypress wildfire, we illustrate the use of discrepancy errors introduced by misspecifying a dynamic model. We end with a discussion in Section 5. For convenience of exposition, proofs of technical results and additional details are provided in the Appendix.

2 Methodology

In Section 2.1, we introduce the notion of deep discrepancy errors. Then in Section 2.2, we introduce the Bayesian hierarchical model (BHM) that allows for deep discrepancy errors, which we call DHGT. Section 2.3 discusses the posterior distribution for the DHGT.

2.1 Deep Discrepancy Errors for Spatio-Temporal Data

We introduce space and time into the notation. Let $\mathbf{s} \in D \subset \mathbb{R}^d$, where D is a spatial domain and $t = 1, \ldots, T$ indexes discrete time. Consider augmenting with J possibly dependent random processes $Y^{(1)}, \ldots, Y^{(J)}$ as follows:

$$
\begin{aligned}
& g\left[E\left\{Z_t(\mathbf{s}) | Y_t^{(1)}(\mathbf{s}), \ldots, Y_t^{(J)}(\mathbf{s})\right\}\right] = Y_t^{(J)}(\mathbf{s}) + \delta_t(\mathbf{s}) \\
& \delta_t(\mathbf{s}) = \sum_{j=1}^{J-1} \left\{Y_t^{(j)}(\mathbf{s}) - Y_t^{(j+1)}(\mathbf{s})\right\}; \ \mathbf{s} \in D, \ t = 1, \ldots, T,
\end{aligned} \tag{2}
$$

where g is an appropriate link function used in generalized linear models (GLM; McCullagh and Nelder, 1989). We have that $g\left[E\left\{Z_t(\mathbf{s}) | Y_t^{(1)}(\mathbf{s}), \ldots, Y_t^{(J)}(\mathbf{s})\right\}\right] = Y_t^{(1)}(\mathbf{s})$, which, similar to (1), arises due to the telescoping nature of $\delta_t(\mathbf{s})$. For prediction of the latent process, we use summaries of $Y_t^{(J)}(\cdot)$, which effectively filters out the error $\delta_t(\cdot)$. The main benefit of introducing augmented processes is that the cross-covariance between the signal and the noise is not necessarily zero, and hence, can be leveraged to improve predictions. That is,

$$
\begin{aligned}
\operatorname{cov}\left\{Y_t^{(J)}(\mathbf{s}), \delta_t(\mathbf{u})\right\} &= \sum_{j=1}^{J-1} \operatorname{cov}\left\{Y_t^{(J)}(\mathbf{s}), Y_t^{(j)}(\mathbf{u}) - Y_t^{(j+1)}(\mathbf{u})\right\} \\
&= \sum_{j=1}^{J-1} \operatorname{cov}\left\{Y_t^{(J)}(\mathbf{s}), Y_t^{(j)}(\mathbf{u})\right\} - \operatorname{cov}\left\{Y_t^{(J)}(\mathbf{s}), Y_t^{(j+1)}(\mathbf{u})\right\} \\
&= \operatorname{cov}\left\{Y_t^{(J)}(\mathbf{s}), Y_t^{(1)}(\mathbf{u}) - Y_t^{(J)}(\mathbf{u})\right\}; \ \mathbf{s}, \mathbf{u} \in D,
\end{aligned}
$$

so that the covariance between the "signal" $Y_t^{(J)}(\mathbf{s})$ and the "noise" $\delta_t(\mathbf{u})$, is equivalent to the covariance between the signal $Y_t^{(J)}(\mathbf{s})$ and the error $Y_t^{(1)}(\mathbf{u}) - Y_t^{(J)}(\mathbf{u})$, which is not necessarily zero. Recall, the augmented processes $Y^{(1)}, \ldots, Y^{(J-1)}$ are interpreted as misspecified, however, $Y^{(j+1)}$ is specified to be from a more realistic model than $Y^{(j)}$. From this perspective the discrepancy error $Y^{(j)} - Y^{(j+1)}$ represents model misspecification error.

Bradley et al. (2020) showed that if discrepancy error covariances are known to be present then one can obtain more precise predictions (i.e., smaller mean squared prediction error). This is not surprising since in general, it is well-known that incorporating a known covariance (e.g., see Cressie, 1993, for examples in spatial statistics) into a statistical model can lead to more precise predictions (i.e., they can be leveraged). The main difference between this model for cross signal to noise covariance and that in Bradley et al. (2020) is that we can leverage additional covariances when $J > 2$. In particular, we assume $\text{cov}\left\{Y_t^{(j)}(\mathbf{s}), Y_t^{(j-1)}(\mathbf{u})\right\}$ to be not necessarily zero for each j and $J > 2$.

2.2 Hierarchical Models with Deep Discrepancy Errors: Deep Hierarchical Generalized Transformation Models

Let h, p and π represent generic probability mass functions (pmf) or probability density functions (pdf). Recall that a traditional BHM can be written as the product of a "data model," "process model," and "parameter model" sometimes written as (e.g., see Cressie and Wikle, 2011, for a standard reference),

$$\begin{aligned} &\text{Data Model}: \ h(\mathbf{z}|\mathbf{y}^{(1)}, \boldsymbol{\theta}^{(1)}) \\ &\text{Process Model}: \ p(\mathbf{y}^{(1)}|\boldsymbol{\theta}^{(1)}) \\ &\text{Parameter Model}: \ \pi(\boldsymbol{\theta}^{(1)}), \end{aligned} \tag{3}$$

where the n-dimensional vector of observed data $\mathbf{z} = \{Z_t(\mathbf{s}_{it}) : i = 1,\ldots,n_t, t = 1,\ldots T\}'$, $\mathbf{s}_{it} \in D$ is the i-th observation at time t, the n-dimensional vector $\mathbf{y}^{(j)} = \{Y_1^{(j)}(\mathbf{s}_{it}) : i = 1,\ldots,n_t, t = 1,\ldots T\}'$ is the vectorized augmented process in (2), n_t is the number of observed data at time t, T is the number of observed time points, $n = \sum_{t=1}^{T} n_t$, and $\boldsymbol{\theta}^{(j)} \in \Omega_j$ is a generic real-valued vector, for $j = 1,\ldots,J$. Note that one can allow elements of $\boldsymbol{\theta}^{(1)}$ to represent random effects, in which case (3) can easily be modified to include those elements in the process model. Now, consider what we

call the "deep hierarchical generalized transformation (DHGT) model," which is defined to be the product of the following:

$$
\begin{aligned}
&\text{First Layer BHM}: \; h(\mathbf{z}|\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)})p(\boldsymbol{y}^{(1)}|\boldsymbol{\theta}^{(1)})\pi(\boldsymbol{\theta}^{(1)}) \\
&\text{Second Layer BHM}: \; h(\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)}|\boldsymbol{y}^{(2)},\boldsymbol{\theta}^{(2)})p(\boldsymbol{y}^{(2)}|\boldsymbol{\theta}^{(2)})\pi(\boldsymbol{\theta}^{(2)}) \; \frac{1}{m_2(\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)})} \\
&\text{Third Layer BHM}: \; h(\boldsymbol{y}^{(2)},\boldsymbol{\theta}^{(2)}|\boldsymbol{y}^{(3)},\boldsymbol{\theta}^{(3)})p(\boldsymbol{y}^{(3)}|\boldsymbol{\theta}^{(3)})\pi(\boldsymbol{\theta}^{(3)}) \; \frac{1}{m_3(\boldsymbol{y}^{(2)},\boldsymbol{\theta}^{(2)})}, \\
&\vdots \\
&J-\text{th Layer BHM}: \; h(\boldsymbol{y}^{(J-1)},\boldsymbol{\theta}^{(J-1)}|\boldsymbol{y}^{(J)},\boldsymbol{\theta}^{(J)})p(\boldsymbol{y}^{(J)}|\boldsymbol{\theta}^{(J)})\pi(\boldsymbol{\theta}^{(J)}) \; \frac{1}{m_J(\boldsymbol{y}^{(J-1)},\boldsymbol{\theta}^{(J-1)})}, \qquad (4)
\end{aligned}
$$

where we assume each pdf/pmf is proper, and the term

$$
m_j(\boldsymbol{y}^{(j-1)},\boldsymbol{\theta}^{(j-1)}) = \int\int h(\boldsymbol{y}^{(j-1)},\boldsymbol{\theta}^{(j-1)}|\boldsymbol{y}^{(j)},\boldsymbol{\theta}^{(j)})p(\boldsymbol{y}^{(j)}|\boldsymbol{\theta}^{(j)})\pi(\boldsymbol{\theta}^{(j)})d\boldsymbol{y}^{(j)}d\boldsymbol{\theta}^{(j)}, \;\; j \geq 2,
$$

guarantees that the joint statistical model is proper; this can be verified by successively integrating out $\boldsymbol{y}^{(J)},\boldsymbol{\theta}^{(J)},\ldots,\boldsymbol{y}^{(2)},\boldsymbol{\theta}^{(2)},\boldsymbol{y}^{(1)},\;\boldsymbol{\theta}^{(1)}$, and $\mathbf{z}$ in the expression of the joint distribution formed by (4). The recently proposed HGT is a special case of the DHGT in (4), which occurs when $J = 2$. Thus, for $J\,(>2)$, the model in (4) is considered "deep." The term $\boldsymbol{y}^{(1)}$ is on the right-hand-side of the "|" symbol in the First Layer BHM, and $\boldsymbol{y}^{(1)}$ is on the left-hand-side of the "|" symbol in the Second Layer BHM. In general, the only way we can change the order of conditional probabilities is through Bayes rule (e.g., see Gelman et al., 2013, among others). Hence, the need for the terms $\frac{1}{m_j(\boldsymbol{y}^{(j-1)},\boldsymbol{\theta}^{(j-1)})}$, which can be interpreted as applications of "Bayes rules," however nested within a single hierarchical model.

Consider the j-th Layer BHM with $j \geq 2$ for discussion,

$$
\begin{aligned}
&j-\text{th Layer Transformed Data Model}: \frac{h(\boldsymbol{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)}|\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)})}{m_j(\boldsymbol{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)})} \\
&j-\text{th Layer Process Model}: \ p(\boldsymbol{y}^{(j)}|\boldsymbol{\theta}^{(j)}) \\
&j-\text{th Layer Parameter Model}: \ \pi(\boldsymbol{\theta}^{(j)}).
\end{aligned} \tag{5}
$$

Notice that in (5), $\boldsymbol{y}^{(j-1)}$ and $\boldsymbol{\theta}^{(j-1)}$ are treated as (transformed) data for the processes and parameters $\boldsymbol{y}^{(j)}$ and $\boldsymbol{\theta}^{(j)}$. Thus, within the DHGT's layers we sequentially treat $(\boldsymbol{y}^{(1)}, \boldsymbol{\theta}^{(1)}), \ldots, (\boldsymbol{y}^{(J-1)}, \boldsymbol{\theta}^{(J-1)})$ as new transformed data. Depending on the specifications of (5), $\text{cov}(\boldsymbol{y}^{(j)}, \boldsymbol{y}^{(j-1)})$ is not necessarily zero, which allows one to leverage these covariances to possibly improve predictions (in terms of mean squared prediction error) provided they are present.

2.3 The Posterior Distribution for the Deep Hierarchical Generalized Transformation Model

The posterior distribution for the DHGT model in (4) is given by (see the Appendix for details),

$$
f(\boldsymbol{y}^{(1)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{y}^{(J)}, \boldsymbol{\theta}^{(J)}|\mathbf{z}) = f(\boldsymbol{y}^{(1)}, \boldsymbol{\theta}^{(1)}|\mathbf{z}) \left\{ \prod_{j=2}^{J} f(\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)}|\boldsymbol{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)}) \right\}. \tag{6}
$$

where $f(\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)}|\boldsymbol{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)})$ is the j-th Layer BHM's posterior distribution,

$$
f(\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)}|\boldsymbol{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)}) = \frac{h(\boldsymbol{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)}|\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)}) p(\boldsymbol{y}^{(j)}|\boldsymbol{\theta}^{(j)}) \pi(\boldsymbol{\theta}^{(j)})}{\int\int h(\boldsymbol{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)}|\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)}) p(\boldsymbol{y}^{(j)}|\boldsymbol{\theta}^{(j)}) \pi(\boldsymbol{\theta}^{(j)}) d\boldsymbol{y}^{(j)} d\boldsymbol{\theta}^{(j)}}.
$$

The distribution $f(\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)}|\boldsymbol{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)})$ is the posterior distribution of the j-th layer BHM that treats $\boldsymbol{y}^{(j-1)}$ and $\boldsymbol{\theta}^{(j-1)}$ as new transformed data. Again, this shows that within the DHGT's layers we sequentially treat $(\boldsymbol{y}^{(1)}, \boldsymbol{\theta}^{(1)}), \ldots, (\boldsymbol{y}^{(J-1)}, \boldsymbol{\theta}^{(J-1)})$ as new transformed data for the next layer. This leads to an extremely efficient (relative to other deep BHMs) dynamic sampling procedure,

Algorithm 1 Step-by-step procedure for sampling from the DHGT posterior distribution given in Equation (6).

1: Set $b = 1$ and initialize $\boldsymbol{y}^{(j)}$, $\boldsymbol{\theta}^{(j)}$, and independent posterior predictive samples $\boldsymbol{y}_{new}^{(j)}$ with $\boldsymbol{y}^{(j)[0]}$, $\boldsymbol{\theta}^{(j)[0]}$, and $\boldsymbol{y}_{new}^{(j)[0]}$.
2: Sample $\boldsymbol{y}^{(1)[b]}$ and $\boldsymbol{\theta}^{(1)[b]}$ from $f(\boldsymbol{y}^{(1)}, \boldsymbol{\theta}^{(1)}|\mathbf{z})$.
3: Sample $\boldsymbol{y}^{(j)[b]}$ and $\boldsymbol{\theta}^{(j)[b]}$ from $f(\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)}|\boldsymbol{y}^{(j-1)[b]}, \boldsymbol{\theta}^{(j-1)[b]})$ for $j = 2, \ldots, J$, which is the posterior distribution associated with the j-th BHM that treats $\boldsymbol{y}^{(j-1)[b]}$ and $\boldsymbol{\theta}^{(j-1)[b]}$ as new transformed data.
4: Generate new posterior predictive samples of the mT_P-dimensional vector $\boldsymbol{y}_{new}^{(j)[b]}$ from $h(\cdot|\boldsymbol{y}^{(j)[b]}, \boldsymbol{\theta}^{(j)[b]})$, which is defined in the j-th Layer BHM. The elements of $\boldsymbol{y}_{new}^{(j)[b]}$ are stacked over times $t = 1 \ldots, T_P$, T_P is the number of time points that are predicted, and the set $D_P \subset D$ consists of a collection of $m \geq n$ pre-specified prediction locations.
5: Set $b = b + 1$.
6: Repeat Steps 2$-$6 until $b = B$ for a prespecified value of B.

which is presented in Algorithm 1.

Statistical inference is based on summaries of $\boldsymbol{y}_{new}^{(J)}$. The posterior mean (and variance) of $\boldsymbol{y}_{new}^{(J)}$ is estimated by averaging (and taking the variance) across b in Step 4, and are used for prediction of the latent process. Using the posterior mean of $\boldsymbol{y}_{new}^{(J)}$ for inference effectively filters out the discrepancy error $\delta_t(\mathbf{s})$ introduced in Equation (2). The steps in Algorithm 1 suggests that the summaries of the process $Y_t^{(j)}(\mathbf{s})$ represents our "j-th estimate" of the value of the latent process, and samples of $Y_t^{(j)}(\mathbf{s})$ will be used as new transformed data at the $(j+1)$-th level BHM. By marginalizing across $Y_t^{(j)}(\mathbf{s})$ for $1 \leq j < J$, we are deciding to keep our final J-th estimate for inference.

Traditional neural network models use dynamic computing (e.g., the output from one layer is passed as input to the next in backward propagation). This sampler has a similar propagation pattern (i.e., the output is used as input in Steps 2 and 3), which arises by the expression of the posterior distribution in (6). Many deep Bayesian models don't have this dynamic computing feature, and thus, can be difficult to implement (e.g., see Papamarkou et al., 2022, for a complete discussion).

Another important feature of the Algorithm in Steps $1-6$ is that the normalizing constants $\frac{1}{m_j(\boldsymbol{y}^{(j-1)},\boldsymbol{\theta}^{(j-1)})}$ never needs to be computed. This is simply because in Equation (6), each layer in the hierarchy draws from a conditional distribution that we will have in closed form. Thus, valid posterior inference using the DHGT only requires the conditional distributions $f(\boldsymbol{y}^{(j)},\boldsymbol{\theta}^{(j)}|\boldsymbol{y}^{(j-1)},\boldsymbol{\theta}^{(j-1)})$ to be well-defined densities or masses.

3 Conjugate DHGT: Exact Bayesian Inference without MCMC

In this section, we provide guidance on a specific class of DHGTs one might consider in practice in more complicated spatio-temporal settings. We refer to this class of DHGTs as the "conjugate DHGT". A motivating feature of the conjugate DHGT is that one can sample directly from the exact posterior distribution in (6) without the use of MCMC. The strategy to obtain an exact sampler is to develop the use of conjugate priors (Diaconis and Ylvisaker, 1979), which can be sampled from directly, in this DHGT framework. The formal specification of the levels of the conjugate DHGT are provided in Sections 3.1 and 3.2.

3.1 First Layer Bayesian Hierarchical Model: Univariate Conjugate Models

The first goal of this section is to define $h(\mathbf{z}|\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)})$, $p(\boldsymbol{y}^{(1)}|\boldsymbol{\theta}^{(1)})$, and $\pi(\boldsymbol{\theta}^{(1)})$ in the First Layer BHM of the conjugate DHGT. The second goal of this section is to provide the First Layer posterior distribution $f(\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)}|\mathbf{z})$ for the conjugate DHGT.

The specification of the First Layer BHM is motivated by the goodness-of-fit of the DHGT. In general, one should assess the goodness-of-fit of a model, and determine if the model oversmooths (i.e., model outputs are too far from the data) or overfits (i.e., model outputs are too close to the data) (Gelman, 1996). This is important when considering the procedure to implement a DHGT, which involves J layers of nested smoothing. Thus, to avoid oversmoothing, we specify $Y^{(1)}$ to overfit the data Z so that samples from $[Y^{(2)}|Y^{(1)}]$ smooths overfitted values (as opposed to

smoothing smoothed values).

We use the same specification of the First Layer BHM used in the HGT introduced by Bradley (2022b). Here, the First Layer BHM is specified to be a saturated model (i.e., there as many parameters as observations). A useful by-product of this perspective is that the conjugate saturated model can be implemented without the use of MCMC. Let $Z_t(\mathbf{s})|Y_t^{(1)}(\mathbf{s})$ be distributed from the exponential family,

$$h(Z_t(\mathbf{s})|Y_t^{(1)}(\mathbf{s}), b_t(\mathbf{s})) = \exp\left\{Z_t(\mathbf{s})Y_t^{(1)}(\mathbf{s}) - b_t(\mathbf{s})\psi(Y_t^{(1)}(\mathbf{s})) + c(Z_t(\mathbf{s}))\right\};\ Z_t(\mathbf{s}) \in \mathscr{Z}, Y_t^{(1)}(\mathbf{s}) \in \mathscr{Y}, \tag{7}$$

where $Z_t(\mathbf{s})$ is conditionally independent of all other processes and parameters given $Y_t^{(1)}(\mathbf{s})$ and $b_t(\mathbf{s})$, $\mathscr{Z}$ is the support of the data, $\mathscr{Y}$ is the support of the process, $b_t(\mathbf{s})$ is a real-valued parameter, both $\psi(\cdot)$ and $c(\cdot)$ are known real-valued functions, and $b_t(\mathbf{s})\psi(\cdot)$ is the log partition function (Lehmann and Casella, 1998). One can also allow for the case where $b_t(\cdot)$ is unknown (e.g., Gaussian with unknown variance). When the univariate conjugate prior distribution exists it is equal to (Diaconis and Ylvisaker, 1979),

$$p(Y_t^{(1)}(\mathbf{s})|\alpha,\kappa) = \mathscr{N}(\alpha,\kappa)\exp\left\{\alpha Y_t^{(1)}(\mathbf{s}) - \kappa\psi(Y_t^{(1)}(\mathbf{s}))\right\};\ Y_t^{(1)}(\mathbf{s}) \in \mathscr{Y}, \frac{\alpha}{\kappa} \in \mathscr{Z}, \kappa > 0, \tag{8}$$

where $\mathscr{N}(\alpha,\kappa)$ is a normalizing constant, and we use the shorthand $Y_t^{(1)}(\mathbf{s})|\alpha,\kappa \sim \text{DY}(\alpha,\kappa)$. The DY distribution is straightforward to simulate from when applied to Gaussian, Poisson, and binomial data, and simply involves simulating from a univariate Gaussian, the log of a univariate gamma, and the logit of a univariate beta distribution, respectively (e.g.,see Bradley, 2022b, for more details). From (7) and (8), we have the posterior is given by

$$Y_t^{(1)}(\mathbf{s})|Z_t(\mathbf{s}),\alpha,\kappa \sim \text{DY}\left(\alpha + Z_t(\mathbf{s}), \kappa + b_t(\mathbf{s})\right),$$

where α and κ are chosen to overfit the data to avoid over-smoothing $\boldsymbol{y}^{(J)}$ when sampling from the

posterior distribution in Equation (6). Summaries from the First Layer BHM can be interpreted as crude initial estimates, which will be improved (in terms of prediction error) by the next augmented process $Y^{(2)}$.

Let

$$\begin{aligned}
\boldsymbol{\beta}^{(1)}|\sigma^2_\beta &\sim N(\mathbf{0}_\ell, \sigma^2_\beta \mathbf{I}_\ell) \\
\boldsymbol{\eta}^{(1)}|\sigma^2_\eta &\sim N(\mathbf{0}_r, \sigma^2_\eta \mathbf{I}_r) \\
\boldsymbol{\xi}^{(1)}|\sigma^2_\xi &\sim N(\mathbf{0}_n, \sigma^2_\xi \mathbf{I}_n),
\end{aligned} \tag{9}$$

where σ^2_β, σ^2_η, and σ^2_ξ are given an inverse-gamma priors, $\mathbf{0}_\ell$ is a ℓ-dimensional vector of zeros, $\mathbf{I}_\ell$ is a $\ell \times \ell$ identity matrix, and $\boldsymbol{\theta}^{(1)} = (\boldsymbol{\beta}^{(1)\prime}, \boldsymbol{\eta}^{(1)\prime}, \boldsymbol{\xi}^{(1)\prime})'$. One can consider different prior and process model specifications for $\boldsymbol{\theta}^{(1)}$ provided that the specifications are proper.

To sample from the First Layer BHM's posterior distribution (i.e., $f(\mathbf{y}^{(1)}, \boldsymbol{\theta}^{(1)}|\mathbf{z})$) we sample $Y_t^{(1)}(\mathbf{s}_{it})$ independently (and directly) from the conjugate distribution DY $(\alpha + Z_t(\mathbf{s}), \kappa + b_t(\mathbf{s}))$ and sample $\boldsymbol{\theta}^{(1)}$ from its distribution $\pi(\boldsymbol{\theta}^{(1)})$. As described in Section 2.3, we marginalize across $Y_t^{(1)}(\mathbf{s}), \ldots, Y_t^{(J-1)}(\mathbf{s})$, and inference on the latent process is done using summaries of $Y_t^{(J)}(\mathbf{s})$. This is particularly important considering $Y_t^{(1)}(\mathbf{s})$ is saturated and will naturally overfit. In the Appendix, we provide some additional technical details on the marginalized DHGT.

Spatio-temporal BHMs often assume $Z_t(\mathbf{s})$ is conditionally independent given $Y_t^{(1)}(\mathbf{s})$ and $Y_t^{(1)}(\mathbf{s})$ is assumed correctly specified and dependent across space and time. Then, upon marginalizing $Y_t^{(1)}(\mathbf{s})$ we obtain data that is dependent over space and time (see Cressie and Wikle, 2011, for a standard reference on this use of conditional independence). However, the process augmentation framework is different from a standard BHM because we include $Y_t^{(1)}(\mathbf{s}), \ldots, Y_t^{(J)}(\mathbf{s})$ for $J > 1$ with $Y_t^{(1)}(\mathbf{s}), \ldots, Y_t^{(J-1)}(\mathbf{s})$ assumed misspecified. To see how dependence is modeled in the data consider the case where $Z_t(\mathbf{s})$ is normally distributed with $\psi(Y_t^{(1)}(\mathbf{s})) = Y_t^{(1)}(\mathbf{s})^2$ and $b_t(\mathbf{s}) = \frac{1}{2\sigma^2_Z}$ with $\sigma^2_Z > 0$. Let $\alpha = 0$ and $\kappa = \frac{1}{2\sigma^2_Y}$ with $\sigma^2_Y > 0$. It follows that that the b-th first layer posterior

sample from $f(Y_t(\mathbf{s})|Z_t^{(1)}(\mathbf{s}),\alpha,\kappa)$ is given by (Gelman et al., 2013),

$$Y_t^{(1)[b]}(\mathbf{s}) = \frac{\sigma_Y^2}{\sigma_Y^2+\sigma_Z^2}Z_t(\mathbf{s}) + \left(\frac{\sigma_Y^2\sigma_Z^2}{\sigma_Y^2+\sigma_Z^2}\right)^{1/2}\varepsilon_Y^{(1)[b]}, \tag{10}$$

where $\varepsilon_Y^{(1)[b]}$ is independently drawn from a standard normal distribution. As $\sigma_Z^2 \to 0$ then $Y_t^{(1)[b]}(\mathbf{s})$ converges to $Z_t(\mathbf{s})$ almost surely. Thus, when $\sigma_Z^2 \approx 0$ we have that $Y_t^{(1)[b]}(\mathbf{s}) \approx Z_t(\mathbf{s})$ so that the new transformed data is equivalent to original data in this limiting case. Thus, when $\sigma_Z^2 \approx 0$ spatio-temporal dependence of the data is effectively modeled in the second layer BHM. This is another reason why overfitting in the first layer is particularly important.

3.2 Remaining Layer Bayesian Hierarchical Models: Spatio-Temporal Mixed Effects Models

The first goal of this section is to define $h(\boldsymbol{y}^{(j-1)},\boldsymbol{\theta}^{(j-1)}|\boldsymbol{y}^{(j)},\boldsymbol{\theta}^{(j)})$, $p(\boldsymbol{y}^{(j)}|\boldsymbol{\theta}^{(j)})$, and $\pi(\boldsymbol{\theta}^{(j)})$ in the j-th Layer BHM of the conjugate DHGT ($j \geq 2$). The second goal of this section is to provide the j-th Layer posterior distribution $h(\boldsymbol{y}^{(j)},\boldsymbol{\theta}^{(j)}|\boldsymbol{y}^{(j-1)},\boldsymbol{\theta}^{(j-1)})$ for the conjugate DHGT.

Assume that $(\mathbf{y}^{(j-1)\prime},\boldsymbol{\beta}^{(j-1)\prime},\boldsymbol{\eta}^{(j-1)\prime},\boldsymbol{\xi}^{(j-1)\prime})'$ is new transformed data for $(\mathbf{y}^{(j)\prime},\boldsymbol{\beta}^{(j)\prime},\boldsymbol{\eta}^{(j)\prime},\boldsymbol{\xi}^{(j)\prime})'$ according to the following model

$$h(\boldsymbol{y}^{(j-1)},\boldsymbol{\theta}^{(j-1)}|\boldsymbol{y}^{(j)},\boldsymbol{\theta}^{(j)}) = Normal\left(\begin{pmatrix}\boldsymbol{y}^{(j)}\\ \boldsymbol{\beta}^{(j)}\\ \boldsymbol{\eta}^{(j)}\\ \boldsymbol{\xi}^{(j)}\end{pmatrix}, \sigma^2\mathbf{I}_{2n+\ell+r}\right);\ j \geq 2, \tag{11}$$

where $Normal(\cdot,\cdot)$ is a shorthand for the normal distribution and $\boldsymbol{\theta}^{(j)} = (\boldsymbol{\xi}^{(j)\prime},\boldsymbol{\beta}^{(j)\prime},\boldsymbol{\eta}^{(j)\prime})'$. Define $\mathbf{y}^{(j)}$ to be a spatio-temporal mixed effects model with large-scale, small-scale, and fine-scale terms

(e.g., see Cressie and Wikle, 2011, among others) so that

$$\boldsymbol{y}^{(j)} = \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)} + \mathbf{G}^{(j)}\boldsymbol{\eta}^{(j)} + \boldsymbol{\xi}^{(j)}, \tag{12}$$

and $p(\mathbf{y}^{(j)}|\boldsymbol{\theta}^{(j)}) = \delta\left\{\boldsymbol{y}^{(j)}, \begin{pmatrix} \mathbf{I}_n & \mathbf{X}^{(j)} & \mathbf{G}^{(j)} \end{pmatrix}\boldsymbol{\theta}^{(j)}\right\}$, where $\mathbf{X}^{(j)}$ be a $n \times \ell$ matrix of covariates, $\mathbf{G}^{(j)}$ be a $n \times r$ matrix of spatio-temporal basis functions (Wikle, 2010), and δ is the Dirac delta function. That is, $\delta\left\{\boldsymbol{y}^{(j)}, \begin{pmatrix} \mathbf{I}_n & \mathbf{X}^{(j)} & \mathbf{G}^{(j)} \end{pmatrix}\boldsymbol{\theta}^{(j)}\right\}$ is equal to one when $\boldsymbol{y}^{(j)}$ equals $\begin{pmatrix} \mathbf{I}_n & \mathbf{X}^{(j)} & \mathbf{G}^{(j)} \end{pmatrix}\boldsymbol{\theta}^{(j)}$ and is zero otherwise. Substituting (12) into (11) leads to,

$$\begin{aligned}
\begin{pmatrix} \boldsymbol{y}^{(j-1)} \\ \boldsymbol{\beta}^{(j-1)} \\ \boldsymbol{\eta}^{(j-1)} \\ \boldsymbol{\xi}^{(j-1)} \end{pmatrix} &= \begin{pmatrix} \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)} + \mathbf{G}^{(j)}\boldsymbol{\eta}^{(j)} + \boldsymbol{\xi}^{(j)} + \boldsymbol{\varepsilon}_y^{(j)} \\ \boldsymbol{\beta}^{(j)} + \boldsymbol{\varepsilon}_\beta^{(j)} \\ \boldsymbol{\eta}^{(j)} + \boldsymbol{\varepsilon}_\eta^{(j)} \\ \boldsymbol{\xi}^{(j)} + \boldsymbol{\varepsilon}_\xi^{(j)} \end{pmatrix} \\
&= \underbrace{\begin{pmatrix} \mathbf{I}_n & \mathbf{X}^{(j)} & \mathbf{G}^{(j)} \\ \mathbf{0}_{p,n} & \mathbf{I}_p & \mathbf{0}_{p,r} \\ \mathbf{0}_{r,n} & \mathbf{0}_{r,p} & \mathbf{I}_r \\ \mathbf{I}_n & \mathbf{0}_{n,p} & \mathbf{0}_{n,r} \end{pmatrix}}_{\mathbf{H}^{(j)}} \underbrace{\begin{pmatrix} \boldsymbol{\xi}^{(j)} \\ \boldsymbol{\beta}^{(j)} \\ \boldsymbol{\eta}^{(j)} \end{pmatrix}}_{\boldsymbol{\theta}^{(j)}} + \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}_y^{(j)} \\ \boldsymbol{\varepsilon}_\beta^{(j)} \\ \boldsymbol{\varepsilon}_\eta^{(j)} \\ \boldsymbol{\varepsilon}_\xi^{(j)} \end{pmatrix}}_{\boldsymbol{\varepsilon}^{(j)}} \\
&= \mathbf{H}^{(j)}\boldsymbol{\theta}^{(j)} + \boldsymbol{\varepsilon}^{(j)}; \ j \geq 2,
\end{aligned} \tag{13}$$

which is simply a (structured) regression model with $\boldsymbol{\varepsilon}^{(j)} \sim Normal(\mathbf{0}_{2n+\ell+r}, \sigma^2\mathbf{I}_{2n+\ell+r})$. We assume an improper "flat" model on $\boldsymbol{\theta}^{(j)}$ (i.e., $\pi(\boldsymbol{\theta}^{(j)}) = 1$) in the j-th Layer BHM for $j \geq 2$. However, more informative priors can be used in the initial BHM to define the model for $\boldsymbol{\theta}^{(1)}$ if available (see Section 3.1).

A natural limitation to the DHGT is that the *marginal process model* for $\mathbf{y}^{(J)}$ (starting with the joint distribution implied by (4)) consists of unknown integrals, and hence can not be interpreted directly. See the Appendix for an expression for the marginal process model for $\mathbf{y}^{(J)}$. However, the improper model specification for $j \geq 2$ allows us to have a heuristic interpretation on a limiting case. In particular, when σ^2 goes to zero, Equation (12) shows that for each $j \geq 2$, $\boldsymbol{\beta}^{(j)}$, $\boldsymbol{\eta}^{(j)}$, and $\boldsymbol{\xi}^{(j)}$ is almost surely equal to $\boldsymbol{\beta}^{(j-1)}$, $\boldsymbol{\eta}^{(j-1)}$, and $\boldsymbol{\xi}^{(j-1)}$ in the limit, where recall $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\eta}^{(1)}$, and $\boldsymbol{\xi}^{(1)}$ are given an interpretable and standard specifications in Section 3.1. Consequently, we choose σ^2 to be "small" fixed value.

Using a straightforward complete the squares argument we have that the j-th layer BHM's posterior distribution is given by,

$$
\begin{aligned}
f(\boldsymbol{\theta}^{(j)}|\boldsymbol{y}^{(j-1)},\boldsymbol{\theta}^{(j-1)}) &= Normal\left((\mathbf{H}^{(j)\prime}\mathbf{H}^{(j)})^{-1}\mathbf{H}^{(j)\prime} \begin{pmatrix} \boldsymbol{y}^{(j-1)} \\ \boldsymbol{\beta}^{(j-1)} \\ \boldsymbol{\eta}^{(j-1)} \\ \boldsymbol{\xi}^{(j-1)} \end{pmatrix}, \sigma^2(\mathbf{H}^{(j)\prime}\mathbf{H}^{(j)})^{-1} \right) \\
f(\boldsymbol{y}^{(j)}|\boldsymbol{\theta}^{(j)},\boldsymbol{y}^{(j-1)},\boldsymbol{\theta}^{(j-1)}) &= \delta\left\{ \boldsymbol{y}^{(j)}, \begin{pmatrix} \mathbf{I}_n & \mathbf{X}^{(j)} & \mathbf{G}^{(j)} \end{pmatrix} \boldsymbol{\theta}^{(j)} \right\}. \qquad (14)
\end{aligned}
$$

It follows that the b-sample from the j-th layer BHM's posterior distribution in (14) is given by

$$
\begin{aligned}
\boldsymbol{\theta}^{(j)[b]} &= (\mathbf{H}^{(j)\prime}\mathbf{H}^{(j)})^{-1}\mathbf{H}^{(j)\prime} \left\{ \begin{pmatrix} \boldsymbol{y}^{(j-1)} \\ \boldsymbol{\beta}^{(j-1)} \\ \boldsymbol{\eta}^{(j-1)} \\ \boldsymbol{\xi}^{(j-1)} \end{pmatrix} + \sigma\boldsymbol{\varepsilon}^{[b]} \right\}, \\
\mathbf{y}^{(j)[b]} &= \begin{pmatrix} \mathbf{I}_n & \mathbf{X}^{(j)} & \mathbf{G}^{(j)} \end{pmatrix} \boldsymbol{\theta}^{(j)[b]}, \qquad (15)
\end{aligned}
$$

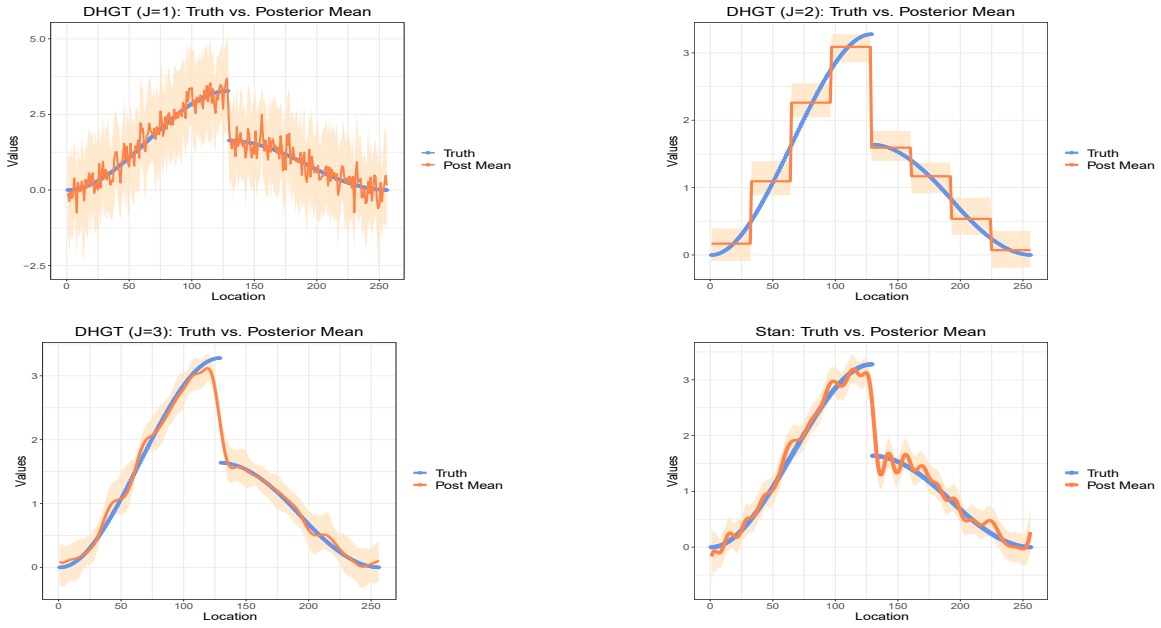

Figure 2: The blue line represents the piecewise polynomial test function, the orange lines indicate the posterior mean, and the transparent region represent 95% credible intervals. The x-axis indicates location s. The title indicates which model is being summarized. See Section 4.1 for more details. There are discontinuities in the credible interval in the top right panel due to the use of Haar wavelets.

where $\boldsymbol{\varepsilon}^{[b]}$ is a $(2n+\ell+r)$-dimensional random vector consisting of iid standard normal random variables. The expression in (15) with $\sigma = 0$ is equivalent to the recently proposed Exact Posterior Regression (EPR) from (Bradley, 2022a), which was derived analytically from a Bayesian generalized linear mixed effects model on a single latent process. The results in Bradley (2022a) allow one to compute (15) efficiently by only inverting $\ell \times \ell$ matrices, $r \times r$ matrices and $n \times n$ diagonal matrices. We make use of these matrix inversion formulas in Section 4.

4 Spatial and Spatio-Temporal Illustrations

In our illustrations, we induce discrepancy error through model misspecification (Bradley et al., 2020), which we leverage to improve predictions. We focus on two potential model misspecifica-

tions that arise in spatio-temporal statistics: in Section 4.1 we induce discrepancy error by choosing too few basis functions to model spatial dependencies (Stein, 2014), and in Section 4.2 we induce discrepancy error by ignoring spatio-temporal dynamics (Wikle and Hooten, 2010).

4.1 Simulations

We consider a standard test function to illustrate the predictive performance of various choices of DHGT in Section 3.1. In particular, we use the "piecewise polynomial" test function from Nason and Silverman (1994) computed using the R package `wavethresh` (Nason and Nason, 2016). We assume the data is Gaussian distributed with mean given by the test function and variance that implies a signal-to-noise ratio equal to ten. In this section, as an illustration, we create discrepancy error by choosing too few basis functions. Low-rank basis function expansions can capture large scale features, but there can be a rather large error (what we refer to as discrepancy error) between the low-rank expansion and the actual process (Stein, 2014). The DHGT gives a way to explicitly account for this error to improve predictions.

The First Layer BHM is given by the saturated conjugate model in Section 3.1, the Second Layer BHM is defined as a mixed effects model in (3.2) with $\sigma^2 \equiv 0$, $\mathbf{X}^{(2)}$ a $\ell = 4$ matrix of Haar wavelets, and $\mathbf{G}^{(2)}$ a $r = 4$ matrix of Haar wavelets. Together $\mathbf{X}^{(2)}$ and $\mathbf{G}^{(2)}$ produce 8 coarse Haar wavelet functions (Novikov et al., 2005). The third Layer BHM is specified with $\mathbf{X}^{(3)}$ as an n-dimensional vector of ones, and $\mathbf{G}^{(3)}$ as an $n \times 40$ matrix of Gaussian radial basis functions. The matrix $\mathbf{G}^{(3)}$ is based on 20 equally spaced knots over 256 equally spaced points in $[0,1]$ with a bandwidth of 0.01 and another 20 equally space knots with a bandwidth 0.001.

The wavelets are purposely misspecified in the sense that we use only eight coarse Haar wavelet functions (Novikov et al., 2005) in the Second Layer BHM. Typically, more wavelets would be used, and hence a discrepancy error is likely present which can be leveraged to improve predictions of $Y^{(3)}$. As an illustration see Figure 2, where $n = 256$ and note that this data can be interpreted

as a one-dimensional spatial process with $s = 1/256, 2/256, \ldots, 1$, $n = 256$, and $T = 1$. Here, the blue line represents the piecewise polynomial test function, the orange lines indicate the posterior mean, and the transparent region represent 95% credible intervals. As we increase the value of J in the subplots of Figure 2, we see that the estimates of the posterior mean are progressively becoming smoother and closer to the true test function, and the credible intervals are becoming successively smaller. The use of coarse Haar wavelets force a clear piece-wise constant pattern at $J = 2$ in Figure 2, however, it follows the general trend of the true test function. In J = 3, we smooth this piece-wise test function when $J = 2$ to obtain an estimate that is closer to the true smooth test function in Figure 2. As a frame of reference we fit a single Gaussian BHM, where the latent mean is modeled with a Gaussian radial basis function expansion implemented via `rstan` (Arezooji, 2020). In Figure 2, we see that the DHGT with $J = 1$ and $J = 2$ are worse than the traditional BHM implemented with `rstan`, however, the traditional BHM implemented with `rstan` performs worse (visually when comparing the estimate to the truth) than the DHGT with $J = 3$. Computationally, all three DHGTs are considerably faster than the traditional BHM for this sample, where DHGT with $J = 2$ was computed in 0.18 seconds, DHGT $J = 3$ was computed in 0.37 seconds, and the BHM implemented via `rstan` took 47.14 seconds.

In Table 1, we provide the average root mean squared error (RMSE) between the posterior mean and the test function over 50 independent replicate datasets to see if these patterns are consistent over multiple samples. We also provide the continuous rank probability score (CRPS) from Gneiting et al. (2005). Here, we see DHGT with $J = 3$ is similar to the traditional BHM in terms of average RMSE with an interval estimate for RMSE overlapping that of the traditional BHM. The DHGT ($J = 3$) is better than the traditional BHM in terms of CRPS with interval estimates that do not overlap. Computationally the DHGT with $J = 3$ is marginally slower than the DHGT with $J = 2$ in a practical sense, and is considerably faster than the traditional BHM implemented with `rstan`. Not only are the DHGT's faster, but simulations from the posterior distribution are exact, *and* do not have any of the computational overhead of MCMC. Hence, we obtain predictions with

Method	Average MSE	CI MSE	CRPS	CI CRPS	Average CPU
DHGT ($J = 1$)	0.324	$(0.319, 0.28)$	0.223	$(0.221, 0.225)$	
DHGT ($J = 2$)	0.236	$(0.235, 0.237)$	0.129	$(0.128, 0.130)$	0.216
DHGT ($J = 3$)	0.147	$(0.144, 0.149)$	0.071	$(0.069, 0.073)$	0.426
Stan	0.151	$(0.148, 0.154)$	0.087	$(0.084, 0.090)$	48.859

Table 1: Fifty independent replicate data sets were generated according to Section 4.1. The column "RMSE" gives the RMSE between the posterior mean and test function, "CI MSE" displays the average RMSE plus or minus two standard deviations. Similarly, the average CPU (in seconds) is given. We do not give the CPU time for DHGT with $J = 1$ since it is in real-time. The continuous rank probability score (CRPS) is computed as defined in Gneiting et al. (2005). The column "CI CRPS" displays the average CRPS plus or minus two standard deviations across the simulated data sets.

similar to better inferential performance as the traditional BHM (with $J = 3$), but is consistently and considerably faster.

This illustration is useful for understanding when to use DHGT. In particular, each layer is purposely misspecified to be increasingly smooth as j increases. Using eight wavelets gives a quick crude approximation of the function in the Second Layer BHM, which is refined in the Third Layer BHM. Thus, when it is computational difficult to incorporate the latent process/function's complexity in a single BHM, the DHGT can be used to successively refine the predictions (as seen in the succession of panels in Figure 2).

4.2 Analysis of the 2017 Haypress Wildfire

The 2017 Haypress fire was the largest out of 19 fires in the 2017 Orleans Complex fire in Siskiyou County California from July 2017 to January 2018 (e.g., see Yoo and Wikle, 2022, who analyzed 22 observed time points using a level-set model). This dataset was downloaded from the GeoMac database (Group, 2019). There are $T = 52$ total time points, and each time point we observe an image of the progressing perimeter of the wildfire. These images are discretized on a 100×100

grid leading to $n = 520000$. We let $Z_t(\mathbf{s}_i) = 1$ if the fire is burning at time t at the i-th grid cell $\mathbf{s}_i$, and $Z_t(\mathbf{s}_i) = 0$ if the fire is not burning at time t at the i-th grid cell $\mathbf{s}_i$. Hence, the data is assumed to be Bernoulli distributed. See the first row of Figure 3, for a plot of the data at two time points $t = 31$ and $t = 32$. These two time points were chosen as there was a large jump in the perimeter of the fire.

The First Layer BHM is given by the saturated conjugate model in Section 3.1, the Second Layer BHM is defined as a mixed effects model in (3.2) with $\sigma^2 \equiv 0$, $\mathbf{X}^{(1)}$ a column vector with all entries equal to one, and $\mathbf{G}^{(2)}$ be a $n \times 27$ matrix of bisquare basis functions (Cressie and Johannesson, 2008) chosen using the R-package `FRK` (Zammit-Mangion and Cressie, 2021), with one resolution, prune option set to 15, and sub-sampling option set to 20000. The Second Layer BHM models the probability (on the logit scale) as a linear combination of a small set of bisquare basis functions, which does not incorporate any time dynamics that are known to be present in an evolving fire perimeter (e.g., see Hooten and Wikle, 2010; Quaife and Speer, 2021, among others). The low-rank specification makes the Second Layer BHM computationally efficient, however, as seen in Section 4, can be problematic. The lack of dynamic structure and low-rank specification in the Second Layer BHM introduces a discrepancy error. Consequently, in the Third Layer BHM we define a type of cellular automata (CA) that incorporates dynamic structure:

$$\begin{aligned} Y_t^{(2)}(\mathbf{s}_i) &= Z_{new,t-1}(\mathbf{s}_i)\beta_1^{(3)} + (1 - I_{\mathscr{N}_{i,t-1}})\beta_2^{(3)} + (1 - Z_{new,t-1}(\mathbf{s}_i))I_{\mathscr{N}_{i,t-1}}\mathbf{g}_t^{(2)\prime}(\mathbf{s}_i)\boldsymbol{\eta}^{(3)} + \xi_{it}^{(3)}(\mathbf{s}_i) \\ &\equiv \mathbf{x}_t^{(3)}(\mathbf{s}_i)'\boldsymbol{\beta}^{(3)} + (1 - Z_{new,t-1}(\mathbf{s}_i))I_{\mathscr{N}_{i,t-1}}\mathbf{g}_t^{(2)\prime}(\mathbf{s}_i)\boldsymbol{\eta}^{(3)} + \xi_t^{(3)}(\mathbf{s}_i), \end{aligned} \tag{16}$$

where $i = 1, \ldots, 10000$, $t = 1, \ldots, T$, $I_{\mathscr{N}_{i,t-1}}$ is equal to 1 if the neighbor of $Z_{t-1}(\mathbf{s}_i)$ is equal to one (and zero otherwise). Hence, the first term indicates when the fire is recorded as burning at pixel i and time $t-1$, the second term indicates when the fire is not burning at pixel i at time $t-1$ and none of the neighboring pixels neighbors are burning at time $t-1$, and the third term indicates when the fire is recorded as not burning at pixel i at time $t-1$ but one of the neighboring pixels

is recorded as burning at time $t-1$ (i.e., the perimeter of the fire). The term $\xi^{(3)}$ represents a generic (Gaussian) error associated with the CA model, and the 2-dimensional vector $\mathbf{x}_t^{(3)}(\mathbf{s}_i) = \{Z_{new,t-1}(\mathbf{s}_i), 1 - I_{\mathcal{N}_{i,t-1}}\}'$. Let $Z_{new,t}(\mathbf{s}_i)$ be an independent sample of $Z_t(\mathbf{s})$, which is assumed to be Bernoulli with probability $\exp\{Y_t^{(2)}(\mathbf{s})\} / \left[1 + \exp\{Y_t^{(2)}(\mathbf{s})\}\right]$. The third Layer BHM defines the (t,i)-th row of the $n \times 2$ matrix $\mathbf{X}^{(3)}$ with $\mathbf{x}_t(\mathbf{s})'$ in (16), and the (t,i)-th row of the $n \times 27$ matrix $\mathbf{G}^{(3)}$ is $(1 - Z_{new,t-1}(\mathbf{s}_i)) I_{\mathcal{N}_{i,t-1}} \mathbf{g}_t^{(2)\prime}(\mathbf{s}_i)$, and $\mathbf{g}_t^{(2)\prime}(\mathbf{s}_i)$ is the (t,i)-th row of $\mathbf{G}^{(2)}$.

There are differences between the CA model in (16) and that in Hooten and Wikle (2010) besides our use of a DHGT (i.e., (16) assumes the autoregressive model on the logit-scale, whereas Hooten and Wikle (2010) impose this structure on the inverse-logit-scale). However, a possibly more crucial difference is that Hooten and Wikle (2010) replaces Z_{new} with Z in a traditional BHM model creating a computationally difficult (in this setting) type of auto-regressive BHM. Our use of posterior predictive data, which is only possible in the DHGT setting, specifies a linear model that mimics this autoregressive structure (provided Z_{new} is close to Z). This illustrates an exciting feature of the DHGT, where one can mimic complicated autoregressive models using a linear model.

In Figure 3, we plot the data, the posterior mean of the probability of a fire at time point 31 from the Second Layer's BHM using time points $t = 1, \ldots, T = 31$ as observed data. We also provide the forecast of the probability of a fire at time-point 32 computed using the posterior mean. Clearly, the Second Layer BHM fails at forecasting because the implied BHM does not include any dynamic structure. Now consider Figure 4, we again plot the data, the posterior mean of the probability of a fire at time point 31 from the Third Layer's BHM using time points $t = 1, \ldots, T = 31$ as observed data. We also provide the forecast of the probability of a fire at time-point 32 computed using the posterior mean. The Third Layer BHM is able to forcast considerably better due to the CA structure. This is verified using the area under the curve (AUC) metric, where the true positives and false positives are computed at locations that are not burning at time point 31. That is, the AUC is 0.65 for the Second Layer BHM and 0.75 for the Third Layer BHM. However, the improvement

when incorporating dynamics is not constant. The above example considers a time where there is a large change in the perimeter of the fire. Now consider using the data from times 1 to $T = 51$ for training and forecasting the perimeter of the fire at time point 52. There is a very small change in the perimeter in this case and the AUC for the Second Layer BHM is 0.79 and the Third Layer BHM has a marginal improvement of an AUC of 0.83. In this case, the discrepancy error is smaller and hence, the improvement in prediction is small.

In Figure 5, we plot the posterior variances of $Y^{(2)}$ and $Y^{(3)}$ at time points 31 and 32. The interior of the fire has small posterior variance, which is intuitively reasonable, since the fire is observed at these locations at these times. There is more variability in our predictions at the perimeter of the fire, which again is reasonable, since there is practical uncertainty in whether or not the fire will spread near the perimeter. Forecast variances are larger than the in-sample variances as expected. The strange circular features arising from model misspecification of $Y^{(2)}$ arise in the posterior variance of $Y^{(2)}$. However, these patterns are not present in the posterior variance for $Y^{(3)}$, since temporal dynamics have been incorporated.

5 Discussion

In this article, we introduce a new Bayesian statistical model that allows for discrepancy errors between a signal term and a noise term in a spatio-temporal additive model, which we refer to as the deep hierarchical generalized transformation model (DHGT). In particular, we extend the process augmentation approach for discrepancy errors in Bradley et al. (2020), but augmenting the process multiple times. The covariances can be written as a telescoping sum in an additive model, which allows one to avoid confounding between the signal term's covariance and the noise term's covariance. Moreover, a dynamic computing procedure naturally arises where posterior predictive sample is used as data in the next layer's posterior distribution. There exists specifications of the DHGT that allows one to simulate directly from its posterior distribution without the use of

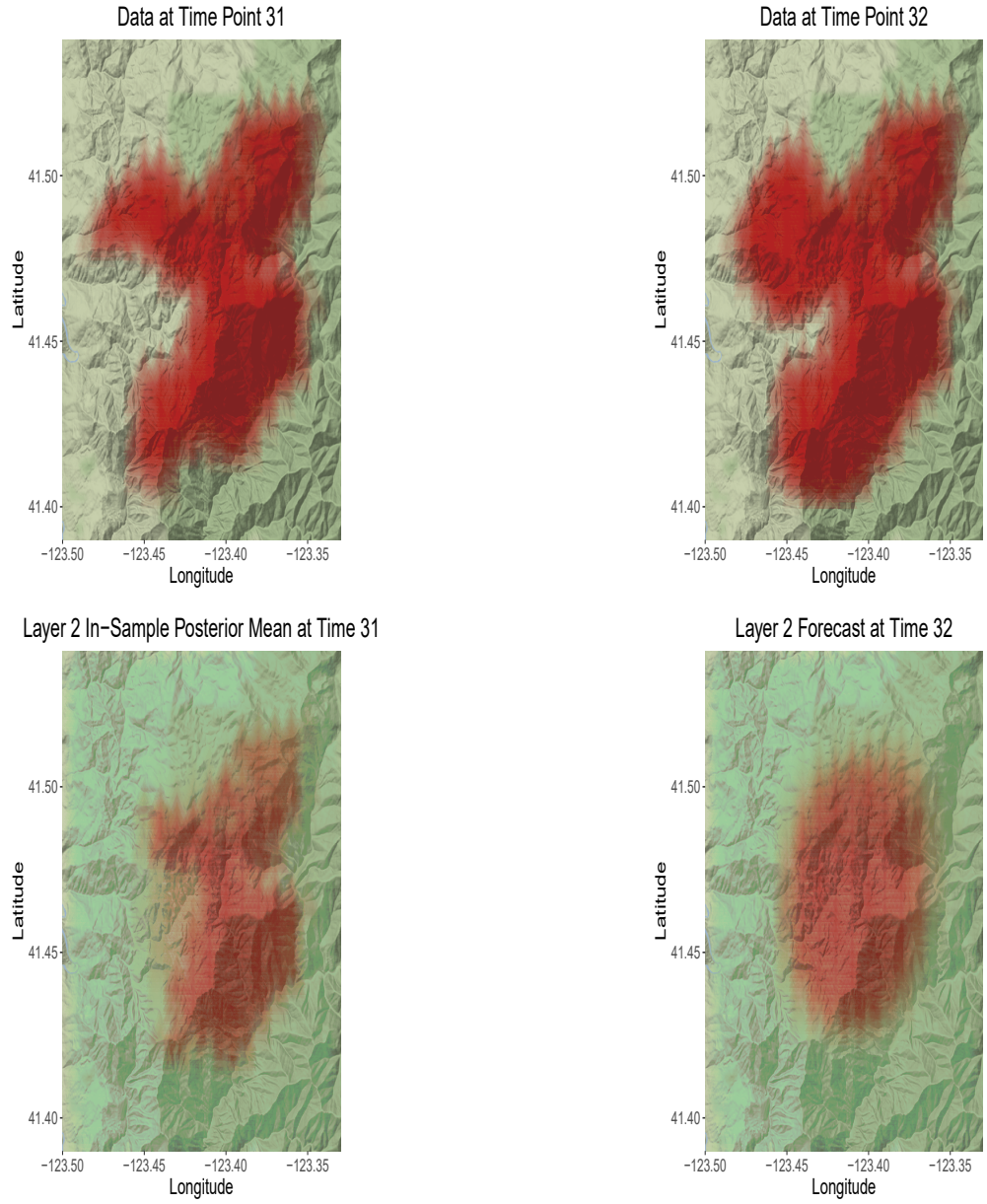

Figure 3: The top row gives the fire perimeter at time points $t = 31$ and $t = 32$. In the bottom left panel we produce the Second Layer BHM predicted (posterior mean) probability of a fire at time point 31 using the time points $1, \ldots, 31$. In the bottom right panel we produce the Second Layer BHM forecast (posterior mean) probability of a fire at time point 32 using the time points $1, \ldots, 31$. The darker red values are closer to the value of one, and lighter-pink values are closer to zero. Notice that edge effects are present from the discretization.

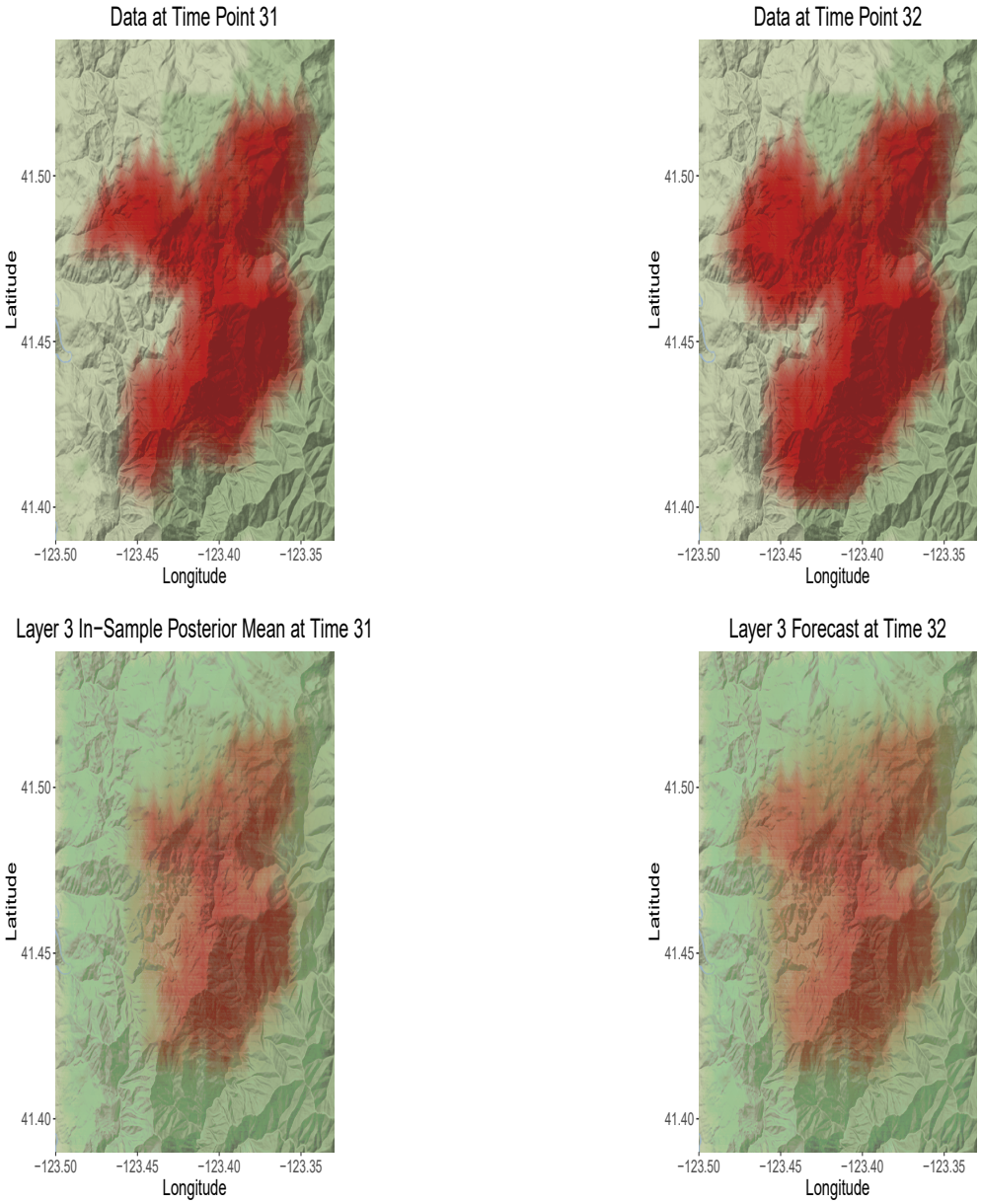

Figure 4: The top row gives the fire perimeter at time points $t = 31$ and $t = 32$. In the bottom left panel we produce the Third Layer BHM predicted (posterior mean) probability of a fire at time point 31 using the time points $1, \ldots, 31$. In the bottom right panel we produce the Third Layer BHM forecast (posterior mean) probability of a fire at time point 32 using the time points $1, \ldots, 31$. The darker red values are closer to the value of one, and lighter-pink values are closer to zero. Notice that edge effects are present from the discretization.

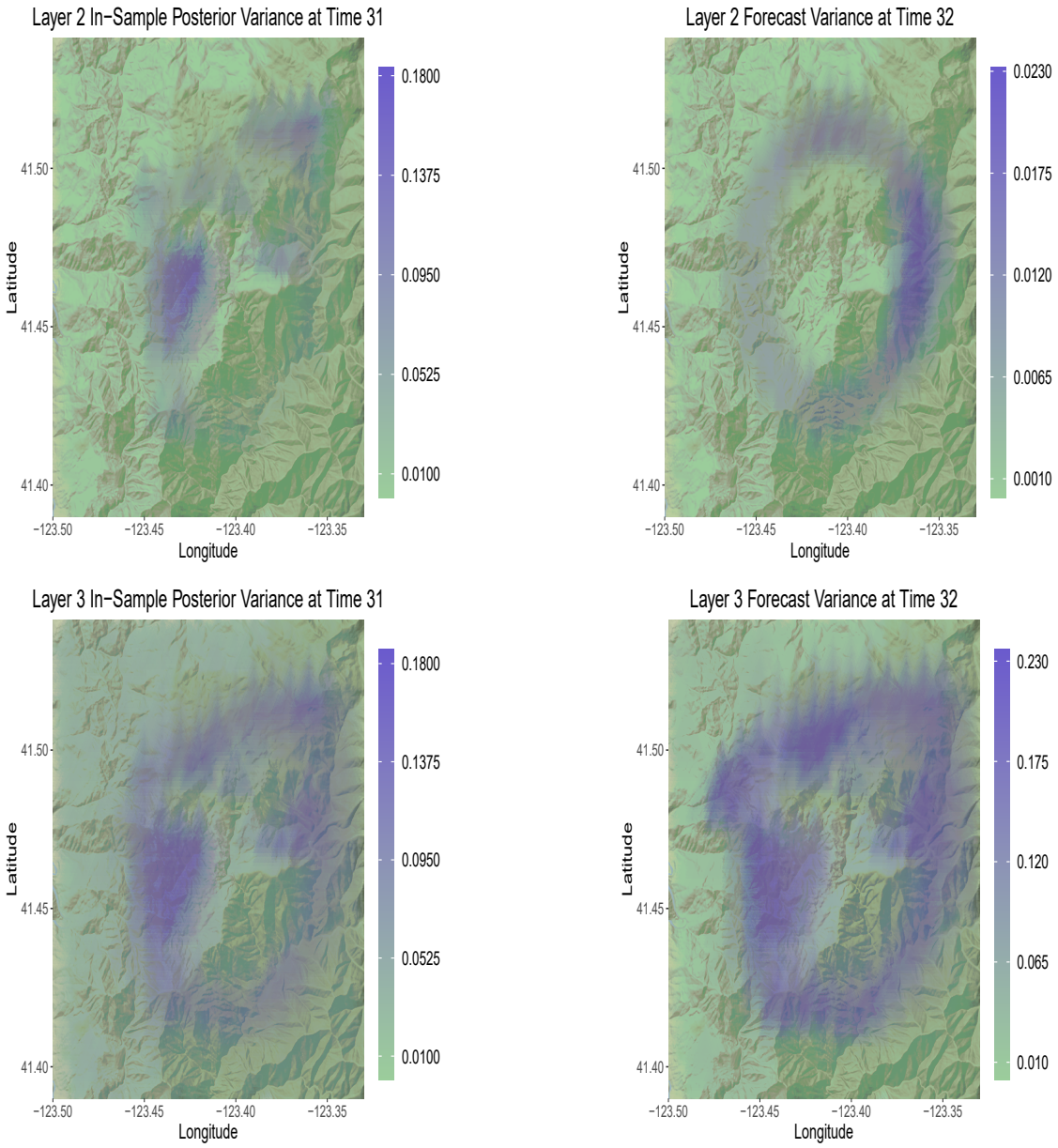

Figure 5: In the top row we produce the Second Layer BHM posterior variance of the probability of a fire at time point 31 and 32, respectively, using the time points $1,\ldots,31$. In the bottom row we produce the Third Layer BHM posterior variance of the probability of a fire at time point 31 and 32, respectively, using the time points $1,\ldots,31$.

MCMC.

In our simulations, the second layer is specified to have a small number of basis functions

leading to a discrepancy error between the true latent process and the model based on a low-rank basis expansion. The results show that the DHGT performs as well to better than a baseline BHM, but is considerably more computationally efficient. In the 2017 Haypress wildfire example, we introduce a discrepancy error between the second and the third layers by ignoring CA dynamics in the second layer and incorporating them in the third. We found that including a third layer that leverages this discrepancy error improves one step ahead forecasting.

A common criticism with deep models is that one may not be able to learn all the parameters in a model. In a Bayesian context this is equivalent to checking whether the data is independent of the parameters so that the posterior distribution is equivalent to the prior distribution, and hence there is no Bayesian learning. In our Bayesian discrepancy error model, the concept of "no Bayesian learning" amounts to the j-th process and parameters being independent of the $(j-1)$-th process and parameters, since the $(j-1)$ layer is used as new transformed data for the j-th Layer BHM in our expression of the posterior distribution in (6). Thus, if these layers are independent then we have that their cross-covariance is zero, and there are no discrepancy error covariance to leverage to improve predictions. This is seen explicitly in the Haypress wildfire, where at a particular time point we would expect the temporal dynamics to be not needed, and as such, we saw small improvements.

There are several developments to the conjugate DHGT methodology to consider in future research. For example, the specification of J is chosen to be 3 in our examples. However, in practice J could be made unknown and prior on J can be considered. Another limitation is that the variance parameters are marginalized out and inference is limited to prediction. One direction that requires significant exploration is to use the j-th layer's variance parameters as new transformed data for the $(j+1)$ layer's variance parameters. There are also several developments to consider for the general DHGT model as well. For example, a key limitation of a general DHGT is that the hierarchical model expression contains integral expressions that may not always have closed form. Finding closed form expressions of these integrals is an important consideration in future

developments.

Appendix: Technical Results

Proof of Equation (6): We start by decomposing the joint posterior distribution into the product of conditional distributions as follows.

$$
\begin{aligned}
f(\boldsymbol{y}^{(J)},\boldsymbol{\theta}^{(J)},\ldots,\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)}|\boldsymbol{z}) &= f(\boldsymbol{y}^{(J)},\boldsymbol{\theta}^{(J)}|\boldsymbol{y}^{(J-1)},\boldsymbol{\theta}^{(J-1)}\ldots,\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)},\boldsymbol{z})f(\boldsymbol{y}^{(J-1)},\boldsymbol{\theta}^{(J-1)},\ldots,\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)}|\boldsymbol{z}) \\
&\vdots \\
&= \left\{\prod_{j=2}^{J} f(\boldsymbol{y}^{(j)},\boldsymbol{\theta}^{(j)}|\mathbf{z},\{\boldsymbol{y}^{(w)}: w<j\},\{\boldsymbol{\theta}^{(w)}: w<j\})\right\} f(\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)}|\mathbf{z}).
\end{aligned} \tag{17}
$$

Thus, we only need to derive $f(\boldsymbol{y}^{(j)},\boldsymbol{\theta}^{(j)}|\mathbf{z},\{\boldsymbol{y}^{(w)}: w<j\},\{\boldsymbol{\theta}^{(w)}: w<j\})$. We split this result into three cases, when $j=1$, $1<j<J$, and $j=J$.

- Case 1, $j=1$:

$$
\begin{aligned}
&f(\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)}|\mathbf{z}) \propto f(\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)},\mathbf{z}) \\
&\propto h(\mathbf{z}|\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)})p(\boldsymbol{y}^{(1)}|\boldsymbol{\theta}^{(1)})\pi(\boldsymbol{\theta}^{(1)}) \\
&\quad \int\int h(\boldsymbol{y}^{(1)}|\boldsymbol{y}^{(2)},\boldsymbol{\theta}^{(2)})p(\boldsymbol{y}^{(2)}|\boldsymbol{\theta}^{(2)})\pi(\boldsymbol{\theta}^{(2)})d\boldsymbol{y}^{(2)}d\boldsymbol{\theta}^{(2)}\frac{1}{m_2(\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)})} \\
&\propto h(\mathbf{z}|\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)})p(\boldsymbol{y}^{(1)}|\boldsymbol{\theta}^{(1)})\pi(\boldsymbol{\theta}^{(1)}) \\
&\propto f(\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)}|\mathbf{z}),
\end{aligned}
$$

where $\int\int h(\boldsymbol{y}^{(1)}|\boldsymbol{y}^{(2)},\boldsymbol{\theta}^{(2)})p(\boldsymbol{y}^{(2)}|\boldsymbol{\theta}^{(2)})\pi(\boldsymbol{\theta}^{(2)})d\boldsymbol{y}^{(2)}d\boldsymbol{\theta}^{(2)}\frac{1}{m_2(\boldsymbol{y}^{(1)},\boldsymbol{\theta}^{(1)})}=1$ by definition of m_2.

- Case 2, $1 < j < J$:

$$
\begin{aligned}
&f(\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)} | \mathbf{z}, \{\boldsymbol{y}^{(w)} : w < j\}, \{\boldsymbol{\theta}^{(w)} : w < j\}) \propto f(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(j)}, \mathbf{z}) \\
&\propto h(\boldsymbol{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)} | \boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)}) p(\boldsymbol{y}^{(j)} | \boldsymbol{\theta}^{(j)}) \pi(\boldsymbol{\theta}^{(j)}) \\
&\int \int h(\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)} | \boldsymbol{y}^{(j+1)}, \boldsymbol{\theta}^{(j+1)}) p(\boldsymbol{y}^{(j+1)} | \boldsymbol{\theta}^{(j+1)}) \pi(\boldsymbol{\theta}^{(j+1)}) d\boldsymbol{y}^{(j+1)} d\boldsymbol{\theta}^{(j+1)} \frac{1}{m_{j+1}(\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)})} \\
&\propto h(\boldsymbol{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)} | \boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)}) p(\boldsymbol{y}^{(j)} | \boldsymbol{\theta}^{(j)}) \pi(\boldsymbol{\theta}^{(j)}) \\
&\propto f(\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)} | \boldsymbol{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)}),
\end{aligned}
$$

  where $\int \int h(\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)} | \boldsymbol{y}^{(j+1)}, \boldsymbol{\theta}^{(j+1)}) p(\boldsymbol{y}^{(j+1)} | \boldsymbol{\theta}^{(j+1)}) \pi(\boldsymbol{\theta}^{(j+1)}) d\boldsymbol{y}^{(j+1)} d\boldsymbol{\theta}^{(j+1)} \frac{1}{m_{j+1}(\boldsymbol{y}^{(j)}, \boldsymbol{\theta}^{(j)})} = 1$ my definition of m_{j+1}.

- Case 3, $j = J$:

$$
\begin{aligned}
&f(\boldsymbol{y}^{(J)}, \boldsymbol{\theta}^{(J)} | \mathbf{z}, \{\boldsymbol{y}^{(w)} : w < J\}, \{\boldsymbol{\theta}^{(w)} : w < J\}) \propto \\
&f(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(J)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(J)}, \mathbf{z}) \\
&\propto h(\boldsymbol{y}^{(J-1)} | \boldsymbol{y}^{(J)}, \boldsymbol{\theta}^{(J)}) p(\boldsymbol{y}^{(J)} | \boldsymbol{\theta}^{(J)}) \pi(\boldsymbol{\theta}^{(J)}) \frac{1}{m_J(\boldsymbol{y}^{(J-1)}, \boldsymbol{\theta}^{(J-1)})} \\
&\propto h(\boldsymbol{y}^{(J-1)} | \boldsymbol{y}^{(J)}, \boldsymbol{\theta}^{(J)}) p(\boldsymbol{y}^{(J)} | \boldsymbol{\theta}^{(J)}) \pi(\boldsymbol{\theta}^{(J)}) \\
&\propto f(\boldsymbol{y}^{(J)}, \boldsymbol{\theta}^{(J)} | \boldsymbol{y}^{(J-1)}, \boldsymbol{\theta}^{(J-1)}),
\end{aligned}
$$

  where we drop $\frac{1}{m_J(\boldsymbol{y}^{(J-1)}, \boldsymbol{\theta}^{(J-1)})}$, since (as a function of $\boldsymbol{y}^{(J)}$ and $\boldsymbol{\theta}^{(J)}$) it is a proportionality constant.

Substituting Case 1 through 3 into the expression in (17) completes the result.

The marginalized DHGT: Discarding samples of $\mathbf{y}^{(j)}$ and $\boldsymbol{\theta}^{(j)}$ $1 \leq j < J$ effectively marginalizes

these quantities from the DHGT. That is, the BHM used for inference is given by

$$\begin{aligned}
&\text{Data Model}: h_m(\mathbf{z}|\mathbf{y}^{(J)}, \boldsymbol{\theta}^{(J)}) \\
&\text{Process Model}: \; p_m(\mathbf{y}^{(J)}|\boldsymbol{\theta}^{(j)}) \\
&\text{Parameter Model}: \; \pi_m(\boldsymbol{\theta}^{(J)}),
\end{aligned} \tag{18}$$

where the subscript "m" stands for "marginal." To derive the marginal data model $h(\mathbf{z}|\mathbf{y}^{(J)}, \boldsymbol{\theta}^{(J)})$ multiply each level in the DHGT in (4) to obtain the joint distribution of the data, processes, and parameters. Then integrate out $\mathbf{y}^{(j)}$ and $\boldsymbol{\theta}^{(j)}$ for $1 \leq j < J$ to obtain the joint distribution of $\mathbf{z}$, $\mathbf{y}^{(J)}$, and $\boldsymbol{\theta}^{(J)}$ as follows:

$$\begin{aligned}
&f(\mathbf{z}, \mathbf{y}^{(J)}, \boldsymbol{\theta}^{(J)}) = \\
&\int \dots \int h(\mathbf{z}|\mathbf{y}^{(1)}, \boldsymbol{\theta}^{(1)}) p(\mathbf{y}^{(1)}|\boldsymbol{\theta}^{(1)}) \pi(\boldsymbol{\theta}^{(1)}) \left\{ \prod_{j=2}^{J-1} \frac{h(\mathbf{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)}|\mathbf{y}^{(j)}, \boldsymbol{\theta}^{(j)}) p(\mathbf{y}^{(j)}|\boldsymbol{\theta}^{(j)}) \pi(\boldsymbol{\theta}^{(j)})}{m_j(\mathbf{y}^{(j-1)}, \boldsymbol{\theta}^{(j-1)})} \right\} \\
&\qquad \times \frac{h(\mathbf{y}^{(J-1)}, \boldsymbol{\theta}^{(J-1)}|\mathbf{y}^{(J)}, \boldsymbol{\theta}^{(J)})}{m_J(\mathbf{y}^{(J-1)}, \boldsymbol{\theta}^{(J-1)})} d\mathbf{y}^{(1)} d\boldsymbol{\theta}^{(1)} \dots d\mathbf{y}^{(J-1)} d\boldsymbol{\theta}^{(J-1)} p(\mathbf{y}^{(J)}|\boldsymbol{\theta}^{(J)}) \pi(\boldsymbol{\theta}^{(J)}).
\end{aligned}$$

This leads to

$$\begin{aligned}
h_m(\mathbf{z}|\mathbf{y}^{(J)}, \boldsymbol{\theta}^{(J)}) &= \frac{f(\mathbf{z}, \mathbf{y}^{(J)}, \boldsymbol{\theta}^{(J)})}{\int f(\mathbf{z}, \mathbf{y}^{(J)}, \boldsymbol{\theta}^{(J)}) d\mathbf{z}} \\
p_m(\mathbf{y}^{(J)}|\boldsymbol{\theta}^{(J)}) &= \frac{\int f(\mathbf{z}, \mathbf{y}^{(J)}, \boldsymbol{\theta}^{(J)}) d\mathbf{z}}{\int \int f(\mathbf{z}, \mathbf{y}^{(J)}, \boldsymbol{\theta}^{(J)}) d\mathbf{z} d\mathbf{y}^{(J)}} \\
\pi_m(\boldsymbol{\theta}^{(1)}) &= \int \int f(\mathbf{z}, \mathbf{y}^{(J)}, \boldsymbol{\theta}^{(J)}) d\mathbf{z}\, d\mathbf{y}^{(J)},
\end{aligned}$$

where appropriate integrals are replaced with sums when $\mathbf{z}$ is discrete.

Acknowledgments

Jonathan R. Bradley's research was partially supported by the U.S. National Science Foundation (NSF) under NSF grant SES-1853099. The author is deeply appreciative to several helpful discussions from Dr. Christopher K. Wikle at the University of Missouri and Drs. Kevin Speer, Bryan Quaife, and Xin Tong at Florida State University Geophysical Fluid Dynamics Institute (GFDI). The authors would also like to thank Mr. Craig Anderson for access to and organization of the 2017 Haypress wildfire data.

References

Achtemeier, G. L. (2003). "Rabbit Rules: an application of Stephen Wolframs New Kind of Science to fire spread modeling." In *Fifth Symposium on Fire and Forest Meteorology, Orlando, Fla*, 16–20.

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

Albinet, G., Searby, G., and Stauffer, D. (1986). "Fire propagation in a 2-D random medium." *Journal de Physique*, 47, 1, 1–7.

Arezooji, D. M. (2020). "A Markov Chain Monte-Carlo Approach to Dose-Response Optimization Using Probabilistic Programming (RStan)." *arXiv preprint arXiv:2011.15034*.

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

Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

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. (2022a). "Efficiently Generating Independent Samples Directly from the Posterior Distribution for a Large Class of Bayesian Generalized Linear Mixed Effects Models." *arXiv preprint arXiv:2203.10028*.

— (2022b). "Joint Bayesian Analysis of Multiple Response-Types Using the Hierarchical Generalized Transformation Model." *Bayesian Analysis*, 17, 1, 127–164.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2020). "Hierarchical models for spatial data with errors that are correlated with the latent process." *Statistica Sinica*, 30, 1, 81–109.

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.

Currie, M., Speer, K., Hiers, J., OBrien, J., Goodrick, S., and Quaife, B. (2019). "Pixel-level statistical analyses of prescribed fire spread." *Canadian Journal of Forest Research*, 49, 1, 18–26.

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

Duarte, J. (1997). "Bushfire automata and their phase transitions." *International Journal of Modern Physics C*, 8, 02, 171–189.

Gelman, A. (1996). "Posterior predictive assessment of model fitness via realized discrepancies." *Statistica Sinica*, 6, 733–807.

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.

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.

Gong, R. (2019). "Exact inference with approximate computation for differentially private data via perturbations." *arXiv preprint arXiv:1909.12237*.

Group, G. M.-A. C. (2019). "Historic perimeters combined 2000-2018 geomac." *Geospatial Multi-Agency Coordination Group*.

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.

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.

Hooten, M. B. and Wikle, C. K. (2010). "Statistical agent-based models for discrete spatio-temporal systems." *Journal of the American Statistical Association*, 105, 489, 236–248.

Katzfuss, M. and Guinness, J. (2021). "A general framework for Vecchia approximations of Gaussian processes." *Statistical Science*, 36, 1, 124–141.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. 2nd ed. New York, NY: Springer.

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

Nandy, S., Holan, S. H., Bradley, J. R., and Wikle, C. K. (2022). "Bayesian Hierarchical Models For Multi-type Survey Data Using Spatially Correlated Covariates Measured With Error." *arXiv preprint arXiv:2211.09797*.

Nason, G. and Nason, M. G. (2016). "Package wavethresh."

Nason, G. P. and Silverman, B. W. (1994). "The discrete wavelet transform in S." *Journal of Computational and Graphical statistics*, 3, 2, 163–191.

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.

Paciorek, C. J. (2007). "Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP package." *Journal of Statistical Software*, 19, 2.

Papamarkou, T., Hinkle, J., Young, M. T., and Womble, D. (2022). "Challenges in Markov chain Monte Carlo for Bayesian neural networks." *Statistical Science*, 37, 3, 425–442.

Quaife, B. and Speer, K. (2021). "A Simple Model for Wildland Fire Vortex–Sink Interactions." *Atmosphere*, 12, 8, 1014.

Quick, H., Banerjee, S., and Carlin, B. P. (2013). "Modeling Temporal Gradients in Regionally Aggregated California Asthma Hospitalization Data." *Annals of Applied Statistics*, 7, 154–176.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association*, 92, 437, 179–191.

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.

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.

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.

Wainwright, M. J., Jordan, M. I., et al. (2008). "Graphical models, exponential families, and variational inference." *Foundations and Trends® in Machine Learning*, 1, 1–2, 1–305.

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. (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.

Wikle, C. K. and Hooten, M. B. (2010). "A general science-based framework for dynamical spatio-temporal models." *Test*, 19, 3, 417–451.

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

Wolfram, S. (1983). "Statistical mechanics of cellular automata." *Reviews of modern physics*, 55, 3, 601.

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

Yoo, M. and Wikle, C. K. (2022). "A Bayesian Spatio-Temporal Level Set Dynamic Model and Application to Fire Front Propagation." *arXiv preprint arXiv:2210.14978*.

Zammit-Mangion, A. and Cressie, N. (2021). "FRK: An R package for spatial and spatio-temporal prediction with large datasets." *Journal of Statistical Software*, 98, 1–48.

Zhang, L., Banerjee, S., and Finley, A. O. (2021). "High-dimensional MultivariateGeostatistics: A Conjugate BayesianMatrix-Normal Approach."