DISSERTATION

STATISTICAL MODELING AND COMPUTING FOR CLIMATE DATA

Submitted by

Joshua Hewitt

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2019

Doctoral Committee:

Advisor: Jennifer A. Hoeting

Daniel S. Cooley
Haonan Wang
Stephanie K. Kampf

ABSTRACT

STATISTICAL MODELING AND COMPUTING FOR CLIMATE DATA

The motivation for this thesis is to provide improved statistical models and approaches to statistical computing for analyzing climate patterns over short and long distances. In particular, information needs for water managers motivate my research. Statistical models and computing techniques exist in a careful balance because climate data are generated by physical processes that can yield computationally intractable statistical models. Simplified or approximate statistical models are often required for practical data analyses. Critically, model complexity is moderated as much by research needs and available data as it is by computational capabilities. I start by developing a weighted likelihood that improves estimates of high quantiles for extreme precipitation (i.e., return levels) from latent spatial extremes models. In my second project, I develop a geostatistical model that accounts for the influence of remotely observed spatial covariates. The model improves prediction of regional precipitation and related climate variables that are influenced by global-scale processes known as teleconnections. I make the model more accessible by providing an `R` package that includes visualization, estimation, prediction, and diagnostic tools. The models from my first two projects require estimating large numbers of latent effects, so their implementations rely on computationally efficient methods. My third project proposes a deterministic, quadrature-based computational approach for estimating hierarchical Bayesian models with many hyperparameters, including those from my first two projects. The deterministic method is easily parallelizable and can require substantially less computational effort than common stochastic alternatives, like Monte Carlo methods. Notably, my quadrature-based method can also improve the computational efficiency of other recent, fast, deterministic approaches for estimating hierarchical Bayesian models, such as the integrated nested Laplace approximation (INLA). I also make the quadrature-based method accessible through an `R` package that provides inference for user-specified hierarchical mod-

els. Throughout my thesis, I demonstrate how improved models, more efficient computational methods, and accessible software allow modeling of large, complex climate data.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Jennifer A. Hoeting, for her countable, but numerous, hours of research insights, mentoring, and advice. This work would not be possible without your support. I would also like to thank many of the faculty and graduate students who also helped introduce me to research topics and skills, including my committee members Daniel S. Cooley, Haonan Wang, Stephanie K. Kampf, Miranda J. Fix, Clint Leach, and the rest of the Hooten-Hoeting reading group—Mevin Hooten, Henry Scharf, John Tipton, Trevor Hefley, and Zachary Weller. Statistics is a broad field, and it has been helpful to get introductions and feedback on ideas from a variety of perspectives. Similarly, I would like to thank Dr. Kristina Quynn, Director of CSU Writes, and members of my CSU Writes peer editing group, Abby Ward, and Kathryn Haggstrom. Your outside opinions and feedback have helped me become a clearer writer. I would also like to thank Zube for keeping my research computer running and answering my Unix questions, especially the basic ones.

My research has also been helped by faculty and researchers outside Colorado State University, including Emeric Thibaud, Mathieu Ribatet, Michael Stein, and Mikael Kuusela. Drs. Thibaud and Ribatet provided code to assist in simulating and fitting latent spatial extremes models (Chapter 2). Dr. Thibaud provided code to simulate Brown-Resnick processes, and Dr. Ribatet provided a development version of the `SpatialExtremes` package, written for the `R` computing language, that implements a Gibbs sampler for the unweighted latent spatial extremes model. Discussions with Drs. Stein and Kuusela helped enrich the interpretation of the RESP model (Chapter 3).

Lastly, I would like to acknowledge and thank the Department of Statistics and National Science Foundation for providing funding to allow me to conduct my research. This material is based upon work supported by the National Science Foundation under grant numbers AGS-1419558 and DMS-1106862. This research utilized the CSU ISTeC Cray HPC System supported by NSF Grant CNS–0923386. This work utilized the RMACC Summit supercomputer, which

is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

DEDICATION

To Kimberley, for supporting my joy in "Number Wang."

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Broad motivation for this work

The motivation for this thesis is to provide improved statistical models and approaches to statistical computing for analyzing climate patterns over short and long distances. Here, we adopt the broad perspective that climate represents distributions of potential weather outcomes. We also refer to all data that are informative of climate as climate data. I present a weighted likelihood that uses information about dependence between observations of extremes to improve estimates of spatially correlated marginal return levels (Chapter 2). I also present a model that accounts for the impact of remote covariates on local precipitation (Chapter 3). As estimation of statistical models for climate data is computationally expensive, I also present a deterministic method for approximate Bayesian inference. The method can be applied to general hierarchical Bayesian models and is faster than commonly used Markov chain Monte Carlo methods (Chapter 4). All of this research is motivated by the need for water managers to have improved forecasts of future climate.

General circulation models (GCMs) can forecast changes in large-scale temperature patterns and other variables, but precipitation is challenging for GCMs to predict and uncertainty is not well quantified (Meehl et al., 2014). Similarly, while GCM output is available at increasingly fine spatial resolutions, the output does not necessarily provide reliable information at spatial scales smaller than 200 km or for extreme events (Maraun et al., 2010). Additionally, GCMs do not explicitly quantify uncertainty. Most GCMs use physics-based models to deterministically simulate future weather given input data and model parameters. Future climate can be estimated by modeling distributions of weather patterns observed in the GCM output. Uncertainty can be indirectly estimated by analyzing simulated weather output from a family of carefully configured GCM simulations, known as climate model ensembles (cf. Kay et al.,

2015). Climate ensembles perturb input data, physics models, or parameters to generate several independent simulations of future weather. Together, the multiple simulations are used to estimate uncertainty in future climate that can be attributed to the chaotic nature of physical processes that generate weather. While statistics offers rich theory and models for quantifying uncertainty, it can be difficult to adapt statistical modeling practices to climate processes.

Climate data are generated by interconnected physical processes that can be difficult to model statistically. In this thesis I develop statistical models, methods, and computational techniques to improve modeling of climate phenomena that arise from complex physics. Physical scientists often use differential equation-based, dynamical systems to model climate phenomena. Statistical estimation for dynamical systems is challenging because dynamical systems often lack distributional forms that allow computationally fast inference (Cressie and Wikle, 2011). As a result, applied statistical analyses of climate data may use "phenomenological" models instead, which focus on using hierarchical Gaussian processes to model first and second-order behaviors of dynamical systems—i.e., variability around a mean state. Such models can still be challenging to fit. As with dynamical system models for climate, many phenomenological models have high-dimensional spatio-temporal state spaces that are difficult to explore computationally. Low-rank, sparse, and approximate Gaussian process models have been proposed in recent years to reduce computational demands for analyses of spatially correlated data (Banerjee et al., 2008; Datta et al., 2016; Furrer et al., 2006; Katzfuss, 2016; Lindgren et al., 2011). However, additional or alternate modeling is required for non-Gaussian climate data, like counts of weather events or extreme precipitation.

Computationally tractable statistical models for climate data, in particular for extreme precipitation data, sometimes need to make simplifying assumptions that limit a model's versatility in order to achieve specific inferential goals. Engineers and urban planners rely on estimates of probabilities for extreme rainfall quantities to design durable roads, buildings, and stormwater runoff systems. Extreme value theory (EVT) provides appropriate statistical models for these information needs by studying asymptotic properties of block maxima and threshold

exceedance data. Extending univariate EVT models to processes defined on continuous spatial domains yields joint probability distributions that, in general, are computationally intractable for datasets that include observations from more than 30 spatial locations (Davison et al., 2012). Thus, EVT models can be difficult to apply to engineering problems with large spatial domains. For some engineering problems, computationally tractable models can be formulated by making the simplifying assumption that data are conditionally independent given parameters for marginal distributions (Cooley et al., 2007). The resulting latent model for spatial extremes uses standard spatial models to model dependence between the parameters of marginal EVT distributions and assumes data are conditionally independent, which limits use of the modeling approach to climate regions without strong extremal dependence. In this thesis, I propose using a weighted likelihood to overcome this limitation.

Like GCMs, statistical models are challenged to characterize future climate while accounting for complex physical processes, uncertainty, and dependence. The remainder of this section briefly introduces two modeling problems for climate data (Sections 1.2 to 1.3) before they are studied in detail (Chapters 2 to 3). In both problems, we propose extended modeling approaches that better account for different types of dependence induced by climate phenomena at long and short distances. In Section 1.4, I describe that while standard Bayesian computational strategies are sufficient for estimating the proposed models for climate data, it will become difficult to fit models to future datasets that are larger and more detailed. Motivated by recent research related to statistical computing for large spatial models, the section briefly introduces general challenges in statistical computing techniques for Bayesian inference before they are studied in more detail (Chapter 4).

1.2 Improving return level estimation

Uncertainty in return level estimates for rare events—i.e., the intensity of low-probability rainfall events—can be extremely high due to short observational records. Large uncertainty makes it difficult to develop strategies to mitigate related hazards, like flooding. Latent spatial

extremes models reduce uncertainty by exploiting spatial dependence in statistical characteristics of extreme events to borrow strength across locations. However, these estimates often underestimate return level uncertainty due to model misspecification: many latent spatial extremes models do not account for extremal dependence, which is spatial dependence in the extreme events themselves. This thesis improves estimates from latent spatial extremes models that make conditional independence assumptions by proposing a weighted likelihood that uses the extremal coefficient to incorporate information about extremal dependence during estimation. This approach differs from, and is simpler than, directly modeling the spatial extremal dependence; for example, by fitting a max-stable process, which is challenging to fit to real, large datasets. This thesis adopts a hierarchical Bayesian framework for inference, use simulation to show the weighted model provides improved estimates of high quantiles, and apply our model to more accurately estimate return level uncertainty for Colorado rainfall events with 1% annual exceedance probability.

1.3 Spatial process models for remote effects

While most spatial data can be modeled with the assumption that distant points are uncorrelated, some problems require dependence at both far and short distances. This thesis introduces a model to directly incorporate dependence in phenomena that influence a distant response. Spatial climate problems often have such modeling needs as data are influenced by local factors in addition to remote phenomena, known as teleconnections. Teleconnections arise from complex interactions between the atmosphere and ocean, of which the El Niño–Southern Oscillation teleconnection is a well-known example. This thesis model extends the standard geostatistical modeling framework to account for effects of covariates observed on a spatially remote domain. The model is framed as an extension of spatially varying coefficient models. Connections to existing methods are highlighted and further modeling needs are addressed by additionally drawing on spatial basis functions and predictive processes. Notably, this approach allows users to model teleconnected data without pre-specifying teleconnection

indices, which other methods often require. This thesis adopts a hierarchical Bayesian framework to conduct inference and make predictions. The method is demonstrated by predicting precipitation in Colorado while accounting for local factors and teleconnection effects with Pacific Ocean sea surface temperatures. The proposed model improves upon standard methods for estimating teleconnection effects, and this thesis discusses the model's utility for climate applications.

1.4 Computationally efficient Bayesian inference

Computationally efficient posterior approximation remains a key challenge in applied Bayesian analyses, especially for hierarchical models. Posterior distributions are challenging to compute because their normalizing constant is often unavailable in closed form. Markov chain Monte Carlo (MCMC) methods avoid this issue using sampling techniques to directly approximate quantities that can be expressed as posterior means. But, MCMC methods are computationally expensive. Laplace approximations, including the integrated nested Laplace approximation (INLA, Rue et al., 2009), are computationally efficient alternatives, in particular for marginal quantities. However, such approximations are often limited to models with relatively few hyperparameters. Sparse grid quadrature methods allow computationally-efficient numerical approximation of high dimensional integrals. We propose using sparse grid quadrature to develop a new method to draw inference about hierarchical Bayesian models. This thesis reformulates Bayesian posterior quantities, including densities and expectations so they may be approximated with sparse grid quadrature methods. The proposed method provides approximate Bayesian inference for hierarchical models, allowing computationally efficient approximation of marginal posterior quantities and normalization constants in computationally challenging models. The approximation framework includes INLA as a special case, but can allow models with greater numbers of hyperparameters and more flexible hierarchical structures. The proposed approximations take the form of weighted mixtures of posterior quantities, such as

conditional means and densities, and the large computational savings relative to MCMC are demonstrated for computationally challenging models.

Chapter 2

Improved return level estimation via a weighted likelihood, latent spatial extremes model[1]

2.1 Introduction

Natural hazards with potentially catastrophic impacts arise as extremes of physical processes that are inherently dependent over space, such as large storms that generate extreme precipitation. Accordingly, the statistical modeling of spatially-referenced extreme values has been an active research area in recent years. To effectively plan mitigation strategies for natural hazards caused by extreme precipitation, it is important to build maps that estimate occurrence probabilities and return levels for extreme precipitation events at individual locations. Return level maps for individual locations inform building safety standards, insurance risks, and surface water runoff requirements for stormwater management systems. However, extreme events are rare by definition, so relevant datasets from networks of environmental monitoring stations typically have relatively short observation lengths. Spatial extremes models allow the tails of probability distributions to be estimated while "borrowing strength" from neighboring time series. Widely used to borrow strength, hierarchical models share statistical information across sampling locations to obtain more accurate and spatially consistent estimates of extreme event characteristics.

Often in extremes studies, the primary interest is in modeling return levels of extreme events at individual locations. Latent spatial extremes models are a flexible and computationally efficient class of models for marginal distributions of spatial extremes and quantities derived from them, like return levels. Latent spatial extremes models use a hierarchical framework to add spatial structure to the *parameters* of an extreme value distribution. Many hierarchi-

[1]Accepted for publication in the Journal of Agricultural, Biological and Environmental Statistics with Fix, M., Hoeting, J. A., & Cooley, D. S.

cal frameworks assume observations of extremes are independent across sampling locations, conditional on the latent spatial processes that specify the data's marginal distributions. Hierarchical spatial layers induce smoothness and correlation in marginal return level estimates across sampling locations, and—critically—allow return level maps to be built using spatial interpolation techniques, like kriging. As such, return level estimates "borrow strength" because estimates balance data at each sampling location with spatial smoothing induced by the latent hierarchical layers. For example, Cooley et al. (2007) use latent Gaussian processes in a hierarchical Bayesian model to capture covariate-driven trends and spatial dependence in precipitation data. Bayesian frameworks allow direct estimation of uncertainties in return levels since the posterior distribution contains this information. Latent spatial Gaussian process models can also be scaled to massive datasets with recent advances in models and computational techniques (Lindgren et al., 2011; Rue et al., 2009). Other recent studies employ latent spatial extremes models in either Bayesian or frequentist paradigms (Cooley and Sain, 2010; Lehmann et al., 2016; Opitz et al., 2018; Sang and Gelfand, 2009). However, due to the conditional independence assumption, these examples of latent spatial extremes model cannot account for extremal dependence, which is dependence in observations of extreme events themselves.

Directly modeling extremal dependence poses theoretical and computational challenges. Classical univariate and multivariate extreme value models are generated via asymptotic arguments about the limiting distributions of appropriately renormalized block maxima. The natural extension to the spatial setting is the max-stable process, which is the limiting process of the componentwise maxima of a sequence of suitably renormalized stochastic processes. Examples include the Smith (1990), Schlather (2002), and Brown-Resnick (Brown and Resnick, 1977; Kabluchko et al., 2009) processes. The advantage of max-stable process modeling is that it directly models spatial dependence in the tail and thus permits inference about joint probabilities in addition to marginal quantities, like return levels. However, full likelihood inference for max-stable processes is only computationally tractable in relatively low-dimensional situations (Castruccio et al., 2016; Davison et al., 2012).

In particular, computationally efficient Bayesian methods for spatially-dependent extremes data remains challenging. Frequentist inference for max-stable processes has typically been based on computationally efficient models that use approximate likelihoods, such as composite likelihoods based on bivariate densities of max-stable processes (Padoan et al., 2010). However, composite likelihood methods are computationally expensive and difficult to implement in hierarchical Bayesian models (Ribatet et al., 2012; Sharkey and Winter, 2018). Some Bayesian models do not need to use approximate likelihoods, but are limited to specific max-stable processes or require additional data for estimation (Reich and Shaby, 2012; Thibaud et al., 2016).

The latent spatial extremes approaches previously introduced address computational issues while providing flexible models for estimating marginal parameters, but raise concerns about the impact of model misspecification on inference. These models make a simplifying conditional independence assumption by defining the likelihood to be the product of each location's marginal density. The misspecification due to the conditional independence assumption can result in unrealistically narrow confidence intervals for return level estimates (Cao and Li, 2018; Zheng et al., 2015). Alternative to assuming conditional independence or using computationally expensive models to account for extremal dependence, we seek a compromise between the two modeling approaches. We want to preserve computationally efficient and flexible models for marginal parameters provided by latent variable models, but also account for extremal dependence in observations.

We propose a method for improving marginal inference that is supported by theory and computationally efficient. We develop a weighted likelihood that uses spatial information to induce an effective sample size correction that accounts for the loss of information due to dependent observations. The likelihood weights improve uncertainty estimates in cases of moderate to strong extremal dependence. The effective sample size motivation differs from previous uses of weighted likelihoods. Weighted likelihoods have previously been used to approximate Bayesian inference and as a method for conducting inference on data sampled from multiple, related populations, for example in Hu and Zidek (2002); Newton and Raferty (1994); Wang

(2006). Weighted likelihoods have also recently been proposed for latent spatial extremes models, but only as they relate to composite likelihood corrections (Sharkey and Winter, 2018). A natural tradeoff in using likelihood weights to better account for estimation uncertainty is that mean squared error can be slightly worse in these cases.

The remainder of the article is organized as follows. Section 2.2 introduces our weighted likelihood and Bayesian implementation. Section 2.5 uses a simulation study to show that the weighted likelihood improves estimates, as compared to several models with similar Bayesian hierarchical structure. As part of our comparisons, we derive the penalized complexity prior for the generalized extreme value (GEV) distribution (Section 2.5.2). Section 2.6 applies the weighted likelihood latent model to daily rainfall observations in Colorado's Front Range of the Rocky Mountains. We conclude with discussions of extensions and other directions for future work (Section 2.7).

2.2 Weighted likelihood latent spatial extremes models

We briefly review extreme value theory for modeling return levels from observations of annual maxima (Section 2.2.1). In particular, we introduce the extremal coefficient, which we will use to build our weights. We then propose and interpret a latent variable model with a weighted likelihood to estimate marginal quantities from spatially-dependent extremes data (Section 2.2.2, Section 2.2.3). When data are dependent, the weighted likelihood accounts for model misspecification in the latent variable modeling approach by Cooley et al. (2007), which assumes data are conditionally independent, given marginal parameters. The model has a hierarchical spatial structure, for which posterior distributions can be approximated via Gibbs sampling (Section 2.2.4, Section 2.3).

2.2.1 Max-stable processes and the extremal coefficient

Max-stable processes for spatially-referenced extremes data arise as the pointwise limit of block maxima, which are pointwise maxima of replications of spatially-referenced processes.

Let $\mathscr{D}$ be a continuous spatial domain and $\{Y_{it}(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$, $t \in \{1,\ldots,m\}$ be m independent replications of a spatial process at time block $i \in \mathscr{T} = \{1,\ldots,T\}$. The size of each block $i \in \mathscr{T}$ is represented by m. As the block size m increases, if the limit

$$Y_i(\boldsymbol{s}) = \lim_{m\to\infty} \frac{\max_{t=1}^{m} Y_{it}(\boldsymbol{s}) - b_m(\boldsymbol{s})}{a_m(\boldsymbol{s})}, \boldsymbol{s} \in \mathscr{D}$$

exists for continuous functions $a_m(\boldsymbol{s}) > 0$ and $b_m(\boldsymbol{s}) \in \mathbb{R}$, then $\{Y_i(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$, $t \in \mathscr{T}$ are independent replications of a max-stable process (De Haan, 1984).

In general, the spatial dependence structure for max-stable processes $\{Y_i(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ is complex, but is often summarized for pairs of random variables $Y_i(\boldsymbol{s})$ and $Y_i(\boldsymbol{t})$ through the extremal coefficient. The extremal coefficient $\theta(d)$ is a function that is traditionally defined implicitly for stationary and isotropic fields such that

$$P\big(Y_i(\boldsymbol{s}) \le y, Y_i(\boldsymbol{t}) \le y\big) = P\big(Y_i(\boldsymbol{s}) \le y\big)^{\theta(d)} \tag{2.1}$$

for pairs of random variables $Y_i(\boldsymbol{s})$ and $Y_i(\boldsymbol{t})$ where $d = \|\boldsymbol{s} - \boldsymbol{t}\|$ (Schlather and Tawn, 2003). The extremal coefficient is interpretable as the effective number of independent random variables among pairs of variables separated by a distance d. As such, it takes values in the closed interval $[1,2]$.

Importantly, the univariate marginal distributions for max-stable processes belong to the generalized extreme value distribution family $Y_i(\boldsymbol{s}) \sim \text{GEV}\big(\boldsymbol{\eta}(\boldsymbol{s})\big)$ with distribution function

$$P\big(Y_i(\boldsymbol{s}) \le y\big) = \begin{cases} \exp\left\{-\left(1+\xi(\boldsymbol{s})\left(\frac{y-\mu(\boldsymbol{s})}{\sigma(\boldsymbol{s})}\right)\right)_+^{-1/\xi(\boldsymbol{s})}\right\} & \xi(\boldsymbol{s}) \neq 0 \\ \exp\left\{-\exp\left\{\frac{y-\mu(\boldsymbol{s})}{\sigma(\boldsymbol{s})}\right\}\right\} & \xi(\boldsymbol{s}) = 0 \end{cases} \tag{2.2}$$

where $a_+ = \max(0,a)$ (De Haan, 1984). The parameter vector $\boldsymbol{\eta}(\boldsymbol{s}) = \big(\mu(\boldsymbol{s}), \log\sigma(\boldsymbol{s}), \xi(\boldsymbol{s})\big)^T$ specifies the distribution's location $\mu(\boldsymbol{s}) \in \mathbb{R}$, scale $\sigma(\boldsymbol{s}) > 0$, and shape $\xi(\boldsymbol{s}) \in \mathbb{R}$ parameters. The GEV quantile function $Q\big(p \big| \boldsymbol{\eta}(\boldsymbol{s})\big)$ is derived from (2.2) and has the closed form

$$Q\big(p\big|\boldsymbol{\eta}(\boldsymbol{s})\big) = \begin{cases} \mu(\boldsymbol{s}) + \frac{\sigma(\boldsymbol{s})}{\xi(\boldsymbol{s})}\Big(\big(-\log p\big)^{-\xi(\boldsymbol{s})} - 1\Big) & \xi(\boldsymbol{s}) \neq 0 \\ \mu(\boldsymbol{s}) - \sigma(\boldsymbol{s})\log\big(-\log p\big) & \xi(\boldsymbol{s}) = 0 \end{cases} \tag{2.3}$$

with $p \in [0,1]$.

Asymptotic convergence justifies use of the GEV distribution as an approximate model for the annual maximum of daily precipitation in year i, which is a block maximum quantity that has large but finite replication $t \in \{1,\ldots,m\}$. The approximation allows marginal return levels for extreme precipitation events to be modeled as high quantiles of the GEV distribution at each location $\boldsymbol{s} \in \mathscr{D}$. Assuming a stationary climate, the quantile $Q\big(p\big|\boldsymbol{\eta}(\boldsymbol{s})\big)$ with $p = 1 - 1/r$ is interpretable as the r-year return level—the amount of precipitation carried by a storm that occurs, on average, once every r years. The quantile $Q\big(p\big|\boldsymbol{\eta}(\boldsymbol{s})\big)$ is also associated with the $1-p$ percent annual exceedance probability; the quantile expresses the amount of precipitation carried by a storm that has a $1-p$ percent chance of occurring in a given year.

2.2.2 Weighted likelihood

We propose a latent variable model that uses a weighted marginal likelihood. In general, weighted likelihoods are missspecified but can improve inference relative to unweighted likelihoods. A correctly-specified likelihood for spatial extremes data would fully account for extremal dependence, but be computationally intractable. Marginal likelihoods assume data are conditionally independent across spatial locations and timepoints, given marginal parameters. When the field $\{Y_i(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ is sampled at N spatial locations $\mathscr{S} = \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_N\} \subset \mathscr{D}$, the weighted marginal likelihood for a finite sample of observations $\big\{y_i\big(\boldsymbol{s}_j\big) : i \in \mathscr{T}, \boldsymbol{s}_j \in \mathscr{S}\big\}$ is defined via

$$L(\boldsymbol{\eta}) = \prod_{j=1}^{N}\prod_{i=1}^{T} f\big(y_i(\boldsymbol{s}_j)\big|\boldsymbol{\eta}(\boldsymbol{s}_j)\big)^{w_{\boldsymbol{s}_j}} \tag{2.4}$$

where $f\big(y_i(\boldsymbol{s}_j)\big|\boldsymbol{\eta}(\boldsymbol{s}_j)\big)$ is the probability density function for the GEV distribution (2.2) and $\boldsymbol{\eta}(\boldsymbol{s})$ is the associated parameter vector. The weighted marginal likelihood (2.4) uses likeli-

hood weights $\left\{w_{\boldsymbol{s}_j} : j = 1,\ldots,N\right\}$ and marginal densities $\left\{f\left(y_i(\boldsymbol{s}_j)\middle|\boldsymbol{\eta}(\boldsymbol{s}_j)\right) : j = 1,\ldots,N\right\}$ to estimate the marginal parameters $\left\{\boldsymbol{\eta}(\boldsymbol{s}_j) \in \mathbb{R}^3 : j = 1,\ldots,N\right\}$ that have been stacked to form the vector $\boldsymbol{\eta} \in \mathbb{R}^{3N}$. During estimation, likelihood weights can be constructed to downweight observations for $y_i(\boldsymbol{s}_j)$ that exhibit strong dependence with neighboring observations. Models assuming conditional independence naively assume the weights are unitary.

We use the extremal coefficient in (2.1) to construct weights that downweight likelihood contributions from locations central to the spatial sampling pattern, where observations tend to be most dependent. We construct each weight $w_{\boldsymbol{s}_j}$ by first mapping extremal coefficients $\theta\left(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|\right)$ for $i \neq j$ to the interval $[1/N, 1]$, then averaging the mapped values, yielding

$$w_{\boldsymbol{s}_j} = \frac{1}{N-1} \sum_{i=1,\, i\neq j}^{N} N^{\theta\left(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|\right)-2}, \tag{2.5}$$

so $w_{\boldsymbol{s}_j} \in [1/N, 1]$. The weights (2.5) are specifically constructed so that the statistical information in the weighted marginal likelihood (2.4) matches the statistical information in non-misspecified likelihoods in two special, limiting cases (Section 2.4.1). In the first limiting case, the field is assumed to be independent, and $w_{\boldsymbol{s}_j} = 1$; in the second limiting case, the field $\{Y_i(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ is assumed to have complete dependence over space, and $w_{\boldsymbol{s}_j} = 1/N$. The field $\{Y_i(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ has complete dependence over space if all potential samples $\{Y_i(\boldsymbol{s}_1),\ldots,Y_i(\boldsymbol{s}_N)\}$ can be represented through a collection of continuous transformations $\left\{g_j : j = 1,\ldots,N\right\}$ of a variable U_i such that

$$(Y_i(\boldsymbol{s}_1),\ldots,Y_i(\boldsymbol{s}_N)) \overset{d}{=} \left(g_1(U_i),\ldots,g_N(U_i)\right).$$

2.2.3 Effective sample size interpretation

From an information-theoretic perspective, we show that the weighted likelihood (2.4) induces an effective sample size that corrects inference on spatially correlated marginal parameters when data are also spatially dependent. Effective sample size statistics quantify the im-

pact that dependence has on estimation uncertainty (e.g., Cressie, 1993, p. 13). We use effective sample size to determine factors that will impact estimator performance in our simulation (Section 2.5). In our application, effective sample size also helps us better interpret losses in statistical efficiency due to dependence in observations of extremes (Section 2.6).

The Fisher information for (2.4) is the block diagonal matrix $I(\boldsymbol{\eta}) \in \mathbb{R}^{Nm \times Nm}$ with j^{th} diagonal block $I(\boldsymbol{\eta}(\boldsymbol{s}_j)) \in \mathbb{R}^{m \times m}$ being

$$I(\boldsymbol{\eta}(\boldsymbol{s}_j)) = w_{\boldsymbol{s}_j} T I_{Y_\bullet(\boldsymbol{s}_j)}(\boldsymbol{\eta}(\boldsymbol{s}_j)), \tag{2.6}$$

where $I_{Y_\bullet(\boldsymbol{s}_j)}(\boldsymbol{\eta}(\boldsymbol{s}_j))$ is the expected Fisher information for each of the independent and identically distributed random variables $\{Y_i(\boldsymbol{s}_j) : i \in \mathcal{T}\}$. Note that the j^{th} block (2.6) is the Fisher information for $w_{\boldsymbol{s}_j} T$ independent observations of the response at $\boldsymbol{s}_j$. Thus, $w_{\boldsymbol{s}_j}$ quantifies the effective proportion of independent observations at location $\boldsymbol{s}_j$ that contribute to inference for the marginal GEV parameters $\boldsymbol{\eta}(\boldsymbol{s}_j)$. As the likelihood weight $w_{\boldsymbol{s}_j}$ decreases, uncertainty increases about the marginal GEV parameters $\boldsymbol{\eta}(\boldsymbol{s}_j)$ and return level $Q(p|\boldsymbol{\eta}(\boldsymbol{s}_j))$. Latent spatial extremes models with unweighted likelihoods can be interpreted as implicitly assigning $w_{\boldsymbol{s}_j} = 1$ for all locations $\boldsymbol{s}_j \in \mathcal{S}$. Such a strategy underestimates parameter uncertainty when data have extremal dependence.

2.2.4 Hierarchical specification

We adopt a hierarchical Bayesian framework to conduct inference on the weighted marginal likelihood, and facilitate spatial interpolation of marginal return levels (2.3). We specify a hierarchical spatial process model for the marginal parameters at each spatial location in the domain $\boldsymbol{\eta}(\boldsymbol{s}) = \left(\mu(\boldsymbol{s}), \log\sigma(\boldsymbol{s}), \xi(\boldsymbol{s})\right)^T \in \mathbb{R}^3$ via

$$\boldsymbol{\eta}(\boldsymbol{s}) = \begin{bmatrix} \boldsymbol{x}_\mu(\boldsymbol{s})^T & & \\ & \boldsymbol{x}_{\log\sigma}(\boldsymbol{s})^T & \\ & & \boldsymbol{x}_\xi(\boldsymbol{s})^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_\mu \\ \boldsymbol{\beta}_{\log\sigma} \\ \boldsymbol{\beta}_\xi \end{bmatrix} + \begin{bmatrix} \varepsilon_\mu(\boldsymbol{s}) \\ \varepsilon_{\log\sigma}(\boldsymbol{s}) \\ \varepsilon_\xi(\boldsymbol{s}) \end{bmatrix}, \tag{2.7}$$

in which $\boldsymbol{x}(\boldsymbol{s})$ and $\boldsymbol{\beta}$ are respectively $p \times 1$ vectors of regression covariates and coefficients, and $\varepsilon(\boldsymbol{s})$ represents spatially-correlated variation in the marginal parameters $\boldsymbol{\eta}(\boldsymbol{s})$. The matrix of covariates in (2.7) is block-diagonal; the blank, off-diagonal entries represent zeros. We use diffuse normal priors for regression coefficients $\boldsymbol{\beta}$. Independent Gaussian processes model the spatially-correlated variation in $\varepsilon_{\mu}(\boldsymbol{s})$, $\varepsilon_{\log\sigma}(\boldsymbol{s})$, and $\varepsilon_{\xi}(\boldsymbol{s})$. Gaussian processes imply finite samples of parameters are jointly-normally distributed and allow estimation of spatially-coherent marginal parameter maps $\{\boldsymbol{\eta}(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ through kriging. Furthermore, stationary isotropic Gaussian processes are sufficient models when departures from stationarity and isotropy are difficult to detect (Cooley et al., 2007).

The Gaussian processes for marginal parameters are fully defined by specifying covariance functions $Cov\left(\varepsilon(\boldsymbol{s}), \varepsilon(\boldsymbol{t})|\boldsymbol{\phi}\right) = \rho\left(\|\boldsymbol{s}-\boldsymbol{t}\|;\boldsymbol{\phi}\right)$ to model the spatial correlation in the parameters between locations $\boldsymbol{s}, \boldsymbol{t} \in \mathscr{D}$. Specific choices for covariance functions ρ and hyperprior distributions for covariance parameters $\boldsymbol{\phi} = (\sigma_0, \lambda_0, \nu_0)^T$ are discussed in Section 2.5.3 and Section 2.6.2. In general, we use weakly informative Gamma priors for covariance range λ_0 and smoothness ν_0 parameters, and weakly informative Inverse-Gamma priors for covariance sill parameters σ_0.

2.3 Bayesian implementation of model

A Gibbs sampler can be constructed for inference on the hierarchical Bayesian model specified in Section 2.2.4, in which likelihood weights (2.5) are updated with the aid of a plug-in estimator for the extremal coefficient. The Bayesian framework allows estimates of return levels $Q\left(p\middle|\boldsymbol{\eta}(\boldsymbol{s})\right)$ to be computed directly from posterior samples of the marginal parameter vector $\boldsymbol{\eta}(\boldsymbol{s})$ since return levels are functions of marginal parameters. Standard hybrid Gibbs sampling approaches are used to sample the marginal GEV parameters, covariance parameters, and regression coefficients. The sampler is described in detail in Section 2.3.2. Estimation of likelihood weights is discussed in Section 2.3.1.

2.3.1 Likelihood weights

Likelihood weights (2.5) are computed with a plug-in estimator $\hat{\theta}(d)$ for the extremal coefficient (Cooley et al., 2006). The plug-in estimator uses sample statistics from the data that have been transformed to have unit Fréchet marginal distributions. Thus, the likelihood weights depend on estimates of the marginal distributions, either estimated through the empirical cumulative distribution function (CDF), or directly through the GEV CDF. Before Gibbs sampling begins, we initialize all of the likelihood weights by using the empirical CDF at each location $\hat{F}(y;\boldsymbol{s}_j) = T^{-1}\sum_{i=1}^{T}\mathbb{1}\{y_i(\boldsymbol{s}_j) \leq y\}$ to transform the data via probability integral transforms. These initial weights may be held fixed and used throughout Gibbs sampling or updated at each Gibbs iteration. To update the weights at each Gibbs iteration, the data may be retransformed by using the GEV CDF (2.2) with the marginal parameters $\boldsymbol{\eta}$ from the previous Gibbs iteration. Updating likelihood weights during Gibbs sampling accounts for uncertainty in the likelihood weights.

2.3.2 Gibbs sampler

Gibbs sampling begins by updating marginal GEV parameters at the sampling locations $\{\boldsymbol{\eta}(\boldsymbol{s}_j) : j = 1,\ldots,N\}$. The parameter vectors are updated sequentially, from $j = 1,\ldots,N$. Separate random walk Metropolis-Hastings steps—with proposal standard deviations s_μ, $s_{\log\sigma}$, and s_ξ—are used to update the entries of each parameter vector $\boldsymbol{\eta}(\boldsymbol{s}_j) = (\mu(\boldsymbol{s}_j), \log\sigma(\boldsymbol{s}_j), \xi(\boldsymbol{s}_j))^T$. Proposal standard deviations are chosen in preliminary test runs of the Gibbs sampler to tune acceptance rates so they are close to 44% (Roberts and Rosenthal, 2001).

Sampling then proceeds to update regression coefficients $\boldsymbol{\beta}$ and spatial covariance parameters $\boldsymbol{\phi} = (\sigma_0, \lambda_0, \nu_0)^T$ for each of the independent Gaussian process priors for the GEV parameter processes $\{\mu(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$, $\{\log\sigma(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$, and $\{\xi(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$. Regression coefficients and spatial covariance parameters determine the mean and covariance structures of the Gaussian processes. Separate random walk Metropolis-Hastings steps update spatial range λ_0 and smoothness ν_0 parameters. The random walk proposal distributions respectively have fixed proposal standard devia-

tions s_{λ_0} and s_{ν_0}, which are specified for each of the Gaussian processes. Regression coefficients $\boldsymbol{\beta}$ and sills σ_0 are sampled from conjugate distributions.

For example, the Gaussian process assumption in Section 2.2.4 implies the collection of GEV location parameters $\boldsymbol{\mu} = [\mu(\boldsymbol{s}_j)]_{j=1}^N \in \mathbb{R}^N$ have the jointly-normal conditional prior distribution $\boldsymbol{\mu}|\boldsymbol{\beta}_\mu, \boldsymbol{\phi}_\mu \sim \mathcal{N}(X_\mu \boldsymbol{\beta}_\mu, \Sigma_\mu)$, where the matrix $X_\mu \in \mathbb{R}^{N \times p_\mu}$ is composed of the N row vectors $\boldsymbol{x}_\mu(\boldsymbol{s}_j)^T \in \mathbb{R}^{p_\mu}$, $j = 1, \ldots, N$, and $\Sigma_\mu \in \mathbb{R}^{N \times N}$ is a spatial covariance matrix specified via entries $(\Sigma_\mu)_{ij} = \rho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|; \boldsymbol{\phi}_\mu)$. Since the regression coefficients have a normal prior distribution $\boldsymbol{\beta}_\mu \sim \mathcal{N}(\mathbf{0}, \Lambda_\mu)$ in which Λ_μ is a fixed prior covariance matrix, the conjugate full conditional posterior distribution for $\boldsymbol{\beta}_\mu$ is $\boldsymbol{\beta}_\mu|\boldsymbol{\mu}, \boldsymbol{\phi}_\mu, \cdot \sim \mathcal{N}(m, \Psi)$ with covariance $\Psi = \left(\Lambda_\mu^{-1} + X_\mu^T \Sigma_\mu X_\mu\right)^{-1}$ and mean $m = \Psi X_\mu^T \Sigma_\mu^{-1} \boldsymbol{\mu}$. The covariance sill has an inverse gamma prior distribution $\sigma_0 \sim \text{IG}(a_\mu, b_\mu)$ and conjugate full conditional posterior distribution that depends on the current iteration of the Gaussian process parameters and values specified via

$$\sigma_0|\boldsymbol{\mu}, \boldsymbol{\beta}_\mu, \lambda_\mu, \nu_\mu, \cdot \sim \text{IG}\left(a_\mu + N/2, b_\mu + e^T(\Sigma_\mu/\sigma_0)^{-1} e/2\right),$$

where $e = \boldsymbol{\mu} - X_\mu \boldsymbol{\beta}_\mu$. The conjugate distributions for the regression and sill parameters of the other Gaussian processes $\{\log\sigma(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ and $\{\xi(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ use similar notation and results for their Gibbs steps.

If the model is being fit with Gibbs-updated likelihood weights or penalties, the weights and penalty tuning parameters are updated next. Likelihood weights are computed from (2.5), in which a plug-in estimator $\hat{\theta}(d)$ is used for the extremal coefficient. The plug-in estimator $\hat{\theta}(d)$ is only a function of the data and marginal parameters, so does not explicitly rely on the dependence parameter γ. The basis for the plug-in estimator is a relationship between the extremal coefficient and the F-madogram $\nu^F(d)$. The F-madogram is an analog of the classical variogram for spatial statistics and measures spatial dependence in stationary max-stable fields. If the marginal GEV shape parameters satisfy $\xi(\boldsymbol{s}) < 1$ (i.e., they are not too large), then the extremal coefficient is related to the F-madogram via $\theta(d) = \left(1 + 2\nu^F(d)\right)/\left(1 - 2\nu^F(d)\right)$ (Cooley et al., 2006). After using marginal parameters $\boldsymbol{\eta}$ to transform data $\{y_i(\boldsymbol{s}_j) : i \in \mathscr{T}, j \in \mathscr{S}\}$ to have

unit Fréchet margins, the sample F-madogram can be estimated in a similar manner as variograms, by working with the differences between pairs of observations separated by a distance d. Likelihood weights may be estimated before Gibbs sampling by using the empirical cumulative distribution function (CDF) to transform the data to have unit Fréchet margins in order to estimate the sample F-madogram. Uncertainty in the likelihood weights (2.5) can also be incorporated by updating them at each Gibbs iteration. Conditional on the data and marginal parameters, the weights are deterministic because the sample F-madogram is deterministic. Thus, the weights do not need to be sampled; weights can be updated by using the marginal parameters $\boldsymbol{\eta}$ to re-transform the data to re-estimate the F-madogram at each Gibbs iteration.

The plug-in estimator $\hat{\theta}(d)$ is updated by using the current Gibbs values of the GEV parameters at the sampling locations $\{\boldsymbol{\eta}(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{S}}$ to transform the data $\{y_i(\boldsymbol{s}_j) : i \in \mathscr{T}, j \in \mathscr{S}\}$ to have unit Fréchet margins. The plug-in estimator $\hat{\theta}(d)$ is recomputed from an estimate of the sample F-madogram, using the transformed data. If the model is using a penalized likelihood, as in Section 2.5.2 and (2.22), the penalty's tuning parameter λ may be updated as well. The penalized complexity prior (2.23) does not have a conjugate distribution, so must be updated with a random walk Metropolis-Hastings step. Unlike the other random walk updates, the sampler uses a basic version of Algorithm 4 from Andrieu and Thoms (2008) to adaptively tune the proposal standard deviation s_λ during estimation so the acceptance rate is close to 44%.

2.4 Weights for completely dependent random variables

2.4.1 Motivation for range of weights

We discuss the two special, limiting cases mentioned in Section 2.2.2 in more detail. If the field $\{Y_i(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ has complete dependence over space, or is spatially independent, then the likelihood weights (2.5) yield the same statistical information about GEV parameters $\boldsymbol{\eta}$ as non-misspecified likelihoods, which fully account for extremal dependence. We justify this claim by showing that our weighted likelihood (2.4) is equivalent to non-misspecified likelihoods (2.8) in

these special, limiting cases. We also make an informal argument that our weighted likelihood will approximate non-misspecified likelihoods in neighborhoods of these limiting cases.

Let the vector $\boldsymbol{\gamma} \in \mathbb{R}^p$ be used generically to parameterize extremal dependence in the field $\{Y_i(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ at time block $i \in \mathscr{T}$. Similarly, let the non-misspecified likelihood for observations $\{y_i(\boldsymbol{s}_j) : i \in \mathscr{T}, \boldsymbol{s}_j \in \mathscr{S}\}$ be defined via

$$L(\boldsymbol{\eta}, \boldsymbol{\gamma}) = \prod_{i=1}^{T} f\big(y_i(\boldsymbol{s}_1), \ldots, y_i(\boldsymbol{s}_N) \big| \boldsymbol{\eta}, \boldsymbol{\gamma}\big). \tag{2.8}$$

Assume the joint density $f\big(y_i(\boldsymbol{s}_1), \ldots, y_i(\boldsymbol{s}_N) \big| \boldsymbol{\eta}, \boldsymbol{\gamma}\big)$ is continuous with respect to $\boldsymbol{\gamma}$, and let the limiting conditions $\|\boldsymbol{\gamma}\| \to 0$ and $\|\boldsymbol{\gamma}\| \to \infty$ respectively parameterize fields that have no extremal dependence, and complete extremal dependence across space.

Alternative to the likelihood weights (2.5) we propose using, likelihood pseudo-weights $\left\{\tilde{w}_{\boldsymbol{s}_j,\boldsymbol{\gamma}} : j = 1, \ldots, N\right\}$ can be explicitly constructed to allow the non-misspecified likelihood (2.8) to be written in a weighted marginal form, such as

$$\prod_{i=1}^{T} f\big(y_i(\boldsymbol{s}_1), \ldots, y_i(\boldsymbol{s}_N) \big| \boldsymbol{\eta}, \boldsymbol{\gamma}\big) = \prod_{j=1}^{N} \prod_{i=1}^{T} f\big(y_i(\boldsymbol{s}_j) \big| \boldsymbol{\eta}(\boldsymbol{s}_j)\big)^{\tilde{w}_{\boldsymbol{s}_j,\boldsymbol{\gamma}}}. \tag{2.9}$$

The pseudo-weights we will construct are purely theoretical tools because they cannot be computed in practice. Using pseudo-weights to express the non-misspecified likelihood (2.8) as a weighted marginal likelihood (2.9) implies the weighted likelihood we propose (2.4) will yield the same inference as non-misspecified likelihoods when our likelihood weights (2.5) are equivalent to the pseudo-weights. We refer to the alternative weights $\left\{\tilde{w}_{\boldsymbol{s}_j,\boldsymbol{\gamma}} : j = 1, \ldots, N\right\}$ as pseudo-weights because we will define them shortly in (2.12) to depend on the joint conditional density $f\big(y_i(\boldsymbol{s}_1), \ldots, y_i(\boldsymbol{s}_N) \big| \boldsymbol{\eta}, \boldsymbol{\gamma}\big)$, which is not computationally tractable for spatially-referenced extremes data with $N > 10$, for example (Davison et al., 2012). Furthermore, computable pseudo-weights imply the joint density is available, thus the non-misspecified likelihood (2.8) may be used directly for inference and weighted likelihoods are unnecessary.

The likelihood pseudo-weights in (2.9) can be constructed in two parts. Begin by defining temporally-indexed weights $\{\tilde{w}_{i,\boldsymbol{\gamma}} : i = 1,\ldots,T\}$ that solve

$$f\big(y_i(\boldsymbol{s}_1),\ldots,y_i(\boldsymbol{s}_N)\big|\,\boldsymbol{\eta},\boldsymbol{\gamma}\big) = \prod_{j=1}^{N} f\big(y_i(\boldsymbol{s}_j)\big|\,\boldsymbol{\eta}(\boldsymbol{s}_j)\big)^{\tilde{w}_{i,\boldsymbol{\gamma}}}$$

for each $i \in \mathcal{T}$ via

$$\tilde{w}_{i,\boldsymbol{\gamma}} = \frac{\ln f\big(y_i(\boldsymbol{s}_1),\ldots,y_i(\boldsymbol{s}_N)\big|\,\boldsymbol{\eta},\boldsymbol{\gamma}\big)}{\sum_{j=1}^{N}\ln f\big(y_i(\boldsymbol{s}_j)\big|\,\boldsymbol{\eta}(\boldsymbol{s}_j)\big)}. \tag{2.10}$$

That is, $\tilde{w}_{i,\boldsymbol{\gamma}}$ is the ratio of the log-likelihood contribution in (2.8) at time i to the log-likelihood contribution from the marginal likelihoods, which assume conditional independence. The temporally-indexed weights (2.10) allow the likelihood (2.8) to be rewritten as

$$L(\boldsymbol{\eta},\boldsymbol{\gamma}) = \prod_{j=1}^{N}\prod_{i=1}^{T} f\big(y_i(\boldsymbol{s}_j)\big|\,\boldsymbol{\eta}(\boldsymbol{s}_j)\big)^{\tilde{w}_{i,\boldsymbol{\gamma}}}. \tag{2.11}$$

The desired likelihood weights $\big\{\tilde{w}_{\boldsymbol{s}_j,\boldsymbol{\gamma}} : j = 1,\ldots,N\big\}$, which are spatially-indexed, allow substitution of the inner product in (2.11) over $i = 1,\ldots,T$ by solving

$$\prod_{i=1}^{T} f\big(y_i(\boldsymbol{s}_j)\big|\,\boldsymbol{\eta}(\boldsymbol{s}_j)\big)^{\tilde{w}_{\boldsymbol{s}_j,\boldsymbol{\gamma}}} = \prod_{i=1}^{T} f\big(y_i(\boldsymbol{s}_j)\big|\,\boldsymbol{\eta}(\boldsymbol{s}_j)\big)^{\tilde{w}_{i,\boldsymbol{\gamma}}}$$

for each $\boldsymbol{s}_j \in \mathcal{S}$ via

$$\tilde{w}_{\boldsymbol{s}_j,\boldsymbol{\gamma}} = \frac{\sum_{i=1}^{T}\tilde{w}_{i,\boldsymbol{\gamma}}\ln f\big(y_i(\boldsymbol{s}_j)\big|\,\boldsymbol{\eta}(\boldsymbol{s}_j)\big)}{\sum_{i=1}^{T}\ln f\big(y_i(\boldsymbol{s}_j)\big|\,\boldsymbol{\eta}(\boldsymbol{s}_j)\big)}. \tag{2.12}$$

The likelihood weights we propose (2.5) converge to the pseudo-weights (2.12) as the extremal dependence approaches the special, limiting cases we consider. The extremal coefficient (2.1), combined with continuity of the joint density $f\big(y_i(\boldsymbol{s}_1),\ldots,y_i(\boldsymbol{s}_N)\big|\,\boldsymbol{\eta},\boldsymbol{\gamma}\big)$ with respect to $\boldsymbol{\gamma}$ imply our likelihood weights (2.5) satisfy $w_{\boldsymbol{s}_j} \to 1$ and $w_{\boldsymbol{s}_j} \to 1/N$, respectively as $\|\boldsymbol{\gamma}\| \to 0$ and $\|\boldsymbol{\gamma}\| \to \infty$. The pseudo-weights satisfy the same properties. In the first special case, con-

vergence $\tilde{w}_{\boldsymbol{s}_j,\boldsymbol{\gamma}} \to 1$ as $\|\boldsymbol{\gamma}\| \to 0$ is immediate because the joint density converges to a product of independent densities $f\big(y_i(\boldsymbol{s}_1),\ldots,y_i(\boldsymbol{s}_N)\big|\boldsymbol{\eta},\boldsymbol{\gamma}\big) \to \prod_{j=1}^{N} f\big(y_i(\boldsymbol{s}_j)\big|\boldsymbol{\eta}(\boldsymbol{s}_j)\big)$. In the second special case, $\|\boldsymbol{\gamma}\| \to \infty$, convergence $\tilde{w}_{\boldsymbol{s}_j,\boldsymbol{\gamma}} \to 1/N$ can be seen since the limiting joint density factors as

$$f\big(y_i(\boldsymbol{s}_1),\ldots,y_i(\boldsymbol{s}_N)\big|\boldsymbol{\eta}\big) = \prod_{j=1}^{N} f\big(y_i(\boldsymbol{s}_j)\big|\boldsymbol{\eta}(\boldsymbol{s}_j)\big)^{1/N} \times \mathbb{1}\left\{F\big(y_i(\boldsymbol{s}_1)\big|\boldsymbol{\eta}(\boldsymbol{s}_1)\big) = \cdots = F\big(y_i(\boldsymbol{s}_N)\big|\boldsymbol{\eta}(\boldsymbol{s}_N)\big)\right\}$$

for certain configurations of marginal density parameters, such as when the data have common marginals. This is a result of Corollary 2.4.1.1, presented and proved in Section 2.4.2.

Convergence of the likelihood weights we propose (2.5) to the pseudo-weights (2.12) implies that the weighted likelihood we propose (2.4) also converges to the non-misspecified likelihood (2.8). Furthermore, convergence of the likelihoods allows us to informally claim that inference based on the two different likelihoods will be similar when data is sampled from a process with extremal dependence $\boldsymbol{\gamma}$ in a neighborhood of the limiting cases $\|\boldsymbol{\gamma}\| \to 0$ and $\|\boldsymbol{\gamma}\| \to \infty$.

2.4.2 Theoretical results

Likelihood weights (2.10) and (2.12) are defined with respect to joint density functions. Completely dependent random variables have been studied in detail in the insurance industry, where they are referred to as comonotonic random variables. However, their joint density is not usually considered. Comonotonic random variables serve as basic models of worst-case scenarios for insurance portfolios, which makes the sum of comonotonic random variables more important than their joint distribution (Dhaene et al., 2002). To compute likelihood weights for completely dependent, or comonotonic variables, we first derive their joint density in Theorem 2.4.1. Proofs for all results are presented in Section 2.4.3.

Theorem 2.4.1. *For any $j \in \{1,\ldots,N\}$, the joint density $g(x_1,\ldots,x_N)$ for a vector of comonotonic random variables $(X_1,\ldots,X_N)$ can be parameterized as*

$$g(x_1,\dots,x_N) = f_j\big(x_j\big)\mathbb{1}(F_1(x_1) = \cdots = F_N(x_N)) \tag{2.13}$$

where F_j and f_j respectively denote the cumulative distribution function and density for X_j relative to Lebesgue measure on $\mathbb{R}$. The density (2.13) is defined with respect to the dominating measure $\lambda_N + \lambda_{\mathscr{C}}$ for which λ_N is Lebesgue measure on $\mathbb{R}^N$ and $\lambda_{\mathscr{C}}$ is Lebesgue measure on $\mathscr{C} = \{(x_1,\dots,x_N) : F_1(x_1) = \cdots = F_N(x_N)\} \subset \mathbb{R}^N$.

While Theorem 2.4.1 allows the likelihood weights (2.10) and (2.12) to be computed, developing intuition requires additional theory because the density (2.13) only explicitly includes one density f_j. Lemma 2.4.1 will allow the likelihood weights to be manipulated by providing a means to express f_i as a rescaling of f_j for $i \neq j$.

Lemma 2.4.1. *For a vector of comonotonic random variables $(X_1,\dots,X_N)$, the marginal density $f_i(x_i)$ for X_i may be re-expressed in terms of $f_j\big(x_j\big)$ using function composition $\circ$ and the quantile density function $q(u) = \frac{\partial}{\partial u}F^{-1}(u)$ through*

$$f_i(x_i) = f_j\big(x_j\big)\frac{\big(q_j \circ F_j\big)\big(x_j\big)}{\big(q_i \circ F_j\big)\big(x_j\big)} \tag{2.14}$$

for any $j = 1,\dots,N$, continuous F_i, F_j, and x_j s.t. $F_i(x_i) = F_j\big(x_j\big)$.

Intuition for the likelihood weights (2.10) and (2.12) follows from algebraic manipulation. In particular, Corollary 2.4.1.1 yields conditions under which likelihood weights are intuitive (e.g., $w_{(j)} = N^{-1}$), such as when $X_1,\dots,X_N$ have common marginals.

Corollary 2.4.1.1. *For any $j \in \{1,\dots,N\}$, the likelihood weight (2.10) for a single vector of comonotonic random variables $(X_1,\dots,X_N)$ is*

$$w_{(j)} = \frac{1}{N + d_{(j)}} \tag{2.15}$$

where

$$d_{(j)} = \frac{1}{\ln f_j(x_j)} \sum_{i=1}^{N} \ln \frac{(q_j \circ F_j)(x_j)}{(q_i \circ F_j)(x_j)}$$

and the subscript highlights the dependence of the weight $w_{(j)}$ on the density f_j used to parameterize the comonotonic density g. The average weight across all parameterizations is

$$\bar{w} = \frac{1}{N} \sum_{j=1}^{N} w_{(j)} = \frac{1}{N}.$$

2.4.3 Proofs of theoretical results

Completely dependent densities: Proof of Theorem 2.4.1

A random vector of comonotonic variables $(X_1, \ldots, X_N)$ has support $\mathscr{C}$ and cumulative distribution function (CDF) given by

$$F(x_1, \ldots, x_N) = \min_{i \in \{1,\ldots,N\}} F_i(x_i)$$

where F_i is the CDF for X_i (Dhaene et al., 2002, Theorem 2). The CDF F and support $\mathscr{C}$ imply the probability measure $\mathscr{P}$ associated with F is absolutely continuous with respect to $\lambda_N + \lambda_{\mathscr{C}}$. Integrating (2.13) over a half-infinite rectangle
$A = \{(y_1, \ldots, y_N) : y_i \leq x_i, i = 1, \ldots, N\}$ yields $F(x_1, \ldots, x_N)$ since

$$\int_A g(y_1, \ldots, y_N) d(\lambda_N + \lambda_{\mathscr{C}}) = \int_{-\infty}^{F_j^{-1}(\min_i F_i(x_i))} f_j(y) dy \tag{2.16}$$

$$= P\left(U \leq \min_i F_i(x_i)\right), \; U \sim U(0,1) \tag{2.17}$$

$$= \min_{i \in \{1,\ldots,N\}} F_i(x_i). \tag{2.18}$$

The integral (2.16) simplifies because g is measure-0 with respect to λ_N and a 1:1 mapping exists between $A \cap \mathscr{C}$ and $\mathbb{R}$ since for any $j \in \{1, \ldots, N\}$

$$
\begin{aligned}
A \cap \mathscr{C} &= \{(y_1, \ldots, y_N) : F_1(y_1) = \cdots = F_N(y_N);\ F_i(y_i) \leq F_i(x_i), i = 1, \ldots, N\} \\
&= \left\{(y_1, \ldots, y_N) : F_1(y_1) = \cdots = F_N(y_N);\ F_j(y_j) \leq \min_i F_i(x_i)\right\} \\
&= \left\{(y_1, \ldots, y_N) : F_1(y_1) = \cdots = F_N(y_N);\ y_j \leq F_j^{-1}\left(\min_i F_i(x_i)\right)\right\}.
\end{aligned}
$$

The probability integral transformation yields (2.17), from which (2.18) naturally follows. The Radon-Nikodym theorem and general properties of distribution functions imply g is a density for $(X_1, \ldots, X_N)$ with respect to $\lambda_N + \lambda_{\mathscr{C}}$.

Rescaled marginal densities: Proof of Lemma 2.4.1

The rescaling (2.14) uses the identity

$$
(f \circ Q)(u)\, q(u) = 1, u \in [0,1] \tag{2.19}
$$

in which $Q(u) = F^{-1}(u)$ is the quantile function for a continuous CDF F (Parzen, 1979, eqn. 2.6). The support constraint $F_i(x_i) = F_j(x_j)$ implies $x_i = (Q_i \circ F_j)(x_j)$ and allows $f_i(x_i)$ to be rewritten as

$$
f_i(x_i) = (f_i \circ Q_i \circ F_j)(x_j). \tag{2.20}
$$

The desired result (2.14) follows from applying the identity (2.19) twice to (2.20) since

$$
\begin{aligned}
f_i(x_i) &= \{(q_i \circ F_j)(x_j)\}^{-1} \\
&= \frac{(f_j \circ Q_j \circ F_j)(x_j)(q_j \circ F_j)(x_j)}{(q_i \circ F_j)(x_j)} \\
&= f_j(x_j) \frac{(q_j \circ F_j)(x_j)}{(q_i \circ F_j)(x_j)}.
\end{aligned}
$$

Completely dependent weights: Proof of Corollary 2.4.1.1

Theorem 2.4.1 implies the likelihood weight (2.10) for a single vector of comonotonic random variables $(X_1, \ldots, X_N)$ is

$$w_{(j)} = \frac{\ln f_j(x_j)}{\sum_{i=1}^{N} \ln f_i(x_i)} \tag{2.21}$$

for all $(x_1, \ldots, x_N)$ that satisfy the support constraint $F_1(x_1) = \cdots = F_N(x_N)$. Lemma 2.4.1 yields the first result (2.15) as it lets us re-express the denominator of (2.21) in terms of $f_j(x_j)$ as

$$\begin{aligned}\sum_{i=1}^{N} \ln f_i(x_i) &= \sum_{i=1}^{N} \left(\ln f_j(x_j) + \ln \frac{(q_j \circ F_j)(x_j)}{(q_i \circ F_1)(x_j)} \right) \\ &= \ln f_j(x_j) \left(N + d_{(j)} \right)\end{aligned}$$

where

$$d_{(j)} = \frac{1}{\ln f_j(x_j)} \sum_{i=1}^{N} \ln \frac{(q_j \circ F_j)(x_j)}{(q_i \circ F_j)(x_j)}.$$

The average weight $\bar{w}$ follows directly from (2.21) since

$$\bar{w} = \frac{1}{N} \sum_{j=1}^{N} \frac{\ln f_j(x_j)}{\sum_{i=1}^{N} \ln f_i(x_i)} = \frac{1}{N}.$$

2.5 Simulation study

We use simulation to show that the weighted marginal likelihood (2.4) improves high quantile estimates on datasets with realistic GEV parameters $\boldsymbol{\eta}(\boldsymbol{s})$, sample sizes, and varying extremal dependence. The simulation compares the weighted likelihood model (Section 2.5.2) to a standard, unweighted latent spatial extremes model and a penalized variation. Penalization is an alternate approach used to correct return level estimates in extreme value models (cf. Opitz et al., 2018; Schliep et al., 2010). Penalized models have hierarchical structures that are similar to our weighted likelihood, so are comparison models with similar computational complexity to our weighted likelihood. We compare models by contrasting properties of estimators of high quantiles, including empirical coverage and mean squared error (Section 2.5.4).

2.5.1 Datasets

We simulate data from four generating models with varying combinations of extremal dependence, and spatial N and temporal T sample sizes. Properties of parameter estimators are empirically approximated using 1,000 datasets simulated from each generating model. Our decision to vary extremal dependence, N, and T is informed by the Fisher information (2.6) and effective sample size discussion (Section 2.2.2), which provide intuition about how extremal dependence and sample size affect estimation. Increasing extremal dependence decreases the amount of statistical information available for parameter estimation, much as occurs with classical spatial dependence (Cressie, 1993, Section 1.3). Similarly, the impact of extremal dependence increases when sampling more spatial locations $\mathscr{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\} \subset \mathscr{D}$ from a fixed domain $\mathscr{D}$. Unweighted latent spatial extremes models are misspecified when data are dependent because they assume the data are conditionally independent given model parameters. The severity of the misspecification increases as the process is observed at more spatial locations N because it becomes more likely that observations from spatially-dependent locations are included in the sample. The Fisher information equation (2.6), however, suggests that statistical information about the marginal parameters increases with the number of replications T despite misspecification, albeit at a slower rate when using likelihood weights.

Simulated data have marginal GEV parameters $\boldsymbol{\eta}(\boldsymbol{s})$ that mimic estimates from observed annual maximum daily precipitation across Colorado's Front Range (Tye and Cooley, 2015). Spatially-dependent GEV parameters $\boldsymbol{\eta}(\boldsymbol{s})$, $\boldsymbol{s} \in \mathscr{D} = [-10, 10]^2$ are sampled from Gaussian processes $\text{GP}\big(m, \rho\big)$ with mean functions $m : \mathscr{D} \to \mathbb{R}$ and powered exponential covariances $\rho : \mathscr{D}^2 \to [0, \infty)$ specified in Table 2.1. Shape parameters $\xi(\boldsymbol{s})$ are resampled until $\xi(\boldsymbol{s}) > 0$ for all $\boldsymbol{s} \in \mathscr{S}$ to ensure data are heavy-tailed. Brown–Resnick processes model extremal dependence in the simulated data (Kabluchko et al., 2009). The semi-variogram $\gamma : \mathscr{D}^2 \to [0, \infty)$ specified in Table 2.1 parameterizes a Brown–Resnick model that induces strong, medium, or weak extremal dependence on $\mathscr{D}$ as measured by the extremal coefficient function $\theta(d)$. For comparison, independent data are also simulated.

Table 2.1: Generating model configurations used to simulate data for comparing the weighted likelihood (2.4) to alternate estimating models (Section 2.5.2). We evaluate model performance with 1,000 datasets for each combination of spatial N and temporal T sample sizes, and extremal dependence.

Spatial sample size	$N \in \{30, 50, 100\}$ sites sampled uniformly on $\mathscr{D} = [-10, 10]^2$	
Temporal sample size	$T \in \{50, 100\}$	
Extremal dependence *(Brown-Resnick parameters)*	Semi-variogram	$\gamma_{(\lambda,\, \alpha)}(\boldsymbol{s}_1, \boldsymbol{s}_2) = (\|\boldsymbol{s}_1 - \boldsymbol{s}_2\| / \lambda)^{\alpha}$
	Independent	$(\lambda = \text{NA},\ \alpha = \text{NA})$
	Weak	$(\lambda = .25,\ \alpha = .75)$
	Moderate	$(\lambda = .5,\ \alpha = .5)$
	Strong	$(\lambda = .75,\ \alpha = .25)$
Prior distributions for GEV parameters $\boldsymbol{\eta}(\boldsymbol{s})$	Covariance function $\rho_{(\sigma_0,\, \lambda_0,\, \nu_0)}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \sigma_0 \exp\left\{-(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\| / \lambda_0)^{\nu_0}\right\}$	
	Gaussian processes	
	$\mu(\boldsymbol{s}) \sim \text{GP}\left(26 + [.5\ \ 0]^T \boldsymbol{s},\ \rho_{(4,\, 20,\, 1)}\right)$	
	$\log\sigma(\boldsymbol{s}) \sim \text{GP}\left(\log(10) + [0\ \ .05]^T \boldsymbol{s},\ \rho_{(.4,\, 5,\, 1)}\right)$	
	$\xi(\boldsymbol{s}) \sim \text{GP}\left(.12,\ \rho_{(.0012,\, 10,\, 1)}\right)$	

2.5.2 Estimating models

The simulation compares estimation of conditionally independent models with weighted (2.4) and unweighted likelihoods (i.e., (2.4) with $w_{\boldsymbol{s}_j} = 1$ for all $\boldsymbol{s}_j \in \mathscr{S}$) and a variation that uses penalized complexity priors as a likelihood penalty (Section 2.5.2). Key differences between the estimating models are summarized in Table 2.2. The comparison models represent different approaches proposed in the extremes literature to improve marginal estimation of GEV parameters and have similar computational complexity.

Penalized complexity prior

Likelihood-based parameter estimates for the univariate GEV distribution are known to perform poorly, but penalized likelihoods can reduce estimation bias (Coles and Dixon, 1999; Martins and Stedinger, 2000). Penalized likelihoods have been incorporated into spatial models for marginal extremes (Opitz et al., 2018; Schliep et al., 2010). Penalization improves estimation of marginal parameters by downweighting estimates of large shape parameters $\xi(\boldsymbol{s})$, which tend to be uncommon in many extreme precipitation data. We adapt a contemporary penalty for use with the GEV distribution as a comparison model.

Penalized complexity (PC) priors have recently been proposed to improve parameter estimation in a related extreme value family—the Generalized Pareto distribution (GPD), which also uses scale $\sigma(\boldsymbol{s}) > 0$, and shape $\xi(\boldsymbol{s}) \in \mathbb{R}$ parameters to model threshold exceedances (Opitz et al., 2018). Penalized complexity priors satisfy several properties that optimize the prior distribution's shape and scale to precisely control the prior's influence over target likelihoods (Simpson et al., 2017). We implement PC priors as penalized likelihoods in our hierarchical spatial model. We derive the penalized complexity prior $\pi(\xi|\lambda)$ for the GEV distribution below and use it with the log-likelihood

$$\ell(\boldsymbol{\eta}) = \sum_{j=1}^{N}\sum_{i=1}^{T} \log f\big(y_i(\boldsymbol{s}_j)\big|\boldsymbol{\eta}(\boldsymbol{s}_j)\big) + \sum_{j=1}^{N} \log \pi\big(\xi(\boldsymbol{s}_j)\big|\lambda\big) \tag{2.22}$$

in place of the log of the unweighted version of the likelihood (2.4), in which $w_{\boldsymbol{s}_j} = 1$ for all $\boldsymbol{s}_j \in \mathscr{S}$.

Bayesian estimation optimizes the PC prior's parameterization by specifying an Inverse-gamma prior distribution for $\lambda \sim \text{IG}(2,1)$. The Inverse-gamma distribution is parameterized to have mean 1 and infinite variance. Prior distributions provide an alternative to cross-validation approaches for optimizing the prior's parameterization, which is computationally infeasible for this simulation study (Hans, 2009; Park and Casella, 2008).

Following Simpson et al. (2017), the penalized complexity prior for the generalized extreme value (GEV) distribution (2.2) is defined through the prior density

$$\pi(\xi \mid \lambda) = \lambda e^{-\lambda d(\xi)} \left| \frac{\partial d(\xi)}{\partial \xi} \right| \tag{2.23}$$

with tuning parameter $\lambda > 0$ and "distance" function $d(\xi) = \sqrt{2\,\text{KLD}\left(f_\xi \| f_{\xi_0}\right)}$. The distance function $d(\xi)$ is based on the Kullback-Leibler divergence $\text{KLD}\left(f_\xi \| f_{\xi_0}\right)$ between the GEV distribution with shape parameter ξ and reference shape parameter ξ_0. The penalized complexity prior encourages shrinkage of the shape parameter ξ toward the reference parameter ξ_0. A natural choice for the reference parameter is $\xi_0 = 0$, the point at which the GEV distribution changes from having a heavy tail ($\xi > 0$) to a light tail ($\xi < 0$). The Kullback-Leibler divergence is

$$\begin{aligned}
\text{KLD}\left(f_\xi \| f_{\xi_0}\right) &= \int_{\mathbb{S}} f_\xi(y) \log \frac{f_\xi(y)}{f_{\xi_0}(y)} dy \\
&= \int_{\mathbb{S}} \frac{1}{\sigma} t_\xi(y)^{\xi+1} \exp\{-t_\xi(y)\} \log\left(\frac{t_\xi(y)^{\xi+1} \exp\{-t_\xi(y)\}}{e^{-(y-\mu)/\sigma} \exp\{-e^{-(y-\mu)/\sigma}\}} \right) dy \\
&= (\xi+1)\psi(1) - 1 + (\Gamma(1-\xi) - 1)/\xi + \exp\{1/\xi\}\, \text{I}_\xi
\end{aligned} \tag{2.24}$$

where $t_\xi(y) = \left(1 + \xi(y-\mu)/\sigma\right)^{-1/\xi}$, $\psi(\cdot)$ is the digamma function, $I_\xi = \int_0^\infty \exp\left\{-\left(\xi s^\xi\right)^{-1}\right\} e^{-s} ds$, and $\mathbb{S} = \left[\mu - \sigma/\xi, \infty\right)$ is the distribution's support for $\xi > 0$. When $\xi < 0$, the support is reversed $\mathbb{S} = \left(-\infty, \mu - \sigma/\xi\right]$. The definite integral I_ξ does not simplify analytically but includes the e^{-s} "weight function" so can be efficiently approximated numerically

with Gauss-Laguerre quadrature (Givens and Hoeting, 2013, Section 5.3). The Kullback-Leibler divergence is trivially zero when $\xi = 0$ but otherwise expands to (2.24), the sum of the integrals (2.25) to (2.28). The first integral (2.25) uses the substitution $s = \log t_\xi(y)$, yielding

$$(\xi+1)\int_{\mathbb{S}} \frac{1}{\sigma} t_\xi(y)^{\xi+1} \exp\{-t_\xi(y)\} \log t_\xi(y)\, dy = (\xi+1)\int_{\mathbb{R}} s \exp\{s - e^s\}\, ds = (\xi+1)\psi(1). \tag{2.25}$$

The transformed integral in (2.25) represents -1 times the expected value for a standard Gumbel random variable, allowing simplification. The second integral (2.26) uses the substitution $s = t_\xi(y)$, yielding

$$-\int_{\mathbb{S}} \frac{1}{\sigma} t_\xi(y)^{\xi+1} \exp\{-t_\xi(y)\}\, t_\xi(y)\, dy = -\int_0^\infty s e^{-s} ds = -1. \tag{2.26}$$

For $\xi < 1$ and $\xi \neq 0$, the third integral (2.27) is exactly equivalent to

$$\mathrm{E}_\xi\left[(y-\mu)/\sigma\right] = (\Gamma(1-\xi) - 1)/\xi. \tag{2.27}$$

The last integral (2.28) also uses the substitution $s = t_\xi(y)$, yielding

$$\int_{\mathbb{S}} \frac{1}{\sigma} t_\xi(y)^{\xi+1} \exp\{-t_\xi(y)\} \exp\{-(y-\mu)/\sigma\}\, dy = \exp\{1/\xi\}\, I_\xi. \tag{2.28}$$

The penalized complexity prior (2.23) also uses the distance function's partial derivative $\frac{\partial}{\partial\xi} d(\xi) = \left(2\,\mathrm{KLD}\left(f_\xi \| f_{\xi_0}\right)\right)^{-1/2} \frac{\partial}{\partial\xi} \mathrm{KLD}\left(f_\xi \| f_{\xi_0}\right)$. While differentiating the Kullback-Leibler divergence (2.24) is straightforward

$$\frac{\partial}{\partial\xi} \mathrm{KLD}\left(f_\xi \| f_{\xi_0}\right) = \psi(1) - \frac{\Gamma(1-\xi)\left(\xi\psi(1-\xi)+1\right) - 1}{\xi^2} + \exp\{1/\xi\}\left(\frac{\partial}{\partial\xi} I_\xi - \frac{I_\xi}{\xi^2}\right),$$

evaluating the derivative also requires Gauss-Laguerre approximation of the definite integral

Table 2.2: Summary of differences between estimating models in simulation study (Section 2.5).

Model	*(Log-)Likelihood*	*Weights*	*Log-Likelihood penalty*
Unweighted	(2.4)	None	None
Weighted	(2.4)	(2.5)	None
PC Prior	(2.22)	None	$\sum_j \log \pi\big(\xi(\boldsymbol{s}_j)\big\vert\, \lambda\big)$

$$\frac{\partial}{\partial \xi} I_\xi = \int_0^\infty \frac{1+\xi \log s}{\xi^2 s^\xi} \exp\left\{-\left(\xi s^\xi\right)^{-1}\right\} e^{-s} ds.$$

2.5.3 Bayesian specification

All models use a hierarchical Bayesian framework in which the GEV parameters $\boldsymbol{\eta}(\boldsymbol{s})$ are estimated as independent latent Gaussian processes with functional forms matching those specified in Table 2.1. Prior distributions for the mean and covariance function parameters are either weakly informative or uninformative, and conjugate where possible. Inference is based on a sample from the posterior distribution, drawn with a Gibbs sampler. Estimators based on the weighted likelihood are evaluated with respect to both fixed and Gibbs-updated weights (See Section 2.3).

Prior distributions for regression coefficients $\boldsymbol{\beta}$ and spatial covariance parameters σ_0 and λ_0 used in the simulation study (Section 2.5) are specified in Table 2.3. The spatial smoothness ν_0 is fixed at the truth. Regression coefficient prior distributions are designed to be uninformative, while the spatial covariance prior distributions are designed to be weakly informative. The spatial covariance prior distributions have infinite or large variation. The distributions are parameterized such that the mean of the priors are centered at the true generating model parameters for the spatial range λ_0; the mode of the priors for the sill parameters σ_0 are centered at the true generating model parameters. The penalized complexity prior parameter has prior distribution $\lambda \sim \text{IG}(2,1)$. The proposal standard deviations for the random walk Metropolis-Hastings samplers are $s_\mu = 1.2$, $s_{\log\sigma} = .08$, $s_\xi = .08$; and s_{λ_0} is .7, .8. and .7, respectively for the GEV location $\{\mu(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$, scale $\{\log\sigma(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$, and shape $\{\xi(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ processes.

Table 2.3: Prior distributions used in simulation study (Section 2.5.2).

	GEV parameter process		
	$\{\mu(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$	$\{\log\sigma(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$	$\{\xi(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$
(Regression coefs.)			
$\boldsymbol{\beta}\sim$	$\mathscr{N}\left(\mathbf{0}, \begin{bmatrix} 400 & \\ & 100 \end{bmatrix}\right)$,	$\mathscr{N}\left(\mathbf{0}, \begin{bmatrix} 400 & \\ & 100 \end{bmatrix}\right)$,	$\mathscr{N}(0,100)$.
(Spatial covariance)			
$\sigma_0\sim$	IG$(1,8)$,	IG$(1,.8)$,	IG$(1,.0024)$.
$\lambda_0\sim$	Gamma$(2,10)$,	Gamma$(2,2.5)$,	Gamma$(2,5)$.

Sample autocorrelation diagnostics indicate the Gibbs sampler mixes slowly, so the sampler was run for 155,000 iterations to ensure Monte Carlo integration error is sufficiently small. The first 5,000 samples were discarded. Posterior inference uses a thinned posterior sample consisting of 10,000 of the remaining 150,000 samples; only every fifteenth sample was saved because the entire posterior sample could not be efficiently stored and manipulated. Thinning reduces statistical efficiency of Markov chain Monte Carlo methods, but can be a necessary tradeoff when the full posterior sample is difficult to store and use to estimate posterior quantities (MacEachern and Berliner, 1994).

2.5.4 Results

Assuming a stationary climate, the 1% annual exceedance probability $Q\big(.99|\boldsymbol{\eta}(\boldsymbol{s})\big)$ from (2.3), also referred to as the 100-year return level, is often used to quantify risk for extreme weather events. The weighted model's results are nearly identical when comparing fixed weights to Gibbs-updated weights. Figure 2.1 presents the empirical coverage of highest posterior density (HPD) intervals for the return level $Q\big(.99|\boldsymbol{\eta}(\boldsymbol{s})\big)$ for each of the models listed in Table 2.2. Figure 2.2 presents mean squared error (MSE) for the same data. Bias is small for all estimators, so MSE mainly quantifies estimator variance. Since the return level $Q\big(.99|\boldsymbol{\eta}(\boldsymbol{s})\big)$ is

greatly influenced by the shape parameter, $\xi(\boldsymbol{s})$, results for return levels and shape parameters are very similar.

Extremal dependence degrades the performance of all marginal models, but the weighted marginal likelihood (2.4) provides the most accurate estimates of uncertainty. Empirical coverage of 95% HPD intervals is closest to the nominal HPD level across all levels of extremal dependence. (Figure 2.1). For the $N = 50$, $T = 50$ simulation with moderate dependence, the weighted model has a coverage rate of 86%, while the unweighted model and penalized complexity prior model have coverage rates of 83% and 82% respectively. In the same scenario, the weighted model also has nearly identical MSE as the other models, although the MSE for the weighted likelihood model is somewhat greater for the simulation with strong dependence (Figure 2.2).

Figure 2.3 through Figure 2.14 present empirical coverage, mean squared error, and relative bias for all GEV parameters $\mu(\boldsymbol{s})$, $\sigma(\boldsymbol{s})$, and $\xi(\boldsymbol{s})$ for all combinations of estimation models and generating model configurations used in the simulation study described in Section 2.5. Relative bias is the estimator bias scaled by the truth, for example

$$\text{Rel. Bias}\big(\mu(\boldsymbol{s})\big) = \frac{E\left[\hat{\mu}(\boldsymbol{s}) - \mu(\boldsymbol{s})\right]}{\mu(\boldsymbol{s})} \times 100\%.$$

2.6 Extreme Colorado precipitation

2.6.1 Data

Previous studies of extreme precipitation in Colorado find that there is weak extremal dependence between locations along the state's Front Range region (Cooley et al., 2007; Tye and Cooley, 2015). We determine the impact the weighted likelihood (2.4) has on estimates of the 1% annual exceedance probability $Q\big(.99|\boldsymbol{\eta}(\boldsymbol{s})\big)$, also referred to as the 100-year return level. Estimates are based on the same subset of annual maxima of daily precipitation Tye and Cooley (2015) use from the Global Historical Climatology Network (GHCN) dataset (Menne et al., 2012). The subset includes annual maxima from 71 stations along the Front Range. Tye and Cooley

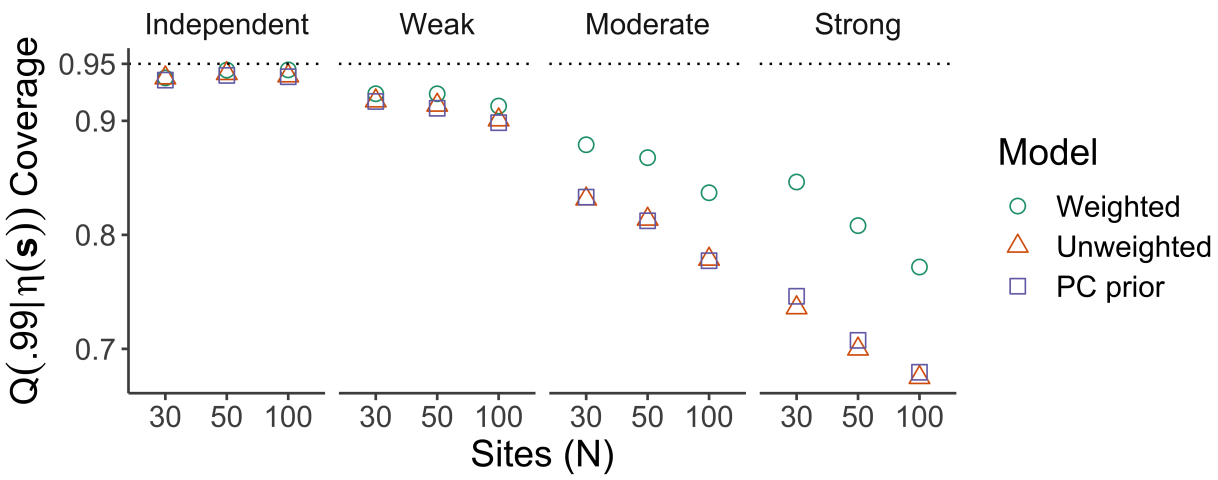

Figure 2.1: Empirical coverage rates of 95% highest posterior density intervals for 100-year return levels $Q\big(.99|\boldsymbol{\eta}(\boldsymbol{s})\big)$ for four levels of extreme dependence across comparison models and simulations with $T = 50$ observations per location. Nominal coverage is marked by the dotted horizontal reference line at .95. While empirical coverage degrades for all estimating models as extremal dependence increases, the weighted model is most robust to model misspecification caused by extremal dependence.

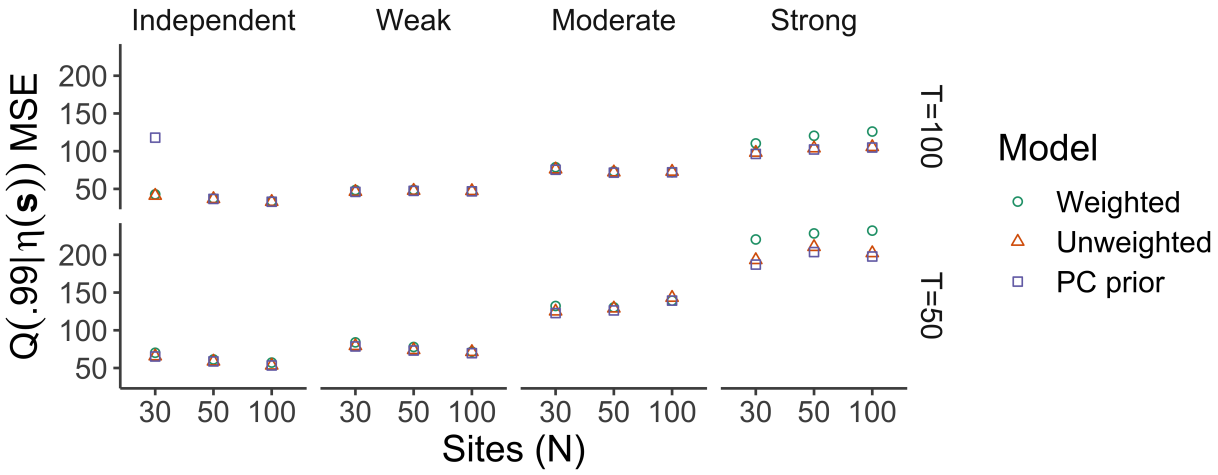

Figure 2.2: Empirical mean squared error (MSE) of posterior estimates for 100-year return levels $Q\big(.99|\boldsymbol{\eta}(\boldsymbol{s})\big)$ for four levels of extreme dependence across comparison models and simulations with $T = 50$ observations per location. The weighted model has similar or better performance than the standard, unweighted model in nearly all simulations. The unweighted model underestimates uncertainty, so has slightly smaller MSE for the simulation with strong extremal dependence.

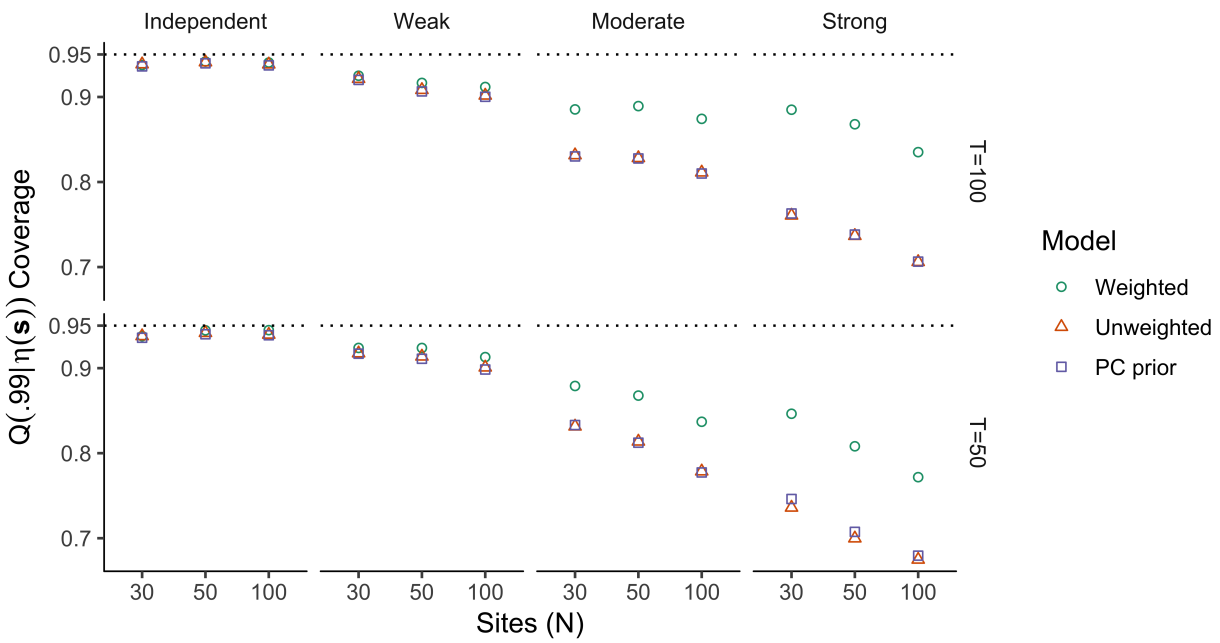

Figure 2.3: Empirical coverage rates of 95% highest posterior density intervals for 100-year return levels $Q\big(.99|\boldsymbol{\eta}(\boldsymbol{s})\big)$ across comparison models and all simulations. Nominal coverage is marked by the dotted horizontal reference line at .95.

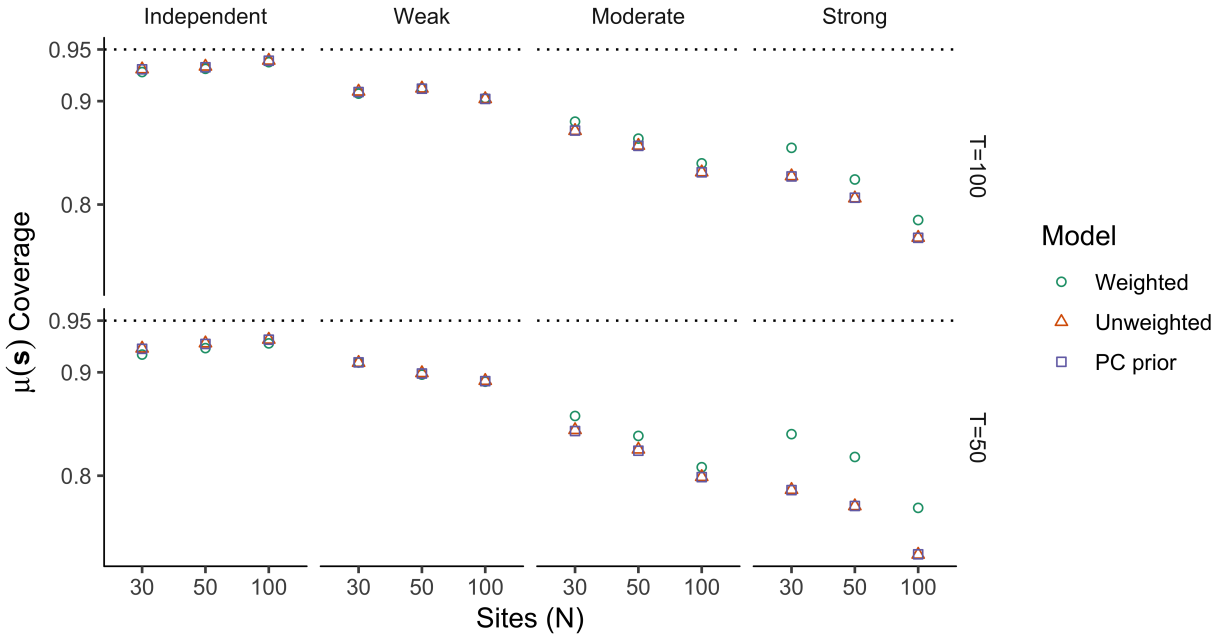

Figure 2.4: Empirical coverage rates of 95% highest posterior density intervals for GEV location parameters $\mu(\boldsymbol{s})$ across comparison models and all simulations. Nominal coverage is marked by the dotted horizontal reference line at .95.

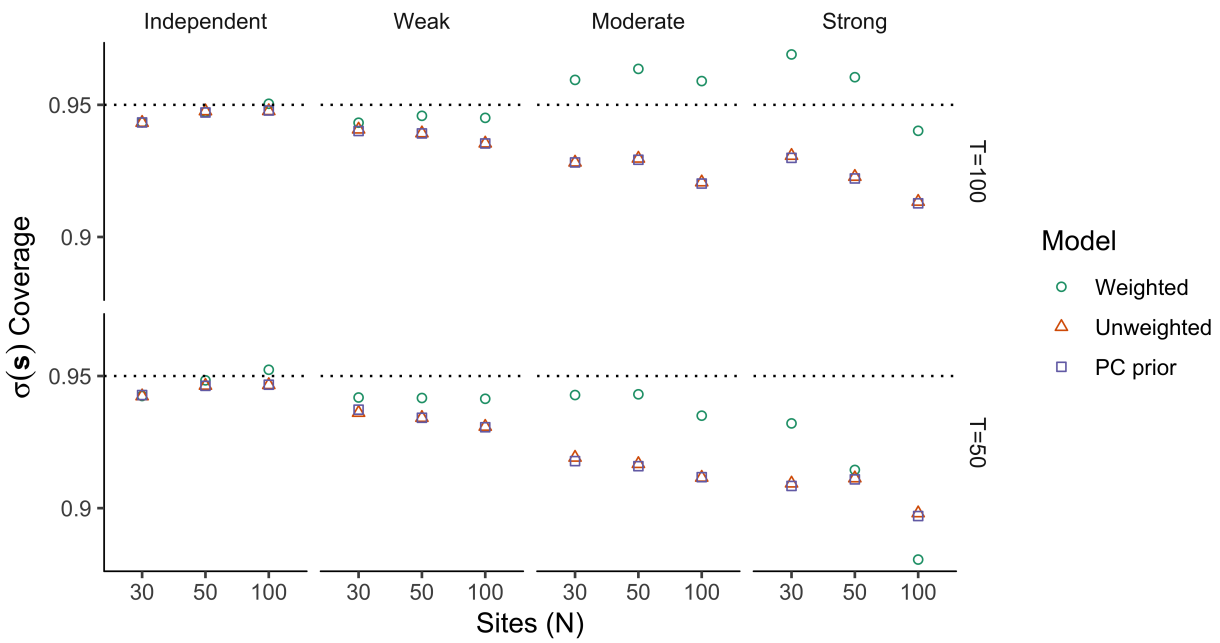

Figure 2.5: Empirical coverage rates of 95% highest posterior density intervals for GEV scale parameters $\sigma(\boldsymbol{s})$ across comparison models and all simulations. Nominal coverage is marked by the dotted horizontal reference line at .95.

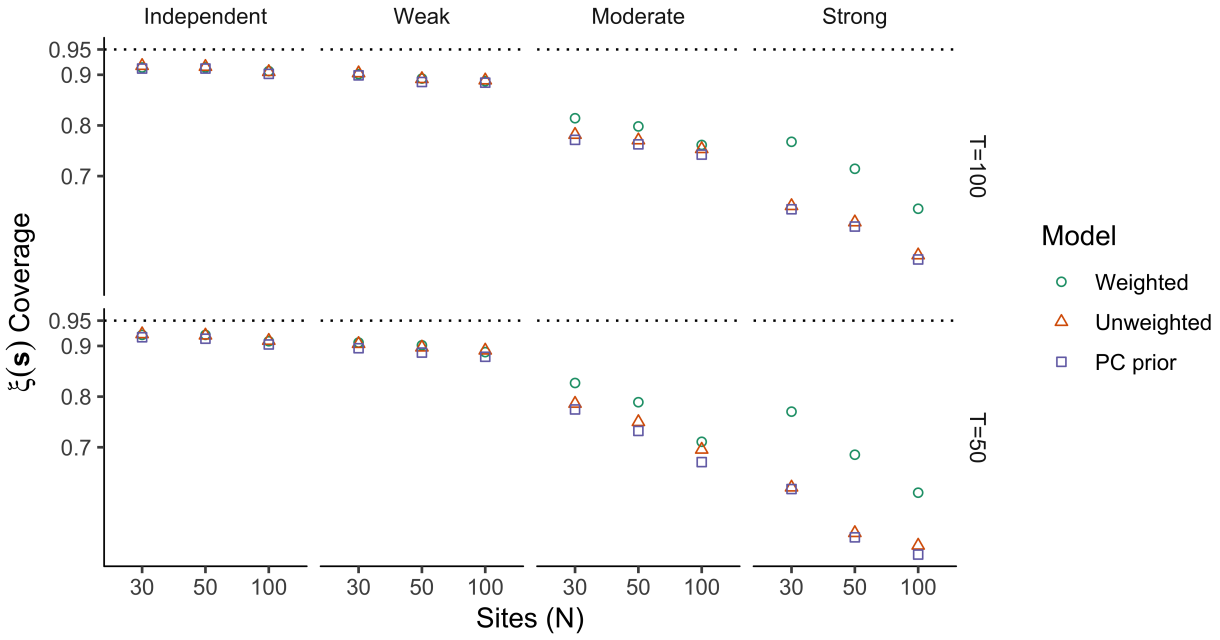

Figure 2.6: Empirical coverage rates of 95% highest posterior density intervals for GEV shape parameters $\xi(\boldsymbol{s})$ across comparison models and all simulations. Nominal coverage is marked by the dotted horizontal reference line at .95.

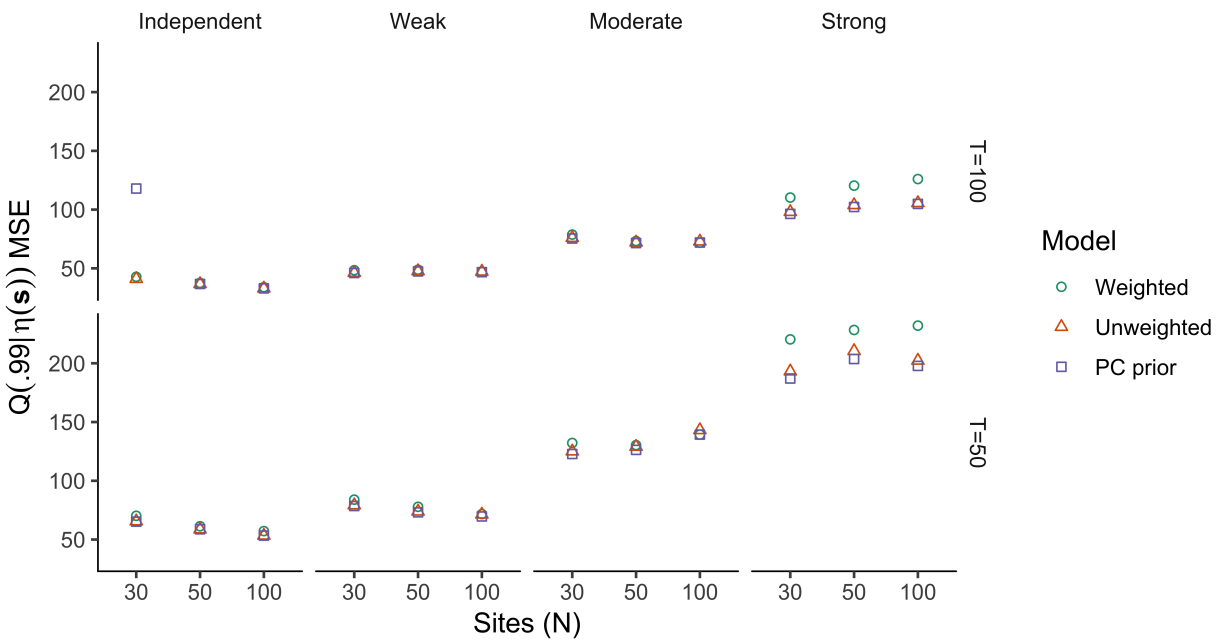

Figure 2.7: Empirical mean square errors (MSE) of posterior estimates for 100-year return levels $Q\big(.99|\boldsymbol{\eta}(\boldsymbol{s})\big)$ across comparison models and all simulations.

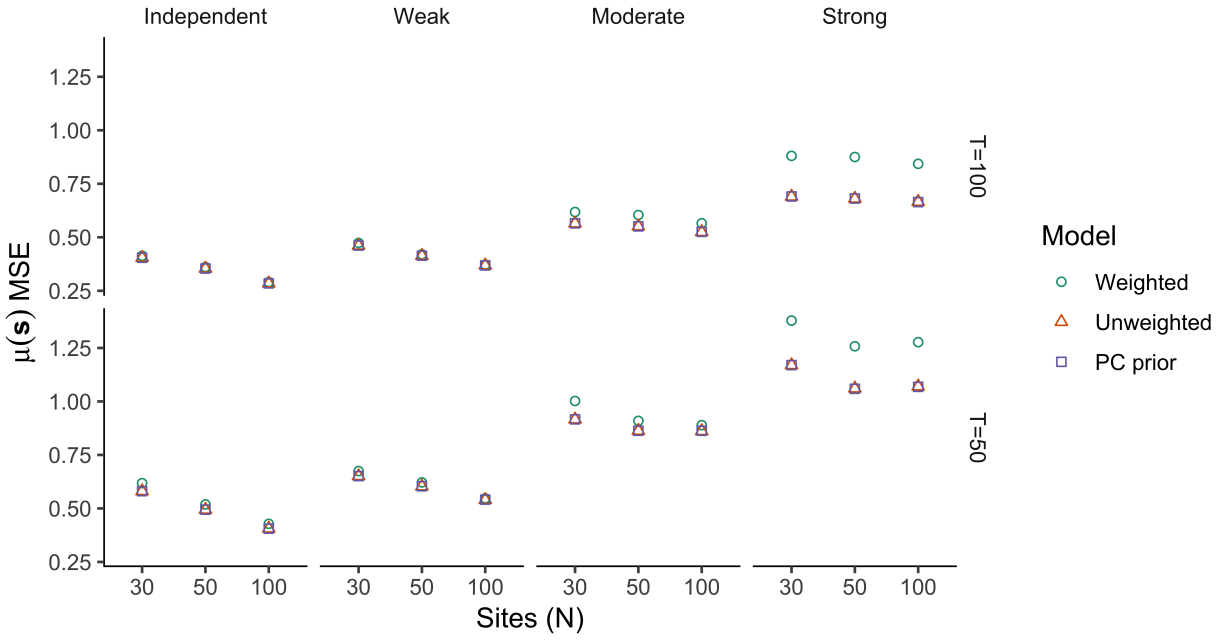

Figure 2.8: Empirical mean square error (MSE) of posterior estimates for GEV location parameters $\mu(\boldsymbol{s})$ across comparison models and all simulations.

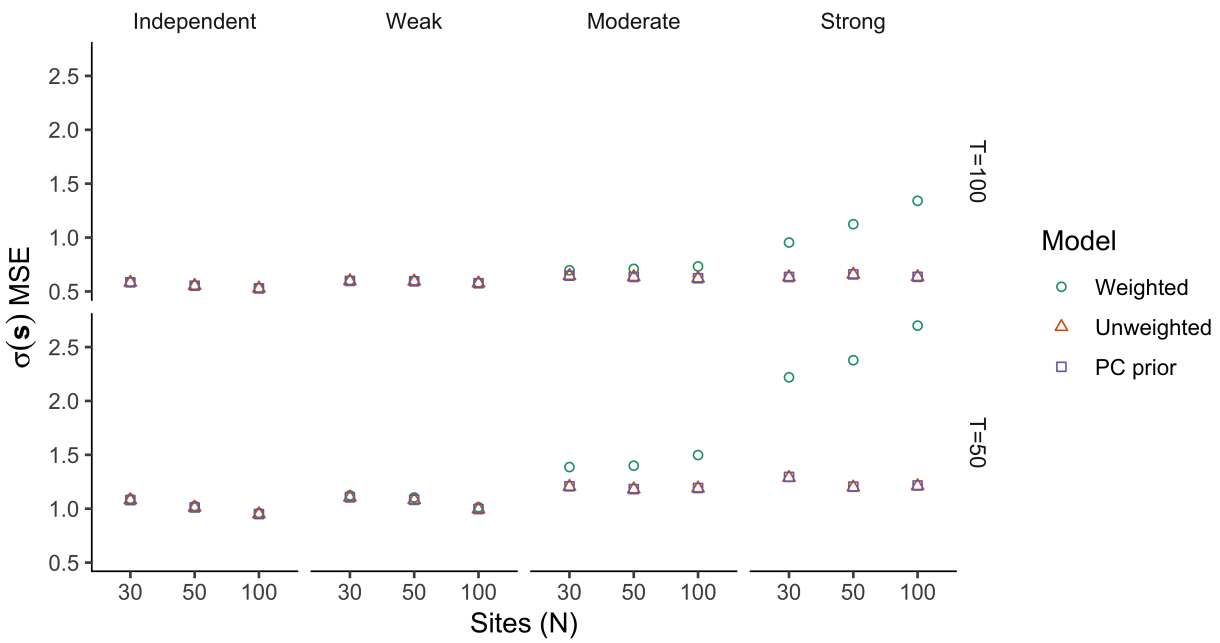

Figure 2.9: Empirical mean square error (MSE) of posterior estimates for GEV scale parameters $\sigma(\boldsymbol{s})$ across comparison models and all simulations.

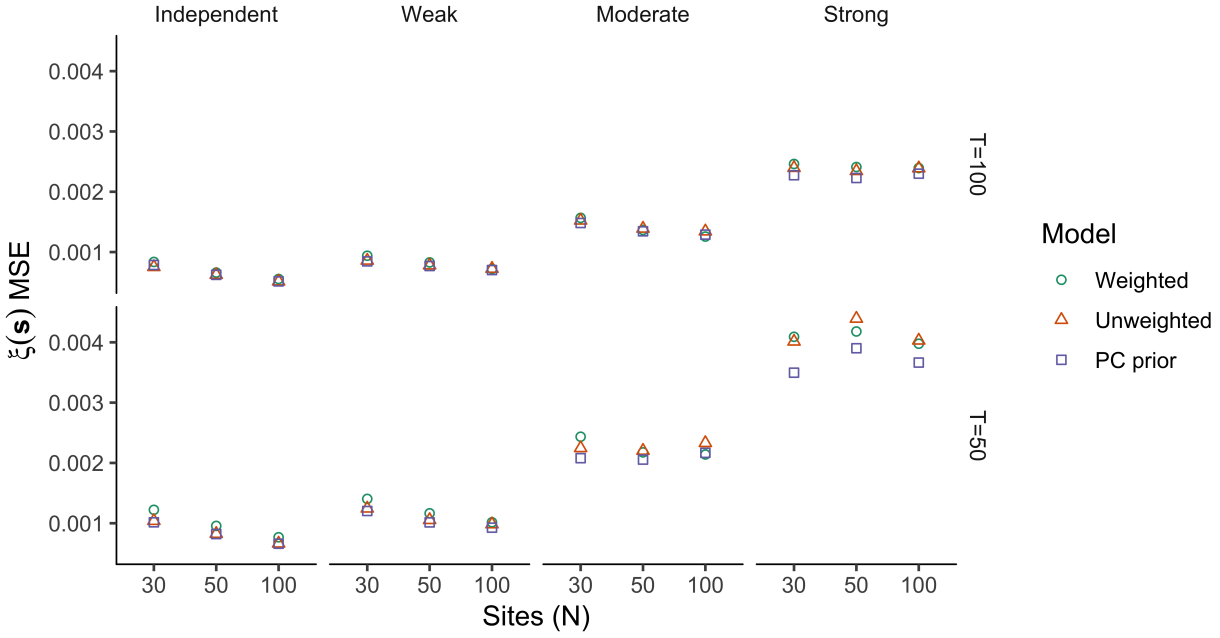

Figure 2.10: Empirical mean square error (MSE) of posterior estimates for GEV shape parameters $\xi(\boldsymbol{s})$ across comparison models and all simulations.

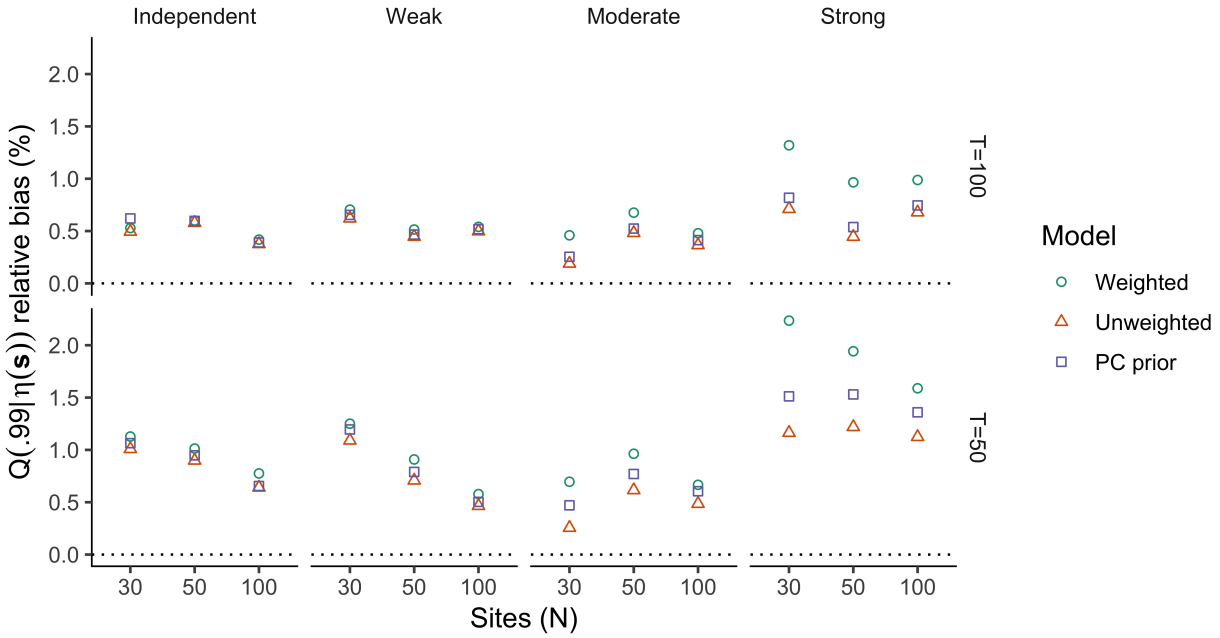

Figure 2.11: Empirical relative bias of posterior estimates for 100-year return levels $Q\big(.99|\boldsymbol{\eta}(\boldsymbol{s})\big)$ across comparison models and all simulations.

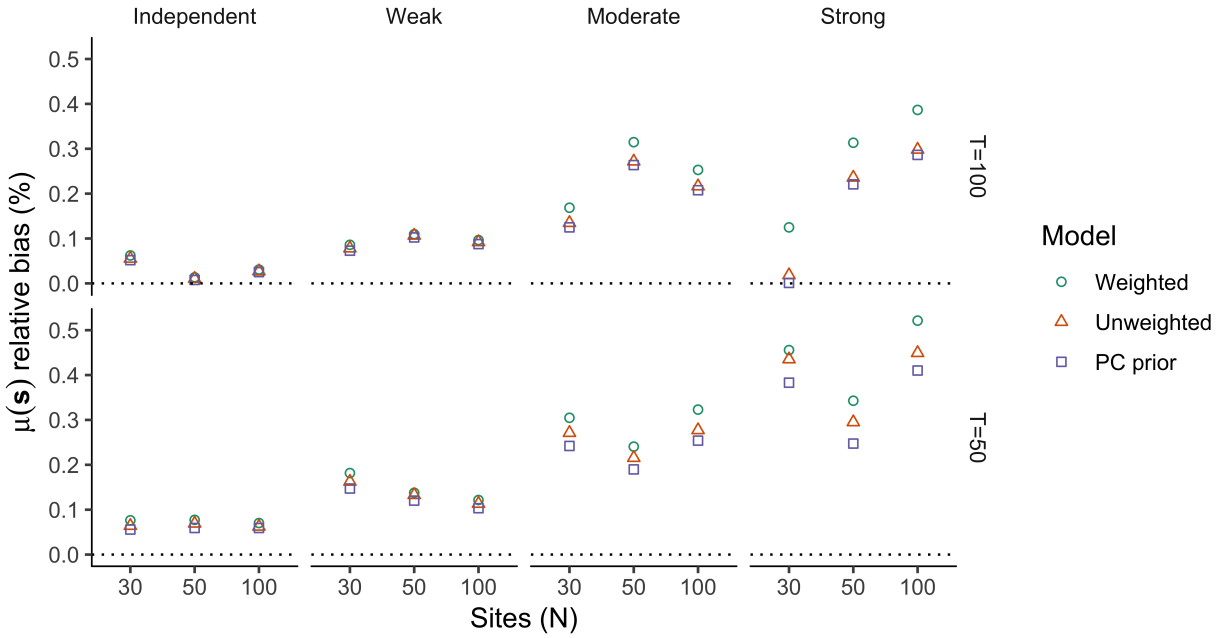

Figure 2.12: Empirical relative bias of posterior estimates for GEV location parameters $\mu(\boldsymbol{s})$ across comparison models and all simulations.

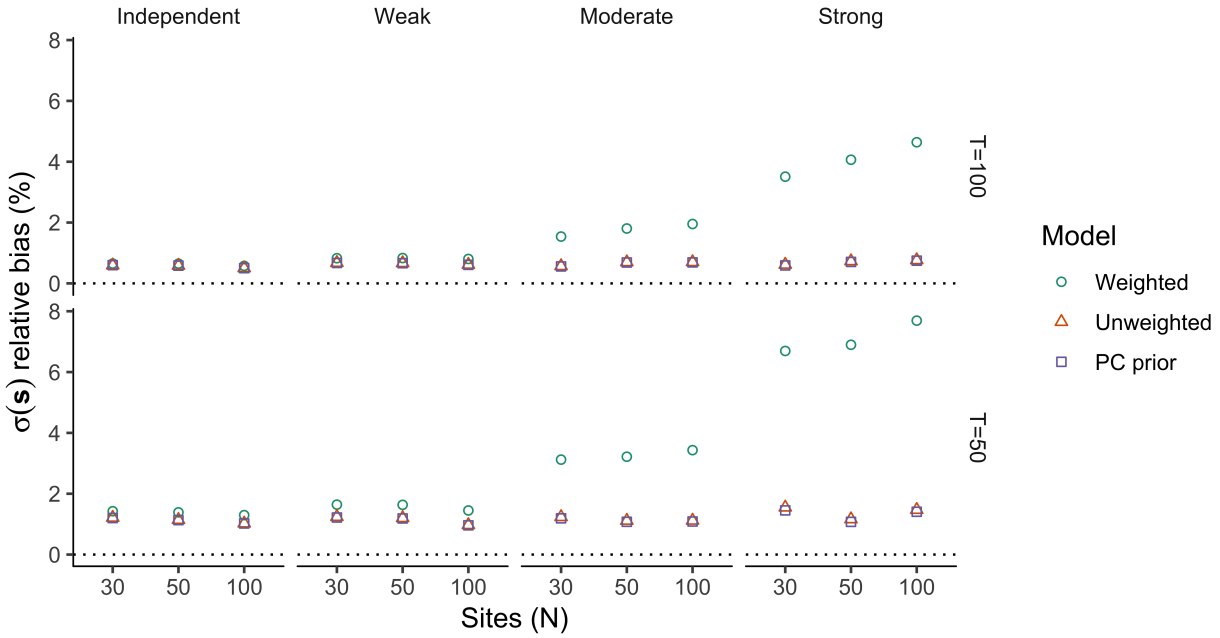

Figure 2.13: Empirical relative bias of posterior estimates for GEV scale parameters $\sigma(\boldsymbol{s})$ across comparison models and all simulations.

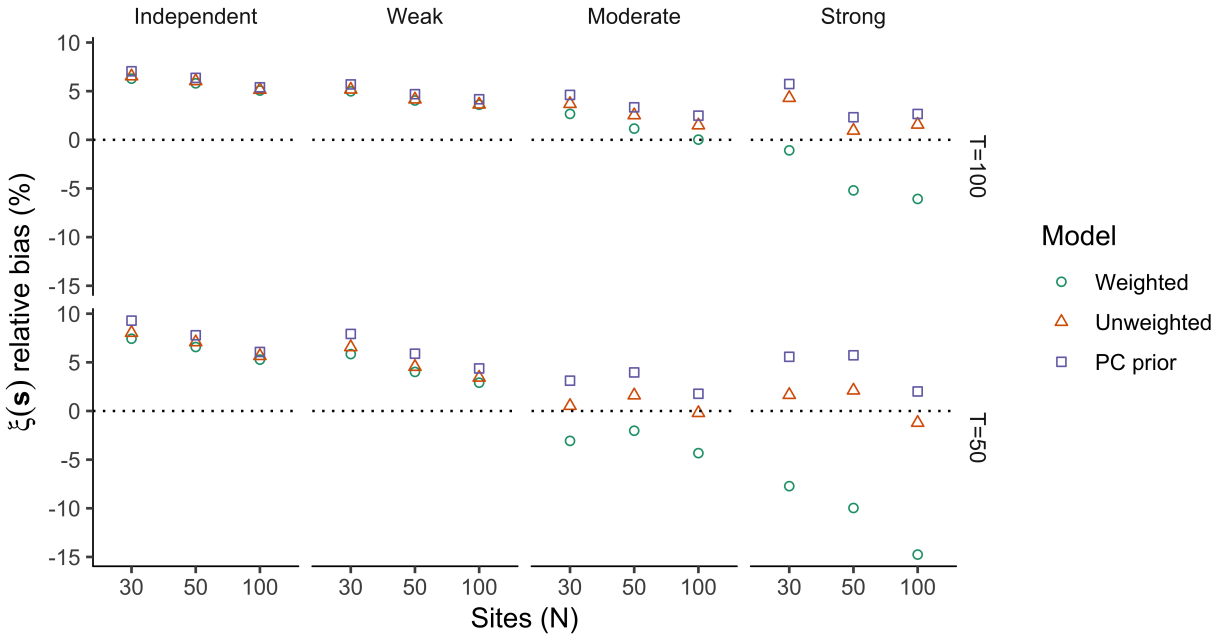

Figure 2.14: Empirical relative bias of posterior estimates for GEV shape parameters $\xi(\boldsymbol{s})$ across comparison models and all simulations.

(2015) fully describe their selection criteria, which, for example, include requirements that stations have been operational for at least 30 years. Additionally, annual maxima of daily precipitation are only analyzed from years with few missing daily records of precipitation. Between 18 and 120 annual maxima are analyzed for each station, with roughly equal representation of all temporal sample sizes.

Exploratory analysis suggests the Front Range GHCN data have between weak and moderate extremal dependence. The estimated extremal coefficient function $\hat{\theta}(d) : (0,\infty) \to [1,2]$ is near-constant between 1.8 and 1.9 for all distances d, which implies the likelihood weights will also have a small range. Schlather and Tawn (2003) also observe a near-constant extremal coefficient function for extreme precipitation in south-west England. The authors remark that the result may have a physical basis because the study region is small relative to the scale of the meteorological systems that generate precipitation, which implies it is likely that no two sites in the region are truly independent. Likelihood weights (2.5) for the GHCN data are similar to weights for simulated data with moderate extremal dependence (Figure 2.21). Since the average number of annual maxima per station ($T = 60$) is also close to our $T = 50$ simulation, we anticipate the weighted likelihood will have closer to nominal coverage and the unweighted likelihood will slightly undercover (Figure 2.1).

2.6.2 Model

As in the simulation, we use the weighted marginal likelihood (2.4) in a hierarchical Bayesian framework in which the GEV parameters $\boldsymbol{\eta}(\boldsymbol{s})$ are estimated as independent latent Gaussian processes. Since the simulation shows that estimators based on fixed and Gibbs-updated weights have similar properties, we use fixed weights during estimation. We use annual mean precipitation from the PRISM precipitation dataset (Daly et al., 2008) as a covariate for each of the GEV parameters, and model the spatial correlation between parameters with the Matérn covariance function. For example, the Matérn specifies the correlation between parameters $\xi(\boldsymbol{s})$ and $\xi(\boldsymbol{t})$ at two locations $\boldsymbol{s}, \boldsymbol{t} \in \mathscr{D}$ via

$$\kappa\left(\boldsymbol{s},\boldsymbol{t};\tau,\rho,\nu\right)=\frac{1}{\tau 2^{\nu-1}\Gamma(\nu)}K_{\nu}\left(\|\boldsymbol{s}-\boldsymbol{t}\|/\rho\right)$$

where K_ν is the modified Bessel function of the second kind with order ν. The Matérn covariance is parameterized through its inverse scale $\tau > 0$, range $\rho > 0$, and smoothness $\nu > 0$ parameters. Annual average precipitation from the PRISM dataset accounts for average weather patterns and orographic effects on precipitation, such as elevation. In general, prior distributions are weakly informative, and prior distributions for spatial covariance parameters are centered around variogram-based estimates of spatial correlation between exploratory estimates of marginal parameters $\boldsymbol{\eta}(\boldsymbol{s})$.

Prior distributions for regression coefficients $\boldsymbol{\beta}$ and spatial covariance parameters σ_0 and λ_0 used in the application to extreme Colorado precipitation (Section 2.6) are specified in Table 2.3. Regression coefficient prior distributions are designed to be uninformative, while the spatial covariance prior distributions are designed to be weakly informative. The spatial covariance prior large variance. The distributions are parameterized such that the prior mode covers least-square variogram estimates for the spatial covariance parameters. Variograms are based on smoothed maximum likelihood fits of the GEV parameters. The proposal standard deviations for the random walk Metropolis-Hastings samplers are $s_\mu = 1.45$, $s_{\log\sigma} = .25$, $s_\xi = .11$; s_{λ_0} is .3, .4. and .6, respectively for the GEV location $\left\{\mu(\boldsymbol{s})\right\}_{\boldsymbol{s}\in\mathscr{D}}$, scale $\left\{\log\sigma(\boldsymbol{s})\right\}_{\boldsymbol{s}\in\mathscr{D}}$, and shape $\{\xi(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ processes; similarly, s_{ν_0} is .12, .1, and .15.

2.6.3 Posterior inference and diagnostics

Inference uses a sample from the posterior distribution, drawn with a Gibbs sampler that was run for 3,002,000 iterations. The first 2,000 samples were discarded. The sampler was run for a large number of iterations because it was slowly mixing. Posterior inference uses 10,000 of the remaining samples; only every 300^{th} sample was saved due to storage constraints. To facilitate model comparison, we also fit the unweighted latent spatial extremes model using the same priors and inference strategy. Diagnostics suggest no significant concerns with conver-

Table 2.4: Prior distributions used in application to extreme Colorado precipitation (Section 2.6).

	GEV parameter process		
	$\{\mu(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$	$\{\log\sigma(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$	$\{\xi(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$
(Regression coefs.)			
$\boldsymbol{\beta}\sim$	$\mathcal{N}(0,100)$,	$\mathcal{N}(0,100)$,	$\mathcal{N}(0,100)$.
(Spatial covariance)			
$\sigma_0\sim$	IG$(2,60)$,	IG$(2,10)$,	IG$(2,.03)$.
$\lambda_0\sim$	Gamma$(2,.5)$,	Gamma$(2,.25)$,	Gamma$(2,.1)$.
$\nu_0\sim$	Gamma$(2,1)$,	Gamma$(2,1)$,	Gamma$(2,1)$.

gence and also that the chain has been run for long enough to control Monte Carlo integration error. Due to the relatively small number of spatial locations in the dataset ($N = 71$), posterior diagnostics indicate the spatial covariance parameters are at least weakly identified by the data. Posterior learning is diagnosed by comparing prior and posterior distributions for the spatial mean and covariance parameters.

Posterior diagnostics do not suggest the Gibbs sampler has not converged. Similarly, posterior diagnostics suggest the sampler has been run for a long enough period of time and is able to identify model parameters from the data. In particular, there is no strong evidence that posterior inference is sensitive to the sampler's initial state. We use potential scale reduction factors (PSRFs) to assess posterior convergence by comparing inference from nine independent copies of our Gibbs sampler. Potential scale reduction factors estimate the potential reduction in uncertainty of posterior means if the Gibbs samplers were allowed to run for an infinitely longer amount of time (Gelman and Rubin, 1992). Each copy of the sampler was randomly initialized by drawing model parameters from the model's prior distribution. The maximum upper confidence limit of PSRFs for return levels at the observation locations is 1.04, suggesting that the posterior uncertainty in return levels is inflated by up to 4% due to Gibbs sampling. A multivariate extension of the PSRFs estimates the largest potential scale reduction factor among all linear combinations of a collection of posterior means (Brooks and Gelman, 1998). For return

levels, this quantity is also 1.04. The multivariate PSRFs for GEV location and scale parameters are each 1.01, and the multivariate PSRF for GEV shape parameters is 1.05.

Posterior traceplots, autocorrelation plots, and effective sample sizes indicate the Gibbs sampler is slowly mixing, but additional diagnostics suggest the sampler has been run for enough samples so as to control Monte Carlo integration error. Estimates of the Monte Carlo integration error are small relative to the magnitude of posterior means of interest, such as posterior means for marginal return levels, latent GEV parameters, and the spatial mean and covariance functions of the spatial processes that model the GEV parameters. In particular, Monte Carlo integration errors are at most .6% of the magnitude of posterior return levels, scale parameters, and location parameters. Monte Carlo integration errors are at most 3.4% the magnitude of posterior shape parameters. The relative sizes of Monte Carlo errors are more variable for the mean and covariance function parameters of the latent GEV parameter fields, but are between .1% and 2.5% for all mean and covariance function parameters except for three parameters. Monte Carlo integration errors are respectively 4.5% and 6.1% of the magnitude of the posterior mean for the smoothness parameters ν_0 of the GEV scale $\left\{\sigma^2(\boldsymbol{s})\right\}_{\boldsymbol{s}\in\mathscr{D}}$ and shape $\{\xi(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ parameter processes. Lastly, Monte Carlo integration error is 37% of the magnitude of the posterior mean for the effect of mean annual precipitation β_1 on shape parameters $\{\xi(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$. However, this ratio is artificially inflated because the parameter is estimated to be small or vanishing ($\hat{\beta}_1 = .001$; 95% highest posterior density interval is $(-.04, .04)$).

Posterior diagnostics also suggest the data at least weakly identify the mean and covariance parameters for the latent GEV parameter fields. The posterior densities for the mean functions of the Gaussian processes used to model the GEV parameters all differ from the prior distributions (Figure 2.15, Figure 2.17, Figure 2.19). Similarly, posterior densities for the parameters of the spatial covariance functions for the latent GEV parameters differ from the prior densities (Figure 2.16, Figure 2.18, Figure 2.20). However, the posterior distributions do not differ dramatically from the priors for the smoothness and range parameters of the covariance function $\rho_{\xi(\boldsymbol{s})}$ of the GEV shape parameters.

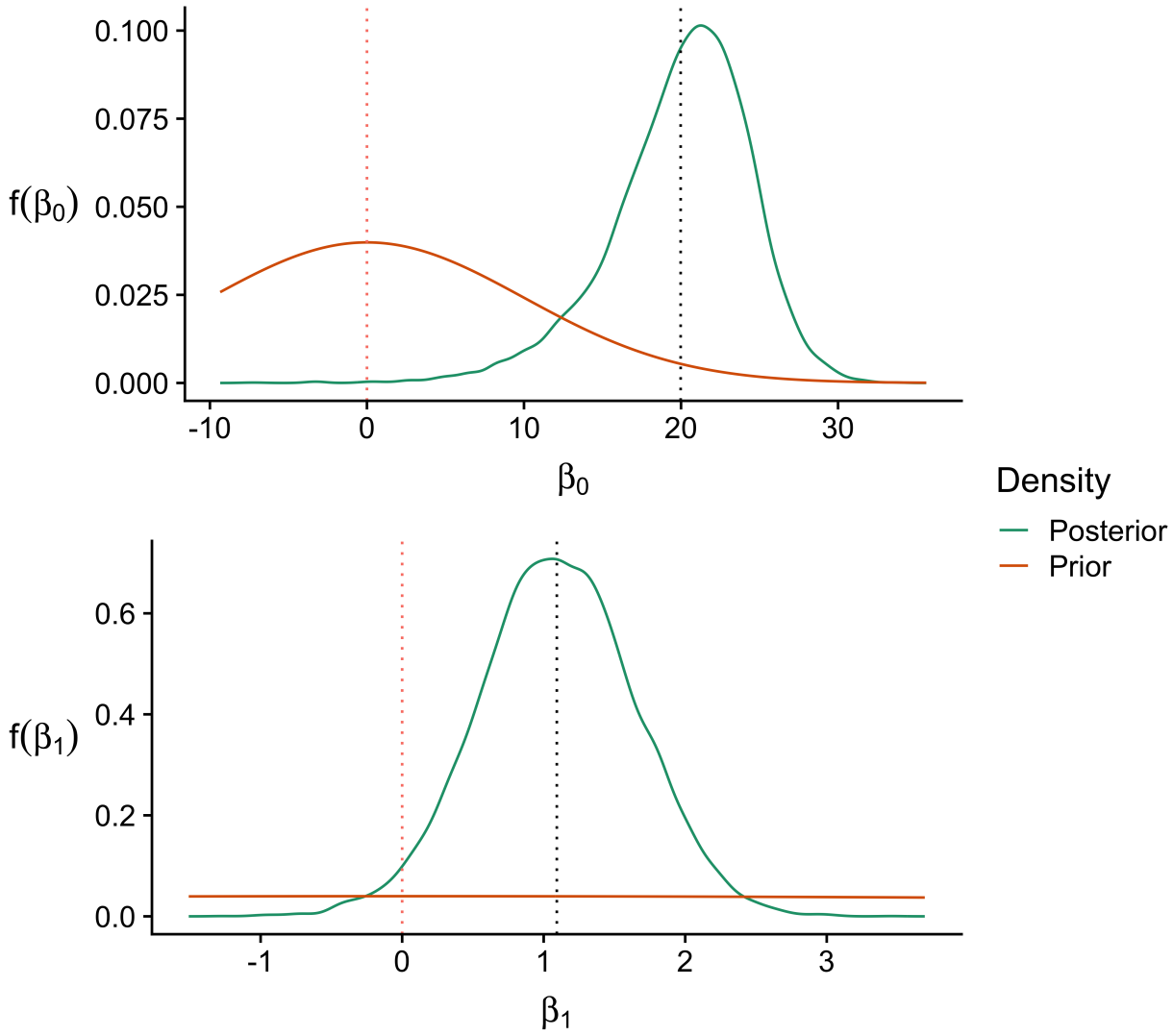

Figure 2.15: Comparison of prior and posterior distributions for the mean function of the latent Gaussian process that models GEV location parameters $\{\mu(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$. Prior and posterior means are marked by vertical dotted lines. The plots show strong posterior learning in both the intercept β_0 and slope parameters β_1, which model a linear trend between annual average precipitation and GEV location parameters.

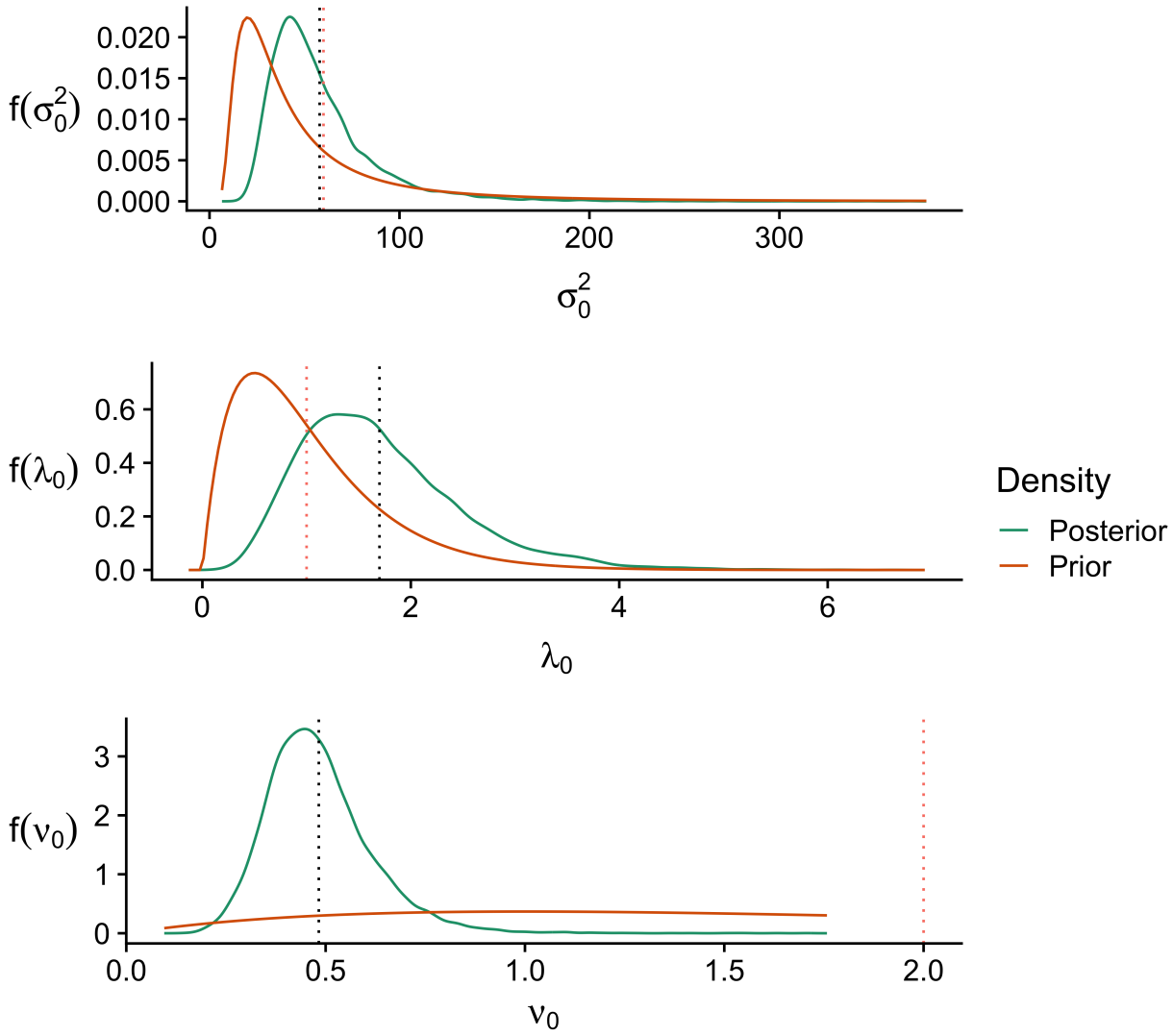

Figure 2.16: Comparison of prior and posterior distributions for the covariance parameters of the latent Gaussian process that models GEV location parameters $\{\mu(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$. Prior and posterior means are marked by vertical dotted lines. The plots show strong posterior learning in the covariance range λ_0 and smoothness ν_0. There is weaker posterior learning in the covariance sill σ_0^2 parameter.

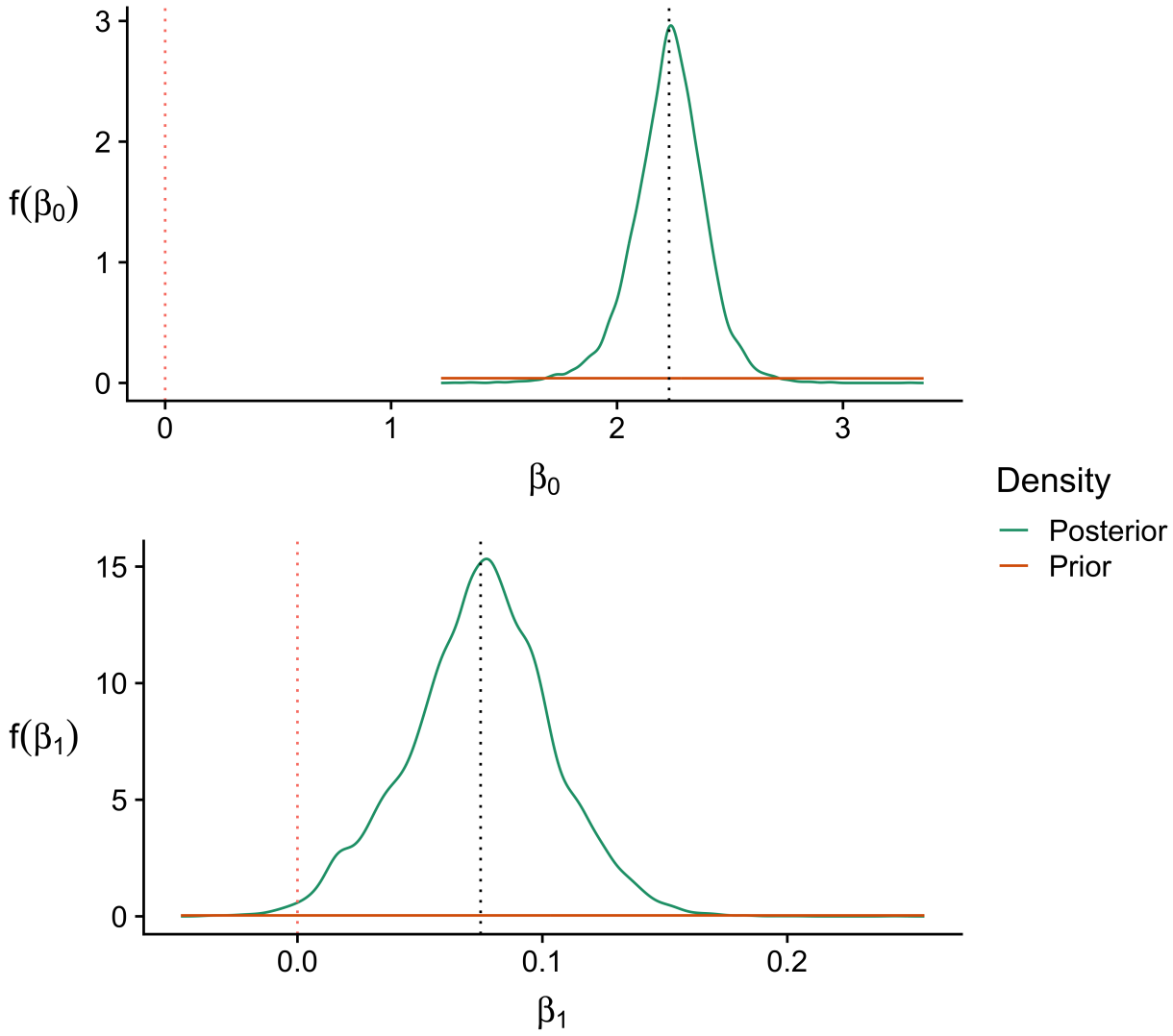

Figure 2.17: Comparison of prior and posterior distributions for the mean function of the latent Gaussian process that models GEV scale parameters $\{\sigma^2(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$. Prior and posterior means are marked by vertical dotted lines. The plots show strong posterior learning in both the intercept β_0 and slope parameters β_1, which model a linear trend between annual average precipitation and GEV scale parameters.

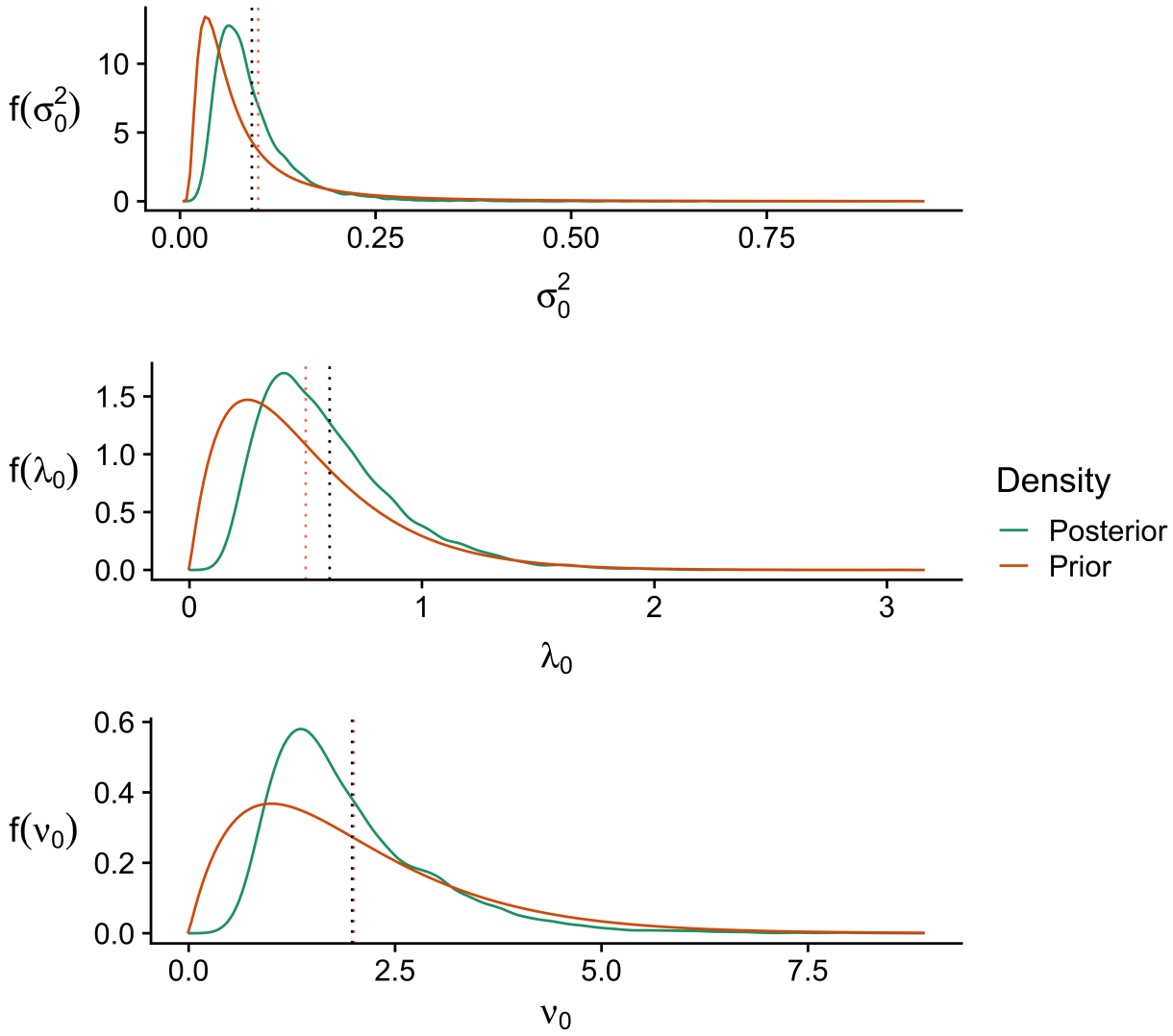

Figure 2.18: Comparison of prior and posterior distributions for the covariance parameters of the latent Gaussian process that models GEV scale parameters $\{\sigma^2(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathcal{D}}$. Prior and posterior means are marked by vertical dotted lines. The plots show moderate posterior learning in the covariange smoothness ν_0 parameter, but weak posterior learning in the covariance sill σ_0^2 and range λ_0 parameters.

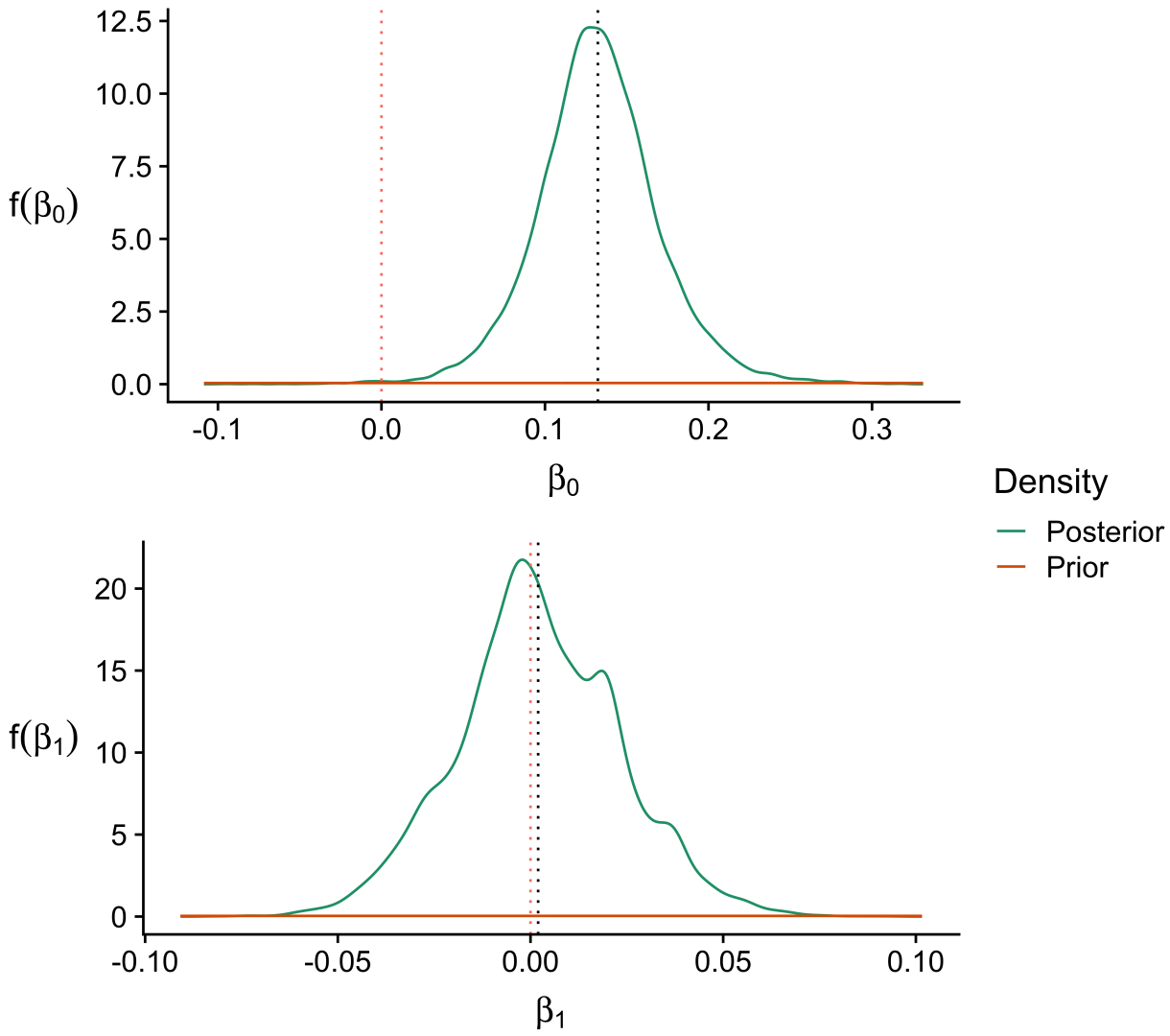

Figure 2.19: Comparison of prior and posterior distributions for the mean function of the latent Gaussian process that models GEV shape parameters $\{\xi(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$. Prior and posterior means are marked by vertical dotted lines. The plots show strong posterior learning in both the intercept β_0 and slope parameters β_1, which model a deterministic trend between annual average precipitation and GEV shape parameters.

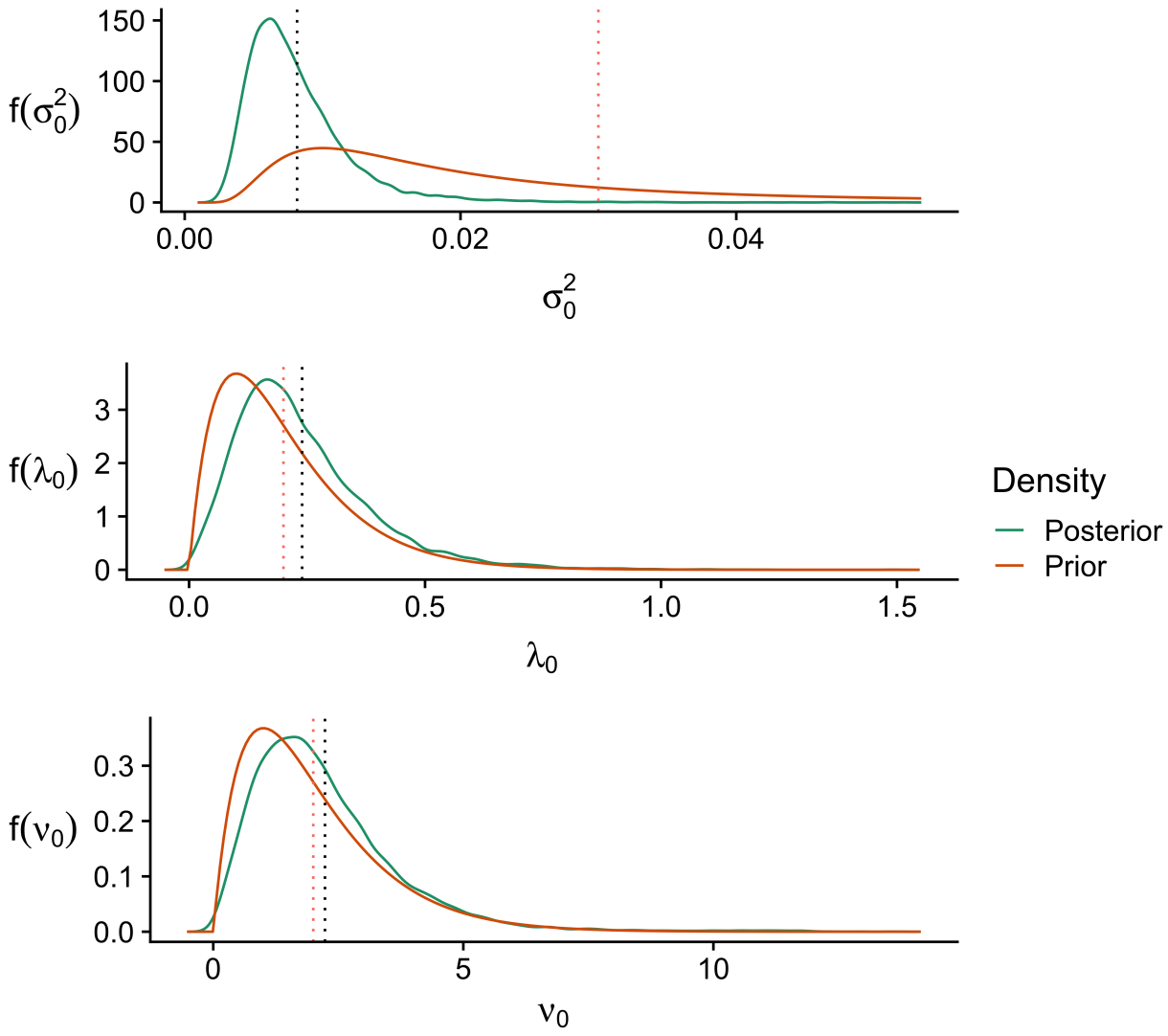

Figure 2.20: Comparison of prior and posterior distributions for the covariance parameters of the latent Gaussian process that models GEV shape parameters $\{\xi(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$. Prior and posterior means are marked by vertical dotted lines. The plots show strong posterior learning in the covariance sill σ_0^2, but almost no posterior learning in the covariance range λ_0 or smoothness λ_0 parameters.

2.6.4 Results

The likelihood weights (2.5) have a spatial pattern and their effect can be interpreted by their impact on the weighted Fisher information (2.6) (Figure 2.22). As expected, stations near the edges of the sampled region tend to have the highest weights because annual maxima observed at these locations are at most weakly dependent with observations at other stations. Annual maxima at distant stations tend to be at most weakly dependent because they tend to experience different large rain events than other stations.

Weighted estimates borrow more strength across locations, which impacts return level estimates. The latent Gaussian processes increase smoothing as more strength is borrowed, shrinking parameter estimates (Figure 2.24). Shrinkage manifests as additional smoothing in maps of return levels (Figure 2.23). In particular, the weighted estimates better match physical features that impact Colorado precipitation. The contours in the weighted return level map have stronger north-south patterns, especially along 105° W—the boundary of the Rocky mountains in the Colorado Front Range region (Figure 2.23 B). The size of the region with elliptical 150–175mm return level contours (■) of extreme precipitation near Boulder, Fort Collins, and Colorado Springs also increase. The larger elliptical regions produced by the weighted model better capture physical effects of the Palmer Divide and the Cheyenne Ridge on Colorado precipitation (Daly et al., 2008; Karr and Wooten, 1976).

We verify that the weighted model's changes are beneficial near the Palmer Divide and Cheyenne Ridge regions by refitting the weighted and unweighted models with a holdout set to test out-of-sample fit. Our holdout set uses data from seven stations (10% of the dataset) near the Palmer Divide and Cheyenne Ridge, and where posterior estimates of return levels differ between the two models (stations marked by diamonds in Figure 2.22). Testing uses the log-score $\ell(\boldsymbol{s}_0)$ at each holdout location $\boldsymbol{s}_0$. Log-scores form strictly proper scoring rules that compare the log-likelihood from both models on data at each holdout location (Gneiting and Raftery, 2007). In our spatial application, we use the posterior kriging distribution to draw a posterior sample of GEV parameters at each test location $\boldsymbol{s}_0$, which we then use to compute the posterior mean

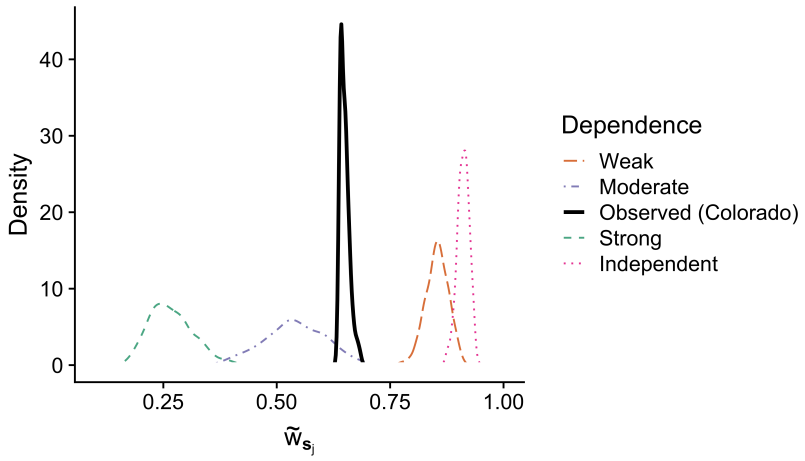

Figure 2.21: Distribution of likelihood weights (2.5) for Colorado data and simulations with $N = T = 50$. The Colorado weights suggest the data have moderate extremal dependence.

log-likelihood at each test location $\ell(\boldsymbol{s}_0)$. Resulting log-scores show that the weighted model improves out-of-sample fit in six out of seven of the holdout locations (Table 2.5). The log-scores also show that neither model fits the data well at Pueblo, CO, the southernmost holdout station. In particular, the data at Pueblo, CO tend to be relatively less extreme. Separate exploratory analysis of Pueblo's data suggests extreme precipitation is associated with a negative shape parameter $\xi(\boldsymbol{s}_0) < 0$. However, the spatial models suggest a positive shape parameter is more appropriate.

The weighted likelihood also model induces shrinkage of the GEV parameters $\boldsymbol{\eta}(\boldsymbol{s})$ and return levels $Q\big(p\big|\boldsymbol{\eta}(\boldsymbol{s})\big)$ (Figure 2.24). In hierarchical models, estimates balance data with smoothness constraints imposed by hierarchical layers. Shrinkage occurs in the weighted model because the weighted model shifts the balance more toward the hierarchical layers.

2.7 Discussion

Estimating marginal return levels is an important step in planning for impacts of natural hazards, especially those caused by precipitation. Extreme precipitation data have dependence, which makes estimation more complicated. Models that explicitly account for dependence in the data have limited ability to scale to large datasets, while models that assume conditional

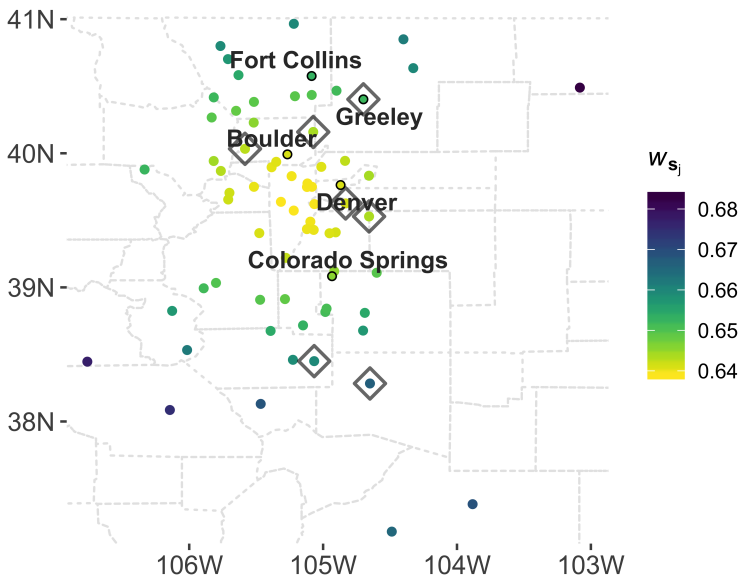

Figure 2.22: Spatial distribution of weights. Weights are smaller for locations central to the spatial sampling pattern, where extremal dependence is more likely to impact data. Cities used in the hold-out model comparison are marked by diamond outlines.

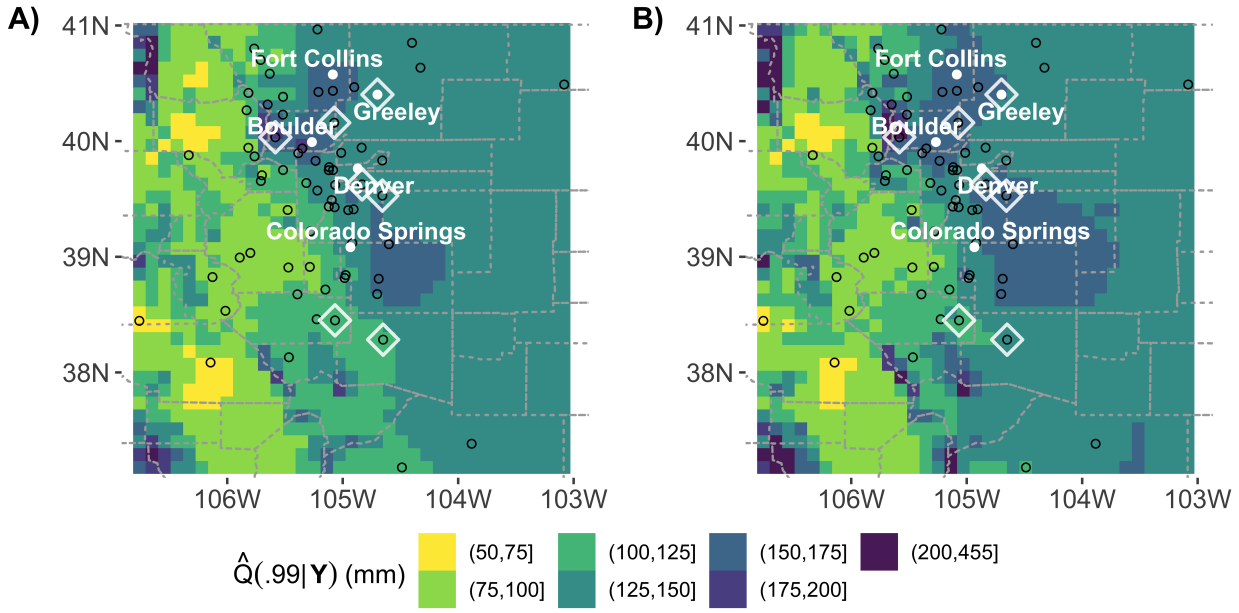

Figure 2.23: Spatially complete estimates $\hat{Q}\big(.99|\boldsymbol{\eta}(\boldsymbol{s})\big)$ of 100-year return levels for daily precipitation in Colorado's Front Range. Estimates are compared from the unweighted (A) and weighted (B) unweighted latent spatial extremes models. The weighted estimates have increased smoothness and spatial range, and overall patterns that better match orographic features in Colorado. The locations of the 71 stations whose data are analyzed are indicated by (∘). For reference, we include the names of several reference cities. Cities used in the hold-out model comparison are marked by diamond outlines.

Table 2.5: Comparison of log-scores for the weighted $\ell_{wtd}(\boldsymbol{s}_0)$ and unweighted models $\ell(\boldsymbol{s}_0)$ at holdout cities; the highest log-score is highlighted for each city. The weighted likelihood model tends to have higher log-scores at holdout cities, suggesting better out-of-sample predictive performance in the targeted regions. The low log-scores in the bottom row also suggest neither model is predictive of extreme precipitation in Pueblo.

Lat.	Lon.	City	$\ell_{wtd}(\boldsymbol{s}_0)$	$\ell(\boldsymbol{s}_0)$
40.4	104.7	Greeley	**−224**	−225
40.2	105.1	Longmont	**−502**	−508
40.0	105.6	Nederland	**−411**	−415
39.6	104.8	Aurora	**−303**	−307
39.5	104.7	Parker	**−262**	−701
38.5	105.1	Penrose	−198	**−196**
38.3	104.7	Pueblo	**−349,848**	−1,779,826

independence in the data can scale well to large datasets, but do not account for dependence. We develop a weighted likelihood that downweights observations from locations central to the spatial sampling pattern in order to better estimate marginal return levels. We use the extremal coefficient in (2.1) to construct weights that downweight likelihood contributions from locations central to the spatial sampling pattern, where observations tend to be most dependent. Simulations confirm that the weighting scheme improved the uncertainty quantification of the return level estimates in situations when data have extremal dependence. In application, estimates from the weighted model better align with expected changes in patterns of extreme precipitation caused by physical features, like mountains.

Since weighted likelihoods are computationally inexpensive, they may be a useful technique to adopt in most settings where latent spatial extremes models are employed. Weighting adds N additional multiplications per likelihood evaluation, whereas alternatives like penalization add N additional function evaluations. Penalization improves estimation for univariate extremes data at a similar computational cost, but its main purpose is to discourage models from exploring unrealistic or undesirable regions of the parameter space, such as those with large shape parameters $\xi(\boldsymbol{s})$. As a result, penalized models underestimate uncertainty almost as much as unweighted models. Composite likelihood corrections are more computationally expensive

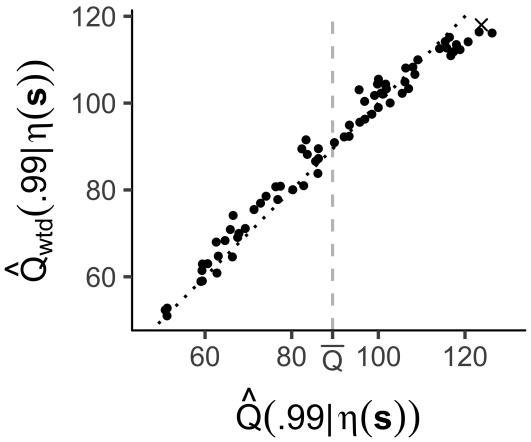

Figure 2.24: Comparison of weighted $\hat{Q}_{wtd}\big(.99|\boldsymbol{\eta}(\boldsymbol{s})\big)$ and unweighted $\hat{Q}\big(.99|\boldsymbol{\eta}(\boldsymbol{s})\big)$ return level estimates plotted against a dotted 1:1 reference line. The weighted model shrinks estimates toward a common return level. Shrinkage occurs as unweighted return level estimates below the unweighted average $\bar{Q}$ tend to increase in the weighted model, while unweighted return level estimates above $\bar{Q}$ tend to decrease.

(Ribatet et al., 2012; Sharkey and Winter, 2018). In practice, weighting encourages borrowing strength across locations to improve estimates at each location.

Refining the likelihood weights (2.5) could further improve the ability for marginal likelihoods to account for extremal dependence when estimating marginal return levels. For example, pairwise densities can be derived for specific max-stable processes (e.g., Padoan et al., 2010). Pairwise densities explicitly model the dependence between pairs of observations, while the extremal coefficient we use to build likelihood weights measures a summary of extremal dependence instead. Empirical Bayes–like procedures could be developed that use likelihood weights based on pairwise densities to further improve the performance of return level estimators. While empirical Bayes procedures will not fully account for estimation uncertainty (e.g., in estimating dependence parameters in bivariate densities), the procedures may still provide a fair compromise between computational complexity and accurate estimation of uncertainty.

Weighting schemes are flexible, so may be extended to accommodate complex issues in modeling and estimation outside extremes applications. While we demonstrate the use of a weighted likelihood for latent spatial extremes models, the theory we develop is more general. The Fisher information interpretation of weighted likelihoods also applies to all weighted like-

lihoods. Similarly, the limiting behaviors of likelihoods for independent data or completely dependent data are largely based on copula theory for arbitrary data, rather than extreme value theory. Importantly, the construction of the weighted likelihood (2.4) can be adapted to other statistical problems where marginal inference is of interest but likelihoods are difficult to evaluate. The construction we propose is based on the idea that a computationally inexpensive measure of dependence between observations can be used to develop a weighted likelihood that better quantifies parameter uncertainty than related unweighted models. The main challenge in adapting our weighted likelihood to other applications is in identifying an appropriate dependence measure that can be used to build likelihood weights.

Chapter 3

Remote effects spatial process models for modeling teleconnections[2]

3.1 Introduction

While most spatial data can be modeled with the assumption that distant points are uncorrelated, some problems require dependence at both far and short distances. Spatial climate data is an example of the latter, as it is influenced by local (i.e., short distance) factors, as well as by remote (i.e., far distance) phenomena called teleconnections. Teleconnections refer to changes in patterns of large-scale atmospheric circulation that can drive changes in temperature and precipitation in distant regions (e.g., Tsonis and Swanson, 2008; Ward et al., 2014). Most teleconnection modeling approaches in the statistical literature do not explicitly estimate dependence within remote phenomena. The statistical literature includes spatially varying coefficient models, analogs, and covariance matrix estimation (Calder et al., 2008; Choi et al., 2015; McDermott and Wikle, 2016; Wikle and Anderson, 2003). Explicitly modeling dependence in remote phenomena can add physically sensible structure that improves prediction accuracy and addresses some modeling challenges. We propose a geostatistical model that addresses this unmet modeling need for teleconnection.

Teleconnections can be forced by changes in sea surface temperature (SST), and there have been many observational and modeling studies studying the link between SSTs, circulation patterns, and impacts on global and regional climate. Several seminal studies connect U.S. precipitation with SST anomalies in the tropical Pacific due to the El Niño–Southern Oscillation teleconnection (ENSO) (Montroy, 1997; Montroy et al., 1998), as well as with SST anomalies in the Pacific (e.g., Dong and Dai, 2015). The ENSO teleconnection has been critical in seasonal

[2]Hewitt, J., Hoeting, J. A., Done, J. M., & Towler, E. (2018). Remote effects spatial process models for modeling teleconnections. *Environmetrics*, 29(8). https://doi.org/10.1002/env.2523.

climate forecasting (Goddard et al., 2001), and decadal variability of sea surface temperature anomalies have been identified as a source of potential skill for decadal predictions that look out one year to a decade (Meehl et al., 2009). In terms of the latter, decadal predictions produced from global climate models (GCMs) have shown skill in reproducing ocean and land temperatures, and less skill in precipitation (Meehl et al., 2014). This is the general finding for GCMs: while GCMs perform poorly in predicting precipitation directly, they can skillfully reproduce surface temperatures and large-scale patterns (Flato et al., 2013). Direct precipitation prediction by GCMs is challenging because of complex and interacting multi-scale physical precipitation processes, resulting in large uncertainty in future precipitation patterns (Deser et al., 2012). As such, this provides a motivating example for demonstrating a teleconnection model that can be used in conjunction with GCM output to estimate impacts on precipitation.

Developing a teleconnection model for application with GCM output has overlaps with the burgeoning field of statistical downscaling. Statistical downscaling methods use large-scale variables to draw inference on regional variables. Similar to what is being proposed here, a type of statistical downscaling called perfect prognosis downscaling (Maraun et al., 2010) develops a statistical relationship between observed large-scale predictors and local-scale weather phenomena (e.g., Bruyere et al., 2012; Towler et al., 2016; Wilby et al., 1998). Common models used for perfect prognosis downscaling do not explicitly model spatial dependence. Maraun et al. (2010) review methods used in the climate literature, which include linear models, analogs, and machine learning techniques like neural networks. Dependence is often indirectly modeled by using principle component or canonical correlation basis functions as predictors and applying various corrections to uncertainties (cf. Karl et al., 1990). After statistical relationships are developed and validated on observed datasets, models can be applied to large-scale GCM output to obtain an estimate of the desired predictant. Clearly, perfect prognosis methods are highly dependent on the selected predictors and model (Fowler et al., 2007).

We propose a remote effects spatial process (RESP) model that extends spatially varying coefficient models to directly model dependence in remote phenomena and address several mod-

eling challenges. Spatially modeling dependence in remote phenomena adds sensible structure to teleconnection models which, in turn, allows better use of the data than standard models. Standard spatially varying coefficient models regress a local response $Y(\boldsymbol{s}, t)$ with spatio-temporal error $w(\boldsymbol{s}, t)$ onto local covariates $\boldsymbol{x}(\boldsymbol{s}, t)$ through

$$Y(\boldsymbol{s}, t) = \boldsymbol{x}(\boldsymbol{s}, t)^T \boldsymbol{\beta} + \boldsymbol{z}(t)^T \boldsymbol{\theta}(\boldsymbol{s}) + w(\boldsymbol{s}, t) \tag{3.1}$$

which includes adjustment for spatially-varying effects $\boldsymbol{\theta}(\boldsymbol{s}) \in \mathbb{R}^k$ associated with a second vector $\boldsymbol{z}(t) \in \mathbb{R}^k$ of k covariates (Banerjee et al., 2015, Section 9.6.2). As applied to teleconnection, the covariate vector $\boldsymbol{z}(t)$ contains one or more indices that quantify the overall strength or state of large-scale patterns, like ENSO or the North Atlantic Oscillation (Calder et al., 2008; Wikle and Anderson, 2003). While effective, the model (3.1) assumes relevant large-scale patterns are known a priori (e.g., ENSO). However, relevant teleconnection indices can depend on the study region and thus be unknown at the start of an analysis (Towler et al., 2016). The spatially varying coefficient model (3.1) will be inefficient if driven by poorly chosen teleconnection indices. Standard formulations of (3.1) also model within-site covariances for spatially varying effects $\Lambda = \text{Cov}(\boldsymbol{\theta}(\boldsymbol{s})) \in \mathbb{R}^{k \times k}$ with non-spatial covariance matrices. While the issue may be less important for orthogonal teleconnection indices, typical indices are defined with respect to different covariates and zonal averages so may not be orthogonal (cf. Ashok et al., 2007; Mantua et al., 1997). Instead, teleconnection indices may have spatial structure induced by remote covariates. The RESP model introduced below directly incorporates remote covariates instead of using teleconnection indices and can offer potential improvement for the a priori and spatial structure concerns (Section 3.2.1). Notably, the RESP model does not lose generality since direct connections can be drawn to standard spatially varying coefficient models (Section 3.2.3).

More generally, the RESP model represents a less-common class of spatial analysis problems that provide rich opportunities for study. We introduce our teleconnection model in the general context of a spatial regression problem involving local and spatially remote covariates (Section 3.2.1). The local and spatially remote covariates are allowed to have different spatial cor-

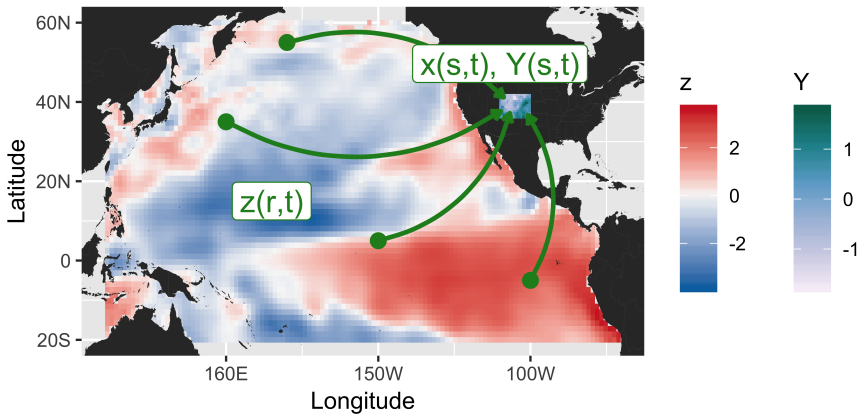

Figure 3.1: Schematic illustration of a teleconnection problem. Colorado precipitation $Y(\boldsymbol{s},t)$ is influenced by both local covariates $\boldsymbol{x}(\boldsymbol{s},t)$ and remote covariates $z(\boldsymbol{r},t)$. The remote covariates shown here are standardized anomalies of average monthly Pacific Ocean sea surface temperatures during Winter, 1982. The data come from the ERA-Interim reanalysis dataset (Dee et al., 2011).

relations structures reflecting their different relationships with the response. fig. 3.1 schematically illustrates the general teleconnection problem in which local $\boldsymbol{x}(\boldsymbol{s},t)$ and remote $z(\boldsymbol{r},t)$ covariates impact a local spatio-temporal response $Y(\boldsymbol{s},t)$. The RESP model accounts for the influence of covariates observed on a geographically remote domain $z(\boldsymbol{r},t)$.

We demonstrate the capacity of the RESP model by validating its ability to predict Colorado winter precipitation in a cross-validation study (Section 3.4). Our study represents a type of perfect prognosis problem in which future precipitation will be studied with covariates that have been simulated by GCMs. Since atmospheric processes have relatively short memory, it is reasonable to assume winter precipitation is conditionally independent across years when local and remote covariates are given. Therefore, we develop the RESP model assuming there is no meaningful temporal dependence. We conclude with discussions of temporal extensions and other directions for future work and further application (Section 3.5).

3.2 A geostatistical model for spatially remote covariates

Teleconnection manifests as an aggregate property of spatially continuous covariates. For example, consider the sea surface temperature (SST) at location $\boldsymbol{r}$ and time t, $z(\boldsymbol{r},t)$. In spatially

varying coefficient models (3.1), it is common to adopt a teleconnection index $z(t) \in \mathbb{R}$ that is defined as the average SST $z(\boldsymbol{r}, t)$ over a region $\mathscr{R} \subset \mathscr{D}_Z$. In (3.1), the spatially varying coefficient term $z(t)\theta(\boldsymbol{s})$ motivates the RESP model through the expansion

$$z(t)\theta(\boldsymbol{s}) = \frac{1}{|\mathscr{R}|}\int_{\mathscr{R}} z(\boldsymbol{r}, t)\theta(\boldsymbol{s})d\boldsymbol{r}. \tag{3.2}$$

The RESP model extends the integral in (3.2) to the entire remote domain $\mathscr{D}_Z$ and allows $\theta(\boldsymbol{s})$ to vary with respect to $\boldsymbol{r}$, distinguishing it from spatially varying coefficient models (Section 3.2.1). Integration is a natural construct for aggregating effects of spatially continuous covariates, represents the conceptual limit of studying teleconnection with increasingly fine subsets of $\mathscr{R}$, and allows study of teleconnection with additional spatial structure and without defining indices a priori.

3.2.1 Model formulation

The remote effects spatial process (RESP) model extends the standard geostatistical setting in which a local response variable $Y(\boldsymbol{s}, t) \in \mathbb{R}$ and known covariate vector $\boldsymbol{x}(\boldsymbol{s}, t) \in \mathbb{R}^p$ are observable at discrete time points $t \in \mathscr{T} = \{t_1, \ldots, t_{n_t}\}$ and at locations $\boldsymbol{s}$ in a continuous domain $\mathscr{D}_Y$. The RESP model includes the effects of known remote covariates $z(\boldsymbol{r}, t) \in \mathbb{R}$, which are observable at locations $\boldsymbol{r}$ in a continuous domain that is spatially disjoint from the local response—i.e., in a continuous $\mathscr{D}_Z$ s.t. $\mathscr{D}_Y \cap \mathscr{D}_Z = \emptyset$. The RESP model is given by

$$Y(\boldsymbol{s}, t) = \boldsymbol{x}^T(\boldsymbol{s}, t)\boldsymbol{\beta} + w(\boldsymbol{s}, t) + \varepsilon(\boldsymbol{s}, t) + \gamma(\boldsymbol{s}, t) \tag{3.3}$$

where the regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$, spatially correlated noise $w(\boldsymbol{s}, t)$, and independent noise $\varepsilon(\boldsymbol{s}, t)$ are standard components for spatial regression models (Banerjee et al., 2015, Chapters 6, 9, 11). In the RESP model the teleconnection effect given by $\gamma(\boldsymbol{s}, t)$ is defined by

$$\gamma(\boldsymbol{s}, t) = \int_{\mathscr{D}_Z} z(\boldsymbol{r}, t)\alpha(\boldsymbol{s}, \boldsymbol{r})d\boldsymbol{r} \tag{3.4}$$

which describes the net effect of the remote covariates $z(\boldsymbol{r}, t)$ on the continuous spatial process $Y(\boldsymbol{s}, t)$ at discrete time t. The integral (3.4) reduces to a sum for finite samples, in which the remote covariates $z(\boldsymbol{r}, t)$ are observed at $n_r < \infty$ locations. Multivariate extensions of (3.4) are discussed in Section 3.5.

The remote coefficients $\alpha(\boldsymbol{s}, \boldsymbol{r})$, also called teleconnection coefficients, are spatially correlated and doubly-indexed by $(\boldsymbol{s}, \boldsymbol{r}) \in \mathscr{D}_Y \times \mathscr{D}_Z$. The spatial correlation and double-indexing of $\alpha(\boldsymbol{s}, \boldsymbol{r})$ represents teleconnection effects that vary regionally in the sense that the response $Y(\boldsymbol{s}, t)$ at one location $\boldsymbol{s} \in \mathscr{D}_Y$ can respond to the remote covariates $z(\boldsymbol{r}, t)$ more strongly than the response $Y(\boldsymbol{s}', t)$ at another location $\boldsymbol{s}' \in \mathscr{D}_Y$. Similarly, the response $Y(\boldsymbol{s}, t)$ at one location $\boldsymbol{s} \in \mathscr{D}_Y$ can respond differently to remote covariates $z(\boldsymbol{r}, t)$ and $z(\boldsymbol{r}', t)$ at distinct remote locations $\boldsymbol{r}, \boldsymbol{r}' \in \mathscr{D}_Z$. Thus, the remote coefficients $\alpha(\boldsymbol{s}, \boldsymbol{r})$ vary spatially and use the remote covariates $z(\boldsymbol{r}, t)$ to provide local adjustment to the mean response. The teleconnection term $\gamma(\boldsymbol{s}, t)$ is well defined because we assume the remote covariates $z(\boldsymbol{r}, t)$ are known and square-integrable over $\mathscr{D}_Z$ at each time point t (Adler and Taylor, 2007, Section 5.2).

The RESP model provides a simple geostatistical approach to modeling teleconnections by extending spatial regression models to incorporate data from spatially remote regions. The teleconnection term $\gamma(\boldsymbol{s}, t)$ distinguishes the RESP model (3.3) from standard geostatistical models, in which—for example—the responses $Y(\boldsymbol{s}, t)$ and $Y(\boldsymbol{s}', t)$ at distinct spatial locations $\boldsymbol{s}, \boldsymbol{s}' \in \mathscr{D}_Y$ are only influenced by distinct covariates $x(\boldsymbol{s}, t)$ and $x(\boldsymbol{s}', t)$. To model the influence of teleconnection phenomena the RESP model lets the remote covariates $z(\boldsymbol{r}, t)$ simultaneously influence the responses $Y(\boldsymbol{s}, t)$ and $Y(\boldsymbol{s}', t)$.

Geostatistical modeling conventions use mean zero Gaussian processes to specify the randomness of the unknown, spatially correlated components $w(\boldsymbol{s}, t)$ and $\alpha(\boldsymbol{s}, \boldsymbol{r})$, and an independent processes to specify the noise $\varepsilon(\boldsymbol{s}, t)$—the nugget. We complete the Gaussian process specifications by defining the covariance functions for the spatially correlated components. Let C_w and C_α respectively be the covariance functions for $w(\boldsymbol{s}, t) + \varepsilon(\boldsymbol{s}, t)$ and $\alpha(\boldsymbol{s}, \boldsymbol{r})$, where

$$C_w\{(\boldsymbol{s},t),(\boldsymbol{s}',t')\} = \big(\kappa(\boldsymbol{s},\boldsymbol{s}';\boldsymbol{\theta}_w) + \sigma_\varepsilon^2 \mathbb{1}(\boldsymbol{s}=\boldsymbol{s}')\big)\mathbb{1}(t=t'), \tag{3.5}$$

$$C_\alpha\{(\boldsymbol{s},\boldsymbol{r}),(\boldsymbol{s}',\boldsymbol{r}')\} = \big(\kappa(\boldsymbol{s},\boldsymbol{s}';\boldsymbol{\theta}_w) + \sigma_\varepsilon^2 \mathbb{1}(\boldsymbol{s}=\boldsymbol{s}')\big)\kappa(\boldsymbol{r},\boldsymbol{r}';\boldsymbol{\theta}_\alpha). \tag{3.6}$$

Our model may be developed with any spatial covariance function κ, but here we choose to work with the stationary Matérn covariance

$$\kappa(\boldsymbol{u},\boldsymbol{v};\boldsymbol{\theta}) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\big(d(\boldsymbol{u},\boldsymbol{v})/\rho\big)^\nu K_\nu\big(d(\boldsymbol{u},\boldsymbol{v})/\rho\big) \tag{3.7}$$

for spatial locations $\boldsymbol{u}$ and $\boldsymbol{v}$, and parameter vector $\boldsymbol{\theta} = \big(\sigma^2,\ \rho,\ \nu\big)^T$. The function $d(\boldsymbol{u},\boldsymbol{v})$ must be an appropriate distance function (e.g., great-circle distances for locations on a sphere), $\sigma^2 > 0$ is a scaling parameter, $\nu > 0$ is a smoothness parameter, $\rho > 0$ is a range parameter, and K_ν is the modified Bessel function of the second kind with order ν. In covariance function definitions (3.5) and (3.6), $\mathbb{1}$ represents the indicator function and σ_ε^2 represents the variance of the nugget process which we specify to be a collection of independent and identically distributed mean zero Gaussian random
variables—i.e., $\varepsilon(\boldsymbol{s},t) \overset{iid}{\sim} \mathcal{N}\big(0,\ \sigma_\varepsilon^2\big)\ \forall(\boldsymbol{s},t) \in \mathscr{D}_Y \times \mathscr{T}$.

While the definitions (3.5) and (3.6) for the local and remote covariances C_w and C_α can be generalized, the definitions restrict our use of the RESP model to working in the perfect prognosis downscaling setting described at the end of Section 3.1. The responses $Y(\boldsymbol{s},t)$ and $Y\big(\boldsymbol{s},t'\big)$ for $t \neq t'$ are independent given covariates and sufficiently separated time points, like successive winters (e.g., winter 1991, winter 1992, etc.). The remote covariates in the teleconnection term (3.4) naturally induce temporal non-stationarity in the response's variance; extensions to accommodate serial dependence are discussed in Section 3.5. The remote covariance C_α also induces a separable structure for the remote coefficients $\alpha(\boldsymbol{s},\boldsymbol{r})$, which constrains the spatial variability of teleconnection effect fields and simultaneously constrains the teleconnection effects $\{\alpha(\boldsymbol{s},\boldsymbol{r}) : \boldsymbol{r} \in \mathscr{D}_Z\}$ and $\big\{\alpha\big(\boldsymbol{s}',\boldsymbol{r}\big) : \boldsymbol{r} \in \mathscr{D}_Z\big\}$ to be similar for nearby locations $\boldsymbol{s},\ \boldsymbol{s}' \in \mathscr{D}_Y$. Simpler covariance structures for the teleconnection effects $\alpha(\boldsymbol{s},\boldsymbol{r})$ may not capture these physical

properties of teleconnection as directly. Similarly, although climate data are often available as gridded data products, we choose to work with geostatistical covariance models (or their discrete approximations, e.g., Lindgren et al., 2011) instead of neighborhood-based spatial models so that we may avoid inducing potentially counterintuitive covariance structures (Assunção and Krainski, 2009; Wall, 2004).

3.2.2 Reduced rank approximation

To apply the RESP model (3.3), additional constraints need to be imposed due to the potential multicollinearity in the covariates. Remote covariates $z(\boldsymbol{r}, t)$ in teleconnection applications will often consist of data that measure ocean properties at high spatial resolution, like sea surface temperature or sea level pressure. This raises concerns for estimating the remote coefficients $\alpha(\boldsymbol{s}, \boldsymbol{r})$ in (3.4) as the main trends in the remote covariates $z(\boldsymbol{r}, t)$ are highly collinear over $\mathscr{D}_Z$. Physically, however, this suggests the remote coefficients should be highly correlated as well. We use predictive processes to mitigate multicollinearity in the remote covariates, which is an alternative motivation for predictive processes. Banerjee et al. (2008) originally introduce predictive processes so that parameters of geostatistical models can be estimated for large spatial datasets, rather than as an approach for mitigating spatial multicollinearity. We consider more general basis expansions of remote coefficients in Section 3.2.3.

We assume the remote coefficients $\alpha(\boldsymbol{s}, \boldsymbol{r})$ can be well represented by weighted averages of remote coefficients $\alpha(\boldsymbol{s}, \boldsymbol{r}^*)$ at knot locations $\boldsymbol{r}_1^*, \ldots, \boldsymbol{r}_k^* \in \mathscr{D}_Z$, so we make the simplifying approximation that a weight function $h(\boldsymbol{r}, \boldsymbol{r}')$ exists and induces $\boldsymbol{h}^*(\boldsymbol{r}) = \left[h\left(\boldsymbol{r}, \boldsymbol{r}_j^*\right)\right]_{j=1}^k \in \mathbb{R}^k$, allowing us to write

$$\alpha(\boldsymbol{s}, \boldsymbol{r}) = \sum_{j=1}^{k} h\left(\boldsymbol{r}, \boldsymbol{r}_j^*\right) \alpha\left(\boldsymbol{s}, \boldsymbol{r}_j^*\right) = \boldsymbol{h}^*(\boldsymbol{r})^T \boldsymbol{\alpha}^*(\boldsymbol{s}), \tag{3.8}$$

where $\boldsymbol{\alpha}^*(\boldsymbol{s}) = \left[\alpha\left(\boldsymbol{s}, \boldsymbol{r}_j^*\right)\right]_{j=1}^k \in \mathbb{R}^k$. The predictive process approach uses kriging to motivate a choice for the weight vector $\boldsymbol{h}^*(\boldsymbol{r})$, which induces a weight function h. Using Gaussian pro-

cesses in Section 3.2.1 to model the remote coefficients implies that $\alpha(\boldsymbol{s},\boldsymbol{r})$ and $\boldsymbol{\alpha}^*(\boldsymbol{s})$ are jointly normally distributed, yielding the conditional expectation for $\alpha(\boldsymbol{s},\boldsymbol{r})$

$$E\left[\,\alpha(\boldsymbol{s},\boldsymbol{r})|\,\boldsymbol{\alpha}^*(\boldsymbol{s})\right] = \boldsymbol{c}^*(\boldsymbol{r})^T R^{*-1}\boldsymbol{\alpha}^*(\boldsymbol{s}) \tag{3.9}$$

in which $\boldsymbol{c}^*(\boldsymbol{r}) = \left[C_\alpha\left\{(\boldsymbol{s},\boldsymbol{r}),\left(\boldsymbol{s},\boldsymbol{r}_j^*\right)\right\}\right]_{j=1}^k \in \mathbb{R}^k$ and $R^* \in \mathbb{R}^{k\times k}$ such that $R_{ij}^* = C_\alpha\left\{\left(\boldsymbol{s},\boldsymbol{r}_i^*\right),\left(\boldsymbol{s},\boldsymbol{r}_j^*\right)\right\}$. Note that the assumption in (3.6) that C_α is stationary means that $\boldsymbol{c}^*(\boldsymbol{r})$ and R^* do not depend on $\boldsymbol{s}$, despite the term appearing in their definitions. The predictive process approach uses the conditional expectation (3.9) to define the weight vector $\boldsymbol{h}^*(\boldsymbol{r}) = R^{*-1}\boldsymbol{c}^*(\boldsymbol{r})$ in the approximation (3.8). Banerjee et al. (2008) show that these types of approximations are reduced rank projections that can capture large-scale spatial structures in data.

Beyond mitigating the statistical issue of multicollinearity in the remote covariates, the predictive process approach relates the RESP model to spatially varying coefficient models (3.1) and also has a scientific interpretation for teleconnection. Using the reduced rank approximation (3.8) to manipulate the integral in (3.3) shows that the reduced rank approximation can be interpreted as inducing transformed covariates $z^*(\boldsymbol{r}^*,t)$ via

$$\begin{aligned}\int_{\mathscr{D}_Z} z(\boldsymbol{r},t)\alpha(\boldsymbol{s},\boldsymbol{r})d\boldsymbol{r} &= \int_{\mathscr{D}_Z} z(\boldsymbol{r},t)\sum_{j=1}^k h\left(\boldsymbol{r},\boldsymbol{r}_j^*\right)\alpha\left(\boldsymbol{s},\boldsymbol{r}_j^*\right)d\boldsymbol{r}\\ &= \sum_{j=1}^k \alpha\left(\boldsymbol{s},\boldsymbol{r}_j^*\right)z^*\left(\boldsymbol{r}_j^*,t\right)\end{aligned}$$

where $z^*\left(\boldsymbol{r}_j^*,t\right) = \int_{\mathscr{D}_Z} z(\boldsymbol{r},t)h\left(\boldsymbol{r},\boldsymbol{r}_j^*\right)d\boldsymbol{r}$. The $z^*\left(\boldsymbol{r}_j^*,t\right)$ and $\alpha\left(\boldsymbol{s},\boldsymbol{r}_j^*\right)$ may be collected into the covariate vector $\boldsymbol{z}(t)$ and spatially varying effects $\boldsymbol{\theta}(\boldsymbol{s})$ in (3.1). We remark that the RESP model differs from standard spatially varying coefficient models in that the transformed covariates $z^*\left(\boldsymbol{r}_j^*,t\right)$ represent induced—rather than a priori—covariates, and the $\alpha\left(\boldsymbol{s},\boldsymbol{r}_j^*\right)$ inherit spatial structure from the model's formulation.

Scientifically, the predictive process approach to addressing multicollinearity reduces the remote covariates $z(\boldsymbol{r},t)$, $\boldsymbol{r}\in\mathscr{D}_Z$ at each time point to k spatially-averaged indices $z^*(\boldsymbol{r}^*,t)$

centered at $\boldsymbol{r}^*$ for $\boldsymbol{r}^* \in \{\boldsymbol{r}_1^*, \ldots, \boldsymbol{r}_k^*\}$. This manipulation is fairly generic and should be applicable to all predictive process models. For teleconnection, this manipulation connects the RESP model to one set of standard teleconnection methodologies in which teleconnection effects are measured with respect to ocean indices based on spatial averages of remote covariates (Ashok et al., 2007; Towler et al., 2016).

3.2.3 Spatial basis function transformation of remote coefficients

The RESP model (3.3) is also related to another set of standard teleconnection methodologies in which teleconnection effects are measured with respect to complex ocean indices such as empirical orthogonal functions (Montroy, 1997; Ting and Wang, 1997). Spatial basis functions provide a means to reparameterize the RESP model and show it can identify and leverage known teleconnections with complex patterns. We use the following reparameterization of the teleconnection effects $\alpha(\boldsymbol{s}, \boldsymbol{r})$ to discuss teleconnection between Pacific Ocean sea surface temperature and Colorado precipitation in Section 3.4.

Complex teleconnection patterns are often based on spatial basis function expansions of the remote covariates $z(\boldsymbol{r}, t)$. If there exist weights $\{a_l(t) : l = 1, \ldots, K;\ t \in \mathcal{T}\}$ such that the remote covariates $z(\boldsymbol{r}, t)$ can be written as a linear combination of continuous, time-invariant basis functions $\{\psi_l(\boldsymbol{r}) : l = 1, \ldots, K;\ \boldsymbol{r} \in \mathcal{D}_Z\}$ via

$$z(\boldsymbol{r}, t) = \sum_{l=1}^{K} a_l(t)\psi_l(\boldsymbol{r}), \tag{3.10}$$

then linearity of the integral in (3.4) and reduced rank approximation (3.8) can induce a reparameterized, reduced-rank teleconnection effect process $\alpha'(\boldsymbol{s}, l)$ for patterns $l = 1, \ldots, K$ by

$$\alpha'(\boldsymbol{s}, l) = \sum_{j=1}^{k} \alpha\left(\boldsymbol{s}, \boldsymbol{r}_j^*\right) \int_{\mathcal{D}_Z} \psi_l(\boldsymbol{r}) h\left(\boldsymbol{r}, \boldsymbol{r}_j^*\right) d\boldsymbol{r}\ . \tag{3.11}$$

Note that the transformation appears naturally because

$$
\begin{aligned}
\int_{\mathscr{D}_Z} z(\boldsymbol{r},t)\alpha(\boldsymbol{s},\boldsymbol{r})d\boldsymbol{r} &= \int_{\mathscr{D}_Z} \sum_{l=1}^{K} a_l(t)\psi_l(\boldsymbol{r}) \sum_{j=1}^{k} h\left(\boldsymbol{r},\boldsymbol{r}_j^*\right)\alpha\left(\boldsymbol{s},\boldsymbol{r}_j^*\right)d\boldsymbol{r} \\
&= \sum_{l=1}^{K} a_l(t) \sum_{j=1}^{k} \alpha\left(\boldsymbol{s},\boldsymbol{r}_j^*\right) \int_{\mathscr{D}_Z} \psi_l(\boldsymbol{r}) h\left(\boldsymbol{r},\boldsymbol{r}_j^*\right) d\boldsymbol{r} \\
&= \sum_{l=1}^{K} a_l(t)\alpha'(\boldsymbol{s},l).
\end{aligned} \tag{3.12}
$$

As with the reduced rank approximation (3.8), the transformation (3.12) also relates the RESP model to spatially varying coefficient models (3.1) and has scientific relevance for teleconnection. The deterministic remote covariate weights $a_l(t)$ and reparameterized remote coefficients $\alpha'(\boldsymbol{s},l)$ may be collected into the covariate vector $\boldsymbol{z}(t)$ and spatially varying effects $\boldsymbol{\theta}(\boldsymbol{s})$ in (3.1). While the covariate weights $a_l(t)$ suggest a priori selection of teleconnection indices, the reparameterization may be applied after model estimation. The $\alpha'(\boldsymbol{s},l)$ additionally inherit spatial structure from the model's formulation. Scientifically, a special case of (3.10) are principal component decompositions or the closely related truncated Karhunen–Lòeve expansions, which are referred to as empirical orthogonal functions (EOFs) in climate science. EOFs are particularly useful expansions for teleconnection because these transformations meaningfully characterize phenomena that impact global climate (Ashok et al., 2007).

3.2.4 Inference

While inference for the RESP model (3.3) can use standard hierarchical Bayesian modeling techniques, the Bayesian framework provides crucial intuition and interpretation for estimates of teleconnection effects (3.8) and (3.11). Full description of model priors and computational techniques for inference are discussed in Section 3.3. The Gaussian process assumption and separable covariance (3.6) for the vector of teleconnection coefficients $\boldsymbol{\alpha}^*(\boldsymbol{s})$ with associated covariance matrix R^* defined in Section 3.2.2 imply the normally-distributed prior $\boldsymbol{\alpha}^*(\boldsymbol{s})|R^* \sim \mathcal{N}(\boldsymbol{0},R^*)$. Gaussian process assumptions for the RESP model's spatial correlation also imply that the likelihood for the vector of responses observed at all n_t timepoints $\boldsymbol{Y}(\boldsymbol{s}) = \left[Y(\boldsymbol{s},t_1),\ldots,Y\left(\boldsymbol{s},t_{n_t}\right)\right]^T \in \mathbb{R}^{n_t}$ is

$$\boldsymbol{Y}(\boldsymbol{s})|\,\boldsymbol{\alpha}^*(\boldsymbol{s}),\boldsymbol{\beta},R^*,\boldsymbol{c}^*,\sigma^2_{\boldsymbol{s}} \sim \mathcal{N}\left(\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}+\boldsymbol{Z}^{*\,T}\boldsymbol{\alpha}^*(\boldsymbol{s}),\ \sigma^2_{\boldsymbol{s}} I_{n_t}\right)$$

with $\sigma^2_{\boldsymbol{s}} = C_w\{(\boldsymbol{s},t),(\boldsymbol{s},t)\}$ and matrices of local covariates $\boldsymbol{X}(\boldsymbol{s}) = \left[\boldsymbol{x}(\boldsymbol{s},t)^T\right]_{t=t_1}^{t_{n_t}} \in \mathbb{R}^{n_t\times p}$ and reduced-rank remote covariates $\boldsymbol{Z}^* \in \mathbb{R}^{k\times n_t}$. The matrix $\boldsymbol{Z}^*$ is comprised of column vectors $\boldsymbol{z}^*_t = R^{*\,-1}\boldsymbol{c}^{*\,T}\boldsymbol{z}_t \in \mathbb{R}^k$ built from remote covariate vectors $\boldsymbol{z}_t = \left[z\left(\boldsymbol{r}_j,t\right)\right]_{j=1}^{n_r} \in \mathbb{R}^{n_r}$. Our formulation of the spatial correlation (3.5) implies the scalar $\sigma^2_{\boldsymbol{s}}$ is constant across time; non-stationary extensions are discussed in Section 3.5. Standard Bayesian linear regression results (Banerjee et al., 2015, Example 5.2) yield the posterior distribution

$$\boldsymbol{\alpha}^*(\boldsymbol{s})\Big|\,\boldsymbol{Y}(\boldsymbol{s}),\boldsymbol{\beta},R^*,\boldsymbol{c}^*,\sigma^2_{\boldsymbol{s}} \sim \mathcal{N}\left(\sigma^{-2}_{\boldsymbol{s}}\boldsymbol{\Psi}\boldsymbol{Z}^*\left(\boldsymbol{Y}(\boldsymbol{s})-\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}\right),\ \boldsymbol{\Psi}\right)$$

for

$$\boldsymbol{\Psi} = \left(R^{*\,-1}+\sigma^{-2}_{\boldsymbol{s}}\boldsymbol{Z}^*\boldsymbol{Z}^{*\,T}\right)^{-1}.$$

The connection to Bayesian linear regression lends intuition for inference on the remote effects $\boldsymbol{\alpha}^*(\boldsymbol{s})$. In particular, the connection provides intuition for using the RESP model when some local covariates $\boldsymbol{x}(\boldsymbol{s},t)$ are also teleconnected with remote covariates $\boldsymbol{z}_t$. Remote coefficients can be interpreted as residual teleconnection effects in the sense that they model the impact of remote covariates on the response after removing local effects $\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}$. Properties of regressions also imply patterns in maps of posterior means for $\boldsymbol{\alpha}^*(\boldsymbol{s})$ may resemble patterns in maps that show pointwise correlations $\mathrm{Cor}_t\left(z^*(\boldsymbol{r}^*,t),Y(\boldsymbol{s},t)\right)$ between remote covariates at $\boldsymbol{r}^*$ and responses at $\boldsymbol{s}$. Similar regression-based interpretations can be derived for the reparameterized teleconnection coefficients (3.11).

3.3 Bayesian implementation of the RESP model

We adopt a hierarchical Bayesian framework and use a hybrid Gibbs sampler for inference for the RESP model (3.3) using the likelihood (3.13). The Bayesian framework allows estimates

for the transformed teleconnection effects (3.11) to be computed directly from posterior samples of $\tilde{\boldsymbol{\alpha}}^*$ by using the definition for $\tilde{\boldsymbol{\alpha}}'$ specified after (3.14) to appropriately transform the sampled teleconnection effects $\tilde{\boldsymbol{\alpha}}^*$. Algebraic manipulation and computational evaluation of the RESP model (3.3) likelihood is simplified through properties of Kronecker products.

3.3.1 Model likelihood

Using Gaussian processes to specify the RESP model's (3.3) randomness implies the data model is jointly normal for finite samples with n_s locations, n_r remote locations, and n_t time points. Let the column vectors $\boldsymbol{Y}_t = [Y(\boldsymbol{s}_i, t)]_{i=1}^{n_s} \in \mathbb{R}^{n_s}$ and $\boldsymbol{z}_t = \left[z\left(\boldsymbol{r}_j, t\right)\right]_{j=1}^{n_r} \in \mathbb{R}^{n_r}$, and the matrix $X_t \in \mathbb{R}^{n_s \times p}$ with row vectors $\boldsymbol{x}(\boldsymbol{s}_i, t)^T$ for $i = 1, \ldots, n_s$ represent the observed response variables and covariates at time t; and let the column vector $\boldsymbol{\alpha}(\boldsymbol{s}) = \left[\alpha\left(\boldsymbol{s}, \boldsymbol{r}_j\right)\right]_{j=1}^{n_r} \in \mathbb{R}^{n_r}$ represent the teleconnection coefficients for location $\boldsymbol{s}$. The reduced rank assumption lets us use the Kriging notation from (3.9) to write $\boldsymbol{\alpha}(\boldsymbol{s}) = \boldsymbol{c}^* R^{*-1} \boldsymbol{\alpha}^*(\boldsymbol{s})$. The matrix $\boldsymbol{c}^* \in \mathbb{R}^{n_r \times k}$ is built with row vectors $\boldsymbol{c}^*(\boldsymbol{r}_i)^T$ for $i = 1, \ldots, n_r$ that contain the covariances among the teleconnection effect $\alpha(\boldsymbol{s}, \boldsymbol{r}_i)$ and the teleconnection effects at knot locations
$\alpha\left(\boldsymbol{s}, \boldsymbol{r}_j^*\right), j = 1, \ldots, k$. This yields the data model for $\boldsymbol{Y} = \left[\boldsymbol{Y}_{t_1}^T \ldots \boldsymbol{Y}_{t_{n_t}}^T\right]^T \in \mathbb{R}^{n_s n_t}$, which is given by

$$\boldsymbol{Y} \mid \boldsymbol{\beta}, \tilde{\boldsymbol{\alpha}}^*, R^*, \boldsymbol{c}^*, \Sigma \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{Y}},\ I_{n_t} \otimes \Sigma\right) \tag{3.13}$$

in which

$$\boldsymbol{\mu}_{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\left(\mathbf{1}_{n_t} \otimes \boldsymbol{\beta}\right) + \tilde{\boldsymbol{Z}}^*\left(\mathbf{1}_{n_t} \otimes \tilde{\boldsymbol{\alpha}}^*\right),$$

where $\otimes$ denotes the Kronecker product, $\tilde{\boldsymbol{X}} = \operatorname{diag}\left\{X_{t_1}, \ldots, X_{t_{n_t}}\right\}$,
$\tilde{\boldsymbol{Z}}^* = \operatorname{diag}\left\{I_{n_s} \otimes \boldsymbol{z}_{t_1}^{*\,T}, \ldots, I_{n_s} \otimes \boldsymbol{z}_{t_{n_t}}^{*\,T}\right\}$, $\boldsymbol{z}_t^{*\,T} = \boldsymbol{z}_t^T \boldsymbol{c}^* R^{*-1} \in \mathbb{R}^{1 \times k}$, $\tilde{\boldsymbol{\alpha}}^* = [\boldsymbol{\alpha}^*(\boldsymbol{s}_i)]_{i=1}^{n_s} \in \mathbb{R}^{n_s k}$, and $\Sigma \in \mathbb{R}^{n_s \times n_s}$ is the local covariance matrix with entries $\Sigma_{ij} = C_w\left\{(\boldsymbol{s}_i, t), (\boldsymbol{s}_j, t)\right\}$. While the covariate matrices $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n_s n_t \times p n_t}$ and $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{n_s n_t \times n_s k n_t}$ are block diagonal, we later introduce alternate notation to make evaluating posterior distributions easier.

Each time-indexed block in the remote effects term $\tilde{\boldsymbol{Z}}^*\left(\mathbf{1}_{n_t} \otimes \tilde{\boldsymbol{\alpha}}^*\right)$ has the form $\left(I_{n_s} \otimes \boldsymbol{z}_t^{*T}\right)\tilde{\boldsymbol{\alpha}}^*$, which helps show how each response $Y(\boldsymbol{s}, t)$ at time t shares the same remote covariates $\boldsymbol{z}_t^*$. Although the remote coefficients $\alpha(\boldsymbol{s}, \boldsymbol{r}^*)$, $\boldsymbol{s} \in \mathscr{D}_Y$ vary spatially across $\mathscr{D}_Y$ for a fixed $\boldsymbol{r}^* \in \mathscr{D}_Z$, the RESP model differs from traditional spatially varying coefficient models (Banerjee et al., 2015, Section 9.6) because all of these remote coefficients share the same covariate $z^*(\boldsymbol{r}^*, t)$.

The likelihood (3.13) changes subtly when reparameterizing the teleconnection effects (3.4) to interpret them with respect to the spatial basis function transformation defined by (3.11). The spatial basis function expansion (3.10) of the remote covariates $z(\boldsymbol{r}, t)$ yields the substitution

$$\tilde{\boldsymbol{Z}}^*\left(\mathbf{1}_{n_t} \otimes \tilde{\boldsymbol{\alpha}}^*\right) = \tilde{\boldsymbol{A}}\left(\mathbf{1}_{n_t} \otimes \tilde{\boldsymbol{\alpha}}'\right) \tag{3.14}$$

in the likelihood (3.13). Where $\tilde{\boldsymbol{A}} = \operatorname{diag}\left\{I_{n_s} \otimes \boldsymbol{A}_{t_1}^T, \ldots, I_{n_s} \otimes \boldsymbol{A}_{t_{n_t}}^T\right\}$ and $\tilde{\boldsymbol{\alpha}}' = \left(I_{n_s} \otimes W^T \boldsymbol{c}^* R^{*-1}\right)\tilde{\boldsymbol{\alpha}}^*$, and $\tilde{\boldsymbol{\alpha}}^* \in \mathbb{R}^{n_s K}$ where the vector $\boldsymbol{A}_t$ and matrix W form the matrix decomposition of the remote covariate vector $\boldsymbol{z}_t$ when expanded by spatial basis functions $\boldsymbol{z}_t = W\boldsymbol{A}_t$. The column vector $\boldsymbol{A}_t = [a_l(t)]_{l=1,\ldots,K} \in \mathbb{R}^K$ contains the weights at time t for the basis functions $\{\psi_l(\boldsymbol{r}) : l = 1, \ldots, K\}$, which are stored in the basis function matrix $W \in \mathbb{R}^{n_r \times K}$ with entries $W_{jl} = \psi_l(\boldsymbol{r}_j)$.

3.3.2 Likelihood marginalization

We note that $cA = A \otimes c$ for $c \in \mathbb{R}$ and $A \in \mathbb{R}^{m \times n}$. This lets us use the mixed product rule for Kronecker products (Banerjee and Roy, 2014, Thm. 14.3) to distribute matrix multiplication across Kronecker products involving vectors. For example, let $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, and $B \in \mathbb{R}^{n \times p}$, then

$$(A \otimes c)B = (A \otimes c)(B \otimes 1) = AB \otimes c.$$

3.3.3 Numerical evaluation of likelihood

Evaluating the RESP likelihood involves computing matrix multiplications that involve Kronecker products. For example, define matrices $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times q}$, and $C \in \mathbb{R}^{nq\times r}$. While computing $(A \otimes B)C$ naively requires $O(mnpqr)$ floating point operations, it can be computed in $O(mpqr + mnqr)$ floating point operations by recognizing that

$$(A \otimes B)C = \begin{bmatrix} B\left(\sum_{j=1}^{n} a_{1j} C_j\right) \\ \vdots \\ B\left(\sum_{j=1}^{n} a_{mj} C_j\right) \end{bmatrix}$$

where C_j represents the j^{th} $q \times r$ block matrix in C, i.e., that $C_j \in \mathbb{R}^{q\times r}$ for $j \in \{1,\ldots,n\}$ and

$$C = \begin{bmatrix} C_1 \\ \vdots \\ C_n \end{bmatrix}.$$

While Banerjee and Roy (2014) discuss a similar idea in Section 14.7, they limit their treatment to the case in which C is a vector. Their discussion also does not present this direct form for numerical evaluation; they present results that rely on the vec$(\cdot)$ operation instead.

3.3.4 Gibbs sampler

We use conjugate prior distributions to specify our Bayesian model where possible, setting $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$ for a fixed prior covariance matrix $\boldsymbol{\Lambda}$ and $\sigma_w^2 \sim IG\left(a_{\sigma_w^2}, b_{\sigma_w^2}\right)$. We use standard choices for weakly informative priors for the remaining parameters: $\sigma_\alpha^2 \sim IG\left(a_{\sigma_\alpha^2}, b_{\sigma_\alpha^2}\right)$, $\sigma_\varepsilon^2 \sim IG\left(a_{\sigma_\varepsilon^2}, b_{\sigma_\varepsilon^2}\right)$, $\rho_w \sim U\left(a_{\rho_w}, b_{\rho_w}\right)$, and $\rho_\alpha \sim U\left(a_{\rho_\alpha}, b_{\rho_\alpha}\right)$ (Banerjee et al., 2008). As Matérn smoothness parameters are difficult to estimate in standard applications, we estimate ν_w and ν_α from sample variograms and treat these parameters as fixed during model fitting.

Bayesian estimation is often more stable after integrating out latent fields (Banerjee et al., 2015, pg. 126). The special case of Kronecker product rules reviewed in Section 3.3.2 facilitates

this integration. The marginalized data likelihood with n_t timepoints after integrating out $\tilde{\boldsymbol{\alpha}}^*$ is given by

$$\boldsymbol{Y} \,|\, \boldsymbol{\beta}, R^*, \boldsymbol{c}^*, \Sigma \sim \mathcal{N}\left(\tilde{\boldsymbol{X}}\left(\mathbf{1}_{n_t} \otimes \boldsymbol{\beta}\right),\ C^{-1} \otimes \Sigma\right) \tag{3.15}$$

where $C^{-1} = I_{n_t} + \boldsymbol{Z}^{*T} R^* \boldsymbol{Z}^*$. Since the Kronecker product is a bilinear operator, the marginalized variance $C^{-1} \otimes \Sigma$ decomposes into the sum $\left(I_{n_t} \otimes \Sigma\right) + \left(\boldsymbol{Z}^{*T} R^* \boldsymbol{Z}^* \otimes \Sigma\right)$ which more clearly highlights how the remote covariates account for some of the spatial variability around the fixed mean $\tilde{\boldsymbol{X}}\left(\mathbf{1}_{n_t} \otimes \boldsymbol{\beta}\right)$.

The data likelihood (3.15) is almost fully identified. The spatial covariance matrix Σ has entries $\Sigma_{ij} = C_w\left\{(\boldsymbol{s}_i, t), (\boldsymbol{s}_j, t)\right\}$ based on the covariance function C_w defined in (3.5). The covariance's scale parameters σ_w^2 and σ_ε^2 are only identifiable with respect to their product $\sigma_w^2 \sigma_\varepsilon^2$. We use parameter expansion to remedy the identifiability issue by reparameterizing the nugget variance as $\sigma_\varepsilon^2 = \sigma_w^2 \tilde{\sigma}_\varepsilon^2$. Therefore, in model fitting, we estimate $\tilde{\sigma}_\varepsilon^2$ instead of estimating σ_ε^2 directly.

Model fitting employs a hybrid Gibbs sampler with adaptive random walk Metropolis steps to estimate parameters for the marginalized likelihood (3.15). Likelihood evaluation uses results detailed in Section 3.3.3 that efficiently implement Kronecker product matrix multiplication. Sampling proceeds by updating the regression coefficient and spatial variance parameters $\boldsymbol{\beta}$ and σ_w^2 by drawing from their full conditional posterior distributions

$$\boldsymbol{\beta}\big|\cdot \ \sim\ \mathcal{N}\left(\Sigma_{\boldsymbol{\beta}|\cdot} \boldsymbol{X}^T \left(C \otimes \Sigma^{-1}\right) \boldsymbol{Y},\ \Sigma_{\boldsymbol{\beta}|\cdot}\right), \tag{3.16}$$

$$\sigma_w^2\big|\cdot \ \sim\ IG\left(a_{\sigma_w^2} + n_s n_t/2,\ b_{\sigma_w^2} + e^T \left[C \otimes \left(\Sigma/\sigma_w^2\right)^{-1}\right] e/2\right) \tag{3.17}$$

where $\boldsymbol{X} \in \mathbb{R}^{n_s n_t \times p}$ is a block row matrix with row blocks X_{t_i} for $i = 1, \ldots, n_t$, the column vector $e = \left(\boldsymbol{Y} - \tilde{\boldsymbol{X}}\left(\mathbf{1}_{n_t} \otimes \boldsymbol{\beta}\right)\right)$ are the model residuals, and

$$\Sigma_{\boldsymbol{\beta}|\cdot} = \left\{\boldsymbol{\Lambda}^{-1} + \boldsymbol{X}^T \left(C \otimes \Sigma^{-1}\right) \boldsymbol{X}\right\}^{-1} \tag{3.18}$$

is the posterior covariance matrix for $\boldsymbol{\beta}$. The remaining parameters ρ_w, ρ_α, σ_α^2, and $\tilde{\sigma}_\varepsilon^2$ are transformed to unconstrained supports and updated using adaptive random walk Metropolis steps with normal proposals. Log transformations are used for the positive parameters σ_α^2 and $\tilde{\sigma}_\varepsilon^2$, and logit transformations are used for the bounded parameters ρ_w and ρ_α. The adaptive proposal variance $\lambda(\cdot)$ differs for each parameter and is tuned at each iteration following a basic version of Algorithm 4 from Andrieu and Thoms (2008). Additional computations, however, are required to estimate the remote coefficients and make predictions.

Composition sampling (Banerjee et al., 2015, p. 126) provides a means to sample from the posterior distribution of the remote coefficients $\tilde{\boldsymbol{\alpha}}^*$ (3.19) as well as from the posterior predictive distribution of the response variables $\boldsymbol{Y}_{t_0}$ at a new timepoint t_0. This allows inference and prediction of these processes, which the Gibbs sampler does not directly study.

The Gaussian process assumption and separable covariance (3.6) imply the remote coefficients have prior distribution

$$\tilde{\boldsymbol{\alpha}}^* \big| \Sigma, R^* \sim \mathcal{N}\big(\mathbf{0},\, \Sigma \otimes R^*\big).$$

The full conditional posterior distribution for $\tilde{\boldsymbol{\alpha}}^*$ is

$$\tilde{\boldsymbol{\alpha}}^* \big| \cdot \sim \mathcal{N}\big(\boldsymbol{\mu}_{\tilde{\boldsymbol{\alpha}}^*|\cdot},\, \Sigma_{\tilde{\boldsymbol{\alpha}}^*|\cdot}\big), \tag{3.19}$$

where "$\cdot$" represents conditioning on all remaining unknown quantities and

$$\begin{aligned}
\boldsymbol{\mu}_{\tilde{\boldsymbol{\alpha}}^*|\cdot} &= \sum_{t \in \mathcal{T}} \left\{ \big(\boldsymbol{Y}_t - X_t \boldsymbol{\beta}\big) \otimes \left(R^{*-1} + \boldsymbol{Z}^* \boldsymbol{Z}^{*T}\right)^{-1} \boldsymbol{z}_t^* \right\}, \\
\Sigma_{\tilde{\boldsymbol{\alpha}}^*|\cdot} &= \Sigma \otimes \left(R^{*-1} + \boldsymbol{Z}^* \boldsymbol{Z}^{*T}\right)^{-1},
\end{aligned}$$

and the matrix $\boldsymbol{Z}^* \in \mathbb{R}^{k \times n_t}$ with column vectors $\boldsymbol{z}_{t_i}^* \in \mathbb{R}^k$ for $i = 1, \ldots, n_t$ is a dense matrix that contains the remote covariates.

3.3.5 Computational approach for conducting inference on remote coefficients

We use standard hierarchical Bayesian spatial modeling techniques to draw inference on $\tilde{\boldsymbol{\alpha}}^*$ through composition sampling (Banerjee et al., 2015, p. 126). Composition sampling generates a posterior sample $\{\tilde{\boldsymbol{\alpha}}^{*(1)}, \ldots, \tilde{\boldsymbol{\alpha}}^{*(G)}\}$ for $\tilde{\boldsymbol{\alpha}}^*$ by using the full conditional posterior distribution (3.19) for $\tilde{\boldsymbol{\alpha}}^*$ with a posterior sample of the model parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}_w$, $\boldsymbol{\theta}_\alpha$, and σ^2_ε. The composition samples may be drawn in parallel to reduce the computation time because composition samples $\tilde{\boldsymbol{\alpha}}^{*(i)}$ are independent given the posterior parameter samples. Drawing inference on $\tilde{\boldsymbol{\alpha}}^*$ also requires computational techniques to reduce memory demands.

The composition sample for $\tilde{\boldsymbol{\alpha}}^*$ requires storing $n_s \times k \times G$ floating point numbers. Even for moderately sized studies with $n_s = 200$, $k = 30$, and $G = 20{,}000$, the composition sample requires 915MB of memory. Although this demand increases linearly in k, n_s, and G, it quickly becomes burdensome for typical personal computers. We therefore estimate $\tilde{\boldsymbol{\alpha}}^*$ using the normal approximation to the posterior. The normal approximation only requires the composition sample's mean $\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\alpha}}|\boldsymbol{Y}} = \frac{1}{G}\sum_{g=1}^{G} \tilde{\boldsymbol{\alpha}}^{*(g)}$ and covariance matrix, the latter defined via

$$\hat{\Sigma}_{\tilde{\boldsymbol{\alpha}}|\boldsymbol{Y}} = \frac{1}{G-1}\sum_{g=1}^{G}\left(\tilde{\boldsymbol{\alpha}}^{*(g)} - \hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\alpha}}|\boldsymbol{Y}}\right)\left(\tilde{\boldsymbol{\alpha}}^{*(g)} - \hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\alpha}}|\boldsymbol{Y}}\right)^T.$$

These objects require storing $(n_s \times k)(n_s \times k + 3)/2$ floating point numbers, which can dramatically reduce memory requirements when $G > (n_s \times k + 3)/2$.

We use strategies from Pébay (2008) to facilitate computing these summary objects with minimal memory requirements. We use partitions of the composition samples and compute $\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\alpha}}|\boldsymbol{Y}}$ and $\hat{\Sigma}_{\tilde{\boldsymbol{\alpha}}|\boldsymbol{Y}}$ in a streaming fashion, which allows estimation of $p(\tilde{\boldsymbol{\alpha}}^* \mid \boldsymbol{Y})$ in parallel and with minimal memory requirements (Pébay, 2008, eqs. 1.1, 1.3, 3.11, & 3.12). These benefits are achieved by recognizing, for example, that a running estimate of $\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\alpha}}|\boldsymbol{Y}}$ based on $\{\tilde{\boldsymbol{\alpha}}^{*(1)}, \ldots, \tilde{\boldsymbol{\alpha}}^{*(g)}\}$ is easy to update when the next composition sample $\tilde{\boldsymbol{\alpha}}^{*(g+1)}$ is drawn, and that the updating equations yield $\hat{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\alpha}}|\boldsymbol{Y}}$ after all G composition samples are processed.

3.4 Climate application: Colorado winter precipitation

The RESP model (3.3) is applied here using remote and local covariates to estimate Colorado winter precipitation. Winter precipitation is important to estimate because it strongly in influences Colorado's annual water supply. We investigate the utility of our RESP model for this application because there is considerable uncertainty regarding precipitation that is directly predicted by GCMs. Further, the RESP model can be applied without specifying teleconnection indices a priori, as many common approaches require. Let $Y(\boldsymbol{s}, t)$ denote average monthly precipitation in winter for location $\boldsymbol{s}$ and year t via

$$Y(\boldsymbol{s}, t) = \big(Y_{Dec}(\boldsymbol{s}, t) + Y_{Jan}(\boldsymbol{s}, t) + Y_{Feb}(\boldsymbol{s}, t)\big)/3$$

in which, for example, $Y_{Dec}(\boldsymbol{s}, t)$ represents the total December precipitation in year t at location $\boldsymbol{s}$. The atmosphere's short memory suggests $Y(\boldsymbol{s}, t)$ is independent from $Y\big(\boldsymbol{s}, t'\big)$ for $t \neq t'$, which is confirmed in an exploratory analysis of Colorado precipitation. Winter precipitation is important to estimate at long time scales because it strongly influences Colorado's annual water supply.

We formulate the problem of estimating precipitation as a need to estimate entire precipitation fields when only covariates are available. We build the RESP model (3.3) with historical data to estimate a statistical relationship between average monthly winter precipitation in Colorado and land and sea surface temperatures. We discuss inference for the RESP model to illustrate that it can estimate teleconnection patterns without specifying teleconnection indices a priori (Section 3.4.5). A leave-one-out cross-validation study validates the model's effectiveness (Section 3.4.5), especially in relation to other common downscaling methods (Section 3.4.3). Although beyond the scope of this study, a next step for future work would be to apply the RESP model to simulated GCM output.

3.4.1 Data

The ERA-Interim reanalysis dataset provides reconstructions of historical sea surface temperatures and local covariates (Dee et al., 2011). The response, precipitation, comes from the PRISM dataset (Daly et al., 2008). We limit our study period to 1981 through 2013 because earlier records of large scale climate are less complete. Both datasets are reanalysis products, which are necessary because working directly with observations can be challenging. Raw data may be from various sources and are often spatially sparse and temporally incomplete. Reanalysis products use statistical techniques and physical relationships to reproduce consistent datasets at regular, gridded locations with complete records after removing or correcting observations that are physically inconsistent or from stations with potential data collection issues.

This study uses data averaged over the boreal winter months (December, January, February) because Northern Hemisphere teleconnections are often strongest in winter (Nigam and Baxter, 2015). We simplify the demonstration using spatially-referenced variables average surface air temperature over Colorado (T) and average Pacific Ocean sea surface temperatures (SST) between 120°E–70°W and 20°S–60°N to predict the spatially-referenced response, average winter precipitation in Colorado (P). We standardize all data to remove the impact of orographic and other location-based effects by removing the pointwise mean from all data and scaling data to have unit variance. We additionally scale the SST values by ${n_r}^{-1}$ to ensure the remote coefficient magnitudes are independent of the resolution at which SST is measured. We standardize our data before conducting the leave-one-out cross-validation study so all of the testing and training data are comparable. Thus, our data are standardized climate anomalies that, for example, represent the number of standard deviations $P(\boldsymbol{s}, t)$ is above or below the time-averaged value $E_t[P(\boldsymbol{s}, t)]$ at location $\boldsymbol{s}$. The data are also spatially aggregated so that $n_s = 240$, 42 km-resolution grid cells cover Colorado and $n_r = 5,252$, 78 km-resolution grid cells cover the Pacific Ocean. Distances between grid cells are measured with great-circle distances. We spatially aggregate the PRISM data to increase the smoothness of the data and to make the problem computationally tractable. We discuss alternate approaches to improve computational tractability

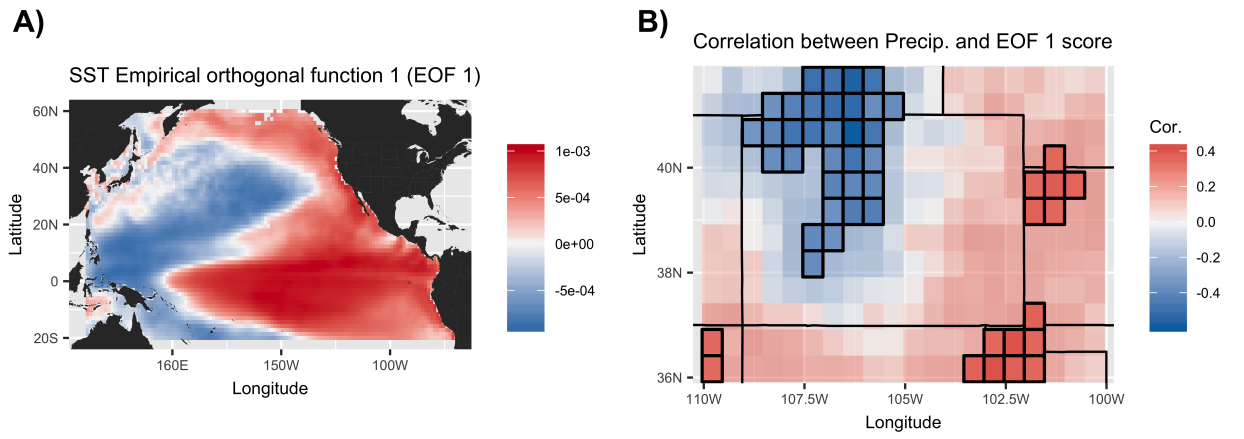

Figure 3.2: Exploratory analysis plots. A) The first empirical orthogonal function (EOF) $\psi_1 : \mathscr{D}_Y \to \mathbb{R}$ for standardized anomalies of Pacific Ocean sea surface temperatures is an indicator of El Niño events, during which sea surface temperatures are anomalously warm in the central and eastern Pacific Ocean tropics but anomalously cool in the western tropics (Ashok et al., 2007). EOF 1 accounts for 30% of the variability in sea surface temperatures. B) Pointwise correlations $\text{Cor}_t(P(\boldsymbol{s},t), a_1(t))$ between Colorado precipitation $P(\boldsymbol{s},t)$ and the EOF 1 score $a_1(t)$ suggest northern and western/central Colorado tends to receive less precipitation than average during El Niño events while eastern Colorado tends to receive more precipitation. Significant correlations (naive independent p-value $< .05$) are highlighted, while non-significant correlations are faded slightly.

in Section 3.5. The spatial aggregation and standardization also increase the normality of the data and provide a scale for precipitation with negative support, making it more appropriate for analysis with the RESP model's Gaussian likelihood.

Pacific Ocean sea surface temperature capture how the ocean influences Colorado precipitation through the El Niño–Southern Oscillation (ENSO) teleconnection (Lukas et al., 2014, Figure 2.4). The ENSO teleconnection is characterized by sea surface temperatures that are anomalously warm in the central and eastern Pacific Ocean tropics but anomalously cool in the western tropics. The first empirical orthogonal function (EOF; i.e., principal component) $\psi_1 : \mathscr{D}_Y \to \mathbb{R}$ for Pacific Ocean sea surface temperature anomalies illustrates this pattern (fig. 3.2). Pointwise correlations $\text{Cor}_t(a_1(t), P(\boldsymbol{s},t))$ between the ENSO teleconnection's strength $a_1(t)$ and Colorado precipitation $P(\boldsymbol{s},t)$ provide standard evidence for teleconnection, suggesting northern and western/central Colorado tend to receive significantly less precipitation than average during ENSO events, which are periods of strong El Niño activity, while plains re-

gions bordering eastern Colorado tend to receive significantly more precipitation than average (fig. 3.2).

3.4.2 RESP model and prior specification

In the RESP model (3.3), we specify a linear relationship between the local covariate T and response P so that $\boldsymbol{\beta}$ in (3.3) has intercept β_0 and slope β_T components $\boldsymbol{\beta} = \left(\beta_0,\ \beta_T\right)^T$. While the RESP model as described in Section 3.2.1 uses a stationary covariance model and precipitation is non-stationary in space, stationary models have comparable predictive performance in Colorado (Paciorek and Schervish, 2006). For the RESP model's remote coefficients, knots are placed at 93 locations that are roughly evenly spaced across the Pacific Ocean and along coastal locations (fig. 3.3). While knot selection can be problematic, Banerjee et al. (2008) find that reasonably dense, regularly spaced grids can yield good results. Since the ENSO teleconnection is scientifically meaningful, we will interpret the transformed teleconnection effects $\alpha'(\boldsymbol{s},1)$ from (3.11), which are associated with ENSO through its connection to the first empirical orthogonal function (EOF) of sea surface temperature anomalies $\psi_1 : \mathscr{D}_Y \to \mathbb{R}$.

We adopt a combination of weakly informative and non-informative prior distributions. A dispersed normal prior is used for the fixed effects $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 10I)$. We use $\sigma_w^2 \sim IG(2,1)$, $\sigma_\varepsilon^2 \sim IG(2,1)$, $\rho_w \sim U(1,600)$, and $\rho_\alpha \sim U(1,2000)$. The Matérn covariance smoothness parameters (3.7) are fixed at $\nu_w = \nu_\alpha = .5$, which correspond to the smoothest well-defined Matérn covariances for Gaussian processes on spheres (Gneiting, 2013). In exploratory analysis, variograms for the local and remote data fit this parameterization well. The prior for σ_α^2 is informative to increase the identifiability of this parameter and the remote range ρ_α (Zhang, 2004). The prior $\sigma_\alpha^2 \sim IG(6,10)$ keeps the model from exploring parameter combinations that would imply very large teleconnection influence relative to the scale of the data $Y(\boldsymbol{s},t)$.

3.4.3 Comparison models

We demonstrate the benefit of remote covariates by comparing the RESP model to RE and SP submodels that, respectively, exclude local and remote covariates. We also show improvement

to statistical downscaling and prediction by comparing RESP model validation scores to spatially varying coefficient (SVC) models (3.1) and other common downscaling models, including a hybrid local and non-local regression using the El-Niño–Southern Oscillation teleconnection (ENSO-T) (van den Dool, 2007, Sections 8.4, 8.5), canonical correlation analysis (CCA) (von Storch and Zwiers, 1999, Chapter 14), and a baseline climatological reference prediction (CLIM) (van den Dool, 2007, Section 8.1).

While analog models provide an alternate means to model teleconnected processes, we do not make comparisons to them in this application because analog models require more temporal replication than our data provide. Analog models require considerable temporal replication because predictions are weighted combinations of past observations, where the weights are based on distances between covariates at the prediction timepoint and all past observations (McDermott and Wikle, 2016). An advantage of analog forecasts, for example, is that the reweighting scheme naturally generates forecasts that have the same spatial patterns as past observations. Without enough past observations, however, the likelihood increases that past observations are not diverse enough to sufficiently approximate future states (Van Den Dool, 1994).

Spatially varying coefficient model (SVC)

To facilitate comparison, the SVC model (3.1) is specified with the same linear relationship between the local covariate T and response P we use with the RESP model. The scores $a_1(t)$ and $a_2(t)$ for the first and second sea surface temperature (SST) anomaly EOFs ψ_1, ψ_2 capture ENSO and ENSO-Modoki teleconnection relationships for Colorado precipitation with bivariate spatially varying coefficients $\boldsymbol{\theta}(\boldsymbol{s}) \in \mathbb{R}^2$. The scores $\{a_i(t) : i = 1,2,\ t \in \mathcal{T}\}$ quantify the strength of ENSO activity and are similar to other measures of ENSO activity (Ashok et al., 2007). The first and second EOFs ψ_1 and ψ_2 respectively account for 30% and 15% of the variability in SST.

We adopt a hierarchical Bayesian framework to estimate the SVC model (Banerjee et al., 2015, Section 9.6.2). An Inverse-Wishart prior $\Lambda \sim IW(I,2)$ is used for $\Lambda = \text{Cov}(\boldsymbol{\theta}(\boldsymbol{s}))$ and a dispersed normal prior is used for the fixed effects $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 10I)$. We use $\sigma^2 \sim IG(2,1)$ and

$\rho \sim U(1,600)$ for the prior distribution of the Matérn covariance with fixed smoothness $\nu = .5$ for the model's spatial correlation.

Hybrid local and non-local regression (ENSO-T)

Pointwise regression models are commonly used to downscale climate data (e.g., Towler et al., 2016). The ENSO-T model predicts precipitation $P(\boldsymbol{s}, t_0)$ at a location $\boldsymbol{s}$ and new time point t_0 by applying a regression of training data $P(\boldsymbol{s}, t)$ onto local surface air temperature $T(\boldsymbol{s}, t)$ and the score $a_1(t)$ for the first sea surface temperature EOF $\psi_1 : \mathscr{D}_Y \to \mathbb{R}$. The ENSO-T downscaler provides a comparison model that accounts for both local and remote effects, but not spatial dependence.

Canonical correlation analysis (CCA)

Canonical correlation analysis uses the empirical correlation structure of sea surface temperature SST and precipitation P vectors to linearly map these variables to a space in which the transformed vectors are maximally correlated (von Storch and Zwiers, 1999, Chapter 14). This mapping may be used in a multivariate regression context with sea surface temperatures at new time points to predict precipitation. The mapping is often developed with some amount of smoothing by removing higher order EOFs from the data. We retain 16 EOFs in our use of CCA because this lets us capture approximately 90% of the variability in the predictors SST and predictand P. The CCA downscaler provides a comparison model that only accounts for remote effects and indirectly accounts for spatial dependence.

Climatological reference (CLIM)

Climatologists use the unconditional distribution of precipitation $P(\boldsymbol{s}, t)$ at a location $\boldsymbol{s}$. When no other information is available, the average value of precipitation $\mathrm{E}_t[P(\boldsymbol{s}, t)]$ is used as a climatological point prediction for precipitation, and the empirical distribution is used for probabilistic predictions. The CLIM downscaler provides a baseline comparison model that does not account for spatial dependence, local, or remote effects.

3.4.4 Implementation of model assessment measures

Variance inflation factors for local effects

The posterior covariance matrix (3.18) for the regression coefficient vector $\boldsymbol{\beta}$ allows us to follow Reich et al. (2006) and define conditional variance inflation factors that can help diagnose multicollinearity between the local and remote covariate matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ via

$$\text{VIF}(\beta_i) = \frac{\left(\left\{\boldsymbol{\Lambda}^{-1} + \boldsymbol{X}^T\left(C \otimes \Sigma^{-1}\right)\boldsymbol{X}\right\}^{-1}\right)_{ii}}{\left(\left\{\boldsymbol{\Lambda}^{-1} + \boldsymbol{X}^T\left(I_{n_t} \otimes \Sigma^{-1}\right)\boldsymbol{X}\right\}^{-1}\right)_{ii}}.$$

The VIF measures the proportional increase in the i^{th} local coefficient's posterior variance caused by adding remote covariates to the model, conditional on the model's covariance parameters. This interpretation follows since the denominator represents the i^{th} local coefficient's posterior covariance in a standard spatial regression model where the local covariates and responses are observed at multiple, independent timepoints. Larger VIF values indicate greater multicollinearity, while the smallest possible VIF value of 1 indicates no multicollinearity. The VIF for β_T is 1.1, which indicates that estimates of the local effects are not impacted by the addition of remote covariates in the case study of Colorado winter precipitation (Section 3.4).

Variance inflation factors for remote effects

As with the local effects, we can use the posterior covariance matrix in (3.19) for the remote effects vector $\tilde{\boldsymbol{\alpha}}^*$ to define conditional variance inflation factors that can help diagnose multicollinearity in teleconnection effects $\{\alpha(\boldsymbol{s}, \boldsymbol{r}_i^*) : \boldsymbol{s} \in \mathscr{D}_Y\}$ associated with the i^{th} knot location $\boldsymbol{r}_i^*$ denoted by

$$\text{VIF}(\boldsymbol{r}_i^*) = \frac{\left(\left(R^{*-1} + \boldsymbol{Z}^*\boldsymbol{Z}^{*T}\right)^{-1}\right)_{ii}}{\left(1/\sigma_\alpha^2 + \boldsymbol{Z}^*_{i,\cdot}\boldsymbol{Z}^{*T}_{i,\cdot}\right)^{-1}}.$$

The notation $\boldsymbol{Z}^*_{i,\cdot}$ indicates the i^{th} row of the matrix $\boldsymbol{Z}^*$, in which row i contains all observations of the remote covariate at location $\boldsymbol{r}_i^*$, should be used in computations. Conditional on

the model's covariance parameters, the VIF measures the proportional increase in the marginal posterior variance of teleconnection effects $\{\alpha(\boldsymbol{s}, \boldsymbol{r}_i^*) : \boldsymbol{s} \in \mathscr{D}_Y\}$ associated with the i^{th} knot location $\boldsymbol{r}_i^*$ that results from adding the remaining remote covariates $\{z(\boldsymbol{r}_j, t) : j \neq i,\ t \in \mathscr{T}\}$ to the model. In the case study of Colorado winter precipitation (Section 3.4), $\text{VIF}(\boldsymbol{r}_i^*)$ ranged between 1.0 and 1.1 for all knot locations $\boldsymbol{r}_i^*$. This indicates the reduced rank approximation and choice of knot locations mitigates potential multicollinearity in estimation of teleconnection effects.

Heidke skill score

The Heidke skill score (HS) evaluates categorical predictions and is commonly used in climate science (von Storch and Zwiers, 1999, Section 18.1). The measure compares the probability the RESP model correctly predicts precipitation p_{RESP} to the probability that a reference model correctly predicts precipitation p_{Ref}. We adopt a standard, naive reference model that assigns equal probability to all precipitation levels, implying $p_{\text{Ref}} = 1/3$ since our precipitation levels represent empirical terciles. We additionally manipulate the Heidke skill score formula (von Storch and Zwiers, 1999, eq. 18.1) to show that it is linear in p_{RESP}. The manipulation also yields an intuitive interpretation of the score:

$$\text{HS}(\text{RESP}) = \frac{p_{\text{RESP}} - p_{\text{Ref}}}{1 - p_{\text{Ref}}} = \left(p_{\text{RESP}}/p_{\text{Ref}} - 1\right)\text{Odds}\left(p_{\text{Ref}}\right).$$

The Heidke skill score scales the odds that the reference model correctly predicts precipitation by the RESP model's relative change in prediction accuracy. Models have positive Heidke skill when they are more accurate than the reference model; the maximum possible score is 1. Similarly, models have negative skill when they are less accurate than the reference model.

The predictive distributions for the RESP, ENSO-T, and CLIM models are continuous, but easily discretized with respect to the category cutpoints that are empirically determined from the leave-one-out training data. Therefore, the posterior mode of these distributions provides a natural choice for categorical point predictions of precipitation. The CCA model only produces continuous point predictions, so its categorical predictions are defined by the category that

matches the "above average", "near average", or "below average" range in which the continuous point prediction lies.

The RESP model (3.3) frequently yields better point predictions than the comparison models (fig. 3.4). The CCA Heidke skill scores are lower, but about as variable as the RESP model, which suggests the CCA model may adequately account for spatial dependence when making predictions, but loses skill by not also incorporating local covariates. Similarly, the ENSO-T model's Heidke skill scores may be more variable than the RESP model's scores since it does not account for spatial dependence.

Ranked probability score

The ranked probability score (RPS) is closely related to the continuous ranked probability score (CRPS), which is a proper scoring rule for probability measures with finite means (Gneiting and Raftery, 2007). Epstein (1969) introduced the RPS to evaluate probabilistic predictions of ordinal variables. Murphy (1971) presented an equivalent formulation of the RPS that shows how the RPS, like the CRPS, sums the squared differences between the predicted and observed cumulative distribution functions for an ordinal response $Y_o(\boldsymbol{s}, t)$ at location $\boldsymbol{s}$ and time t. Since the RPS is defined pointwise, we follow common practice in climatological applications and average RPS scores over n_s locations at which we observe the response $Y_o(\boldsymbol{s}, t)$ (Hersbach, 2000), yielding the RPS score at time t for a model $\mathscr{M}$ that predicts the cumulative distribution $\hat{F}_t\big(j; \boldsymbol{s}\big)$ for $j \in \{1, \ldots, k\}$ ordered categories of the response at location $\boldsymbol{s}$

$$\text{RPS}(\mathscr{M}, t) = \frac{1}{n_s} \sum_{i=1}^{n_s} \sum_{j=1}^{k} \left(\hat{F}_t\big(j; \boldsymbol{s}_i\big) - \mathbb{1}\left\{Y_o(\boldsymbol{s}_i, t) \leq j\right\}\right)^2.$$

3.4.5 Results

Model results are based on 20,000 samples from the posterior distribution after a burn in period of 1,000 samples. Convergence was assessed by examining trace plots, autocorrelation plots, and effective sample sizes in addition to comparing results from multiple runs with randomly initialized parameters. Model adequacy was assessed using residual and qq-normal

plots. These diagnostics suggest there are no serious violations of the convergence and distributional assumptions. Variance inflation factors (VIFs) that account for the RESP model design also show no concern for multicollinearity in the fitted model Section 3.4.4.

Inference

The parameter estimates for the RESP model yield reasonable scientific interpretations (table 3.1). The sign of the regression coefficient β_T for the temperature covariate T is consistent with physical processes that influence precipitation (Daly et al., 2008). The local covariance range parameter ρ_w implies the dependence between locations $\boldsymbol{s} \in \mathscr{D}_Y$ has an effective range between 500 and 570 km, which is the distance between locations beyond which the Matérn correlation (3.7) is small ($\leq .05$). This length scale is in the size range of mesoscale weather processes that produce precipitation (Parker, 2015). The remote covariance range parameter ρ_α implies the dependence between locations $\boldsymbol{r} \in \mathscr{D}_Z$ has an effective range between 720 and 2,200 km, which is roughly the size of the mid-sized structures seen in the EOF patterns in fig. 3.2 A. Since local temperature T is teleconnected with sea surface temperatures SST, remote effects must be interpreted as residual teleconnection effects, as described at the end of Section 3.2.4. Significant remote effects suggest Colorado's teleconnection with the Pacific Ocean cannot be represented through a linear relationship with temperature alone; the teleconnection likely involves non-linear relationships and additional variables or interactions. Posterior estimates for the transformed remote effects $\{\alpha'(\boldsymbol{s}, 1) : \boldsymbol{s} \in \mathscr{D}_Y\}$ associated with $\psi_1 : \mathscr{D}_Y \to \mathbb{R}$ (fig. 3.5) largely match the exploratory pointwise correlations between $P(\boldsymbol{s}, t)$ and $a_1(t)$ found in the exploratory plot (fig. 3.2), indicating the RESP model (3.3) is capturing known Colorado teleconnections. Fewer locations have significant teleconnection, however, as the estimates incorporate more uncertainty due to spatial correlation; significance is determined with respect to evaluating highest posterior density intervals, separately for each location $\boldsymbol{s} \in \mathscr{D}_Y$.

Table 3.1: Posterior mean estimates and 95% highest posterior density (HPD) intervals for the RESP model's parameters, which include an intercept β_0 and temperature effect β_T on the mean response (see equation (3.3)), and covariance scale σ^2 and range ρ parameters for the local w and remote α spatial dependence and nugget effect ε (see (3.5) and (3.6)). The smoothness parameters ν_w and ν_α were fixed (Section 3.4.2).

		Posterior mean	95% HPD
Local effects	β_0	−0.00	(−0.14, 0.14)
	β_T	−0.18	(−0.24, −0.12)
Covariance	σ^2_w	0.55	(0.49, 0.62)
	σ^2_α	6.05	(1.04, 14.81)
	σ^2_ε	0.01	(0.01, 0.01)
	ρ_w	248.00	(220, 280)
	ρ_α	509.00	(266, 799)

Model validation

Leave-one-out cross-validation scores demonstrate the RESP model benefits from including remote covariates and offers improvement over comparison models in the intended prediction-like setting of perfect prognosis downscaling (fig. 3.6). The RESP and comparison models are trained on all but one year of available data, then used to predict the responses $\{P(\boldsymbol{s}, t) : \boldsymbol{s} \in \mathscr{D}_Y\}$ for the test year t to mimic the perfect prognosis downscaling setting in which a climate variable must be completely inferred from covariate data only. The process is repeated with all years of available data. While the RESP and comparison models yield continuous predictive distributions, we discretize the distributions before assessing them. Climate forecasts are often discretized because it is inherently difficult to develop more precise climate predictions at seasonal and longer time scales (Mason, 2012; van den Dool, 2007, Section 9.6). We use the empirical terciles $\hat{q}(1/3; P(\boldsymbol{s}, \cdot))$ and $\hat{q}(2/3; P(\boldsymbol{s}, \cdot))$ to discretize the predictive distribution $f(P(\boldsymbol{s}, t_0) | \boldsymbol{P})$ at each location $\boldsymbol{s} \in \mathscr{D}_Y$ into "below average", "near average", and "above average" categories. While it is possible to directly fit discrete models to the data, doing so is not necessarily helpful. For example, a probit-link RESP or SVC model would require re-estimation of observed continuous data $P(\boldsymbol{s}, t)$ as latent fields (Higgs and Hoeting, 2010).

We use ranked probability scores (RPS) to assess probabilistic forecasts for ordinal variables, giving lower scores to models that generate predictive distributions that better match the true distribution (Gneiting and Raftery, 2007). The CCA model only yields point predictions since predictive uncertainties are difficult to obtain. Thus, the CCA's validation scores are inflated since its discretized predictive distribution is defined by a point mass on the category that matches the tercile in which the point prediction lies.

The RESP model (3.3) frequently yields better probabilistic predictions than the comparison models. In particular, the RESP model performs better than the RE or SP submodels which highlights the advantage of combining local and remote information. Sample maps of predictions and uncertainties are presented in Section 3.4.5. The RESP model also tends to perform better than the SVC model which highlights the advantage of not specifying teleconnection indices a priori and adding additional spatial structure to estimates of teleconnection effects. Similar results are obtained using Heidke skill scores to compare models. Heidke skill scores are commonly used in climate science to measure a model's misclassification rate for categorical point predictions (von Storch and Zwiers, 1999, Section 18.1). Formulas and details for RPS and Heidke skill scores can be found in Section 3.4.4.

Model validation maps

Section 3.4.5 builds support for the RESP model by comparing it to submodels and alternatives (fig. 3.6). fig. 3.7 and fig. 3.8 show continuous and discretized (categorical) predictions and uncertainties for the 1982 validation set. Shrinkage and uncertainty in the continuous predictions can be anticipated because even though teleconnection effects contain predictive information, their overall influence on precipitation tends to be relatively weak (fig. 3.2 B).

The categorical predictions for average monthly precipitation in winter across Colorado are better determined near teleconnected regions. Posterior logits (fig. 3.8 C) for the categorical forecasts (fig. 3.8 B) quantify uncertainty on an interpretable scale in this application. Since the modes of the discretized posterior predictive distributions are the categorical forecasts, odds compare the the forecasted category probabilities to the other categories. The logit (log-odds)

allows zero to be a natural reference point for comparing uncertainties. Since we discretize posterior predictive distributions into three categories, the probability for each categorical forecast is at least 1/3. Non-negative logits indicate at least 50% greater certainty since their categorical forecast probability is at least 1/2. Regions with non-negative logits occur near locations with significant teleconnection effects (fig. 3.5).

3.5 Discussion

The RESP model (3.3) expands geostatistical frameworks that incorporate the effect of both local and remote covariates on spatially correlated responses, like precipitation, but can be extended to address additional spatio-temporal modeling needs. For example, while we use the RESP model to draw inference on entire response fields, the model's process-formulation also allows it to be applied to more standard spatial interpolation problems as well. Since there is great uncertainty in global climate model (GCM) predictions of future precipitation, statistical downscaling methods have been widely used in regional climate change studies. Validating the RESP model on historical data marks an improvement on existing approaches and implies it can be used with GCM predictions of surface temperatures and large-scale patterns to infer predictions for precipitation from covariate data only. By comparison with the RESP model, other models directly model less of the spatial structure in teleconnected data, but other models have been studied in broader statistical contexts. Fortunately, it is possible to formulate the RESP model more broadly.

Many scientific disciplines work with spatially-referenced non-Gaussian data, for which the RESP model can be adapted. For example, the RESP model could be adapted to study teleconnective effects on the number of large rain events, which are important for many ecological systems and sectors of society. Following approaches common to generalized linear models for spatial data, the existing RESP response $Y(\boldsymbol{s}, t)$ may be reinterpreted as a latent Gaussian field that helps parameterize the distribution for non-Gaussian observations (Diggle et al., 1998;

Higgs and Hoeting, 2010). The primary technical challenge for Bayesian implementations of such models is to develop efficient estimation procedures since conjugacy is lost.

Modeling effects for multivariate remote covariates or data on large spatial domains could both be facilitated by modeling spatial dependence with sparse geostatistical models. Inference and prediction for many geostatistical models involves matrix operations with $O(n_s^3)$ computational complexity. Sparse geostatistical models can avoid these costs on large spatial domains, for example, by using Gaussian Markov random field approximations to specific classes of Gaussian fields with Matérn covariances (Lindgren et al., 2011), covariance tapering to generate spatial covariance matrices with banded structure (Furrer et al., 2006), multiresolution covariance models (Katzfuss, 2016), or hierarchical nearest neighbor models (Datta et al., 2016). While computational savings may be minimal for small spatial domains like Colorado, they may offset computational costs of estimating teleconnection effects for multiple sets of remote covariates. The RESP model may naturally be extended to include multiple teleconnection effects (3.4) to model impacts from Pacific and Atlantic Ocean temperatures, for example. Multivariate teleconnection effects can also be used to model impacts from a vector $\boldsymbol{z}(\boldsymbol{r}, t) \in \mathbb{R}^m$ of m remote covariates at location $\boldsymbol{r} \in \mathscr{D}_Z$. Both extensions require sensibly modifying the remote coefficient covariance function (3.6) and will yield likelihood structures similar to the RESP model (3.3), especially if relationships between additional teleconnection effects are modeled with separable covariances.

Non-stationary covariance models and temporal extensions can also allow the RESP model to be applied to more diverse data and problems. While the teleconnection term (3.4) admits temporal non-stationarity moderated by the remote covariates, modeling temporal dependence across timepoints can allow the RESP model to be used in more traditional forecasting problems. Similarly, modeling spatial non-stationarity can potentially improve model fit and prediction at unobserved spatial locations. In particular, nonstationary covariances could allow the remote coefficients to vary temporally. This extension may be relevant because Mason and Goddard (2001) find that teleconnection effects can vary across seasons. As in Choi

et al. (2015), however, changes over time may be difficult to detect because the effects tend to be weak.

Without considering any extensions, however, the RESP model yields additional discussion about spatial modeling. The RESP model's inclusion of dependence at both long and short distances echoes descriptions of the screening effect (Stein, 2015). Carefully studying spectral densities of covariance functions show that if they decay quickly enough, then spatial predictions are primarily driven by data from nearby locations. While the RESP model allows distant locations to influence spatial prediction, the RESP model does not contradict the screening effect because it explicitly models long range dependence through the teleconnection term (3.4) and the screening effect is a property of local covariance functions (3.5). Of similar subtlety, maps of estimated teleconnection effects (fig. 3.5) raise discussion about uncertainty estimates for spatial patterns. Significance in fig. 3.5 is determined pointwise with respect to the posterior distribution for $\alpha'(\boldsymbol{s}, 1)$ at each location so can provide inference for teleconnection effects at individual points. Here, pointwise significance can help individual municipalities determine whether they are strongly impacted by teleconnection effects and may benefit from use of the RESP model. Determining uncertainty for entire regions is a multiple testing problem not considered in this study (Bolin and Lindgren, 2015; French and Hoeting, 2016). Uncertainties for entire regions are more important, for example, when trying to estimate boundaries for polluted areas.

There is potential for more diverse application of the RESP model because teleconnections exist in other fields, like ecology (Brierley et al., 1999) and human geography (Seto et al., 2012). The model's general introduction in Section 3.2 as a spatial regression problem highlights a less-common class of spatial analysis problems because it addresses problems that require dependence at both long and short distances, at odds with typical assumptions that data at distant points are effectively independent. While the RESP model assumes the response and remote covariates are defined on disjoint spatial domains, it suggests even broader classes of problems in which overlapping domains characterize dependence between distant locations, or in which

teleconnected domains are not known a priori and need to be estimated. The latter problem is reminiscent of general covariance or graphical model structure estimation problems, which may provide possible directions for future spatial statistics research topics.

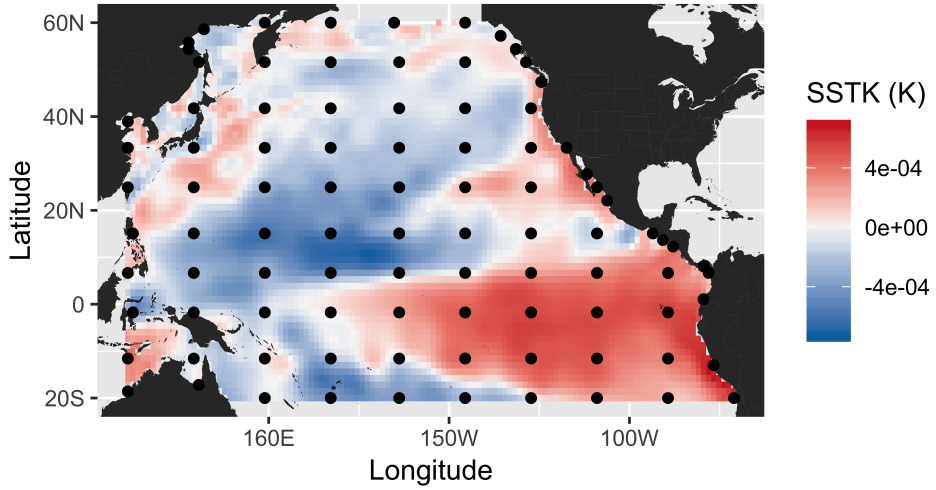

Figure 3.3: Average monthly sea surface temperature standardized anomalies from Winter, 1982. Black dots mark knot locations.

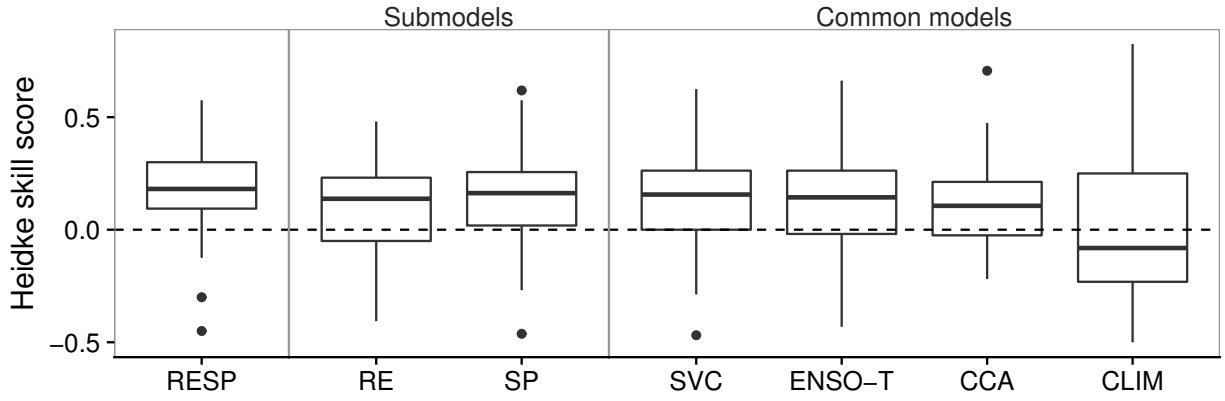

Figure 3.4: Comparison of Heidke skill scores for categorical point predictions on the leave-one-out test datasets for the RESP, ENSO-T, CCA, and CLIM models. The dashed line at 0 marks the Heidke skill for a naive reference model that produces random point predictions The RESP model generally has better (i.e., higher) and less variable skill than the comparison models.

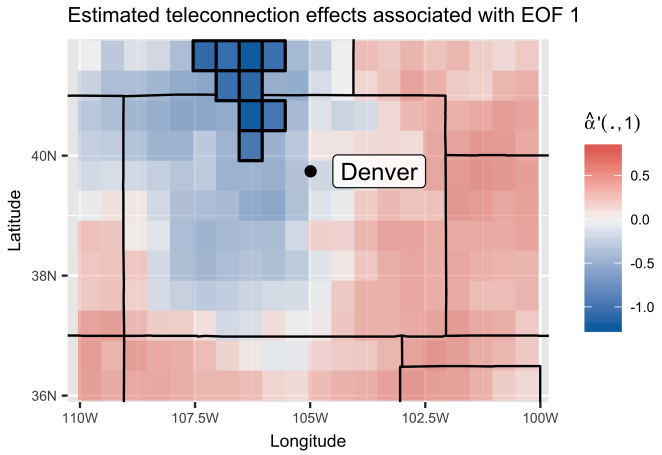

Figure 3.5: Estimated teleconnection effects $\hat{\alpha}'(\boldsymbol{s}, 1)$ for EOF 1 $\psi_1 : \mathscr{D}_Y \to \mathbb{R}$. The overall patterns yield similar interpretations as those made with the fig. 3.2 exploratory plots, however, the RESP model reduces the regions in which evidence exists for significant teleconnection. Significant teleconnection effects, as determined using 95% highest posterior density intervals, are highlighted.

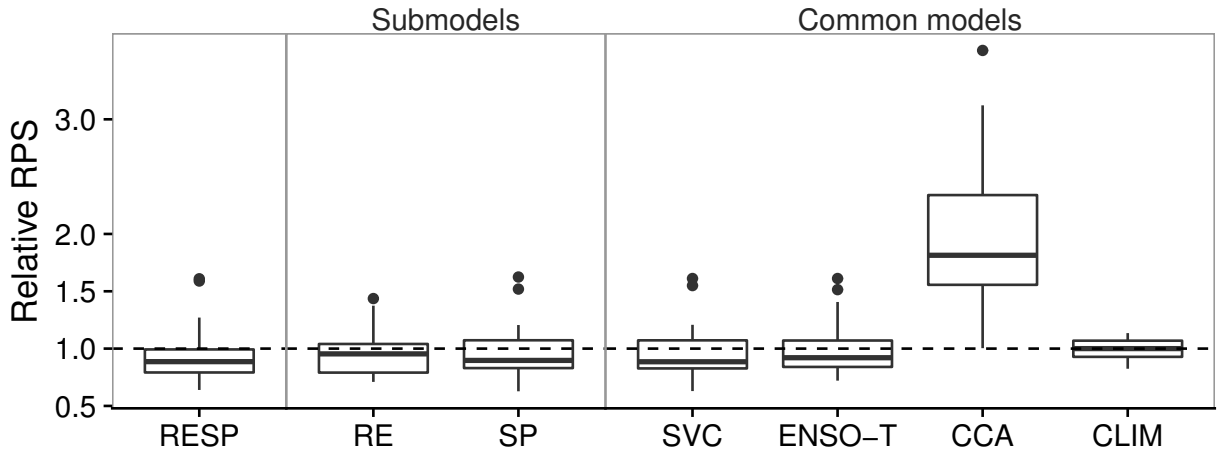

Figure 3.6: Comparison of Ranked probability scores (RPS) for probabilistic categorical predictions on the leave-one-out test datasets for the RESP and comparison models. RPS scores are reported relative to the median RPS for the CLIM reference model's unconditional predictions. The RESP model generally has better (i.e., lower) and slightly less variable skill than the "Sub" and "Common" comparison models.

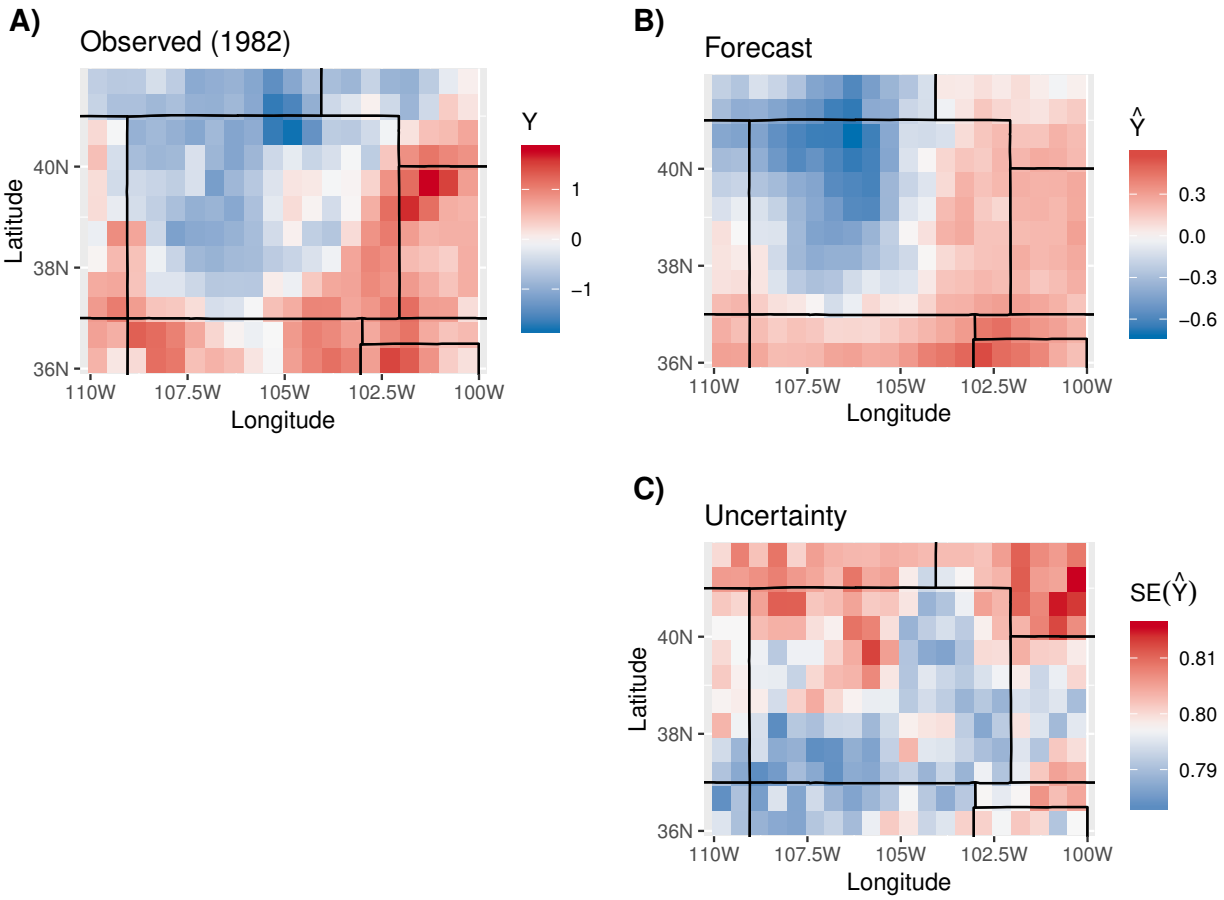

Figure 3.7: Comparison of predictions for the 1982 validation set in the leave-one-out cross-validation study. The pattern of the posterior predictive means (B) matches the PRISM responses (A) well, but the color scale indicates shrinkage of the forecasted magnitudes. Posterior predictive standard errors (C) indicate uncertainty for the forecast.

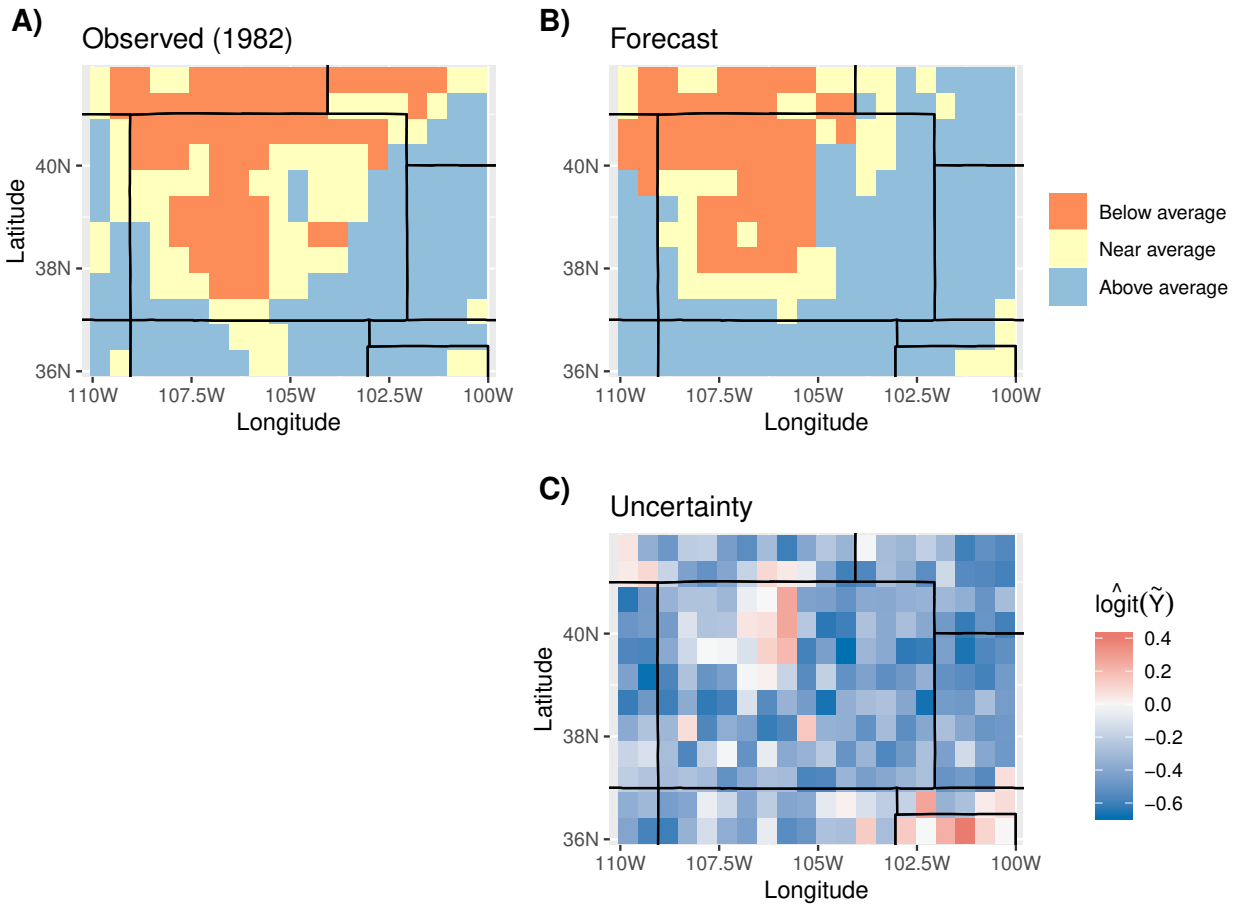

Figure 3.8: Comparison of discretized (categorical) predictions for the 1982 validation set in the leave-one-out cross-validation study. The pattern of the posterior predictive modes (B) matches the PRISM responses (A) well. Logits for the posterior predictive modes (C) indicate higher uncertainty for the forecasts (blue), especially in regions without significant teleconnection effects.

Chapter 4

Approximate Bayesian Inference via Sparse grid Quadrature Evaluation for hierarchical models[3]

4.1 Introduction

Computationally efficient posterior approximation remains a key challenge and concern in applied Bayesian analyses, especially for hierarchical models. Hierarchical Bayesian models allow flexible modeling of complex data, but make posterior inference challenging because simple, conjugate distributions are typically unavailable. Posterior densities, expectations, and other quantities involve computing integrals that often require numerical approximation. The required approximations can be computationally expensive or challenging since many hierarchical models include many unknown parameters, thus integrals are defined over high dimensional state spaces. Sampling-based approaches, like Markov chain Monte Carlo (MCMC) methods, are widely used because they are generally reliable and relatively simple to implement (Gelfand and Smith, 1990). However, MCMC approximations can be computationally expensive for many models as many full conditional posterior distributions have high correlation or are difficult to sample. As a result, if n dependent samples are drawn via MCMC methods, the stochastic approximation error rate can often be higher than the error $\mathcal{O}_p\left(n^{-1/2}\right)$ for direct Monte Carlo approximations, which are often infeasible since many full posterior distributions cannot be sampled directly. Alternate approximation is available via a range of stochastic and deterministic methods, including Laplace and Integrated Nested Laplace approximations (Rue et al., 2009; Tierney and Kadane, 1986), classical quadrature-based approximations (Naylor and Smith, 1982), Variational Bayes (Attias, 2000), and Approximate Bayesian Computing (Rubin, 1984; Tavare et al., 1997) to just name a few. Generally, each method is motivated by com-

[3] In preparation for submission with J. A. Hoeting.

putational issues and structures found in different classes of models, so no method is necessarily well-suited for all hierarchical models. In particular, technical limitations of Integrated Nested Laplace approximations (INLA) and classical quadrature motivate us to develop a strategy to yield approximate Bayesian Inference via Sparse grid Quadrature Evaluation (BISQuE) for a wider range of hierarchical models.

INLA approximates marginal posterior distributions by using a discrete numerical integration grid of hyperparameters to average over Laplace approximations of conditional posterior densities. The method is developed for models that link observations to latent Gaussian variables through link functions, similar to generalized linear models. The approximation enables fast inference for a wide range of scientifically relevant models. However, it can sometimes be difficult to reparameterize, reformulate, or otherwise embed models without latent Gaussian structures to the INLA framework. Additionally, the numerical integration can become computationally infeasible for models with too many hyperparameters. The latter issue is a limitation shared by classical quadrature-based approximations for posterior quantities.

Classical quadrature methods can approximate marginal posterior distributions and expectations for general Bayesian models, but like INLA, the models must have relatively small dimension (Naylor and Smith, 1982). Quadrature methods approximate an integral by evaluating its integrand at deterministic locations, then weighting the results. Locations and weights are chosen using known information about the shape of the integrand. However, classical quadrature methods have limited practical use for approximate Bayesian inference. Classical methods integrate over all unknown parameters—not just hyperparameters—and the size of the integration grids suffer from the curse of dimensionality, growing exponentially as parameters are added to models.

More recent quadrature literature formalized theory and methods that yield sparse integration grids, thereby mitigating the curse of dimensionality for quadrature approximations of high dimensional integrals (Gerstner and Griebel, 1998; Novak and Ritter, 1996,9). In statistics, sparse grid quadrature methods have been used to approximate likelihoods that involve

high dimensional integrals, as can arise from econometric models (Heiss and Winschel, 2008). Sparse grid quadrature has also been used to approximate posterior expectations, densities, and integration constants for non-linear inverse problems with normal errors (Emery and Johnson, 2012; Schillings and Schwab, 2013), estimate Kullback-Leibler information gains to solve Bayesian experimental design problems (Long et al., 2013), and to accelerate computations for specific non-linear Kalman filters (Arasaratnam and Haykin, 2009; Jia et al., 2012). By comparison, we consider approximate Bayesian posterior inference more generally.

We propose reformulating Bayesian posterior quantities, such as densities and expectations, so that they can be efficiently approximated by combining conditioning techniques with sparse grid quadrature methods. Our reformulation lets us apply sparse grid quadrature methods to hierarchical Bayesian models with non-Gaussian structures and potentially many hyperparameters. The resulting computational approach greatly reduces computation time as compared to MCMC approaches for many models, including fully non-Gaussian models. Our framework can also potentially be combined with INLA to allow fast inference for latent Gaussian models with many hyperparameters.

We briefly review quadrature and sparse grid methods (Section 4.2), then introduce the Bayesian Inference via Sparse grid Quadrature Evaluation (BISQuE) strategy to yield approximate inference for hierarchical Bayesian models (Section 4.3). Our method reduces the computational effort required to approximate posterior densities, means, and variances in examples where traditional MCMC methods are relatively slow (Section 4.5). We conclude with discussions of extensions and other directions for future work (Section 4.6).

4.2 Quadrature and Sparse grid methods

Let $f(\boldsymbol{x})$ be a map from a d-dimensional space $\mathscr{S}$ onto the real line $\mathbb{R}$, and $w(\boldsymbol{x})$ be a weight function with the same support. The integral

$$I(f) = \int_{\mathscr{S}} f(\boldsymbol{x})\, w(\boldsymbol{x})\, d\boldsymbol{x} \tag{4.1}$$

may be approximated via the weighted sum

$$\hat{I}(f) = \sum_{\ell=1}^{k_i} f\left(\boldsymbol{x}^{(i,\ell)}\right) w^{(i,\ell)} \tag{4.2}$$

for some choice of summation length $k_i \in \mathbb{N}$, nodes $\mathscr{A}^i = \{\boldsymbol{x}^{(i,\ell)} : \ell = 1,\ldots,k_i\} \subset \mathscr{S}$, and weights $\mathscr{W}^i = \{w^{(i,\ell)} : \ell = 1,\ldots,k_i\} \subset \mathbb{R}^{k_i}$. We will use the index i shortly. The approximation (4.2) is called a quadrature rule if the integration domain $\mathscr{S}$, weight function w, and desired approximation accuracy or computational cost are used with specific procedures to specify k_i, $\mathscr{A}^i$, and $\mathscr{W}^i$ (Givens and Hoeting, 2013, Section 5.3). The number of nodes and weights k_i balances the approximation error in (4.2) with the approximation's computational cost. Large k_i can yield more accurate approximation (or even exact evaluation) of (4.1), but at potentially high computational cost. In practice, sequences of increasingly accurate quadrature rules defined by $(k_1, \mathscr{A}^1, \mathscr{W}^1)$, $(k_2, \mathscr{A}^2, \mathscr{W}^2)$, … such that $k_1 < k_2 < \ldots$ can be used to estimate and control approximation error (Laurie, 1985). Quadrature rules can yield highly accurate approximations for integrals $I(f)$ of smooth functions f defined on $\mathscr{S}$, but computational efficiency is difficult to achieve if $\mathscr{S}$ has high dimension.

The simplest quadrature rules to construct for multidimensional $\mathscr{S}$ are product rules, but these suffer from the curse of dimensionality. Product rules are formed by iteratively applying univariate quadrature rules along each dimension of $\mathscr{S}$ to approximate (4.1); they are aptly named because their nodes $\mathscr{A}^i$ are a Cartesian product of nodes from the underlying univariate quadrature rules (cf. Novak and Ritter, 1996). To be precise, let $\mathscr{S}$ be the product space $\mathscr{S} = \mathscr{S}_1 \times \cdots \mathscr{S}_d$ of one-dimensional, σ-finite measure spaces $\mathscr{S}_1,\ldots,\mathscr{S}_d$, and let the weight function $w(\boldsymbol{x})$ be the product $w(\boldsymbol{x}) = \prod_{i=1}^d w_i(x_i)$ of weight functions $w_1(x_1),\ldots,w_d(x_d)$ that are respectively defined on $\mathscr{S}_1,\ldots,\mathscr{S}_d$. If the target integral (4.1) is well defined, then Fubini's theorem implies (4.1) may be evaluated as an iterated integral. Iterated integration allows approximation by applying univariate quadrature rules along each dimension of $\mathscr{S}$. Define $U_1^{i_1},\ldots,U_d^{i_d}$ to be univariate quadrature rules that respectively approximate integrals on $\mathscr{S}_1,\ldots,\mathscr{S}_d$ with $k_{i_1},\ldots,k_{i_d}$ nodes $\mathscr{A}_1^{i_1},\ldots,\mathscr{A}_d^{i_d}$ and weights $\mathscr{W}_1^{i_1},\ldots,\mathscr{W}_d^{i_d}$. The product rule that approximates

(4.1) is defined via

$$\left(U_1^{i_1} \otimes \cdots \otimes U_d^{i_d}\right)(f) = \sum_{\ell_1=1}^{k_{i_1}} \cdots \sum_{\ell_d=1}^{k_{i_d}} f\left(x_1^{(i_1,\ell_1)}, \ldots, x_d^{(i_d,\ell_d)}\right) w_1^{(i_1,\ell_1)} \ldots w_d^{(i_d,\ell_d)}. \tag{4.3}$$

Note that the product rule (4.3) is a special case of the general approximation form (4.2). The product rule (4.3) requires evaluation of f at $\left|\mathscr{A}_1^{i_1} \times \cdots \times \mathscr{A}_d^{i_d}\right| = k_{i_1} \cdots k_{i_d}$ nodes. The number of quadrature nodes grows exponentially as $d \uparrow \infty$ if f is explored equally in all dimensions, i.e., if $k_{i_1} = \cdots = k_{i_d}$. The curse of dimensionality for product rules can be partially mitigated by exploring f unequally in different dimensions, but this approach is only practical if f is extremely smooth in some dimensions.

By comparison, sparse grid quadrature rules are computationally efficient approximations for integrals on multidimensional $\mathscr{S}$. Novak and Ritter (1996,9) use the Smolyak (1963) formula to combine univariate quadrature rules $U_1^{i_1}, \ldots, U_d^{i_d}$ in a computationally efficient approximation (4.2) of (4.1). The Smolyak formula specifies a linear combination $A(q,d)$ of product rules (4.3) that approximates (4.1) via

$$A(q,d)(f) = \sum_{q-d+1 \le |\boldsymbol{i}| \le q} (-1)^{q-|\boldsymbol{i}|} \binom{d-1}{q-|\boldsymbol{i}|} \left(U_1^{i_1} \otimes \cdots \otimes U_d^{i_d}\right)(f), \tag{4.4}$$

in which $q \ge d$ and $|\boldsymbol{i}| = i_1 + \cdots + i_d$. Note that the Smolyak rule (4.4) is also a special case of the general approximation form (4.2). The constant $q \in \mathbb{N}$ is called the rule's *level* and most directly controls the accuracy and computational cost of the approximation in applications. The Smolyak rule (4.4) is called a sparse grid quadrature rule if each of the $j = 1, \ldots, d$ univariate quadrature rules have nested nodes in the sense that $\mathscr{A}_j^1 \subset \mathscr{A}_j^2 \subset \cdots$. The rule (4.4) requires evaluation of f at the nodes

$$\mathscr{A}(q,d) = \bigcup_{q-d+1 \le |\boldsymbol{i}| \le q} \mathscr{A}_1^{i_1} \times \cdots \times A_d^{i_d}.$$

Adopting the convention that $A_j^0 = x_j^0$ for some base point $x_j^0 \in \mathscr{S}_j$, nesting implies $\mathscr{A}(q,d)$ is a sparse subset of the nodes used by the product rule $\left(U_1^q \otimes \cdots \otimes U_d^q\right)(f)$.

The sparse grid quadrature rule (4.4) mitigates the curse of dimensionality by creating sparse integration grids relative to product rules, but requires f to satisfy stricter smoothness properties in exchange. Novak and Ritter (1999) present growth rates, bounds, and approximations for $k = \left|\mathscr{A}(q,d)\right|$ under different scenarios. Novak and Ritter (1996) also show that the approximation's order of convergence is

$$\left|I(f) - A(q,d)(f)\right| = \mathscr{O}\left(k^{-r}(\log k)^{(d-1)(r/d+1)}\right)$$

if f has a bounded mixed derivative $f^{(r,\ldots,r)}$. Even more precisely, Novak and Ritter (1999) show that $I(f) = A(q,d)(f)$ if f is a polynomial with bounded total degree, i.e., that the approximation (4.4) is *exact* for the integral (4.1). The specific bound depends on exactness properties of the underlying univariate quadrature rules $U_1^{i_1}, \ldots, U_d^{i_d}$. The total degree of a polynomial is the maximum degree of its monomials, and the total degree of each monomial is the sum of the exponents of the variables that appear in it. For example, the total degree of the polynomial $x_1^3 + x_1^2 x_2^3 + x_2^4$ is 5. In practice, the sparse grid quadrature rule (4.4) is most computationally efficient for functions f that behave approximately as polynomials with relatively low total degree. In statistical contexts, this is similar to saying that the rule (4.4) is most useful for polynomial surfaces f that are mainly driven by main effects and low order interaction terms. We will satisfy this requirement for computational efficiency in our application by appealing, in part, to the Bayesian central limit theorem to claim that many posterior surfaces and other quantities can be well approximated by the product of a Gaussian weight function $w(\boldsymbol{x})$ with a relatively low-order correction term f.

4.3 Posterior inference via weighted mixtures

We combine conditioning techniques with sparse grid quadrature rules to develop specialized, computationally efficient formulas like (4.4) that approximate Bayesian posterior inference for marginal quantities. For example, when used to approximate marginal posterior densities, our method will yield a weighted mixture of full conditional posterior distributions. We briefly motivate the Bayesian Inference via Sparse grid Quadrature Evaluation (BISQuE) approximation strategy by arguing that it can be computationally inefficient to use sparse grid quadrature rules to directly approximate posterior quantities. First, our motivation simultaneously highlights the general strategy used to apply sparse grid quadrature rules to Bayesian models as well as key technical issues BISQuE addresses. Then, the remainder of Section 4.3 defines the family of posterior quantities to which BISQuE applies (Section 4.3.1), the BISQuE approximation (Section 4.4.3), and a nested integration technique that is useful for applying BISQuE to models that lack closed form expressions of posterior densities (Section 4.4.4).

Consider a generic hierarchical Bayesian model. Let $\boldsymbol{X} \in \Omega_0$ be a sample of continuous, discrete, or mixed random variables from an arbitrary process. Define a conditional probability model for $\boldsymbol{X}$ such that

$$\begin{aligned} \boldsymbol{X}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 &\sim f(\boldsymbol{X}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \\ (\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) &\sim f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \end{aligned} \tag{4.5}$$

for parameters $\boldsymbol{\theta}_1 \in \Omega_1$ and $\boldsymbol{\theta}_2 \in \Omega_2$. Many Bayesian models can be written like (4.5). For example, many hierarchical Bayesian models add conditional independence assumptions and hierarchical structure to (4.5) so that

$$\begin{aligned} f(\boldsymbol{X}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) &= f(\boldsymbol{X}|\boldsymbol{\theta}_1) \\ f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) &= f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)f(\boldsymbol{\theta}_2). \end{aligned}$$

Non-hierarchical models also fit within our framework (4.5). For example, Bayesian formulations of some linear regression models specify prior independence between regression coefficients $\boldsymbol{\theta}_1$ and variance components $\boldsymbol{\theta}_2$, thus define $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = f(\boldsymbol{\theta}_1) f(\boldsymbol{\theta}_2)$.

The marginal posterior density $f(\boldsymbol{\theta}_1 | \boldsymbol{X})$ is often of interest in posterior inference. The density may be computed by integrating $\boldsymbol{\theta}_2$ out of the joint posterior density

$$f(\boldsymbol{\theta}_1 | \boldsymbol{X}) = \int f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{X}) d\boldsymbol{\theta}_2. \tag{4.6}$$

Sparse grid quadrature rules (4.4) yield weighted-sum approximations (4.2) of (4.6) by introducing a weight function $w(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{X})$ and proceeding via

$$f(\boldsymbol{\theta}_1 | \boldsymbol{X}) = \int \frac{f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{X})}{w(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{X})} w(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{X}) d\boldsymbol{\theta}_2 \approx \sum_{\ell=1}^{k_i} \frac{f\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^{(i,\ell)} \middle| \boldsymbol{X}\right)}{w\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^{(i,\ell)}, \boldsymbol{X}\right)} w^{(i,\ell,\boldsymbol{\theta}_1)}, \tag{4.7}$$

in which quadrature nodes $\boldsymbol{\theta}_2^{(i,\ell)}$ and weights $w^{(i,\ell,\boldsymbol{\theta}_1)}$ are determined by applying the Smolyak formula (4.4) to a collection of univariate quadrature rules that are appropriate for the support of $\boldsymbol{\theta}_2$. For fixed $\boldsymbol{\theta}_1 \in \Omega_1$, the Gaussian approximation to $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{X})$ will often be a sensible default choice for the weight function $w(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{X})$ since the weight ratio f/w in (4.7) accounts for deviations from normality in $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{X})$.

The direct marginal posterior density approximation (4.7) has two key inefficiencies that the BISQuE approximation completely avoids or minimizes. First, the weight function w depends on $\boldsymbol{\theta}_1$, which implies a separate weight function must be used to approximate $f(\boldsymbol{\theta}_1 | \boldsymbol{X})$ at each $\boldsymbol{\theta}_1 \in \Omega_1$. Second, the approximation (4.7) assumes $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{X})$ is computable. Oftentimes, the joint posterior density $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{X})$ is only known in closed form up to a proportionality constant because the density's integration constant requires numerical approximation for many Bayesian models. While sparse grid quadrature rules could approximate the integration constant, BISQuE is able to avoid or reduce cost of the approximation.

4.3.1 Targeted posterior quantities

We develop BISQuE to approximate marginal posterior quantities $h(\boldsymbol{\theta}_1; \boldsymbol{X})$ of hierarchical models (4.5) that are defined implicitly with respect to a function or random variable $h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X})$ via

$$h(\boldsymbol{\theta}_1; \boldsymbol{X}) = \int h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X}) f(\boldsymbol{\theta}_2 | \boldsymbol{X}) d\boldsymbol{\theta}_2. \tag{4.8}$$

For example, the construction (4.8) defines the marginal posterior density $h(\boldsymbol{\theta}_1; \boldsymbol{X}) = f(\boldsymbol{\theta}_1 | \boldsymbol{X})$ when $h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X}) = f(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2, \boldsymbol{X})$. The posterior marginal density $f(\boldsymbol{\theta}_2 | \boldsymbol{X})$ and all other marginal posterior quantities may be formed by switching the roles of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. In comparison to the definition (4.6) used in the direct sparse grid approximation (4.7), the BISQuE construction (4.8) uses conditioning results to express the joint posterior density in conditional form, as $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{X}) = f(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2, \boldsymbol{X}) f(\boldsymbol{\theta}_2 | \boldsymbol{X})$. The construction (4.8) allows us to develop sparse grid quadrature rules with weight functions $w(\boldsymbol{\theta}_2, \boldsymbol{X})$ that only depend on $\boldsymbol{\theta}_2$ (Section 4.4.3), thus addresses the first technical issue described at the end of the Section 4.3 introduction.

The BISQuE construction (4.8) allows one set of quadrature nodes and weights to be reused to approximate many posterior quantities. For example, (4.8) defines the posterior mean $h(\boldsymbol{\theta}_1; \boldsymbol{X}) = E\left[g(\boldsymbol{\theta}_1) \middle| \boldsymbol{X} \right]$ when $h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X}) = \mathrm{E}\left[g(\boldsymbol{\theta}_1) \middle| \boldsymbol{\theta}_2, \boldsymbol{X} \right]$. Again, the approach relies on conditioning as

$$\begin{aligned} \mathrm{E}\left[g(\boldsymbol{\theta}_1) \middle| \boldsymbol{X} \right] &= \mathrm{E}_{\boldsymbol{\theta}_2 | \boldsymbol{X}} \left\{ \mathrm{E}\left[g(\boldsymbol{\theta}_1) \middle| \boldsymbol{\theta}_2, \boldsymbol{X} \right] \right\} \\ &= \int \mathrm{E}\left[g(\boldsymbol{\theta}_1) \middle| \boldsymbol{\theta}_2, \boldsymbol{X} \right] f(\boldsymbol{\theta}_2 | \boldsymbol{X}) d\boldsymbol{\theta}_2. \end{aligned}$$

Posterior predictive distributions, variances and higher central moments, cumulative distribution functions, and model selection criteria such as the deviance information criteria (DIC, Spiegelhalter et al., 2002) and the Watanabe-Akaike information criterion (WAIC, Watanabe, 2010) can also be expressed through one or more applications of (4.8). To be precise, the posterior variance $\mathrm{Var}\left(g(\boldsymbol{\theta}_1) \middle| \boldsymbol{X} \right)$ can be approximated by using the law of total variance to introduce

expectations with respect to $f(\boldsymbol{\theta}_2|\boldsymbol{X})$ via

$$\begin{aligned}\text{Var}\big(g(\boldsymbol{\theta}_1)\big|\boldsymbol{X}\big) =& \text{E}_{\boldsymbol{\theta}_2|\boldsymbol{X}}\big[\text{Var}\big(g(\boldsymbol{\theta}_1)\big|\boldsymbol{\theta}_2,\boldsymbol{X}\big)\big] + \\ & \text{E}_{\boldsymbol{\theta}_2|\boldsymbol{X}}\Big[\big(\text{E}\big[g(\boldsymbol{\theta}_1)\big|\boldsymbol{\theta}_2,\boldsymbol{X}\big] - \text{E}\big[g(\boldsymbol{\theta}_1)\big|\boldsymbol{X}\big]\big)^2\Big],\end{aligned} \tag{4.9}$$

for which

$$h(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2;\boldsymbol{X}) = \text{Var}\big(g(\boldsymbol{\theta}_1)\big|\boldsymbol{\theta}_2,\boldsymbol{X}\big) + \big(\text{E}\big[g(\boldsymbol{\theta}_1)\big|\boldsymbol{\theta}_2,\boldsymbol{X}\big] - \text{E}\big[g(\boldsymbol{\theta}_1)\big|\boldsymbol{X}\big]\big)^2. \tag{4.10}$$

Note that the marginal posterior expectation $\text{E}\big[g(\boldsymbol{\theta}_1)\big|\boldsymbol{X}\big]$ must be approximated before (4.9). We present expressions for the other quantities mentioned in Section 4.3.2.

4.3.2 Additional posterior quantities

We briefly formulate additional posterior quantities as integrals of functions with respect to the posterior density $f(\boldsymbol{\theta}_2|\boldsymbol{X})$, which is required for our construction (4.8).

Posterior predictive distributions

Posterior predictive distributions $f(\boldsymbol{X}_0|\boldsymbol{X})$ naturally fit into the framework described in Section 3.2.1 because

$$\begin{aligned}f(\boldsymbol{X}_0|\boldsymbol{X}) &= \int f(\boldsymbol{X}_0,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\boldsymbol{X})\,d(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \\ &= \int f(\boldsymbol{X}_0|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{X})\,f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\boldsymbol{X})\,d(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2).\end{aligned} \tag{4.11}$$

The posterior predictive distribution (4.11) is exactly a marginal posterior quantity as in (4.8) for a hierarchical model like (4.5) in which $\boldsymbol{\theta}_1' = \boldsymbol{X}_0$ and $\boldsymbol{\theta}_2' = (\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$.

Higher order central moments

Posterior variances $\text{Var}(\boldsymbol{\theta}_1|\boldsymbol{X})$ can be computed with assistance from the law of total variance (4.9), which uses conditional variances and expectations to facilitate computation. The

decomposition is convenient as conditional variances and expectations may be available in closed form. However, decompositions similar to the law of total variance are not available for general higher order central moments. Approximations must be constructed from the definition of higher order central moments via

$$\mathrm{E}\left[(\boldsymbol{\theta}_1 - \mathrm{E}[\theta_1 | \boldsymbol{X}])^n \middle| \boldsymbol{X} \right] = \int \mathrm{E}_{\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2, \boldsymbol{X}} \left[(\boldsymbol{\theta}_1 - \mathrm{E}[\theta_1 | \boldsymbol{X}])^n \right] f(\boldsymbol{\theta}_2 | \boldsymbol{X}) d\boldsymbol{\theta}_2.$$

The integrand $\mathrm{E}_{\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2, \boldsymbol{X}} \left[(\boldsymbol{\theta}_1 - \mathrm{E}[\theta_1 | \boldsymbol{X}])^n \right]$ does not represent a conditional central moment because the moment is centered around the posterior mean $\mathrm{E}[\theta_1 | \boldsymbol{X}]$ while the expectation is taken with respect to the conditional posterior density $f(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2, \boldsymbol{X})$. If the moment cannot be computed in closed form, approximation strategies may depend on the hierarchical model in question. For example, sparse grid quadrature rules could directly approximate the conditional expectation $\mathrm{E}_{\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2, \boldsymbol{X}} \left[(\boldsymbol{\theta}_1 - \mathrm{E}[\theta_1 | \boldsymbol{X}])^n \right]$, or it may also be possible to use Laplace approximations.

Cumulative distribution functions

Marginal cumulative distribution functions (CDFs) may be formulated as a weighted average of conditional CDFs. The posterior density $f(\boldsymbol{\theta}_1 | \boldsymbol{X})$ may be expressed as an integral with respect to $f(\boldsymbol{\theta}_2 | \boldsymbol{X})$, and Fubini's theorem allows an exchange of integrals that yield the result via

$$\begin{aligned} F(\boldsymbol{\theta}_1 \leq \boldsymbol{t} | \boldsymbol{X}) &= \int_{-\infty}^{\boldsymbol{t}} \left(\int f(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2, \boldsymbol{X}) f(\boldsymbol{\theta}_2 | \boldsymbol{X}) d\boldsymbol{\theta}_2 \right) d\boldsymbol{\theta}_1 \\ &= \int F(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2, \boldsymbol{X}) f(\boldsymbol{\theta}_2 | \boldsymbol{X}) d\boldsymbol{\theta}_2. \end{aligned}$$

Information criteria

The Deviance information criteria (DIC, Spiegelhalter et al., 2002) allows for model comparison and is based on the deviance, defined via

$$D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = -2 \ln f(\boldsymbol{X} | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + C$$

for a constant C that depends on the data. The DIC is defined via

$$\text{DIC} = \text{E}[D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|\boldsymbol{X}] + p_D,$$

in which $p_D = \text{E}[D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|\boldsymbol{X}] - D(\text{E}[\boldsymbol{\theta}_1|\boldsymbol{X}], \text{E}[\boldsymbol{\theta}_2|\boldsymbol{X}])$. Only the posterior expectation $\text{E}[D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|\boldsymbol{X}]$ requires additional formulation. The law of total expectation yields an integral with respect to $f(\boldsymbol{\theta}_2|\boldsymbol{X})$ via

$$\text{E}[D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|\boldsymbol{X}] = \int \text{E}_{\boldsymbol{\theta}_1|\boldsymbol{\theta}_2,\boldsymbol{X}}[D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)] f(\boldsymbol{\theta}_2|\boldsymbol{X}) d\boldsymbol{\theta}_2.$$

Similar to the formulation of higher order central moments (Section 4.3.2), closed form expressions may be available for the integrand $\text{E}_{\boldsymbol{\theta}_1|\boldsymbol{\theta}_2,\boldsymbol{X}}[D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)]$, or it may need to be approximated directly via sparse grid quadrature rules or via Laplace approximations. The Watanabe-Akaike information criterion (WAIC) uses similar quantities as the DIC, so may be similarly approximated (Watanabe, 2010).

4.4 Additional computational techniques

BISQuE approximations use weighted sums of densities and likelihoods, however, it is often more numerically stable to evaluate log-densities and log-likelihoods. This section reviews techniques that allow log-densities and log-likelihoods to be used to compute weighted sums of densities.

4.4.1 Evaluating an unnormalized density

Let $f(x)$ be a probability density function such that $f(x) \propto g(x)$ for some unnormalized density function $g(x)$. If $g(x)$ or $\ln g(x)$ are known, then one strategy to evaluate $f(x)$ is to first compute the integration constant $C = \int g(x)dx$. In special cases, quadrature techniques can efficiently approximate the integration constant C. However, the numerical stability of such approximations are often better when $\ln g(x)$ is used to compute kC for some scale factor $k > 0$. An

m point quadrature rule with quadrature nodes $\{x^{(i)} : i = 1, \ldots, m\}$ and weights $\{w_i : i = 1, \ldots, m\}$ approximates the scaled constant kC via

$$kC = \int \exp\{\ln g(x) + \ln k\}\, dx \approx \sum_{i=1}^{m} \exp\left\{\ln g\left(x^{(i)}\right) + \ln k\right\} w_i. \tag{4.12}$$

Choosing k can be difficult, but the approximation (4.12) suggests that k such that $\ln k = -m^{-1} \sum_{i=1}^{m} \ln g\left(x^{(i)}\right)$ will often be a reasonable choice since $\ln k$ centers the shifted unnormalized log-densities $\left\{\ln g\left(x^{(i)}\right) + \ln k : i = 1, \ldots, m\right\}$ around 0. The integration constant C can be recovered via $C = \exp\{\ln kC - \ln k\}$ after kC and $\ln k$ are numerically evaluated.

4.4.2 Evaluating mixture densities

Let $f(x)$ be a mixture of densities $\{f_i(x) : i = 1, \ldots, m\}$ with weights $\{w_i : i =, \ldots, m\}$ specified via

$$f(x) = \sum_{i=1}^{m} f_i(x)\, w_i.$$

One strategy for evaluating $f(x)$ is to again introduce a scale factor $k > 0$. This allows for numerically stable evaluation of $f(x)$ via

$$f(x) = \frac{kf(x)}{k} = \frac{\sum_{i=1}^{m} \exp\{\ln f_i(x) + \ln k\}\, w_i}{k}.$$

Choosing k can be difficult, but as in Section 4.4.1, k such that $\ln k = -m \sum_{i=1}^{m} \ln f_i(x)$ will often be reasonable.

4.4.3 Approximate posterior inference

We specialize the integral form (4.1) and use sparse grid quadrature rules (4.4) to approximate marginal posterior quantities (4.8) of hierarchical Bayesian models (4.5). While we define marginal posterior quantities by integrating functions over the posterior density $f(\boldsymbol{\theta}_2 | \boldsymbol{X})$, numerical integration methods often use transformations to increase computational stability and

efficiency. Thus, we develop quadrature rules that integrate over $f(\boldsymbol{\nu}|\boldsymbol{X})$ where $\boldsymbol{\nu} = T(\boldsymbol{\theta}_2) \in \mathbb{R}^p$ is defined by a monotone transformation to a real coordinate space $T : \Omega_2 \to \mathbb{R}^p$. Change of variable results imply the transformed density is specified via

$$f(\boldsymbol{\nu}|\boldsymbol{X}) = f\left(T^{-1}(\boldsymbol{\nu})\middle|\boldsymbol{X}\right)\left|J\left(T^{-1}(\boldsymbol{\nu})\right)\right|,$$

in which $\left|J\left(T^{-1}(\boldsymbol{\nu})\right)\right|$ is the determinant of the Jacobian for the transformation T^{-1}. We propose using sparse grid quadrature rules (4.4) to derive quadrature nodes and weights that approximate marginal posterior quantities (4.8) via the BISQuE approximation

$$\begin{aligned} h(\boldsymbol{\theta}_1;\boldsymbol{X}) &= \int h\left(\boldsymbol{\theta}_1, T^{-1}(\boldsymbol{\nu});\boldsymbol{X}\right)\frac{f(\boldsymbol{\nu}|\boldsymbol{X})}{w(\boldsymbol{\nu},\boldsymbol{X})}w(\boldsymbol{\nu},\boldsymbol{X})d\boldsymbol{\nu} \\ &\approx \sum_{\ell=1}^{k_i} h\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^{(i,\ell)};\boldsymbol{X}\right)\tilde{w}^{(i,\ell)}, \end{aligned} \tag{4.13}$$

in which

$$\tilde{w}^{(i,\ell)} = \frac{f\left(\boldsymbol{\nu}^{(i,\ell)}\middle|\boldsymbol{X}\right)}{w\left(\boldsymbol{\nu}^{(i,\ell)},\boldsymbol{X}\right)}w^{(i,\ell)},$$

$w(\boldsymbol{\nu},\boldsymbol{X})$ is a weight function; and $w^{(i,\ell)}$, $\boldsymbol{\nu}^{(i,\ell)}$, and $\boldsymbol{\theta}_2^{(i,\ell)} = T^{-1}\left(\boldsymbol{\nu}^{(i,\ell)}\right)$ are respectively quadrature weights, nodes, and back-transformed nodes. Software libraries, including the `mvQuad` package for `R` and the SGMGA libraries for C and C++ (Burkardt, 2007; Weiser, 2016), contain tables and routines that compute sparse grid quadrature nodes and weights if $w(\boldsymbol{\nu},\boldsymbol{X})$ is a member of a standard family of weight functions (Givens and Hoeting, 2013, Table 5.6).

Sparse grid quadrature theory implies the computational efficiency of the approximation (4.13) relies on several statistical and numerical assumptions. The weight function $w(\boldsymbol{\nu},\boldsymbol{X})$ should approximate the transformed density $f(\boldsymbol{\nu}|\boldsymbol{X})$ well and have known, computationally efficient, nested quadrature rules. In particular, such quadrature rules have been developed for Gaussian weight functions (Genz and Keister, 1996). Thus, we appeal to Bayesian analogs of the central limit theorem if sample size is large and the dimension of the model is fixed to

justify proposing the Gaussian approximation $f^G(\boldsymbol{\nu}|\boldsymbol{X})$ at the posterior mode of $f(\boldsymbol{\nu}|\boldsymbol{X})$ as a sensible default choice for a weight function for many Bayesian models (Berger, 1985, pg. 224–225). Sparse grid quadrature rules will also be most efficient if the modified integrand $h\big(\boldsymbol{\theta}_1, T^{-1}(\boldsymbol{\nu}); \boldsymbol{X}\big) f(\boldsymbol{\nu}|\boldsymbol{X}) / w(\boldsymbol{\nu}, \boldsymbol{X})$ in (4.13) can be well-approximated by a low-order polynomial in $\boldsymbol{\nu}$. This requirement is easier to satisfy if the weight function $w(\boldsymbol{\nu}, \boldsymbol{X})$ approximates $f(\boldsymbol{\nu}|\boldsymbol{X})$ well and $h\big(\boldsymbol{\theta}_1, T^{-1}(\boldsymbol{\nu}); \boldsymbol{X}\big)$ is slowly varying with respect to $\boldsymbol{\nu}$.

Standardizing the BISQuE approximation (4.13) weights $\tilde{w}^{(i,\ell)}$ can address part of the second technical issue described at the end of the Section 4.3 introduction. For example, it is possible to have Bayesian models in which both the joint $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\boldsymbol{X})$ and marginal $f(\boldsymbol{\theta}_2|\boldsymbol{X})$ posterior densities are known up to a proportionality constant while the full conditional posterior $f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{X})$ is completely known (Section 4.5). In such cases, using standardized weights $\tilde{w}_*^{(i,\ell)} = \tilde{w}^{(i,\ell)} / \sum_{j=1}^{k_i} \tilde{w}^{(i,j)}$ that sum to one can approximate marginal posterior quantities $h(\boldsymbol{\theta}_1; \boldsymbol{X})$ like $f(\boldsymbol{\theta}_1|\boldsymbol{X})$ by implicitly cancelling the unknown integration constants. The result borrows ideas from importance sampling (Givens and Hoeting, 2013, pg. 181). An alternate definition for posterior quantities,

$$h(\boldsymbol{\theta}_1; \boldsymbol{X}) = \frac{\int h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X}) f(\boldsymbol{\theta}_2|\boldsymbol{X}) d\boldsymbol{\theta}_2}{\int f(\boldsymbol{\theta}_2|\boldsymbol{X}) d\boldsymbol{\theta}_2}, \tag{4.14}$$

is equivalent to the original construction (4.8) since $\int f(\boldsymbol{\theta}_2|\boldsymbol{X}) d\boldsymbol{\theta}_2 = 1$. Plugin BISQuE approximations (4.13) for the numerator and denominator in (4.14) yield quadrature approximations with standardized weights via

$$\frac{\int h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X}) f(\boldsymbol{\theta}_2|\boldsymbol{X}) d\boldsymbol{\theta}_2}{\int f(\boldsymbol{\theta}_2|\boldsymbol{X}) d\boldsymbol{\theta}_2} \approx \frac{\sum_{\ell=1}^{k_i} h\Big(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^{(i,\ell)}; \boldsymbol{X}\Big) \tilde{w}^{(i,\ell)}}{\sum_{j=1}^{k_i} \tilde{w}^{(i,j)}} = \sum_{\ell=1}^{k_i} h\Big(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^{(i,\ell)}; \boldsymbol{X}\Big) \tilde{w}_*^{(i,\ell)}. \tag{4.15}$$

Standardization also allows approximations of $f(\boldsymbol{\theta}_1|\boldsymbol{X})$ to integrate exactly to one.

Table 4.1 summarizes the procedures outlined in this section as they would be applied when using a Gaussian approximation to the transformed posterior density to approximate posterior quantities (4.8).

Table 4.1: Summary of steps to develop a BISQuE approximation.

1. Write posterior quantity of interest in BISQuE form (4.8).

   Computable approximations or exact expressions must exist for the components $h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X})$ *and* $f(\boldsymbol{\theta}_2 | \boldsymbol{X})$. *Section 4.4.4 proposes nested integration strategies (4.17) and (4.18) if approximation is necessary; nested Laplace approximations can also be used for components in latent Gaussian models (cf. Rue et al., 2009).*

2. Select transformation $\boldsymbol{\nu} = T(\boldsymbol{\theta}_2)$ to map $\boldsymbol{\theta}_2 \in \Omega_2$ to $\boldsymbol{\nu} \in \mathbb{R}^p$.

   Favor transformations T *that yield an approximately Gaussian posterior density* $f(\boldsymbol{\nu} | \boldsymbol{X})$.

3. Apply the BISQuE approximation that uses unstandardized (4.13) or standardized (4.15) weights.

   The level $q \in \mathbb{N}$ *of the underlying sparse grid quadrature rule (4.4) determines the integration nodes* $\boldsymbol{\nu}^{(i,\ell)}$ *and weights* $w^{(i,\ell)}$.

4. Increase the level q of underlying quadrature rule (4.4) until the approximation (4.13) or (4.15) converges.

   Nested quadrature rules allow the level q *approximation to reduce computational cost by reusing quadrature nodes and weight ratios from the level* $q - 1$ *approximation.*

4.4.4 Nested integration strategies

While hierarchical Bayesian models (4.5) typically have closed form expressions for the likelihood $f(\boldsymbol{X}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ and prior $f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$, many models do not have closed form expressions for the posterior densities $f(\boldsymbol{\theta}_2|\boldsymbol{X})$ and $f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2,\boldsymbol{X})$. Lack of closed form expressions is a concern related to the second technical issue described at the end of the Section 4.3 introduction. We propose a nested numerical integration scheme to address the concern and allow application of BISQuE to a wider range of models. Recall that for a fixed dataset $\boldsymbol{X}$, the joint posterior density $f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\boldsymbol{X})$ is often only known up to a proportionality constant since

$$f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\boldsymbol{X}) = \frac{f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{X})}{f(\boldsymbol{X})} \propto f(\boldsymbol{X}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$$

and the marginal density $f(\boldsymbol{X})$ often requires prohibitively expensive numerical approximation.

The densities $f(\boldsymbol{\theta}_2|\boldsymbol{X})$ and $f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2,\boldsymbol{X})$ may be derived (and ultimately approximated) indirectly, by factoring the joint density $f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{X})$ into components $g_1(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2;X)$ and $g_2(\boldsymbol{\theta}_2;\boldsymbol{X})$ such that

$$f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{X}) = g_1(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2;\boldsymbol{X}) g_2(\boldsymbol{\theta}_2;\boldsymbol{X}). \tag{4.16}$$

The factored joint density (4.16) implies

$$f(\boldsymbol{\theta}_2|\boldsymbol{X}) = \int f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\boldsymbol{X}) d\boldsymbol{\theta}_1 = \frac{g_2(\boldsymbol{\theta}_2;\boldsymbol{X}) C_1(\boldsymbol{\theta}_2)}{f(\boldsymbol{X})} \tag{4.17}$$

and

$$f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2,\boldsymbol{X}) = \frac{f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\boldsymbol{X})}{f(\boldsymbol{\theta}_2|\boldsymbol{X})} = \frac{g_1(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2;\boldsymbol{X})}{C_1(\boldsymbol{\theta}_2)}, \tag{4.18}$$

for which the integration constant $C_1(\boldsymbol{\theta}_2)$ must be approximated numerically and is specified via

$$C_1(\boldsymbol{\theta}_2) = \int g_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X}) d\boldsymbol{\theta}_1. \tag{4.19}$$

The alternate expressions (4.17) and (4.18) allow BISQuE to approximate posterior inference for models that lack closed form expressions for the densities $f(\boldsymbol{\theta}_2|\boldsymbol{X})$ and $f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{X})$. Standardized BISQuE weights $\tilde{w}_*^{(i,\ell)}$ implicitly cancel the unknown factor $f(\boldsymbol{X})$, and standard quadrature techniques can efficiently approximate the integration constant (4.19) when the parameter vector $\boldsymbol{\theta}_1$ has small dimension. The parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ can often be defined or repartitioned to satisfy this requirement because the hierarchical model (4.5) places few restrictions on the parameters; we use this flexibility in Section 4.5. The added computational cost that the nested integration (4.19) adds to the BISQuE approximation is minimized as the integration constant (4.19) only needs to be approximated relatively few times, specifically, at the quadrature nodes and when developing the weight function—e.g., the Gaussian approximation at the posterior mode.

4.5 Examples

We demonstrate the benefits of the BISQuE approximation (4.13) on data that are typically analyzed with standard, Gibbs sampling techniques for approximate Bayesian posterior inference. We approximate posterior inference for a fully non-Gaussian capture-recapture model (Section 4.5.1), a spatial Gaussian process model (Section 4.5.2), and a more complex, applied spatial Gaussian process model for climate teleconnection (Section 4.5.3). Posterior distributions in the first and third examples respectively require integration over 8 and 5-dimensional parameter vectors $\boldsymbol{\theta}_2$. Posterior approximations for the second and third examples have computational complexity that is $\mathcal{O}\left(MN^3\right)$ in the number of spatial observations N and M points at which the posterior distribution is explored, thus computational strategies like BISQuE that reduce the number of points required for posterior approximation can be extremely beneficial.

We compare posterior inference and computational effort between standard Gibbs sampling techniques and BISQuE. Computational effort is measured indirectly with respect to com-

putation time. All computations are conducted on a modest workstation with eight logical processors. We use parallelization to compute the BISQuE approximation's k_i mixture components, and to draw posterior predictive samples via composition sampling in the spatial examples (cf. Banerjee et al., 2015, pg. 126). For each posterior quantity, the level q for the underlying sparse grid quadrature rule (4.4) is chosen to be the smallest value (i.e., the simplest approximation) such that the posterior density approximations have converged. The number of Gibbs steps used in each approximation is similarly chosen. The BISQuE approximation also requires specification of univariate quadrature rules, for which we choose nested Gauss-Hermite rules (Genz and Keister, 1996).

4.5.1 Fur seals

Data and model

Givens and Hoeting (2013, example 7.7) analyze data from a capture-recapture study conducted in New Zealand. The study's research goal was to estimate the total number of pups in a fur seal colony $N \in \mathbb{N}$. Researchers visited the colony $I = 7$ times throughout the course of a single season. In each visit, the researchers captured and marked all of the fur seal pups present, noting the total number of pups captured in each visit $\boldsymbol{c} = (c_1, \ldots, c_I) \in \mathbb{N}^I$ in addition to the number of newly captured pups $m_1, \ldots, m_I \in \mathbb{N}$. The data are analyzed using a Bayesian model for capture-recapture data (4.20), and posterior distributions are approximated with a Gibbs sampler. Gibbs sampling is particularly inefficient as one pair of hyperparameters has high posterior correlation and are only weakly identified by the data. By comparison, the BISQuE strategy (4.13) approximates posterior quantities for this model with substantially less computational effort.

The model (4.20) assumes N remains fixed during the time period of the study (i.e., the model assumes a closed population). Let $r = \sum_{i=1}^{I} m_i$ be the total number of pups captured during the study. Givens and Hoeting (2013) introduce a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I) \in [0,1]^I$ with capture probabilities for each census attempt and discuss modeling the data with the hierar-

chical model

$$
\begin{aligned}
f(\boldsymbol{c}, r \mid N, \boldsymbol{\alpha}) &\propto \frac{N!}{(N-r)!} \prod_{i=1}^{I} \alpha_i^{c_i} (1-\alpha_i)^{N-c_i} \\
f(N) &\propto 1/N \\
f(\alpha_i \mid \theta_1, \theta_2) &\sim \text{Beta}(\theta_1, \theta_2) \text{ for } i = 1, \ldots, I \\
f(\theta_1, \theta_2) &\propto \exp\{-(\theta_1 + \theta_2)/1000\},
\end{aligned}
\tag{4.20}
$$

in which (θ_1, θ_2) are hyperparameters for the capture probabilities. We use the Beta distribution's mean–sample size parameterization to increase the identifiability of the hyperparameters. Specifically, let $U_1 = \text{logit}(\theta_1/(\theta_1+\theta_2))$ and $U_2 = \log(\theta_1+\theta_2)$ and fix $U_2 = 5.5$.

Derivations

We derive components required for the BISQuE approximations of posterior quantities for the fur seals example, as specified in Table 4.2.

Joint posterior

The joint posterior density is known up to a proportionality constant via

$$
\begin{aligned}
f(N, \boldsymbol{\alpha}, \theta_1, \theta_2 \mid \boldsymbol{c}, r) &\propto f(\boldsymbol{c}, r \mid N, \boldsymbol{\alpha}) f(N) f(\boldsymbol{\alpha} \mid \theta_1, \theta_2) f(\theta_1, \theta_2) \\
&\propto \frac{(N-1)!}{(N-r)!} \frac{\exp\{-(\theta_1+\theta_2)/1000\}}{B(\theta_1, \theta_2)^I} \prod_{i=1}^{I} \alpha_i^{c_i+\theta_1-1} (1-\alpha_i)^{N-c_i+\theta_2-1},
\end{aligned}
\tag{4.21}
$$

in which $B(\theta_1, \theta_2) = \Gamma(\theta_1)\Gamma(\theta_2)/\Gamma(\theta_1+\theta_2)$ is the beta function.

Population size

The BISQuE approximation for $f(N \mid \boldsymbol{c}, r)$ uses the two conditional posterior densities $f(N-r \mid \boldsymbol{\alpha}, \theta_1, \theta_2, \boldsymbol{c}, r)$ and $f(\boldsymbol{\alpha}, \theta_1, \theta_2 \mid \boldsymbol{c}, r)$, which are derived from the joint posterior density (4.21). Factoring (4.21) yields

$$f(N|\boldsymbol{\alpha},\theta_1,\theta_2,\boldsymbol{c},r) \propto \frac{(N-1)!}{(N-r)!}\prod_{i=1}^{I}(1-\alpha_i)^N,$$

which implies the change of variable $k = N - r$ yields the result

$$f(N-r|\boldsymbol{\alpha},\theta_1,\theta_2,\boldsymbol{c},r) \sim \text{Neg. Bin.}\left(r, 1-\prod_{i=1}^{I}(1-\alpha_i)\right).$$

The posterior density $f(\boldsymbol{\alpha},\theta_1,\theta_2|\boldsymbol{c},r)$ may be computed by marginalizing (4.21) with respect to N, via

$$\begin{aligned} f(\boldsymbol{\alpha},\theta_1,\theta_2|\boldsymbol{c},r) &\propto \sum_{N=r}^{\infty} f(N,\boldsymbol{\alpha},\theta_1,\theta_2|\boldsymbol{c},r) \\ &\propto h(\boldsymbol{\alpha})\frac{\exp\{-(\theta_1+\theta_2)/1000\}}{B(\theta_1,\theta_2)^I}\prod_{i=1}^{I}\alpha_i^{c_i+\theta_1-1}(1-\alpha_i)^{\theta_2-c_i-1}, \end{aligned} \tag{4.22}$$

in which

$$\begin{aligned} h(\boldsymbol{\alpha}) &= \sum_{N=r}^{\infty}\frac{(N-1)!}{(N-r)!}(1-p)^N \\ &\propto (1-p)^r p^{-r}\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}(1-p)^k p^r \\ &\propto (1-p)^r p^{-r} \end{aligned}$$

for $1-p = \prod_{i=1}^{I}(1-\alpha_i)$ and the change of variable $k = N - r$. The marginalized posterior may be simplified further since (4.22) contains kernels for Beta distributions. Thus,

$$\begin{aligned} f(\boldsymbol{\alpha},\theta_1,\theta_2|\boldsymbol{c},r) \propto \frac{\exp\{-(\theta_1+\theta_2)/1000\}}{B(\theta_1,\theta_2)^I}\times \\ p^{-r}\prod_{i=1}^{I}B(\theta_1+c_i,\theta_2+r-c_i)f(\alpha_i|\theta_1+c_i,\theta_2+r-c_i). \end{aligned} \tag{4.23}$$

Capture probabilities

The BISQuE approximation for $f(\alpha_i|\boldsymbol{c},r)$ uses the posterior densities $f(\alpha_i|N,\theta_1,\theta_2)$ and $f(N,\theta_1,\theta_2|\boldsymbol{c},r)$. Factoring (4.21) immediately yields

$$f(\alpha_i|N,\theta_1,\theta_2) \sim \text{Beta}(\theta_1+c_i,\theta_2+N-c_i).$$

Similarly, marginalizing (4.21) with respect to $\boldsymbol{\alpha}$ immediately yields

$$f(N,\theta_1,\theta_2|\boldsymbol{c},r) \propto \frac{(N-1)!}{(N-r)!}\frac{\exp\{-(\theta_1+\theta_2)/1000\}}{B(\theta_1,\theta_2)^I}\prod_{i=1}^{I} B(\theta_1+c_i,\theta_2+N-c_i).$$

Hyperparameters

The BISQuE approximation for $f(U_1|\boldsymbol{c},r)$ uses the posterior densities $f(\theta_1,\theta_2|\boldsymbol{\alpha},\boldsymbol{c},r)$ and $f(\boldsymbol{\alpha}|\boldsymbol{c},r)$. Factoring (4.23) yields

$$f(\theta_1,\theta_2|\boldsymbol{\alpha},\boldsymbol{c},r) \propto \frac{\exp\{-(\theta_1+\theta_2)/1000\}}{B(\theta_1,\theta_2)^I}\prod_{i=1}^{I}\alpha_i^{\theta_1}(1-\alpha_i)^{\theta_2}.$$

The marginal posterior $f(\boldsymbol{\alpha}|\boldsymbol{c},r)$ must be approximated via nested integration, and is specified via

$$f(\boldsymbol{\alpha}|\boldsymbol{c},r) \propto p^{-r}\left(\prod_{i=1}^{I}\alpha_i^{c_i-1}(1-\alpha_i)^{r-c_i-1}\right)\int f(\theta_1,\theta_2|\boldsymbol{\alpha},\boldsymbol{c},r)d(\theta_1,\theta_2).$$

Posterior inference and results

Givens and Hoeting (2013) use standard Gibbs-sampling approaches to draw posterior samples for model parameters. The full conditional posterior distributions $f(N|\boldsymbol{c},r,\boldsymbol{\alpha},\theta_1,\theta_2)$ and $f(\boldsymbol{\alpha}|\boldsymbol{c},r,N,\theta_1,\theta_2)$ are conjugate and easy to sample. Posterior samples for U_1 are drawn using Metropolis steps. The sampler is run for 100,000 iterations, taking 298 seconds to complete; posterior inference uses the final 50,000 samples.

We use the BISQuE strategy to approximate the posterior marginal densities $f(N|\boldsymbol{c},r)$, $f(\alpha_i|\boldsymbol{c},r)$, and $f(U_1|\boldsymbol{c},r)$. Table 4.2 connects this example's notation to that used with BISQuE.

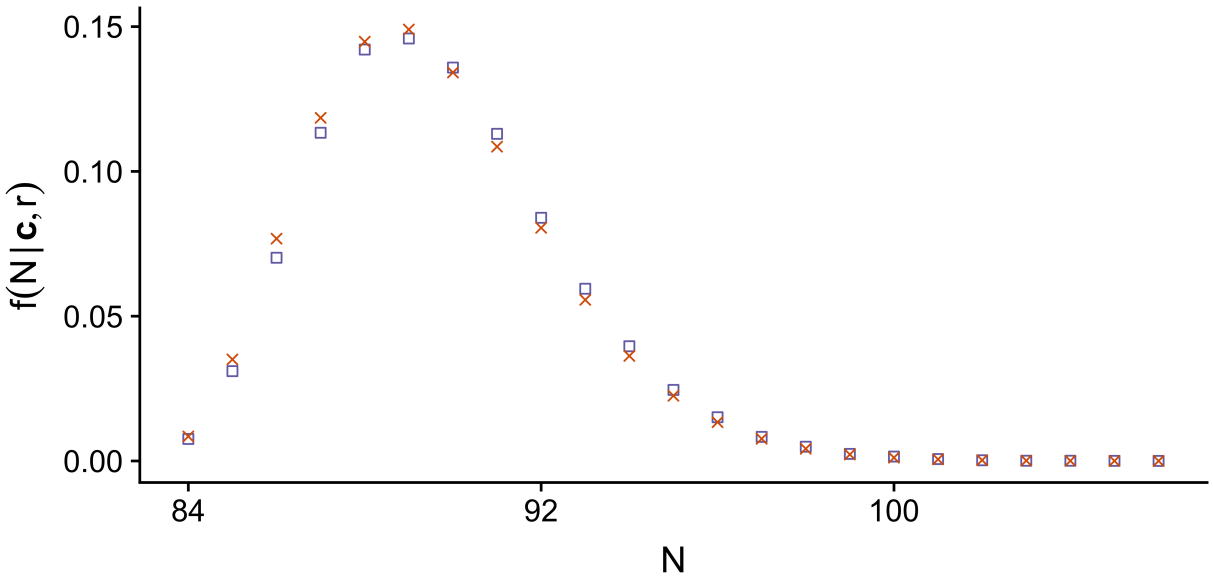

Figure 4.1: BISQuE (x) and Gibbs (□) approximations to the posterior density for total number of fur seal pups $f(N|\boldsymbol{c},r)$ are nearly identical.

When used as the BISQuE conditioning variable $\boldsymbol{\theta}_2$, we map parameters to the real line by using log transforms with $N-r$ and logit transforms with the capture probabilities $\boldsymbol{\alpha}$. We also rely on the Gaussian approximation to the negative binomial distribution in order to justify using N as a conditioning variable $\boldsymbol{\theta}_2$ in BISQuE. Almost all conditional and marginal posterior densities required for BISQuE are computable in closed form up to a proportionality constant (Givens and Hoeting (2013, eqs. 7.16, 7.17) and Section 4.5.1). The posterior for $f(U_1|\boldsymbol{c},r)$ requires approximation via nested integration strategies (Section 4.4.4).

Posterior inference via BISQuE is effectively identical to posterior inference via Gibbs sampling, but is computed with substantially less effort. Gibbs sampling takes 298 seconds to complete on our test machine, whereas the BISQuE approximations require a total of 5 seconds (Table 4.2), and posterior densities are nearly identical (Figures 4.1 to 4.3).

Table 4.2: Definitions of the parameters and posterior quantities for the BISQuE approximations in Section 4.5. $\boldsymbol{X} = (\boldsymbol{c}, r)$ for the fur seals example (Section 4.5.1), and $\boldsymbol{X} = \boldsymbol{Y}$ for the Remote effects spatial process model example (RESP, Section 4.5.3). The marginal posterior densities for the covariance parameters (σ^2, ρ, ν) in the spatial example (Section 4.5.2) are computed using sparse-grid quadrature methods (4.4) to directly marginalize the joint posterior distribution $f(\sigma^2, \rho, \nu | \boldsymbol{X})$ at each evaluation point. Computation times are also presented. For the RESP example, let $\boldsymbol{\theta}^* = (\sigma_w^2, \sigma_\varepsilon^2, \sigma_\alpha^2, \rho_w, \rho_\alpha)$ and $I(\boldsymbol{s}) = (c_{i-1}(\boldsymbol{s}), c_i(\boldsymbol{s}))$.

					Time (sec.)	
Example	$h(\boldsymbol{\theta}_1; \boldsymbol{X})$	$h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X})$	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	BISQuE	Gibbs
Fur seals	$f(N \mid \boldsymbol{c}, r)$	$f(N \mid \boldsymbol{\theta}_2, \boldsymbol{c}, r)$	N	$(\boldsymbol{\alpha}, U_1)$	0.3	298
	$f(\alpha_i \mid \boldsymbol{c}, r)$	$f(\alpha_i \mid \boldsymbol{\theta}_2, \boldsymbol{c}, r)$	α_i	(N, U_1)	0.1	298
	$f(U_1 \mid \boldsymbol{c}, r)$	$f(U_1 \mid \boldsymbol{\theta}_2, \boldsymbol{c}, r)$	U_1	$\boldsymbol{\alpha}$	5.0	298
Spatial	$\mathrm{E}[\boldsymbol{X}_0 \mid \boldsymbol{X}]$	$\mathrm{E}[\boldsymbol{X}_0 \mid \boldsymbol{\theta}_2, \boldsymbol{X}]$	$\boldsymbol{X}_0$	(σ^2, ρ, ν)	6	2,651
	$\mathrm{Var}(\boldsymbol{X}_0 \mid \boldsymbol{X})$	(4.10)	$\boldsymbol{X}_0$	(σ^2, ρ, ν)	6	2,651
	$f(\boldsymbol{X}_0 \mid \boldsymbol{X})$	$f(\boldsymbol{X}_0 \mid \boldsymbol{\theta}_2, \boldsymbol{X})$	$\boldsymbol{X}_0$	(σ^2, ρ, ν)	6	2,651
	$f(\sigma^2 \mid \boldsymbol{X})$	N/A	σ^2	(ρ, ν)	74	2,043
	$f(\rho \mid \boldsymbol{X})$	N/A	ρ	(σ^2, ν)	74	2,043
	$f(\nu \mid \boldsymbol{X})$	N/A	ν	(σ^2, ρ)	74	2,043
RESP	$f(\boldsymbol{Y}_0 \mid \boldsymbol{Y})$	$f(\boldsymbol{Y}_0 \mid \boldsymbol{\theta}_2, \boldsymbol{Y})$	$\boldsymbol{Y}_0$	$\boldsymbol{\theta}^*$	118	9,086
	$f(\tilde{\boldsymbol{Y}}_0 \mid \boldsymbol{Y})$	$P(Y_0(\boldsymbol{s}, t) \in I(\boldsymbol{s}) \mid \boldsymbol{\theta}_2, \boldsymbol{Y})$	$\boldsymbol{Y}_0$	$\boldsymbol{\theta}^*$	118	9,086

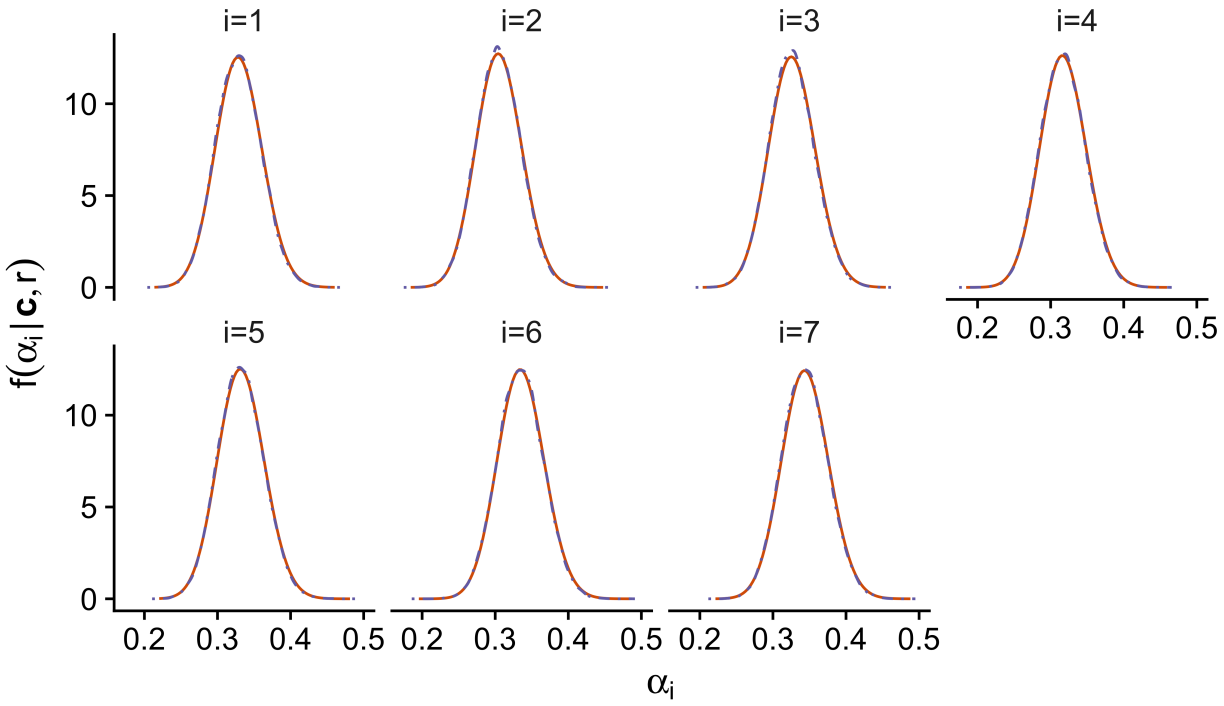

Figure 4.2: BISQuE (—) and Gibbs (- - -) approximations to the posterior densities $f(\alpha_i|\boldsymbol{c},r)$ are nearly identitcal.

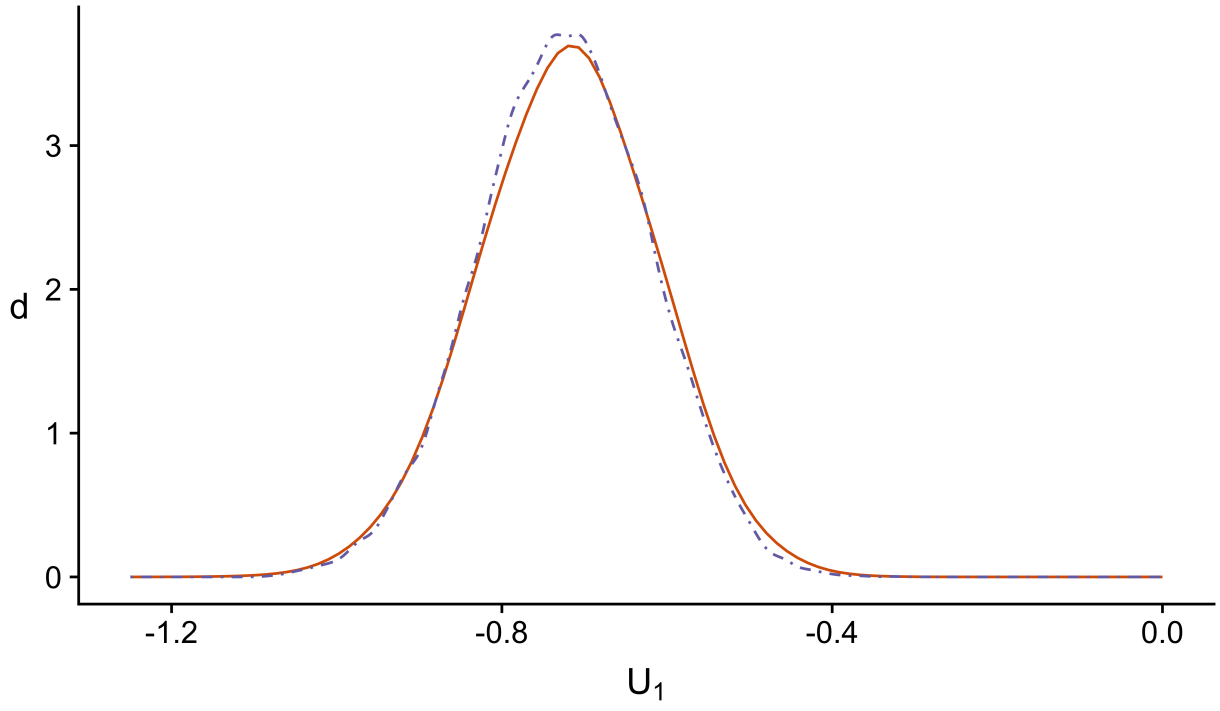

Figure 4.3: BISQuE (—) and Gibbs (- - -) approximations to the joint posterior density $f(U_1|\boldsymbol{c},r)$ are nearly identitcal.

4.5.2 Spatial

Simulated data and model

We work with data simulated from a geostatistical spatial model. Gibbs sampling is computationally expensive for such models because it involves decomposing spatially-structured covariance matrices in $\mathbb{R}^{N\times N}$ at each Gibbs iteration, where N is the number of observations. Let $\{X(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ be a random field, whose stochasticity is defined by a mean-zero Gaussian process on a continuous spatial domain $\mathscr{D}\subset\mathbb{R}^2$. Let the covariance $Cov(X(\boldsymbol{s}),X(\boldsymbol{t}))$ between random variates $X(\boldsymbol{s})$, $X(\boldsymbol{t})$ be specified by the isotropic Matérn covariance function, defined via

$$\kappa\big(\boldsymbol{s},\boldsymbol{t};\sigma^2,\rho,\nu\big)=\frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\big(\|\boldsymbol{s}-\boldsymbol{t}\|/\rho\big)^{\nu}K_{\nu}\big(\|\boldsymbol{s}-\boldsymbol{t}\|/\rho\big),$$

in which $\|\cdot\|$ is the Euclidean norm, K_ν is the modified Bessel function of the second kind with order $\nu>0$, which governs the smoothness of the process; $\sigma^2>0$ is a scaling parameter; and $\rho>0$ is a range parameter. Gaussian processes imply that the vector of observations $\boldsymbol{X}=(X(\boldsymbol{s}_1),\ldots,X(\boldsymbol{s}_N))^T\in\mathbb{R}^N$ at the finite collection of sampling locations $\mathscr{S}=\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_N\}\subset\mathscr{D}$ is normally distributed $\boldsymbol{X}\sim\mathscr{N}(\boldsymbol{0},\Sigma)$. The covariance matrix $\Sigma\in\mathbb{R}^{N\times N}$ is spatially-structured, with entries $\Sigma_{ij}=\kappa\big(\boldsymbol{s}_i,\boldsymbol{s}_j;\sigma^2,\rho,\nu\big)$. The Gaussian process assumption allows estimation of the field $\{X(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ at unobserved locations $\mathscr{S}_0=\{\boldsymbol{s}_{01},\ldots,\boldsymbol{s}_{0M}\}\subset\mathscr{D}$ via kriging, which uses conditional normal distributions for the unobserved responses. Standard Bayesian hierarchical modeling techniques for spatial data (e.g., Banerjee et al., 2015, Chapter 6) use conjugate or weakly informative priors for the covariance parameters, specified via

$$\begin{aligned}\sigma^2&\sim\text{Inverse-Gamma}(a,b),\\ \rho&\sim\text{Uniform}(L_0,U_0),\\ \nu&\sim\text{Uniform}(L_1,U_1).\end{aligned}$$

We simulate one dataset with $N = 300$ locations, sampled uniformly from the unit square $\mathscr{D} = [0,1]^2$ and with covariance parameters $\left(\sigma^2, \rho, \nu\right) = (1, .3, .5)$. We then estimate the covariance parameters as well as the field $\{X(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ at $M = 400$ unobserved, gridded locations $\mathscr{S}_0 \subset \mathscr{D}$. The priors are specified via $(a, b, L_0, U_0, L_1, U_1) = (2, 1, 0, 1, 0, 1)$.

Posterior inference and results

Standard techniques approximate posterior distributions with a Gibbs sampler and composition sampling (e.g., Banerjee et al., 2015, Chapter 6). Conjugate distributions are used to sample the scale σ^2 and unobserved field values $\boldsymbol{X}_0 = (X(\boldsymbol{s}_{01}), \ldots, X(\boldsymbol{s}_{0M})) \in \mathbb{R}^M$, but Metropolis steps are used for the range ρ and smoothness ν parameters. The Gibbs sampler is used to draw 60,000 posterior samples for the covariance parameters, taking 2,043 seconds to complete; posterior inference uses the final 30,000 iterations. After drawing posterior samples for the covariance parameters, composition sampling is used to draw samples for the unobserved field values $\boldsymbol{X}_0$ in parallel, taking 608 seconds to complete (Banerjee et al., 2015, pg. 126).

We use the BISQuE strategy to approximate the posterior density $f(\boldsymbol{X}_0 | \boldsymbol{X})$. Sparse grid quadrature techniques are used to directly approximate the marginal posterior covariance densities $f\left(\sigma^2 \middle| \boldsymbol{X}\right)$, $f\left(\rho \middle| \boldsymbol{X}\right)$, and $f(\nu | \boldsymbol{X})$. Table 4.2 connects this example's notation to that used with BISQuE. When used as the BISQuE conditioning variable $\boldsymbol{\theta}_2$, we map covariance parameters to the real line by log-transforming the scale parameter σ^2, and logit-transforming the range ρ and smoothness ν parameters. All conditional and marginal posterior densities required for BISQuE are computable in closed form up to a proportionality constant; refer to Banerjee et al. (2015, eqs. 2.15–16) for details.

Posterior inference via BISQuE and sparse grid quadrature is effectively identical to posterior inference via Gibbs sampling, but is computed with substantially less effort. Drawing posterior covariance parameter samples takes 2,043 seconds and composition sampling takes an additional 608 seconds, whereas the BISQuE and sparse grid quadrature approximations take a total of 238 seconds (Table 4.2), and posterior inference is nearly identical (Figures 4.4 to 4.6).

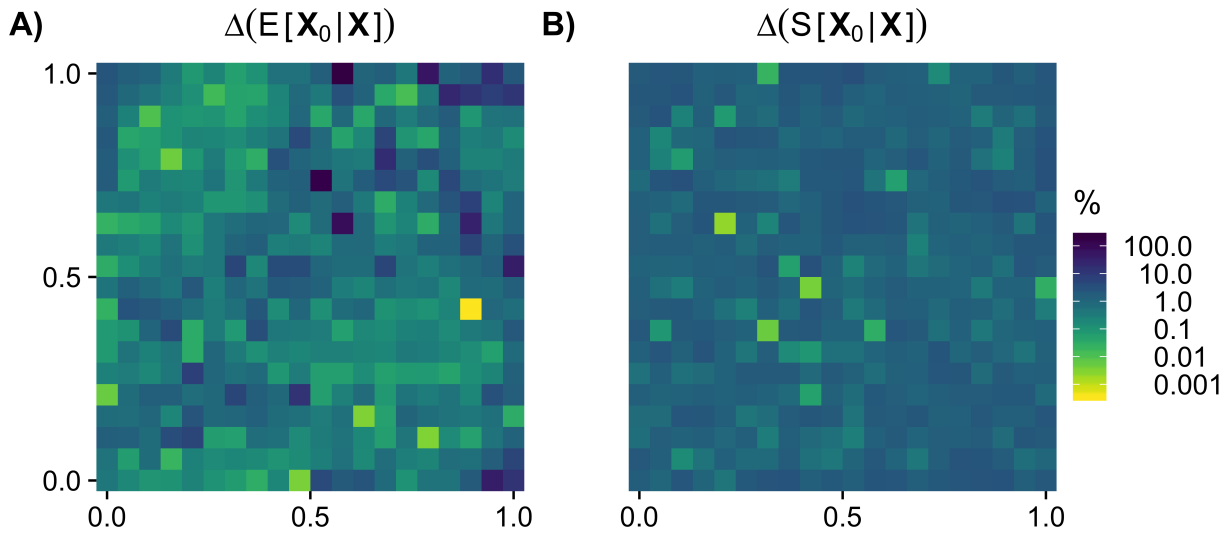

Figure 4.4: Relative differences $\Delta(Y) = \left(Y_{BISQuE} - Y_{Gibbs}\right)/Y_{Gibbs} \times 100\%$ between BISQuE and Gibbs approximations to the posterior predictive means (A) and standard errors (B) for the field $\{X(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathscr{D}}$ at unobserved locations $\mathscr{S}_0$. Nearly all (95%) relative differences in the posterior mean (A) are less than 5.5% (median=0.4%); relative differences in the mean are artificially large in regions where the posterior mean is near 0. All relative differences in the posterior standard errors (B) are below 3.3% (median=1.4%).

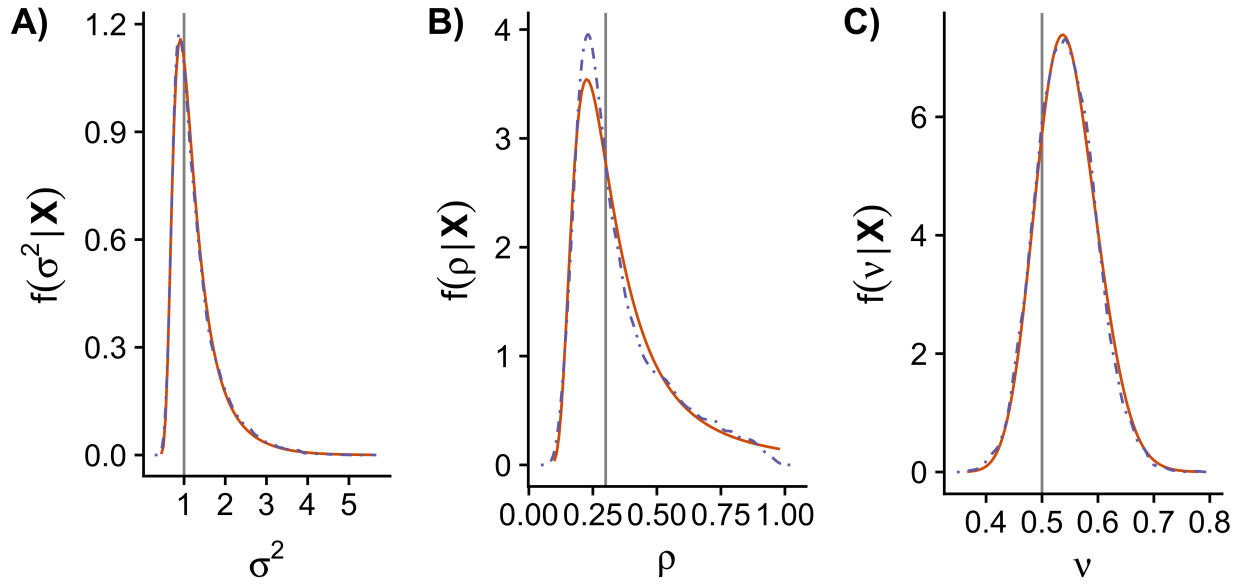

Figure 4.5: Sparse grid quadrature (—) and Gibbs (- - -) approximations to the posterior densities for the spatial covariance parameters $\left(\sigma^2, \rho, \nu\right)$ are nearly identitcal. The true values of the parameters are marked by grey vertical lines.

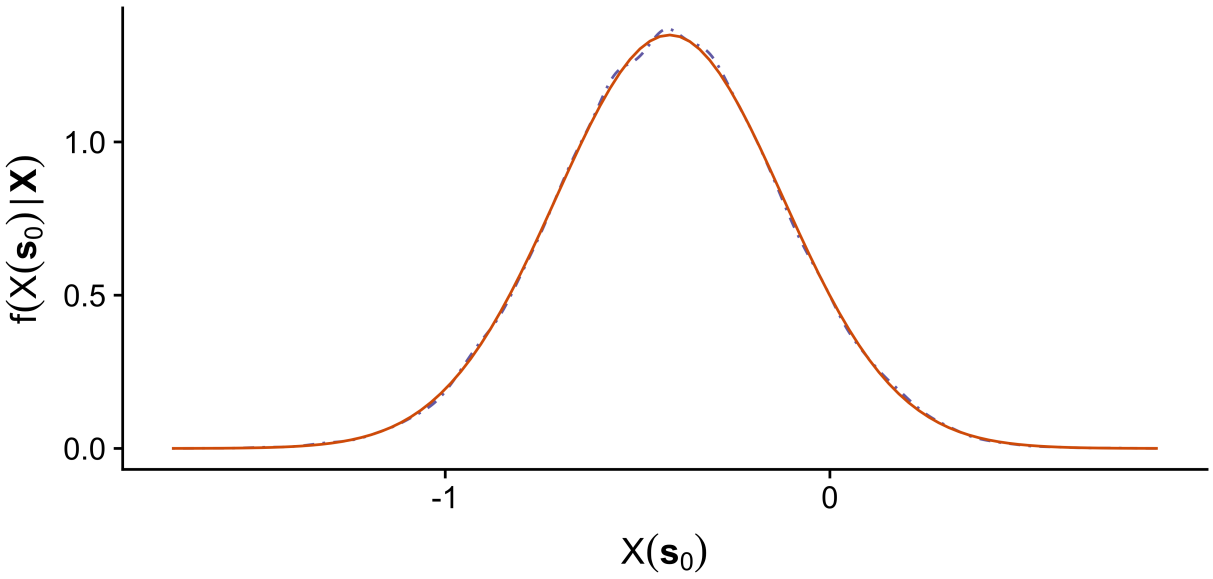

Figure 4.6: BISQuE (—) and Gibbs (···) approximations to the posterior density for $X(\boldsymbol{s}_0)$ is nearly identical at $\boldsymbol{s}_0 = (.5, .2)$, for example.

4.5.3 Remote effects spatial process models

Data and model

While most spatial data can be modeled with the assumption that distant points are uncorrelated, large-scale atmospheric circulations can induce dependence between fields separated by large distances. The resulting climate phenomena, known as teleconnection, may be modeled using remote effects spatial process (RESP) models, which can improve teleconnection-based predictions of seasonal precipitation (Hewitt et al., 2018). The RESP model is given by

$$Y(\boldsymbol{s}, t) = \boldsymbol{x}^T(\boldsymbol{s}, t)\boldsymbol{\beta} + w(\boldsymbol{s}, t) + \gamma(\boldsymbol{s}, t), \tag{4.24}$$

which uses a stochastic teleconnection term

$$\gamma(\boldsymbol{s}, t) = \int_{\mathscr{D}_Z} z(\boldsymbol{r}, t)\alpha(\boldsymbol{s}, \boldsymbol{r})d\boldsymbol{r} \tag{4.25}$$

to extend standard geostatistical regression models for a process $\{Y(\boldsymbol{s}, t) : \boldsymbol{s} \in \mathscr{D}_Y, t \in \mathscr{T}\}$ defined on a continuous spatial domain $\mathscr{D}_Y$ for discrete times $\mathscr{T}$. Regression coefficients $\boldsymbol{\beta}$ and

spatially-correlated variation $w(\boldsymbol{s}, t)$ are augmented by (4.25), which uses doubly-indexed random effects $\alpha(\boldsymbol{s}, \boldsymbol{r})$ to aggregate the impact of remote covariates $\{z(\boldsymbol{r}, t) : \boldsymbol{r} \in \mathscr{D}_Z, t \in \mathscr{T}\}$, such as sea surface temperatures, on a distant response, such as the standardized deviation $Y(\boldsymbol{s}, t)$ from mean seasonal precipitation. The authors adopt the climate science convention that mean precipitation is treated as known, and the standardized deviation $Y(\boldsymbol{s}, t)$ is the scientifically interesting response variable to model.

The RESP model uses two Matérn covariances $\kappa\big(\boldsymbol{s}, \boldsymbol{s}'; \sigma_w^2, \rho_w, \nu_w\big)$, $\kappa\big(\boldsymbol{r}, \boldsymbol{r}'; \sigma_\alpha^2, \rho_\alpha, \nu_\alpha\big)$, and a nugget effect σ_ε^2 to define Gaussian processes that model the spatial variation $\{w(\boldsymbol{s}, t) : \boldsymbol{s} \in \mathscr{D}_Y\}$ and teleconnection effects $\{\alpha(\boldsymbol{s}, \boldsymbol{r}) : \boldsymbol{s} \in \mathscr{D}_Y, \boldsymbol{r} \in \mathscr{D}_Z\}$. The Matérn smoothness parameters ν_w and ν_α are treated as fixed, and standard priors are used to model the remaining regression coefficients $\boldsymbol{\beta}$ and covariance parameters σ_w^2, ρ_w, σ_ε^2, σ_α^2, and ρ_α (cf. Section 4.5.2).

We follow Hewitt et al. (2018) and use the RESP model to analyze Colorado precipitation data in a statistical downscaling-like scenario. The RESP model regresses standardized deviations $Y(\boldsymbol{s}, t)$ from mean Colorado precipitation observed at 240 locations $\boldsymbol{s} \in \mathscr{D}_Y$ onto local surface temperatures $\boldsymbol{x}(\boldsymbol{s}, t)$ and Pacific Ocean sea surface temperatures $z(\boldsymbol{r}, t)$. The model is fit to Winter averages from 1981–2012 and an ordinal response $\tilde{Y}(\boldsymbol{s}, t) \in \{v_1, \ldots, v_m\}$ is predicted for Winter 2013, given the covariate values $\boldsymbol{x}(\boldsymbol{s}, t)$ and $z(\boldsymbol{r}, t)$ for $t = 2013$. The distribution for the ordinal responses $\tilde{Y}(\boldsymbol{s}, t)$ is induced by known cut points $c_0(\boldsymbol{s}), \ldots, c_m(\boldsymbol{s})$ and defined such that $P(\tilde{Y}(\boldsymbol{s}, t) = v_i) = P(c_{i-1}(\boldsymbol{s}) < Y(\boldsymbol{s}, t) < c_i(\boldsymbol{s}))$. In this application, the ordinal response $\tilde{Y}(\boldsymbol{s}, t)$ represents below average v_1, about average v_2, or above average precipitation v_3.

Posterior inference and results

Hewitt et al. (2018) construct a Gibbs sampler that approximates posterior distributions for the RESP model (4.24). Gibbs sampling is computationally expensive for the RESP model because two spatially-structured covariance matrices must be decomposed at each Gibbs iteration. Let $\boldsymbol{Y}$ denote all observations $Y(\boldsymbol{s}, t)$ from $t = 1981, \ldots, 2012$; $\boldsymbol{Y}_0$ denote all unobserved responses $Y(\boldsymbol{s}, t)$ at $t = 2013$; and $\tilde{\boldsymbol{Y}}_0$ denote all unobserved ordinal responses $\tilde{Y}(\boldsymbol{s}, t)$ at $t = 2013$. Conjugate distributions are used to sample the regression parameters $\boldsymbol{\beta}$, scales σ_w^2 and σ_α^2, and

continuous predictions $\boldsymbol{Y}_0$; and Metropolis steps are used for the ranges ρ_w and ρ_α. The Gibbs sampler is used to draw 41,000 posterior samples for the regression and covariance parameters, taking 8,331 seconds to complete; posterior inference discards the first 1,000 iterations as the chain mixes quickly, but requires many iterations to control Monte Carlo integration error. Composition sampling is then used to draw samples for the predicted response $\boldsymbol{Y}_0$ in parallel, taking 755 seconds to complete. The continuous posterior predictive density $f(\boldsymbol{Y}_0|\boldsymbol{Y})$ is discretized after sampling to approximate $f(\tilde{\boldsymbol{Y}}_0|\boldsymbol{Y})$ by using the empirical quantiles of historical precipitation as cut points $c_0(\boldsymbol{s}),\ldots,c_3(\boldsymbol{s})$.

We use the BISQuE strategy to approximate the posterior predictive densities $f(\boldsymbol{Y}_0|\boldsymbol{Y})$ and $f(\tilde{\boldsymbol{Y}}_0|\boldsymbol{Y})$. In particular, we use the BISQuE strategy to directly approximate $f(\tilde{\boldsymbol{Y}}_0|\boldsymbol{Y})$ by letting $h(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2;\boldsymbol{X})$ in (4.13) be the conditional cumulative distribution function for $\boldsymbol{Y}_0$. Table 4.2 connects this example's notation to that used with BISQuE. When used as the BISQuE conditioning variable $\boldsymbol{\theta}_2$, we map covariance parameters to the real line by log-transforming scale parameters σ^2 and logit-transforming range parameters ρ. All conditional and marginal posterior densities required for BISQuE are computable in closed form up to a proportionality constant; refer to Hewitt et al. (2018) for distributional results.

Posterior inference via BISQuE is effectively identical to posterior inference via Gibbs sampling, but is computed with substantially less effort. Drawing posterior covariance parameter samples takes 8,331 seconds and composition sampling takes an additional 755 seconds, whereas the BISQuE approximations take a total of 118 seconds (Table 4.2), and posterior inference is nearly identical (e.g., Figure 4.7). The approximate BISQuE and Gibbs posterior masses $\hat{P}(\tilde{Y}_0(\boldsymbol{s},t)=v_i|\boldsymbol{Y})$ agree to at least two decimal places for all 240 locations $\boldsymbol{s}\in\mathscr{D}_Y$ and values v_1,v_2,v_3; additional computing effort can further reduce approximation errors, but offers limited practical benefit because the discretization is coarse.

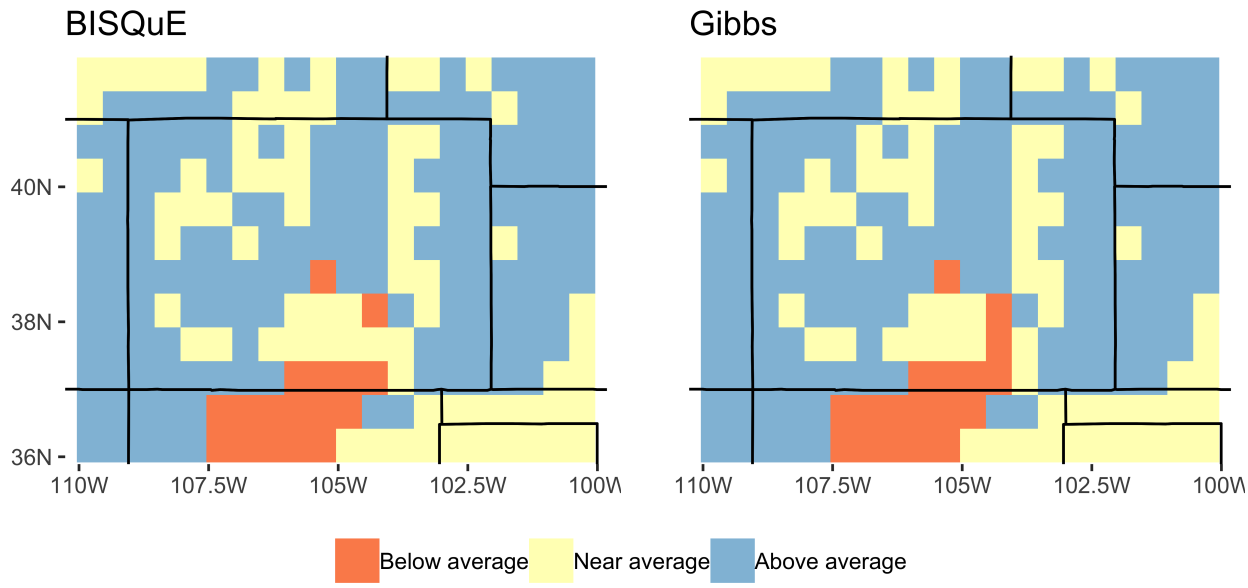

Figure 4.7: BISQuE and Gibbs approximations to the mode of the discretized posterior predictive distributions $f(\tilde{\boldsymbol{Y}}_0|\boldsymbol{Y})$ are nearly identical.

4.6 Discussion

We combine conditioning with sparse grid quadrature rules to approximate Bayesian Inference via Sparse grid Quadrature Evaluation (BISQuE). Approximations (4.13) are developed by reformulating Bayesian posterior quantities, such as densities and expectations, so that they may be approximated as weighted mixtures of conditional quantities $h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X})$. The integration nodes and weights from sparse grid quadrature rules are used to build mixing weights $w^{(i,\ell)}$ and conditioning values $\boldsymbol{\theta}_2^{(i,\ell)}$. In a similar manner as general quadrature techniques and importance sampling methods, the final BISQuE approximation weights $\tilde{w}^{(i,\ell)}$ use weight ratios $f(\boldsymbol{v}^{(i,\ell)}|\boldsymbol{X}) / w(\boldsymbol{v}^{(i,\ell)}, \boldsymbol{X})$ to align the "theoretical distribution" $f(\boldsymbol{v}|\boldsymbol{X})$ with the "sampling distribution" $w(\boldsymbol{v}, \boldsymbol{X})$ (Givens and Hoeting, 2013, pgs. 143, 181). Nested integration strategies can help compute BISQuE approximations (4.13) when models do not have closed form expressions for required components (Section 4.4.4). Posterior approximation via BISQuE is deterministic and computationally efficient, offering faster computation than MCMC methods for a wide range of models (4.5) and posterior quantities (4.8). In our applications, we find that

BISQuE often reduces overall computing time by two orders of magnitude and yields nearly identical inference to standard MCMC approaches (Section 4.5).

The BISQuE approximation is similar to, and can be combined with Integrated Nested Laplace approximations (INLA) for latent Gaussian models (Rue et al., 2009). Combining the BISQuE approximation with INLA can yield an approximation technique that scales better to models with more hyperparameters. Similar to INLA, our framework will be most efficient when used to approximate low-dimensional posterior quantities, like marginal densities or joint densities with computationally tractable closed form expressions (e.g., $f(\boldsymbol{X}_0|\boldsymbol{X})$ in Section 4.5.2). However, BISQUE does not require that a model have a latent Gaussian structure and is thus applicable to a broad class of models such as the population estimation model of Section 4.5.1.

We can combine the BISQuE approximation (4.13) and INLA because both methods use conditioning and integration grids to yield fast deterministic posterior approximation. In terms of the general hierarchical model (4.5), INLA specifies a hierarchical parameter model such that $f(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)f(\boldsymbol{\theta}_2)$ in which $f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ is Gaussian and $f(\boldsymbol{\theta}_2)$ is a prior distribution for relatively low-dimensional hyperparameters $\boldsymbol{\theta}_2$. Rue et al. (2009) define $\boldsymbol{\theta}_1 = (\theta_{11},\ldots,\theta_{1i},\ldots,\theta_{1n})$, develop an integration grid, and use Laplace approximations for $f(\boldsymbol{\theta}_{1i}|\boldsymbol{\theta}_2,\boldsymbol{X})$ and $f(\boldsymbol{\theta}_2|\boldsymbol{X})$ to approximate the marginal posterior density $f(\theta_{1i}|\boldsymbol{X})$. The nested Laplace approximations can be embedded in the BISQuE approximation (4.13), yielding posterior approximation that uses an alternate integration grid to INLA. The embedding can be beneficial because sparse grid quadrature rules allow for more computationally efficient approximation in models with higher dimensional hyperparameters $\boldsymbol{\theta}_2$. Specifically, Rue et al. (2009) suggest creating integration grids for models with high-dimensional $\boldsymbol{\theta}_2$ by using central composite design (CCD) methods—an experimental design and response surface technique for approximating second order surfaces with relatively few function evaluations (Box and Wilson, 1951). When integration is the main concern, sparse grid quadrature methods can require substantially fewer integration nodes in high dimensions (Novak and Ritter, 1999, Table 2, $\ell = 3$) than CCD-based grids (Sanchez and Sanchez, 2005, Table 3).

Our BISQuE approximation advances Bayesian computing for hierarchical models, but open questions remain for wider application of the method. Notably, our approximation requires the ability to evaluate $h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X})$ quickly, so may often be limited to marginal posterior inference for $\boldsymbol{\theta}_1$ with relatively small dimension. Our approximation also relies on the availability of nested quadrature rules for $\boldsymbol{\theta}_2$. It is difficult to develop quadrature rules for discrete variables, thus practical use of our approximation may be limited to models with parameters $\boldsymbol{\theta}_2$ defined on continuous spaces Ω_2. Fast convergence of our approximation also relies on the availability of accurate approximations to $f(\boldsymbol{\theta}_2|\boldsymbol{X})$. If the BISQuE approximation (4.13) has not converged, intuition about numerical integration suggests the resulting approximation will likely underestimate posterior variability (Rue et al., 2009). However, Rue et al. (2009, Section 6.5) also point out that $f(\boldsymbol{\theta}_2|\boldsymbol{X})$ often becomes increasingly Gaussian as the dimension of $\boldsymbol{\theta}_2$ grows since the Bayesian structure will increase variability and regularity will the dimension, which will help accelerate convergence.

The BISQUE methodology suggests continued develop in several areas. More thorough diagnostics should also be developed for wider practical application of the BISQuE approximation (4.13). The approximation's convergence can be monitored by checking the approximation's stability as the level q of the underlying sparse grid quadrature rule (4.4) is increased (Laurie, 1985). However, this does not necessarily provide a diagnostic that can assess how well conditioned a model (4.5) or posterior quantity (4.8) is for use with BISQuE. Drawing from importance sampling, studying the weight ratio $f(\boldsymbol{\nu}^{(i,\ell)}|\boldsymbol{X})/w(\boldsymbol{\nu}^{(i,\ell)}, \boldsymbol{X})$ in (4.13) at quadrature nodes $\boldsymbol{\nu}^{(i,\ell)}$ may help diagnose practical issues. Theoretical smoothness properties of $h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{X})$ or concentration of the posterior density $f(\boldsymbol{\theta}_2|\boldsymbol{X})$ may also provide insight into the conditioning for specific models.

Chapter 5

Conclusions and Future work

The research in this thesis improves geostatistical models to better account for dependence induced by both local and global-scale phenomena that impact climate data. This thesis also proposes a computational method to make estimation and prediction for Bayesian models more computationally efficient. Advancing statistical models and computing can improve predictions of future climate to allow policy makers, resource managers, engineers, and other consumers of climate forecasts to make more informed decisions about how to manage risk of varied impacts on communities, water resources, and agriculture. I briefly discuss directions for continuing the research in this dissertation.

5.1 Improving efficiency of weighted likelihood latent spatial extremes models

Chapter 2 proposes a weighted likelihood that improves coverage for estimates of marginal return levels from latent spatial extremes models applied to data with extremal dependence. The weighted likelihood correction allows latent spatial extremes models to be applied to data that violate the model's underlying conditional independence assumption, but uncertainty is still underestimated when extremal dependence is strong. Proposing alternate weight functions can improve this issue. The likelihood weights in Chapter 2 are based on the extremal coefficient, which is an exploratory measure of pairwise extremal dependence. Unlike dependence in spatially-correlated data arising from Gaussian processes, dependence in spatially correlated extremes data is not uniquely characterized by pairwise dependence functions. Defining a weight function that uses higher-order exploratory measures or borrows information from composite likelihoods can allow more accurate measurement of return level uncertainty.

The likelihood weights could also be extended via non-stationary weight functions. The weighted likelihood latent spatial extremes modeling approach is attractive because the latent Gaussian processes employed can scale to large spatial datasets with data collected from thousands of monitoring stations. However, the proposed weight function assumes extremal dependence is stationary across space. As the spatial domain analyzed grows, stationarity becomes harder to justify. Non-stationary weight functions could be developed from "local" versions of the extremal coefficient. If weight functions can be developed from composite likelihoods for stationary extremes processes, then it may be possible to extend such methods using composite likelihoods for non-stationary extremes processes, such as those proposed in Huser and Genton (2016).

5.2 Extending RESP models for non-Gaussian data and broader application

Chapter 3 proposes a remote effects spatial process (RESP) model for spatially-correlated climate data impacted by teleconnection effects, which requires modeling dependence at both long and short distances. The model is developed for normally distributed data, but should be extended for analysis of non-Gaussian data and temporally-varying teleconnection effects. Mason and Goddard (2001) find that teleconnection effects may change over time, and annual counts of tornadoes and large storms have been linked to teleconnection effects (Timm et al., 2011; Wikle and Anderson, 2003). Temporally-varying teleconnection effects may be modeled using methods for modeling dynamical spatio-temporal models (Cressie and Wikle, 2011). Modeling non-Gaussian data is more complicated.

While it is easy to formulate the RESP model as a generalized linear model (GLM) for non-Gaussian data influenced by spatial random effects, estimation is computationally challenging. Posterior inference via integrated nested Laplace approximations (INLA) is possible, but is difficult to implement and effectively requires the RESP model to replace its latent Gaussian processes with Gaussian Markov Random Fields (GMRF, Rue and Held, 2005). Markov chain Monte

Carlo (MCMC) is often used for estimating spatial GLMs, but the MCMC samplers can have slow convergence properties since conjugacy of the spatial effects is lost (Diggle et al., 1998; Higgs and Hoeting, 2010). Computationally efficient MCMC samplers may be feasible by combining stochastic partial differential equation (SPDE) approximations to Gaussian processes with one-block Metropolis-Hastings proposals (Lindgren et al., 2011; Rue and Held, 2005). The SPDE approximation to Gaussian processes yields a GMRF proposal distribution for the MCMC sampler, for which it is easy to sample the entire latent field. All model parameters may be updated in a single Gibbs sampling block using a Metropolis-Hastings step to jointly accept or reject the GMRF-based proposals.

5.3 Further development and application of weighted mixtures approximations

This thesis proposes a method for weighted mixtures approximations to posterior distributions, in particular using sparse-grid quadrature methods. To further demonstrate the wide applicability of the methodology proposed in Chapter 4, I plan to apply it to a wider set of hierarchical models. Application is especially important because detailed comparisons between MCMC methods and product-grid quadrature methods are difficult to make using theory alone. Theoretical comparisons can be difficult to make because the theoretical accuracy for each technique is developed from different assumptions. For example, the accuracy of sparse-grid quadrature methods and product-grid quadrature methods is studied for integrals of different classes of integrands. Additionally, quadrature approximations are deterministic, while MCMC approximations are stochastic.

The weighted mixtures approximation can be studied further to develop diagnostics and refined theory for models that do not have closed form expressions for the weighting density $f(\boldsymbol{\theta}_2|\boldsymbol{X})$. This thesis proposes one solution for such models, but implicit assumptions and potential limitations are not thoroughly explored. Similarly, diagnostics can be developed to help identify when the weighted mixture approximation is inaccurate. Diagnostics for weighted mix-

ture approximations may be able to be adapted from importance sampling techniques. The weighted mixtures approximation uses weight ratios, which also appear in importance samplers (Givens and Hoeting, 2013, Section 6.3.1). Diagnostics for importance samplers look for outliers in histograms of weight ratios because importance samplers are known to perform poorly when the empirical distribution of sampling weights includes several very large weights. Weight ratio diagnostics could help identify if the weighting density $f(\boldsymbol{\theta}_2|\boldsymbol{X})$ is not well approximated by a Gaussian density. The weighted mixture approximation also requires the integrand $h(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2;\boldsymbol{X})$ to be smooth relative to $\boldsymbol{\theta}_2$. Empirical coefficients of variation may provide a simple way to study the relative smoothness of $h(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2;\boldsymbol{X})$ with respect to $\boldsymbol{\theta}_2$.

References

Adler, R. J. and Taylor, J. E. (2007). *Random Fields and Geometry*. Springer Science + Business Media, LLC, New York.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18:343–373.

Arasaratnam, I. and Haykin, S. (2009). Cubature Kalman Filters. *IEEE Transactions on Automatic Control*, 54(6):1254–1269.

Ashok, K., Behera, S. K., Rao, S. A., Weng, H., and Yamagata, T. (2007). El Nino Modoki and its possible teleconnection. *Journal of Geophysical Research*, 112:1–27.

Assunção, R. and Krainski, E. (2009). Neighborhood Dependence in Bayesian Spatial Models. *Biometrical Journal*, 51(5):851–869.

Attias, H. (2000). A Variational Bayesian Framework for Graphical Models. *Advances in neural information processing systems*, pages 209–215.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, Boca Raton, FL, second edition.

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: Statistical Methodology*, 70(4):825–848.

Banerjee, S. and Roy, A. (2014). *Linear Algebra and Matrix Analysis for Statistics*. CRC Press, Boca Raton, FL.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Science+Business Media, LLC, New York, second ed. edition.

Bolin, D. and Lindgren, F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:85–106.

Box, G. E. P. and Wilson, K. B. (1951). On the Experimental Attainment of Optimum Conditions. *Journal of the Royal Statistical Society Series B*, 13(1):1–45.

Brierley, A. S., Demer, D. A., Watkins, J. L., and Hewitt, R. P. (1999). Concordance of interannual fluctuations in acoustically estimated densities of Antarctic krill around South Georgia and Elephant Island: Biological evidence of same-year teleconnections across the Scotia Sea. *Marine Biology*, 134:675–681.

Brooks, S. P. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Brown, B. M. and Resnick, S. I. (1977). Extreme values of independent stochastic processes. *Journal of Applied Probability*, 14(4):732–739.

Bruyere, C. L., Holland, G. J., and Towler, E. (2012). Investigating the Use of a Genesis Potential Index for Tropical Cyclones in the North Atlantic Basin. *Journal of Climate*, 25:8611–8626.

Burkardt, J. (2007). Sparse Grid Mixed Growth Anisotropic Rules. *https://people.sc.fsu.edu/~jburkardt/cpp_src/sgmga*.

Calder, C. A., Craigmile, P. F., and Mosley-Thompson, E. (2008). Spatial variation in the influence of the North Atlantic Oscillation on precipitation across Greenland. *Journal of Geophysical Research*, 113.

Cao, Y. and Li, B. (2018). Assessing models for estimation and methods for uncertainty quantification for spatial return levels. *Environmetrics*.

Castruccio, S., Huser, R., and Genton, M. G. (2016). High-Order Composite Likelihood Inference for Max-Stable Distributions and Processes. *Journal of Computational and Graphical Statistics*, 25(4):1212–1229.

Choi, I., Li, B., Zhang, H., and Li, Y. (2015). Modelling space-time varying ENSO teleconnections to droughts in North America. *Stat*, 4(1):140–156.

Coles, S. G. and Dixon, M. J. (1999). Likelihood-Based Inference for Extreme Value Models. *Extremes*, 2(1):5–23.

Cooley, D., Naveau, P., and Poncet, P. (2006). Variograms for spatial max-stable random fields. In Bertail, P., Doukhan, P., and Soulier, P., editors, *Dependence in Probability and Statistics*, pages 373–390. Springer Science+Business Media, LLC, New York, NY.

Cooley, D., Nychka, D., and Naveau, P. (2007). Bayesian Spatial Modeling of Extreme Precipitation Return Levels. *Journal of the American Statistical Association*, 102(479):824–840.

Cooley, D. and Sain, S. R. (2010). Spatial hierarchical modeling of precipitation extremes from a regional climate model. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(3):381–402.

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

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons, Inc., Hoboken, NJ, revised edition.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J., and Pasteris, P. P. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, 28(15):2031–2064.

Datta, A., Banerjee, S., Finley, A., and Gelfand, A. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012). Statistical Modeling of Spatial Extremes. *Statistical Science*, 27(2):161–186.

De Haan, L. (1984). A Spectral Representation for Max-stable Processes. *The Annals of Probability*, 12(4):1194–1204.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P.,

Tavolato, C., Thépaut, J. N., and Vitart, F. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597.

Deser, C., Phillips, A., Bourdette, V., and Teng, H. (2012). Uncertainty in climate change projections : the role of internal variability. *Climate Dynamics*, 38:527–546.

Dhaene, J., Denuit, M., Goovaerts, M. J., Kaas, R., and Vyncke, D. (2002). The concept of comonotonicity in actuarial science and finance : theory. *Insurance: Mathematics and Economics*, 31:3–33.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-Based Geostatistics. *Journal of the Royal Statistical Society, Series C*, 47(3):299–350.

Dong, B. and Dai, A. (2015). The influence of the Interdecadal Pacific Oscillation on Temperature and Precipitation over the Globe. *Climate Dynamics*, 45:2667–2681.

Emery, A. F. and Johnson, K. C. (2012). Practical considerations when using sparse grids with Bayesian inference for parameter estimation. *Inverse Problems in Science and Engineering*, 20(5):591–608.

Epstein, E. S. (1969). A Scoring System for Probability Forecasts of Ranked Categories. *Journal of Applied Meteorology*, 8:985–987.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chan Chou, S., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M. (2013). Evaluation of Climate Models 9. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, chapter 9, pages 741–882. Cambridge University Press, Cambridge, United Kingdom.

Fowler, H. J., Blenkinsop, S., and Tebaldi, C. (2007). Linking climate change modelling to impacts studies : recent advances in downscaling techniques for hydrological. *International Journal of Climatology*, 27:1547–1578.

French, J. P. and Hoeting, J. A. (2016). Credible regions for exceedance sets of geostatistical data. *Environmetrics*, 27:4–14.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–511.

Genz, A. and Keister, B. D. (1996). Fully Symmetric Interpolatory Rules for Multiple Integrals over Infinite Regions with Gaussian Weight. *J. Comp. Appl. Math.*, 71:299–309.

Gerstner, T. and Griebel, M. (1998). Numerical Integration Using Sparse Grids. *Numerical Algorithms*, 18:209–232.

Givens, G. H. and Hoeting, J. A. (2013). *Computational Statistics*. John Wiley & Sons, Inc., Hoboken, NJ, second edition.

Gneiting, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19(4):1327–1349.

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goddard, L., Mason, S. J., Zebiak, S. E., Ropelewski, C. F., Basher, R., and Cane, M. A. (2001). Current approaches to seasonal-to-interannual climate predictions. *International Journal of Climatology*, 21:1111–1152.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Heiss, F. and Winschel, V. (2008). Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics*, 144(1):62–80.

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, 15:559–570.

Hewitt, J., Hoeting, J. A., Done, J. M., and Towler, E. (2018). Remote effects spatial process models for modeling teleconnections. *Environmetrics*.

Higgs, M. D. and Hoeting, J. A. (2010). A clipped latent variable model for spatially correlated ordered categorical data. *Computational Statistics and Data Analysis*, 54(8):1999–2011.

Hu, F. and Zidek, J. V. (2002). The weighted likelihood. *The Canadian Journal of Statistics*, 30(3):347–371.

Huser, R. and Genton, M. G. (2016). Non-Stationary Dependence Structures for Spatial Extremes. *Journal of Agricultural, Biological, and Environmental Statistics*, 21(3):470–491.

Jia, B., Xin, M., and Cheng, Y. (2012). Sparse-grid quadrature nonlinear filtering âIJl'. *Automatica*, 48:327–341.

Kabluchko, Z., Schlather, M., and de Haan, L. (2009). Stationary Max-Stable Fields Associated to Negative Definite Functions. *The Annals of Probability*, 37(5):2042–2065.

Karl, T. R., Wang, W.-C., Schlesinger, M. E., Knight, R. W., and Portman, D. (1990). A Method of Relating General Circulation Model Simulated Climate to the Observed Local Climate. Part I: Seasonal Statistics. *Journal of Climate*, 3:1053–1079.

Karr, T. W. and Wooten, R. L. (1976). Summer Radar Echo Distribution Around Limon, Colorado. *Monthly Weather Review*, 104:728–734.

Katzfuss, M. (2016). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112(517):201–214.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and Vertenstein, M. (2015). The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability. *Bulletin of the American Meteorological Society*, 96(8):1333–1349.

Laurie, D. P. (1985). Practical error estimation in numerical integration. *Journal of Computational and Applied Mathematics*, 12:425–431.

Lehmann, E. A., Phatak, A., Stephenson, A., and Lau, R. (2016). Spatial modelling framework for the characterisation of rainfall extremes at different durations and under climate change. *Environmetrics*, 27(4):239–251.

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: Statistical Methodology*, 73(4):423–498.

Long, Q., Scavino, M., Tempone, R., and Wang, S. (2013). Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, 259:24–39.

Lukas, J., Barsugli, J., Doesken, N., Rangwala, I., and Wolter, K. (2014). *Climate Change in Colorado*. University of Colorado Boulder, second edition.

MacEachern, S. N. and Berliner, L. M. (1994). Subsampling the Gibbs Sampler. *The American Statistician*, 48(3):188–190.

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., and Francis, R. C. (1997). A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production. *Bulletin of the American Meteorological Society*, 78(6):1069–1079.

Maraun, D., Wetterhall, F., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themeßl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., and Thiele-Eich, I. (2010). Precipitation downscaling under climate change: Recent developements to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(RG3003).

Martins, E. S. and Stedinger, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, 36(3):737–744.

Mason, S. J. (2012). Seasonal and longer-range forecasts. In Jolliffe, I. T. and Stephenson, D. B., editors, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, chapter 11, pages 204–220. John Wiley & Sons, Ltd., Oxford, second edition.

Mason, S. J. and Goddard, L. (2001). Probabilistic precipitation anomalies associated with ENSO. *Bulletin of the American Meteorological Society*, 82:619–638.

McDermott, P. L. and Wikle, C. K. (2016). A model-based approach for analog spatio-temporal dynamic forecasting. *Environmetrics*, 27:70–82.

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., Corti, S., Danabasoglu, G., Doblas-Reyes, F., Hawkins, E., Karspeck, A., Kimoto, M., Kumar, A., Matei, D., Mignot, J., Msadek, R., Navarra, A., Pohlmann, H., Rienecker, M., Rosati, T., Schneider, E., Smith, D., Sutton, R., Teng, H., Van Oldenborgh, G. J., Vecchi, G., and Yeager, S. (2014). Decadal climate prediction: An update from the trenches. *Bulletin of the American Meteorological Society*, 95(2):243–267.

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M. A., Greene, A. M., Hawkins, E., Hegerl, G., Karoly, D., Keenlyside, N., Kimoto, M., Kirtman, B., Navarra, A., Pulwarty, R., Smith, D., Stammer, D., and Stockdale, T. (2009). Decadal Prediction. *Bulletin of the American Meteorological Society*, 90(10):1467–1485.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29(7):897–910.

Montroy, D. (1997). Linear Relation of Central and Eastern North American Precipitation to Tropical Pacific Sea Surface Temperature Anomalies. *Journal of Climate*, 10:541–558.

Montroy, D., Richman, M. B., and Lamb, P. J. (1998). Observed Nonlinearities of Monthly Teleconnections between Tropical Pacific Sea Surface Temperature Anomalies and Central and Eastern North American Precipitation. *Journal of Climate*, 11:1812–1835.

Murphy, A. H. (1971). A Note on the Ranked Probability Score. *Journal of Applied Meteorology*, 10:155–156.

Naylor, J. C. and Smith, A. F. M. (1982). Applications of a Method for the Efficient Computation of Posterior Distributions. *Journal of the Royal Statistical Society, Series C*, 31(3):214–225.

Newton, M. A. and Raferty, A. E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society Series B*, 56(1):3–48.

Nigam, S. and Baxter, S. (2015). Teleconnections. In *Encyclopedia of Atmospheric Sciences 2nd Edition*, volume 3, pages 90–109. Elsevier Ltd., second edition.

Novak, E. and Ritter, K. (1996). High dimensional integration of smooth functions over cubes. *Numerische Mathematik*, 75:79–97.

Novak, E. and Ritter, K. (1999). Simple Cubature Formulas with High Polynomial Exactness. *Constructive Approximation*, 15:499–522.

Opitz, T., Huser, R., Bakka, H., and Rue, H. (2018). INLA goes extreme : Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes*.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506.

Padoan, S. A., Ribatet, M., and Sisson, S. A. (2010). Likelihood-Based Inference for Max-Stable Processes. *Journal of the American Statistical Association*, 105(489):263–277.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Parker, D. (2015). Mesoscale Meteorology. In *Encyclopedia of Atmospheric Sciences 2nd Edition*, volume 3, pages 316–322. Elsevier Ltd.

Parzen, E. (1979). Nonparametric Statistical Data Modeling. *Journal of the American Statistical Association*, 74(365):105–121.

Pébay, P. (2008). Formulas for Robust, One-Pass Parallel Computation of Covariances and Arbitrary-Order Statistical Moments. *Sandia Report SAND2008-6212, Sandia National Laboratories*, 94.

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(December):1197–1206.

Reich, B. J. and Shaby, B. A. (2012). A Hierarchical Max-Stable Spatial Model for Extreme Precipitation. *Annals of Applied Statistics*, 6(4):1430–1451.

Ribatet, M., Cooley, D., and Davision, A. S. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22(2):813–845.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367.

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12(4):1151–1172.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, Boca Raton, FL.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society Series B*, 71(2):319–392.

Sanchez, S. M. and Sanchez, P. J. (2005). Very large fractional factorials and central composite designs. *ACM Transactions on Modeling and Computer Simulation*, 15(4):362–377.

Sang, H. and Gelfand, A. E. (2009). Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics*, 16(3):407–426.

Schillings, C. and Schwab, C. (2013). Sparse, adaptive Smolyak quadratures for Bayesian inverse problems. *Inverse Problems*, 29.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes*, 5(1):33–44.

Schlather, M. and Tawn, J. A. (2003). A Dependence Measure for Multivariate and Spatial Extreme Values: Properties and Inference. *Biometrika*, 90(1):139–156.

Schliep, E. M., Cooley, D., Sain, S. R., and Hoeting, J. A. (2010). A comparison study of extreme precipitation from six different regional climate models via spatial hierarchical modeling. *Extremes*, 13:219–239.

Seto, K. C., Reenberg, A., Boone, C. G., Fragkias, M., Haase, D., Langanke, T., Marcotullio, P., Munroe, D. K., Olah, B., and Simon, D. (2012). Urban land teleconnections and sustainability. *Proceedings of the National Academy of Sciences*, 109(20):7687–7692.

Sharkey, P. and Winter, H. C. (2018). A Bayesian spatial hierarchical model for extreme precipitation in Great Britain. *Environmetrics*.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising Model Component Complexity : A Principled , Practical Approach to Constructing Priors. *Statistical Science*, 32(1):1–28.

Smith, R. L. (1990). Max-Stable Processes and Spatial Extremes. *Unpublished manuscript*.

Smolyak, S. A. (1963). Quadrature and interpolation formulas for tensor products of certain classes of functions. *Doklady Akademii Nauk SSSR*, 148(5):1042–1045.

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(4):583–639.

Stein, M. L. (2015). When does the screening effect not hold? *Spatial Statistics*, 11:65–80.

Tavare, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring Coalescence Times From DNA Sequence Data. *Genetics*, 145:505–518.

Thibaud, E., Aalto, J., Cooley, D. S., Davison, A. C., and Heikkinen, J. (2016). Bayesian inference for the BrownâĂŞResnick process, with an application to extreme low temperatures. *Annals of Applied Statistics*, 10(4):2303–2324.

Tierney, L. and Kadane, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81(393):82–86.

Timm, O. E., Diaz, H. F., Giambelluca, T. W., and Takahashi, M. (2011). Projection of changes in the frequency of heavy rain events over Hawaii based on leading Pacific climate modes. *Journal of Geophysical Research*, 116(D04109).

Ting, M. F. and Wang, H. (1997). Summertime U.S. precipitation variability and its Relation to Pacific Sea Surface Temperature. *Journal of Climate*, 10(8):1853–1873.

Towler, E., PaiMazumder, D., and Holland, G. (2016). A framework for investigating large-scale patterns as an alternative to precipitation for downscaling to local drought. *Climate Dynamics*, pages 1–12.

Tsonis, A. A. and Swanson, K. L. (2008). On the Role of Atmospheric Teleconnections in Climate. *Journal of Climate*, 21:2990–3001.

Tye, M. R. and Cooley, D. (2015). A spatial model to examine rainfall extremes in Colorado's Front Range. *Journal of Hydrology*, 530:15–23.

van den Dool, H. (2007). *Empirical Methods in Short-Term Climate Predictions*. Oxford University Press, Oxford.

Van Den Dool, H. M. (1994). Searching for analogues, how long must we wait? *Tellus*, 46A(314-324).

von Storch, H. and Zwiers, F. W. (1999). *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge.

Wall, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, 121:311–324.

Wang, X. (2006). Approximating Bayesian inference by weighted likelihood. *The Canadian Journal of Statistics*, 34(2):279–298.

Ward, P. J., Jongman, B., Kummu, M., Dettinger, M. D., Sperna Weiland, F. C., and Winsemius, H. C. (2014). Strong influence of El Nino Southern Oscillation on flood risk around the world. *Proceedings of the National Academy of Sciences*, 111(44):15659–15664.

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11:3571–3594.

Weiser, C. (2016). mvQuad: Methods for Multivariate Quadrature. *http://cran.r-project.org/package=mvQuad*.

Wikle, C. K. and Anderson, C. J. (2003). Climatological analysis of tornado report counts using a hierarchical Bayesian spatiotemporal model. *Journal of Geophysical Research*, 108(D24).

Wilby, R. L., Wigley, T. M. L., Conway, D., Jones, P. D., Hewitson, B. C., Main, J., and Wilks, D. S. (1998). Statistical downscaling of general circulation model output : A comparison of methods. *Water Resources Research*, 34(11):2995–3008.

Zhang, H. (2004). Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

Zheng, F., Thibaud, E., Leonard, M., and Westra, S. (2015). Assessing the performance of the independence method in modeling spatial extreme rainfall. *Water Resources Research*, 51:7744–7758.