ABSTRACT

MORRIS, SAMUEL ALAN. Doing Statistics. (Under the direction of Brian Reich.)

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui.

Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Doing Statistics

by
Samuel Alan Morris

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2016

APPROVED BY:

Alyson Wilson	Arnab Maity
Adam Terando	Brian Reich Chair of Advisory Committee

DEDICATION

To my parents.

BIOGRAPHY

The author was born in a small town …

ACKNOWLEDGEMENTS

I would like to thank my advisor for his help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

This is where the introduction will go...

CHAPTER

2

A SPATIAL-TIME SKEW-T MODEL FOR THRESHOLD EXCEEDANCES

2.1 Introduction

Epidemiological studies have linked air quality to public health concerns regarding morbidity and mortality [Sam00]. As a result, the Environmental Protection Agency (EPA) has developed a set of standards to help reduce air pollution thereby improving air quality. Our study is motivated by an air pollution application where the focus is not on the average behavior, but instead the behavior over a fixed threshold determined by government regulation. More specifically, we consider the case of compliance for ozone. A site is said to be in compliance if the fourth highest daily maximum 8-hour concentration averaged over 3 years does not exceed 75 parts per billion (ppb). Figure 2.1 shows the ozone levels from July 10, 2005, at 1089 stations across the United States. We see a large area above the compliance level in the midwest covering Ohio, Indiana, Illinois, and parts of the surrounding states.

A spatial model for threshold exceedances warrants special consideration and standard spatial methods are likely to perform poorly. First, because we are interested only in high values, we want to "let the tail speak for itself". That is, if we fit a model to the entire data set, low-to-moderate values

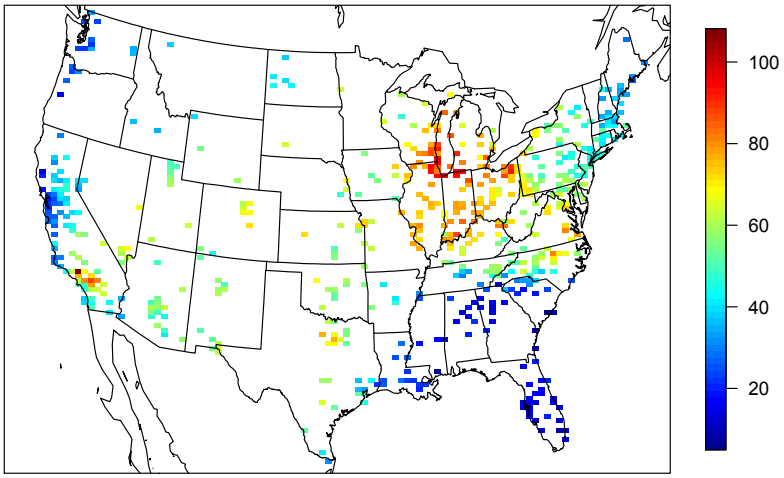

Figure 2.1 Ozone values (ppb) on July 10, 2005

would influence the fit of the overall model. As there are more of these values, they can unduly influence the distribution at the higher levels about which we are interested. Our inference method will only use data which exceed a threshold, and will impute data below the threshold, thereby tailoring the fit to the levels of interest. Second, likelihood-based spatial modeling typically assumes a Gaussian process, which is appropriate when mean behavior is of interest. However, the Gaussian distribution is light-tailed and symmetric, and therefore may be inappropriate for modeling data which does not share this tail behavior. Third, we aim to capture the dependence structure when ozone is at high levels, and dependence at these levels may not be well-represented by covariances which focus again on mean behavior. Asymptotic dependence/independence (see Section 2.2.2) are notions which describe how two random variables' probability of simultaneous exceedance of an extreme threshold behaves as one increases the threshold. The Gaussian distribution always exhibits asymptotic independence, except in the case of perfect dependence, thus is an inappropriate model for data which exhibits asymptotic dependence. To allow for more flexibility in the marginal tail and to allow for asymptotic dependence, the skew-t distribution forms the basis for our model.

Our approach differs from threshold modeling approaches based on extreme value distributions. There has been extensive work on threshold modeling in the field of extreme value statistics where

extreme events are naturally defined in terms of exceedances over a high threshold. Davison & Smith [DS90] considered modeling threshold exceedances of univariate time series by the generalized Pareto distribution. Bivariate threshold models for extreme value distributions were considered by Ledford & Tawn [LT96] who introduced a censored approach that provides a way to deal with different types of exceedances of a bivariate threshold in terms of only one or both components. These threshold models were extended to spatial models for extremes by Wadsworth & Tawn [WT12] and Thibaud et al. [Thi13] who fit various models to spatial extremes using a censored pairwise likelihood [Pad10] based on the approach of Ledford & Tawn [LT96]. Huser & Davison [HD14] further extended this to space-time modeling. Thibaud & Opitz [TO15], Engelke et al. [Eng14], and Wadsworth & Tawn [WT14], introduced more efficient inference for threshold exceedances of extremal spatial processes with full likelihood methods. The previous approaches to threshold modeling are motivated by extreme value theory and assume the threshold is high enough that extremal models are valid for the data, and for extrapolation beyond the range of observed values. Moreover, these approaches are computationally intensive and limited to rather small datasets. For example, Wadsworth & Tawn [WT14] present a simulation study with observations at 16 sites on a regular grid, and Engelke et al. [Eng14] analyze a dataset with observations at 35 meteorological stations. Our application with ozone data does not fit into this framework because we do not focus on exceedances of a very high threshold, but on exceedances of a fixed threshold. Furthermore, in our application, we have observations at over 1,000 ozone monitoring locations.

We propose a new spatiotemporal threshold exceedance model based on the skew-t process [Pad11]. Our model is a threshold exceedance model for the multivariate skew-t distribution for a fixed threshold. In this setting, we describe the threshold as fixed because it is specified in advance by regulatory compliance. This differs from the more traditional extremes literature where a threshold is selected to be the value beyond which an extremal model is appropriate for the data. We use a skew-t distribution because of its flexibility to model asymmetry and heavy-tailed data with the aim of predicting the probability of exceeding a high fixed threshold at an unobserved location. Need to add in more references here to skew-t used in a spatial setting.

Our model allows for inference and predictions using the full likelihood with computing on the order of Gaussian models. This allows us to use Bayesian methods to impute data below the threshold as well as to make predictions at unobserved locations. In a spatial setting, the multivariate skew-t distribution demonstrates asymptotic dependence between observations at all sites regardless of the distance between the sites. In order to address this concern, we introduce a random spatial partition similar to the method used by Kim et al. [Kim05] for non-stationary Gaussian data. This partitioning alleviates the asymptotic spatial dependence present in the skew-t distribution for sites that are far apart.

The paper is organized as follows. Section 2.2 is a brief review of the spatial skew-t process. In Section 2.3, we build upon the traditional skew-t process by incorporating censoring to focus on tails, partitioning to remove long-range asymptotic dependence, and extending the model to space-time data. The computing is described in Section 2.4. In Section 2.5, we present a simulation study that examines the predictive capabilities of this model compared to Gaussian and max-stable methods. We then compare our method to Gaussian and max-stable methods with a data analysis of ozone measurements from 800 sites throughout the US in Section 2.6. The final section provides brief discussion and direction for future research.

2.2 Spatial skew processes

The skew-elliptical family of distributions provides models that are mathematically tractable while introducing a slant parameter to account for asymmetric data. A brief review of the additive process [AC03; AC14; Ber16] by which a skew-t process is created is given here.

2.2.1 Skew-t process

Let $Y(\mathbf{s})$ be the observation at spatial location $\mathbf{s}$ in a spatial domain of interest $\mathscr{D} \in \mathscr{R}^2$. The spatial skew-$t$ process can be written

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + \lambda\sigma|z| + \sigma\, v(\mathbf{s}) \tag{2.1}$$

where $\mathbf{X}(\mathbf{s}) = [\mathbf{X}(\mathbf{s}_1)', \ldots, \mathbf{X}(\mathbf{s}_n)']$ is a set of spatial covariates at site $\mathbf{s}$, $\boldsymbol{\beta}$ is the vector of regression parameters, $\lambda \in \mathscr{R}$ is a parameter controlling skew, $z \sim N(0,1)$, $\sigma^2 \sim \text{IG}(a/2, b)$ is random scale parameter, IG is the distribution function of an inverse gamma random variable, a is the degrees of freedom, b controls the precision of the process, and $v(\mathbf{s})$ is a spatial Gaussian process with mean zero, variance one, and a positive definite correlation function $\rho(\mathbf{s}_1, \mathbf{s}_2)$. Although any positive definite correlation function could be used, we choose to use the stationary isotropic Matérn correlation with

$$\rho(h) = (1-\gamma)I(h=0) + \gamma \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\sqrt{2\nu}\frac{h}{\varphi}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}\frac{h}{\varphi}\right) \tag{2.2}$$

where φ is the spatial range, ν is the smoothness, γ is the proportion of variance accounted for by the spatial variation, K_ν is a modified Bessel function of the second kind, and $h = ||\mathbf{s}_1 - \mathbf{s}_2||$ is the distance between sites $\mathbf{s}_1$ and $\mathbf{s}_2$.

For a finite collection of locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, denote the vector of observations $\mathbf{Y} = [\,Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)]'$.

After marginalizing over both z and σ, using the notation from Azzalini & Capitanio [AC14, p. 176],

$$\mathbf{Y} \sim \mathrm{ST}_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, a), \tag{2.3}$$

that is, $\mathbf{Y}$ follows an n-dimensional skew-t distribution with location $\mathbf{X}\boldsymbol{\beta}$, variance matrix $\boldsymbol{\Omega}$, slant parameters $\boldsymbol{\alpha}$ and degrees of freedom a, where $\mathbf{X}$ is defined above,

$$\begin{aligned}
\boldsymbol{\Omega} &= \boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega}, \\
\boldsymbol{\omega} &= \mathrm{diag}\left(\sqrt{\frac{a(1+\lambda^2)}{b}}, \ldots, \sqrt{\frac{a(1+\lambda^2)}{b}}\right), \\
\bar{\boldsymbol{\Omega}} &= \frac{1}{1+\lambda^2}(\boldsymbol{\Sigma} + \lambda^2 \mathbf{1}\mathbf{1}') \\
\boldsymbol{\alpha} &= \lambda(1+\lambda^2)^{1/2}(1+\lambda^2 \mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1/2}\boldsymbol{\Sigma}^{-1}\mathbf{1},
\end{aligned}$$

and $\boldsymbol{\Sigma}$ is the positive definite correlation matrix which is obtained from $\rho(h)$. This process is desirable because of its flexible tail that is controlled by the skewness parameter λ and degrees of freedom a. Furthermore, the marginal distributions at each location also follow a univariate skew-t distribution [AC14]. For comparisons with other parameterizations, please see Appendix A.5.

2.2.2 Extremal dependence

Our interest lies in spatial dependence in the tail of the skew-t process. One measure of extremal dependence is the χ statistic [Col99]. For a stationary and isotropic spatial process, the χ statistic for locations $\mathbf{s}_1$ and $\mathbf{s}_2$ separated by distance $h = ||\mathbf{s}_1 - \mathbf{s}_2||$ with identical marginal distributions is

$$\chi(h) = \lim_{c \to c^*} \Pr[Y(\mathbf{s}_1) > c \,|\, Y(\mathbf{s}_2) > c] \tag{2.4}$$

where c^* is the upper limit of the support of Y. For the skew-t distribution, $c^* = \infty$. If $\chi(h) = 0$, then observations are asymptotically independent at distance h. For Gaussian processes, $\chi(h) = 0$ regardless of the distance h, so they are not suitable for modeling asymptotically dependent extremes. Unlike the Gaussian process, the skew-t process is asymptotically dependent (the explicit expression for $\chi(h)$ is given in Appendix A.4). However, one problem with the spatial skew-t process is that $\lim_{h\to\infty} \chi(h) > 0$. This occurs because all observations, both near and far, share the same z and σ terms. Therefore, this long-range dependence feature of the skew-t process is not ideal for spatial analysis of large geographic regions where we expect only local spatial dependence. We propose a solution to this in Section 2.3.2.

2.3 Spatiotemporal skew-t model for threshold exceedances

In this section, we propose extensions to the skew-t process to model spatial extremes over a large geographic region by introducing censoring to focus on tail behavior and a random partition to remove long-range asymptotic dependence. For notational convenience, we introduce the model for a single replication, and then extend this model to the spatiotemporal setting in Section 2.3.3.

2.3.1 Censoring to focus on the tails

As mentioned previously, we propose to use a censored approach because we are interested in high values and do not want the low-to-moderate values to influence the fit of the overall model. Let

$$\widetilde{Y}(\mathbf{s}) = \begin{cases} Y(\mathbf{s}) & \delta(\mathbf{s}) = 1 \\ T & \delta(\mathbf{s}) = 0 \end{cases} \tag{2.5}$$

be the censored observation at site $\mathbf{s}$ where $Y(\mathbf{s})$ is the uncensored observation, $\delta(\mathbf{s}) = I[Y(\mathbf{s}) > T]$, and T is a pre-specified threshold value. We impute the censored values as a step in the MCMC algorithm used to fit the model described in Section 2.4.1. Need to elaborate more here about the bias-variance tradeoff for threshold selection.

2.3.2 Partitioning to remove long-range asymptotic dependence

The motivation for the partition is that for a large spatial domain, it may not be reasonable to assume sites that are far apart demonstrate asymptotic dependence. Modeling different levels of asymptotic dependence was discussed by Wadsworth & Tawn [WT12] with a hybrid spatial dependence model. Huser & Davison [HD14] also allow for varying asymptotic dependence across both space and time with a partition structure represented by random discs moving across the space for a random duration with a random velocity and random radius. We handle the problem of long-range asymptotic dependence with a random partition. As discussed in Section 2.2, the source of long-range dependence is the shared z and σ. Therefore, to alleviate this dependence, we allow z and σ to vary by site. The model becomes

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + \lambda\sigma(\mathbf{s})|z(\mathbf{s})| + \sigma(\mathbf{s})v(\mathbf{s}). \tag{2.6}$$

To model spatial variation, consider a set of spatial knots $\mathbf{w}_1, \ldots, \mathbf{w}_K$ from a homogeneous Poisson process with intensity μ over spatial domain $\mathscr{D} \in \mathscr{R}^2$. The knots define a random partition of $\mathscr{D}$ by

subregions $P_1, \ldots, P_K$ defined as

$$P_k = \{\mathbf{s} : k = \arg\min_\ell ||\mathbf{s} - \mathbf{w}_\ell||\}. \tag{2.7}$$

In other words, P_1 is composed of all sites for which the closest knot is $\mathbf{w}_1$. For all $\mathbf{s} \in P_k$, with $k = 1, 2, \ldots, K$, the functions $z(\mathbf{s})$ and $\sigma(\mathbf{s})$ are equal to the constants z_k and σ_k respectively. and the z_k and σ_k^2 are distributed as $z_k \overset{iid}{\sim} N(0,1)$ and $\sigma_k^2 \overset{iid}{\sim} \text{IG}(a,b)$. So, within each partition, $Y(\mathbf{s})$ follows the spatial skew-t process defined in Section 2.2. Across partitions, the $Y(\mathbf{s})$ remain correlated via the correlation function for $\nu(\mathbf{s})$ because $\nu(\mathbf{s})$ spans all partitions.

The partitioning model remove long-range dependence. Conditional on knots $\mathbf{w}_1, \ldots, \mathbf{w}_K$, the χ statistic for two sites $\mathbf{s}_1$ and $\mathbf{s}_2$ in partitions k_1 and k_2 respectively is

$$\begin{aligned} \chi(h) &= I(k_1 = k_2)\chi_{\text{skew-}t}(h) + I(k_1 \neq k_2)\chi_{\text{Gaus}}(h) \\ &= I(k_1 = k_2)\chi_{\text{skew-}t}(h) \end{aligned} \tag{2.8}$$

where $I(\cdot)$ is an indicator function, $\chi_{\text{skew-}t}(h)$ is the χ statistic for a skew-t process given in (A.11), $\chi_{\text{Gaus}}(h)$ is the χ statistic for a Gaussian process, and $h = ||\mathbf{s} - \mathbf{t}||$. Therefore, sites in different subregions are asymptotically independent because $\chi_{\text{Gaus}}(h) = 0$ for all h. Marginally, over the knots, $\chi(h) = \pi(h)\chi_{\text{skew}-t}(h)$, where $\pi(h) = \Pr(k_1 = k_2)$ is the probability that two sites separated by distance h are in the same partition. In Appendix A.3, we show that $\lim_{h\to\infty} \pi(h) = 0$, assuming that the knots follow a homogeneous Poisson process. This implies that $\lim_{h\to\infty} \chi(h) = 0$. In Figure 2.2, we give $\chi(h)$ for $K = 1, 3, 5, 10$ partitions for a skew-t distribution with $\alpha = 10$, and 3 degrees of freedom. In Figure 2.2 we estimate $\pi(h)$ through simulation.

2.3.3 Extension to space-time data

When using daily measurements, the assumption of temporal independence is often inappropriate. In this section, we extend (2.6) to the spatiotemporal setting. There are several places where temporal dependence could be incorporated in the model, including the residuals $\nu_t(\mathbf{s})$. However, we choose to allow for temporal dependence in the $\mathbf{w}$, z, and σ terms because these terms dictate the tail behavior which is our primary focus. Let

$$Y_t(\mathbf{s}) = \mathbf{X}_t(\mathbf{s})'\boldsymbol{\beta} + \lambda\sigma_t(\mathbf{s})|z_t(\mathbf{s})| + \sigma_t(\mathbf{s})\nu_t(\mathbf{s}), \tag{2.9}$$

where $t \in \{1, \ldots, T\}$ denotes the day of each observation. Let $\mathbf{w}_{tk} = (w_{tk1}, w_{tk2})$ be a spatial knot on day t, and let $\mathbf{w}_{t1}, \ldots, \mathbf{w}_{tK}$ be a collection of spatial knots on day t. As in Section 2.3.2, these knots

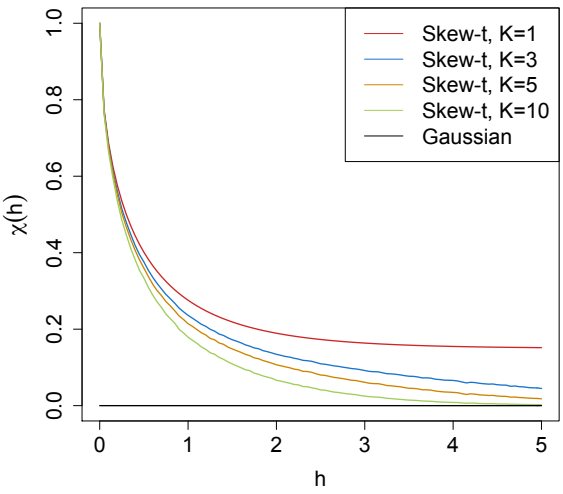

Figure 2.2 Extremal dependence measure $\chi(h)$, as a function of distance, h, for $K = 1, 3, 5$, and 10 knots.

define a daily partition $P_{t1}, \ldots, P_{tK}$, and for $\mathbf{s} \in P_{tk}$,

$$z_t(\mathbf{s}) = z_{tk} \quad \text{and} \quad \sigma_t(\mathbf{s}) = \sigma_{tk}. \tag{2.10}$$

We allow the partition structure to vary from day to day in order to account for sharp spikes in a response that may not be present every day (e.g. the impact of a forest fire on ozone levels).

We use an AR(1) time series model for $\mathbf{w}_{tk}$, z_{tk}, and σ_{tk}. The time series model must be specified after a transformation to preserve the skew-t process at each time point. For each time-varying parameter, we transform to obtain a standard normal marginal distribution, place a Gaussian prior with autocorrelation on the transformed parameter, and then transform back to the appropriate marginal distribution for the skew-t process. We first transform the spatial knots from $\mathscr{D}$ to $\mathscr{R}^2$ as follows. Let

$$w^*_{tki} = \Phi^{-1}\left[\frac{w_{tki} - \min(\mathbf{s}_i)}{\max(\mathbf{s}_i) - \min(\mathbf{s}_i)}\right], \quad i = 1, 2 \tag{2.11}$$

where Φ is a univariate standard normal density function and $\mathbf{s}_i = [s_{1i}, \ldots, s_{ni}]$. Then the transformed knots $\mathbf{w}^*_{tk} \in \mathscr{R}^2$. We use a transformation of a Gaussian random variable on $z_t(\mathbf{s})$ to ensure that the

marginal distributions of $z_t(\mathbf{s})$ are half-normal. Let

$$z_t^*(\mathbf{s}) = \Phi^{-1}\{\text{HN}[z_t(\mathbf{s})]\} \tag{2.12}$$

where HN is the distribution function of a half-normal random variable. We also use a transformation of a Gaussian random variable on $\sigma_t^2(\mathbf{s})$ to ensure that the marginal distributions of $\sigma_t^2(\mathbf{s})$ are inverse gamma. Let

$$\sigma_t^{2*}(\mathbf{s}) = \Phi^{-1}\left\{\text{IG}[\sigma_t^2(\mathbf{s})]\right\} \tag{2.13}$$

where IG is defined as before. The AR(1) process for each tail parameter is $\mathbf{w}_{1k}^* \sim N_w(0,1)$, $z_{1k}^* \sim N(0,1)$, $\sigma_{1k}^{2*} \sim N(0,1)$, and for $t > 1$ the time series is modeled as

$$\mathbf{w}_{tk}^* | \mathbf{w}_{t-1,k}^* \sim N_2\left[\phi_w \mathbf{w}_{t-1,k}^*, (1-\phi_w^2)\right] \tag{2.14}$$

$$z_{tk}^* | z_{t-1,k}^* \sim N\left[\phi_z z_{t-1,k}^*, (1-\phi_z^2)\right] \tag{2.15}$$

$$\sigma_{tk}^{2*} | \sigma_{t-1,k}^{2*} \sim N\left[\phi_\sigma \sigma_{t-1,k}^{2*}, (1-\phi_\sigma^2)\right] \tag{2.16}$$

where $|\phi_w|$, $|\phi_z|$, $|\phi_\sigma| < 1$. These are stationary time series models with marginal distributions $\mathbf{w}_k^* \sim N_2(0,1)$, $z_k^* \sim N(0,1)$, and $\sigma_k^{2*} \sim N(0,1)$. After transformation back to the original space, $\mathbf{w}_{tk} \sim \text{Unif}(\mathscr{D})$, $z_{tk} \sim HN(0,1)$, and $\sigma_{tk}^2 \sim \text{IG}(a,b)$. We then create the partition for day t using $\mathbf{w}_{t1}, \ldots, \mathbf{w}_{tk}$. For each day, the model is identical to the spatial-only model in (2.6) by construction.

2.4 Hierarchical model

Conditioned on $z_{tk}(\mathbf{s})$, $\sigma_{tk}^2(\mathbf{s})$, and P_{tk}, the marginal distributions of $Y(\mathbf{s})$ are Gaussian and the joint distribution of $Y(\mathbf{s})$ is multivariate Gaussian. However, we do not fix the partitions, they are treated as unknown and updated in the MCMC algorithm. We model this with a Bayesian hierarchical model as follows. Let $\mathbf{w}_{t1}, \ldots, \mathbf{w}_{tK}$ be a set of daily spatial knots in a spatial domain of interest, $\mathscr{D}$, and P_{tk} as defined in (2.7). One approach would be to allow K to be unknown and follow a Poisson process prior, but this would lead to onerous computing. Therefore, we elect to treat K as a tuning parameter for the MCMC by fixing it at different levels and assessing its impact on prediction as

described in Section 2.5.2. Then

$$
\begin{aligned}
Y_t(\mathbf{s}) \mid z_t(\mathbf{s}), \sigma_t^2(\mathbf{s}), P_{tk}, \Theta &= \mathbf{X}_t(\mathbf{s})'\beta + \lambda\sigma_t(\mathbf{s})|z_t(\mathbf{s})| + \sigma_t(\mathbf{s})v_t(\mathbf{s}) \qquad (2.17)\\
z_t(\mathbf{s}) &= z_{tk} \text{ if } \mathbf{s} \in P_{tk}\\
\sigma_t^2(\mathbf{s}) &= \sigma_{tk}^2 \text{ if } \mathbf{s} \in P_{tk}\\
\lambda &\sim N(0, \sigma_\lambda^2)\\
v_t(\mathbf{s}) \mid \Theta &\sim \text{Matérn}\,(0, \Sigma)\\
z_{tk}^* \mid z_{t-1,k}^* &\sim N(\phi_z z_{t-1,k}^*, (1-\phi_z^2))\\
\sigma_{tk}^{2*} \mid \sigma_{t-1,k}^{2*} &\sim N(\phi_\sigma \sigma_{t-1,k}^{2*}, (1-\phi_\sigma^2))\\
\mathbf{w}_{tk}^* \mid \mathbf{w}_{t-1,k}^* &\sim N_2(\phi_w \mathbf{w}_{t-1,k}^*, (1-\phi_w^2))
\end{aligned}
$$

where $\Theta = \{\varphi, \nu, \gamma\lambda, \beta\}$, and Σ is a Matérn covariance matrix as described in Section 2.2.1.

2.4.1 Computation

We use MCMC methods to explore the posterior. At each MCMC iteration, we first impute values below the threshold conditional on observations above the threshold. This is feasible for large datasets with our model because for a single day, conditional on the model parameters, we only need to draw from a truncated multivariate normal distribution. We can use Gibbs sampling to update $Y_t(\mathbf{s})$ for censored observations that are below the threshold T. After conditioning on λ, $z_t(\mathbf{s})$ and non-censored observations, $Y_t(\mathbf{s})$ has truncated normal full conditionals. So we sample $Y_t(\mathbf{s}) \sim N_{(-\infty,T)}(\mathbf{X}_t'(\mathbf{s})\beta + \lambda|z_t(\mathbf{s})|, \mathbf{\Sigma})$.

Then, we update model parameters, Θ, using a Metropolis-Hastings algorithm with Gibbs sampling when needed. The final step of the computation is to use Bayesian Kriging to generate a predictive distribution for $Y_t(\mathbf{s}^*)$ at prediction location $\mathbf{s}^*$. This step is similar to the imputation for censored observations except that the full conditionals are no longer truncated at T. See Appendices A.1 and A.2 for details regarding the MCMC algorithm.

2.5 Simulation study

In this section, we present the results from a simulation study to investigate how the number of partitions and the level of thresholding impact the accuracy of predictions made by the model and to compare with Gaussian and max-stable methods.

2.5.1 Design

For all simulation designs, we generated data from model (2.6) in Section 2.3.2 using $n_s = 144$ sites and $n_t = 50$ independent days. The sites were generated Uniform($[0,10] \times [0,10]$). We generated data from 5 different simulation designs:

(1) Gaussian marginal, $K = 1$ knot

(2) Skew-t marginal, $K = 1$ knots

(3) Skew-t marginal, $K = 5$ knots

(4) Max-stable

(5) A Brown-Resnick process with range 1 and smoothness 0.5

In the first three designs, the $\nu_t(\mathbf{s})$ terms were generated using a Matérn covariance with smoothness parameter $\nu = 0.5$ and spatial range $\varphi = 1$. For the covariance matrices in designs 1 – 3, the proportion of the variance accounted for by the spatial variation was $\gamma = 0.9$ while the proportion of the variance accounted for by the nugget effect was 0.1. In the first design, $\sigma^2 = 2$ was used for all days which results in a Gaussian distribution. For designs 2 and 3, $\sigma^2_{tk} \overset{iid}{\sim} \text{IG}(3,8)$ to give a t distribution with 6 degrees of freedom. For design 1, we set $\lambda = 0$. For designs 2 and 3, $\lambda = 3$ was used as to simulate moderate skewness, and the z_t were generated as described in (**??**). In designs 1 – 3, the mean $\mathbf{X}'\boldsymbol{\beta} = 10$ was assumed to be constant across space. In the fourth design, we generated from a spatial max-stable distribution [RS12]. In this design, data have marginal distributions that follow a generalized extreme value distribution with location parameter 1, scale parameter 1, and shape parameter 0.2. In this model, a random effect was used to induce spatial dependence using 144 spatial knots on a regular lattice in the square $[1,9] \times [1,9]$. For this setting, $\gamma_{\text{RS}} \in (0,1)$, is similar to γ in that it controls the relative strength of the nugget effect. We set $\gamma_{\text{RS}} = 0.5$, which represents moderate spatial dependence.

$M = 50$ data sets are generated for each design. For each data set we fit the data using six models

(1) Gaussian marginal, $K = 1$ knots

(2) Skew-t marginal, $K = 1$ knots, $T = -\infty$

(3) Symmetric-t marginal, $K = 1$ knots, $T = q(0.80)$

(4) Skew-t marginal, $K = 5$ knots, $T = -\infty$

(5) Symmetric-t marginal, $K = 5$ knots, $T = q(0.80)$

(6) A max-stable model based on Reich & Shaby [RS12] thresholded at $T = q(0.80)$

where $q(0.80)$ is the 80th sample quantile of the data. The design matrix $\mathbf{X}$ includes an intercept with a first-order spatial trend with priors of $\beta_{\text{int}}, \beta_{\text{lat}}, \beta_{\text{long}} \overset{iid}{\sim} \text{N}(0, 10)$. The spatial covariance parameters have priors $\log(\nu) \sim \text{N}(-1.2, 1)$, $\gamma \sim \text{Unif}(0, 1)$, $\varphi \sim \text{Unif}(15)$. The skewness parameter has prior $\lambda \sim N(0, 20)$. The residual variance terms have priors $\sigma_t^2(\mathbf{s}) \sim \text{IG}(a, b)$, where a has a discrete uniform prior on a mesh from 0.2 to 20 with spacing of 0.1 and b has a Gamma(0.1, 0.1) prior. The knots have priors $\mathbf{w} \sim \text{Unif}(\mathscr{D})$. We tried also fitting the skew-t marginals for the thresholded models, but it is very challenging for the MCMC to properly identify the skewness parameter with a censored left tail. Each chain of the MCMC ran for 20,000 iterations with a burn-in period of 10,000 iterations. Parameters appear to converge properly; however, in the models with multiple partitions (i.e. models 4 and 5) it is hard to assess the convergence of $\mathbf{w}$, $z(\mathbf{s})$, and $\sigma^2(\mathbf{s})$ because of partition label switching throughout the MCMC.

2.5.2 Cross validation

Models were compared using cross validation, with 100 sites used as training sites to fit the models, and 44 sites withheld for testing the predictions. Because one of the primary goals of this model is to predict exceedances over a fixed threshold, we use Brier scores to compare the models [GR07]. The Brier score for predicting exceedance of a threshold c is given by $[e(c) - P(c)]^2$ where $e(c) = I[y > c]$ is an indicator function indicating that a test set value, y, has exceeded the threshold, c, and $P(c)$ is the predicted probability of exceeding c. We average the Brier scores over all test sites and days. For the Brier score, a lower score indicates a better fit.

2.5.3 Results

We compared the Brier scores for exceeding 4 different thresholds for each dataset. The thresholds used for the Brier scores are extreme quantiles from the simulated data for $q(0.90)$, $q(0.95)$, $q(0.98)$, and $q(0.99)$. Figure 2.3 gives the Brier score relative to the Brier score for the Gaussian method calculated as

$$\text{BS}_{\text{rel}} = \frac{\text{BS}_{\text{method}}}{\text{BS}_{\text{Gaussian}}}. \tag{2.18}$$

We analyzed the results for the simulation study using a Friedman test at $\alpha = 0.05$ to see if at least one method had a significantly different Brier score. For Friedman tests that came back with a significant p-value, we conducted a Wilcoxon-Nemenyi-McDonald-Thompson [Hol14] test to see which of the

methods had different results. The full results for the Wilcoxon-Nemenyi-McDonald-Thompson tests are given in Appendix A.6.

The results show that when the data are generated from a Gaussian process, our method performs comparably to a Gaussian approach. In general, when the underlying process is not Gaussian, our method results in an improvement over both the max-stable and Gaussian methods. One exception to this is the case when the generative process is max-stable. In this case, the max-stable method outperforms our method; however, for predictions at high quantile levels, the differences between the max-stable method and our method decrease. The non-thresholded methods tend to outperform the thresholded methods, but this is not surprising given that in most cases, the data are generated directly from the model used in the method. In summary, our method provides more flexibility for data that demonstrate some level of asymmetry or heavy tails, while still performing comparably to Gaussian methods when the data are symmetric and have light tails.

2.6 Data analysis

We consider daily observations of maximum 8-hour ozone measurements for the 31 days of July 2005 at 1,089 Air Quality System (AQS) monitoring sites in the United States as the response (see Figure 2.1). For each site, we also have covariate information containing the estimated ozone from the Community Multi-scale Air Quality (CMAQ) modeling system. Initially, we fit a linear regression with $\mathbf{X}_t(\mathbf{s}) = [1, \text{CMAQ}_t(\mathbf{s})]'$. Figure 2.4 shows a Q-Q plot of the residuals compared to a skew-t distribution with 10 d.f. and $\alpha = 1$.

Standard exploratory data analysis techniques for extremal dependence are very challenging with only 31 days worth of data because it is difficult to estimate extreme quantiles at each site to obtain empirical estimates of χ. Despite the fact that there is only one month of data, we can get some sense of extremal dependence between sites by looking at joint occurrences of high sample quantiles. For example, Figure 2.5 suggests there is more agreement between sites that are close to one another than sites that are far from one another. Another aspect that distinguishes our approach from more traditional extremes analyses, is how the threshold is selected. In our example, a threshold of 75 ppb which corresponds to $q(92)$ for all observations, but marginally it represents anywhere from $q(0.06)$ to $q(1)$.

2.6.1 Model comparisons

We fit the model using Gaussian and skew-t marginal distributions with $K = 1, 5, 6, 7, 8, 9, 10, 15$ partitions. We choose to censor $Y(\mathbf{s})$ at $T = 0$, $T = 50$ (0.42 sample quantile), and $T = 75$ (0.92

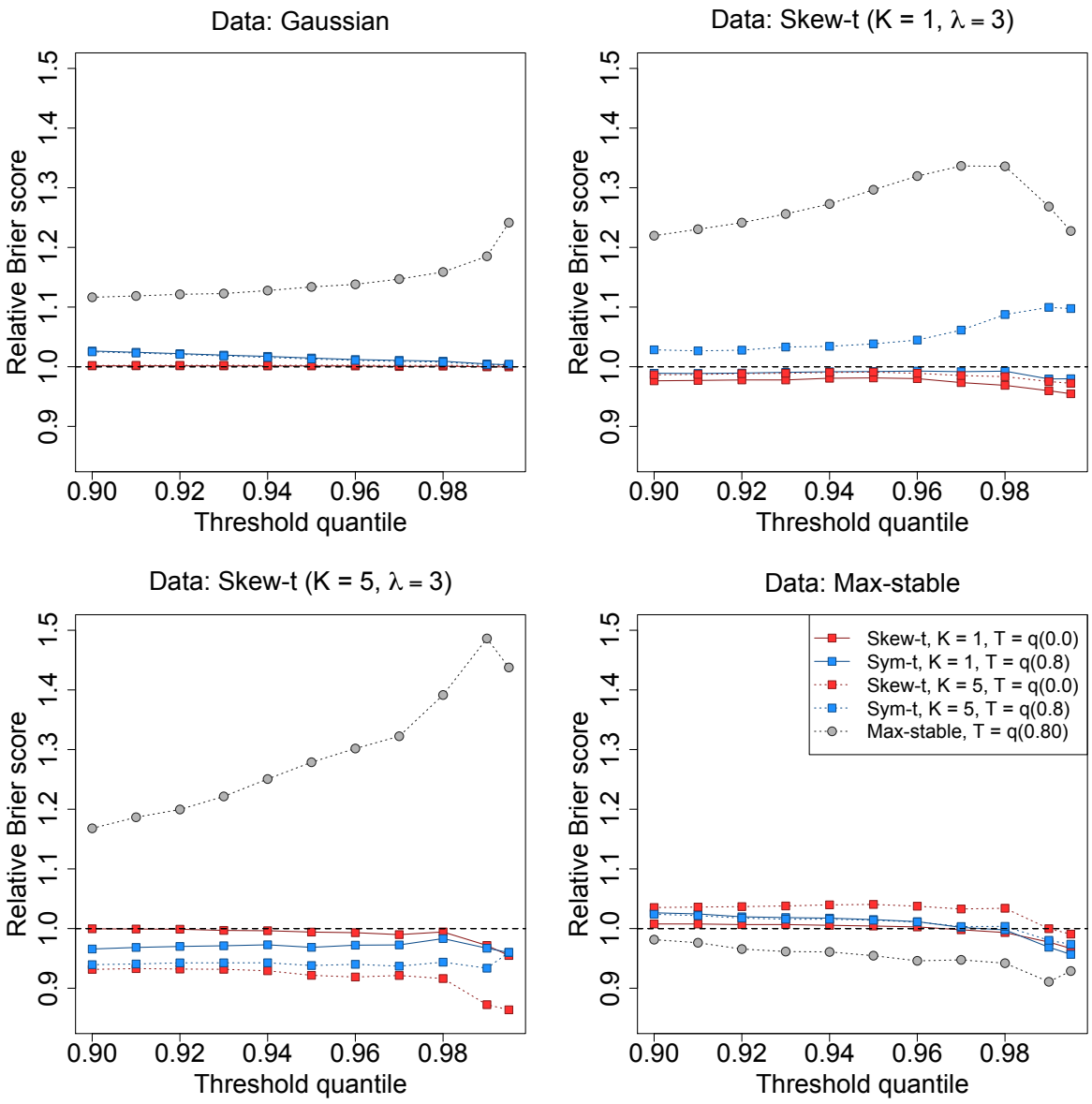

Figure 2.3 Brier scores relative to the Gaussian method for simulation study results. A ratio lower than 1 indicates that the method outperforms the Gaussian method.

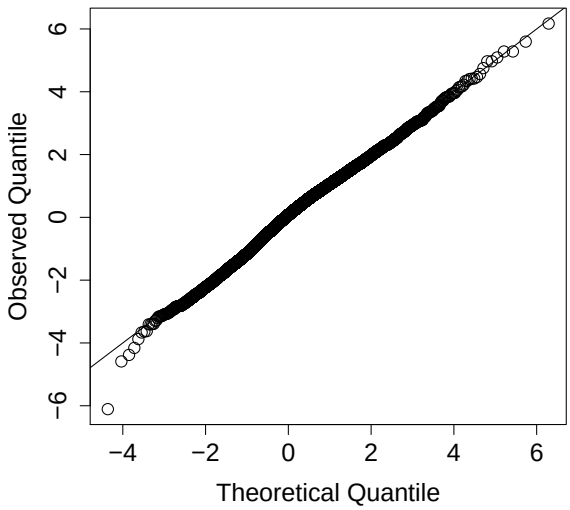

Figure 2.4 Q-Q plot of the residuals for a skew-t distribution with 10 d.f. and $\alpha = 1$ (right)

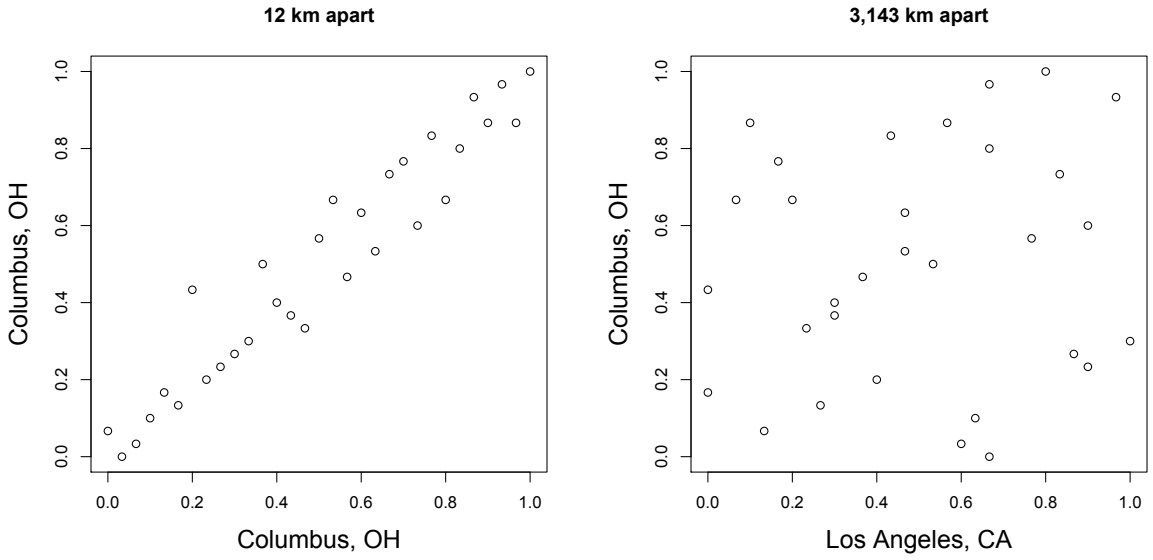

Figure 2.5 Daily quantiles for two monitoring locations near Columbus, OH (left) and daily quantiles for a monitoring location in Los Angeles, CA and Columbus, OH (right)

sample quantile) ppb in order to compare results from no, moderate, and high censoring. The upper threshold of 75 ppb was used because the current air quality standard is based on exceedance of 75 ppb. As with the simulation study, for models with a threshold of $T = 75$, we use a symmetric-t marginal distribution. We also compare models with no time series to models that include the time series. Finally, as a comparison to max-stable methods, we fit the model using the hierarchical max-stable model of Reich & Shaby [RS12] with the data thresholded at $T = 75$. All methods assume $\mathbf{X}_t(\mathbf{s}) = [1, \text{CMAQ}_t(\mathbf{s})]'$. To ensure that the max-stable method runs in a reasonable amount of time, we use a stratified sub-sample of 800 sites. We conduct two-fold cross validation using 400 training sites and 400 validation sites as described in Section 2.5.2

Each chain of the MCMC ran for 30,000 iterations with a burn-in period of 25,000 iterations. We use the same priors for the spatial covariance parameters, skewness parameter, and knots as in the simulation study. The prior for the residual variance terms is $\sigma_t^2(\mathbf{s}) \sim text IG(a,b)$ where a is the same as the simulation study, but b has a Gamma$(1,1)$ prior. Parameters appear to converge properly; however, as before, for models with multiple partitions it is hard to assess the convergence of $\mathbf{w}$, $z(\mathbf{s})$, and $\sigma^2(\mathbf{s})$ because of partition label switching throughout the MCMC. For each model, Brier scores were averaged over all sites and days to obtain a single Brier score for each dataset. At a particular threshold or quantile level, the model that fits the best is the one with the lowest score. We then compute the relative (to Gaussian) Brier scores (see Section 2.5.3) to compare each model.

2.6.2 Results

The results suggest that the skew-t, thresholded, partitioned, and time series models all give an improvement in predictions over the Gaussian model, whereas the max-stable method results in relative Brier scores between 1.07 and 1.15 indicating poorer performance than the Gaussian model. The plots in Figure 2.6 show the relative Brier scores for time-series and non-time-series models, using $K = 1, 7$, and 15 knots at thresholds $T = 0, 50$, and 75 ppb. Most of the models perform similarly across all the Brier scores; however, for single-partition models without thresholding, performance tends to diminish in the extreme quantiles. The results also suggest that thresholding improves performance for estimates in the extreme quantiles. Both plots have similar features suggesting that most settings do reasonably well. In particular, for all extreme quantiles, selecting a moderate number of knots (e.g. $K = 5, \ldots, 10$) tends to give the best results. Table 2.1 shows the best two models for selected extreme quantiles.

We illustrate the predictive capability of our model in Figure 2.7 by plotting the 99th quantile for South Carolina and Georgia, a subset of the spatial domain, in order to study local features. The four methods used are

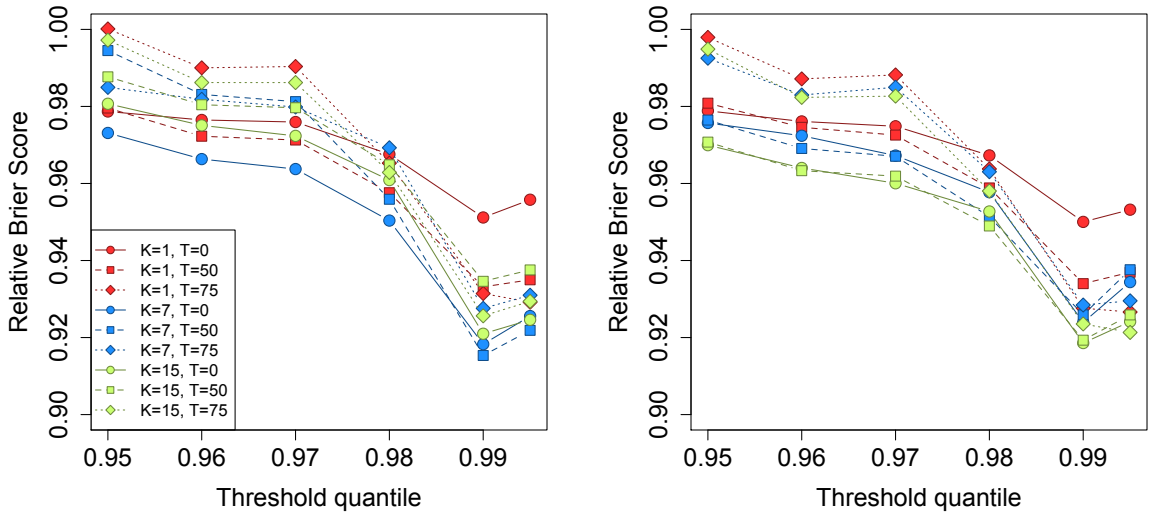

Figure 2.6 Relative Brier scores for time-series models (left) and non-time-series models (right). Relative brier score for the max-stable model is between 1.07 and 1.15

Table 2.1 Top two performing models for ozone analysis at extreme quantiles with Relative Brier score

	1st				2nd			
$q(0.90)$	No time series	$K=7$	$T=0$	BS: 0.980	No time series	$K=9$	$T=0$	BS: 0.980
$q(0.95)$	No time series	$K=15$	$T=50$	BS: 0.970	No time series	$K=9$	$T=50$	BS: 0.970
$q(0.98)$	No time series	$K=5$	$T=50$	BS: 0.945	No time series	$K=10$	$T=50$	BS: 0.946
$q(0.99)$	Time series	$K=10$	$T=75$	BS: 0.912	Time series	$K=6$	$T=75$	BS: 0.913
$q(0.995)$	Time series	$K=6$	$T=75$	BS: 0.917	Time series	$K=10$	$T=75$	BS: 0.918

(1) Gaussian

(2) Skew-t, $K=1$ knot, $T=0$, no time series

(3) Skew-t, $K=5$ knots, $T=50$, no time series

(4) Symmetric-t, $K=10$ knots, $T=75$, time series.

In the bottom two plots, we plot the differences between method 4 and methods 1 and 2. The most noticeable differences between the reference methods and the comparison methods is that the comparison methods tend to give higher estimates of the 99th quantile along the I-85 corridor between Charlotte and Atlanta.

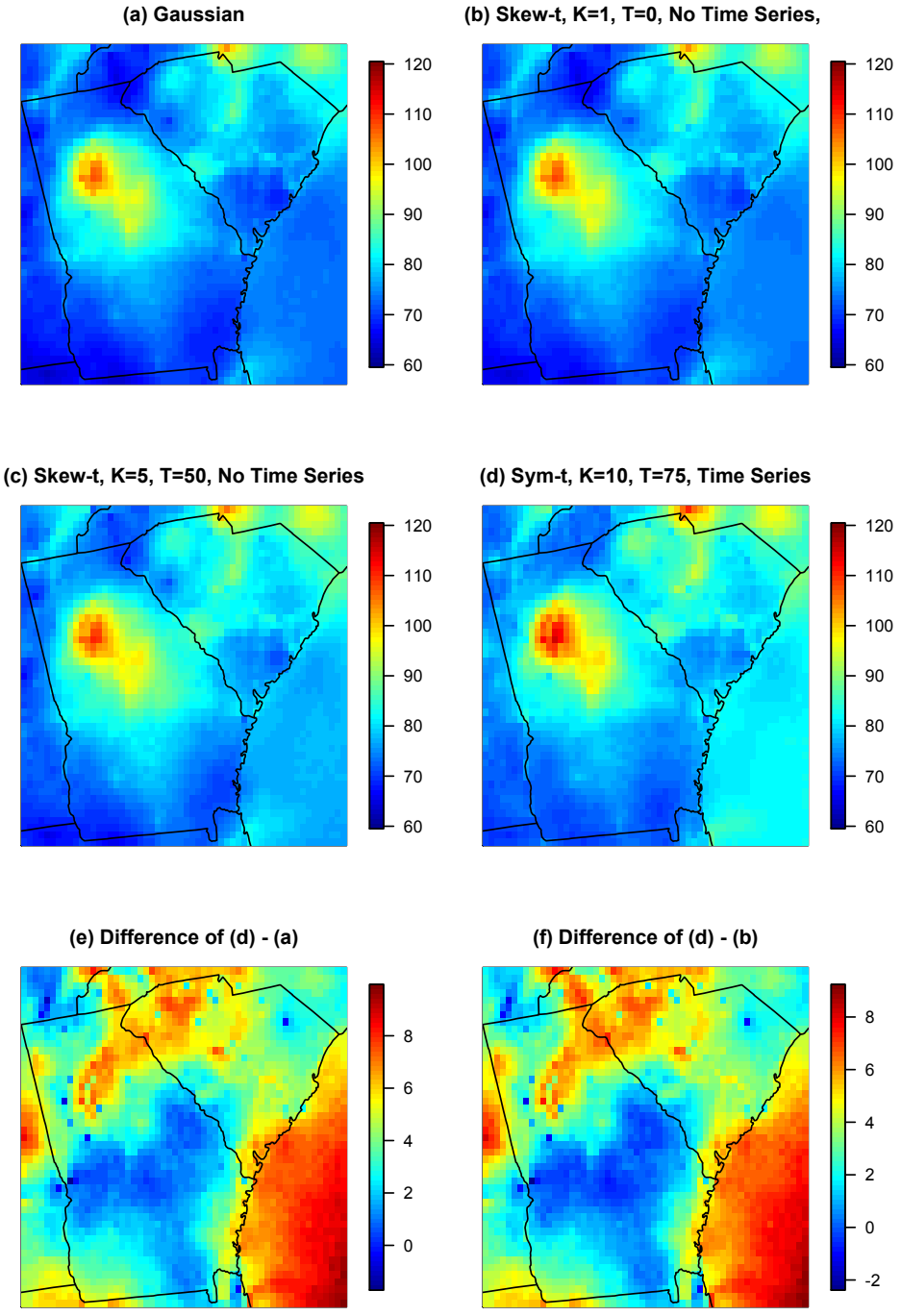

Figure 2.7 Panels (a) – (d) give the posterior predictive $\widehat{q}(0.99)$ for the month of July under four different models, panel (e) gives the difference between $\widehat{q}(0.99)$ in panels (d) and (a), panel (f) gives the difference between $\widehat{q}(0.99)$ in panels (d) and (b).

2.7 Discussion

In this paper we propose a new threshold exceedance approach for spatiotemporal modeling based on the skew-t process. The proposed model gives flexible tail behavior, demonstrates asymptotic dependence for observations at sites that are near to one another, and has computation on the order of Gaussian models for large space-time datasets. In the simulation study, we demonstrate that this model shows statistically significant improvements over a naïve Gaussian approach and in most cases, a max-stable approach. In both the simulation study, and the application to ozone data, we find that incorporating a partition in the model can improve extreme predictions. Furthermore the results from the data analysis suggest that thresholding can improve performance when predicting in the extreme tails of the data.

This model presents new avenues for future research. One possibility is the implementation of a different partition structure. We choose to define the random effects for a site by using an indicator function based on closeness to a knot. However, this indicator function could be replaced by kernel function that would allow for multiple knots to impact each site, with the weight of each knot to be determined by some characteristic such as distance. Another area that should be explored is the temporal dependence in the model. Instead of implementing a time series on the random effects, a three-dimensional covariance structure on the residuals could be implemented to address temporal dependence. Finally, we acknowledge that by specifying the number of knots, we may be underestimating the uncertainty in the model. This could be incorporated by treating the number of knots as a model parameter instead of fixing it to be a specific value.

CHAPTER

3

A SPATIAL MODEL FOR RARE BINARY EVENTS

3.1 Introduction

The goal in binary regression is to relate a latent variable to a response using a link function. Two common examples of binary regression include logistic regression and probit regression. The link functions for logistic and probit regression are symmetric, so they may not be well-suited for asymmetric data. An asymmetric alternative to these link functions is the complementary log-log (cloglog) link function. More recently, Wang & Dey [WD10] introduced the generalized extreme value (GEV) link function for rare binary data. The GEV link function introduces a new shape parameter to the link function that controls the degree of asymmetry. The cloglog link is a special case of the GEV link function when the shape parameter is 0.

Want to make the case in this paragraph that spatial logistic and probit models are not appropriate because asymptotic dependence is 0. Spatial logistic and probit models are commonly presented using a hierarchical model citation. In the hierarchical framework, spatial dependence is typically modeled with an underlying latent Gaussian process, and conditioned on this process, observations are independent. However, if the latent variable is assumed to follow a GEV marginally,

then a Gaussian process may not be appropriate to describe the dependence due to the fact that they do not demonstrate asymptotic dependence regardless of the strength of the dependence in the bulk of the data. As an alternative to the Gaussian process, we propose using a latent max-stable process because it allows for asymptotic dependence citation. Max-stable processes are extremely flexible, but are often challenging to work with because very few finite dimensional representations exist in more than two or three dimensions.

Paragraph outlining the structure of the paper

3.2 Binary regression using the GEV link

Here, we provide a brief review of the the GEV link of Wang & Dey [WD10]. Let $Y_i \in \{0,1\}, i = 1,\ldots,n$ be a collection of i.i.d. binary responses. It is assumed that $Y_i = I(z_i > 0)$ where $I(\cdot)$ is an indicator function, $z_i = [1-\xi \mathbf{X}_i \boldsymbol{\beta}]^{1/\xi}$ is a latent variable following a GEV(1,1,1) distribution, $\mathbf{X}_i$ be the associated p-vector of covariates with first element equal to one for the intercept, and $\boldsymbol{\beta}$ is a p-vector of regression coefficients. So, the marginal probability of an event is given by

$$\pi_i = 1 - \exp\left(-\frac{1}{z_i}\right). \tag{3.1}$$

Although this link was selected by Wang & Dey based on its ability to handle asymmetry, the GEV distribution is one of the primary distributions used for modeling extremes. Traditionally, analysis of extreme events is done using block maxima or occurrences over a suitably high threshold. Because extreme events are rare, it is therefore reasonable to use similar methods when analyzing rare binary data.

3.3 Spatial dependence for binary regression

In many binary regression applications, spatial dependence is handled using a hierarchical model assuming an latent spatial process. Let $Y(\mathbf{s})$ be the observation at spatial location $\mathbf{s}$ in a spatial domain of interest $\mathscr{D} \in \mathscr{R}^2$. We assume $Y(\mathbf{s}) = I[Z(\mathbf{s}) > 0]$ where $Z(\mathbf{s})$ is a latent spatial process. In spatial logistic and probit regression, the latent spatial process is assumed to be a Gaussian process. A Gaussian process may not be appropriate when describing dependence in the tails of the distribution because it always exhibits asymptotic independence, except in the case of perfect dependence. Because we use extreme values analysis as the foundation for rare binary analysis, we propose useing a max-stable process to model the latent spatial process

A max-stable process has generalized extreme value marginal distributions with location $\mathbf{X}^T(\mathbf{s})\boldsymbol{\beta}$,

scale σ, and shape ξ. For identifiability purposes we fix $\sigma = 1$. Although $\boldsymbol{\beta}$ and ξ could be permitted to vary across space, we assume that they are constant across $\mathscr{D}$.

For a finite collection of locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, denote the vector of observations $\mathbf{Y} = [Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)]^T$. The spatial dependence is determined by the joint distribution of $\mathbf{Z} = [Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)]^T$,

$$G(\mathbf{z}) = \mathrm{P}[Z(\mathbf{s}_1) < z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n) < z(\mathbf{s}_n)] = \exp\left\{-\sum_{l=1}^{L}\left[\sum_{i=1}^{n}\left(\frac{w_l(\mathbf{s}_i)}{z(\mathbf{s}_i)}\right)^{1/\alpha}\right]^{\alpha}\right\}, \tag{3.2}$$

where $w_l(\mathbf{s}_i)$ are a set of L weights that determine the spatial dependence structure, and $\alpha \in (0,1)$ determines the strength of dependence, with α near zero giving strong dependence and $\alpha = 1$ giving joint independence. This is a special case of the multivariate GEV distribution with asymmetric Laplace dependence function [Taw90]. The weights $w_l(\mathbf{s}_i)$ in (3.2) vary smoothly across space to induce spatial dependence. Many weight functions are possible, but the weights must be constrained so that $\sum_{l=1}^{L} w_l(\mathbf{s}_i) = 1$ for all $i = 1, \ldots, n$ to preserve the marginal GEV distribution. For example, [RS12] take the weights to be scaled Gaussian kernels with knots $\mathbf{v}_l$,

$$w_l(\mathbf{s}_i) = \frac{\exp\left[-0.5\left(||\mathbf{s}_i - \mathbf{v}_l||/\rho\right)^2\right]}{\sum_{j=1}^{L}\exp\left[-0.5\left(||\mathbf{s}_i - \mathbf{v}_j||/\rho\right)^2\right]}. \tag{3.3}$$

The kernel bandwidth $\rho > 0$ determines the spatial range of the dependence, with large ρ giving long-range dependence and vice versa.

One nice feature to this representation for the max-stable process is that the lower-dimensional marginal distributions also follow a multivariate extreme value distribution. More importantly, at a single site i, the marginal distribution gives $P[Y(\mathbf{s}_i) = 1] = 1 - \exp\left[-\frac{1}{z(\mathbf{s}_i)}\right]$ which is the same as the marginal distribution given by Wang & Dey [WD10].

The joint likelihood of Y is computationally challenging to obtain. Therefore, to incorporate spatial dependence into the model, we consider the hierarchical max-stable process of Reich & Shaby [RS12]. Consider a set of $A_1, \ldots, A_l \overset{iid}{\sim} \mathrm{PS}(\alpha)$ random effects associated with spatial knots $\mathbf{v}_1, \ldots, \mathbf{v}_L$. The hierarchical model is given by

$$\begin{aligned} Z(\mathbf{s})|A_l, \ldots, A_l &\overset{indep}{\sim} \mathrm{GEV}[\theta(\mathbf{s}), \alpha\theta(\mathbf{s}), \alpha] \\ A_l &\overset{iid}{\sim} \mathrm{PS}(\alpha) \end{aligned} \tag{3.4}$$

where $\theta(\mathbf{s}) = \left[\sum_{l=1}^{L} A_l\, w_l(\mathbf{s})^{1/\alpha}\right]^{\alpha}$.

3.4 Joint distribution

In section 3.4.1, we give an exact expression in the case where there are only two spatial locations which is useful for constructing a pairwise composite likelihood and studying spatial dependence. For more than two locations, we are also able to compute the exact likelihood when the number of locations is large but the number of events is small, as might be expected for very rare events (see Appendix .1.1).

3.4.1 Bivariate distribution

Then in a bivariate setting, the probability mass function as a function of α is

$$\mathrm{P}[Y(\mathbf{s}_i), Y(\mathbf{s}_j)] = \begin{cases} \exp\left\{-\sum_{l=1}^{L}\left[\left(\frac{w_l(\mathbf{s}_i)}{z(\mathbf{s}_i)}\right)^{1/\alpha}+\left(\frac{w_l(\mathbf{s}_j)}{z(\mathbf{s}_j)}\right)^{1/\alpha}\right]^{\alpha}\right\} & Y(\mathbf{s}_i)=0, Y(\mathbf{s}_j)=0 \\ \frac{1}{z(\mathbf{s}_i)}-\exp\left\{-\sum_{l=1}^{L}\left[\left(\frac{w_l(\mathbf{s}_i)}{z(\mathbf{s}_i)}\right)^{1/\alpha}+\left(\frac{w_l(\mathbf{s}_j)}{z(\mathbf{s}_j)}\right)^{1/\alpha}\right]^{\alpha}\right\} & Y(\mathbf{s}_i)=1, Y(\mathbf{s}_j)=0 \\ 1-\exp\left\{-\frac{1}{z(\mathbf{s}_i)}\right\}-\exp\left\{-\frac{1}{z(\mathbf{s}_j)}\right\}+\exp\left\{-\sum_{l=1}^{L}\left[\left(\frac{w_l(\mathbf{s}_i)}{z(\mathbf{s}_i)}\right)^{1/\alpha}+\left(\frac{w_l(\mathbf{s}_j)}{z(\mathbf{s}_j)}\right)^{1/\alpha}\right]^{\alpha}\right\} & Y(\mathbf{s}_i)=1, Y(\mathbf{s}_j)=1 \end{cases} \tag{3.5}$$

3.5 Quantifying spatial dependence

I still need to incorporate Brian's suggestions here In the literature on extremes, one common metric to describe the bivariate dependence is the χ statistic of Coles et al. [Col99]. The χ statistic between two observations z_1 and z_2 is given by

$$\chi(\mathbf{s}_1,\mathbf{s}_2) = \lim_{c\to\infty} = P(Z_1 > c \,|\, Z_2 > c). \tag{3.6}$$

However, in this latent variable approach, $\lim_{c\to\infty}$ may not be the most reasonable metric because the observed data are a series of zeros and ones. Therefore, we chose the κ statistic of Cohen [Coh60] defined by

$$\kappa = \frac{P(A)-P(E)}{1-P(E)} \tag{3.7}$$

where $P(A)$ is the joint probability of agreement and $P(E)$ is the joint probability of agreement under an assumption of independence. We believe this measure of dependence to be reasonable because,

$$\lim_{\beta_0\to\infty} \kappa(h) = \chi(h) = 2-\vartheta(\mathbf{s}_i,\mathbf{s}_j) \tag{3.8}$$

where β_0 is the intercept from $\mathbf{X}^T\boldsymbol{\beta}$ and $\vartheta(\mathbf{s}_i,\mathbf{s}_j)=\sum_{l=1}^{L}\left[w_l(\mathbf{s}_i)^{1/\alpha}+w_l(\mathbf{s}_j)^{1/\alpha}\right]^{\alpha}$ is the pairwise extremal coefficient given by Reich & Shaby [RS12] (see Appendix .1.2). In the case of complete dependence, $\kappa=1$, and in the case of complete independence, $\kappa=0$.

3.6 Computation

For small K we can evaluate the likelihood directly. When K is large, we use MCMC methods with the random effects model to explore the posterior distribution. This is possible because the expression for the joint density, conditional on $A_1,\ldots,A_L$, is given by

$$P[Y(\mathbf{s}_1)=y(\mathbf{s}_1),\ldots,Y(\mathbf{s}_n)=y(\mathbf{s}_n)]=\prod_{i=1}^{n}\pi(\mathbf{s}_i)^{1-Y_i}[1-\pi(\mathbf{s}_i)]^{Y_i}. \tag{3.9}$$

where

$$\pi(\mathbf{s}_i)=\exp\left\{-\sum_{l=1}^{L}A_l\left(\frac{w_l(\mathbf{s}_i)}{z(\mathbf{s}_i)}\right)^{1/\alpha}\right\}. \tag{3.10}$$

The model parameters and random effects are updated using a combination of Metropolis Hastings (MH) and Hamiltonian Monte Carlo (HMC) update steps. To overcome challenges with evaluating the positive stable density, we follow Reich & Shaby [RS12] and incorporate the auxiliary variable technique of Stephenson [Ste09].

3.7 Simulation study

For our simulation study, we generate $n_m=50$ datasets under 6 different settings to explore the impact of sample size and misspecification of link function. We generate data assuming three possible types of underlying process. For each process, we consider two sample sizes $n_s=650$ and $n_s=1300$.

The first of these processes is a max-stable process that uses the GEV link described in 3.4 with knots on a 21×21 grid on $[0,1]\times[0,1]$. For this process, we set $\alpha=0.3$, $\rho=0.025$, $\xi=0$ for idenfiability purposes, and β_0 is set for each dataset to give 5% rarity. We then set $Y(\mathbf{s})=I[z(\mathbf{s})>0]$ where $I[\cdot]$ is an indicator function.

For the second process, we generate a latent variable from a spatial Gaussian process with a

mean of logit(0.05) ≈ -2.9444 and an exponential covariance given by

$$\text{cov}(\mathbf{s}_1, \mathbf{s}_2) = \tau^2_{\text{Gau}} * \exp\left\{-\frac{||\mathbf{s}_1 - \mathbf{s}_2||}{\rho_{\text{Gau}}}\right\} \tag{3.11}$$

where $\tau_{\text{Gau}} = 7$ and $\rho_{\text{Gau}} = 0.10$. The mean of the Gaussian process is set to give approximately 5% rarity. Finally, we generate $Y(\mathbf{s}) \overset{ind}{\sim} \text{Bern}[\pi(\mathbf{s})]$ where $\pi(\mathbf{s}) = \exp\left\{\frac{z(\mathbf{s})}{1+z(\mathbf{s})}\right\}$

For the third process, we generate data using a hotspot method. For this process, we first generate hotspots throughout the space, and then set the probability of occurrence to be higher when a site is within a circle of radius $\rho = 0.05$ from a hotspot location. More specifically, generate $K \sim \text{Poisson}(9)$ hotspot locations, $\mathbf{v}^*_1, \ldots, \mathbf{v}^*_K$, from a uniform distribution over $[0,1] \times [0,1]$. If $||\mathbf{s}_i - \mathbf{v}^*_k|| < 0.05$ for any k, then $\pi(\mathbf{s}_i) = 0.70$ otherwise, $\pi(\mathbf{s}_i) = 0.01$. We then generate $Y(\mathbf{s}_i) \overset{ind}{\sim} \text{Bern}[\pi(\mathbf{s}_i)]$.

For each dataset, we fit the model using three different methods, spatial logistic regression, spatial probit regression, and the proposed spatial GEV method.

3.7.1 Spatial logistic and probit methods

Because logistic and probit methods represent two of the more common spatial techniques for binary data, we chose to compare our method to them. One way these methods differ from our proposed method is that they assume the underlying process is Gaussian. In this case, we assume that $Z(\mathbf{s})$ follows a Gaussian process with mean $\mathbf{X}(s)^T\boldsymbol{\beta}$ and exponential covariance function. The marginal distributions are given by

$$P(Y=1) = \begin{cases} \frac{\exp[\mathbf{X}^T\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\alpha}]}{1+\exp[\mathbf{X}^T\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\alpha}]}, & \text{logistic} \\ \phi(\mathbf{X}^T\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\alpha}), & \text{probit} \end{cases} \tag{3.12}$$

where $\boldsymbol{\alpha}$ are Gaussian random effects at the knot locations, and the $\mathbf{W}$ are basis functions to recreate the Gaussian process at all sites.

3.7.2 Cross validation

For each dataset, we fit the model using 500 of the observations as a training set, and the remaining observations are used as a validation set to assess the model's predictive power. Because our goal is to predict a the occurrence of an event, we use Brier scores to compare the models [GR07]. The Brier score for predicting an occurrence at site $\mathbf{s}$ is given by $\{I[Y(\mathbf{s}) = 1] - P[Y(\mathbf{s}) = 1]\}^2$ where $I[Y(\mathbf{s}) = 1]$ is an indicator function indicating that an event occurred at site $\mathbf{s}$, and $P[Y(\mathbf{s}) = 1]$ is obtained by taking the median of the posterior distribution. We average the Brier scores over all test sites, and a

Table 3.1 Relative Brier scores for GEV and Probit methods

	GEV	Probit
Setting 1	0.9047	0.9754
Setting 2	0.7885	0.9804
Setting 3	1.0275	1.0018
Setting 4	1.0264	1.0089
Setting 5	1.0458	0.9963
Setting 6	1.0565	0.9945

Table 3.2 Relative AUC for GEV and Probit methods

	GEV	Probit	Logit
Setting 1	0.8998	0.8973	0.8897
Setting 2	0.9458	0.9399	0.9356
Setting 3	0.7288	0.7371	0.7157
Setting 4	0.7906	0.8056	0.8115
Setting 5	0.8426	0.8458	0.8388
Setting 6	0.8756	0.8686	0.8765

lower score indicates a better fit.

We also consider the receiver operating characteristic (ROC) curve and look at the area under the curve (AUC) for the different methods and settings. AUC is commonly used as a metric to determine how well a model can classify data. When AUC = 1, the model perfectly classifies all observations in the test set, and when AUC = 0, the model misclassifies all observations in the test set. AUC is computed using the `roc` function in the `pROC` package of `R` using the median of the posterior predictive distribution at the testing locations. We then average AUCs across all datasets for each method and setting to obtain a single AUC for each combination of method and setting.

3.7.3 Results

Table 3.1 gives the Brier score relative to the Brier score for the spatial logistic method calculated as

$$\mathrm{BS}_{\mathrm{rel}} = \frac{\mathrm{BS}_{\mathrm{method}}}{\mathrm{BS}_{\mathrm{logistic}}}. \tag{3.13}$$

Table 3.2 gives the AUC relative to the AUC for the spatial logistic method calculated in similarly to the relative Brier score.

We analyzed the results for this simulation study using a Friedman test at $\alpha = 0.05$ to see if at least one method had a significantly different Brier score or AUC. For any setting that yielded a

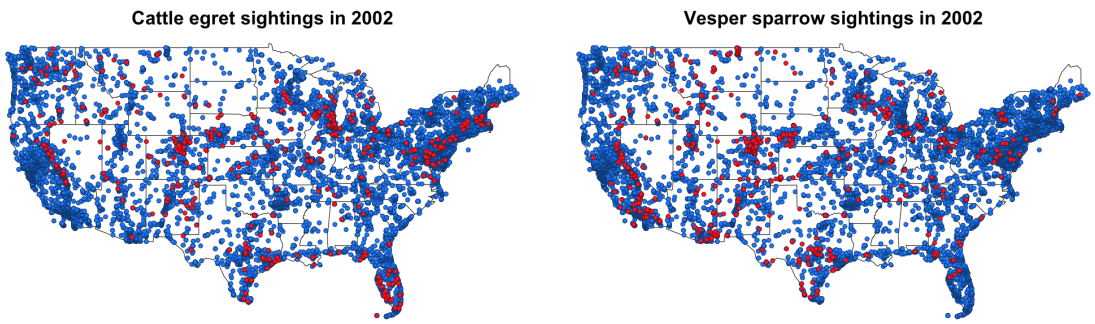

Figure 3.1 Reported sighting for Cattle egret (left) and Vesper sparrow (right) in 2002

significant p-value, we conducted a Wilcoxon-Nemenyi-McDonald-Thompson test to see which of the methods had different results. The full results for the Wilcoxon-Nemenyi-McDonald-Thompson tests are given in Appendix B.1. For all settings, we find significant results for the Friedman test comparing the Brier scores for the methods. Specifcally, we see a statistically significant reduction in Brier score using the GEV compared to logit for settings 1 and 2 and compared to probit for setting 2. However, in the other settings, the logit and probit methods tend to perform better than the GEV method.

The results using AUC are much less conclusive with only settings 1 and 4 demonstrating significant differences between the methods at $\alpha = 0.05$. As with the Brier scores, the GEV method shows a statistically significant increase in AUC over the logit method for setting 1, and for setting 4, the both the probit and logit methods show a statistically significant improvement in AUC over the GEV method.

3.8 Data analysis

For the data analysis, we consider data from the eBirds dataset, a citizen-based observation network of bird sitings in the United States [Sul09]. The data are publicly available from `http://ebird.org`. We use data from 2002, and focus on 10 different species. Figure 3.1 shows the sighting data for cattle egrets and vesper sparrows

b

3.9 Conclusions

Acknowledgments

APPENDICES

.1 Appendices

.1.1 Derivation of the likelihood

We use the hierarchical max-stable spatial model given by Reich & Shaby [RS12]. If at each margin, $Z_i \sim \text{GEV}(1,1,1)$, then $Z_i|\theta_i \overset{indep}{\sim} \text{GEV}(\theta, \alpha\theta, \alpha)$. We reorder the data such that $Y_1 = \ldots = Y_K = 1$, and $Y_{K+1} = \ldots = Y_n = 0$. Then the joint likelihood conditional on the random effect θ is

$$\begin{aligned}
P(Y_1 = y_1, \ldots, Y_n = y_n) &= \prod_{i \le K} \left\{ 1 - \exp\left[-\left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \right\} \prod_{i > K} \exp\left[-\left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \\
&= \exp\left[-\sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] - \exp\left[-\sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \sum_{i=1}^{K} \exp\left[-\left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \\
&\quad + \exp\left[-\sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \sum_{1 < i < j \le K} \left\{ \exp\left[-\left(\frac{\theta_i}{z_i} \right)^{1/\alpha} - \left(\frac{\theta_j}{z_j} \right)^{1/\alpha} \right] \right\} \\
&\quad + \cdots + (-1)^K \exp\left[-\sum_{i=1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \qquad (14)
\end{aligned}$$

Finally marginalizing over the random effect, we obtain

$$\begin{aligned}
P(Y_1 = y_1, \ldots, Y_n = y_n) &= \int G(\mathbf{z}|\mathbf{A}) p(\mathbf{A}|\alpha) d\mathbf{A}. \\
&= \int \exp\left[-\sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] - \exp\left[-\sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \sum_{i=1}^{K} \exp\left[-\left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \\
&\quad + \exp\left[-\sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] \sum_{1 < i < j \le K} \left\{ \exp\left[-\left(\frac{\theta_i}{z_i} \right)^{1/\alpha} - \left(\frac{\theta_j}{z_j} \right)^{1/\alpha} \right] \right\} \\
&\quad + \cdots + (-1)^K \exp\left[-\sum_{i=1}^{n} \left(\frac{\theta_i}{z_i} \right)^{1/\alpha} \right] p(\mathbf{A}|\alpha) d\mathbf{A}. \qquad (15)
\end{aligned}$$

Consider the first term in the summation,

$$\int \exp\left\{-\sum_{i=K+1}^{n}\left(\frac{\theta_i}{z_i}\right)^{1/\alpha}\right\} p(\mathbf{A}|\alpha)d\mathbf{A} = \int \exp\left\{-\sum_{i=K+1}^{n}\left(\frac{\left[\sum_{l=1}^{L} A_l\, w_l(\mathbf{s}_i)^{1/\alpha}\right)^{\alpha}\right]^{1/\alpha}}{z_i}\right\} p(\mathbf{A}|\alpha)d\mathbf{A}$$

$$= \int \exp\left\{-\sum_{i=K+1}^{n}\sum_{l=1}^{L} A_l\left(\frac{w_l(\mathbf{s}_i)}{z_i}\right)^{1/\alpha}\right\} p(\mathbf{A}|\alpha)d\mathbf{A}$$

$$= \exp\left\{-\sum_{l=1}^{L}\left[\sum_{i=K+1}^{n}\left(\frac{w_l(\mathbf{s}_i)}{z_i}\right)^{1/\alpha}\right]^{\alpha}\right\}. \tag{16}$$

The remaining terms in equation (15) are straightforward to obtain, and after integrating out the random effect, the joint density for $K = 0, 1, 2$ is given by

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \begin{cases} G(\mathbf{z}) & K = 0 \\ G(\mathbf{z}_{(1)}) - G(\mathbf{z}) & K = 1 \\ G(\mathbf{z}_{(12)}) - G(\mathbf{z}_{(1)}) - G(\mathbf{z}_{(2)}) + G(\mathbf{z}) & K = 2 \end{cases} \tag{17}$$

where

$$G[\mathbf{z}_{(1)}] = P[Z(\mathbf{s}_2) < z(\mathbf{s}_2), \ldots, Z(\mathbf{s}_n) < z(\mathbf{s}_n)]$$
$$G[\mathbf{z}_{(2)}] = P[Z(\mathbf{s}_1) < z(\mathbf{s}_1), Z(\mathbf{s}_3) < z(\mathbf{s}_3), \ldots, Z(\mathbf{s}_n) < z(\mathbf{s}_n)]$$
$$G[\mathbf{z}_{(12)}] = P[Z(\mathbf{s}_3) < z(\mathbf{s}_3), \ldots, Z(\mathbf{s}_n) < z(\mathbf{s}_n)].$$

Similar expressions can be derived for all K, but become cumbersome for large K.

.1.2 Derivation of the χ statistic

$$
\begin{aligned}
\chi &= \lim_{p\to 0} \mathrm{P}(Y_i = 1 | Y_j = 1) \\
&= \lim_{p\to\infty} \frac{p + p - \left(1 - \exp\left\{-\sum_{l=1}^{L}\left[\left(-\log(1-p)w_l(\mathbf{s}_i)\right)^{1/\alpha} + \left(-\log(1-p)w_l(\mathbf{s}_j)\right)^{1/\alpha}\right]^{\alpha}\right\}\right)}{p} \\
&= \lim_{p\to 0} \frac{2p - \left(1 - \exp\left\{\log(1-p)\sum_{l=1}^{L}\left[w_l(\mathbf{s}_i)^{1/\alpha} + w_l(\mathbf{s}_j)^{1/\alpha}\right]^{\alpha}\right\}\right)}{p} \\
&= \lim_{p\to 0} \frac{2p - \left(1 - (1-p)^{\sum_{l=1}^{L}\left[\left(w_l(\mathbf{s}_i)\right)^{1/\alpha} + \left(w_l(\mathbf{s}_j)\right)^{1/\alpha}\right]^{\alpha}}\right)}{p} \\
&= \lim_{p\to 0} 2 - \sum_{l=1}^{L}\left[w_l(\mathbf{s}_i)^{1/\alpha} + w_l(\mathbf{s}_j)^{1/\alpha}\right]^{\alpha}(1-p)^{-1+\sum_{l=1}^{L}\left[w_l(\mathbf{s}_i)^{1/\alpha} + w_l(\mathbf{s}_j)^{1/\alpha}\right]^{\alpha}} \\
&= 2 - \sum_{l=1}^{L}\left[w_l(\mathbf{s}_i)^{1/\alpha} + w_l(\mathbf{s}_j)^{1/\alpha}\right]^{\alpha}. \qquad (18)
\end{aligned}
$$

APPENDICES

.1 Simulation study pairwise difference results

The following tables show the methods that have significantly different Brier scores when using a Wilcoxon-Nemenyi-McDonald-Thompson test. In each column, different letters signify that the methods have significantly different Brier scores.

Table 3 Pairwise BS comparisons

	Setting 1	Setting 2	Setting 3	Setting 4	Setting 5	Setting 6
Method 1	A	A	A	C	B	B
Method 2	A B	B	A	B	A	A
Method 3	B	B	A	A	A B	A

APPENDIX

A

EMPIRICAL BASIS FUNCTIONS FOR MAX-STABLE SPATIAL DEPENDENCE

A.1 Introduction

A.2 Model

Let $Y_t(\mathbf{s})$ be the observation at spatial location $\mathbf{s}$ and time t. We temporarily drop the subscript t and describe the model for the process $Y(\mathbf{s})$ for a single time point, but return to the spatiotemporal setting in Section A.3. To focus attention on the extreme values, we emphasize the statistical model for exceedances above a location-specific threshold $T(\mathbf{s})$. We begin by specifying a spatial model for the complete data $Y(\mathbf{s})$ and then use the censored likelihood defined by $T(\mathbf{s})$ for inference as described in Section A.4. Although the model presented implements a censored likelihood, the model also can fit uncensored data (such as block-maxima) by setting $T(\mathbf{s}) = -\infty$.

Spatial dependence is captured by modeling $Y(\mathbf{s})$ as a max-stable process (ref). Max-stable processes have generalized extremal value (GEV; see Appendix A.1) marginal distribution. The GEV has three parameters: location $\mu(\mathbf{s})$; scale $\sigma(\mathbf{s})$; and shape $\xi(\mathbf{s})$. Spatial dependence is present both

in the GEV parameters but also the standardized residual process

$$Z(\mathbf{s}) = \left\{1 + \frac{\xi(\mathbf{s})}{\sigma(\mathbf{s})}\big[Y(\mathbf{s}) - \mu(\mathbf{s})\big]\right\}^{1/\xi(\mathbf{s})}, \tag{A.1}$$

which has unit Fréchet (i.e., GEV with location, scale, and shape all equal one) marginal distribution for all $\mathbf{s}$.

Our objective is to identify a low-rank model for the spatial dependence of $Z(\mathbf{s})$. The spectral representation theorem (ref) states that any max-stable process can be written

$$Z(\mathbf{s}) = \sup_l B(\mathbf{s}, \mathbf{t}_l) A_l \tag{A.2}$$

where the function B satisfies $B(\mathbf{s},\mathbf{t}) > 0$ for all $(\mathbf{s},\mathbf{t})$ and $\int B(\mathbf{s},\mathbf{t})d\mathbf{t} = 1$ for all $\mathbf{s}$, and $(\mathbf{t}_l, A_l)$ for $l = 1, ..., \infty$ are a Poisson process with intensity measure $dAd\mathbf{t}/A^2$. This representation provides a means of truncation. Ref propose the max-linear model

$$Z(\mathbf{s}) = \bigvee_{l=1}^{L} B_l(\mathbf{s}) A_l \tag{A.3}$$

where $B_l(\mathbf{s}) > 0$, $\int B_l(\mathbf{s})d\mathbf{s} = 1$ for all L, and A_l are independent Fréchet random variables.

The assumption that $Z(\mathbf{s})$ equals exactly the maximum of a small number of functions is unrealistic, especially for data measured with error. We therefore follow the Reich and Shaby (ref) and decompose $Z(\mathbf{s})$ as $Z(\mathbf{s}) = \theta(\mathbf{s})\varepsilon(\mathbf{s})$ where $\theta(\mathbf{s})$ is a spatial process and $\varepsilon(\mathbf{s}) \overset{iid}{\sim} \text{GEV}(1,\alpha,\alpha)$ is independent error. The spatial component is

$$\theta(\mathbf{s}) = \left(\sum_{l=1}^{L} B_l(\mathbf{s})^{1/\alpha} A_l\right)^{\alpha}. \tag{A.4}$$

If $B_l(\mathbf{s}) > 0$, $\sum_{l=1}^{L} B_l(\mathbf{s}) = 1$ for all $\mathbf{s}$, and the A_l have positive stable (PS; Appendix A.1) distribution $A_l \overset{iid}{\sim} \text{PS}(\alpha)$, then $Z(\mathbf{s})$ is max-stable and has unit Fréchet marginal distributions.

Extremal spatial dependence can be summarized by the extremal coefficient (EC; ref) $\vartheta(\mathbf{s},\mathbf{t}) \in [1,2]$, where

$$\text{Prob}[Z(\mathbf{s}) < c, Z(\mathbf{t}) < c] = \text{Prob}[Z(\mathbf{s}) < c]^{\vartheta(\mathbf{s},\mathbf{t})}. \tag{A.5}$$

For the PS random effects model the EC has the form

$$\vartheta(\mathbf{s},\mathbf{t}) = \sum_{l=1}^{L}\big[B_l(\mathbf{s})^{1/\alpha} + B_l(\mathbf{s})^{1/\alpha}\big]^{\alpha}. \tag{A.6}$$

In particular, $\vartheta(\mathbf{s},\mathbf{s}) = 2^{\alpha}$ for all $\mathbf{s}$.

A.3 Estimating the spatial dependence function

To estimate the extremal coefficient function, we consider the process at n_s spatial locations $\mathbf{s}_1, ..., \mathbf{s}_{n_s}$ and n_t times $t = 1, ..., n_t$. Denote $Y_t(\mathbf{s}_i) = Y_{it}$, $B_l(\mathbf{s}_i) = B_{il}$, $T(\mathbf{s}_i) = T_i$, and $\vartheta(\mathbf{s}_i, \mathbf{s}_j) = \vartheta_{ij}$. In this section we develop an algorithm to estimate the spatial dependence parameter α and the $n_s \times L$ matrix $\mathbf{B} = \{B_{il}\}$. Given these parameters, we insert them into our model and proceed with Bayesian analysis as described in Section A.4. Our algorithm has the following steps:

(1) Obtain an initial estimate of the extremal coefficient for each pair of locations, $\hat{\vartheta}_{ij}$.

(2) Spatially smooth these initial estimates $\hat{\vartheta}_{ij}$ using kernel smoothing to obtain $\tilde{\vartheta}_{ij}$.

(3) Estimate the spatial dependence parameters by minimizing the difference between model-based coefficients, ϑ_{ij}, and smoothed coefficients, $\tilde{\vartheta}_{ij}$.

The first-stage estimates are obtained using the approach of citation for pairwise estimates. To estimate the spatial dependence we first remove variation in the marginal distribution. Let $U_{it} = \sum_{k=1}^{n_t} I[Y_{ik} < Y_{it}]/n_t$, so that the U_{it} are approximately uniform at each location. Then for some extreme probability $q \in (0,1)$, solving (A.5) suggests the estimate

$$\hat{\vartheta}_{ij}(q) = \frac{\log[Q_{ij}(q)]}{\log(q)}, \tag{A.7}$$

where $Q_{ij}(q) = \sum_{t=1}^{n_t} I[U_{it} < q, U_{jt} < q]/n_t$ is the sample proportion of the time points at which both sites are less than q. Since all large q give valid estimates, we average over a grid of q with $q_1 < ... < q_{n_q}$

$$\hat{\vartheta}_{ij} = \frac{1}{n_q}\sum_{j=1}^{n_q} \hat{\vartheta}_{ij}(q_j). \tag{A.8}$$

Assuming the true EC is smooth over space, the initial estimates $\hat{\vartheta}_{ij}$ can be improved by smoothing. Let

$$\tilde{\vartheta}_{ij} = \frac{\sum_{u=1}^{n_s}\sum_{v=1}^{n_s} w_{iu} w_{jv} \hat{\vartheta}_{uv}}{\sum_{u=1}^{n_s}\sum_{v=1}^{n_s} w_{iu} w_{jv}}, \tag{A.9}$$

where $w_{iu} = \exp[-(||\mathbf{s}_i - \mathbf{s}'_u||/\phi)^2])$ is the Gaussian kernel function with bandwidth ϕ. The elements $\hat{\vartheta}_{ii}$ do not contribute any information as $\hat{\vartheta}_{ii} = 1$ for all i by construction. To eliminate the influence

of these estimates we set $w_{ii} = 0$. However, this approach does give imputed values $\tilde{\vartheta}_{ii}$, which provide information about small-scale spatial variability.

The dependence parameters are estimated by comparing estimates $\tilde{\vartheta}_{ij}$ with the model-based values ϑ_{ij}. For all i, $\vartheta_{ii} = 2^{\alpha}$, and therefore we set α to $\hat{\alpha} = \log_2(\sum_{i=1}^{n_s} \tilde{\vartheta}_{ii}/n_s)$. Given $\alpha = \hat{\alpha}$, it remains to estimate $\mathbf{B}$. The estimate $\hat{\mathbf{B}}$ is the minimizer of

$$\sum_{i<j} \left(\tilde{\vartheta}_{ij} - \vartheta_{ij}\right)^2 = \sum_{i<j} \left(\tilde{\vartheta}_{ji} - \sum_{l=1}^{L} [B_{il}^{1/\hat{\alpha}} + B_{jl}^{1/\hat{\alpha}}]^{\hat{\alpha}} \right)^2 \tag{A.10}$$

under the restrictions that $B_{il} \geq 0$ for all i and l and $\sum_{l=1}^{L} B_{il} = 1$ for all i. Since the minimizer of (A.10) does not have a closed form, we use block coordinate descent to obtain $\hat{\mathbf{B}}$. We cycle through spatial locations and update the vectors $(B_{i1}, ..., B_{iL})$ conditioned on the values for the other location and repeat until convergence. At each step, we use the restricted optimization routine in the R (ref) function `optim`. This algorithm gives estimates of the B_{il} at the n_s data locations, but is easily extended to all $\mathbf{s}$ for spatial prediction. The kernel smoothing step ensures that the estimates for $\hat{B}_{il}$ are spatially smooth, and thus interpolation of the $\hat{B}_{il}$ gives spatial functions $\hat{B}_l(\mathbf{s})$.

The relative contribution of each term can be measured by

$$v_l = \frac{1}{n_s} \sum_{i=1}^{n_s} \hat{B}_{il}. \tag{A.11}$$

Since $\sum_{l=1}^{L} \hat{B}_{il} = 1$ for all i, we have $\sum_{l=1}^{L} v_l = 1$. Therefore, terms with large v_l are the most important. The order of the terms is arbitrary, and so we reorder the terms so that $v_1 \geq ... \geq v_L$.

A.4 Bayesian implementation details

For our data analysis in Section A.5 we allow the GEV location and scale parameters, denoted μ_{it} and scale σ_{it} respectively, to vary with space and time. The GEV shape parameter ξ is held constant over space and time because this parameter is notoriously difficult to estimate (ref). Collectively, let the marginal GEV parameters at location i and time t be $\Theta_{it} = \{\mu_{it}, \sigma_{it}, \xi\}$. The GEV location and scale vary according to covariates $\mathbf{X}_{it}$ with $\mu_{it} = \mathbf{X}_{it}^T \boldsymbol{\beta}_1$ and $\log(\sigma_{it}) = \mathbf{X}_{it}^T \boldsymbol{\beta}_2$.

As shown in Reich and Shaby (ref), the uncensored responses Y_{it} are conditionally independent given the spatial random effects, with conditional distribution

$$Y_{it} | \theta_{it}, \Theta_{it} \overset{indep}{\sim} GEV(\mu_{it}^*, \sigma_{it}^*, \xi^*), \tag{A.12}$$

where $\mu^*_{it} = \mu_{it} + \frac{\sigma_{it}}{\xi}(\theta_{it}^{\xi} - 1)$, $\sigma^*_{it} = \alpha\sigma_{it}\theta_{it}^{\xi}$, and $\xi^* = \alpha\xi$. Therefore, the conditional likelihood conveniently factors across observations; marginalizing over the random effect θ_{it} induces extremal spatial dependence. To focus on the extreme values above the local threshold T_i, we use the censored likelihood

$$d(y;\theta_{it},\Theta_{it},T_i) = \begin{cases} F(y;\mu^*_{it},\sigma^*_{it},\xi^*) & y \le T_i \\ f(y;\mu^*_{it},\sigma^*_{it},\xi^*) & y > T_i, \end{cases} \tag{A.13}$$

where F and f are the GEV distribution and density functions, respectively, defined in Appendix A.1.

In summary, given the estimates of α and $\mathbf{B}$, the hierarchical model is

$$\begin{aligned} Y_{it}|\theta_{ij} &\overset{indep}{\sim} d(y;\theta_{it},\Theta_{it},T_i) \\ \theta_{it} &= \left(\sum_{l=1}^{L} \hat{B}_{il}^{1/\hat{\alpha}} A_{lt}\right)^{\hat{\alpha}} \quad \text{where} \quad A_{lt} \overset{iid}{\sim} PS(\hat{\alpha}) \\ \mu_{it} &= \mathbf{X}_{it}^T\boldsymbol{\beta}_1 \quad \text{and} \quad \log(\sigma_{it}) = \mathbf{X}_{it}^T\boldsymbol{\beta}_2. \end{aligned} \tag{A.14}$$

To complete the Bayesian model, we select independent normal priors with mean zero and variance σ_1^2 and σ_2^2 for the components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ respectively, and $\xi \sim \text{Normal}(0, 0.5^2)$. We use InvGamma(1, 1) priors for σ_1^2 and σ_2^2. We estimate parameters $\Theta = \{A_{lt}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \xi, \sigma_1^2, \sigma_2^2\}$ using Markov chain Monte Carlo methods. We use a Metropolis-Hastings algorithm to update the model parameters with random walk candidate distributions for all parameters except σ_1^2 and σ_2^2 which we update using Gibbs sampling. The PS density is challenging to evaluate as it does not have a closed form. One technique to avoid this complication is to incorporate auxiliary random variables (Stephenson et al, ref), but we opt for a numerical approximation to the integral as described in Appendix which one.

The first-stage estimate of the extremal coefficients has three tuning parameters: the quantile thresholds $q_1, ..., q_{n_q}$, the kernel bandwidth ϕ, and the number of terms L. In Section A.5 we explore a few possibilities for ϕ and L and discuss sensitivity to these choices. The second-stage Bayesian analysis requires selecting thresholds $T_i, ..., T_{n_s}$. For this we use spatially smoothed sample quantiles. That is, we set T_i to the 0.95 quantile of the Y_{it} and Y_{jt} for sites j with $||\mathbf{s}_i - \mathbf{s}_j|| < r$, where r is set to XXX?

A.5 Data analysis

In this section, we illustrate our method using both points above a threshold and block maxima. In Section A.5.1, we present an analysis using annual acreage burned due to wildfires in Georgia from

1965 – 2014. This is followed in Section A.5.4 by an analysis of precipitation data in the eastern U.S.

A.5.1 Analysis of extreme Georgia fires

The dataset used for our application is composed of yearly acreage burned due to wildfires for each county in Georgia from 1965 – 2014 (`http://weather.gfc.state.ga.us/FireData/`). Figure A.1 shows the time series of log(acres burned) for 25 randomly selected counties. Based on this plot and other exploratory analysis, we see no evidence of non-linear trends and proceed with linear time trends for the GEV location and scale parameters. For covariates we use the standardized linear time trend $t^* = (t - n_t/2)/n_t$, and L bivariate Gaussian kernel functions $\widetilde{B}_{il}$, and their interactions: $\mathbf{X}_{it} = (1, t^*, \widetilde{B}_{i1}, ..., \widetilde{B}_{iL}, t^*\widetilde{B}_{i1}, ..., t^*\widetilde{B}_{iL})^T$. For the bivariate Gaussian kernels, we select L knot locations, $\mathbf{v}_1, \ldots, \mathbf{v}_L$ from the county centroids, using a space-filling design (ref add to bibtex Johnson, M.E., Moore, L.M., and Ylvisaker, D. ,1990). Then

$$\widetilde{B}_{il} = \exp\left\{-\frac{||\mathbf{s}_i - \mathbf{v}_l||^2}{\rho^2}\right\} \tag{A.15}$$

where ρ is included as an unknown parameter with a $U(\rho_l, \rho_u)$ prior where ρ_l is the 5th quantile of $||\mathbf{s}_i - \mathbf{v}_l||^2$ and ρ_u is the 95th quantile of $||\mathbf{s}_i - \mathbf{v}_l||^2$ basically I took the 5th and 95th quantile of the squared distances between the sites and the knots.

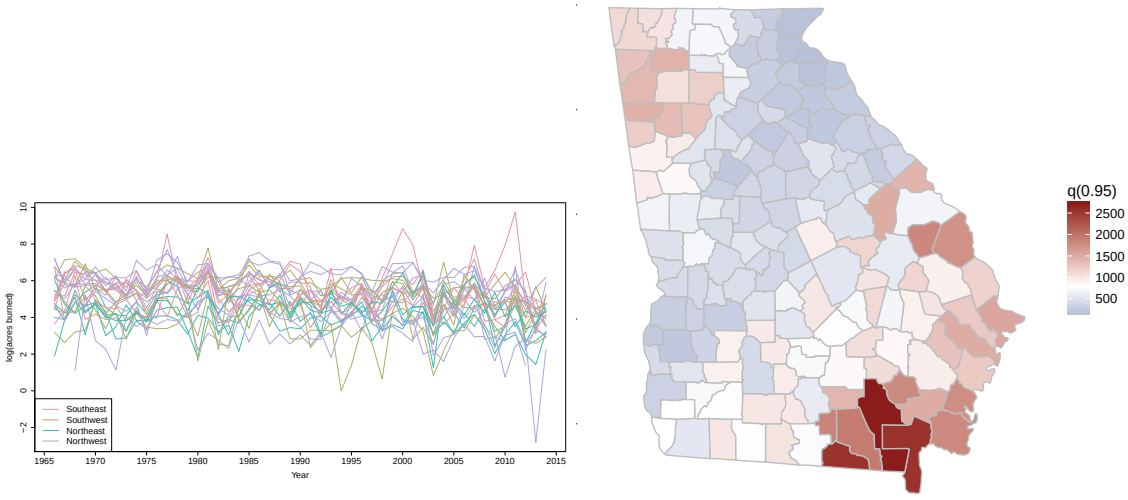

Figure A.1 Time series of log acres burned for 25 randomly selected counties with colors coding the county's quadrant (left), and spatially smoothed threshold values, T_i for each county (right).

We estimate the extremal coefficient function $\hat{\theta}_{ij}$ by setting $q_1 = 0.90$ and using $n_q = 100$. With more data, it would possible to increase q_1, but we set $q_1 = 0.90$ to increase the stability when estimating $\hat{\theta}_{ij}$.

Because these data are not block-maxima, we select a site-specific threshold T_i to use in the analysis with the following algorithm. Without some adjustment to the data, it is challenging to borrow information across sites to inform the threshold selection. We first standardize the data, separated by county, by subtracting the site's median and dividing by the site's interquartile range. Denote the standardized data by $\widetilde{\mathbf{Y}}_i$. Then we combine all sites together and plot a mean residual plot for $\widetilde{Y}_{it}, i = 1,\ldots,n_s$ and $t = 1,\ldots,n_t$. The mean residual plot is given in Figure A.2. Based upon the mean residual plot, we select the 95th percentile for the threshold. To calculate T_i for each county, we use the 95th percentile for the combined data for county i and its five closest counties.

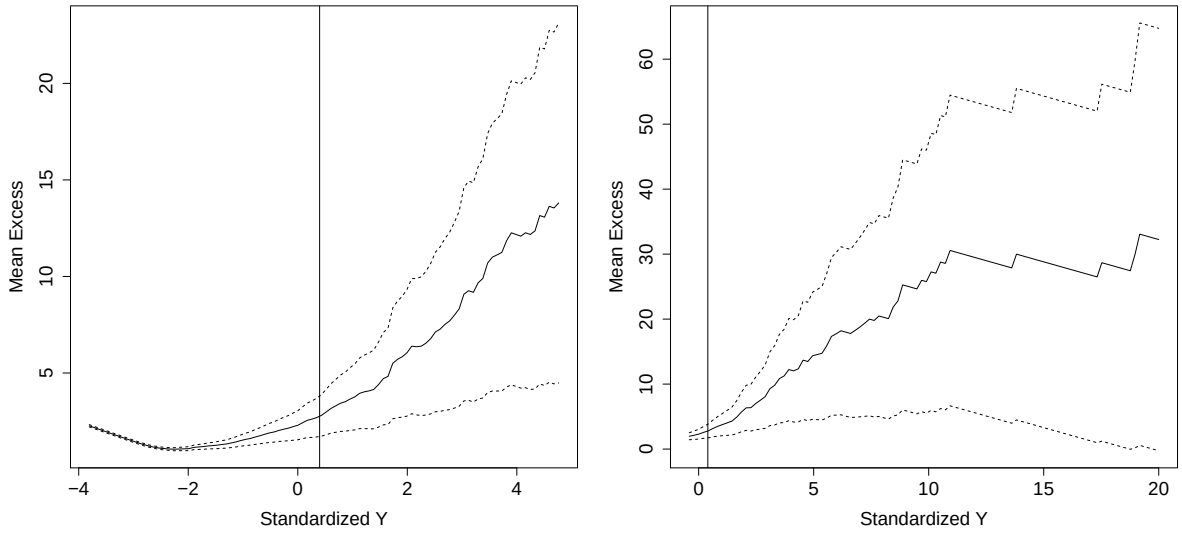

Figure A.2 Mean residual plot for the data pooled across counties after standardizing using the county's median and interquartile range. The two panels show different ranges on the x-axis and include a vertical line at the sample 95th percentile.

A.5.2 Results for fire analysis

We use 10-fold cross-validation to assess the predictive performance of a model. For each method, we randomly select 90% of the observations across counties and years to be used as a training set to fit the model. The remaining 10% of sites and years are withheld for testing model predictions. To assess the predictions for the test set, we use quantile scores and Brier scores citation. The quantile score is given by give formula. The Brier score is given by give formula. For both of these methods, we use a negative orientation, so a lower score indicates a better fit. The Brier and quantile scores for the fire analysis are given in Table A.1.

Based on the Brier scores and quantiles scores, we run a full analysis using all of the data with $L = 35$. Posterior summaries for each county's β_{time} coefficient are given in Figures A.3 and A.4. These plots both seem to catch similar features with some differences particularly in the posterior distribution of the county-specific $\beta_{\mu,\text{time}}$.

Table A.1 Average Brier scores (×100) for selected thresholds and quantile scores for selected quantiles for fire analysis Need timings for L = 35, 40 to be added in

		Brier Scores (×100)		Quantile Scores		Time (in hours)
	Process	$q(0.95)$	$q(0.99)$	$q(0.95)$	$q(0.99)$	
L = 5	EBF	5.640	2.265	135.685	80.471	1.1
	GSK	5.726	2.301	134.419	78.639	1.1
L = 10	EBF	5.329	2.130	127.313	75.974	1.8
	GSK	5.311	2.142	127.593	75.123	1.9
L = 15	EBF	4.997	2.043	128.277	68.946	2.7
	GSK	4.907	2.034	124.537	**59.266**	2.7
L = 20	EBF	4.930	2.036	122.394	66.413	3.7
	GSK	4.864	2.043	121.145	62.172	3.7
L = 25	EBF	4.776	**1.920**	116.944	61.704	4.8
	GSK	4.740	1.921	113.872	59.524	4.7
L = 30	EBF	4.745	1.923	114.878	62.020	5.9
	GSK	4.719	1.936	114.918	61.905	5.7
L = 35	EBF	4.761	1.920	115.696	62.581	
	GSK	4.767	1.933	114.026	60.805	
L = 40	EBF	4.722	1.935	115.213	62.039	
	GSK	**4.716**	1.921	**113.362**	60.300	

We need a plot of the first 5 basis functions, and we also talked about the % variability. I wasn't sure if this should be for $L = 35$ or a different number of knots.

Based upon the cross-validation results, we reran the full data analysis using $L = 15$ basis

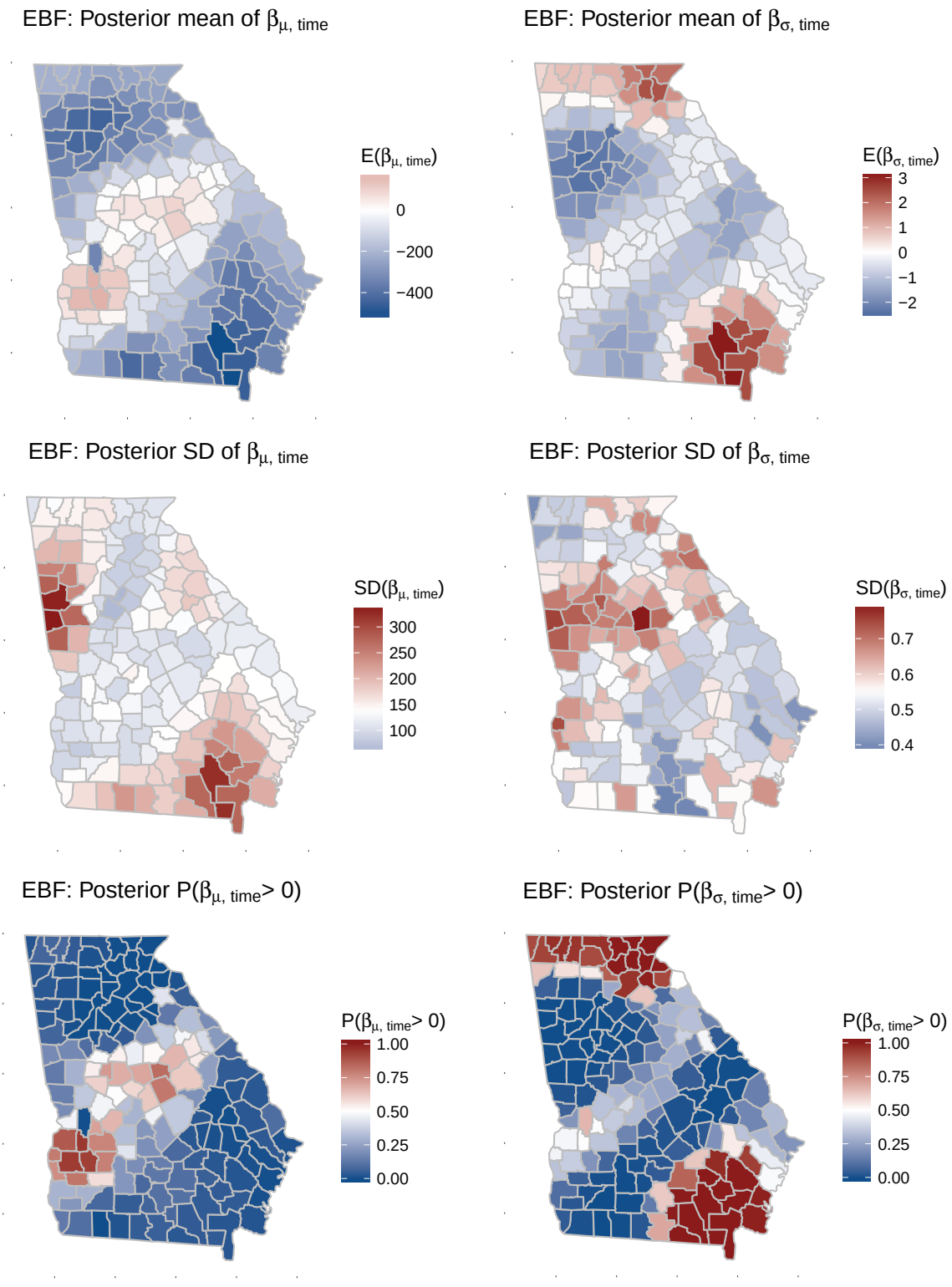

Figure A.3 Posterior summaries of β_{time} when using EBF for the spatial process with $L = 35$.

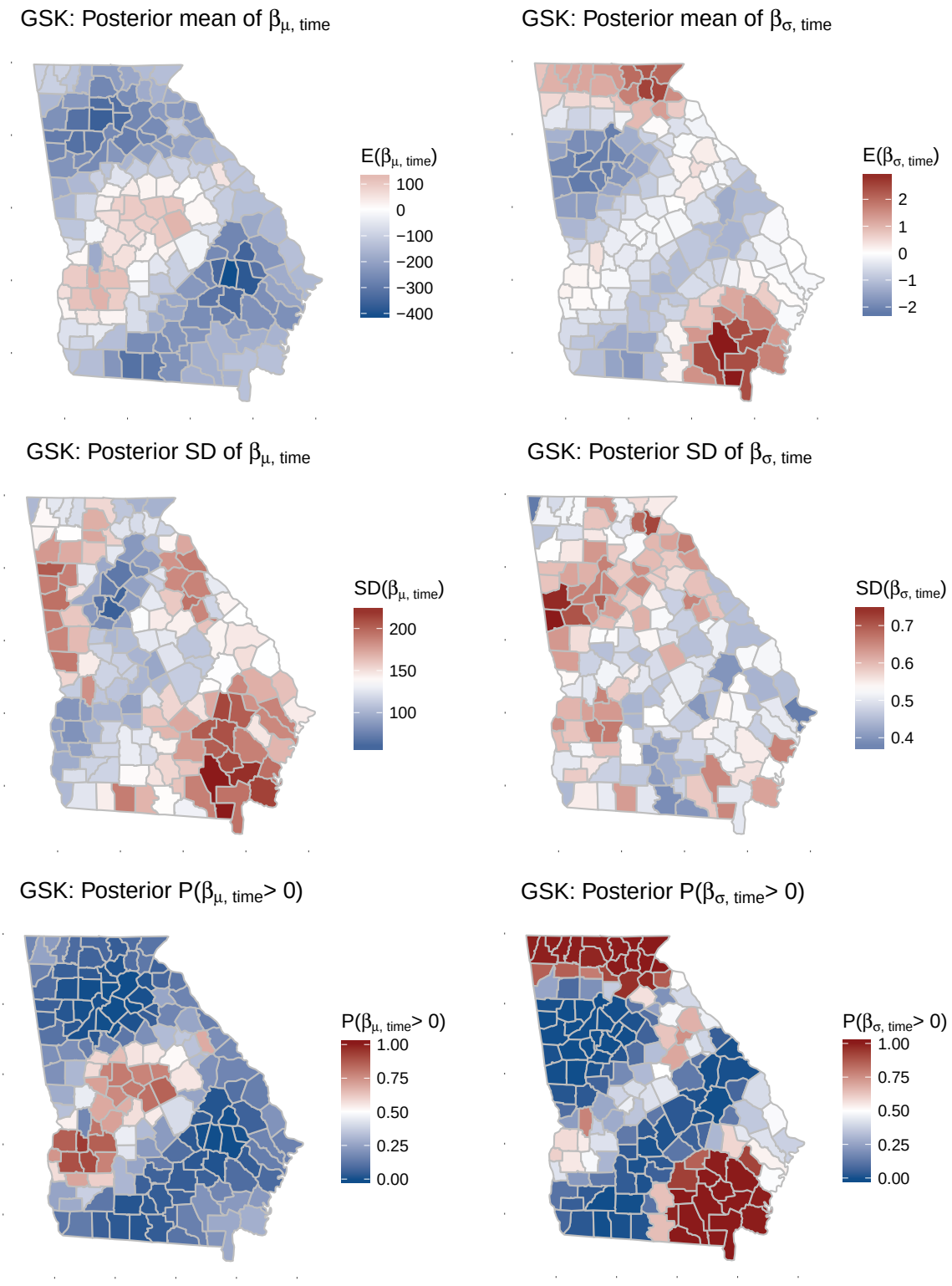

Figure A.4 Posterior summaries of β_{time} when using GSK for the spatial process with $L = 35$.

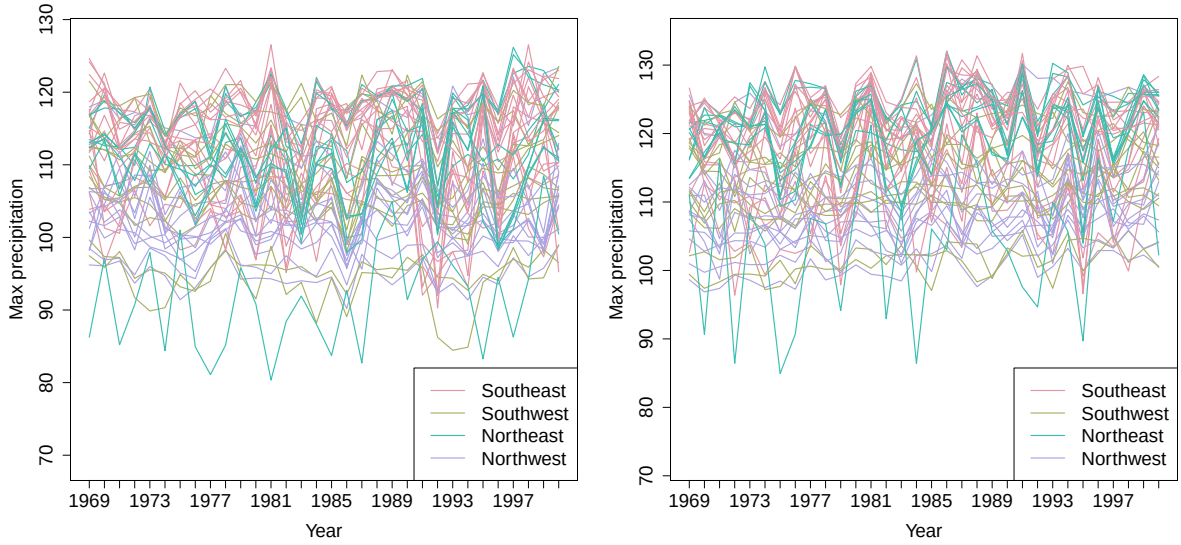

Figure A.5 Time series of yearly max precipitation for current (1969 – 2000) (left). Time series of yearly max precipitation for future (2039 – 2070) (right).

functions. Figure here with panel of location & scale: mean, sd, and P($\beta_t > 0$)

A.5.3 Model checking and sensitivity analysis

A.5.4 Analysis of annual precipitation

We also conduct an analysis of the precipitation data presented in [RS12]. The data are climate model output from the North American Regional Climate Change Assessment Program (NARCCAP). This data consists of $n_s = 697$ grid cells at a 50km resolution in the eastern US, and includes historical data (1969 – 2000) as well as future conditions (2039 – 2070).

Include figures of locations of grid cells

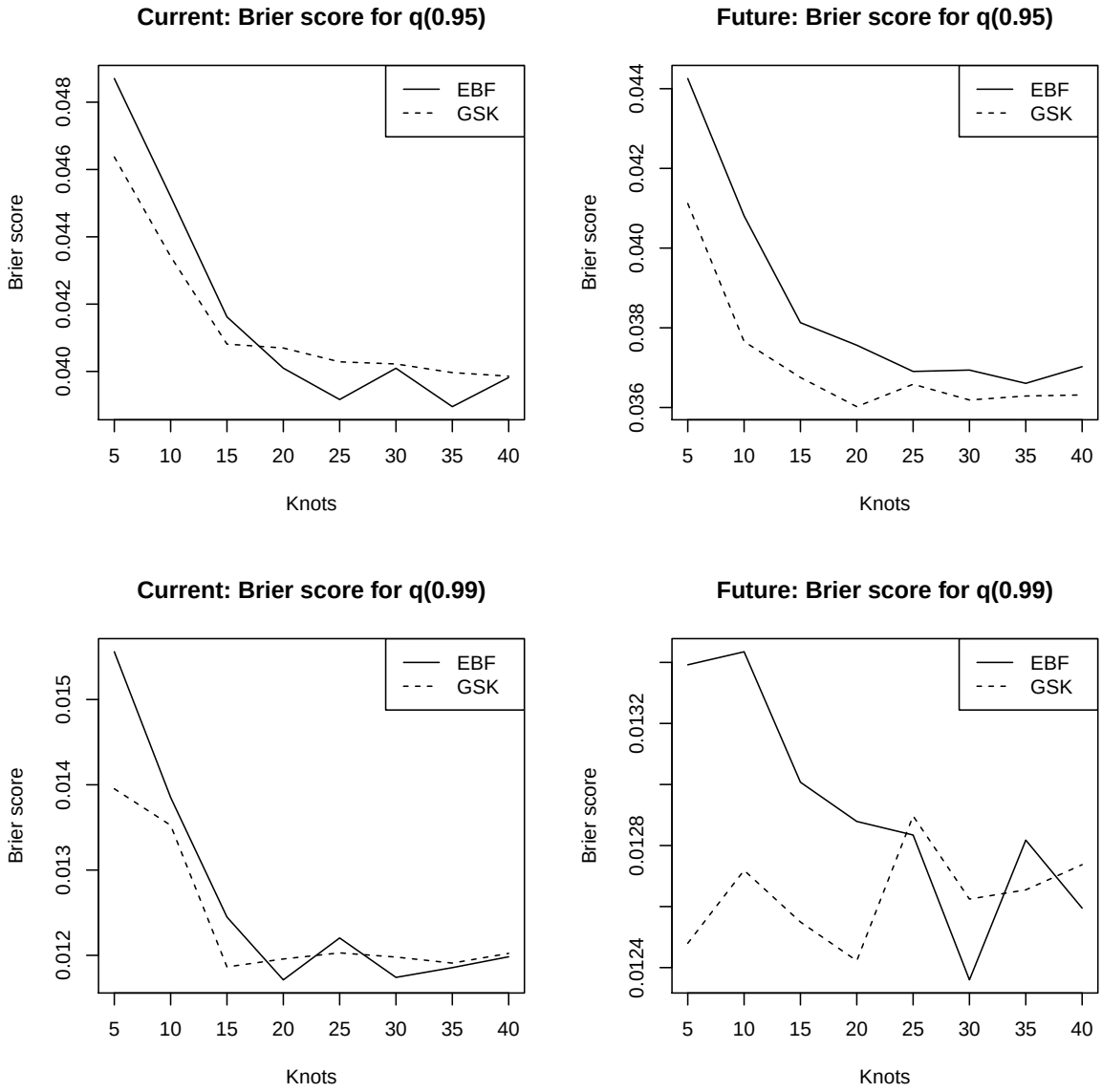

Figure A.6 Brier scores for current and future precipitation analysis.

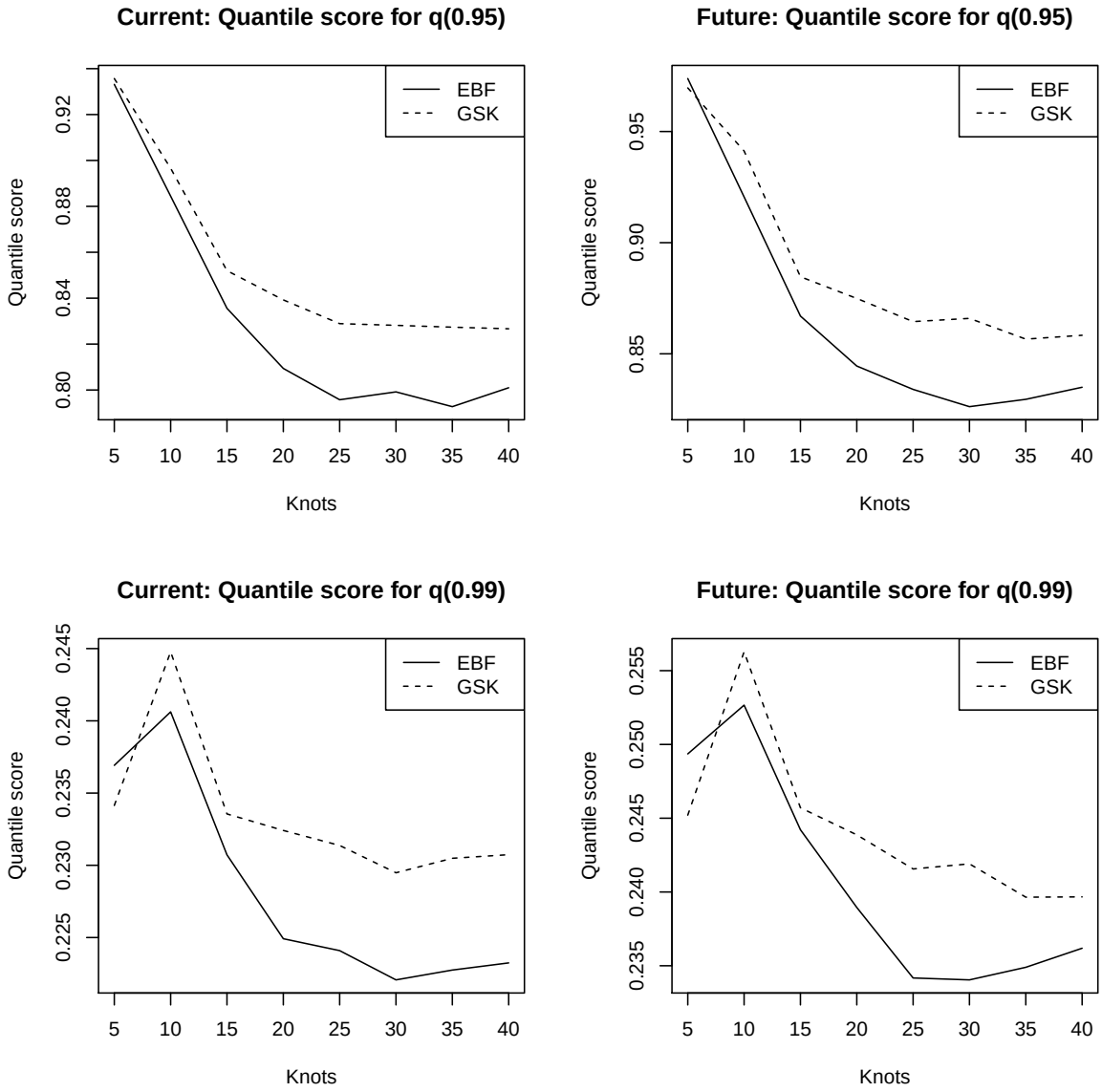

Figure A.7 Quantile scores for current and future precipitation analysis.

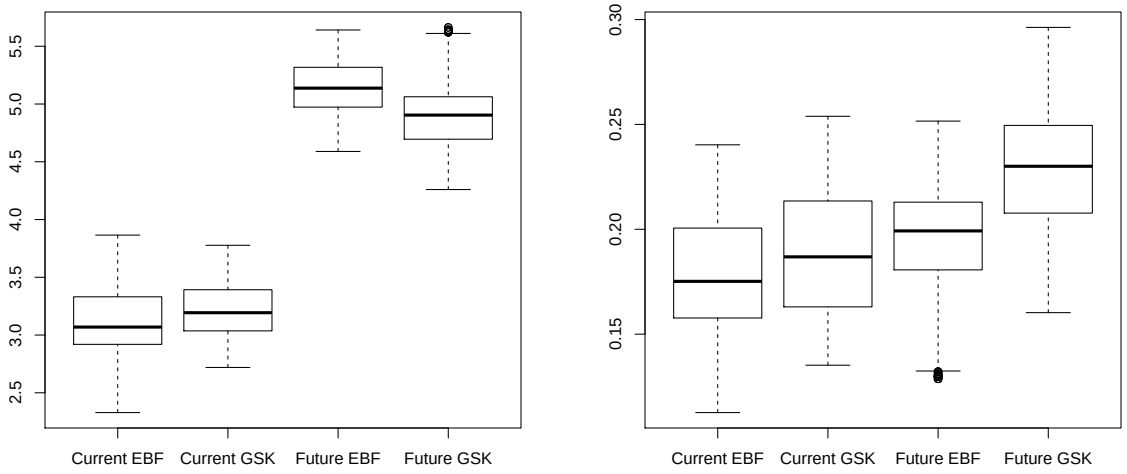

Figure A.8 Posterior distributions for β_{time} for μ (left) and $\log(\sigma)$ (right).

A.5.5 Results for precipitation analysis

A.6 Conclusions

Acknowledgements

A.1 Extreme value distributions

Define (1) GEV density f and CDF F; (2) PS pdf and the grid approximation to the integral.

A.2 Gradient for β

BIBLIOGRAPHY

[AC14] Azzalini, A. & Capitanio, A. *The Skew-Normal and Related Families*. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2014.

[AC03] Azzalini, A. & Capitanio, A. "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**.2 (2003), pp. 367–389.

[Ber16] Beranger, B. et al. "Models for extremal dependence derived from skew-symmetric families". *ArXiv e-prints* (2016). arXiv:1507.00108. arXiv: 1507.00108.

[Coh60] Cohen, J. "A Coefficient of Agreement for Nominal Scales". *Educational and Psychological Measurement* **20**.1 (1960), pp. 37–46.

[Col99] Coles, S. et al. "Dependence Measures for Extreme Value Analyses". *Extremes* **2**.4 (1999), pp. 339–365.

[DS90] Davison, A. C. & Smith, R. L. "Models for exceedances over high thresholds (with Discussion)". *Journal of the Royal Statistical Society. Series B (Methodological)* **52**.3 (1990), pp. 393–442.

[Eng14] Engelke, S. et al. "Estimation of Hüsler-Reiss distributions and Brown-Resnick processes". *Journal of the Royal Statistical Society. Series B: Statistical Methodology* (2014), pp. 239–265. arXiv: arXiv:1207.6886v2.

[GR07] Gneiting, T. & Raftery, A. E. "Strictly Proper Scoring Rules, Prediction, and Estimation". *Journal of the American Statistical Association* **102**.477 (2007), pp. 359–378.

[Hol14] Hollander, M. et al. "Nonparametric Statistical Methods". 3rd. Wiley, 2014, pp. 316–322.

[HD14] Huser, R. & Davison, A. C. "Space-time modelling of extreme events". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**.2 (2014), pp. 439–461. arXiv: 1201.3245.

[Kim05] Kim, H.-M. et al. "Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes". *Journal of the American Statistical Association* **100**.470 (2005), pp. 653–668.

[LT96] Ledford, A. W. & Tawn, J. a. "Statistics for near independence in multivariate extreme values." *Biometrika* **83** (1996), pp. 169–187.

[Pad10] Padoan, S. A. et al. "Likelihood-Based Inference for Max-Stable Processes". *Journal of the American Statistical Association* **105**.489 (2010), pp. 263–277.

[Pad11] Padoan, S. A. "Multivariate extreme models based on underlying skew- and skew-normal distributions". *Journal of Multivariate Analysis* **102**.5 (2011), pp. 977–991.

[RS12] Reich, B. J. & Shaby, B. A. "A hierarchical max-stable spatial model for extreme precipitation". *The Annals of Applied Statistics* **6**.4 (2012), pp. 1430–1451.

[Sam00] Samet, J. M. et al. *The National Morbidity, Mortalityand Air Pollution Study Part I : Methods and Methodologic Issues*. Tech. rep. 94. 2000.

[Ste09] Stephenson, A. G. "High-Dimensional Parametric Modelling of Multivariate Extreme Events". *Australian & New Zealand Journal of Statistics* **51**.1 (2009), pp. 77–88.

[Sul09] Sullivan, B. L. et al. "eBird: A citizen-based bird observation network in the biological sciences". *Biological Conservation* **142**.10 (2009), pp. 2282–2292.

[Taw90] Tawn, J. A. "Modelling multivariate extreme value distributions". *Biometrika* **77**.2 (1990), pp. 245–253.

[Thi13] Thibaud, E. et al. "Threshold modeling of extreme spatial rainfall". *Water Resources Research* **49**.8 (2013), pp. 4633–4644.

[TO15] Thibaud, E. & Opitz, T. "Efficient inference and simulation for elliptical Pareto processes". *Biometrika* **102**.4 (2015), pp. 855–870. arXiv: `arXiv:1401.0168v1`.

[WT14] Wadsworth, J. L. & Tawn, J. a. "Efficient inference for spatial extreme value processes associated to log-Gaussian random functions". *Biometrika* **101**.1 (2014), pp. 1–15.

[WT12] Wadsworth, J. L. & Tawn, J. a. "Dependence modelling for spatial extremes". *Biometrika* **99**.2 (2012), pp. 253–272.

[WD10] Wang, X. & Dey, D. K. "Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption". *The Annals of Applied Statistics* **4**.4 (2010), pp. 2000–2023.

APPENDICES

APPENDIX

A

SPACE-TIME SKEW-T MODEL

A.1 MCMC details

The MCMC sampling for the model in Section 2.4 is done using R (http://www.r-project.org). Whenever possible, we select conjugate priors (see Appendix A.2); however, for some of the parameters, no conjugate prior distributions exist. For these parameters, we use a random walk Metropolis-Hastings update step. In each Metropolis-Hastings update, we tune the algorithm during the burn-in period to give acceptance rates near 0.40.

Spatial knot locations

For each day, we update the spatial knot locations, $\mathbf{w}_1,\ldots,\mathbf{w}_K$, using a Metropolis-Hastings block update. Because the spatial domain is bounded, we generate candidate knots using the transformed knots $\mathbf{w}_1^*,\ldots,\mathbf{w}_K^*$ (see Section 2.3.3) and a random walk bivariate Gaussian candidate distribution

$$\mathbf{w}_k^{*(c)} \sim \mathrm{N}(\mathbf{w}_k^{*(r-1)}, s^2 I_2)$$

where $\mathbf{w}_k^{*(r-1)}$ is the location for the transformed knot at MCMC iteration $r-1$, s is a tuning parameter, and I_2 is an identity matrix. After candidates have been generated for all K knots, the acceptance

ratio is

$$R = \left\{ \frac{l[Y_t(\mathbf{s})|\mathbf{w}_1^{(c)},\ldots,\mathbf{w}_K^{(c)},\ldots)]}{l[Y_t(\mathbf{s})|\mathbf{w}_1^{(r-1)},\ldots,\mathbf{w}_K^{(r-1)},\ldots)]} \right\} \times \left\{ \frac{\prod_{k=1}^K \phi(\mathbf{w}_k^{(c)})}{\prod_{k=1}^K \phi(\mathbf{w}_k^{(r-1)})} \right\} \times \left\{ \frac{\prod_{k=1}^K p(\mathbf{w}_k^{*(c)})}{\prod_{k=1}^K p(\mathbf{w}_k^{*(r-1)})} \right\}$$

where l is the likelihood given in (2.17), and $p(\cdot)$ is the prior either taken from the time series given in (2.3.3) or assumed to be uniform over $\mathscr{D}$. The candidate knots are accepted with probability $\min\{R,1\}$.

Spatial random effects

If there is no temporal dependence amongst the observations, we use a Gibbs update for z_{tk}, and the posterior distribution is given in Appendix A.2. If there is temporal dependence amongst the observations, then we update z_{tk} using a Metropolis-Hastings update. Because this model uses $|z_{tk}|$, we generate candidate random effects using the z_{tk}^* (see Section 2.3.3) and a random walk Gaussian candidate distribution

$$z_{tk}^{*\,(c)} \sim \mathrm{N}(z_{tk}^{*\,(r-1)}, s^2)$$

where $z_{tk}^{*\,(r-1)}$ is the value at MCMC iteration $r-1$, and s is a tuning parameter. The acceptance ratio is

$$R = \left\{ \frac{l[Y_t(\mathbf{s})|z_{tk}^{(c)},\ldots]}{l[Y_t(\mathbf{s})|z_{tk}^{(r-1)}]} \right\} \times \left\{ \frac{p[z_{tk}^{(c)}]}{p[z_{tk}^{(r-1)}]} \right\}$$

where $p[\cdot]$ is the prior taken from the time series given in Section 2.3.3. The candidate is accepted with probability $\min\{R,1\}$.

Variance terms

When there is more than one site in a partition, then we update σ_{tk}^2 using a Metropolis-Hastings update. First, we generate a candidate for σ_{tk}^2 using an $\mathrm{IG}(a^*/s, b^*/s)$ candidate distribution in an independence Metropolis-Hastings update where $a^* = (n_{tk}+1)/2 + a$, $b^* = [Y_{tk}'\Sigma_{tk}^{-1}Y_{tk} + z_{tk}^2]/2 + b$, n_{tk} is the number of sites in partition k on day t, and Y_{tk} and Σ_{tk}^{-1} are the observations and precision matrix for partition k on day t. The acceptance ratio is

$$R = \left\{ \frac{l[Y_t(\mathbf{s})|\sigma_{tk}^{2\,(c)},\ldots]}{l[Y_t(\mathbf{s})|\sigma_{tk}^{2\,(r-1)}]} \right\} \times \left\{ \frac{l[z_{tk}|\sigma_{tk}^{2\,(c)},\ldots]}{l[z_{tk}|\sigma_{tk}^{2\,(r-1)},\ldots]} \right\} \times \left\{ \frac{p[\sigma_{tk}^{2\,(c)}]}{p[\sigma_{tk}^{2\,(r-1)}]} \right\} \times \left\{ \frac{c[\sigma_{tk}^{2\,(r-1)}]}{c[\sigma_{tk}^{2\,(c)}]} \right\}$$

where $p[\cdot]$ is the prior either taken from the time series given in Section 2.3.3 or assumed to be IG(a,b), and $c[\cdot]$ is the candidate distribution. The candidate is accepted with probability $\min\{R,1\}$.

Spatial covariance parameters

We update the three spatial covariance parameters, $\log(\rho)$, $\log(\nu)$, γ, using a Metropolis-Hastings block update step. First, we generate a candidate using a random walk Gaussian candidate distribution

$$\log(\rho)^{(c)} \sim \mathrm{N}(\log(\rho)^{(r-1)}, s^2)$$

where $\log(\rho)^{(r-1)}$ is the value at MCMC iteration $r-1$, and s is a tuning parameter. Candidates are generated for $\log(\nu)$ and γ in a similar fashion. The acceptance ratio is

$$R = \left\{ \frac{\prod_{t=1}^{T} l[Y_t(\mathbf{s})|\rho^{(c)}, \nu^{(c)}, \gamma^{(c)}, \ldots]}{\prod_{t=1}^{T} l[Y_t(\mathbf{s})|\rho^{(r-1)}, \nu^{(r-1)}, \gamma^{(r-1)}, \ldots]} \right\} \times \left\{ \frac{p[\rho^{(c)}]}{p[\rho^{(r-1)}]} \right\} \times \left\{ \frac{p[\nu^{(c)}]}{p[\nu^{(r-1)}]} \right\} \times \left\{ \frac{p[\gamma^{(c)}]}{p[\gamma^{(r-1)}]} \right\}.$$

All three candidates are accepted with probability $\min\{R,1\}$.

A.2 Posterior distributions

Conditional posterior of $z_{tk} \mid \ldots$

If knots are independent over days, then the conditional posterior distribution of $|z_{tk}|$ is conjugate. For simplicity, drop the subscript t, let $\tilde{z}_{tk} = |z_{tk}|$, and define

$$R(\mathbf{s}) = \begin{cases} Y(\mathbf{s}) - X(\mathbf{s})\beta & s \in P_l \\ Y(\mathbf{s}) - X(\mathbf{s})\beta - \lambda \tilde{z}(\mathbf{s}) & s \notin P_l \end{cases}$$

Let

$$\begin{aligned} R_1 &= \text{the vector of } R(\mathbf{s}) \text{ for } s \in P_l \\ R_2 &= \text{the vector of } R(\mathbf{s}) \text{ for } s \notin P_l \\ \Omega &= \Sigma^{-1}. \end{aligned}$$

Then

$$\begin{aligned}\pi(z_l|\ldots) &\propto \exp\left\{-\frac{1}{2}\left[\begin{pmatrix} R_1-\lambda\tilde{z}_l\mathbf{1} \\ R_2 \end{pmatrix}'\begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}\begin{pmatrix} R_1-\lambda\tilde{z}_l\mathbf{1} \\ R_2 \end{pmatrix}+\frac{\tilde{z}_l^2}{\sigma_l^2}\right]\right\} I(z_l>0) \\ &\propto \exp\left\{-\frac{1}{2}\left[\Lambda_l\tilde{z}_l^2-2\mu_l\tilde{z}_l\right]\right\}\end{aligned}$$

where

$$\begin{aligned}\mu_l &= \lambda(R_1'\Omega_{11}+R_2'\Omega_{21})\mathbf{1} \\ \Lambda_l &= \lambda^2\mathbf{1}'\Omega_{11}\mathbf{1}+\frac{1}{\sigma_l^2}.\end{aligned}$$

Then $\tilde{Z}_l|\ldots \sim N_{(0,\infty)}(\Lambda_l^{-1}\mu_l,\Lambda_l^{-1})$

Conditional posterior of $\beta \mid \ldots$

Let $\beta \sim \mathrm{N}_p(0,\Lambda_0)$ where Λ_0 is a precision matrix. Then

$$\begin{aligned}\pi(\beta\mid\ldots) &\propto \exp\left\{-\frac{1}{2}\beta'\Lambda_0\beta-\frac{1}{2}\sum_{t=1}^{m}\left[\mathbf{Y}_t-\mathbf{X}_t\beta-\lambda|z_t|\right]'\Omega\left[\mathbf{Y}_t-\mathbf{X}_t\beta-\lambda|z_t|\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\beta'\Lambda_\beta\beta-2\sum_{t=1}^{'}[\beta' X_t'\Omega(\mathbf{Y}_t-\lambda|z_t|)]\right]\right\} \\ &\propto \mathrm{N}(\Lambda_\beta^{-1}\mu_\beta,\Lambda_\beta^{-1})\end{aligned}$$

where

$$\begin{aligned}\mu_\beta &= \sum_{t=1}^{T}\left[X_t'\Omega(\mathbf{Y}_t-\lambda|z_t|)\right] \\ \Lambda_\beta &= \Lambda_0+\sum_{t=1}^{T}X_t'\Omega X_t.\end{aligned}$$

Conditional posterior of $\sigma^2 \mid \ldots$

In the case where $L = 1$ and temporal dependence is negligible, then σ^2 has a conjugate posterior distribution. Let $\sigma_t^2 \overset{iid}{\sim} \text{IG}(\alpha_0, \beta_0)$. For simplicity, drop the subscript t. Then

$$\begin{aligned}
\pi(\sigma^2 \mid \ldots) &\propto (\sigma^2)^{-\alpha_0 - 1/2 - n/2 - 1} \exp\left\{ -\frac{\beta_0}{\sigma^2} - \frac{|z|^2}{2\sigma^2} - \frac{(\mathbf{Y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\mu})}{2\sigma^2} \right\} \\
&\propto (\sigma^2)^{-\alpha_0 - 1/2 - n/2 - 1} \exp\left\{ -\frac{1}{\sigma^2} \left[\beta_0 + \frac{|z|^2}{2} + \frac{1}{2} (\mathbf{Y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{Y} - \mu) \right] \right\} \\
&\propto \text{IG}(\alpha^*, \beta^*)
\end{aligned}$$

where

$$\begin{aligned}
\alpha^* &= \alpha_0 + \frac{1}{2} + \frac{n}{2} \\
\beta^* &= \beta_0 + \frac{|z|^2}{2} + \frac{1}{2} (\mathbf{Y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\mu}).
\end{aligned}$$

In the case that $L > 1$, a random walk Metropolis Hastings step will be used to update σ_{lt}^2.

Conditional posterior of $\lambda \mid \ldots$

For convergence purposes we model $\lambda = \lambda_1 \lambda_2$ where

$$\lambda_1 = \begin{cases} +1 & \text{w.p.0.5} \\ -1 & \text{w.p.0.5} \end{cases} \tag{A.1}$$

$$\lambda_2^2 \sim IG(\alpha_\lambda, \beta_\lambda). \tag{A.2}$$

$$\tag{A.3}$$

Then

$$\begin{aligned}
\pi(\lambda_2 \mid \ldots) &\propto \lambda_2^{2(-\alpha_\lambda - 1)} \exp\left\{ -\frac{\beta_\lambda}{\lambda_2^2} \right\} \prod_{t=1}^{T} \prod_{k=1}^{K} \frac{1}{\lambda_2} \exp\left\{ -\frac{z_{tk}^2}{2\lambda_2^2 \sigma_{tk})^2} \right\} \\
&\propto \lambda_2^{2(-\alpha_\lambda - kt - 1)} \exp\left\{ -\frac{1}{\lambda_2^2} \left[\beta_\lambda + \frac{z^2}{2\sigma_{tk}^2} \right] \right\}
\end{aligned}$$

Then $\lambda_2 \mid \ldots \sim IG\left(\alpha_\lambda + kt, \beta_\lambda + \frac{z^2}{2\sigma_{tk}^2} \right)$

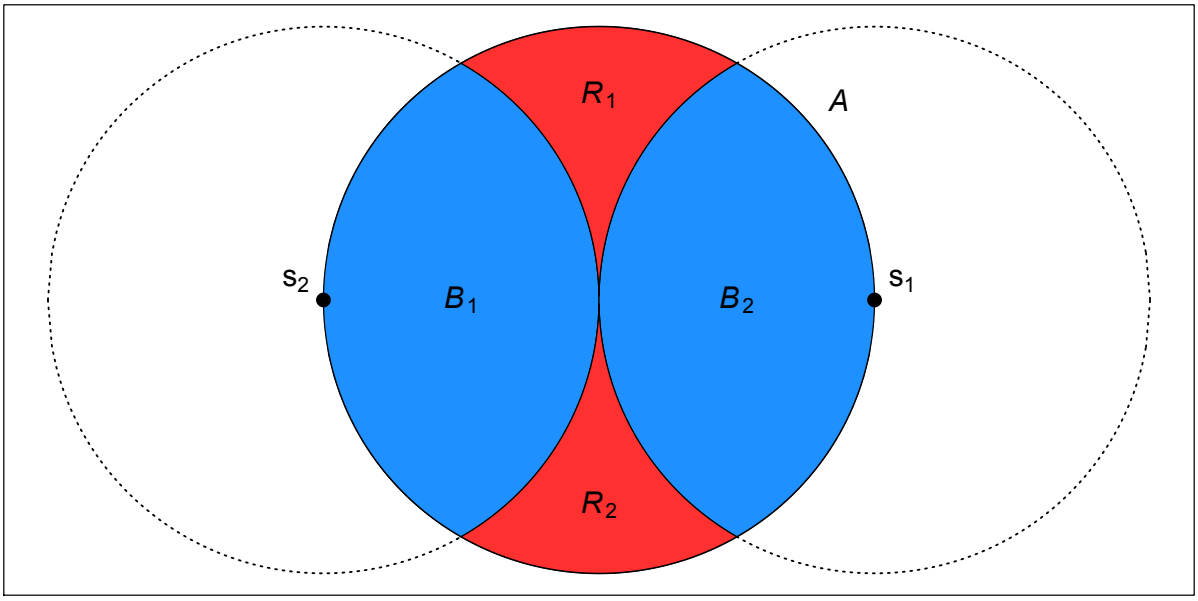

Figure A.1 Illustration of the partition of A.

A.3 Proof that $\lim_{h\to\infty}\pi(h)=0$

Let c be the midpoint of $\mathbf{s}_1$ and $\mathbf{s}_2$. Define A as the circle centered at c with radius $h/2$ where $h=||\mathbf{s}_1-\mathbf{s}_2||$ is the distance between sites $\mathbf{s}_1$ and $\mathbf{s}_2$. Consider a homogeneous spatial Poisson process over A with intensity given by

$$\mu(A)=\lambda_{PP}|A|=\lambda_{PP}\pi\left(\frac{h}{2}\right)^2=\lambda^*_{PPA}h^2.$$

Consider a partition of A into four regions, B_1, B_2, R_1, R_2 as seen in Figure A.1. Let N_j be the number of knots in B_j, $j=1,2$. Then

$$P(\mathbf{s}_1\in P_i,\mathbf{s}_2\in P_{j\neq i})\geq P(N_1>0,N_2>0) \tag{A.4}$$

since knots in both B_1 and B_2 is sufficient, but not necessary, to ensure that $\mathbf{s}_1$ and $\mathbf{s}_2$ are in different partition sets. By definition of a Poission process, N_1 and N_2 are independent and thus $P(N_1>0,N_2>$

$0) = P(N_1 > 0)^2$, and the intensity measure over B_1 is given by

$$\begin{aligned}\mu(B_1) &= \lambda_{PP}|B_1| = \lambda_{PP}\frac{h^2}{4}\left(\frac{2\pi}{3} - \frac{\sqrt{3}}{2}\right) \\ &= \lambda^*_{PPB1}h^2. \end{aligned} \tag{A.5}$$

So,

$$P(\mathbf{s}_1 \in P_i, \mathbf{s}_2 \in P_{j\neq i}) >= P(N_1 > 0)^2 = [1 - P(N_1 = 0)]^2 = [1 - \exp\left(-\lambda^*_{PPB1}h^2\right)]^2 \tag{A.6}$$

which goes to 1 as h goes to infinity.

A.4 Skew-t distribution

Univariate skew-t distribution

We say that Y follows a univariate extended skew-t distribution with location $\xi \in \mathscr{R}$, scale $\omega > 0$, skew parameter $\alpha \in \mathscr{R}$, and degrees of freedom ν if has distribution function

$$f_{\text{EST}}(y) = 2f_T(z;\nu)F_T\left[\alpha z\sqrt{\frac{\nu+1}{\nu+z^2}};\nu+1\right] \tag{A.7}$$

where $f_T(t;\nu)$ is a univariate Student's t with ν degrees of freedom, $F_T(t;\nu) = P(T < t)$, and $z = (y-\xi)/\omega$.

Multivariate skew-t distribution

If $\mathbf{Z} \sim \text{ST}_d(0, \bar{\mathbf{\Omega}}, \boldsymbol{\alpha}, \eta)$ is a d-dimensional skew-t distribution, and $\mathbf{Y} = \xi + \boldsymbol{\omega}\mathbf{Z}$, where $\boldsymbol{\omega} = \text{diag}(\omega_1, \ldots, \omega_d)$, then the density of Y at y is

$$f_y(\mathbf{y}) = d\,e\,t(\boldsymbol{\omega})^{-1}f_z(\mathbf{z}) \tag{A.8}$$

where

$$f_z(\mathbf{z}) = 2\,t_d(\mathbf{z};\bar{\mathbf{\Omega}},\eta)T\left[\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\eta+d}{\nu+Q(\mathbf{z})}};\eta+d\right] \tag{A.9}$$

$$\mathbf{z} = \boldsymbol{\omega}^{-1}(\mathbf{y}-\xi) \tag{A.10}$$

where $t_d(\mathbf{z};\bar{\mathbf{\Omega}},\eta)$ is a d-dimensional Student's t-distribution with scale matrix $\bar{\mathbf{\Omega}}$ and degrees of freedom η, $Q(z)=\mathbf{z}'\bar{\Omega}^{-1}\mathbf{z}$ and $T(\cdot;\eta)$ denotes the univariate Student's t distribution function with η degrees of freedom [AC14].

Extremal dependence

For a bivariate skew-t random variable $\mathbf{Y}=[Y(\mathbf{s}),Y(\mathbf{t})]'$, the $\chi(h)$ statistic [Pad11] is given by

$$\chi(h)=\bar{F}_{\text{EST}}\left\{\frac{[x_1^{1/\eta}-\varrho(h)]\sqrt{\eta+1}}{\sqrt{1-\varrho(h)^2}};0,1,\alpha_1,\tau_1,\eta+1\right\}+\bar{F}_{\text{EST}}\left\{\frac{[x_2^{1/\eta}-\varrho(h)]\sqrt{\eta+1}}{\sqrt{1-\varrho(h)^2}};0,1,\alpha_2,\tau_2,\eta+1\right\}, \tag{A.11}$$

where $\bar{F}_{\text{EST}}$ is the univariate survival extended skew-t function with zero location and unit scale, $\varrho(h)=\text{cor}[y(\mathbf{s}),y(\mathbf{t})]$, $\alpha_j=\alpha_i\sqrt{1-\varrho^2}$, $\tau_j=\sqrt{\eta+1}(\alpha_j+\alpha_i\varrho)$, and $x_j=F_T(\bar{\alpha}_i\sqrt{\eta+1};0,1,\eta)/F_T(\bar{\alpha}_j\sqrt{\eta+1};0,1,\eta)$ with $j=1,2$ and $i=2,1$ and where $\bar{\alpha}_j=(\alpha_j+\alpha_i\varrho)/\sqrt{1+\alpha_i^2[1-\varrho(h)^2]}$.

Proof that $\lim_{h\to\infty}\chi(h)>0$

Consider the bivariate distribution of $\mathbf{Y}=[Y(\mathbf{s}),Y(\mathbf{t})]'$, with $\varrho(h)$ given by (2.2). So, $\lim_{h\to\infty}\varrho(h)=0$. Then

$$\lim_{h\to\infty}\chi(h)=\bar{F}_{\text{EST}}\left\{\sqrt{\eta+1};0,1,\alpha_1,\tau_1,\eta+1\right\}+\bar{F}_{\text{EST}}\left\{\sqrt{\eta+1};0,1,\alpha_2,\tau_2,\eta+1\right\}. \tag{A.12}$$

Because the extended skew-t distribution is not bounded above, for all $\bar{F}_{\text{EST}}(x)=1-F_{\text{EST}(x)}>0$ for all $x<\infty$. Therefore, for a skew-t distribution, $\lim_{h\to\infty}\chi(h)>0$.

A.5 Other parameterizations

Various forms of multivariate skew-normal and skew-t distributions have been proposed in the literature. In this section, we make a connection between our parameterization in (2.1) of the main text and another popular version. Azzalini & Capitanio [AC14] and Beranger et al. [Ber16] define a skew-normal process as

$$\tilde{X}(\mathbf{s})=\tilde{\lambda}|z|+(1-\tilde{\lambda}^2)^{1/2}v(\mathbf{s}) \tag{A.13}$$

where $\tilde{\lambda}\in(-1,1)$, z $N(0,1)$, $v(\mathbf{s})$ is a Gaussian process with mean zero, variance one, and spatial correlation function ρ. To extend this to the skew-t distribution, Azzalini & Capitanio [AC03] take

$\tilde{Y}(\mathbf{s}) = W\tilde{X}(\mathbf{s})$ where $W^{-2} \sim \text{Gamma}(a/2, a/2)$. Returning to the proposed parameterization, let $W^{-2} = \frac{2b}{a}\sigma^{-2} \sim \text{Gamma}(a/2, a/2)$ so that (2.1) becomes

$$Y(\mathbf{s}) = W\left[\lambda\left(\frac{a}{2b}\right)^{1/2}|z| + \left(\frac{a}{2b}\right)^{1/2} v(\mathbf{s})\right]. \tag{A.14}$$

Clearly setting $a = \nu/2 > 0$, $b = \frac{\nu}{(1-\tilde{\lambda}^2)} > 0$, and $\lambda = \tilde{\lambda}/(1-\tilde{\lambda}^2)^{1/2} \in (-\infty, \infty)$ resolves the difference in parameterizations. We note that our parameterization has three parameters (a, b, λ) compared to the to parameters of the alternative parameterization $(a, \tilde{\lambda})$. Since we have assumed that both $v(\mathbf{s})$ and z have unit scale, the additional b parameter in our parameterization controls the scale.

A.6 Simulation study results

The following tables show the methods that have significantly different Brier scores when using a Wilcoxon-Nemenyi-McDonald-Thompson test. In each column, different letters signify that the methods have significantly different Brier scores. For example, there is significant evidence to suggest that method 1 and method 4 have different Brier scores at $q(0.90)$, whereas there is not significant evidence to suggest that method 1 and method 2 have different Brier scores at $q(0.90)$. In each table group A represents the group with the lowest Brier scores. Groups are significant with a familywise error rate of $\alpha = 0.05$.

Table A.1 Setting 1 – Gaussian marginal, $K = 1$ knot

	$q(0.90)$	$q(0.95)$	$q(0.98)$	$q(0.99)$
Method 1	A	A	A	A B
Method 2	A	A	A	A
Method 3	B	B	C	B
Method 4	A	A	A B	A B
Method 5	B	B	B C	A B
Method 6	C	C	D	C

Table A.2 Setting 2 – Skew-t marginal, $K = 1$ knot

	$q(0.90)$	$q(0.95)$	$q(0.98)$	$q(0.99)$
Method 1	C	B	B C	B
Method 2	A	A	A	A
Method 3	B C	A B	A B	A B
Method 4	A B	B	B	A
Method 5	D	C	C	B
Method 6	E	D	D	C

Table A.3 Setting 3 – Skew-t marginal, $K = 5$ knots

	$q(0.90)$	$q(0.95)$	$q(0.98)$	$q(0.99)$
Method 1	B	C	B	B
Method 2	B	C	B	B
Method 3	A	B	B	B
Method 4	A	A	A	A
Method 5	A	A	A	A
Method 6	C	D	C	C

Table A.4 Setting 4 – Max-stable

	$q(0.90)$	$q(0.95)$	$q(0.98)$	$q(0.99)$
Method 1	A B	B	B	C
Method 2	B	B C	B	B C
Method 3	C D	C	B	B
Method 4	D	D	C	C
Method 5	C	C	B	B C
Method 6	A	A	A	A

Table A.5 Setting 5 – Transformation below $T = q(0.80)$

	$q(0.90)$	$q(0.95)$	$q(0.98)$	$q(0.99)$
Method 1	C	B	C	C
Method 2	B	B	B	A B
Method 3	A	A	A	A
Method 4	B C	B	B	B C
Method 5	B	B	B C	C
Method 6	D	C	D	D

APPENDIX

B

RARE-BINARY

B.1 Simulation study results

The following tables show the methods that have significantly different Brier scores when using a Wilcoxon-Nemenyi-McDonald-Thompson test. In each column, different letters signify that the methods have significantly different Brier scores.

Table B.1 Pairwise BS comparisons

	Setting 1	Setting 2	Setting 3	Setting 4	Setting 5	Setting 6
Method 1	A	A	A	C	B	B
Method 2	A B	B	A	B	A	A
Method 3	B	B	A	A	A B	A

APPENDIX

C

SPACE-TIME SKEW-T MODEL

C.1 Extreme value distributions

Define (1) GEV density f and CDF F; (2) PS pdf and the grid approximation to the integral.

C.2 Gradient for β