A spatial-skew model for threshold exceedances

May 24, 2015

1 Introduction

In many climatological applications, researchers are interested in learning about the average behavior of different climate variables (e.g. ozone, temperature, rainfall). 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 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).

Traditional geostatistical modeling is based upon the assumption that observations come from a Gaussian process, a process that is fully defined by its mean and covariance functions. In the limit of the Gaussian distribution, all observations are independent regardless of the strength of the correlation in the bulk of the data. Furthermore, the Gaussian distribution is light-tailed and symmetric. Therefore, it is inappropriate to use standard geostatistical methods when trying to describe dependence in the tails of the distribution.

Threshold modeling is popular in the field of extreme value statistics where extreme events are naturally defined in terms of exceedances over a high threshold. Davison and Smith (1990) considered modeling threshold exceedances of univariate time series by the generalized Pareto distribution. Bivariate threshold models for extreme value distributions were considered by Ledford and Tawn (1996) 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 and Tawn (2012) and Thibaud et al. (2013) who fit various models to spatial extremes using a censored pairwise likelihood (Padoan et al., 2010) based on the approach of Ledford and Tawn (1996). Huser

1

and Davison (2014) further extended this to spate-time modeling. Engelke et al. (2014), Wadsworth and Tawn (2014), and Thibaud and Opitz (2013) 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 to assume 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. 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.

Instead, we propose a new spatiotemporal threshold exceedance model based on the skew-t process (Padoan, 2011). Our model is a threshold exceedance model for the multivariate skew-t distribution that uses imputation for values below a fixed threshold. 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.

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. (2005) for non-stationary Gaussian data. This partition alleviates the asymptotic spatial dependence present in the skew-t distribution for sites that are far apart. Finally, our model allows for inference and predictions using the full likelihood with computing on the order of Gaussian models for large space-time datasets.

The paper is organized as follows. Section 2 is a brief review of the spatial skew-t process. In Section 3.3, we build upon the traditional skew-t 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 4. In Section 5, we present a simulation study that examines the predictive capabilities of this

2

model compared with a naïve Gaussian method. We then compare our method to Gaussian and max-stable

methods with a data analysis of ozone measurements from throughout the US in section 6. The final section

provides brief discussion and direction for future research.

2 Spatial skew processes

Many types of data demonstrate some level of skewness and therefore should be modeled with distributions

that allow for asymmetry. The skew-elliptical family of distributions provides models that are mathemati-

cally tractable while introducing a slant parameter to account for asymmetric data (Genton, 2004). A brief

review of the additive process by which a skew-t process is created is given here.

2.1 Skew-t process

Let $Y(\mathbf{s})$ be the observation at spatial location $\mathbf{s} = (s_1, s_2)$ in a spatial domain of interest $\mathcal{D} \in \mathcal{R}^2$. The

spatial skew-t process can be written

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T \boldsymbol{\beta} + \lambda \sigma |z| + \sigma v(\mathbf{s}) \tag{1}$$

where $\mathbf{X}(\mathbf{s})$ is a set of spatial covariates at site $\mathbf{s}$, $\boldsymbol{\beta}$ is the vector of regression parameters, $\lambda \in \mathcal{R}$ is a

parameter controlling skew, $z \sim N(0, 1)$, $\sigma^2 \sim \mathrm{IG}(a, b)$ is an inverse gamma random variable, and $v(\mathbf{s})$ is

a spatial Gaussian process with mean zero, variance one, and a positive definite correlation function.

For a finite collection of locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, consider observations $\mathbf{Y} = [Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)]^T$. After

marginalizing over both z and σ,

$$\mathbf{Y} \sim \mathrm{ST}_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, 2a), \tag{2}$$

3

that is, $\mathbf{Y}$ follows an n-dimensional skew-t distribution with location $\mathbf{X}\boldsymbol{\beta}$, correlation matrix $\boldsymbol{\Omega}$, slant param-

eters $\boldsymbol{\alpha}$ and degrees of freedom $2a$, where $\mathbf{X} = [\mathbf{X}(\mathbf{s}_1)^T, \ldots, \mathbf{X}(\mathbf{s}_n)^T]$, $\boldsymbol{\Omega} = \boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega}$, $\boldsymbol{\omega} = \mathrm{diag}\left(\frac{1}{\sqrt{ab}}, \ldots, \frac{1}{\sqrt{ab}}\right)$,

$\bar{\boldsymbol{\Omega}} = (\boldsymbol{\Sigma} + \lambda^2 \mathbf{1}\mathbf{1}^T)$, $\boldsymbol{\Sigma}$ is the positive definite correlation matrix of $[v(\mathbf{s}_1), \ldots, v(\mathbf{s}_n)]$, $\boldsymbol{\alpha} = \lambda(1 + \lambda^2 \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1})^{-1/2} \mathbf{1}^T \boldsymbol{\Sigma}^{-1}$

is a vector of slant parameters. Although $\boldsymbol{\Sigma}$ can be any positive definite correlation function, we choose to

use the stationary isotropic Matérn correlation with

$$\mathrm{cor}[v(\mathbf{s}), v(\mathbf{t})] = \gamma I(\mathbf{s} = \mathbf{t}) + (1 - \gamma)\frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\sqrt{2\nu}\frac{h}{\rho}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}\frac{h}{\rho}\right) \tag{3}$$

where ρ 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 $h = ||\mathbf{s} - \mathbf{t}||$. This process is

desirable because of its flexible tail that is controlled by the skewness parameter λ and degrees of freedom

$2a$. Furthermore, the marginal distributions at each location also follow a univariate skew-t distribution

(Azzalini and Capitanio, 2014).

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 (Coles et al., 1999). For a stationary and isotropic spatial process, the χ statistic for two

location $\mathbf{s}$ and $\mathbf{t}$ separated by distance $h = ||\mathbf{s} - \mathbf{t}||$ with identical marginal distributions is

$$\chi(h) = \lim_{c \to c^*} \Pr[Y(\mathbf{s}) > c | Y(\mathbf{t}) > c] \tag{4}$$

where c^* is the upper limit of the support of Y. If $\chi(h) = 0$, then observations are asymptotically in-

dependent at distance h. For Gaussian processes, $\chi(h) = 0$ regardless of the distance, so they are not

suitable for modeling asymptotically dependent extremes. Unlike the Gaussian process, the skew-t process

4

is asymptotically dependent (see 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. The explicit expression for $\chi(h)$ is given in Appendix A.4.

3 Spatiotemporal skew-t model for extremes

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 convenice, we introduce the model for a single replication, and then extend this model to the spatiotemporal setting in Section 3.3.

3.1 Censoring to focus on the tails

Because one of our goals is to model the dependence of the distribution in the tails of the data, we choose to censor values below threshold. Let

$$
\widetilde{Y}(\mathbf{s}) = \begin{cases} Y(\mathbf{s}) & \delta(\mathbf{s}) = 1 \\ T & \delta(\mathbf{s}) = 0 \end{cases} \tag{5}
$$

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. Then, assuming the uncensored data $Y(\mathbf{s})$ are observations from a skew-t process, we update values censored below the threshold using standard Bayesian missing data methods as described in Section 4.

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 and Tawn (2012) with a hybrid spatial dependence model. Huser and Davison (2014) also allow for 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, 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

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T\boldsymbol{\beta} + \lambda\sigma(\mathbf{s})|z(\mathbf{s})| + \sigma(\mathbf{s})v(\mathbf{s}). \tag{6}$$

Let $\mathbf{w} = (w_1, w_2)$ be the location of a spatial knot. 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 $\mathcal{D} \in \mathcal{R}^2$. The knots define a random partition of $\mathcal{D}$ by subregions $P_1, \ldots, P_K$ defined as

$$P_k = \{\mathbf{s} : k = \arg\min_\ell ||\mathbf{s} - \mathbf{w}_\ell||\}. \tag{7}$$

All $z(\mathbf{s})$ and $\sigma(\mathbf{s})$ for sites in subregion k are assigned common values

$$z(\mathbf{s}) = z_k \quad \text{and} \quad \sigma(\mathbf{s}) = \sigma_k \tag{8}$$

and the z_k and σ_k^2 are distributed as $z_k \overset{iid}{\sim} N(0,1)$ and $\sigma^2 \overset{iid}{\sim} \text{IG}(a,b)$ where IG is the distribution function of an inverse gamma random variable. So, within each partition, $Y(\mathbf{s})$ follows the spatial skew-t process

6

defined in Section 2. Across partitions, the $Y(\mathbf{s})$ remain correlated via the correlation function for $v(\mathbf{s})$ because $v(\mathbf{s})$ spans all partitions.

When incorporating the random partition, conditional on knots $\mathbf{w}_1, \ldots, \mathbf{w}_K$, the χ statistic for two sites $\mathbf{s}$ and $\mathbf{t}$ in partitions k_s and k_t respectively is

$$\chi(h) = I(k_s = k_t)\chi_{\text{skew-}t}(h) + I(k_s \neq k_t)\chi_{\text{Gaus}}(h)$$

$$= I(k_s = k_t)\chi_{\text{skew-}t}(h) \tag{9}$$

where $I(\cdot)$ is an indicator function, $\chi_{\text{skew-}t}(h)$ is the χ statistic for a skew-t process, $\chi_{\text{Gaus}}(h)$ 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 $\mathbf{w}_1, \ldots, \mathbf{w}_K$, $\chi(h) = \pi(h)\chi_{\text{skew}-t}(h)$, where $\pi(h) = \Pr(k_s = k_t)$ is the probability that two sites separated by distance h are in the same partition. So, to show that $\lim_{h \to \infty} \chi(h) = 0$, we need only know that $\lim_{h \to \infty} \pi(h) = 0$. A proof of this is given in Appendix A.3.

In Figure 1, we give $\chi(h)$ for $K = 1, 3, 5, 10$ partitions for a skew-t distribution with $\alpha = 10$, and 3 degrees of freedom. To estimate $\pi(h)$, we generate 500 sites uniformly over the unit-square. We then randomly generate 400 different sets of partitions using $K = 3$, 5, and 10. For each set of knots, we take $\pi(h)$ to be the proportion of sites in the same partition that are separated by distance h. This plot demonstrates how partitioning helps to reduce extremal dependence as h increases.

3.3 Extension to space-time data

When using daily measurements, the assumption of temporal independence is often inappropriate. There are several places where temporal dependence could be incorporated in the model, including the residual $v_t(\mathbf{s})$. However, we choose to allow for temporal dependence in the $\mathbf{w}$, z, and σ terms because these terms dictate
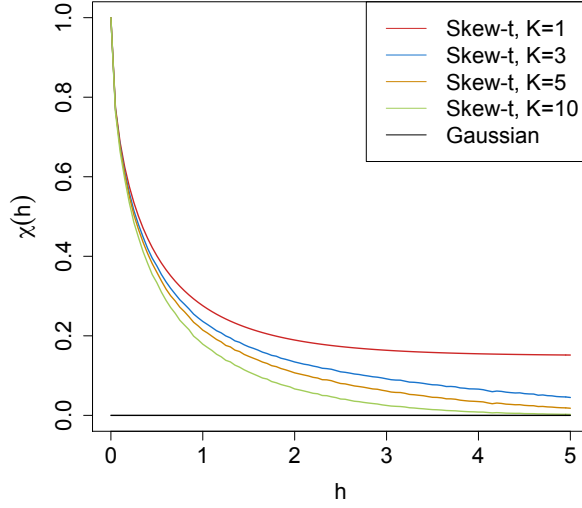
7

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

the tail behavior which is our primary focus. In this section, we extend (6) to the spatiotemporal setting. Let

$$Y_t(\mathbf{s}) = \mathbf{X}_t(\mathbf{s})^T \boldsymbol{\beta} + \lambda \sigma_t(\mathbf{s})|z_t(\mathbf{s})| + \sigma_t(\mathbf{s})v_t(\mathbf{s}), \tag{10}$$

where $t \in \{1, \dots, T\}$ denotes the day of each observation. Let $\mathbf{w}_{tk} = (w_{tk1}, w_{tk2})$ be a spatial knot on day t, and let $w_{t1}, \dots, w_{tK}$ be a collection of spatial knots on day t. As in section 3.2, these knots define a daily partition $P_{t1}, \dots, P_{tK}$, and for $\mathbf{s} \in P_{tk}$,

$$z_t(\mathbf{s}) = z_{tk} \quad \text{and} \quad \sigma_t(\mathbf{s}) = \sigma_{tk}. \tag{11}$$

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

We use an AR(1) time series model for 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

8

transform to obtain a standard normal marginal distribution, place a Gaussian prior with autocorrelation on the transformed parameter, and then transform back to obtain the marginal distribution required to preserve the skew-t process. We first transform the spatial knots from $\mathcal{D}$ to $\mathcal{R}^2$ as follows. Let

$$w_{tki}^* = \Phi^{-1}\left[\frac{w_{tki} - \min(\mathbf{s}_i)}{\max(\mathbf{s}_i) - \min(\mathbf{s}_i)}\right], \quad i = 1, 2 \tag{12}$$

where Φ is a univariate standard normal density function, and $\mathbf{s}_i = [s_{1i}, \ldots, s_{ni}]$. Then the transformed knots $\mathbf{w}_{tk}^* \in \mathcal{R}^2$. We use a copula on $\sigma_t^2(\mathbf{s})$ to ensure that the marginal distributions of $\sigma_t^2(\mathbf{s})$ are inverse gamma. Let

$$\sigma_t^{2*}(\mathbf{s}) = \Phi^{-1}\left\{\text{IG}[\sigma_t^2(\mathbf{s})]\right\} \tag{13}$$

where IG is defined as before. We also use a copula on $z_t(\mathbf{s})$ to ensure that the marginal distributions of $z_t(\mathbf{s})$ are half-normal. Let

$$z_t^*(\mathbf{s}) = \Phi^{-1}\left\{\text{HN}[z_t(\mathbf{s})]\right\} \tag{14}$$

where HN is the distribution function of a half-normal random variable. The AR(1) process for each tail parameter is $\mathbf{w}_{1k}^* \sim N_w(0, 1)$, $z_{1k}^* \sim N(0, \sigma_{1k}^2)$, $\sigma_{1k}^{2*} \sim N(0, 1)$, and for $t > 1$ the time series is modeled as

$$\mathbf{w}_{tk}^*|\mathbf{w}_{t-1,k}^* \sim N_2\left[\phi_w \mathbf{w}_{t-1,k}^*, (1 - \phi_w^2)\right] \tag{15}$$

$$z_{tk}^*|z_{t-1,k}^* \sim N\left[\phi_z z_{t-1,k}^*, \sigma_{tk}^2(1 - \phi_z^2)\right] \tag{16}$$

$$\sigma_{tk}^{2*}|\sigma_{t-1,k}^{2*} \sim N\left[\phi_\sigma \sigma_{t-1,k}^{2*}, (1 - \phi_\sigma^2)\right] \tag{17}$$

9

148 where $|\phi_w|, |\phi_z|, |\phi_\sigma| < 1$. These are stationary time series models with marginal distributions $\mathbf{w}_k^* \sim N_2(0, 1)$,

149 $z_k^* \sim N(0, \sigma_k^2)$, and $\sigma_k^{2*} \sim N(0, 1)$. After transformation back to the original space, $\mathbf{w}_{tk} \sim \text{Unif}(\mathcal{D})$,

150 $z_{tk} \sim HN(0, \sigma_{tk}^2)$ $\sigma_{tk}^2 \sim \text{IG}(a, b)$. For each day, the model is identical to the spatial-only model in (6)

151 by construction.

4 Computation

153 First, we impute values below the threshold conditional on observations above the threshold. This is feasible

154 for large datasets with our model because for a single day, conditional on the model parameters, we only

155 need to draw from a truncated multivariate normal distribution. Then, we update model parameters, Θ,

156 using Metropolis Hastings or Gibbs sampling when appropriate. Finally, we make spatial predictions using

157 conditional multivariate normal results and the fact that the distribution of $Y_t(\mathbf{s}) \mid \Theta, z(\mathbf{s})$ is the usual

158 multivariate normal distribution with a Matérn spatial covariance structure.

159 We can use Gibbs sampling to update $Y_t(\mathbf{s})$ for censored observations that are below the threshold T.

160 After conditioning on λ, $z_t(\mathbf{s})$ and non-censored observations, $Y_t(\mathbf{s})$ has truncated normal full conditionals.

161 So we sample $Y_t(\mathbf{s}) \sim N_{(-\infty, T)}(\mathbf{X}_t^T(\mathbf{s})\beta + \lambda|z_t(\mathbf{s})|, \mathbf{\Sigma})$. After imputing the censored observations, we

162 update the model parameters. To update the model parameters, we use standard Gibbs updates for parame-

163 ters when possible. In the case Gibbs sampling is not possible, parameters are updated using a random-walk

164 Metropolis Hastings algorithm. See Appendices A.1 and A.2 for details regarding the MCMC. The final step

165 of the computation is to use Bayesian Kriging to generate a predictive distribution for $Y_t(\mathbf{s}^*)$ at prediction

166 location $\mathbf{s}^*$. This step is similar to the imputation for censored observations except that the full conditionals

167 are no longer truncated at T.

10

4.1 Hierarchical model

169 Conditioned on $z_{tk}(\mathbf{s})$, $\sigma_{tk}^2(\mathbf{s})$, and P_{tk}, the marginal distributions are Gaussian and the joint distribution

170 multivariate Gaussian. However, we do not fix the partitions, they are treated as unknown and updated in the

171 MCMC. We model this with a Bayesian hierarchical model as follows. Let $\mathbf{w}_{t1}, \ldots, \mathbf{w}_{tK}$ be a set of daily

172 spatial knots in a spatial domain of interest, $\mathcal{D}$, and P_{tk} as defined in (7). In practice, we fix K at many

173 different levels, and assess the impact of fit as described in 5.2. Then

$$Y_t(\mathbf{s}) \mid z_t(\mathbf{s}), \sigma_t^2(\mathbf{s}), P_{tk}, \Theta = \mathbf{X}_t(\mathbf{s})^T \beta + \lambda |z_t(\mathbf{s})| + \sigma_t(\mathbf{s}) v_t(\mathbf{s}) \qquad (18)$$

$$z_t(\mathbf{s}) = z_{tk} \text{ if } \mathbf{s} \in P_{tk}$$

$$\sigma_t^2(\mathbf{s}) = \sigma_{tk}^2 \text{ if } \mathbf{s} \in P_{tk}$$

$$\lambda = \lambda_1 \lambda_2$$

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

$$\lambda_2^2 \sim IG(a, b)$$

$$v_t(\mathbf{s}) \mid \Theta \sim \text{Matérn}\,(0, \Sigma)$$

$$z_{tk}^* \mid z_{t-1,k}^*, \sigma_{tk}^2 \sim N(\phi_z z_{t-1,k}^*, \sigma_{tk}^2(1 - \phi_z^2))$$

$$\sigma_{tk}^{2*} \mid \sigma_{t-1,k}^{2*} \sim N(\phi_\sigma \sigma_{t-1,k}^{2*}, (1 - \phi_\sigma^2))$$

$$\mathbf{w}_{tk}^* \mid \mathbf{w}_{t-1,k}^* \sim N_2(\phi_w \mathbf{w}_{t-1,k}^*, (1 - \phi_w^2))$$

174 where $\Theta = \{\rho, \nu, \gamma\,\lambda, \beta\}$, and Σ is a Matérn covariance matrix as described in Section 2.1. We parameterize

175 $\lambda = \lambda_1 \lambda_2$ to help with convergence in the MCMC.

5 Simulation study

In this section, we conduct a simulation study to investigate how the number of partitions and the level of thresholding impact the accuracy of predictions made by the model.

5.1 Design

For all simulation designs, we generate data from the model in Section 3.2 using $n_s = 144$ sites and $n_t = 50$ independent days. The sites are generated Uniform$([0, 10] \times [0, 10])$. We generate 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. Transformation below $T = q(0.80)$

In the first three designs, the $v_t(\mathbf{s})$ terms are generated using a Matérn covariance with smoothness parameter $\nu = 0.5$ and spatial range $\rho = 1$. For the covariance matrices in designs $1 - 3$, the proportion of the variance accounted for by the spatial variation is $\gamma = 0.9$ while the proportion of the variance accounted for by the nugget effect is 0.1. In the first design, $\sigma^2 = 2$ is used for all days which results in a Gaussian distribution. For designs 2 and 3, $\sigma_{tk}^2 \overset{iid}{\sim} \text{IG}(3, 8)$ to give a t distribution with 6 degrees of freedom. For designs 1, we set $\lambda = 0$. For designs 2 and 3, $\lambda = 3$ was used as to simulate moderate skewness, and the z_t are generated as described in (8). In the fourth design, we generate from a spatial max-stable distribution (Reich and Shaby, 2012). In this design, data have marginal distributions that follow a generalized extreme value distribution with parameters $\mu = 1, \sigma = 1, \xi = 0.2$. In this model, a random effect is used to induce spatial dependence using 144 spatial knots on a regular lattice in the square $[1, 9] \times [1, 9]$. For this setting, we set $\gamma = 0.5$. In

12

the final design, we generate $\tilde{y}$ using the setting from design two, and then consider the data

$$
y = \begin{cases} \tilde{y}, & \tilde{y} > T \\[2em] T \exp\{\tilde{y} - T\}, & \tilde{y} \leq T \end{cases} \tag{19}
$$

where $T = q(0.80)$ is the 80th sample quantile of the data. In all five designs, the mean $\mathbf{X}^T \boldsymbol{\beta} = 10$ is assumed to be constant across space.

$M = 50$ data sets are generated for each design. For each data set we fit the data using five 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 and Shaby (2012) 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 $\log(\nu) \sim \text{N}(-1.2, 1)$, $\gamma \sim \text{Unif}(0, 1)$, $\rho \sim \text{Unif}(15)$. The skewness parameter has prior $\lambda_2 \sim \text{IG}(0.1, 0.1)$. The residual variance terms have priors $\sigma_t^2(\mathbf{s}) \sim \text{IG}(0.1, 0.1)$. The knots have priors $\mathbf{w} \sim \text{Unif}(\mathcal{D})$. 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 only one tail worth of data. Each chain of the MCMC ran for 20000 iterations with a burn-in period of 10000 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.

13

5.2 Cross validation

Models were compared using cross validation with 100 sites used as training sites and 44 sites witheld for testing. The model was fit using the training set, and predictions were generated at the testing site locations. Because one of the primary goals of this model is to predict exceedances over a fixed threshold, we use Brier scores to select the model that best fits the data (Gneiting and Raftery, 2007). 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.

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 gives the Brier score relative to the Brier score for the Gaussian method calculated as

$$\text{BS}_{\text{rel}} = \frac{\text{BS}_{\text{method}}}{\text{BS}_{\text{Gaussian}}}. \tag{20}$$

We analyzed the results for the simulation study using a Friedman test at $\alpha = 0.05$. If the Friedman test came back with a significant results, we conducted a Wilcoxon-Nemenyi-McDonald-Thompson test to see which methods had different results. The full results for the Wilcoxon-Nemenyi-McDonald-Thompson tests are given in Appendix A.5.

Figure 2 shows that when the data come from a Gaussian process, our methods perform comparably to the Gaussian method. For data settings with skew-t marginals (settings $2 - 3$), we find significant improve-

14

ment over the Gaussian method. Furthermore in these data settings, we find the best performance occurs when the number of knots used in the method matches the number of knots used for data generation. The non-thresholded methods tend to outperform the thresholded methods, but this is not surprising given that the data are generated directly from the model used in the method. For the max-stable data, we see that for low-extreme quantiles, the Gaussian method performs better, for more extreme quantiles, the single-partition method, both thresholded and non-thresholded, perform significantly better than the Gaussian. Finally, for setting 5, although the thresholded version of the single-partition model tends to perform the best across all of the extreme quantiles, the difference between the thresholded and non-thresholded methods is no longer significant in the more extreme quantiles.

6 Data analysis

To illustrate this method, we consider the daily maximum 8-hour ozone measurements for July 1 - 31, 2005 at 1089 Air Quality System (AQS) monitoring sites in the United States as the response (see Figure 3). 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 assuming a mean function of

$$\mathbf{X}_t^T(\mathbf{s})\boldsymbol{\beta} = \beta_0 + \beta_1 \cdot \mathrm{CMAQ}_t(\mathbf{s}). \tag{21}$$

The data from July 10 are shown in Figure 3 along with a Q-Q plot of the residuals compared to a skew-t distribution with 10 d.f. and $\alpha = 1$. Exploratory data analysis indicates that there is dependence in the high quantile levels of the residuals beyond what we expect in the case of independence.
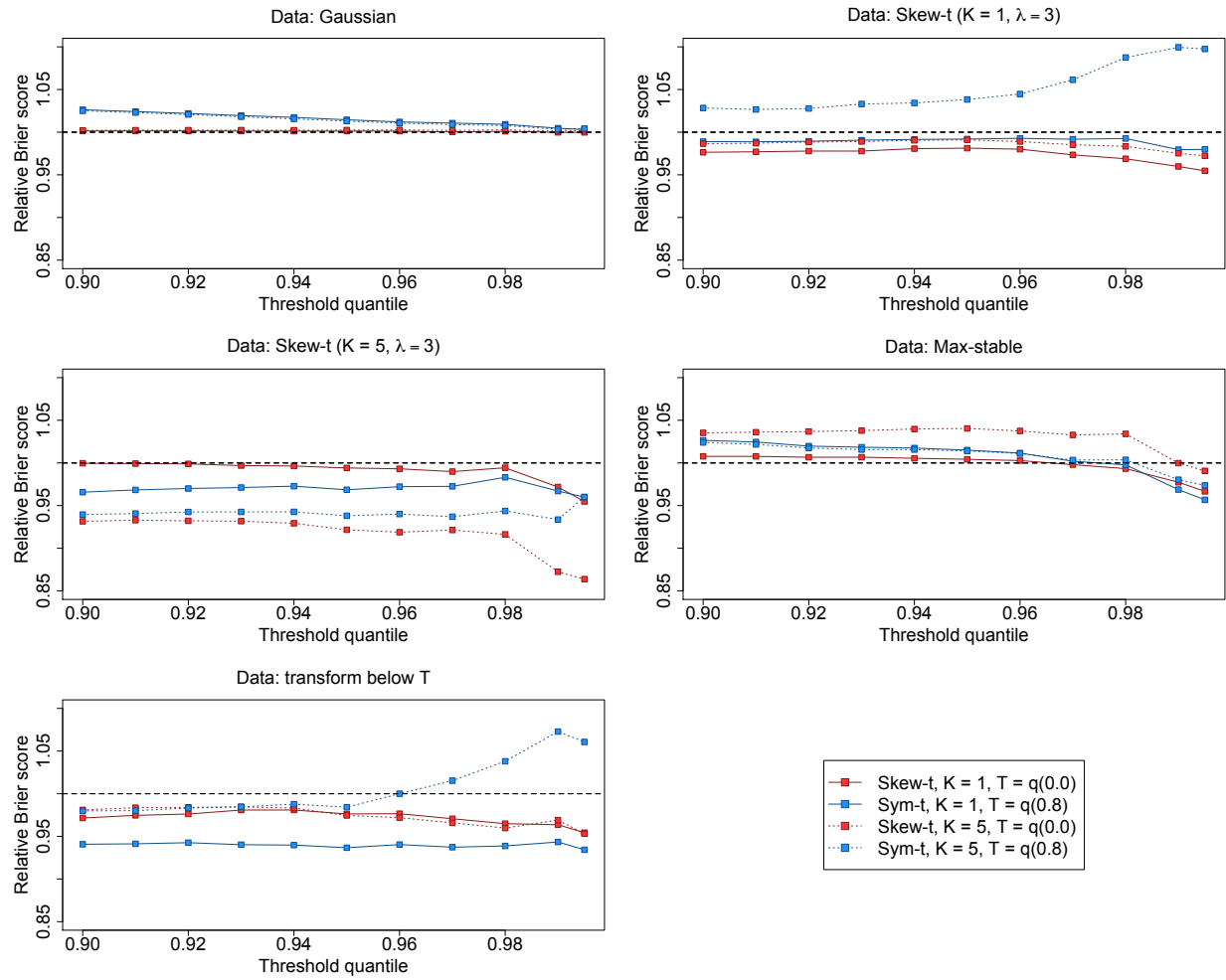
15

Figure 2: Brier scores relative to the Gaussian method for simulation study results. A ratio lower than 1 indicates that the method outperforms the Gaussian method.
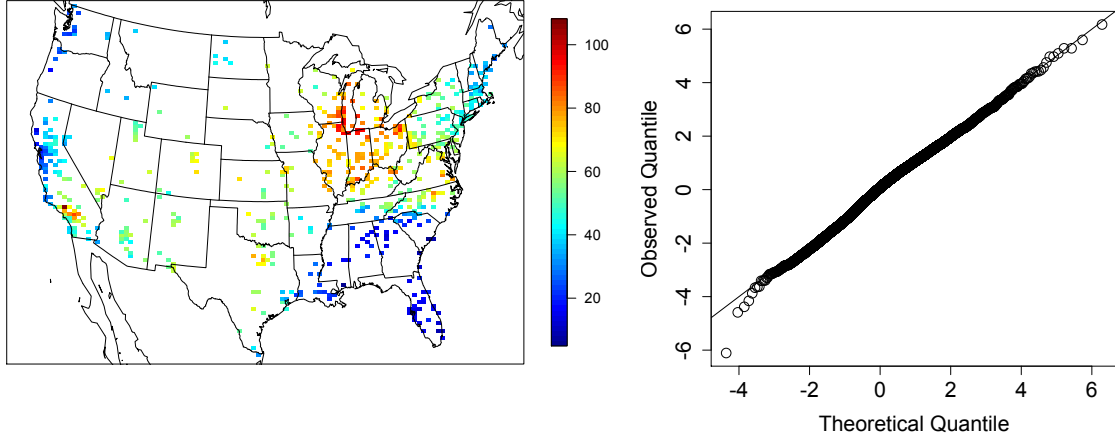
Figure 3: Ozone values on 10 July 2005 (left) Q-Q plot of the residuals (right)

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 $T = 0, 50$ (0.42 sample quantile), and 75 (0.92 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 and Shaby (2012) with the data thresholded at $T = 75$. All methods assume the mean function given in (21). To ensure that the max-stable method runs in a reasonable amount of time, we take a stratified sample of the sites to get 800 sites and consider this our new dataset. We conduct two-fold cross validation using 400 training sites and 400 validation sites as described in Section 5.2

Each chain of the MCMC ran for 30000 iterations with a burn-in period of 25000 iterations. Parameters appear to converge properly; however, as before, for models with multiple partitions it is hard to assess the

17

268 convergence of $\mathbf{w}$, $z(\mathbf{s})$, and $\sigma^2(\mathbf{s})$ because of partition label switching throughout the MCMC. For each

269 model, Brier scores were averaged over all sites and days to obtain a single Brier score for each dataset. At

270 a particular threshold or quantile level, the model that fits the best is the one with the lowest score. We then

271 compute the relative (to Gaussian) Brier scores (see Section 5.3) to compare each model.

6.2 Results

273 The results suggest that the skew-t, thresholded, partitioned, and time series models all give an improvement

274 in predictions over the Gaussian model, whereas the max-stable method results in relative Brier scores

275 between 1.07 and 1.15 indicating poorer performance than the Gaussian model. The plots in Figure 4

276 show the relative Brier scores for time-series and non-time-series models, using $K = 1$, 7, and 15 knots at

277 thresholds $T = 0$, 50, and 75 ppb. Most of the models perform similarly across all the Brier scores; however,

278 for single-partition models without thresholding, performance tends to diminish in the extreme quantiles.

279 The results also suggest that thresholding improves performance for estimates in the extreme quantiles. Both

280 plots have similar features suggesting that most settings do reasonably well. In particular, for all extreme

281 quantiles, selecting a moderate number of knots (e.g. $K = 5, \ldots, 10$) tends to give the best results. Table 1

282 shows the best two models for selected extreme quantiles.

283   We illustrate the predictive capability of our model in Figure 5 by plotting the 99th quantile of the

284 posterior predictive density for July in South Carolina and Georgia. We fit the model using four methods,

285 two reference and two that performed better. These four methods are

286   1. Gaussian (reference)

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

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

289   4. Symmetric-t, $K = 10$ knots, $T = 75$, time series (comparison).
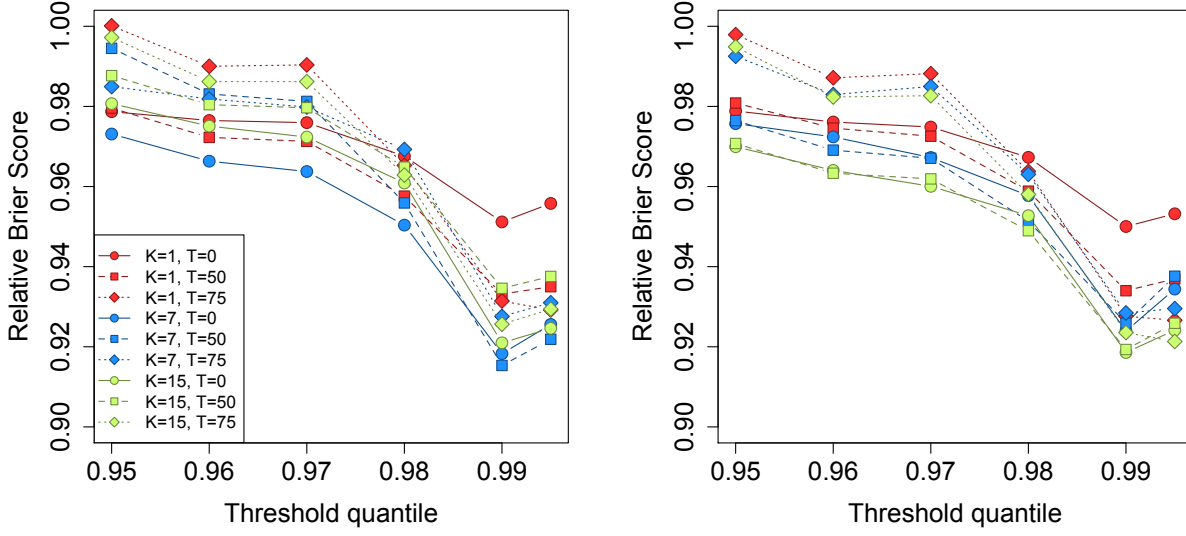
18

Figure 4: 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 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

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.

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
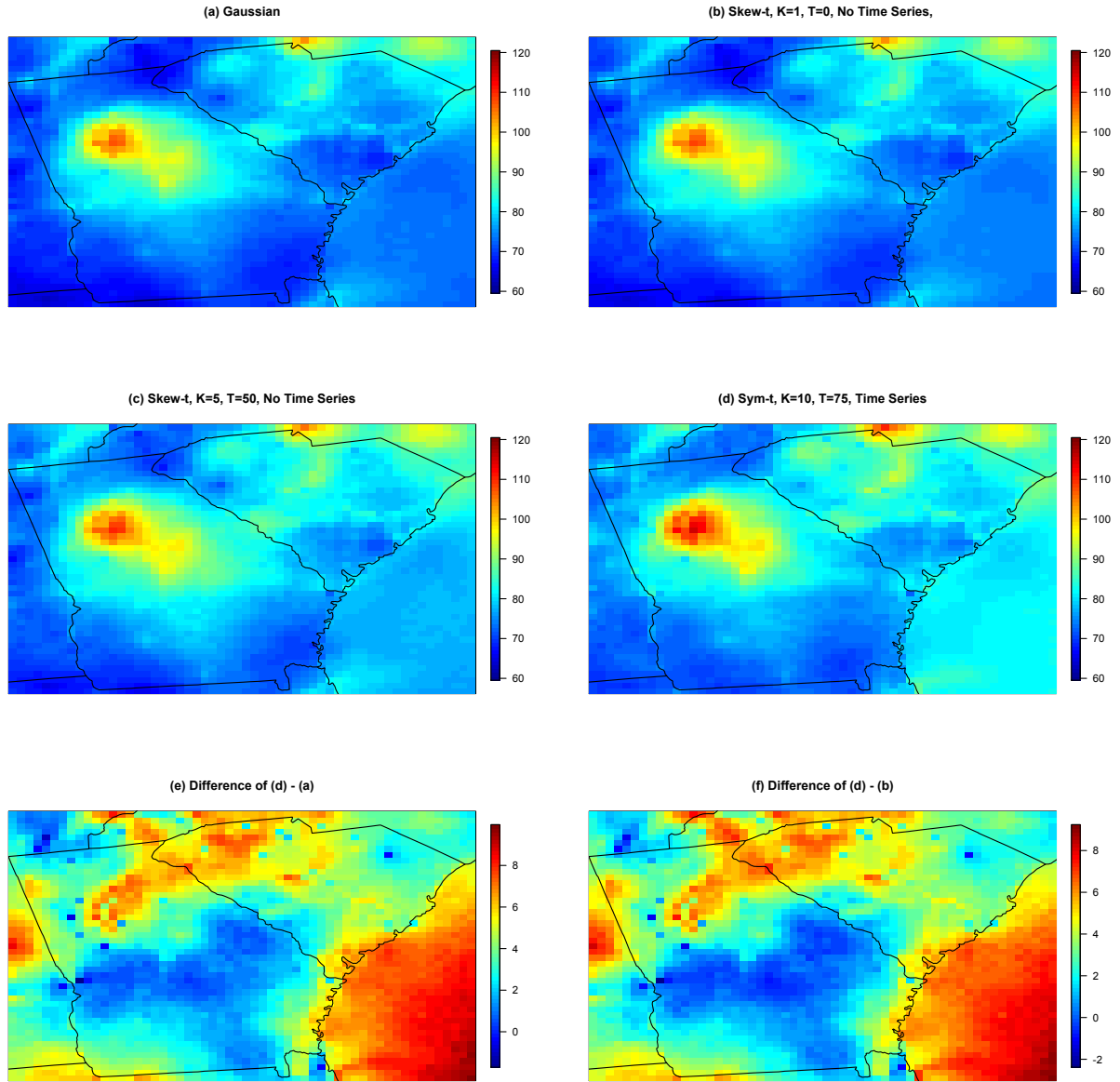
Figure 5: (a) – (d) give give the posterior predictive $\widehat{q}(0.99)$ for the month of July under four different models, (e) gives the difference between $\widehat{q}(0.99)$ in plots (d) and (a), (f) gives the difference between $\widehat{q}(0.99)$ in plots (d) and (b).

for large space-time datasets. In the simulation study, we demonstrate that this model shows statistically significant improvements over a naïve Gaussian approach. In both the simulation study, and the application to ozone data, we find that incorporating a partition in the model improves extreme prediction. 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 incoporated by treating the number of knots as a model parameter instead of fixing it to be a specific value.

Acknowledgments

A   Appendices

A.1   MCMC details

The MCMC sampling for the model 4.1 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. When no conjugate prior distribution exists, we use a random walk Metropolis Hastings update step. In each Metropolis Hastings update, we tune the algorithm 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 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 (18), and $p(\cdot)$ is the prior either taken from the time series given in (3.3) or assumed to be uniform over $\mathcal{D}$. 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 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 3.3) and a random walk Gaussian candidate distribution

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

22

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 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 $\text{IG}(a^*/s, b^*/s)$ candidate distribution in an independence Metropolis Hastings update where $a^* = (n_{tk} + 1)/2 + a$, $b^* = [Y_{tk}^T \Sigma_{tk}^{-1} Y_{tk} + z_{tk}^2]/2 + b$, 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 3.3 or assumed to be $\text{IG}(a, b)$, 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

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

23

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

$$R_1 = \text{the vector of } R(\mathbf{s}) \text{ for } s \in P_l$$

$$R_2 = \text{the vector of } R(\mathbf{s}) \text{ for } s \notin P_l$$

$$\Omega = \Sigma^{-1}.$$

24

Then

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

$$\propto \exp\left\{ -\frac{1}{2}\left[\Lambda_l \tilde{z}_l^2 - 2\mu_l \tilde{z}_l \right] \right\}$$

where

$$\mu_l = \lambda(R_1^T \Omega_{11} + R_2^T \Omega_{21})\mathbf{1}$$

$$\Lambda_l = \lambda^2 \mathbf{1}^T \Omega_{11} \mathbf{1} + \frac{1}{\sigma_l^2}.$$

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

Conditional posterior of $\beta \mid \ldots$

Let $\beta \sim N_p(0, \Lambda_0)$ where Λ_0 is a precision matrix. Then

$$\pi(\beta \mid \ldots) \propto \exp\left\{ -\frac{1}{2}\beta^T \Lambda_0 \beta - \frac{1}{2}\sum_{t=1}^{T}[\mathbf{Y}_t - X_t\beta - \lambda|z_t|]^T \Omega[\mathbf{Y}_t - X_t\beta - \lambda|z_t|] \right\}$$

$$\propto \exp\left\{ -\frac{1}{2}\left[\beta^T \Lambda_\beta \beta - 2\sum_{t=1}^{T}[\beta^T X_t^T \Omega(\mathbf{Y}_t - \lambda|z_t|)] \right] \right\}$$

$$\propto N(\Lambda_\beta^{-1}\mu_\beta, \Lambda_\beta^{-1})$$

25

358 where

$$\mu_\beta = \sum_{t=1}^{T} \left[X_t^T \Omega (\mathbf{Y}_t - \lambda |z_t|) \right]$$

$$\Lambda_\beta = \Lambda_0 + \sum_{t=1}^{T} X_t^T \Omega X_t.$$

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

360 In the case where $L = 1$ and temporal dependence is negligible, then σ^2 has a conjugate posterior distribu-

361 tion. Let $\sigma_t^2 \overset{iid}{\sim} \text{IG}(\alpha_0, \beta_0)$. For simplicity, drop the subscript t. Then

$$\pi(\sigma^2 \mid \ldots) \propto (\sigma^2)^{-\alpha_0 - 1/2 - n/2 - 1} \exp \left\{ -\frac{\beta_0}{\sigma^2} - \frac{|z|^2}{2\sigma^2} - \frac{(\mathbf{Y} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\mu})}{2\sigma^2} \right\}$$

$$\propto (\sigma^2)^{-\alpha_0 - 1/2 - n/2 - 1} \exp \left\{ -\frac{1}{\sigma^2} \left[\beta_0 + \frac{|z|^2}{2} + \frac{1}{2} (\mathbf{Y} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\mu}) \right] \right\}$$

$$\propto \text{IG}(\alpha^*, \beta^*)$$

362 where

$$\alpha^* = \alpha_0 + \frac{1}{2} + \frac{n}{2}$$

$$\beta^* = \beta_0 + \frac{|z|^2}{2} + \frac{1}{2} (\mathbf{Y} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\mu}).$$

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

Conditional posterior of $\lambda \mid \dots$

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

$$\lambda_1 = \begin{cases} +1 & \text{w.p.0.5} \\ \\ -1 & \text{w.p.0.5} \end{cases} \tag{22}$$

$$\lambda_2^2 \sim IG(\alpha_\lambda, \beta_\lambda). \tag{23}$$

$$\tag{24}$$

Then

$$\pi(\lambda_2 \mid \dots) \propto \lambda_2^{2(-\alpha_\lambda - 1)} \exp\left\{-\frac{\beta_\lambda}{\lambda_2^2}\right\} \prod_{t=1}^{T} \prod_{k=1}^{K} \frac{1}{\lambda_2} \exp\left\{-\frac{z_{tk}^2}{2\lambda_2^2 \sigma_{tk})^2}\right\}$$

$$\propto \lambda_2^{2(-\alpha_\lambda - kt - 1)} \exp\left\{-\frac{1}{\lambda_2^2}\left[\beta_\lambda + \frac{z^2}{2\sigma_{tk}^2}\right]\right\}$$

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

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

Let $N(A)$ be the number of knots in A, the area between sites $\mathbf{s}_1$ and $\mathbf{s}_2$. Consider a spatial Poisson process with intensity $\mu(A)$. So,

$$P[N(A) = k] = \frac{\mu(A)^k \exp\{-\mu(A)\}}{k!}.$$

Then for any finite k, $\lim_{h \to \infty} P[N(A) = k] = 0$ because $\lim_{h \to \infty} \mu(A) = \infty$. With each additional knot in A, the chance that $\mathbf{s}_1$ and $\mathbf{s}_2$ will be be in the same partition will decrease, because partition membership is defined by the closest knot to a site. Therefore, $\lim_{h \to \infty} \pi(h) = 0$.

27

A.4   Skew-t distribution

Univariate extended skew-t distribution

We say that Y follow a univariate extended skew-t distribution with location $\xi \in \mathcal{R}$, scale $\omega > 0$, skew

parameter $\alpha \in \mathcal{R}$, extended parameter $\tau \in \mathcal{R}$, and degrees of freedom ν if has distribution function

$$f_{\text{EST}}(y) = \omega^{-1} \frac{f_T(z; \nu)}{F_T(\tau/\sqrt{1+\alpha^2}; \nu)} F_T \left[(\alpha z + \tau) \sqrt{\frac{\nu+1}{\nu+z^2}}; 0, 1, \nu+1 \right] \tag{25}$$

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

In the case that $\tau = 0$, then Y follows a univariate skew-t distribution.

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

$$f_y(\mathbf{y}) = det(\boldsymbol{\omega})^{-1} f_z(\mathbf{z}) \tag{26}$$

where

$$f_z(\mathbf{z}) = 2t_d(\mathbf{z}; \bar{\mathbf{\Omega}}, \eta) T \left[\boldsymbol{\alpha}^T \mathbf{z} \sqrt{\frac{\eta+d}{\nu+Q(\mathbf{z})}}; \eta+d \right] \tag{27}$$

$$\mathbf{z} = \boldsymbol{\omega}^{-1}(\mathbf{y} - \xi) \tag{28}$$

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, $Q(z) = \mathbf{z}^T \bar{\Omega}^{-1} \mathbf{z}$ and $T(\cdot; \eta)$ denotes the univariate Student's t distribution function with η degrees of

freedom (Azzalini and Capitanio, 2014).

Extremal dependence

For a bivariate skew-t random variable $\mathbf{Y} = [Y(\mathbf{s}), Y(\mathbf{t})]^T$, the $\chi(h)$ statistic (Padoan, 2011) is given by

$$\chi(h) = \bar{F}_{\text{EST}}\left\{\frac{[x_1^{1/\eta} - \varrho(h)]\sqrt{\eta+1}}{\sqrt{1-\varrho(h)^2}}; 0, 1, \alpha_1, \tau_1, \eta+1\right\} + \bar{F}_{\text{EST}}\left\{\frac{[x_2^{1/\eta} - \varrho(h)]\sqrt{\eta+1}}{\sqrt{1-\varrho(h)^2}}; 0, 1, \alpha_2, \tau_2, \eta+1\right\},$$
(29)

where $\bar{F}_{\text{EST}}$ is the univariate survival extended skew-t function with zero location and unit scale, $\varrho(h) = \text{cor}(y_1, y_2)$,

$\alpha_j = \alpha_i\sqrt{1-\varrho^2}$, $\tau_j = \sqrt{\eta+1}(\alpha_j + \alpha_i\varrho)$, 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 $\mathbf{Y} = [Y(\mathbf{s}), Y(\mathbf{t})]^T$, with $\varrho(h)$ given by (3). So, $\lim_{h\to\infty} \varrho(h) = 0$.

Then

$$\lim_{h\to\infty} \chi(h) = \bar{F}_{\text{EST}}\left\{\sqrt{\eta+1}; 0, 1, \alpha_1, \tau_1, \eta+1\right\} + \bar{F}_{\text{EST}}\left\{\sqrt{\eta+1}; 0, 1, \alpha_2, \tau_2, \eta+1\right\}.$$
(30)

Because the extended skew-t distribution is not bounded above, for all $\bar{F}_{\text{EST}}(x) = 1 - F_{\text{EST}} > 0$ for all

$x < \infty$. Therefore, for a skew-t distribution, $\lim_{h\to\infty} \chi(h) > 0$.

A.5   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 signifi-

cantly 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

29

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 2: 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
Method 2	A B	A B	A	A
Method 3	C	C	C	B
Method 4	B	B	B	B
Method 5	C	C	C	B

Table 3: Setting 2 – Skew-t marginal, $K = 1$ knot

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

Table 4: Setting 3 – Skew-t marginal, $K = 5$ knots

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

References

Azzalini, A. and Capitanio, A. (2014) *The Skew-Normal and Related Families*. Institute of Mathematical Statistics Monographs. Cambridge University Press. URLhttps://books.google.com/books?id=-tkaAgAAQBAJ.

Coles, S., Heffernan, J. and Tawn, J. (1999) Dependence Measures for Extreme Value Analyses. *Extremes*, **2**, 339–365. URLhttp://link.springer.com/article/10.1023/A:1009963131610.

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

Table 5: Setting 4 – Max-stable

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

Table 6: 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	A B	A B
Method 3	A	A	A	A
Method 4	B C	B	B	B C
Method 5	B C	B	C	C

Engelke, S., Malinowski, A., Kabluchko, Z. and Schlather, M. (2014) Estimation of Hüsler-Reiss distributions and Brown-Resnick processes. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 239–265.

Genton, M. G. (2004) *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Statistics (Chapman & Hall/CRC). Taylor & Francis. URLhttp://books.google.com/books?id=R9hBnwEACAAJ.

Gneiting, T. and Raftery, A. E. (2007) Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102**, 359–378. URLhttp://www.tandfonline.com/doi/abs/10.1198/016214506000001437.

Huser, R. and Davison, A. C. (2014) Space-time modelling of extreme events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**, 439–461. URLhttp://arxiv.org/abs/1201.3245 http://doi.wiley.com/10.1111/rssb.12035.

Kim, H.-M., Mallick, B. K. and Holmes, C. C. (2005) Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes. *Journal of the American Statistical Association*, **100**, 653–668.

Ledford, A. W. and Tawn, J. a. (1996) Statistics for near independence in multivariate extreme values. *Biometrika*, **83**, 169–187. URLhttp://biomet.oxfordjournals.org/cgi/content/abstract/83/1/169.

Padoan, S. A. (2011) Multivariate extreme models based on underlying skew- and skew-normal distributions. *Journal of Multivariate Analysis*, **102**, 977–991. URLhttp://linkinghub.elsevier.com/retrieve/pii/S0047259X11000157.

Padoan, S. A., Ribatet, M. and Sisson, S. A. (2010) Likelihood-Based Inference for Max-Stable Processes. *Journal of the American Statistical Association*, **105**, 263–277. URLhttp://www.tandfonline.com/doi/abs/10.1198/jasa.2009.tm08577.

Reich, B. J. and Shaby, B. A. (2012) A hierarchical max-stable spatial model for extreme precipitation. *The Annals of Applied Statistics*, **6**, 1430–1451. URLhttp://projecteuclid.org/euclid.aoas/1356629046.

Thibaud, E., Mutzner, R. and Davison, a. C. (2013) Threshold modeling of extreme spatial rainfall. *Water Resources Research*, **49**, 4633–4644.

Thibaud, E. and Opitz, T. (2013) Efficient inference and simulation for elliptical Pareto processes. 1–19.

Wadsworth, J. L. and Tawn, J. a. (2012) Dependence modelling for spatial extremes. *Biometrika*, **99**, 253–272.

— (2014) Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika*, **101**, 1–15. URLhttp://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/ast042.