BIOMETRICS 000, 1–20 DOI: 000
000 0000

A space-time skew-t model for threshold exceedances

Samuel A Morris[1,*], Brian J Reich[1], Emeric Thibaud[2], and Daniel Cooley[2]

[1]Department of Statistics, North Carolina State University, Raleigh, North Carolina, U.S.A.

[2]Department of Statistics, Colorado State University, Fort Collins, Colorado, U.S.A.

**email:* samorris@ncsu.edu

SUMMARY: To assess the compliance of air quality regulations, the Environmental Protection Agency (EPA) must know if a site exceeds a pre-specified level. In the case of ozone, the level for compliance is fixed at 75 parts per billion, which is high, but not extreme at all locations. We present a new space-time model for threshold exceedances based on the skew-t process. Our method incorporates a random partition to permit long-distance asymptotic independence while allowing for sites that are near one another to be asymptotically dependent, and we incorporate thresholding to allow the tails of the data to speak for themselves. We also introduce a transformed AR(1) time-series to allow for temporal dependence. Finally, our model allows for high-dimensional Bayesian inference that is comparable in computation time to traditional geostatistical methods for large datasets. We apply our method to an ozone analysis for July 2005, and find that our model improves over both Gaussian and max-stable methods in terms of predicting exceedances of a high level.

KEY WORDS: Extreme value analysis, Markov chain Monte Carlo, Random partition, Skew-t , Spatio-temporal modeling

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Epidemiological studies have linked air quality to public health concerns regarding morbidity and mortality (Samet et al., 2000). 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 on the behavior over a high level 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 three years does not exceed 75 parts per billion (ppb). Figure 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. We analyze these data with the goals of spatial prediction for unmonitored locations and to map the probability of extreme events.

[Figure 1 about here.]

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 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 pre-selected threshold and will censor 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) are notions which describe the probability that two random variables simultaneously exceed of an extremely high level. 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 and Smith (1990) considered modeling threshold exceedances of univariate time series by the generalized Pareto distribution. Threshold based inference for multivariate extreme value distributions was considered by Ledford and Tawn (1996) who introduced a censored approach that provides a way to deal with different types of exceedances of a threshold. These models were extended to spatial models for threshold exceedances 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 and Davison (2014) further extended this to space-time modeling. Wadsworth and Tawn (2014), Engelke et al. (2015), and Thibaud and Opitz (2015) 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. Our application with ozone data does not fit into this framework because we do not focus on exceedances of a very high level, and we have observations at 1,089 ozone monitoring locations.

We propose a new spatiotemporal threshold exceedance model based on the skew-t process

(Padoan, 2011). We use a skew-t distribution because of its flexibility to model asymmetry and heavy-tailed data with the aim of modeling exceedances of a high fixed level at an unobserved location. 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, which we use to fit the model, handle censored data below the threshold, and make predictions at unobserved locations. The multivariate skew normal distribution was introduced by Azzalini and Dalla Valle (1996), and this was extended to the multivariate skew-t by Branco and Dey (2001). These skew-elliptical distributions have been used in the spatial setting (Genton, 2004; Kim and Mallick, 2004). Zhang and El-Shaarawi (2010) propose the skew-Gaussian process as a class of stationary processes that have skewed marginal distributions. Padoan (2011) examined the usage of skew-Gaussian and skew-t distributions for multivariate extremes. 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.

The paper is organized as follows. Section 2 is a brief review of the spatial skew-t process. In Section 3, we build upon the traditional skew-t process by incorporating censoring, partitioning, 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 model compared to Gaussian and max-stable methods. We compare our method to Gaussian and max-stable methods with a data analysis of ozone measurements throughout the US in Section 6.

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 (Azzalini and Capitanio, 2003; Azzalini and Capitanio, 2014, p. 129; Beranger et al., 2016) by which a skew-t process is created is given here.

2.1 *Skew-t process*

Let $Y(\mathbf{s})$ be a spatial process defined for spatial location $\mathbf{s}$ in a spatial domain of interest $\mathcal{D} \in \mathcal{R}^2$. The spatial skew-$t$ process can be written as

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

where $\mathbf{X}(\mathbf{s})$ is the observed covariate vector at site $\mathbf{s}$, $\boldsymbol{\beta}$ is the p-vector of regression parameters, $\lambda \in \mathcal{R}$ is a parameter controlling skewness, $z \sim N(0,1)$, $\sigma^2 \sim \text{IG}(a/2, b/2)$ 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 standard Gaussian process with positive definite correlation function $\text{Cor}\left[Y(\mathbf{s}_1), Y(\mathbf{s}_2)\right] = \rho(\mathbf{s}_1, \mathbf{s}_2)$. Although any positive definite correlation function could be used, we choose to use the stationary isotropic Matérn correlation with

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

where $I(\cdot)$ is an indicator function, $\varphi > 0$ is the spatial range, $\nu > 0$ is the smoothness, $\gamma \in [0,1]$ 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 Euclidean 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 $\mathbf{Y} = [Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)]^\top$, and the covariate matrix $\mathbf{X}_{n\times p} = [\mathbf{X}(\mathbf{s}_1), \ldots, \mathbf{X}(\mathbf{s}_n)]^\top$. After marginalizing over both z and σ, using the notation from Azzalini and Capitanio (2014, p. 176),

$$\mathbf{Y} \sim \text{ST}_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, a), \tag{3}$$

that is, $\mathbf{Y}$ follows an n-dimensional skew-t distribution with location $\mathbf{X}\boldsymbol{\beta} \in \mathcal{R}^n$; covariance matrix $\boldsymbol{\Omega}_{n\times n} = \boldsymbol{\omega}\left[\frac{1}{1+\lambda^2}(\boldsymbol{\Sigma} + \lambda^2\mathbf{1}\mathbf{1}^\top)\right]\boldsymbol{\omega}$, $\boldsymbol{\Sigma}_{n\times n}$ is the positive definite correlation matrix which is obtained from $\rho(h)$, and $\boldsymbol{\omega}_{n\times n} = \text{diag}\left(\sqrt{\frac{b}{a(1+\lambda^2)}}, \ldots, \sqrt{\frac{b}{a(1+\lambda^2)}}\right)$; slant parameters $\boldsymbol{\alpha} \in \mathcal{R}^n = \lambda(1+\lambda^2)^{1/2}(1+\lambda^2\mathbf{1}^\top\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1/2}\boldsymbol{\Sigma}^{-1}\mathbf{1}$, and degrees of freedom a. Furthermore, the marginal distributions at each location also follow a univariate skew-t distribution (Azzalini and Capitanio,

2014). This process is desirable because it is heavy tailed with tail index a, and the shape of the distribution is controlled by the skewness parameter. For a comparison with other parameterizations, see Web Appendix E.

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 locations separated by distance h is

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

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 Web Appendix D). 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 desirable for spatial analysis of large geographic regions where we expect only local spatial dependence. We propose a solution to this in Section 3.2.

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 3.3.

3.1 *Censoring to focus on the tail*

We do not want the low-to-moderate values to influence the fit of the model. We propose the use of a censored approach to fit threshold exceedances only. More specifically, we assume our skew-t model $Y(\mathbf{s})$ is valid at each location $\mathbf{s}$ above a threshold T, and censor the values below T for which we don't assume the model to be valid. We define our partially censored observations as $\widetilde{\mathbf{Y}} = [\widetilde{Y}(\mathbf{s}_1), \ldots, \widetilde{Y}(\mathbf{s}_n)]^\top$ where $\widetilde{Y}(\mathbf{s}) = \max\{Y(\mathbf{s}), T\}$, and fit the skew-$t$ process to these $\widetilde{\mathbf{Y}}$. In our Bayesian framework, inference can be easily performed by imputing censored observations below T (see Section 4.1).

As our goal is to model exceedances above a high level L, we should select a value for $T \leqslant L$. For example, in predicting ozone exceedances, we might set $T = 50$ ppb in order to predict exceedances of $L = 75$ ppb. Selecting T too small may lead to bias in estimating the tail parameters; selecting T too large increases variance. We impute the censored values as a step in the algorithm used to fit the model described in Section 4.1, and use cross-validation to select T.

3.2 *Partitioning to remove long-range asymptotic dependence*

For a large spatial domain, it may not be reasonable to assume sites that are far apart demonstrate asymptotic dependence. 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 using a partitioning approach. The model becomes

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})^\top \boldsymbol{\beta} + \lambda\sigma(\mathbf{s})|z(\mathbf{s})| + \sigma(\mathbf{s})v(\mathbf{s}). \tag{5}$$

To model spatial variation, consider a set of spatial knots $\mathbf{w}_1, \ldots, \mathbf{w}_K$ from a homogeneous Poisson process 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

$$P_k = \{\mathbf{s} : k = \arg\min_\ell ||\mathbf{s} - \mathbf{w}_\ell||\}. \tag{6}$$

In other words, P_k is composed of all sites for which the closest knot is $\mathbf{w}_k$. 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 $\sigma_k^2 \overset{iid}{\sim} \text{IG}(a/2, b/2)$. So, within each partition, $Y(\mathbf{s})$ follows the spatial skew-t process defined in Section 2. Across partitions, the $Y(\mathbf{s})$ remain dependent via the correlation function for $v(\mathbf{s})$ because $v(\mathbf{s})$ spans all partitions. However, the bivariate distribution for sites in different partitions is neither Gaussian nor skew-t and does not have asymptotic dependence.

The partitioning model removes 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

$$\chi(h) = I(k_1 = k_2)\chi_{\text{skew-}t}(h) \tag{7}$$

where $\chi_{\text{skew-}t}(h)$ is the χ statistic for a skew-t process given in equation (8) of Web Appendix D, and $h = ||\mathbf{s}_1 - \mathbf{s}_2||$. 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 Web Appendix C, we show that assuming the knots follow a homogeneous Poisson process, $\lim_{h\to\infty} \pi(h) = 0$, and thus long-range dependence is removed. In practice we fix K at a finite value and use a uniform distribution for the knots $\mathbf{w}_1, \ldots, \mathbf{w}_K$. In Figure 2, we show $\chi(h)$ for $K = 1, 3, 5, 10$ partitions for a skew-t distribution with $\alpha = 10$, and 3 degrees of freedom.

[Figure 2 about here.]

3.3 *Extension to space-time data*

When using daily measurements, the assumption of temporal independence is often inappropriate. In this section, we extend (5) to the spatiotemporal setting. There are several places where temporal dependence could be incorporated in the model, including the Gaussian process $v_t(\mathbf{s})$. However, we choose to allow for temporal dependence in the $\mathbf{w}$, z, and σ terms because these terms dictate the tail behavior of the process which is our primary focus (see Web Appendix F for a discussion

of the induced temporal asymptotic dependence). Let

$$Y_t(\mathbf{s}) = \mathbf{X}_t(\mathbf{s})^\top \boldsymbol{\beta} + \lambda \sigma_t(\mathbf{s})|z_t(\mathbf{s})| + \sigma_t(\mathbf{s}) v_t(\mathbf{s}), \tag{8}$$

where $t \in \{1, \ldots, n_t\}$ 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 the collection of spatial knots on day t. As in Section 3.2, these knots define a daily partition $P_{t1}, \ldots, P_{tK}$, and for $\mathbf{s} \in P_{tk}$,

$$z_t(\mathbf{s}) = z_{tk} \quad \text{and} \quad \sigma_t(\mathbf{s}) = \sigma_{tk}. \tag{9}$$

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 the parameter 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 $\mathcal{D}$ to $\mathcal{R}^2$ as follows. Let

$$w^*_{tki} = \Phi^{-1}\left[\frac{w_{tki} - \min(\mathbf{s}_i)}{\max(\mathbf{s}_i) - \min(\mathbf{s}_i)}\right], \quad i = 1, 2 \tag{10}$$

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 transform $z_t(\mathbf{s})$ to the Gaussian scale using the probability integral transform to ensure that the marginal distributions of $z_t(\mathbf{s})$ are half-normal. Let

$$z^*_t(\mathbf{s}) = \Phi^{-1}\left\{\text{HN}[z_t(\mathbf{s})]\right\} \tag{11}$$

where HN is the distribution function of a half-normal random variable. We also transform $\sigma^2_t(\mathbf{s})$ to the Gaussian scale using the probability integral transform to ensure that the marginal distributions of $\sigma^2_t(\mathbf{s})$ are inverse gamma. Let

$$\sigma^{2*}_t(\mathbf{s}) = \Phi^{-1}\left\{\text{IG}[\sigma^2_t(\mathbf{s})]\right\} \tag{12}$$

where IG is defined as before. The AR(1) process for each tail parameter is $\mathbf{w}^*_{1k} \sim N_2(\mathbf{0}, \mathbf{I}_2)$ where $\mathbf{I}_2 = \text{diag}(1,1)$, $z^*_{1k} \sim N(0,1)$, $\sigma^{2*}_{1k} \sim N(0,1)$, and for $t > 1$ the time series is modeled as

$$\mathbf{w}^*_{tk}|\mathbf{w}^*_{t-1,k} \sim N_2\left[\phi_w \mathbf{w}^*_{t-1,k}, (1-\phi_w^2)\mathbf{I}_2\right] \tag{13}$$

$$z^*_{tk}|z^*_{t-1,k} \sim N\left[\phi_z z^*_{t-1,k}, (1-\phi_z^2)\right] \tag{14}$$

$$\sigma^{2*}_{tk}|\sigma^{2*}_{t-1,k} \sim N\left[\phi_\sigma \sigma^{2*}_{t-1,k}, (1-\phi_\sigma^2)\right] \tag{15}$$

where $|\phi_w|$, $|\phi_z|$, $|\phi_\sigma| < 1$. These are stationary time series models with marginal distributions $\mathbf{w}^*_k \sim N_2(\mathbf{0}, \mathbf{I}_2)$ $z^*_k \sim N(0,1)$, and $\sigma^{2*}_k \sim N(0,1)$. After transformation back to the original space, $\mathbf{w}_{tk} \sim \text{Unif}(\mathcal{D})$, $z_{tk} \sim \text{HN}(0,1)$, and $\sigma^2_{tk} \sim \text{IG}(a/2, b/2)$. 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 (5) by construction.

4. Hierarchical model

We define a Bayesian hierarchical model based on the skew-t process and use a Markov chain Monte Carlo (MCMC) algorithm to fit the data. We model our data as partially censored observations (see Section 3.1) from the skew-t model defined in Section 3. In the first step of the MCMC algorithm, we impute censored values of $\widetilde{Y}$ such that the estimation of model parameters can be based on the completed $\mathbf{Y}$. Conditioned on $z_{tk}(\mathbf{s})$, $\sigma^2_{tk}(\mathbf{s})$, and P_{tk}, joint distribution of $\mathbf{Y}$ is multivariate Gaussian. We do not fix the partitions; instead, the locations of the K knots are treated as unknown and random. 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 values and assessing its impact on prediction as

described in Section 5.2. Then the hierarchical model is given as

$$Y_t(\mathbf{s}) \mid z_t(\mathbf{s}), \sigma_t^2(\mathbf{s}), P_{tk}, \Theta = \mathbf{X}_t(\mathbf{s})^\top \beta + \lambda \sigma_t(\mathbf{s}) |z_t(\mathbf{s})| + \sigma_t(\mathbf{s}) v_t(\mathbf{s}) \tag{16}$$

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

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

$$\lambda \sim N(0, \sigma_\lambda^2)$$

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

where $\Theta = \{\varphi, \nu, \gamma\, \lambda, \boldsymbol{\beta}\}$, $\mathbf{\Sigma}$ is a Matérn covariance matrix as described in Section 2.1, and the priors on $\mathbf{w}_{tk}^*$, z_{tk}^*, and σ_{tk}^{2*} are given in (13) – (15).

4.1 *Computation*

We use MCMC methods to estimate the posterior distribution of the model parameters. At each MCMC iteration, we first impute values below the threshold conditional on observations above the threshold. After conditioning on λ, $z_t(\mathbf{s})$ and non-censored observations, $Y_t(\mathbf{s})$ has truncated normal full conditionals $Y_t(\mathbf{s}) \sim N_{(-\infty,T)}(\mathbf{X}_t^\top(\mathbf{s})\beta + \lambda|z_t(\mathbf{s})|, \mathbf{\Sigma})$. To impute the censored observations, at each MCMC iteration, for censored site i and day t, we use the conditional distribution of $\widetilde{Y}_t(\mathbf{s}_i)$ given the current imputed values for all censored sites and true values for all non-censored sites $\widetilde{Y}_t^{(c)}(\mathbf{s}_{-i})$. Conditional on $\widetilde{Y}_t^{(c)}(\mathbf{s}_{-i})$, the imputation of $\widetilde{Y}_t(\mathbf{s}_i)$ is done using a univariate truncated normal likelihood with mean and standard deviation from the conditional multivariate normal distribution of $\widetilde{Y}_t(\mathbf{s})$.

We update model parameters Θ using Gibbs sampling with Metropolis-Hastings steps when needed. In our case, we also wish to be able to make predictions at sites where we do not have data. We can easily implement Bayesian Kriging as a part of the algorithm 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.

5. Simulation study

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

5.1 *Design*

For all simulation designs, we generated data from model (5) in Section 3.2 using $n_s = 144$ sites and $n_t = 50$ independent days. The sites were generated uniformly on the square $[0, 10] \times [0, 10]$. We generated data from five different simulation designs:

(1) Gaussian, $K = 1$ knot

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

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

(4) Reich and Shaby (2012) max-stable process

(5) Brown-Resnick max-stable process (Kabluchko et al., 2009)

In the first three designs, the realizations from $v_t(\mathbf{s})$ were generated using a Matérn covariance with smoothness parameter $\nu = 0.5$, spatial range $\varphi = 1$ and $\gamma = 0.9$. 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 Section 3.2. In designs 1 – 3, the mean $\mathbf{X}^\top \boldsymbol{\beta} = 10$ was assumed to be constant across space. In the fourth design, we generated from the max-stable model of Reich and Shaby (2012). The marginal distributions follow a generalized extreme value distribution with location parameter 1, scale parameter 1, and shape parameter 0.2. Spatial dependence in the form an asymmetric logistic dependence function is induced by random effects for kernel basis functions associated with 144 spatial knots defined on a square grid on $[1, 9] \times [1, 9]$. We set the dependence parameter (α in Reich and Shaby, 2012) to 0.5 which represents moderate spatial dependence. For the final design, we generated data from

a Brown-Resnick max-stable process using `rmaxstab` in the `SpatialExtremes` package of `R` (Ribatet, 2015). For this design we fixed unit Fréchet margins, and we used a range of 1 and smoothness 0.5.

$M = 50$ data sets were 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) Reich and Shaby (2012) max-stable model thresholded at $T = q(0.80)$

where $q(0.80)$ is the 80th sample quantile of the data. All methods were fit using a fully-Bayesian approach that simultaneously estimates marginal and spatial dependence parameters. The design matrix $\mathbf{X}$ includes an intercept with a first-order spatial trend with priors of $\beta_{\text{int}}, \beta_{\text{lat}}, \beta_{\text{long}}, \overset{iid}{\sim}$ N$(0, 10)$ although only the intercept is used in the data generation. The spatial covariance parameters have priors $\log(\nu) \sim$ N$(-1.2, 1)$, $\gamma \sim$ Unif$(0, 1)$, $\varphi \sim$ Unif$(0, 15)$. The skewness parameter has prior $\lambda \sim N(0, 20)$. The residual variance terms have priors $\sigma_t^2(\mathbf{s}) \sim$ IG$(a/2, b/2)$, 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. As described in Section 2.1, these priors are meant to be fairly uninformative to allow the data to dictate both the degrees of freedom a and precision b of the process. The knots have priors $\mathbf{w} \sim$ 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 a censored left tail. Each chain of the MCMC ran for 20,000 iterations which includes a burn-in period of 10,000 iterations. Although the goal of this simulation study is not to assess parameter estimation, for design (2) and method (2) where data are generated and fit with a skew-t distribution, the samples

converges for λ and a, and the empirical coverage of posterior 95% intervals is near the nominal level (96% coverage for λ and 90% for a). It should be noted that 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; however, we are not interested in these parameters but rather spatial predictions and tail probabilities which converge well. Finally, we did not fit a Brown-Resnick model because we cannot use a Bayesian approach with so many sites.

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 high level, we use Brier scores to compare the models (Gneiting and Raftery, 2007). The Brier score for predicting exceedance of a level L is given by $[e(L) - P(L)]^2$ where $e(L) = I[y > L]$ is an indicator function indicating that a test set value, y, has exceeded the level, L, and $P(L)$ is the predicted probability of exceeding L. 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 four different high levels for each dataset. The levels used for the Brier scores are extreme quantiles from the simulated data for $L = q(0.90)$, $q(0.95)$, $q(0.98)$, $q(0.99)$. Figure 3 gives the Brier score relative to the Brier score for the Gaussian method calculated as

$$\text{BS}_{\text{rel}} = \frac{\text{BS}_{\text{method}}}{\text{BS}_{\text{Gaussian}}}. \tag{17}$$

We analyzed the results for the simulation study using a Friedman (Hollander et al., 2014) 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 (Hollander et al., 2014) test to see which of the methods had different results. The full results for the Wilcoxon-Nemenyi-McDonald-Thompson tests are given in Web Appendix J.

In general, we find that when the method to fit the data matches the data generation scheme, there is some improvement over other methods. 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. We also see that 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. Finally, in the case where the data are generated from a Brown-Resnick process, we find that our method is competitive with using a max-stable model. In summary, our method provides great flexibility for data that demonstrate some level of asymmetry and heavy tails, while still performing comparably to Gaussian methods when the data are symmetric and have light tails.

[Figure 3 about here.]

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 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 $\mathbf{X}_t(\mathbf{s}) = [1, \text{CMAQ}_t(\mathbf{s})]^\top$. Figure 4 shows a Q-Q plot of the residuals compared to a skew-t distribution with $a = 10$ and $\lambda = 1$, suggesting the data are heavy tailed.

[Figure 4 about here.]

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, in Web Appendix G, we present a plot that suggests there is more agreement between sites that are close to one another than sites that are far from one another.

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 censored $Y(\mathbf{s})$ at $T = 0$, $T = 50$ (0.42 sample quantile), and $T = 75$ (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 used a symmetric-t marginal distribution. We also compared models with no time series to models that included 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 assumed $\mathbf{X}_t(\mathbf{s}) = [1, \text{CMAQ}_t(\mathbf{s})]^\top$. To ensure that the max-stable method ran in a reasonable amount of time, we used a stratified sub-sample of 800 sites. We conducted 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 30,000 iterations which includes a burn-in period of 25,000 iterations. We used the same priors for the spatial covariance parameters, skewness parameter, and knots as in the simulation study. The prior for the residual variance terms was $\sigma^2_t(\mathbf{s}) \sim \text{IG}(a/2, b/2)$ where a was the same as the simulation study, but b had a $\text{Gamma}(1, 1)$ prior. Parameters appeared to converge properly; however, as before, for models with multiple partitions it was 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, we averaged Brier scores over all sites and days to obtain a single Brier score for each dataset. At a particular level, the model that fit the best was the one with the lowest score. We then computed the relative (to Gaussian) Brier scores (see Section 5.3) to compare each model.

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 1.13 and 1.18 indicating poorer performance than the Gaussian model. The plots in Figure 5 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. See Web Appendix H for the performance of the best two models for selected extreme quantiles.

We illustrate the predictive capability of our model in Figure 6 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

(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. Among these methods, the fourth method demonstrates the best performance. For a map of Brier scores for the 99th quantile between Gaussian and the fourth method, see Web Appendix I.

[Figure 5 about here.]

[Figure 6 about here.]

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.

8. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2.1, 2.2, 3.2, 3.3, 5.3, 6, and 6.2 are available with this paper at the Biometrics website on Wiley Online Library. R code implementing the spatial skew-t model for threshold exceedances is available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgments

The authors' work was partially supported by grants from the Department of the Interior (14-1-04-9), National Institutes of Health (R21ES022795-01A1), the US Environmental Protection Agency (R835228), the National Science Foundation (1107046), and the Research Network for Statistical Methods for Atmospheric and Oceanic Sciences (STATMOS). The authors also received high-performance computing support from Yellowstone (ark:/85065/d7wd3xhc) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.

References

Azzalini, A. and Capitanio, A. (2003). 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,** 367–389.

Azzalini, A. and Capitanio, A. (2014). *The Skew-Normal and Related Families*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83,** 715–726.

Beranger, B., Padoan, S. A., and Sisson, S. A. (2016). Models for extremal dependence derived from skew-symmetric families. *ArXiv e-prints* .

Branco, M. D. and Dey, D. K. (2001). A General Class of Multivariate Skew-Elliptical Distributions. *Journal of Multivariate Analysis* **79,** 99–113.

Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence Measures for Extreme Value Analyses. *Extremes* **2,** 339–365.

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.

Engelke, S., Malinowski, A., Kabluchko, Z., and Schlather, M. (2015). Estimation of Hüsler-Reiss distributions and Brown-Resnick processes. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **77,** 239–265.

Genton, M. G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Statistics (Chapman & Hall/CRC). Taylor & Francis.

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

Hollander, M., Wolfe, D. A., and Chicken, E. (2014). *Nonparametric Statistical Methods*. Wiley, 3rd edition.

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.

Kabluchko, Z., Schlather, M., and de Haan, L. (2009). Stationary max-stable fields associated to negative definite functions. *Annals of Probability* **37,** 2042–2065.

Kim, H. M. and Mallick, B. K. (2004). A Bayesian prediction using the skew Gaussian distribution. *Journal of Statistical Planning and Inference* **120,** 85–101.

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.

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

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.

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.

Ribatet, M. (2015). *SpatialExtremes: Modelling Spatial Extremes*. R package version 2.0-2.

Samet, J. M., Dominici, F., Zeger, S. L., Schwartz, J., and Dockery, D. W. (2000). The National Morbidity, Mortalityand Air Pollution Study Part I : Methods and Methodologic Issues. Technical Report 94.

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. (2015). Efficient inference and simulation for elliptical Pareto processes. *Biometrika* **102,** 855–870.

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

Wadsworth, J. L. and Tawn, J. A. (2014). Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika* **101,** 1–15.

Zhang, H. and El-Shaarawi, A. (2010). On spatial skewGaussian processes and applications. *Environmetrics* **21,** 33–47.

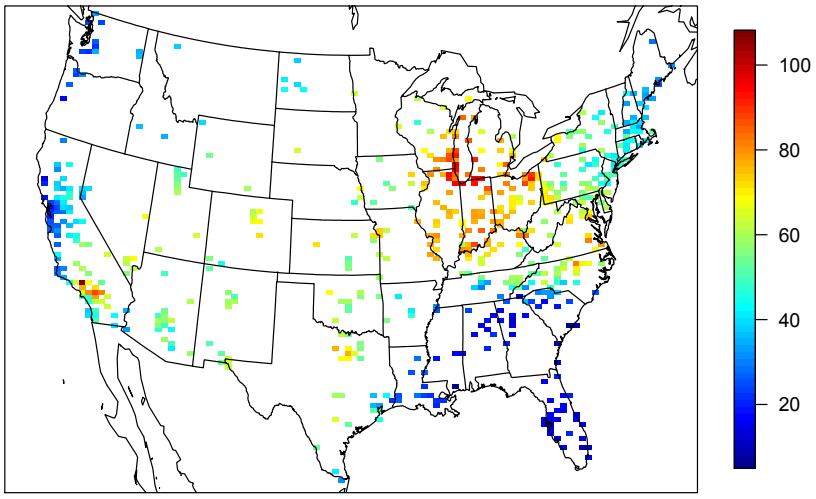

Figure 1. Ozone values (ppb) on July 10, 2005

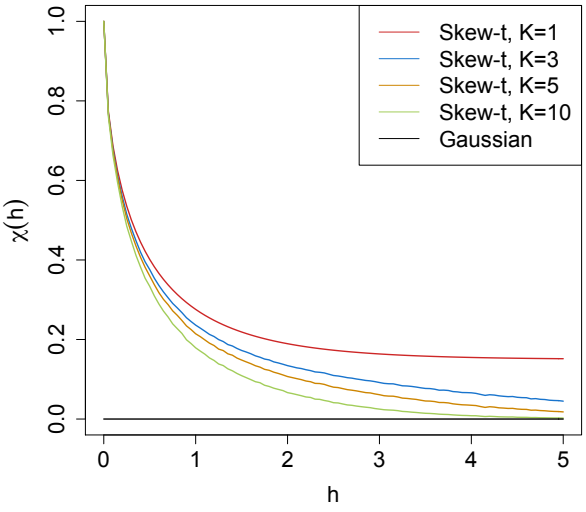

Figure 2. Extremal dependence measure $\chi(h)$, as a function of distance, h, for $K = 1, 3, 5$, and 10 knots.

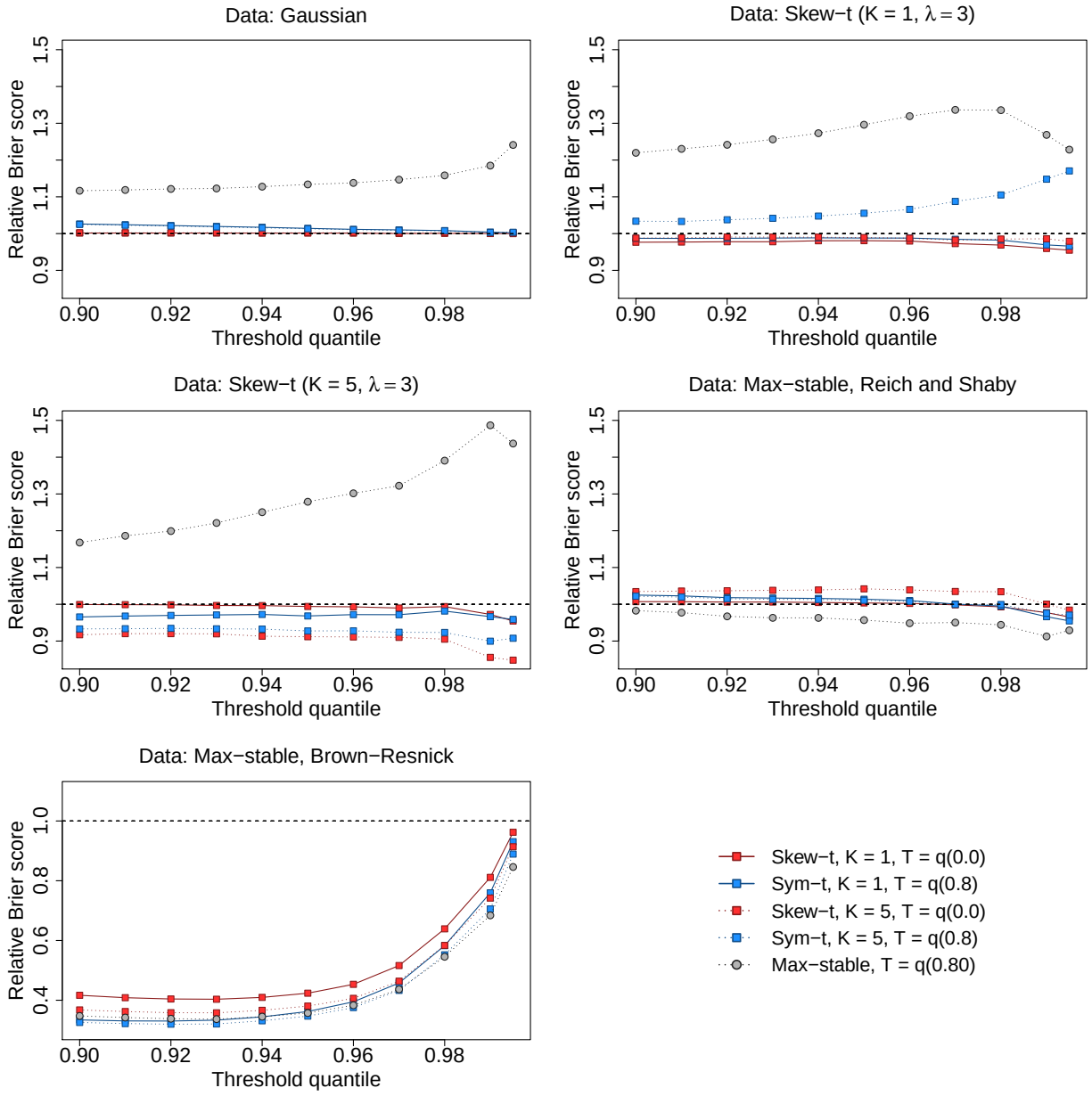

Figure 3. Brier scores relative to the Gaussian method for simulation study results. A ratio lower than 1 indicates that the method outperforms the Gaussian method.

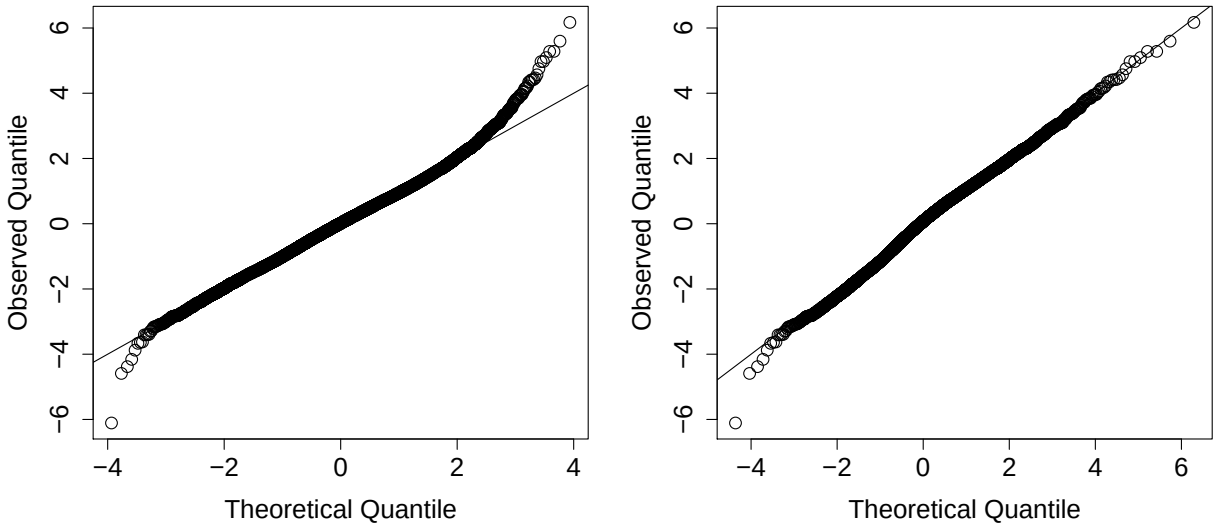

Figure 4. Gaussian Q-Q plot (left) and skew-t with $a = 10$ and $\lambda = 1$ Q-Q plot (right) of the residuals.

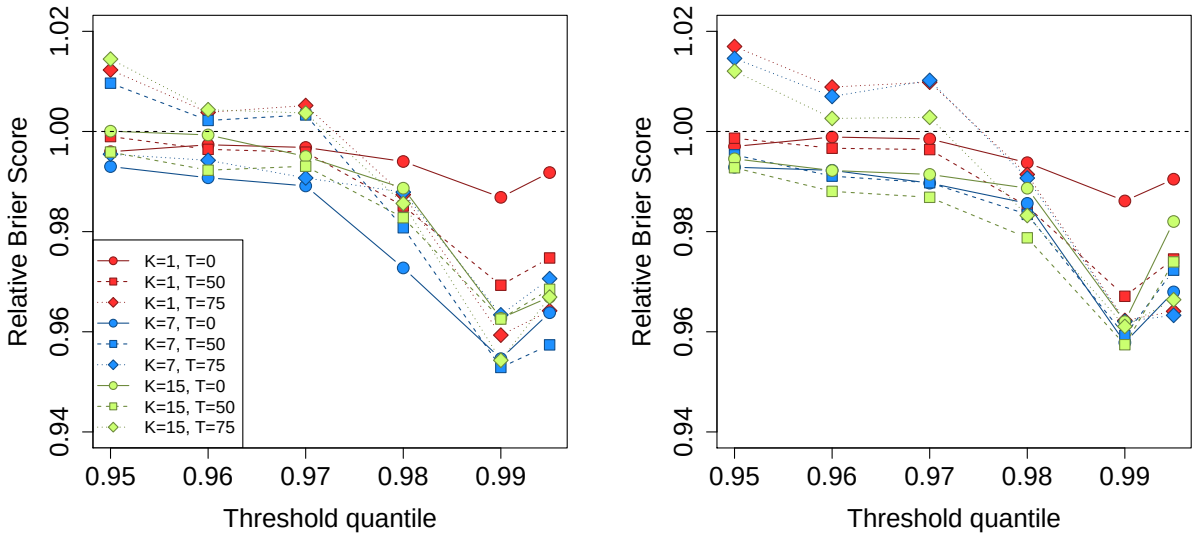

Figure 5. Relative Brier scores for time-series models (left) and non-time-series models (right). Relative brier score for the max-stable model is between 1.13 and 1.18

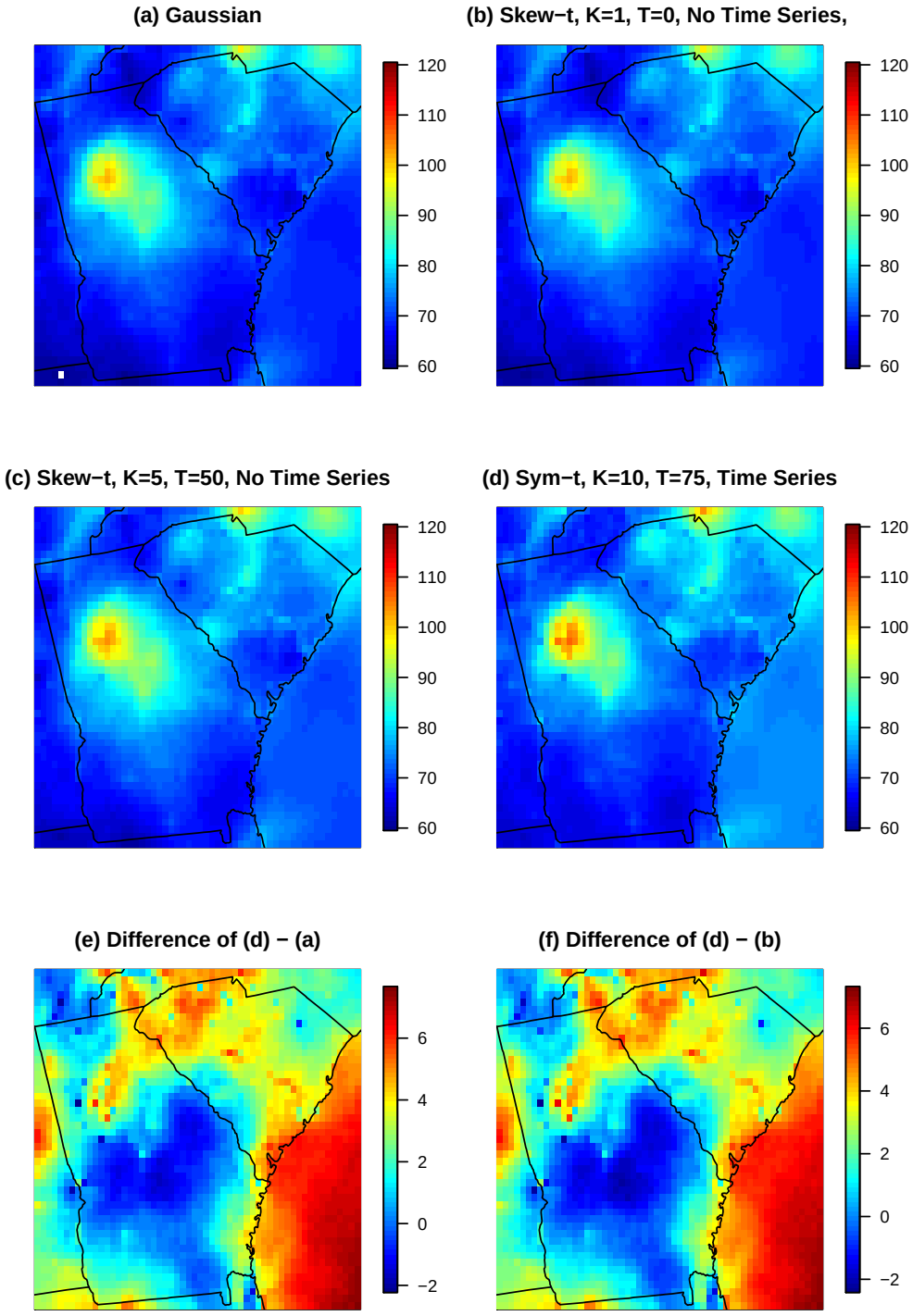

Figure 6. Panels (a) – (d) give the posterior predictive $\widehat{q}(0.99)$ for the month of July under four different models, panel (e) gives the difference between $\widehat{q}(0.99)$ in panels (d) and (a), panel (f) gives the difference between $\widehat{q}(0.99)$ in panels (d) and (b).