PCA for extremes

Sam Morris[1], Brian J Reich[1], Emeric Thibauld[2], and Dan Cooley[2]

June 7, 2016

Abstract

words...

Key words: Max-stable process.

[1]North Carolina State University
[2]Colorado State University

1 Introduction

2 Model

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

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

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

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

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

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

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

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

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

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

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

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

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

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

For the PS random effects model the EC has the form

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

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

3 Estimating the spatial dependence function

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

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

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

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

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

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

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

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

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

Let

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

where $w_{iu} = \exp[-(||\mathbf{s}_i - \mathbf{s}'_u||/\phi)^2])$ is the Gaussian kernel function with bandwidth ϕ. The elements $\hat{\vartheta}_{ii}$ do not contribute any information as $\hat{\vartheta}_{ii} = 1$ for all i by construction. To eliminate the influence of these estimates we set $w_{ii} = 0$. However, this approach does give imputed values $\tilde{\vartheta}_{ii}$, which provide information about small-scale spatial variability.

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

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

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

The relative contribution of each term can be measured by

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

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

4 Bayesian implementation details

For our data analysis in Section 5 we allow the GEV location and scale parameters, denoted $\mu_t(\mathbf{s})$ and scale $\sigma_t(\mathbf{s})$ respectively, to vary with space and time. The GEV shape parameter ξ is held constant over space and time because this parameter is notoriously difficult to estimate (ref). We place a Gaussian process prior on $\mu_t(\mathbf{s})$ and $\log[\sigma_t(\mathbf{s})]$, with

$$\mu_t(\mathbf{s}) \sim MVN\left[\mathbf{X}_{1t}(\mathbf{s})^\top \boldsymbol{\beta}_1,\ \sigma_{1t}^2 \boldsymbol{\Sigma}\right] \tag{12}$$

$$\log[\sigma_t(\mathbf{s})] \sim MVN\left[\mathbf{X}_{2t}(\mathbf{s})^\top \boldsymbol{\beta}_2,\ \sigma_{2t}^2 \boldsymbol{\Sigma}\right] \tag{13}$$

where $\mathbf{X}_{\cdot t}(\mathbf{s})$ is a p-vectors of covariates at site $\mathbf{s}$ for time t, $\boldsymbol{\beta}_\cdot^2$ is a p-vector of regression coefficients, $\sigma_{\cdot t}^2 \sim IG(a, b)$ is the variance for time t, and $\boldsymbol{\Sigma}$ is an exponential spatial correlation matrix obtained from $\rho(h) = \exp\{-\frac{h}{\phi}\}$ where $h = ||\mathbf{s}_1 - \mathbf{s}_2||$ is the Euclidean distance between sites $\mathbf{s}_1$ and $\mathbf{s}_2$. Collectively, let the marginal GEV parameters at location i and time t be $\Theta_{it} = \{\mu_{it}, \sigma_{it}, \xi\}$ where $\mu_{it} = \mu_t(\mathbf{s}_i)$ and $\sigma_{it} = \sigma_t(\mathbf{s}_i)$.

As shown in Reich and Shaby (2012), the uncensored responses $Y_t(\mathbf{s})$ are conditionally independent given the spatial random effects, with conditional distribution

$$Y_{it}|\theta_{it}, \Theta_{it} \overset{indep}{\sim} GEV(\mu_{it}^*, \sigma_{it}^*, \xi^*), \tag{14}$$

where $\mu_{it}^* = \mu_{it} + \frac{\sigma_{it}}{\xi}(\theta_{it}^\xi - 1)$, $\sigma_{it}^* = \alpha\sigma_{it}\theta_{it}^\xi$, and $\xi^* = \alpha\xi$. Therefore, the conditional likelihood con-

veniently factors across observations; marginalizing over the random effect θ_{it} induces extremal spatial dependence. To focus on the extreme values above the local threshold T_i, we use the censored likelihood

$$d(y;\theta_{it},\Theta_{it},T_i) = \begin{cases} F(y;\mu^*_{it},\sigma^*_{it},\xi^*) & y \leq T_i \\ f(y;\mu^*_{it},\sigma^*_{it},\xi^*) & y > T_i, \end{cases} \tag{15}$$

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

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

$$\begin{aligned} Y_{it}|\theta_{ij} &\overset{indep}{\sim} d(y;\theta_{it},\Theta_{it},T_i) \\ \theta_{it} &= \left(\sum_{l=1}^{L} \hat{B}_{il}^{1/\hat{\alpha}} A_{lt}\right)^{\hat{\alpha}} \quad \text{where} \quad A_{lt} \overset{iid}{\sim} PS(\hat{\alpha}) \\ \mu_{it} &= \mathbf{X}_{it}^\top \boldsymbol{\beta}_1 \quad \text{and} \quad \log(\sigma_{it}) = \mathbf{X}_{it}^\top \boldsymbol{\beta}_2. \end{aligned} \tag{16}$$

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

The first-stage estimate of the extremal coefficients has three tuning parameters: the quantile thresholds $q_1, ..., q_{n_q}$, the kernel bandwidth ϕ, and the number of terms L. In Section 5 we explore a few possibilities for ϕ and L and discuss sensitivity to these choices. The second-stage Bayesian analysis requires selecting

thresholds $T_i, ..., T_{n_s}$. For this we use spatially smoothed sample quantiles. That is, we set T_i to the 0.95 quantile of the Y_{it} and Y_{jt} for sites j with $||\mathbf{s}_i - \mathbf{s}_j|| < r$, where r is set to XXX?

5 Data analysis

In this section, we illustrate our method using both points above a threshold and block maxima. In Section 5.1, we present an analysis using annual acreage burned due to wildfires in Georgia from 1965 – 2014. This is followed in Section 5.4 by an analysis of precipitation data in the eastern U.S.

5.1 Analysis of extreme Georgia fires

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

$$\widetilde{B}_{il} = \exp\left\{-\frac{||\mathbf{s}_i - \mathbf{v}_l||^2}{\rho^2}\right\} \tag{17}$$

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

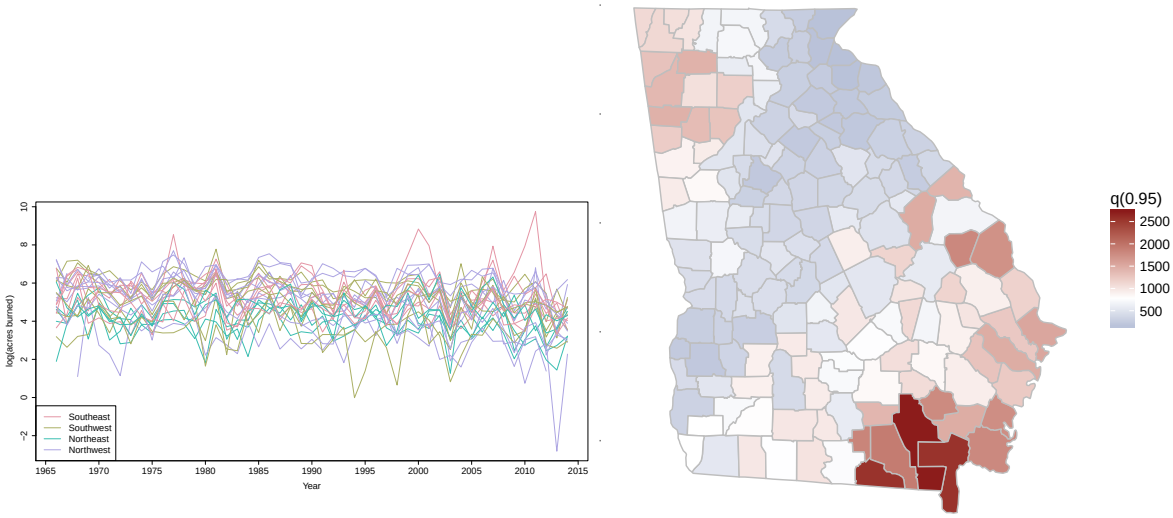

Figure 1: Time series of log acres burned for 25 randomly selected counties with colors coding the county's quadrant (left), and spatially smoothed threshold values, T_i for each county (right).

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

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

5.2 Results for fire analysis

We use 10-fold cross-validation to assess the predictive performance of a model. For each method, we randomly select 90% of the observations across counties and years to be used as a training set to fit the

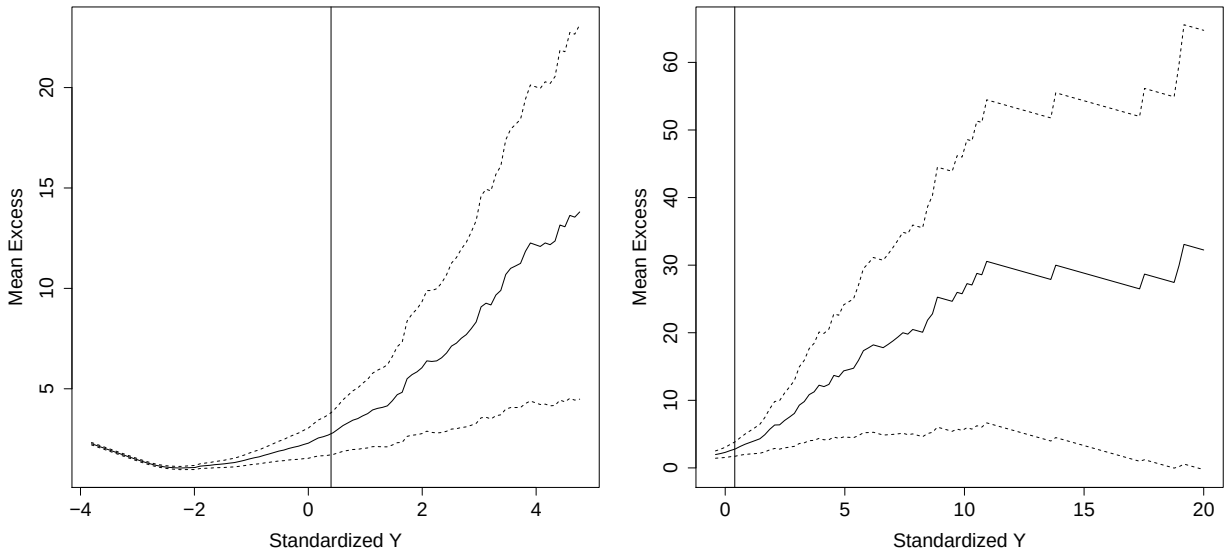

Figure 2: Mean residual plot for the data pooled across counties after standardizing using the county's median and interquartile range. The two panels show different ranges on the x-axis and include a vertical line at the sample 95th percentile.

model. The remaining 10% of sites and years are withheld for testing model predictions. To assess the predictions for the test set, we use quantile scores and Brier scores citation. The quantile score is given by give formula. The Brier score is given by give formula. For both of these methods, we use a negative orientation, so a lower score indicates a better fit. The Brier and quantile scores for the fire analysis are given in Table 1.

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

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

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

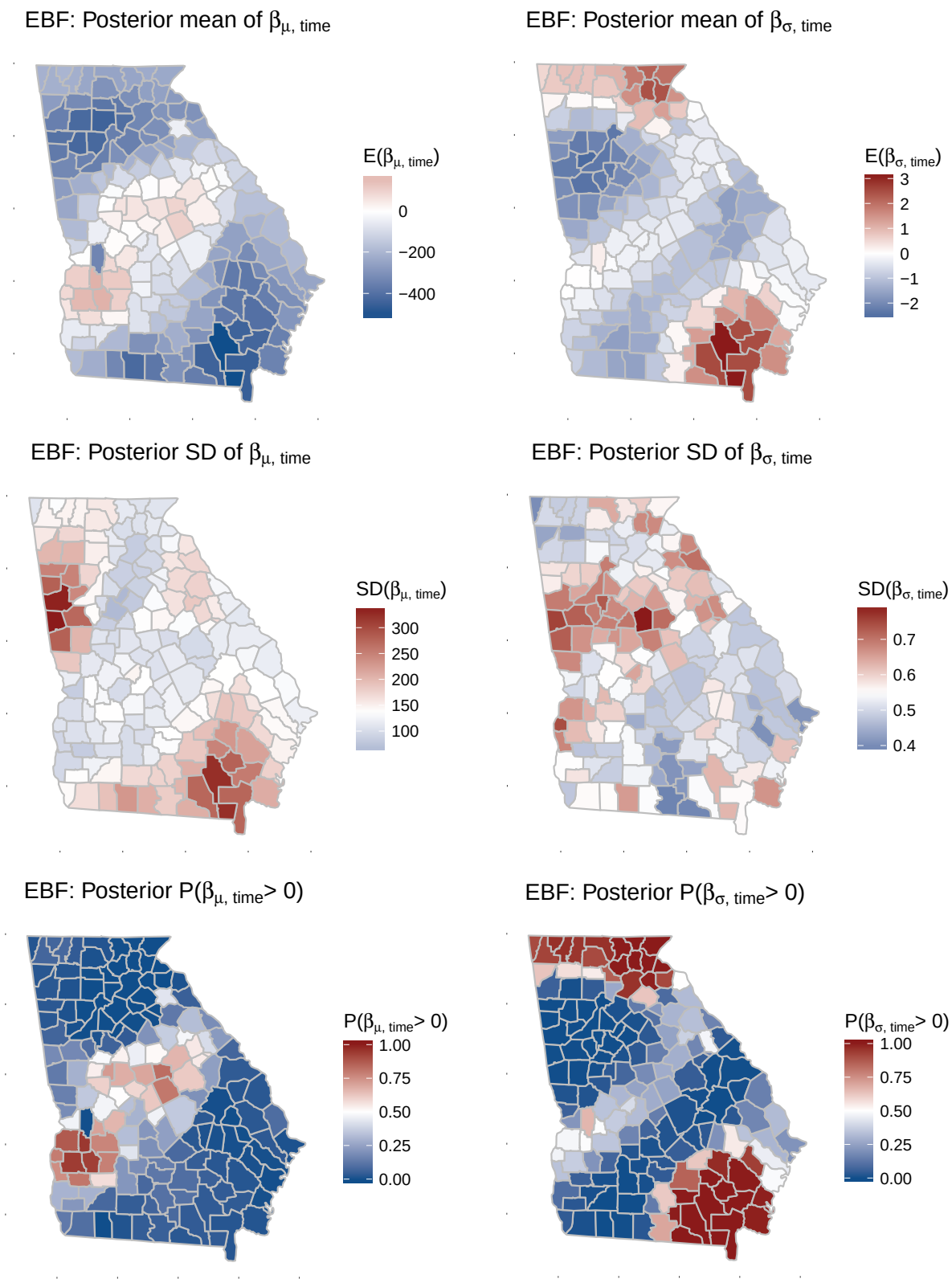

Figure 3: Posterior summaries of β_{time} when using EBF for the spatial process with $L = 35$.

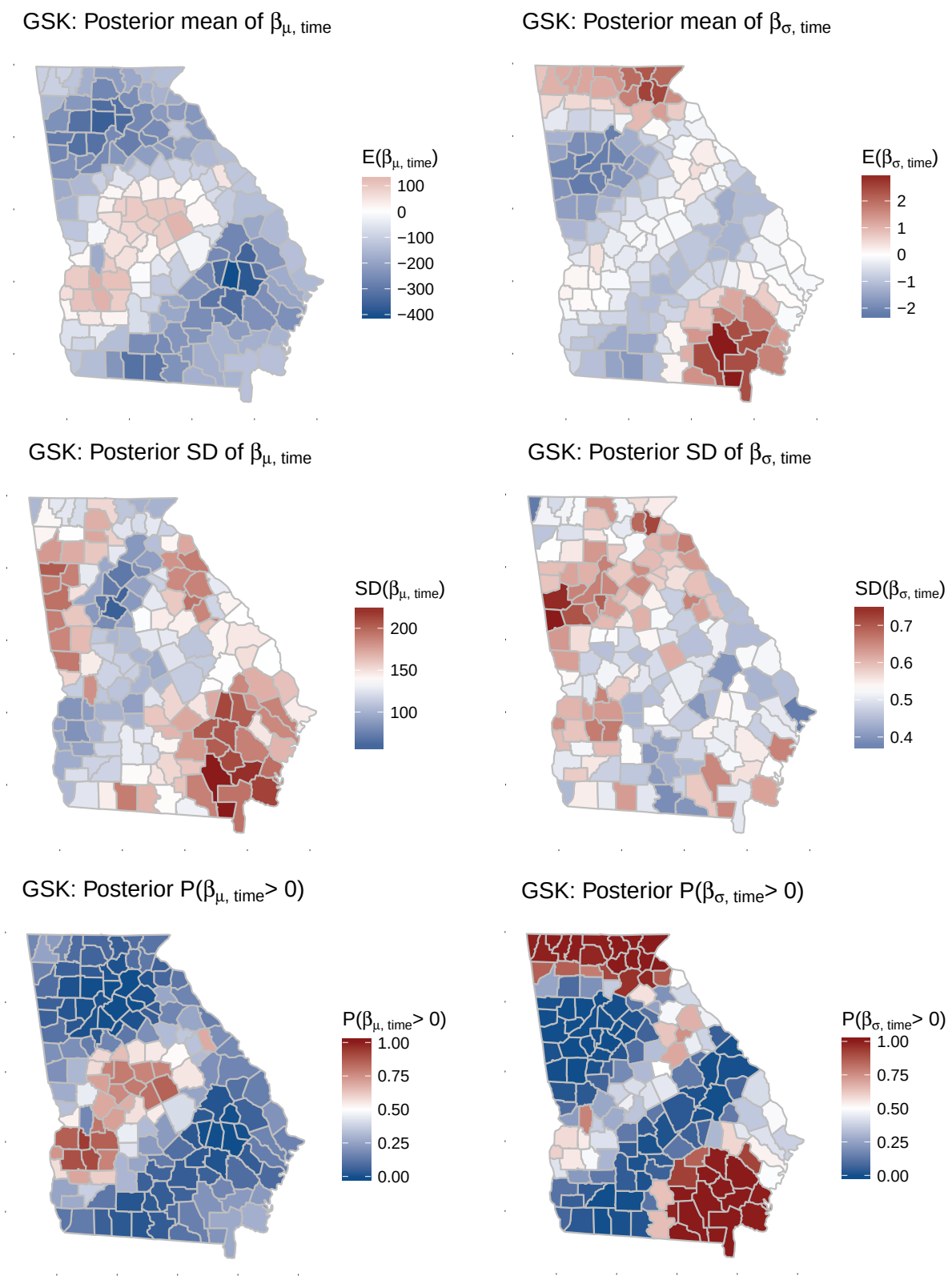

Figure 4: Posterior summaries of β_{time} when using GSK for the spatial process with $L = 35$.

Table 1: Average Brier scores ($\times 100$) for selected thresholds and quantile scores for selected quantiles for fire analysis Need timings for L = 35, 40 to be added in

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

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

5.3 Model checking and sensitivity analysis

5.4 Analysis of annual precipitation

We also conduct an analysis of the precipitation data presented in (Reich and Shaby, 2012). The data are climate model output from the North American Regional Climate Change Assessment Program (NARCCAP). This data consists of $n_s = 697$ grid cells at a 50km resolution in the eastern US, and includes historical data (1969 – 2000) as well as future conditions (2039 – 2070).

Include figures of locations of grid cells

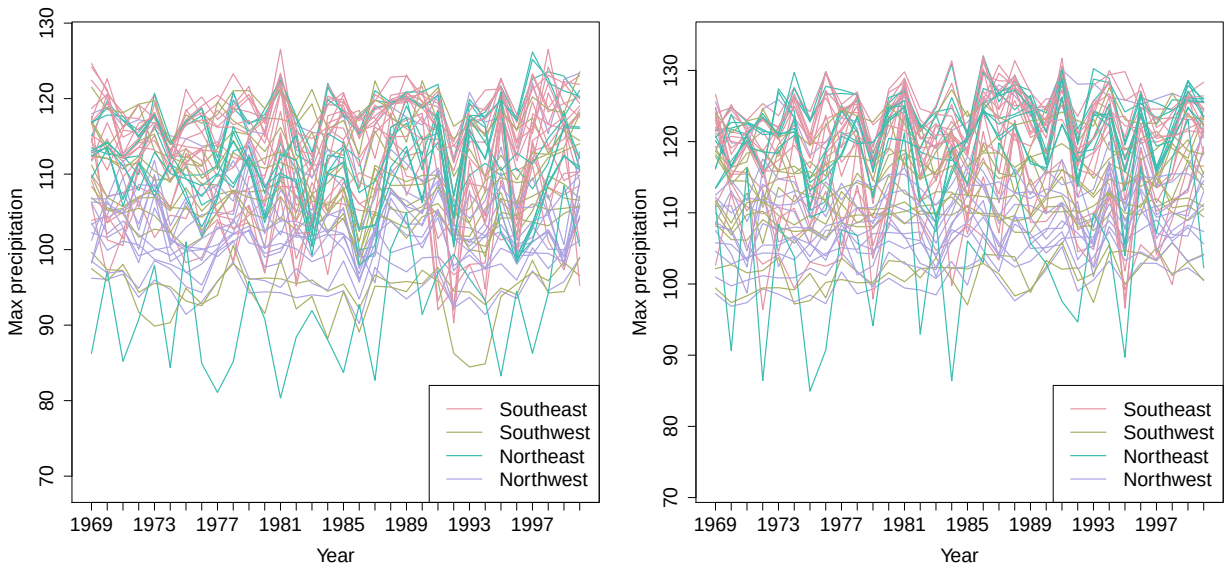

Figure 5: Time series of yearly max precipitation for current (1969 – 2000) (left). Time series of yearly max precipitation for future (2039 – 2070) (right).

5.5 Results for precipitation analysis

6 Conclusions

Acknowledgements

A.1 Extreme value distributions

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

A.2 Gradient for β

References

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.

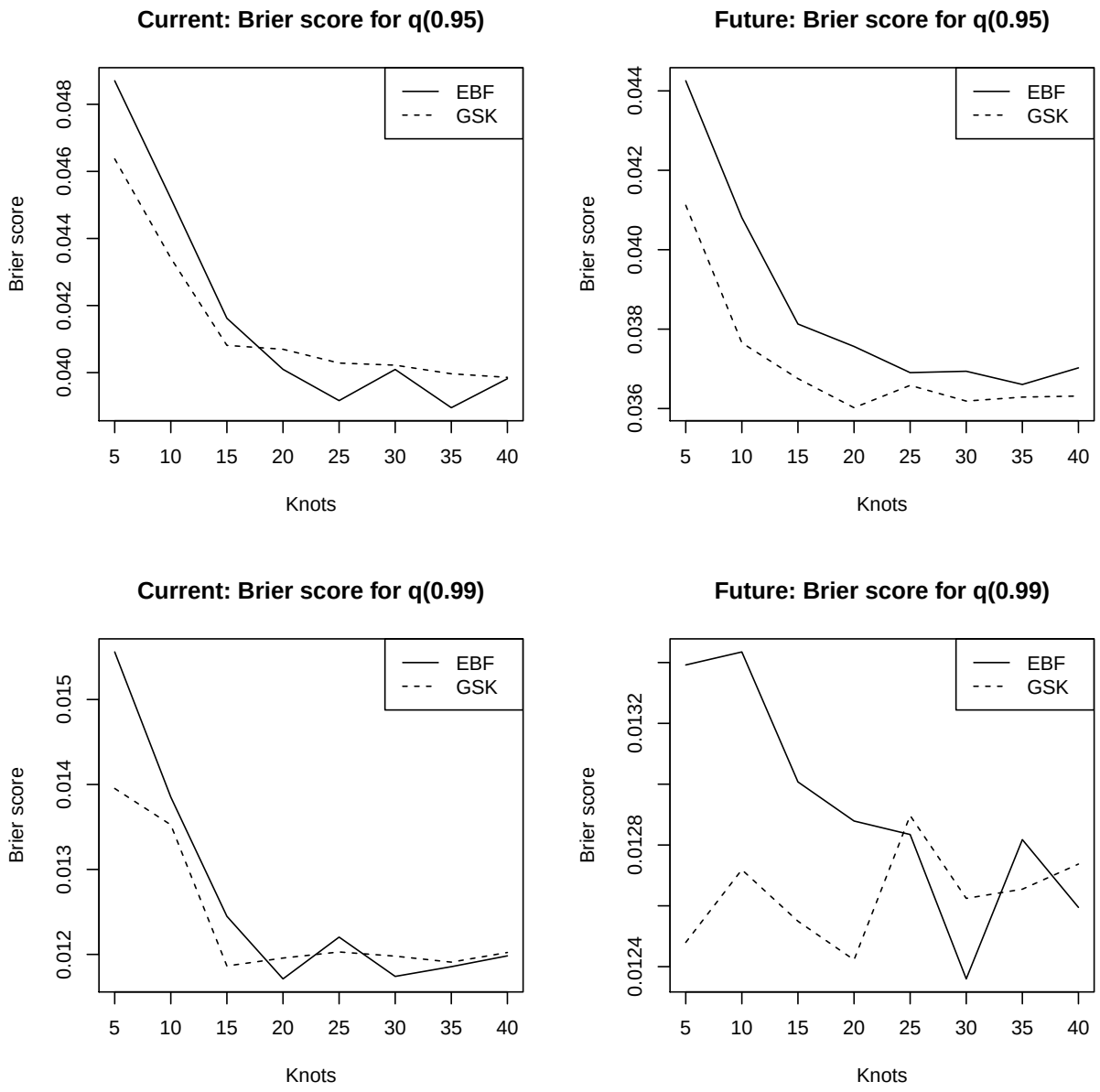

Figure 6: Brier scores for current and future precipitation analysis.

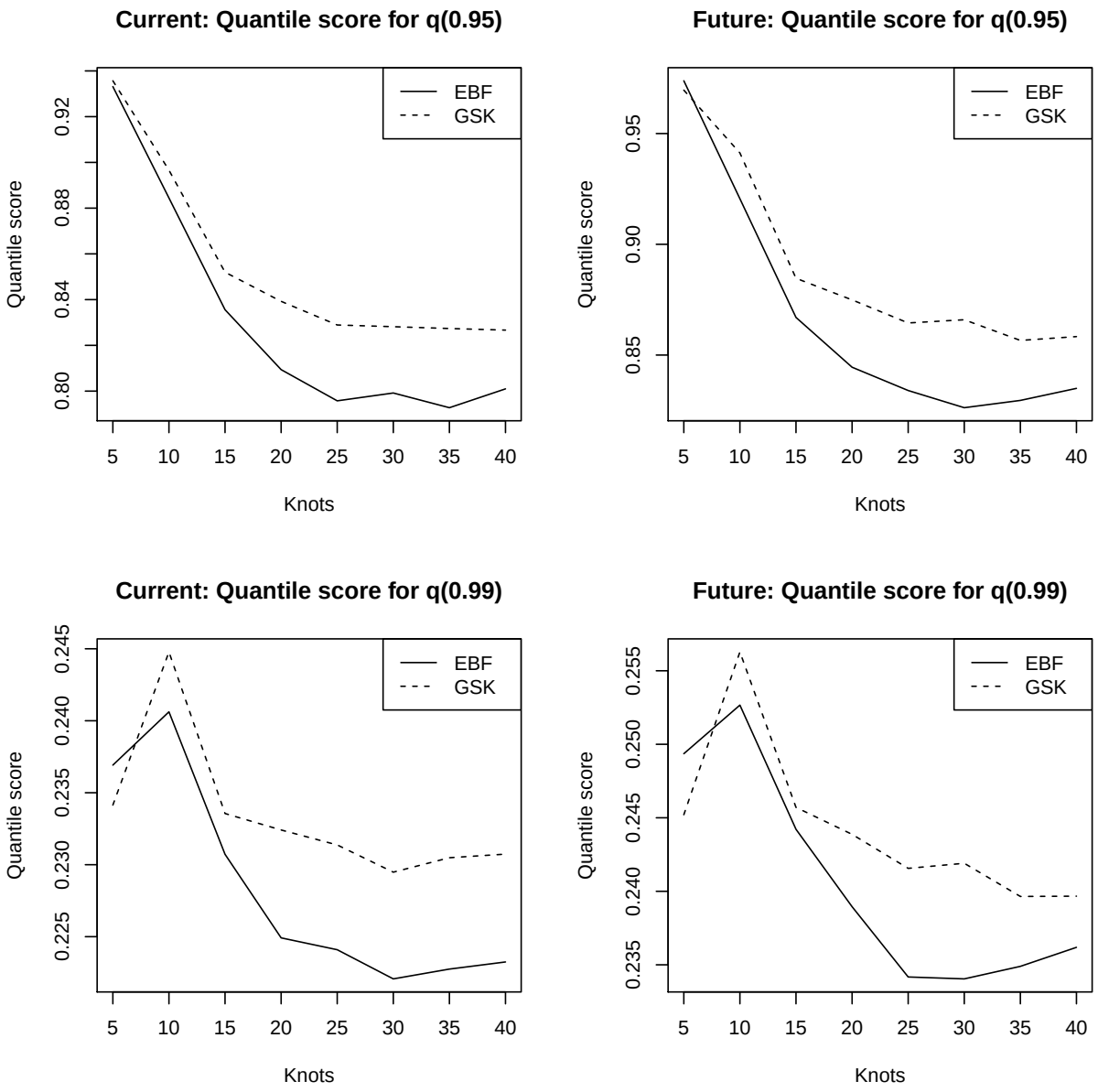

Figure 7: Quantile scores for current and future precipitation analysis.

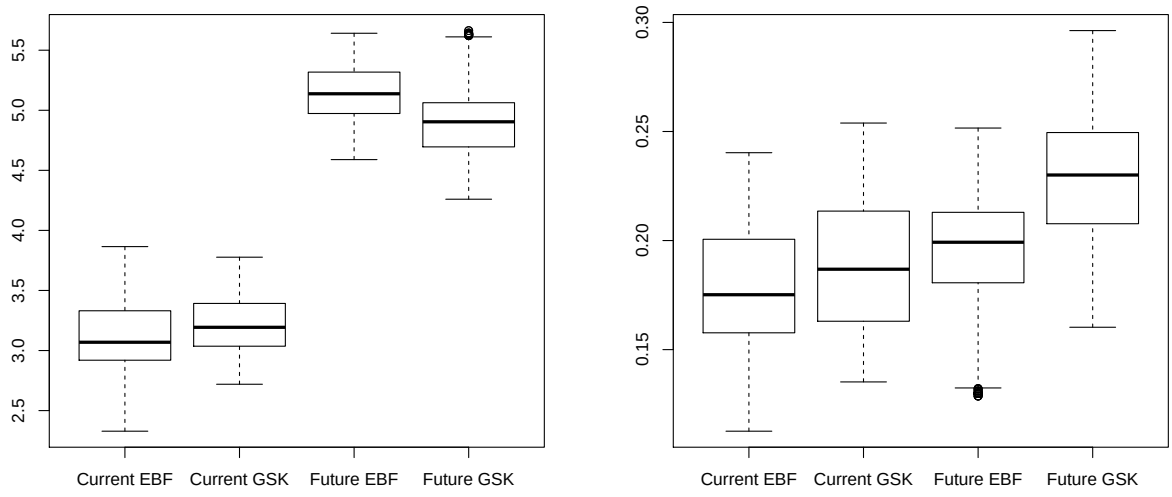

Figure 8: Posterior distributions for β_{time} for μ (left) and $\log(\sigma)$ (right).