PCA for extremes

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

April 27, 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 4. Although the model presented implements a censored likelihood, it should be noted that the model also can fit max-stable data 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 μ_{it} and scale σ_{it} respectively, to vary with space and time. The GEV shape parameter ξ is held constant over space and time because this parameter is notoriously difficult to estimate (ref). Collectively, let the marginal GEV parameters at location i and time t be $\Theta_{it} = \{\mu_{it}, \sigma_{it}, \xi\}$. The GEV location and scale vary according to covariates $\mathbf{X}_{it}$ with $\mu_{it} = \mathbf{X}_{it}^T \boldsymbol{\beta}_1$ and $\log(\sigma_{it}) = \mathbf{X}_{it}^T \boldsymbol{\beta}_2$.

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

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

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

$$d(y; \theta_{it}, \Theta_{it}) = \begin{cases} F(y; \mu_{it}^*, \sigma_{it}^*, \xi^*) & y \leq T_i \\ f(y; \mu_{it}^*, \sigma_{it}^*, \xi^*) & y > T_i, \end{cases} \tag{13}$$

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

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

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

To complete the Bayesian model, we select independent normal priors with mean zero and variance σ_1^2 and σ_2^2 for the components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ respectively, and $\xi \sim \text{Normal}(0, 0.5^2)$. We use inverse gamma (1, 1) priors for σ_1^2 and σ_2^2. 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. We set $\{q_1, ..., q_{n_q}\} = \{0.95, 0.96, ..., 0.99\}$. 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?

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, σ_2^2 which we update using Gibbs sampling. The positive stable density can be challenging to evaluate due to the integral inside the density function. One technique to avoid this complication is to incorporate auxiliary i.i.d. Uniform(0, 1) random variables (Stephenson et al, ref). We opt for a numerical approximation to the integral using 51 evenly spaced quantiles for a Beta(0.5, 0.5) distribution.

5 Data analysis

In this section, we illustrate our method using both censored and max-stable datasets. 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. These applications are chosen because they demonstrate the suitability of our model for both censored and max-stable data.

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.stat.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: $\mathbf{X}_{it} = (1, t^*, \widetilde{B}_{i1}, ..., \widetilde{B}_{iL}, t^*\widetilde{B}_{i1}, ..., t^*\widetilde{B}_{iL})^T$. For the bivariate Gaussian kernels, we select L knot locations, $\mathbf{v}_1, \ldots, \mathbf{v}_L$ from the county centroids, using the space-filling design by (ref add to bibtex Johnson, M.E., Moore, L.M., and Ylvisaker, D. ,1990). Then

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

where ρ is included as a parameter in the MCMC. We also tried using the estimated basis functions, $\hat{B}_1, \ldots, \hat{B}_L$ as covariates, but they did not perform as well as the bivariate Gaussian kernels. Note: I decided not to standardize the basis functions because it allows for an intercept which is easier to come up with a starting value for.

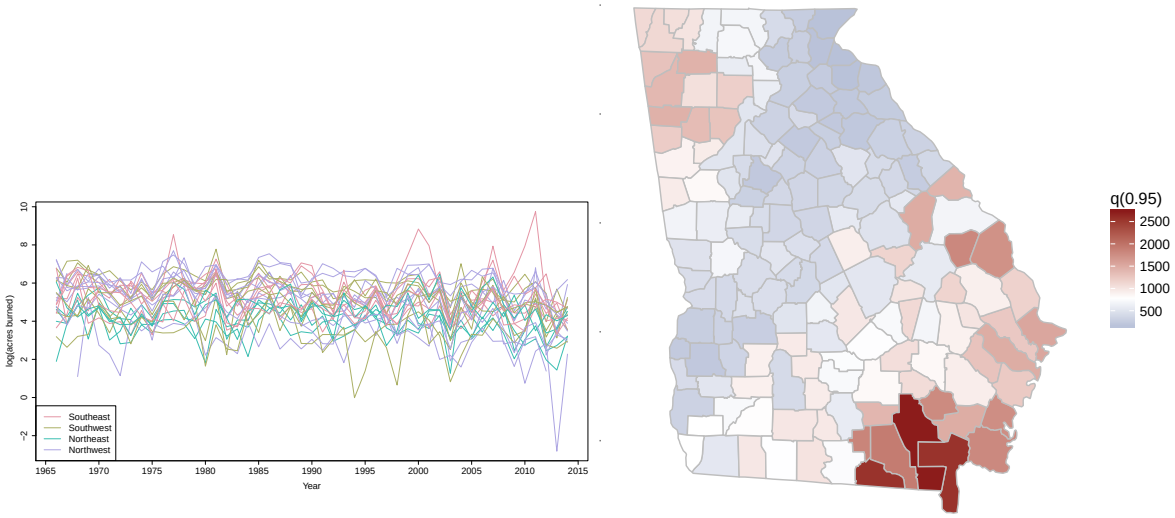

Figure 1: Time series of log acres burned for 25 randomly selected counties (left). need to adjust time series plot Spatially smoothed threshold values for each county (right).

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 max-stable, 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 compute

$$\widetilde{\mathbf{Y}}_i = \frac{\mathbf{Y}_i - \text{med}(\mathbf{Y}_i)}{\text{IQR}(\mathbf{Y}_i)} \tag{16}$$

where $\text{med}(\cdot)$ is the median, and $\text{IQR}(\cdot)$ is the inter-quartile range. Then we combine all sites together and plot a mean residual plot for $\widetilde{\mathbf{Y}}_i, i = 1, \ldots, n_s$. The mean residual plot is given in Figure 2, with a vertical line indicating the quantile we use for the county-specific values $\mathbf{T}$. Based upon the mean residual plot, we select $q(0.95)$ for the spatially smoothed threshold. To calculate T_i for each county, we find $\hat{q}(0.95)$ by taking the $95th$ quantile for county i and the five closest counties.

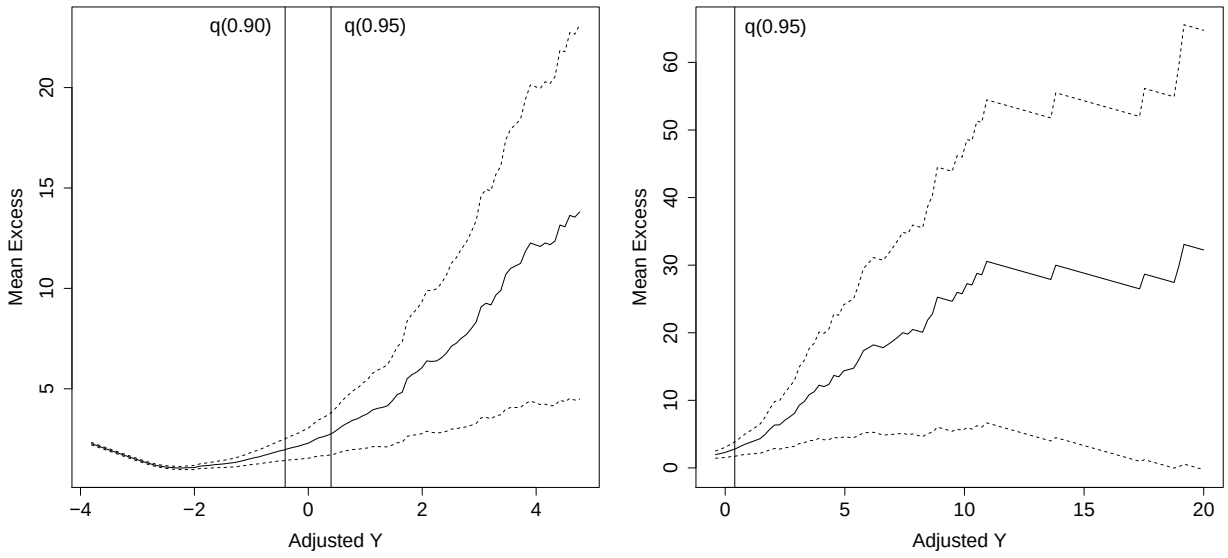

Figure 2: Mean residual plot from min(Adj Y) to q(0.995) (left). Mean residual plot from q(0.90) to 20 (right). Vertical lines for q(0.90) and q(0.95) are given in the plots.

5.2 Results

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

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

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

Table 1: Average quantile scores for selected quantiles and Brier scores ($\times 100$) for selected thresholds Need update with L = 30 and 159, timings 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	5.709	2.268	131.92	79.525	
	gsk	5.808	2.303	133.37	79.985	
L = 10	ebf	5.066	2.101	130.21	72.486	
	gsk	5.192	2.126	130.67	75.501	
L = 15	ebf	4.947	2.041	130.95	64.586	
	gsk	5.078	2.074	130.79	69.610	
L = 20	ebf	4.989	2.035	122.64	66.102	
	gsk	4.942	2.030	121.71	65.267	
L = 25	ebf	4.911	**1.999**	**119.71**	63.302	
	gsk	**4.842**	2.017	119.94	**62.192**	
L = 30	ebf					
	gsk					
L = 159	gsk					

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

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.

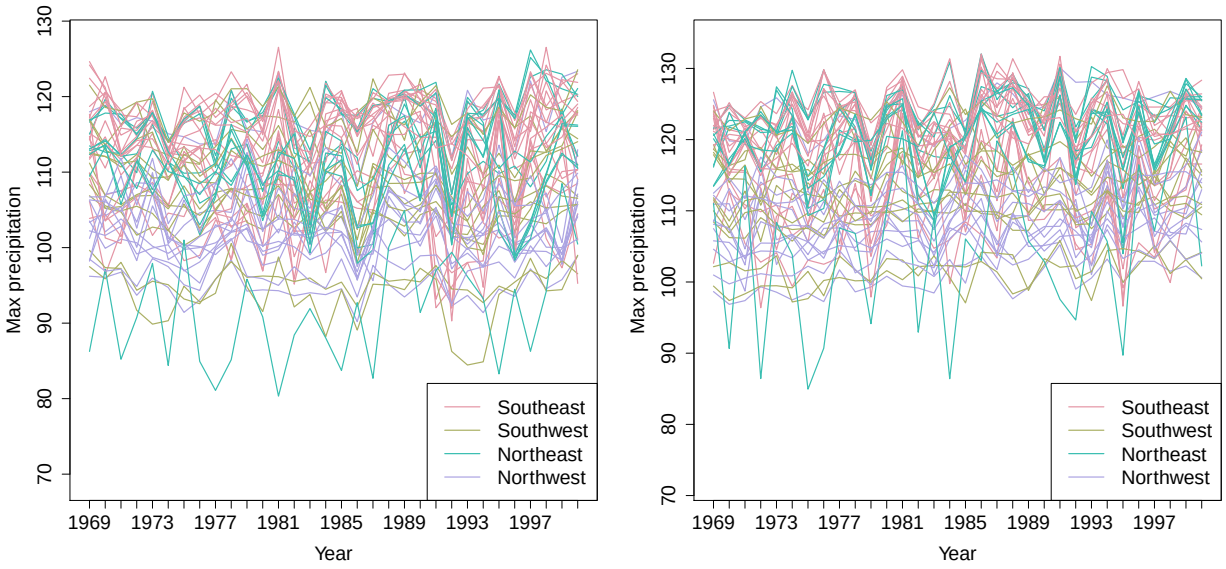

Figure 3: Time series of yearly max precipitation for current (1969 – 2000) (left). Time series of yearly max precipitation for future (2039 – 2070) (right).

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.