A spatial Markov model for climate extremes

Brian Reich (NCSU) and Ben Shaby (PSU)

April 20, 2017
Importance of extremes in climate research

- From heat waves to hurricanes, often the environmental processes that are the most critical to understand probabilistically are extreme events.

- There is a large literature on EVA.

- There are some beautiful mathematical results.

- The statistical methodology is unique.

- There are many open statistical problems.
Common objectives in EVA

▶ Estimate the 1,000 year return level, i.e., the value that occurs on average once every 1,000 years

▶ Identify environmental covariates that drive extremes

▶ Test the hypothesis that the likelihood of an extreme event is changing over time

▶ Determine if two locations are asymptotically dependent

▶ Project the change in the 99.9th percentile in 2050
Unique statistical challenges

▶ Most of our intuition and methodology are built around the mean and deviation from the mean

▶ In EVA concepts like mean and variance are irrelevant because they don’t speak to the tail of the distribution

▶ Similarly, correlation isn’t the best measure of dependence because it is based on deviation around the means

▶ We need new ways to describe distributions and dependence between random variables
Isolating the extremes

- The first step in classic EVA is to separate extreme observations from the bulk of the distribution.

- For example, in a daily time series of precipitation in FL these correspond to very different weather regimes.
  - Bulk: Thunderstorms
  - Tails: Hurricanes

- Mean regression focus on thunderstorms and treats hurricanes as outliers.

- If you want to estimate the 100-year storm, you should focus only on hurricanes.

- Two common ways to isolate the extremes: block maxima and points above a threshold.
Block maxima (BM) in Cheeseboro

*Block* is a year and the block maximum is annual maximum.
Points above a threshold (POT) for Cheeseboro

The *threshold* is 50 and we analyze the points in red
Pros/cons of analyzing BM

Pros:
► Can evoke EVA theory and use a simple model
► It removes dependence with block

Cons:
► Excludes some large values (second highest each year)
► Must pick the block size: too big and you lose data; too small and you can’t use EVA theory
Pros/cons of analyzing POT

Pros:
▶ Can evoke EVA theory and use a simple model
▶ Retains all large values in the analysis

Cons:
▶ Must deal with dependence within block
▶ Setting the threshold is really difficult
Let $Y_1, \ldots, Y_n$ be the $n$ independent and identically distributed values in a block (say $n = 365$ days in a year).

In certain conditions, for large $n$ the sample (annual) mean

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

is approximately normally distributed.

Holds with some forms of dependence and nonstationarity.

The underlying data $Y_i$ do not have to be Gaussian.

For example, the mean of 10 uniforms is $\approx$ normal.

So if you are analyzing data that are constructed as means, then a normal distribution is a good start.

You should still check the assumption of normality.
A similar result holds for the block maximum

\[ \tilde{Y}_n = \max\{ Y_1, \ldots, Y_n \} \]

Under certain conditions, for large \( n \), \( \tilde{Y}_n \) approximately follows the Generalized Extreme Value (GEV) distribution.

Holds with some forms of dependence and nonstationarity.

So if you are analyzing data that are constructed as say annual maximums, then GEV is a good start.

You should still of course check the GEV fit.
The GEV distribution

- Location: $\mu$
- Scale: $\sigma > 0$
- Shape: $\xi$

The shape defines three special cases:
- Weibull: $\xi < 0$ and the distribution is bounded above
- Gumbel: $\xi = 0$ and the distribution is unbounded
- Frechet: $\xi > 0$ and the distribution is bounded below
Spatial extremes

- EVA can benefit greatly from spatial methods

- Spatial methods can map risk and borrow strength over space to estimate rare-event probabilities

- Accounting for spatial dependence is necessary for valid inference

- Methods and software in this area are developing rapidly to meet a growing demand
Gaussian data: Geostatistical vs areal models

- **Geostats**: we sample $n$ of an uncountable number of potential spatial locations
  - Example: precipitation monitors
  - Common methods: Matern correlation, Kriging, etc.

- **Areal**: the entire domain is partitioned into $n$ regions
  - Example: county-level disease rates
  - Common methods: Conditionally autoregressive (CAR) model
Gaussian data: Geostatistical vs areal models

- Compared to geostat models, CAR models have computational advantages because they are defined locally.
- In extremes, max-stable processes are the analogy of geostat models.
- Max-stable processes are far more complex (conceptually and computationally) than Geostat models.
- Climate data are often areal.
- To our knowledge there is currently no analogy of CAR models for extremes.
Example 1 – Forest fires in GA
Example 2 – Climate model precip output
Areal data model - Marginal distributions

- $Y_{it}$ is the annual maximum in region $i$ and year $t$

- The marginals are GEV with spatiotemporal parameters

- Location: $\mu_{it} = \sum_{l=1}^{L} X_{lt} \beta_{il}$, where
  - $X_{lt}$ are known B-spline basis functions of time
  - $\beta_{il}$ are unknown basis coefficients

- Scale: $\log(\sigma_{it}) = \beta_{i0}$

- Shape: $\xi_{it} = \xi$
Areal data model - Marginal distributions

- GEV parameters $\beta_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{iL})^T$ have MCAR prior

- $\beta_i | \beta_j, j \neq i \sim \text{Normal} \left( \gamma + \rho (\bar{\beta}_i - \gamma), \frac{1}{m_i} \Sigma \right)$
  - $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_L)^T$ is the mean vector
  - $\bar{\beta}_i$ is the mean of region $i$'s neighbors
  - $\rho \in (0, 1)$ controls the strength of spatial dependence
  - $\Sigma$ is an $L + 1 \times L + 1$ covariance matrix
Residual dependence

- To produce reliable Bayesian inference we must account for residual dependence

- Let $Z_{it} = \left[1 + \xi (Y_{it} - \mu_{it}) / \sigma_i \right]^{1/\xi}$

- The residuals have unit Frechet distribution $Z_{it} \sim GEV(1, 1, 1)$

- We will specify a spatial Markov model for $Z_{it}$ as a function of neighboring $Z_{jt}$
Residual dependence

- For notational simplicity, we temporarily omit the temporal subscript, $Z_{it} \rightarrow Z_i$.
- Spatial dependence is introduced via a random partition of the $n$ regions.
- The adjacency structure of the regions determines the partition probabilities.
- The random partition model captures spatial clusters of extreme events.
- Within these clusters, we assume a multivariate GEV (MGEV) distribution to model extremal dependence.
Residual dependence

- Assume the regions are partitioned into $K$ clusters
- $g_i = k$ indicates that region $i$ is allocated to cluster $k$
- Observations in different clusters are independent
- Observations within a cluster follow an exchangeable MGEV distribution
- We use the symmetric logistic extremal measure with dependence parameter $\alpha \in (0, 1)$
- Small $\alpha$ gives dependence, $\alpha = 1$ gives independence
Residual dependence

- The cluster labels follow the spatial Potts model

\[ p(g_1, \ldots, g_n | \phi) \propto \exp \left( \sum_{i \sim j} \phi I(g_i = g_j) \right) \]

- \( \phi > 0 \) determines the strength of spatial dependence

- This leads to a Markov model for the labels

\[ \text{Prob}(g_i = k | g_j, j \neq i) \propto \exp \left[ \phi \sum_{j \sim i} I(g_j = k) \right] \]

- The log odds of label \( k \) increases by \( \phi \) for each neighbor in cluster \( k \)
The following slides show realizations of the process with $\mu_i = 0$, $\sigma_i = 1$ and $\xi_i = 0.1$.

We plot both the labels, $g_i$, and the responses, $Y_i$.

The spatial dependence parameter $\phi$ ranges from small to large.

The MGEV parameter $\alpha$ ranges from small (dependence) to large (independence).
Small $\phi$, small $\alpha$
Small $\phi$, small $\alpha$
Medium $\phi$, small $\alpha$

Cluster label, $g$

Reich + Shaby
A spatial Markov model for climate extremes
27/61
Medium $\phi$, small $\alpha$

Response, Y

A spatial Markov model for climate extremes
Large $\phi$, small $\alpha$
Large $\phi$, small $\alpha$
Large $\phi$, large $\alpha$
Large $\phi$, large $\alpha$
Asymptotic properties

- Asymptotic dependence is often measured by
  \[ \chi_{ij} = \lim_{z \to \infty} \text{Prob} (Z_i > z | Z_j > z) \]

- \(Z_i\) and \(Z_j\) are asymptotically independent if \(\chi_{ij} = 0\)

- For the Potts/MGEV model, \(\chi_{ij} = \pi_{ij}(\phi)(2 - 2^\alpha)\)

- \(\pi_{ij}(\phi) = \text{Prob}(g_i = g_j) \in (1/K, 1)\) from the Potts model
Asymptotic properties

- The following slides plot $\chi_{ij}$ for a linear one-dimensional grid of $n = 50$ locations with first-order neighbors.

- The plots give $\pi_{25,k}(\phi) = \text{Prob}(g_{25} = g_k)$ and extremal dependence measure $\chi_{25,k}$.

- The Potts probabilities do not have a closed form.

- We use Monte Carlo simulation.
Asymptotic properties - $\pi_{25,k}$

Reich + Shaby
A spatial Markov model for climate extremes
Asymptotic properties - $\chi_{25,k}$

\[\chi(25, k)\]

- $\phi = 5$, $K = 50$, $\alpha = 0.5$
- $\phi = 5$, $K = 50$, $\alpha = 0.1$
- $\phi = 3$, $K = 10$, $\alpha = 0.5$
- $\phi = 3$, $K = 50$, $\alpha = 0.5$
Computation - overview

- We primarily use MCMC

- The Potts parameter is hard to estimate

- After a reparameterization, the likelihood factors across sites

- All MCMC updates are local and fast
Computation - Estimating the Potts parameter

- Updating $\phi$ is difficult because the Potts distribution has an intractible normalizing constant

- We use a plug-in estimator

- Recall $\chi_{ij} = \pi_{ij}(\phi)[2 - 2^\alpha]$

- We fix $\phi$ at the value that maximizes correlation between $\pi_{ij}(\phi)$ and empirical estimates of $\chi_{ij}$
Simulation study

A spatial Markov model for climate extremes
Computation - random effects

The MGEV distribution with symmetric logistic dependence has a mixture representation

Let \( A_k \sim \text{generalized inverse gamma} \) be a random effect for cluster \( k \)

Then for sites in cluster \( k \),

\[
Z_i \mid g_i = k \sim \text{GEV}(A_k^\alpha, \alpha A_k^\alpha, \alpha)
\]

independent over \( i \)

The likelihood now factors across all sites
Computation - sketch of MCMC

- $\beta_i$: independence sampler from prior
- $g_{kt}$: Gibbs
- $A_{kt}$: Gibbs/Metropolis
- The rest is straightforward Gibbs
CLIMDEX data

- The CLIMDEX/GHCNDEX data repository contains a suite of gridded climate indices.

- Each index is calculated annually over the period from 1950–2015.

- Data are provided on the 2.5 x 2.5 degree grid of $n = 509$ locations.

- We study eight indices.
# CLIMDEX data

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TXx</td>
<td>Annual maximum of daily max temp (°C)</td>
</tr>
<tr>
<td>TNx</td>
<td>Annual maximum of daily min temp (°C)</td>
</tr>
<tr>
<td>TXn</td>
<td>Annual minimum of daily max temp (°C)</td>
</tr>
<tr>
<td>TNn</td>
<td>Annual minimum of daily min temp (°C)</td>
</tr>
<tr>
<td>Rx1day</td>
<td>Annual maximum of daily precip (mm)</td>
</tr>
<tr>
<td>Rx5day</td>
<td>Annual maximum of 5-day average precip</td>
</tr>
<tr>
<td>CDD</td>
<td>Maximum length of dry spell</td>
</tr>
<tr>
<td>CWD</td>
<td>Maximum length of wet spell</td>
</tr>
</tbody>
</table>
CLIMDEX data - number of indicies per site
We take $K = n$ possible clusters

All other priors are uninformative

We compare $L = 4, 6, 8$ temporal basis functions using CV
Cross validation MAD (of standardized data)

<table>
<thead>
<tr>
<th>$L$</th>
<th>TXx</th>
<th>TNx</th>
<th>TXn</th>
<th>TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.081</td>
<td>0.083</td>
<td>0.030</td>
<td>0.033</td>
</tr>
<tr>
<td>6</td>
<td>0.080</td>
<td>0.081</td>
<td>0.029</td>
<td>0.033</td>
</tr>
<tr>
<td>8</td>
<td>0.079</td>
<td>0.081</td>
<td>0.028</td>
<td>0.032</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$L$</th>
<th>Rx1day</th>
<th>Rx5day</th>
<th>CDD</th>
<th>CWD</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.259</td>
<td>0.281</td>
<td>0.272</td>
<td>0.414</td>
</tr>
<tr>
<td>6</td>
<td>0.257</td>
<td>0.281</td>
<td>0.271</td>
<td>0.413</td>
</tr>
<tr>
<td>8</td>
<td>0.259</td>
<td>0.280</td>
<td>0.270</td>
<td>0.408</td>
</tr>
</tbody>
</table>
### Coverage of 80% intervals

<table>
<thead>
<tr>
<th>$L$</th>
<th>TXx</th>
<th>TNx</th>
<th>TXn</th>
<th>TNn</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>82.7</td>
<td>82.0</td>
<td>87.3</td>
<td>86.1</td>
</tr>
<tr>
<td>6</td>
<td>81.6</td>
<td>81.4</td>
<td>87.1</td>
<td>87.1</td>
</tr>
<tr>
<td>8</td>
<td>81.6</td>
<td>81.3</td>
<td>86.6</td>
<td>86.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$L$</th>
<th>Rx1day</th>
<th>Rx5day</th>
<th>CDD</th>
<th>CWD</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>81.3</td>
<td>81.1</td>
<td>79.9</td>
<td>79.8</td>
</tr>
<tr>
<td>6</td>
<td>80.9</td>
<td>81.2</td>
<td>80.0</td>
<td>79.6</td>
</tr>
<tr>
<td>8</td>
<td>81.0</td>
<td>81.1</td>
<td>79.5</td>
<td>79.6</td>
</tr>
</tbody>
</table>
## Coverage of 90% intervals

<table>
<thead>
<tr>
<th>$L$</th>
<th>TXx</th>
<th>TNx</th>
<th>TXn</th>
<th>TNn</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>91.7</td>
<td>91.1</td>
<td>95.1</td>
<td>94.3</td>
</tr>
<tr>
<td>6</td>
<td>91.3</td>
<td>91.5</td>
<td>94.9</td>
<td>94.4</td>
</tr>
<tr>
<td>8</td>
<td>91.4</td>
<td>91.0</td>
<td>94.5</td>
<td>94.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$L$</th>
<th>Rx1day</th>
<th>Rx5day</th>
<th>CDD</th>
<th>CWD</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>91.0</td>
<td>91.4</td>
<td>90.2</td>
<td>89.7</td>
</tr>
<tr>
<td>6</td>
<td>90.4</td>
<td>91.4</td>
<td>90.2</td>
<td>89.3</td>
</tr>
<tr>
<td>8</td>
<td>90.8</td>
<td>91.2</td>
<td>90.3</td>
<td>89.3</td>
</tr>
</tbody>
</table>
## Coverage of 95% intervals

<table>
<thead>
<tr>
<th>$L$</th>
<th>TXx</th>
<th>TNx</th>
<th>TXn</th>
<th>TNn</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>96.0</td>
<td>96.0</td>
<td>97.8</td>
<td>97.5</td>
</tr>
<tr>
<td>6</td>
<td>95.8</td>
<td>95.9</td>
<td>97.7</td>
<td>97.7</td>
</tr>
<tr>
<td>8</td>
<td>95.8</td>
<td>95.7</td>
<td>97.7</td>
<td>97.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$L$</th>
<th>Rx1day</th>
<th>Rx5day</th>
<th>CDD</th>
<th>CWD</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>95.2</td>
<td>95.7</td>
<td>95.1</td>
<td>94.5</td>
</tr>
<tr>
<td>6</td>
<td>95.1</td>
<td>95.9</td>
<td>94.9</td>
<td>94.6</td>
</tr>
<tr>
<td>8</td>
<td>95.2</td>
<td>95.7</td>
<td>95.2</td>
<td>94.5</td>
</tr>
</tbody>
</table>
CLIMDEX data - Reliability plot with $L = 4$

4 degrees of freedom

Empirical quantiles of $u$

- $TX_x$
- $TN_x$
- $TX_n$
- $TN_n$

Unif(0,1) quantiles

- $Rx_{1\text{day}}$
- $Rx_{5\text{day}}$
- $CDD$
- $CWD$

Reich + Shaby
A spatial Markov model for climate extremes
CLIMDEX data

- We summarize the results using the posterior distribution of the decadal average change.
- We map posterior means and posterior probability the change is positive.
- We also plot the data versus fitted values for several pixels of interest.
- These plots illustrate the non-linear fit of the GEV location.
CDD - mean change

(e) Change in GEV location per decade
CDD - prob change $>0$

(f) Prob change $> 0$
Time series plots

Fresno, CA

Primary site

Adjacent sites
Time series plots

Miami, FL

Primary site
Adjacent sites

Year


TNN

−10 −8 −6 −4 −2 0

Reich + Shaby
A spatial Markov model for climate extremes
Time series plots

New Orleans, LA

Year

Rx1day

Primary site

Adjacent sites

Reich + Shaby

A spatial Markov model for climate extremes
Time series plots

New Orleans, LA

Year

CWD

Primary site

Adjacent sites


Reich + Shaby
A spatial Markov model for climate extremes
Time series plots

Houston, TX

- Primary site
- Adjacent sites

Year


Txx

36 38 40 42
Time series plots

Houston, TX

Year

TNx

Primary site

Adjacent sites

Reich + Shaby
A spatial Markov model for climate extremes
Summary

▶ We have proposed a Markov model for extremes

▶ Can we find a max-stable Markov model?

▶ We’d like to be able to compute the full posterior of $\phi$

▶ Points over a threshold extension should be easy

▶ Support: NSF, DOI, EPA

▶ Thanks!