Introduce Myself - Elizabeth Shamseldin

- PhD Candidate under Richard Smith at UNC-Chapel Hill
- Working with Bayesian methods in spatial statistics
- Opportunity to work on extreme value project at NCAR
Extreme Precipitation:
An Application Modeling N-Year Return Levels at the Station Level

Presented by: Elizabeth Shamseldin
Joint work with: Richard Smith, Doug Nychka, Steve Sain, Dan Cooley
Statistics Group, IMAGe
National Center for Atmospheric Research

April 27, 2006
Motivating Question

Extreme precipitation under various climate scenarios

- Can regional climate model return level estimations be used to obtain return level predictions at the station level?

- Relationship between extremes of grid-cell data (re-analysis NCEP) and weather station data (NCDC)

- Leads to relationship between climate model data and point source data
**Data**

- Point-source observational data from NCDC, originally obtained from Dr. Pavel Groisman

- Covers period 1950-1999 over 5873 stations.

- The data are daily rainfall values; units are tenths of a millimeter.

- Grid-cell data are from NCEP

- Covering period 1948–2003 with no missing data on 288 2.5° grid cells, converted to the same units as the NCDC data.

- Rainfall values are considered over the four seasons

- Threshold values are determined by the 95\textsuperscript{th} and 97\textsuperscript{th} percentiles

- Clustering method is used to define peaks
Outline of Techniques

• Relationship of grid cell n-year return levels to the n-year return levels at point locations is explored

• Patterns are similar, scales differ

• Need model to translate between grid level and point level returns
Outline of Techniques

Tail of the Generalized Extreme Value distribution (GEV) is fit to the grid cell data above a given threshold

\[ \Pr\{Y \leq y\} = \exp \left\{ - \left( 1 + \xi \frac{y - \mu}{\psi} \right)^{-1/\xi} \right\} \] (1)

where \( Y \) is a random variable, \( \mu \) is a location parameter, \( \psi \) a scale parameter and \( \xi \) is the extreme-value shape parameter; \( \mu \) and \( \xi \) can take any value in \((-\infty, \infty)\) but \( \psi \) has to be \( > 0 \).
Technique

- Peaks Over Threshold model: considers \( \Pr\{X \leq u + x \mid X > u\} \) for a given threshold \( u \); the parameters are directly tied to the threshold value.

- Point process approach is similar to the peaks over threshold method: all observations over a given threshold are considered. Tail of GEV is estimated.

- However, in the PP approach, the parameters are not tied to the threshold value.
Theory

Provided the model fits the data:

- PP approach should produce equivalent parameter values as the POT approach

- Parameters are independent of the threshold (adjusting for estimation error)

- Ideal threshold determined by considering where the parameter values stabilize
Theory

Choose to model n-year return due to interpretability. Event so extreme it is expected to happen once every n years: \( \left( \frac{1}{n} \right) \text{ probability/yr} \)

The n-year return values can be directly obtained using the estimated GEV parameters. Define \( y_n \) by the equation:

\[
\left( 1 + \xi \frac{y_n - \mu}{\psi} \right)^{-1/\xi} = \frac{1}{n}
\]

which leads to the formula:

\[
y_n = \begin{cases} 
\mu + \psi \frac{n^\xi - 1}{\xi} & \text{if } \xi \neq 0, \\
\mu + \psi \log n & \text{if } \xi = 0. 
\end{cases}
\]  

(2)
Recap

- GEV parameter estimates are obtained and checked through various diagnostics

- GEV parameter estimates are used to generate n-year return levels at grid and station levels

- Various models are explored to predict point location n-year return from grid cell n-year return
GEV Model Fit - Mu Parameter: 95 Grid - Point vs 97 Grid - Point
GEV Model Fit - LogPsi Parameter: 95 Grid - Point vs 97 Grid - Point

Grid LogPsi DJF-95

Grid LogPsi DJF-97

Point LogPsi DJF-95

Point LogPsi DJF-97
GEV Model Fit - Xi Parameter: 95 Grid - Point vs 97 Grid - Point
100-Year Return: Point vs Grid - LogPoint vs Grid

100-Yr Ret Fitted Values DJF

100-Yr Ret Student Residuals - DJF

100-Yr Ret LogPt (W) Fitted Values - DJF

100-Yr Ret LogPt (W) Student Resid - DJF
100-Year Return: LogPoint + Elevation vs Grid
### Model Results and Comparison

Univariate Regression Predicting Point Locations 100-Year Return Level Using Grid Cell

100-Year Return

<table>
<thead>
<tr>
<th>Season</th>
<th>Model Type</th>
<th>OBS</th>
<th>AIC</th>
<th>Intercept</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WINTER - 100</strong></td>
<td>log(Pt) Grid+Elev</td>
<td>4025</td>
<td>3425</td>
<td>5.3186</td>
<td>0.0030</td>
<td>-0.000124</td>
</tr>
<tr>
<td></td>
<td>97 log(Pt) Grid+Elev</td>
<td>4013</td>
<td>3356</td>
<td>5.3074</td>
<td>0.0030</td>
<td>-0.000113</td>
</tr>
<tr>
<td><strong>SPRING - 100</strong></td>
<td>log(Pt) Grid+Elev</td>
<td>4138</td>
<td>1924</td>
<td>5.8960</td>
<td>0.0023</td>
<td>-0.000258</td>
</tr>
<tr>
<td></td>
<td>97 log(Pt) Grid+Elev</td>
<td>4145</td>
<td>2282</td>
<td>5.9736</td>
<td>0.0021</td>
<td>-0.000281</td>
</tr>
</tbody>
</table>
100-Year Return: $\text{LogPoint} \sim \text{Grid} + \text{Elevation Spatial Trend}$

**Fitted 100-Yr Return Values**

$\text{LogPt-Grid+Elev Fitted DJF-95}$

**Residuals**

$\text{LogPt-Grid+Elev StuResid DJF-95}$
100-Year Return: $\text{LogPt} \sim \text{Grid} + \text{Elevation Including Lat and Lon}$

**Quadratic Model**

**Residuals**

**Cubic Model**

**Residuals**
100-Year Return: LogPt $\sim$ Grid + Elevation Including Lat and Lon

Variograms of Log(Pt)+Elevation

**Quadratic Variogram**

**Cubic Variogram**
Other Models Considered

Cubic vs Quartic

- Cubic and quartic models show little difference in the remaining spatial trend

- Cubic model: all terms and interaction terms are significant in the model

- Quartic model: some terms are no longer significant, including elevation

- AIC vs BIC

Elevation

- Quadratic models in elevation: did not resolve the spatial correlation among station return levels as well as the higher order models in latitude and longitude
Conclusion and Future Work

- Modeling the tail of the GEV Distribution appears to produce stable GEV parameter estimates and model coefficients within seasons and across 95% and 97% thresholds

- 100 year return levels are successfully modeled by season at the point (station) level using grid-level return values, station elevation, and station latitude and longitude coordinates

- The regression relationship between climate model predictions and future modeled extremes is best expressed through a cubic model in latitude and longitude to account for the spatial correlation, with \( R^2 = 0.77 \) and standard errors within 1.008 and 1.056 tenths of a millimeter.

- Future work includes plans to test grid-point models on RCM and CCSM data
References


Theory

To take account of these deficiencies, an alternative class of methods has been developed, often known as the *Peaks Over Thresholds* (POT) approach. For the distribution of the excess values, a common family of probability density functions is the *Generalized Pareto Distribution* (GPD), introduced by Pickands (1975), and given by

\[
\Pr\{X \leq u + x \mid X > u\} = 1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-1/\xi}.
\]

One drawback in the POT model is that the parameters are directly tied to the threshold value, \(u\).

A third approach, the *point process approach* (Smith 1989, 2003, Coles 2001), although operationally very similar to the POT approach, uses a representation of the probability distribution that leads directly to the GEV parameters \((\mu, \psi, \xi)\).
The Point Process method considers $N$ peaks, $Y_1...Y_N$, observed at times $T_1...T_N$. Pairs are viewed as points in the space $[0, T] \times (u, \infty)$ ($u=$threshold), which form a nonhomogeneous Poisson process with intensity measure:

$$\lambda(t, y) = \frac{1}{\psi} \left( 1 + \xi \frac{(y - \mu)}{\psi} \right)^{-\frac{1}{\xi} - 1}$$
Theory

By standard formulae for a Poisson Process, the likelihood is of the form:

\[
L(\mu, \psi, \xi) = \prod_{i=1}^{N} \lambda(T_i, Y_i) \cdot \exp \left\{ - \int_0^T \int_0^\infty \lambda(t, y) dt dy \right\} \\
= \prod_{i=1}^{N} \lambda(T_i, X_i) \cdot \exp \left\{ -T \left( 1 + \xi \frac{u - \mu}{\psi} \right)^{-1/\xi} \right\}. \tag{4}
\]
Theory

In practice we work with the negative log likelihood, $\ell = -\log L$, which leads to

$$\ell(\mu, \psi, \xi) = N \log \psi + \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{N} \log \left(1 + \xi \frac{Y_i - \mu}{\psi}\right) +$$

$$+ T \left(1 + \xi \frac{u - \mu}{\psi}\right)^{-1/\xi} +$$

where $T$ is the length of the observation period in years and the $\ldots (+$ symbols essentially mean that the expression are only evaluated if $1 + \xi \frac{u - \mu}{\psi} > 0$ and $1 + \xi \frac{Y_i - \mu}{\psi} > 0$ for each $i$ (if these constraints are violated, $L$ is automatically set to 0).
The basic method of estimation is therefore to choose the parameters \((\mu, \psi, \xi)\) to maximize (??) or equivalently to minimize (5).

This is performed by numerical nonlinear optimization.

In practice it is convenient to replace \((\mu, \psi, \xi)\) by \((\theta_1, \theta_2, \theta_3)\) where \(\theta_1 = \mu, \theta_2 = \log \psi, \theta_3 = \xi\) (defining \(\theta_2\) to be \(\log \psi\) rather than \(\psi\) itself makes the algorithm more numerically stable, and has the advantage that we don’t have to build the constraint \(\psi > 0\) explicitly into the optimization procedure).