Climate Extremes, What to do with 8000 histograms?

Douglas Nychka, National Center for Atmospheric Research
Summary

- Precipitation extremes
- Regional Climate models
- Adding a spatial element
- What about Yellowstone and Cheyenne?

Components: NonGaussian distributions, functional data, sparse and embarrassingly parallel methods

Credits:
Whitney Huang, Dorit Hammerling, Sophia Chen, and Nathan Lenssen.
Precipitation extremes for Boulder, CO

Daily precipitation amounts for Boulder

25 year daily return level:
In any given year daily precipitation has a 1/25 chance of exceeding this level.

How does this vary over space?

How well does a model simulate this variable?
PART 1:
Estimates of climate extremes

- Generalized Pareto pdf
- Nonparametric density estimates

Suzanne Beard
Generalized Pareto Fit:

Fit to observations > 2 cm with 95% CI for 25 year return level

Generalized Pareto: pdf( x) depends on three parameters:

\[ pdf(x) \sim \left( 1 + \xi \frac{(x - \mu)}{\sigma} \right)^{-\frac{1}{\xi}} \] for \( x \geq \mu \)

- (1) scale (\( \sigma \)) , (2) shape( \( \xi \)) and (3) probability of exceeding threshold (\( P(Z > \mu) \)) .

- With these one can find all quantiles, means and return levels.

D. Nychka Climate Extremes, Computing Extremes
Beyond the Pareto

Probability density function:

\[ pdf(x) = e^{g(x)} \]

\( g \) is the log density function

- Estimate \( g \) as a flexible spline function and in the scale of log precipitation. i.e. \( x = \log(\text{precip}) \)
Approximate, but fast, log densities

- Apply a Possion generalized linear model to a finely binned histogram of counts
- Use a penalized, cubic spline smoother and estimate the smoothing parameter by approximate cross validation.
- Normalize estimate of $g$ to integrate to one.

**log Penalized likelihood,**

$$
\min_g \left( \sum_{j=1}^{N} y_j g_j - e^{g_j} + \log(y_j!) \right) - \lambda \left( \int_{[x_1,x_N]} (g''(x))^2 dx \right)
$$

$x_j$ bin midpoints, $y_j$ bin counts, $g_j = g(x_j)$

- Maximization is easy using iteratively reweighted least squares
- No easy way to sample from "posterior"
Details ...

- Constrain $g$ to extrapolate as a linear function

linear $g \rightarrow$ pdf has exponential tail $\log(\text{precip})$

$\rightarrow$ polynomial tail precip

**Off the shelf tools:**

- Chong Gu spline density estimate
- Adapt *gam, mgcv* – S. Woods R packages
Fit to Boulder data

Three different smoothing parameters:

Log scale

Cross validation choice for $\lambda$ is effected by discretization at small precipitation amounts.
log densities

log density

log precip

log spline rough, log spline smooth, Generalized Pareto
PART 2: Extremes from regional climate models

We Share a Dream, Kaarina Kaikkonen
Modeling strategy

- Nest a fine-scale weather model in part of a global model’s domain.

- Consider different regional models to characterize model uncertainty.

- North American Regional Climate Change and Assessment Program (NARCCAP)
  a large set of numerical experiments to explore uncertainty.

A snapshot from the 3-dimensional RSM3 model (right) forced by global observations (left)
Four regional models (MM5I, RCM3, WRFP, ECPC) that are driven by observed atmosphere at the boundaries of the NARCCAP domain.

Just look at part of Rocky Mountain region – about 800 grid points.

20 years of daily downscaled/simulated weather for each model.

**How do extremes of daily summer rainfall vary over space and over climate models?**
Functional boxplots of log densities

log spline pdfs for four models at all 800+ grid points

See Sun and Genton (2011) for more on functional boxplots

D. Nychka Climate Extremes, Computing Extremes
Principle components

First three principle components of log densities

\[ g(x) = a_1 \phi_1(x) + a_2 \phi_2(x) + a_3 \phi_3(x) \]

\((a_1, a_2, a_3)\) vary for every grid box and every regional model \(4 \times 800 \times 3\) coefficients
Use these as basis functions to refit models using standard GLM maximum likelihood

\[ g(x) = a_1\phi_1(x) + a_2\phi_2(x) + a_3\phi_3(x) \]

- Still cheating by fitting bins counts and not normalizing as a density function
- This may be a way to introduce covariates in a simple way
- Local likelihood fitting over space to smooth.
The spatial problem

*Coefficients vary over space, are noisy and are correlated.*

We have 4 Models $\times$ 3 coefficients = 12 spatial fields.

**First coefficient for MM5I**

- Transform each climate models coefficients to be uncorrelated.
- Smooth transformed coefficients using spatial statistics.
PART 3:
Spatial stats for large data
LatticeKrig: spatial smoother

Representing the surface: \[ g(x) = \sum_j \phi_j(x)c_j \]

Fix the basis and estimate the coefficients from data.

\[ y = Xc + e \quad c \sim N(0, Q^{-1}) \]

\[ X_{i,j} = \phi_j(x_i) \]
More about Q

Some coefficients:
\[
\begin{array}{cccc}
\ldots & \ldots & \ldots & \ldots \\
\ldots & c_1 & \ldots \\
\ldots & c_2 & c_* & c_3 \\
\ldots & c_4 & \ldots \\
\ldots & \ldots & \ldots & \ldots \\
\end{array}
\]

Some weights:
\[
\begin{array}{cccc}
\ldots & \ldots & \ldots & \ldots \\
\ldots & -1/4 & \ldots \\
-1/4 & \alpha & -1/4 \\
-1/4 & \ldots \\
\ldots & \ldots & \ldots & \ldots \\
\end{array}
\]

The filter:
\[
\alpha c_* - 1/4 (c_1 + c_2 + c_3 + c_4) = \text{white noise}
\]

\[\alpha \geq 1.\]

- Can exploit sparse linear algebra for the "Kriging" computation
- Multiresolution version approximates standard spatial covariance functions.
First coefficient for MM5I

Original coefficients.

Smooth component

Elevation

Fitted values
Reconstructing the Boulder grid box

MMI5 model, log spline, GLM with 3 basis functions, smoothed coefficients
Uncertainty

Density

log precip

0.00 0.05 0.10 0.15 0.20 0.25 0.30

0 2 4

-2

D. Nychka Climate Extremes, Computing Extremes
Boulder grid box 25 year return

Return Level (years)

Precip (cm)
25 year return surface

"posterior mode" for MM5I model.
PART 4:
Data analysis on Yellowstone

If I have to wait too long for my answer I forget my question.  – Rich Loft
The Yellowstone supercomputer.

≈72K cores = 4536 (nodes) × 16 (cores) and each core with 2Gb memory
16 Pb parallel file system

- Core-hours are available to the NSF research community.
- Simple application process for graduate student allocations.
- Supports R in both interactive and batch mode.

*Cheyene* will be running January 2017 and will have 3 times capacity of *Yellowstone*. 
Using the `Rmpi` package.

In R ...

```r
library(Rmpi)
# Spawn 4 workers
mpi.spawn.Rworkers(nworkers=4)
  # Broadcast an R function to all workers

mpi.bcast.Robj2worker(doStats)
  # apply this function to 100 tasks (each worker will get about 25)

output <- mpi.iapplyLB(1:100, doStats)
```

*output is a list (100 components) with the result for each case.*
Are many R workers processes feasible?

- Time to initiate 100 - 1000 workers nearly constant at 3 seconds
- Workers lose little time reading common data files.
- Median execution time of task per worker is nearly constant.
- Successfully used for fitting extreme value distributions, spatial fields, covariance models.
Summary

- Nonparametric methods are available for estimating the tail behavior of climate distributions.
- Borrowing strength from spatial neighbors and dimension reduction help to make them work.
- Methods can be easily migrated to large computing systems.
Thank you
Regional simulations for N. America

North American Regional Climate Change and Assessment Program (NARCCAP)

4GCMS × 6RCMs:
12 runs – balanced half fraction design
Global observations × 6RCMs
X High resolution global atmosphere

<table>
<thead>
<tr>
<th>GLOBAL MODEL</th>
<th>REGIONAL MODELS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MM5I</td>
</tr>
<tr>
<td>GFDL</td>
<td></td>
</tr>
<tr>
<td>HADCM3</td>
<td>○</td>
</tr>
<tr>
<td>CCSM</td>
<td>○</td>
</tr>
<tr>
<td>CGCM3</td>
<td></td>
</tr>
<tr>
<td>Reanalysis</td>
<td>■</td>
</tr>
</tbody>
</table>

NCAR grid over land is ≈ 8-9K grid points.
Study region

NARCCAP domain and Rocky Mountain MM5I grid cells.

(About 800 grid points in subregion.)

How do extremes of daily summer rainfall vary over space and over climate models?