Analyzing Extreme Events using Extreme Value Theory

Erin Towler & Mari Tye

C3WE, NCAR

July 26, 2019

Weather & Climate Extremes Tutorial

NCAR

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Yesterday, in the *Extremes: State of the Science* lecture, we learned:

There is no exact definition of an extreme.

2 (Main) Ways to Define Climate Extremes

1) "Extreme indices" (Abby's talk)
   - Based on probability of occurrence (relative) – e.g., 90th percentile of observed Tmax.
   - Based on specific (possibly impact-related) threshold (absolute).
   - These are "moderate extremes", use up to 5% of the sample.

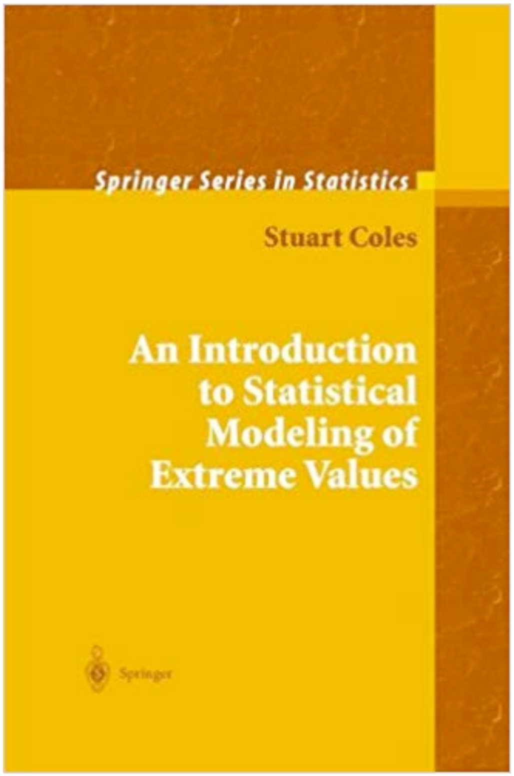
2) Extreme value Theory (EVT) (This talk!)
   - For more "extreme extremes", need EVT because of sampling issues (typically < 1-5% of total sample)
   - Can be use to estimate probabilities of "values never seen".

And earlier, we learned about the statistical models commonly used for perfect prognosis (PP) downscaling:

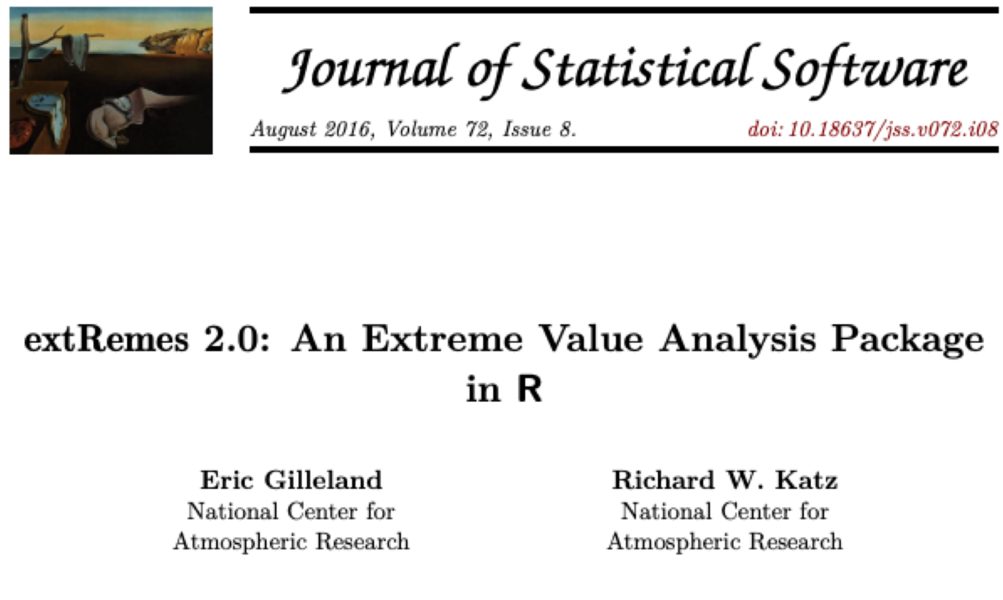
1. Linear Regression
2. Generalized linear models (GLM)
3. Nonlinear regression
4. Analogue method
5. Extreme value statistics

(See Maraun et al 2010 for additional models)

Several good resources for EVT:



Coles (2001)



Journal of Statistical Software

August 2016, Volume 72, Issue 8. doi: 10.18637/jss.v072.i08

extRemes 2.0: An Extreme Value Analysis Package in R

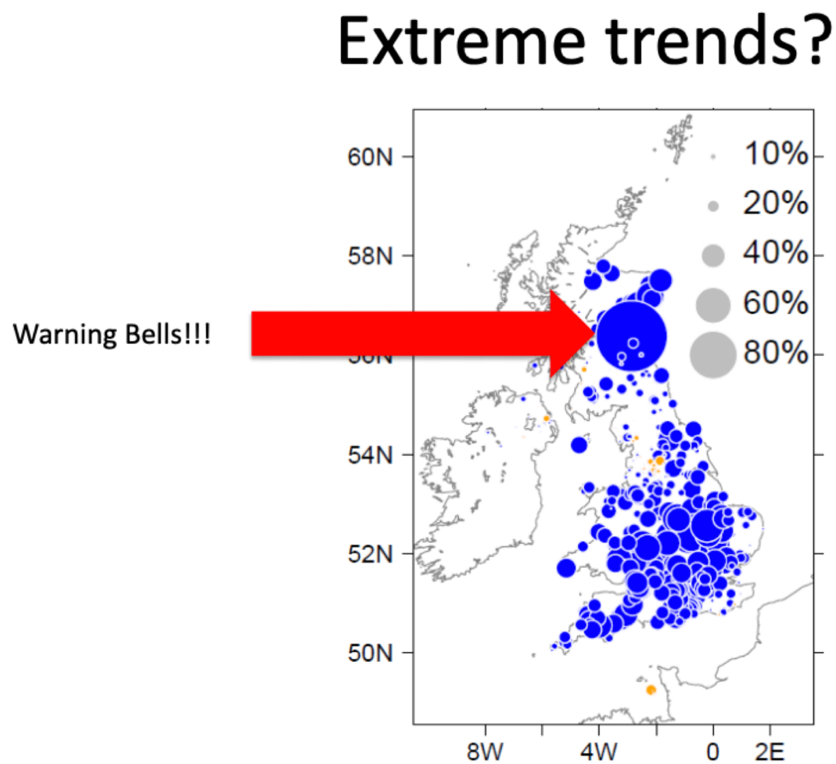
Eric Gilleland
National Center for Atmospheric Research

Richard W. Katz
National Center for Atmospheric Research

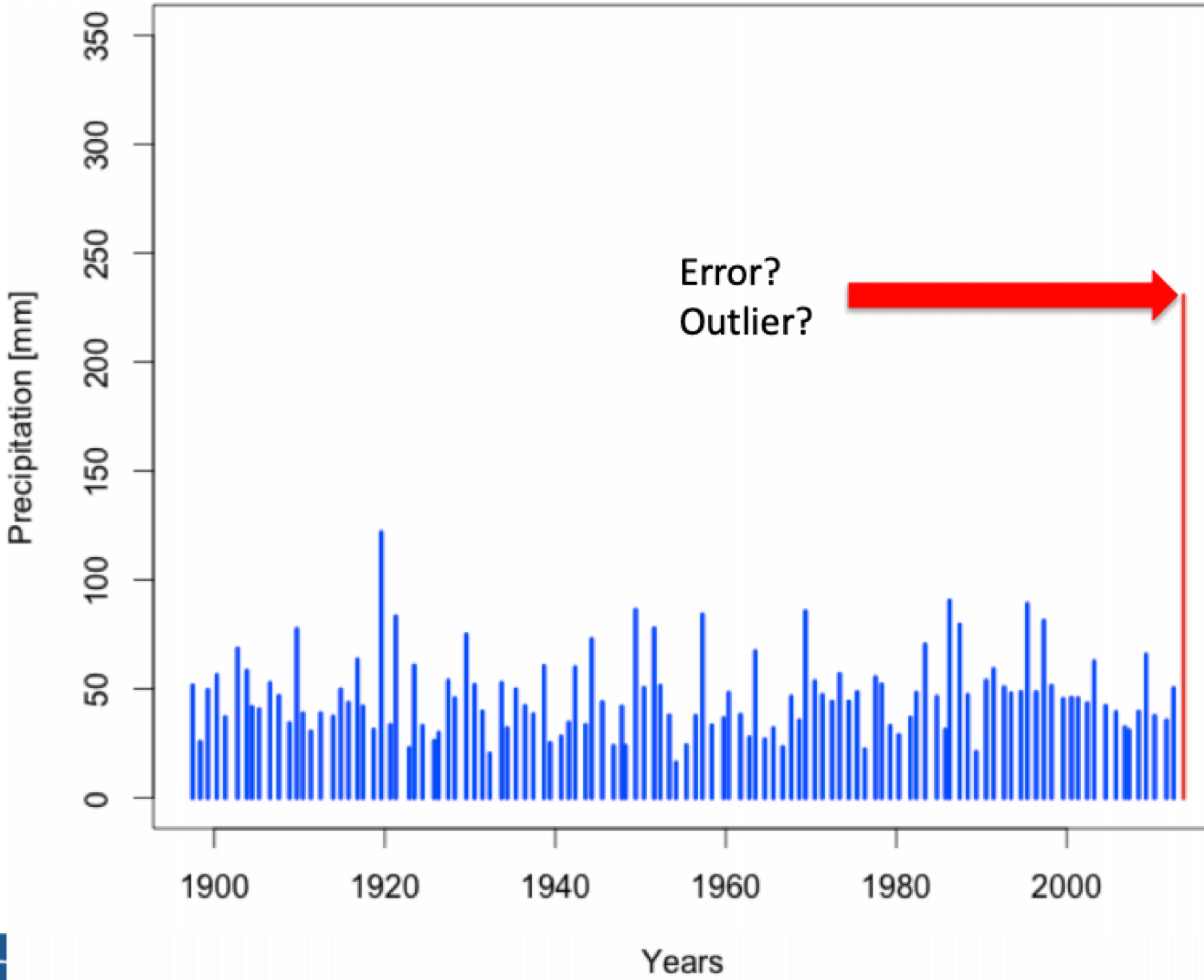
Gilleland and Katz (2016)

But check your data, is it really an extreme?

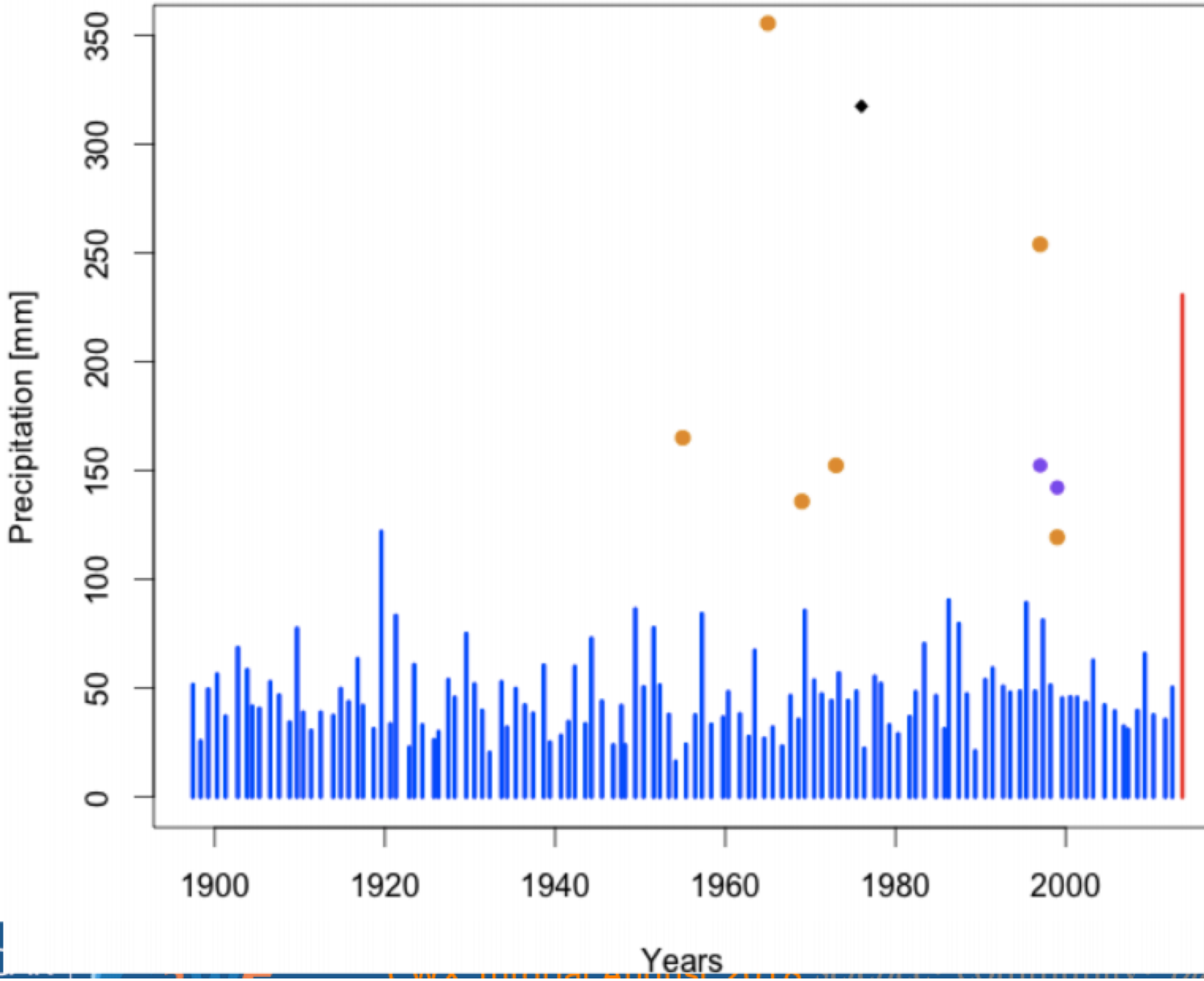
- Extremes always on 1st or last of month?
- Always rains on a Monday?

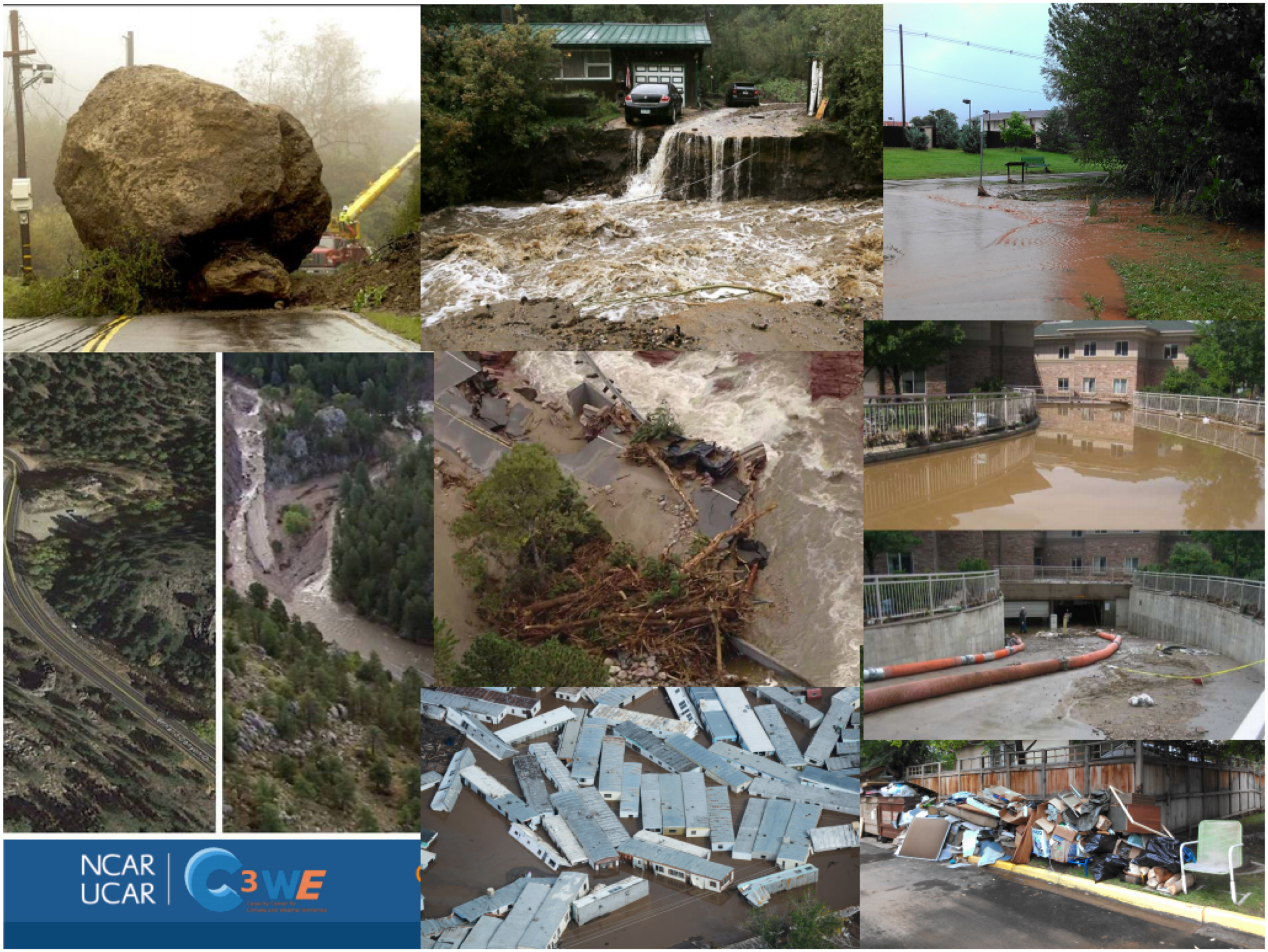


An extreme by any other name



An extreme by any other name



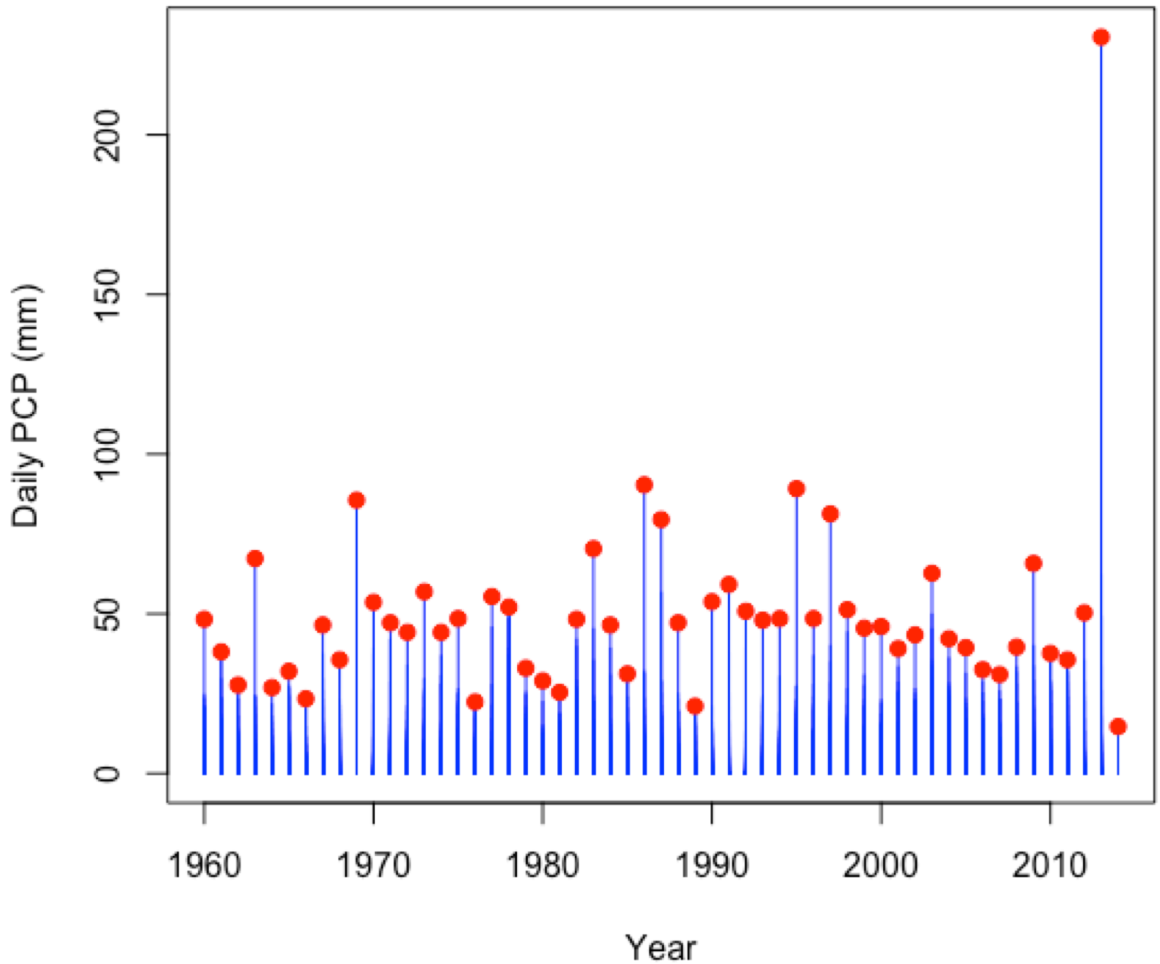


2 main approaches to analyzing extremes:

1. Block maxima

2. Peaks over threshold

Block maxima approach extracts maximum values for a given time block (e.g., month, season, year).



Block maxima can be fit using the generalized extreme value (GEV) distribution function, which has three fitted parameters

$$G(z) = exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

Block maxima can be fit using the generalized extreme value (GEV) distribution function, which has three fitted parameters

$$G(z) = exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

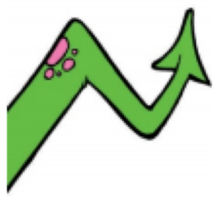
3 Model parameters:

Location parameter, μ, where distribution is centered, $-\infty > \mu < \infty$.

Scale parameter, σ, spread of the distribution, $\sigma > 0$.

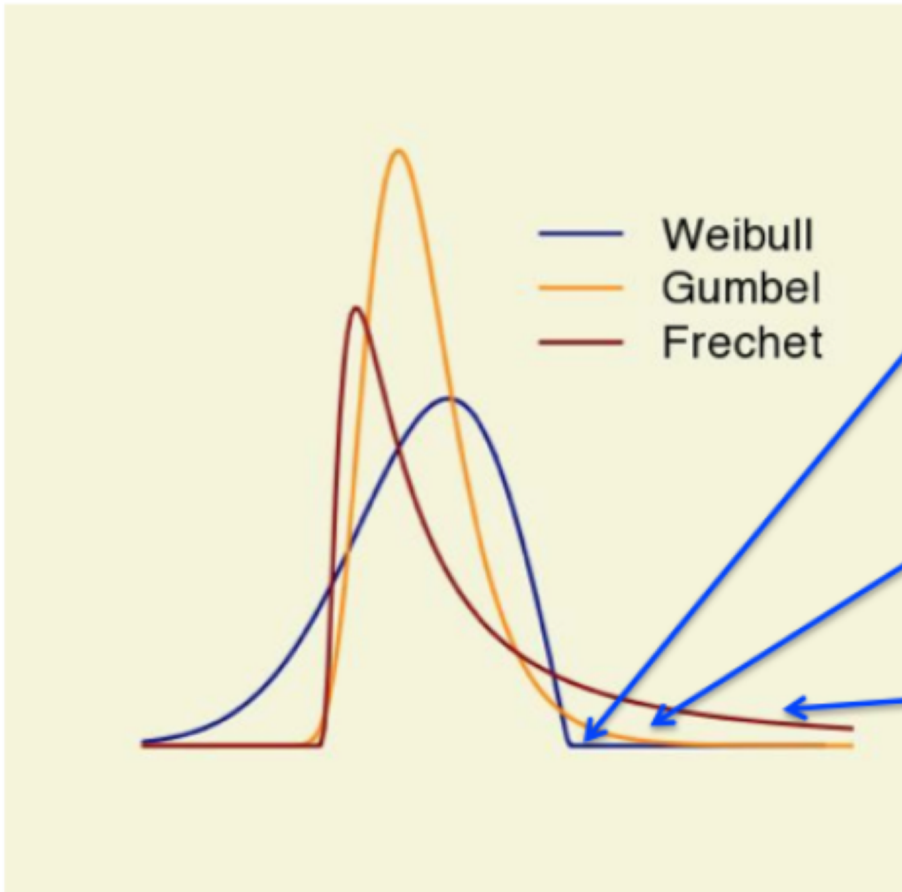
Shape parameter, ξ, behavior of distribution tail.

All about the Tails....



With different shapes and scales

The shape parameter determines the three types of GEV distributions



Weibull; bounded upper tail $\xi < 0$

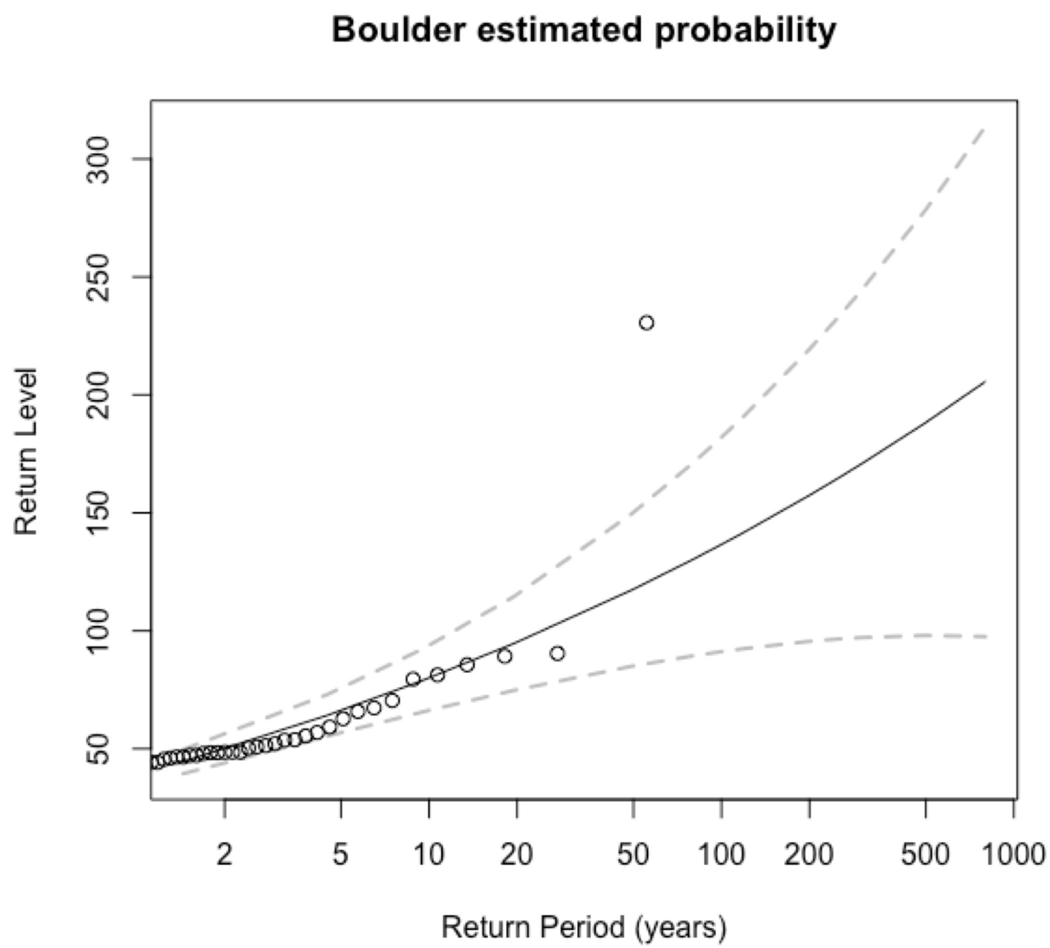
Gumbel; light tail $\xi \longrightarrow 0$

Fréchet; heavy tail $\xi > 0$

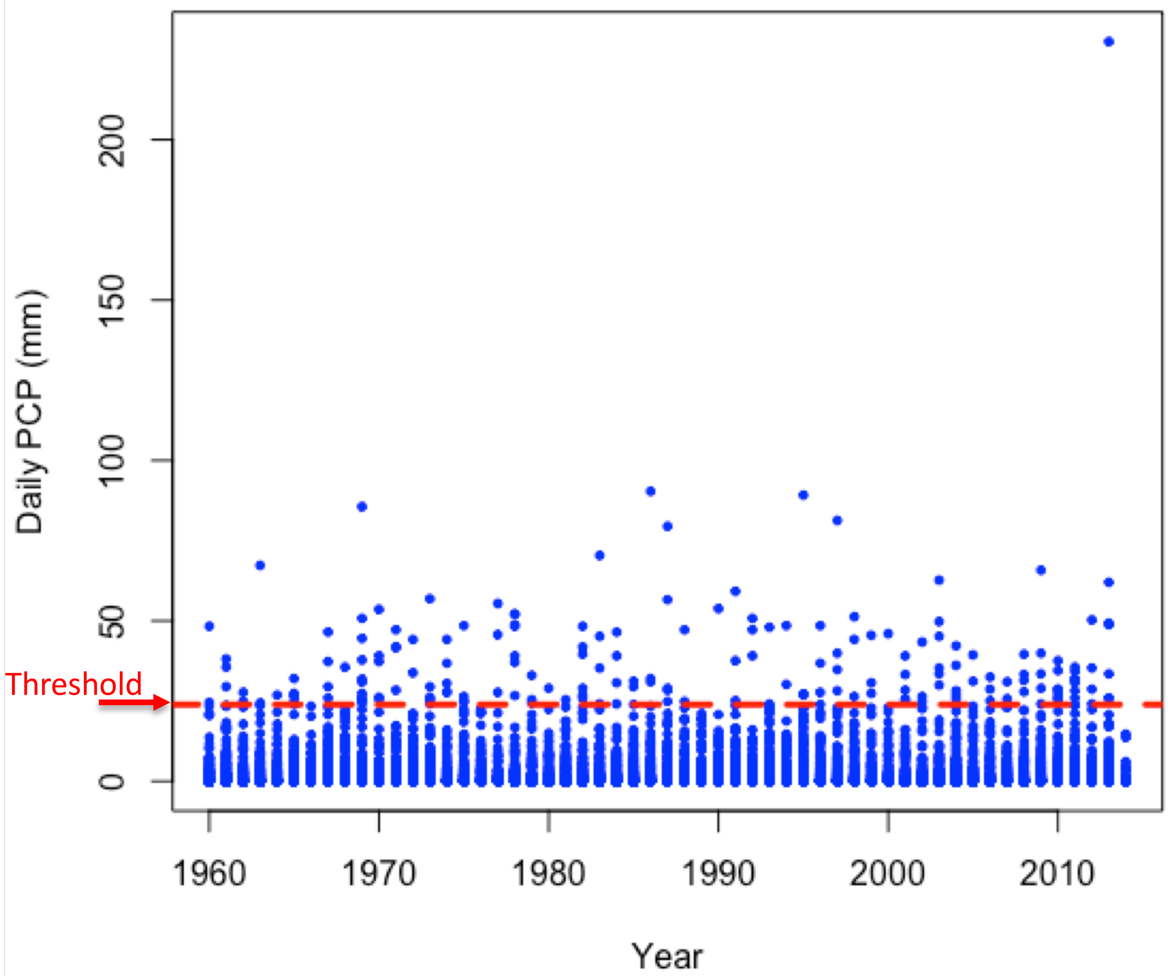
Extremal Types Theorem

- Weibull Type: *Negative* shape parameter
  and bounded upper tail
  - temperature, wind speed, sea level
- Gumbel Type: Zero shape parameter
  light upper tail
  - domain of attraction for exponential family
- Fréchet Type: *positive* shape parameter
  heavy upper tail
  - precipitation, stream flow, economic damages

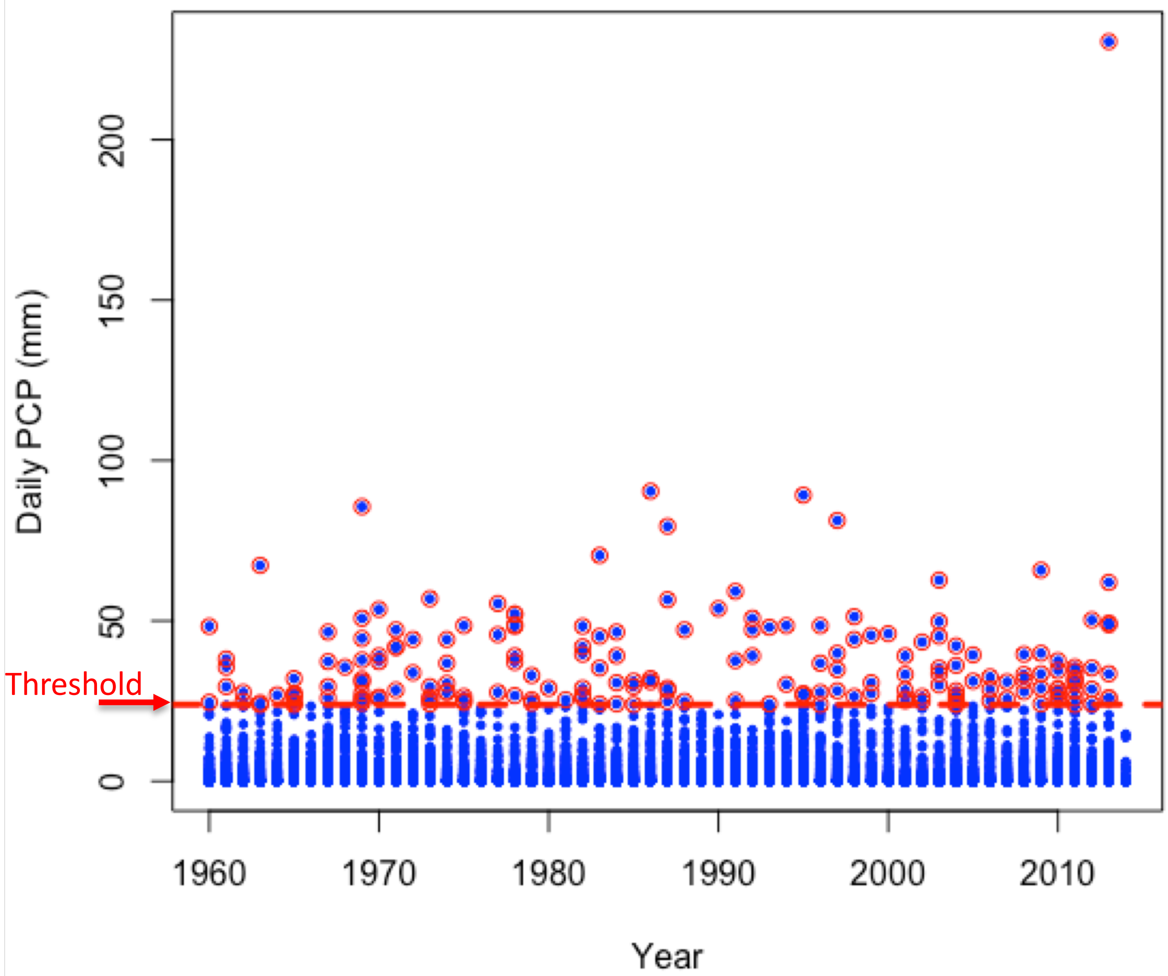
GEV can be used to obtain return levels and confidence intervals



Peaks over threshold (POT) extracts values above a high threshold



Peaks over threshold (POT) extracts values above a high threshold



- Considers more extremes per time block
- BUT:
    - Need to select a threshold (subjective)
    - Data may be temporally clustered (not independent)

POT can be fit using the generalized Pareto (GP) distribution, which is analogous to GEV.

$$\mathrm{H}(x) = 1 - \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}$$

Select threshold, u

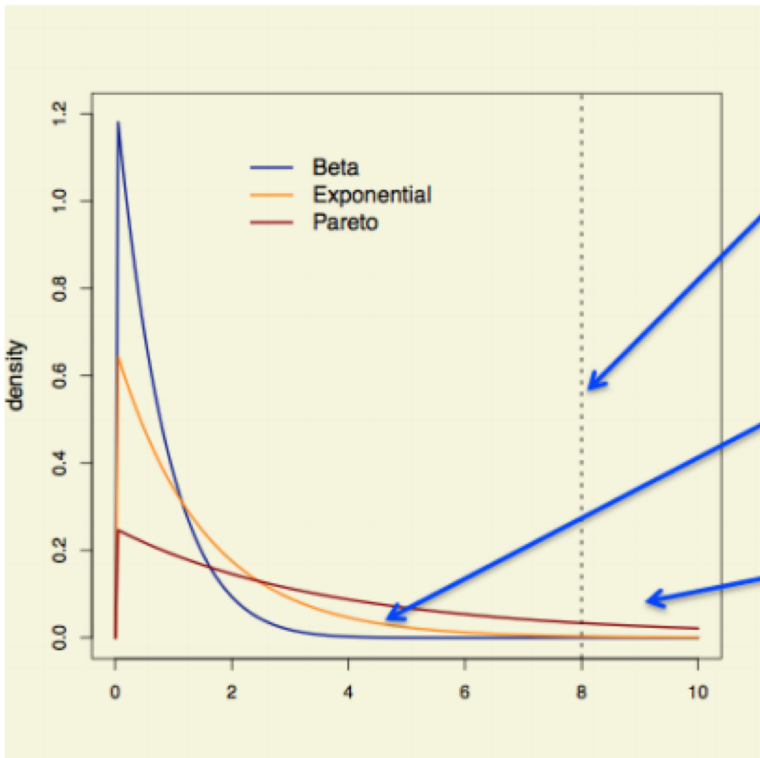
2 Model parameters:

Scale parameter, σ, spread of the distribution, $\sigma > 0$.

Shape parameter, ξ, behavior of distribution tail.

Generalized Pareto Distribution

Three types – as for GEV



Beta df;
bounded
upper tail

Exponential;
light tail

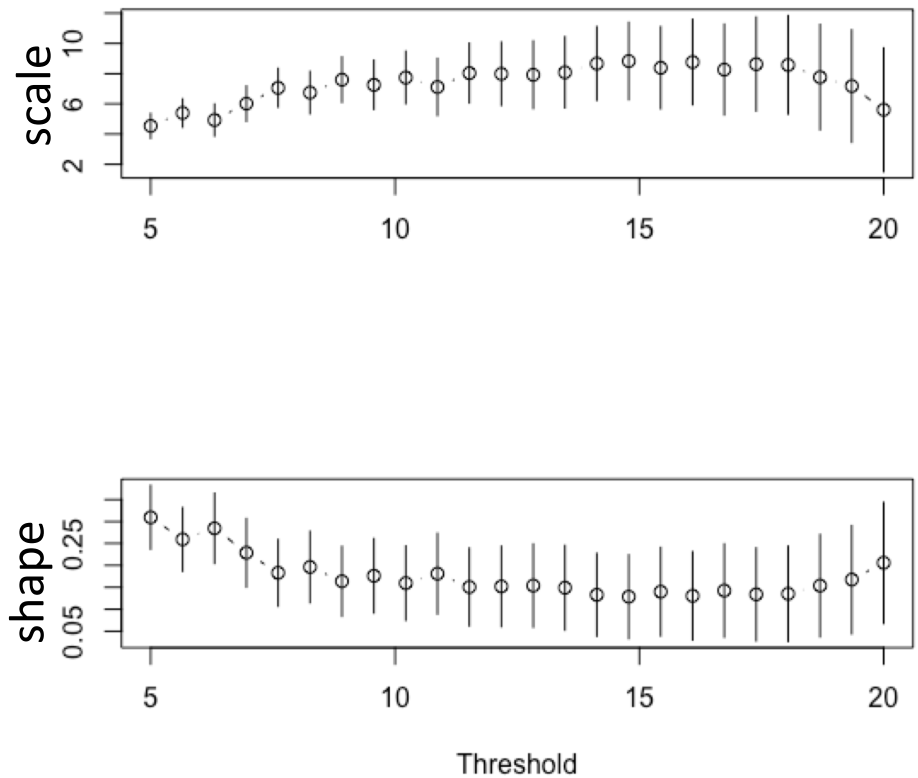
Pareto;
heavy tail

Threshold selection is a tradeoff between bias and variance:

- Higher threshold gives better GP approximation (lower bias, but less values means higher variance)
- Lower threshold improves confidence (more values means less variance, but higher bias)
- No real automation procedure

Threshold selection can be guided by where GP parameter estimates stabilize

- Need to check parameter stability for changing threshold (Δu)
- Ultimate decision is subjective and requires trial-and-error

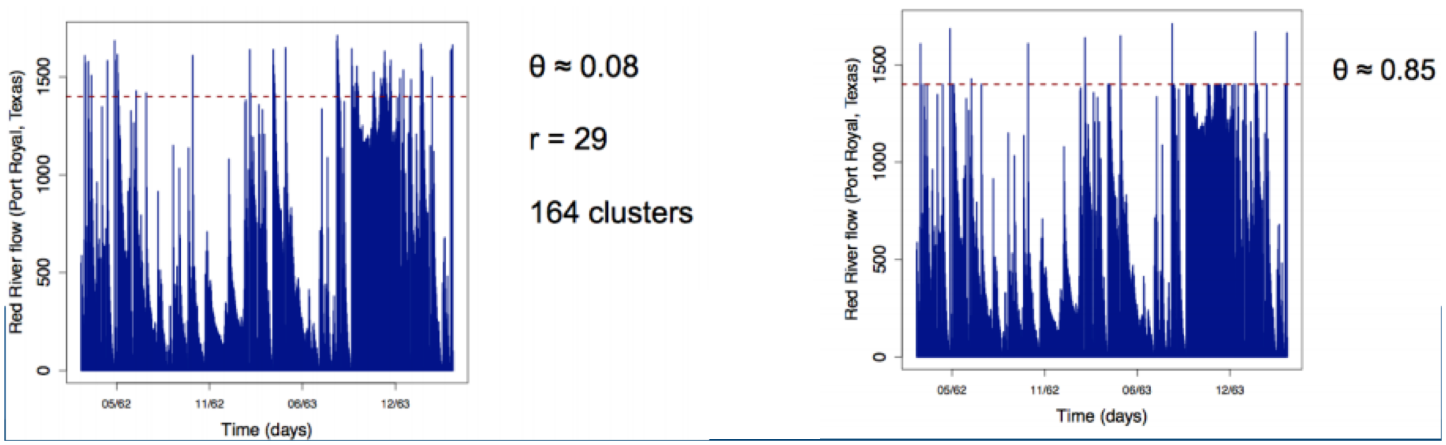


An assumption of the GP is that the data are independent, which is often violated by climate and weather data.

- Remove dependence in exceedances by declustering
  - Test extremal index

$0 < \theta \leq 1$ Mean cluster size $1/\theta$

$\theta = 1$ implies no clustering



GEV versus GP

Generalized Extreme Value	Generalized Pareto
Use single value per time block	Doesn't ignore other maxima in time block
Can assume independence	Clustered data can cause problems
Direct estimation of frequency & intensity	Only estimates intensity, unless combined with Poisson distribution for Point-Process

Model evaluation

- Test the model fit
  - Akaike Information Criterion, Deviance Statistic, Negative Log Likelihood
- Diagnostic plots
  - Quantile-quantile plots
- Check the significance of nested models
  - Likelihood ratio test.

To account for non-stationarity, the parameters can vary with covariates, or predictors.

e.g., temporal trend in location, μ

$\mu(t) = \mu_0 + \mu_1 \mathrm{t}$

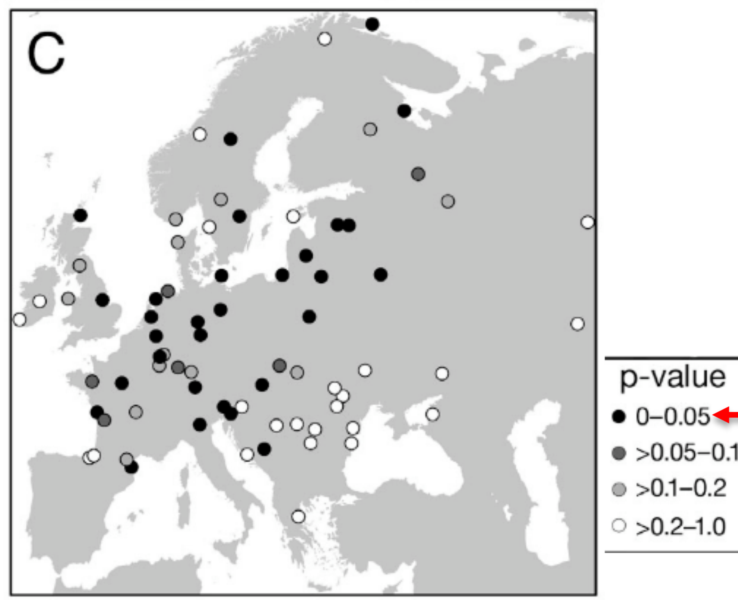
$$G(z) = exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

$\ln\sigma(t) = \phi_0 + \phi_1 \mathrm{t}$

Incorporating non-stationarity can improve statistics or be used for downscaling

Example covariates:

- Atmospheric driver (ENSO, PDO, NAO, etc)
- Trends (urbanization, climate)
- Seasonal cycle

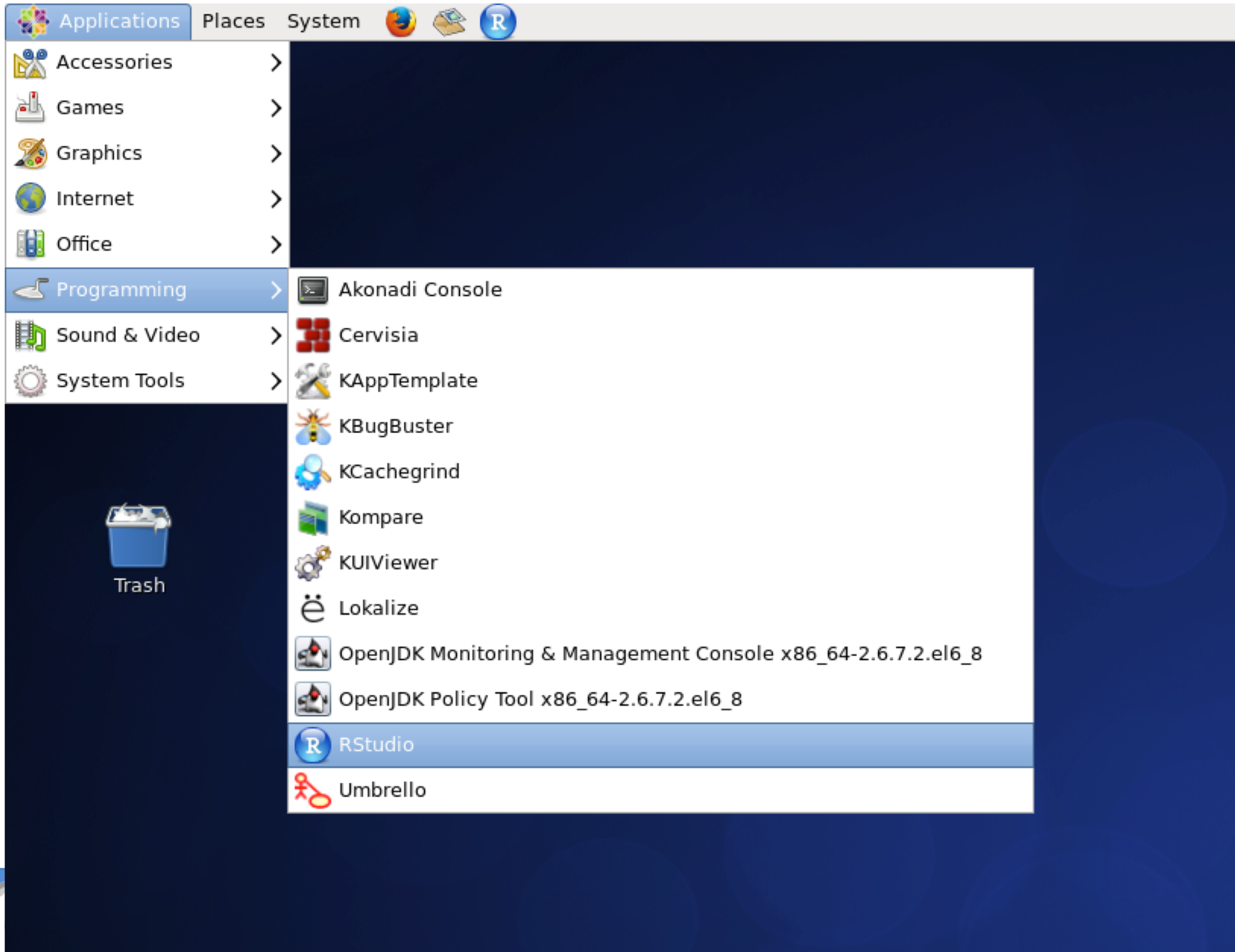


$\mu = f(NAO)$ for hot spells.

Photiadou et al., 2014
10.3354/cr01191

Questions?

Opening R Studio on classroom machines:



Why use R ?

- Free and open source
- Works on every platform
- Cutting edge tools
- Create repeatable work
- Huge support community (18,000+ users, LinkedIn Group, RUsers, Meetups, twitter, stackoverflow)