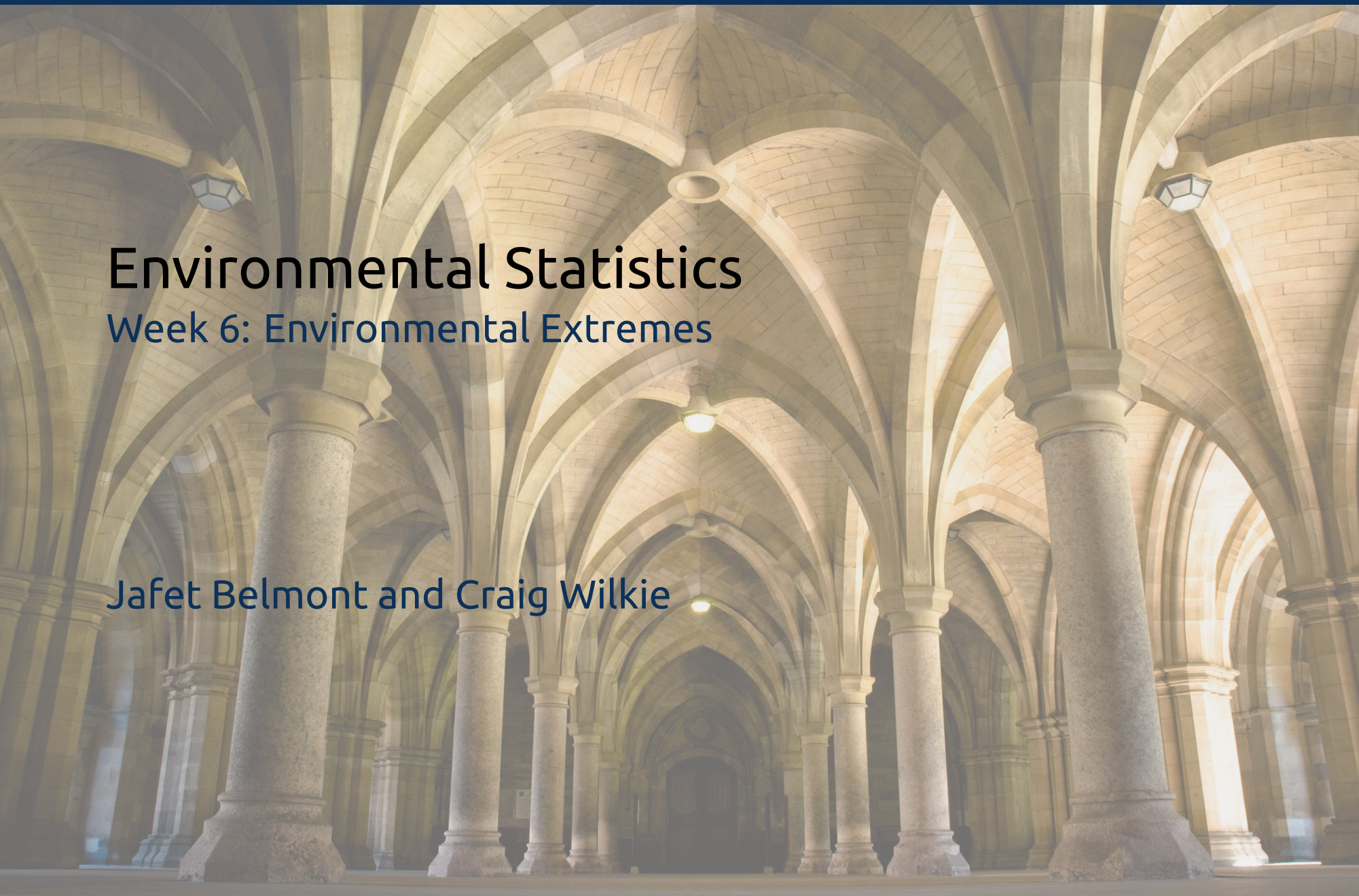
University of Glasgow



Environmental Statistics

Week 6: Environmental Extremes

Jafet Belmont and Craig Wilkie

Recap

- For the last two weeks, we have looked at time series, and how appropriate models can be fitted that deal with the temporal structures in the data.
- These models allow us to make inference on the mean of the data, but sometimes we are more interested in very high or very low values.
- In some cases, we want to specifically look at the maxima or minima of a particular environmental system.
- This week, we will look at methods for modelling these statistical **extremes**.

Extremes

- We are trying to model rare events - by their very definition we won't have a lot of data on these.
- The bulk of the data in any statistical distribution will be in the centre.
- Standard density estimation techniques (eg the normal distribution) work well where the data have the greatest density, but that's not what we need here.
- We need to use a statistical model which is good at estimating the **tails** of our distribution.

Motivation

Statistical modelling of extreme environmental phenomena has a very practical motivation:

reliability - anything we build needs to have a good chance of surviving the weather/environment for the whole of its working life.

This has obvious implications for civil engineers and planners. They need to know:

- how strong to make buildings;
- how high to build sea walls;
- how tall to build reservoire dams;
- how much fuel to stockpile;

This motivates the need to estimate what the:

- strongest wind;
- highest tide;
- heaviest rainfall;
- most severe cold-spell;

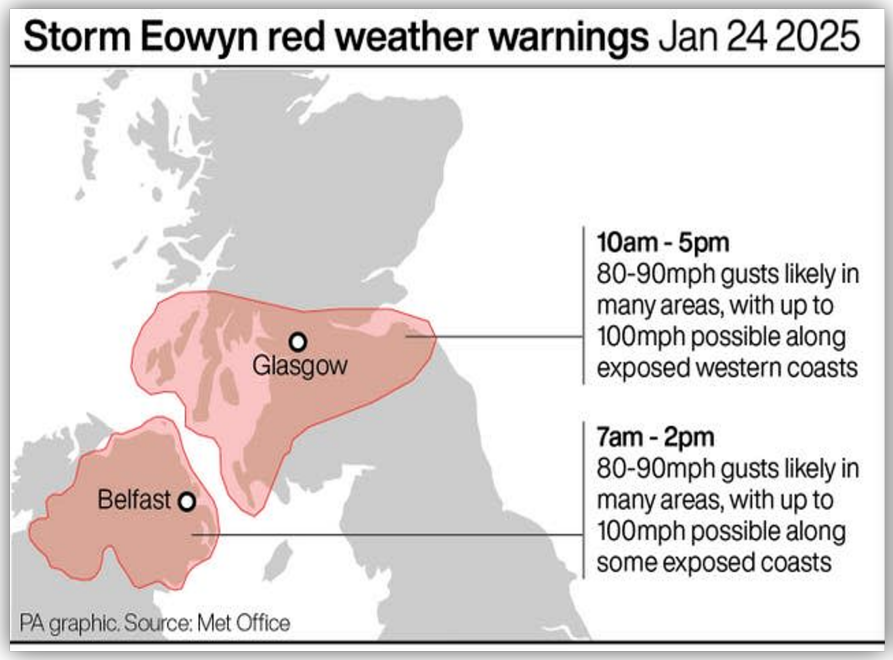
etc. will be over some fixed period of future time.

The only sensible way to do this is to use data on the variable of interest (wind, rain etc.) and fit an appropiate statistical model.

The models themselves are motivated by asymptotic theory, and this is our starting point.

Environmental extremes in our life

Environmental extreme events are closer than you think...



Environmental extremes in our life

Extreme weather expected to cause food price volatility in 2025 after cost of cocoa and coffee doubles

Trend towards more extreme-weather events will continue to hit crop yields and create price spikes, Inverto says

Chocolate being added to cups of coffee. The prices for cocoa and coffee rose 163% and 103% respectively in the year to January. Photograph: Luca Bruno/AP

Extreme weather events are expected to lead to volatile food prices throughout 2025, supply chain analysts have said, after cocoa and coffee prices more than doubled over the past year.

The highest price rises were for cocoa and coffee, up 163% and 103% respectively, due to a combination of higher than average rainfall and temperatures in producing regions, according to the research.

Weather tracker: extreme cold and heavy rainfall batters US

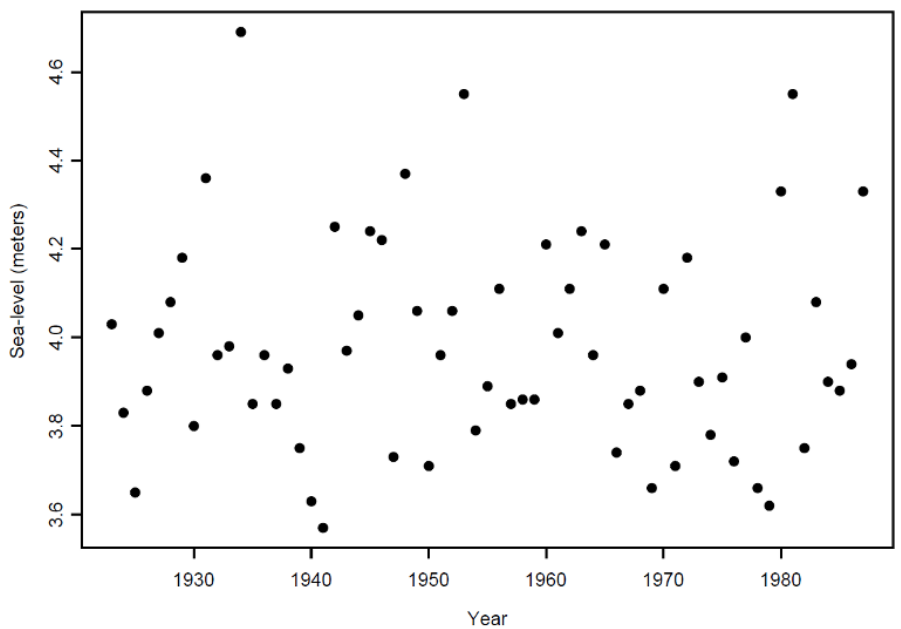
Parts of the Midwest have seen temperatures 15-30C below the climate average, while Australia temperatures hit almost 50C

Snowstorms have caused havoc in states like Michigan. Photograph: Don Campbell/AP

Disruptive weather has continued to affect the US this week, with a mixture of winter hazards, heavy rainfall and extreme temperatures across the country. Extreme cold warnings have affected more than 90 million people, with parts of the midwest seeing temperatures about 15-30C below the climate average.

Defining an extreme

- The first thing we have to consider is what actually represents an *extreme* observation.
- This will vary depending on the context of the dataset.



- We could just look at the biggest (or smallest) values.
- If so, over what time period? Annual maximum? Over the whole dataset? Both?
- Other times it may be all observations above a certain threshold.

Extreme Value theory

Extreme value modelling has a central theoretical result, analagous to the Central Limit Theorem...

- Suppose we have a series of random variables $X_1, \ldots, X_n$, each with cumulative distribution function F, where $F(x) = \mathrm{P}(X \leq x)$.
- We can define the maximum of this set of random variables as $M_n = \max\{X_1, \ldots, X_n\}$.
- Then we can show that $\mathrm{P}(M_n \leq x) = P(X_1 \leq x, \ldots X_n \leq x) = \{F(x)\}^n = F^n(x)$
- The block maxima approach focuses on understanding and estimating this function $F(x)$

Extreme Value Theory

- The true cumulative distribution function, $F(x)$, is unknown.
- We could replace it with an estimate, but any small differences between the estimate and the truth for $F(x)$ could lead to large differences in $F^n(x)$.
- Instead, we tend to focus on the limiting distribution of $F^n(x)$ as $n \to \infty$.
- This is a distribution $G(x)$ such that, for constants, $a_n > 0$ and b_n,

$$\mathrm{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \to G(x) \text{ as } n \to \infty$$

G belongs to the type of one of three distribution functions.

Extreme Value Distributions

- There are three main families of extreme value distribution which have our desired properties as limiting distributions of $F^n(x)$.

  Gumbel: $G(x) = \exp\left(-\exp\left[\frac{-(x-a_n)}{b_n}\right]\right)$

  Frechet: $G(x) = \exp\left(-\left[\frac{(x-a_n)}{b_n}\right]^{-\alpha}\right)$ for $z > b_n$, 0 otherwise.

  Weibull: $G(x) = \exp\left(-\left[\frac{-(x-a_n)}{b_n}\right]^{\alpha}\right)$ for $z > b_n$, 1 otherwise.

- Here, a_n is a location parameter, b_n is a scale parameter and α is a shape parameter.

Block Maxima

Block Maxima

- A lot of investigation of environmental extremes will focus on time series data.
- Within time series data, we typically have natural groupings or blocks of observations (days, months, years etc)
- Therefore a common approach for modelling extremes focuses on the idea of **block maxima** - identifying the maximum (or minimum) value in each block.
- For example, if we have daily temperature data measured over 100 years, we could look at the highest temperature in each year.

Block Maxima approach

- Break up our sequence $X_1, X_2, \ldots$ into blocks of size n (with n reasonably large), and extract only the maximum observation from each block.
- Now we fit $G(x)$ to the sequence of extracted maxima $M_{(1)}, M_{(2)}, \ldots, M_{(N)}$ and use this as the basis for statistical inference.
- For example, consider the annual maxima of daily rainfall. Here our blocks have $n = 365$ observations,which is reasonably large,so we fit our model to N annual maxima (where N is the number of years).
- This rough and ready approach has shown itself to be surprisingly robust

The GEV Distribution

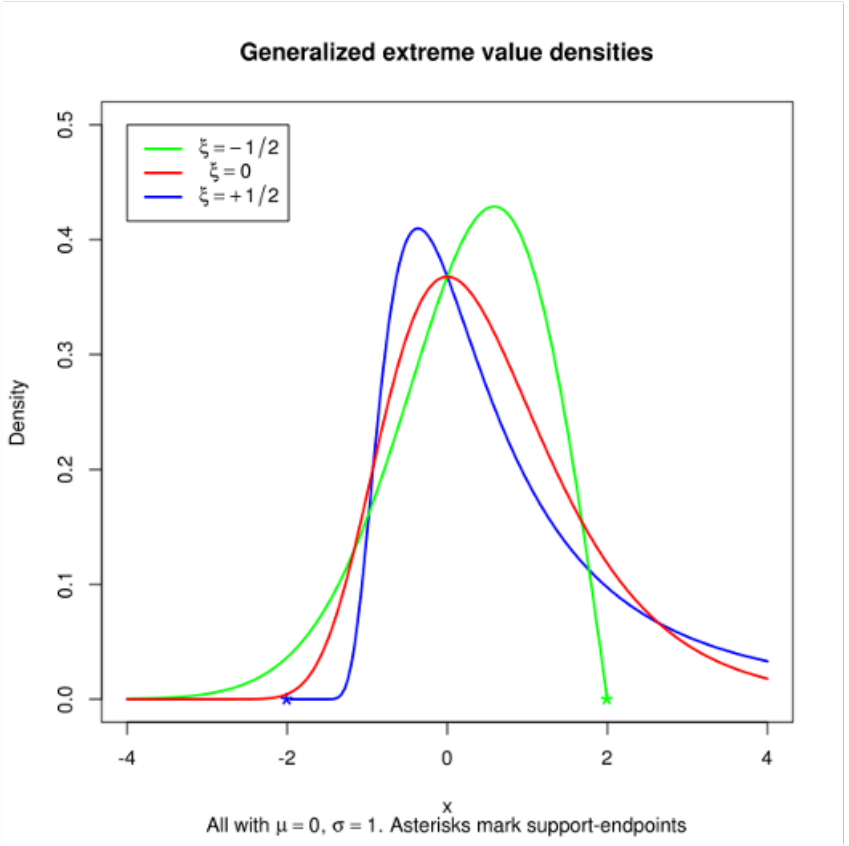
- More generally, we can model the maxima using the Generalised Extreme Value (GEV) distribution

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

- Here, μ is the location parameter, σ is the scale parameter and ξ is the shape parameter.
- The Gumbel, Frechet and Weibull distributions are all special cases depending on the value of ξ.

The GEV Distribution

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



- If $\xi < 0$ then we have the Weibull distribution.
- If $\xi > 0$ then we have the Frechet distribution.
- As $\xi \to 0$ then we have the Gumbel distribution.

Maxima and Minima

- Communication of extremes typically focuses on maxima (or minima).
- Environmental or climate events are often described as the “worst/highest/lowest in X years”.

Europe's drought the worst in 500 years - report

23 August 2022

Europe heatwaves

BBC

A boat trapped in the dried-out shore where the French-Swiss Lac des Brenets lake should be

INDEPENDENT

Bangladesh faces ‘worst flooding in 100 years’ while climate talks in Bonn falter

Experts call the record-breaking floods ‘unprecedented, but not unpredictable’, with millions inundated just as climate talks in Germany end on a disappointing note

Stuti Mishra • Monday 20 June 2022 13:45 • 2 Comments

Return Levels

- In statistics, this idea of the "highest in X years" can be related to the idea of a **return level** and **return period**.
- The return level z_p is the value we would expect to be exceeded once every p years, where $\frac{1}{p}$ is the return period.
- The return level can also be thought of as the value which has probability $\frac{1}{p}$ of being exceeded in a given year.
- Now consider the statement: *"The temperature in Glasgow will reach 20 degrees once every 50 years"*.
- Here, we have return period $\frac{1}{50} = 0.02$ and return level $z_p = 20$.

Return Levels and the GEV

- The return level z_p is the $(1 - \frac{1}{p})$ quantile of the GEV distribution, since we have a probability $\frac{1}{p}$ of the maximum exceeding that value, i.e., $P(M_n > z_p) = \frac{1}{p}$.
- Recall that the GEV takes the form

$$G(z_p) = \begin{cases} \exp\left(-\left[1 + \frac{\xi(z_p - \mu)}{\sigma}\right]^{-\frac{1}{\xi}}\right) & , \xi \neq 0 \\ \exp(-\exp(-\frac{z_p - \mu}{\sigma})) & , \xi = 0 \end{cases}$$

$$G(z_p) = P(M_n \leq z_p) = 1 - P(M_n > z_p)$$

- Therefore the return level can be obtained by inverting this distribution to obtain

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1 - \frac{1}{p})\}^{\xi}\right] & \xi \neq 0 \\ \mu - \sigma \log\{-\log(1 - \frac{1}{p})\} & \xi = 0 \end{cases}$$

Example - Venice Sea Levels

- Sea levels in Venice are rising.
- The city floods on a regular basis.
- What sea level can we expect in the next 5, 10, 100 years?

theguardian

home › world › europe US americas asia australia africa middle eas all

Cities Resilient cities

Inside Venice's bid to hold back the tide

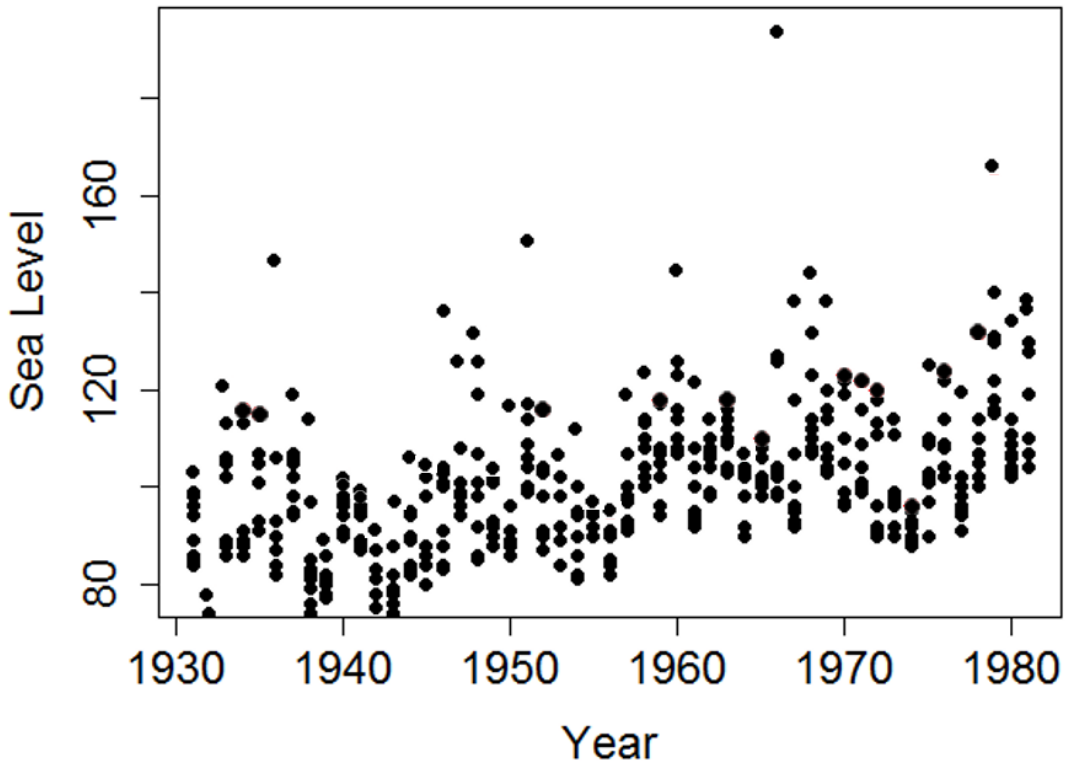
With the city slipping slowly into the lagoon and sea levels continuing to rise, the delayed €5.4bn Mose flood barrier project will be fighting the laws of nature

A history of flooding in the sinking city of Venice - in pictures

A woman walks through Piazza San Marco during a flood. Waters regularly rise above 130cm in the winter. Photograph: Andrea Pattaro/AFP/Getty

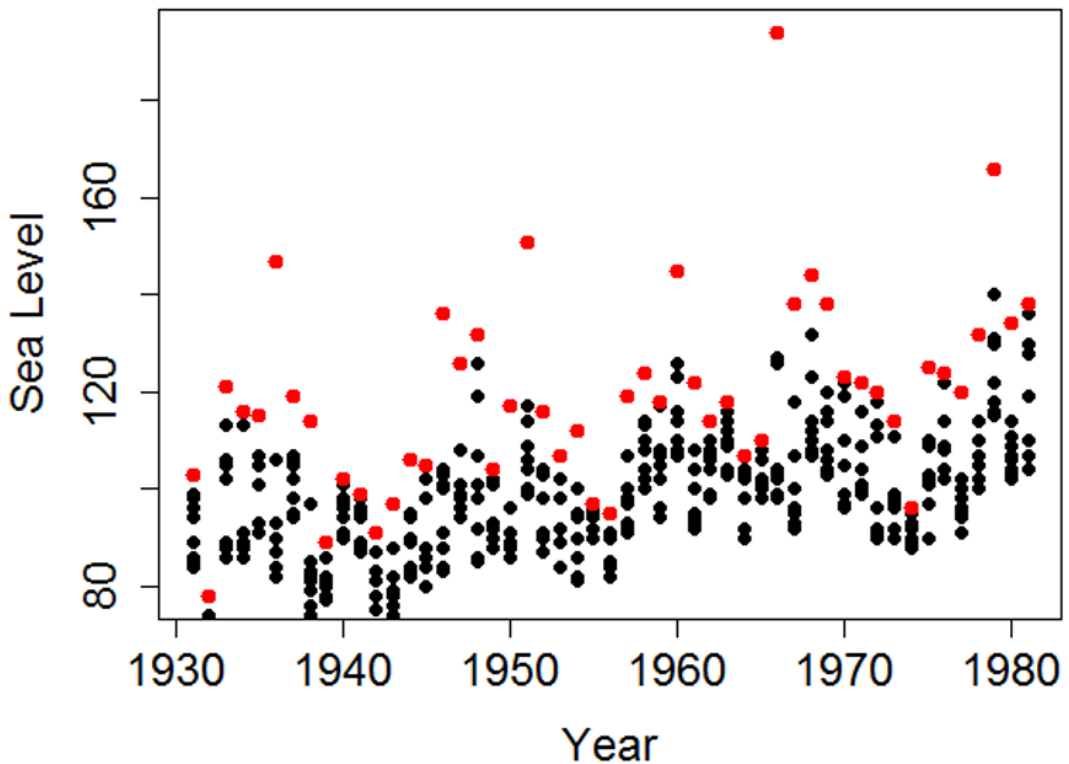
Example - Venice Sea Levels

- We have daily sea level measurements from 1931-1981.
- The plot below shows the 10 highest sea level measurements from each year.



Example - Venice Sea Levels

- We can apply a block maxima approach, treating each year as a block.
- This requires us to identify and model the maximum values every year - highlighted in red.



Fitting the GEV in R

- We use the `ismev` package to fit a GEV distribution in R.
- The `gev.fit()` simply takes a data vector and provides parameter estimates using maximum likelihood esitmation.

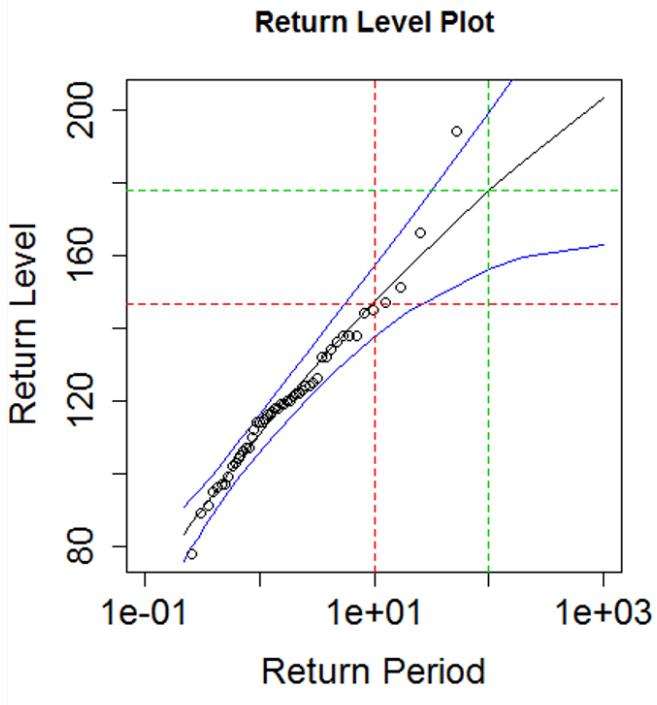
```
mod1 <- gev.fit(venice)

mod1$mle
[1] 111.09925486  17.17548761  -0.07673265
```

- From the output we can see that $\mu = 111.1$, $\sigma = 17.2$ and $\xi = -0.077$.

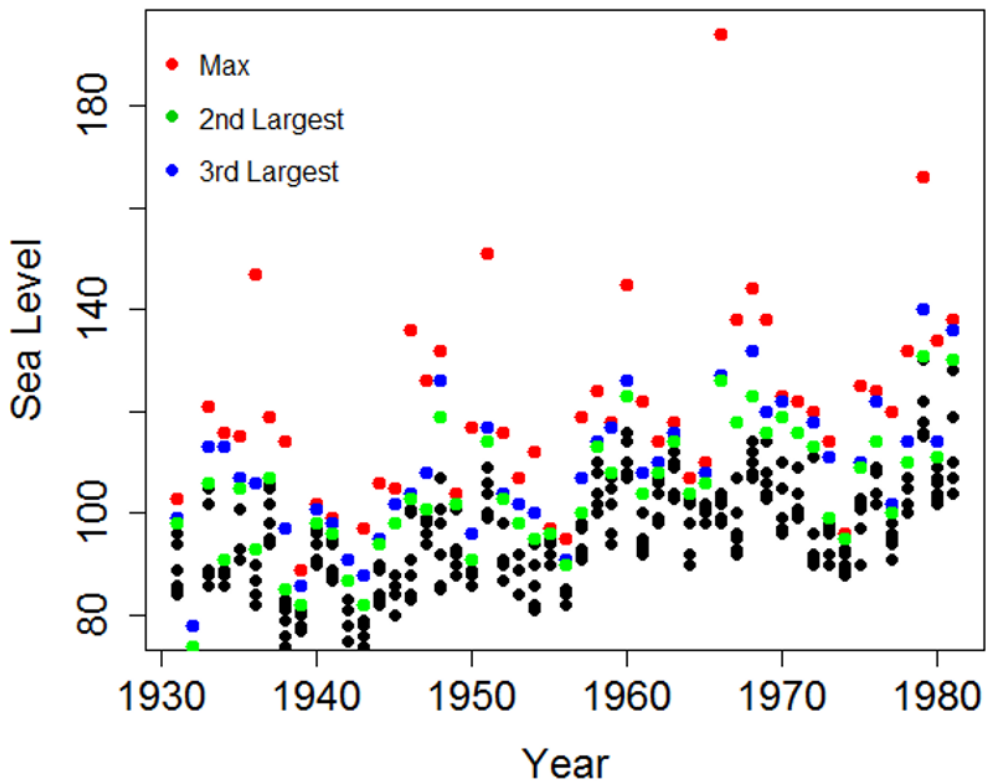
Return Level Plot

- We can assess the suitability of the GEV distribution using a return level plot.
- This involves commuting the return level at various return periods, and comparing it to the theoretical result under the GEV.
- The observed points lie along the theoretical line.
- Our proposed GEV distribution is appropriate.
- The Venice maxima follow a Gumbel distribution.



Extending Block Maxima - *k*-largest

- Looking at just the maxima means we throw away a lot of data, making it harder to accurately estimate parameters.
- We could instead follow an approach which models the *k* largest values within a block.



Choice of block size

- We have to make difficult subjective choices when fitting a block maxima model.
- What do we choose as our block? Week? Month? Year? Decade?
- Bigger blocks means we have fewer data points, but smaller blocks mean our ‘extremes’ might not be extreme at all, violating the assumptions of the GEV distribution.
- If we use a k-largest approach, we have a similar decision on what value of k to choose.

Peak Over Threshold

Drawbacks of Block Maxima

- Block maxima methods work well in many situations, and take advantage of natural blocks in the data.
- However, it does not work well if there is a lot of variability between blocks.
- In that scenario, some blocks may have many more large counts than others, and much of the data will be discarded.
- We can overcome this by using a threshold approach, which models all observations which exceed some pre-specified value.

Peak Over Threshold

- This approach is known as **peak over threshold (POT)** modelling .
- Again, we assume we have data represented by a time series, and some threshold u.
- We need a statistical model for the values which are above u, known as **exceedances**.
- Sometimes we may also wish to model the *number* of exceedances.

Peak Over Threshold

- Again, let $X_1, \ldots, X_n$ be a sequence of independent random variables with a common distribution function F.
- We can consider our extreme values in terms of their **threshold excess** (how much they exceed the threshold by).
- For an extreme value $X > u$, its threshold excess is given as $y = X - u$.
- The probability of threshold excess of size y is given by

$$\mathrm{P}(X > u + y | X > u) = \frac{1 - F(u + y)}{1 - F(u)} \quad \text{where } y > 0.$$

Generalised Pareto

- The function F is still unknown, but the distribution of all threshold excesses can be approximated by a **Generalised Pareto distribution (GPD)**.
- The cdf of the Generalised Pareto distribution is given by

$$G(y) = \begin{cases} 1 - \left(1 + \frac{\xi(y-\mu)}{\sigma}\right)^{-\frac{1}{\xi}} & \xi \neq 0 \\ 1 - \exp\left(-\frac{y-\mu}{\sigma}\right) & \xi = 0 \end{cases}$$

- Again, μ is the location parameter, σ is the scale parameter and ξ is the shape parameter.

Return Levels for POT models

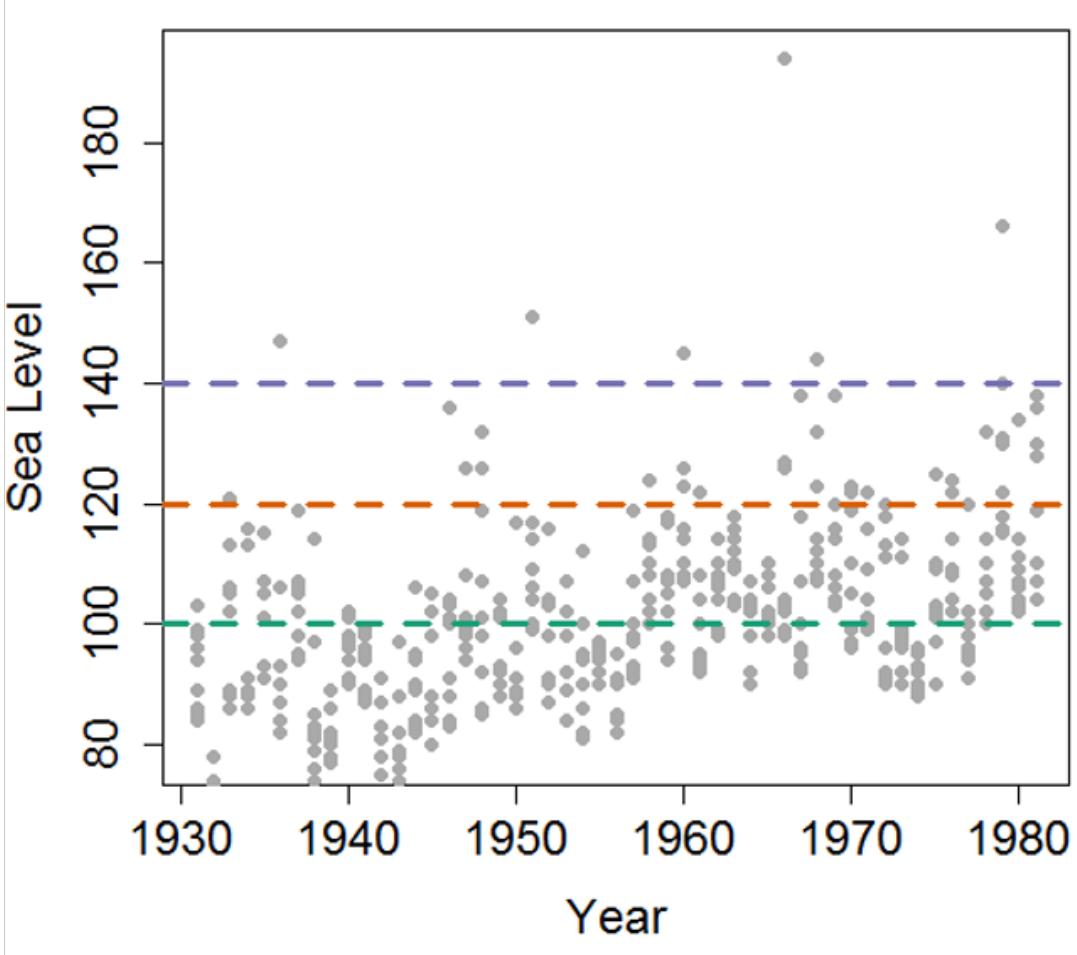
- We can define a return level for POT models in a roughly similar way to block maxima models.
- The m-observation return level, x_m is defined as the level expected to be exceeded once every m observations, with

$$x_m = \begin{cases} u + \frac{\sigma}{\xi}\left[(m\mathsf{P}(X > u))^{\xi} - 1\right] & \xi \neq 0 \\ u + \sigma \log\left(m\mathsf{P}(X > u)\right) & \xi = 0 \end{cases}$$

- For any given observation, the probability of exceeding x_m is simply $\frac{1}{m}$.

Choosing a threshold

- We need a threshold low enough that we have sufficient data, but high enough that values above it are genuinely extreme.

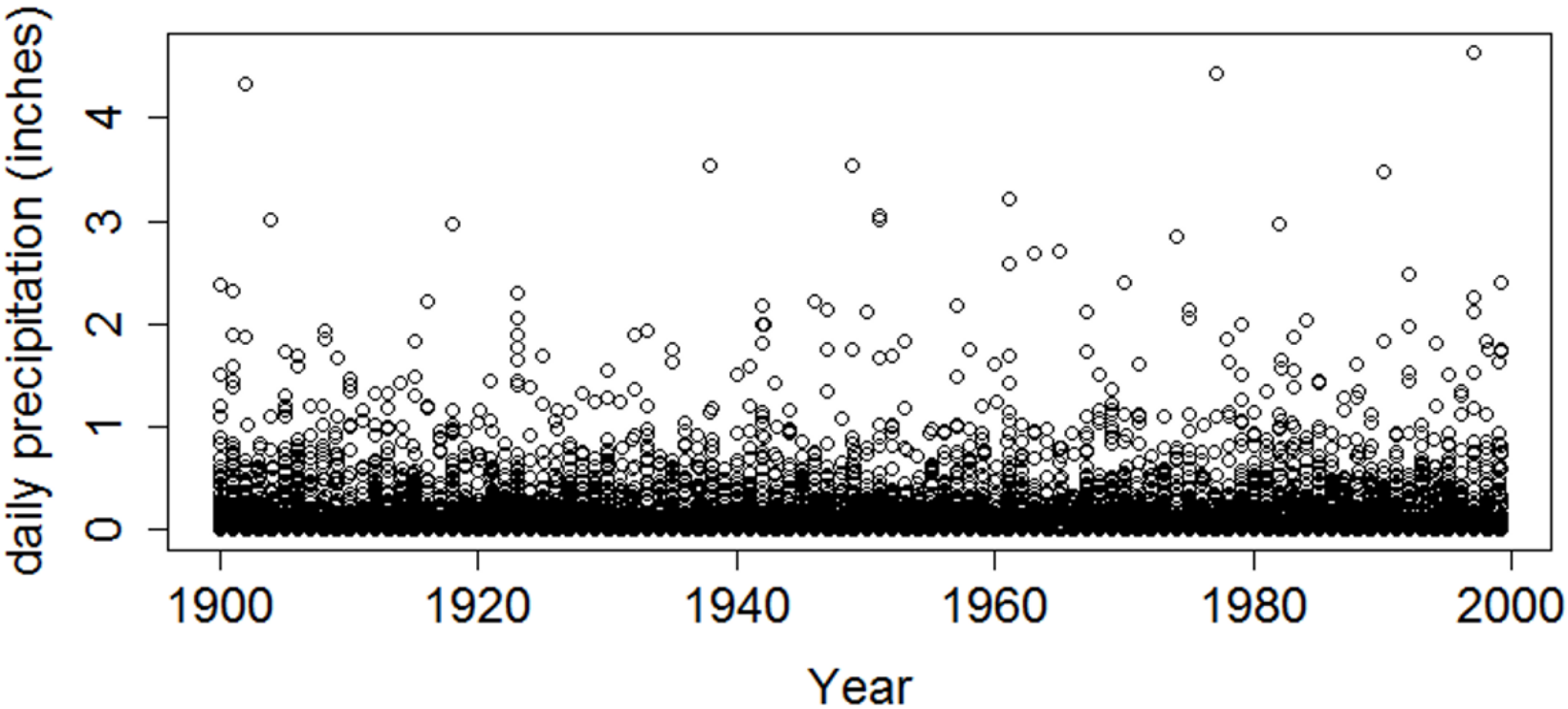


Choosing a threshold

- Occasionally there is a natural choice of threshold (eg a legal limit for a pollutant), but generally we need to choose it.
- One approach is to use a **mean residual life plot**, which plots the sample mean excess (mean of $x > u$) at a variety of thresholds u.
- If the GPD is appropriate, the mean excess should be linearly related to the threshold.
- Therefore, we can identify a suitable threshold as one which lies with an area of linearity on this plot.

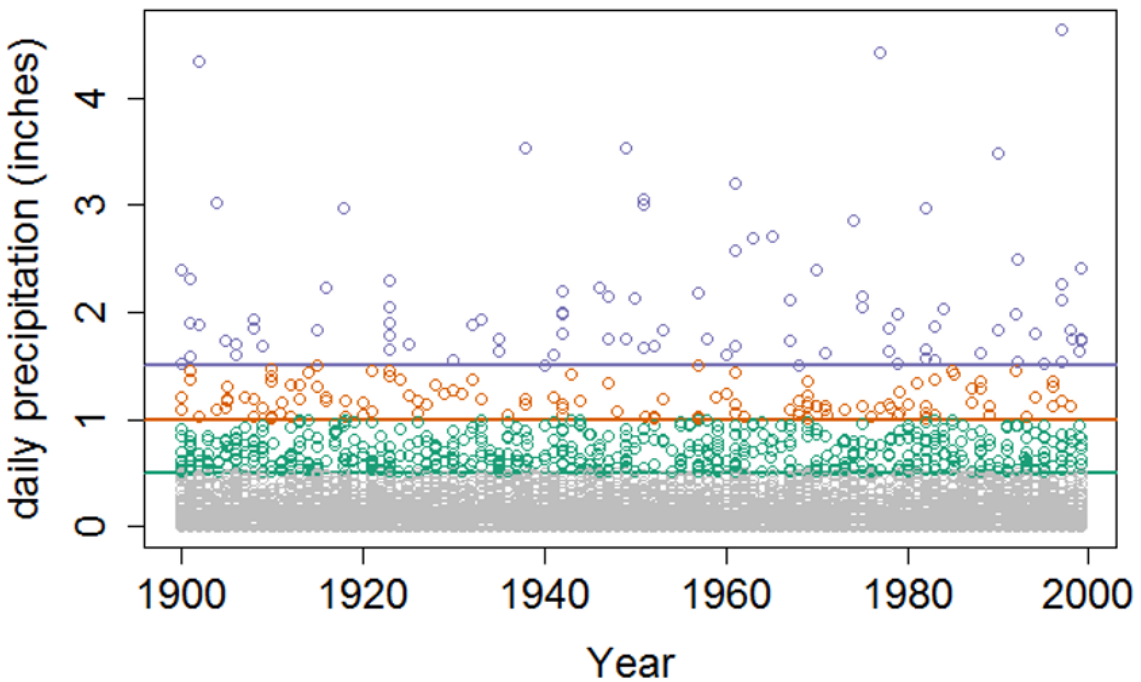
Example - Fort Collins, Colorado

- We have daily precipitation data from 1900-1999, obtained from a rain gauge in Fort Collins, Colorado, taken from Katz et al, 2002.



Example - Fort Collins, Colorado

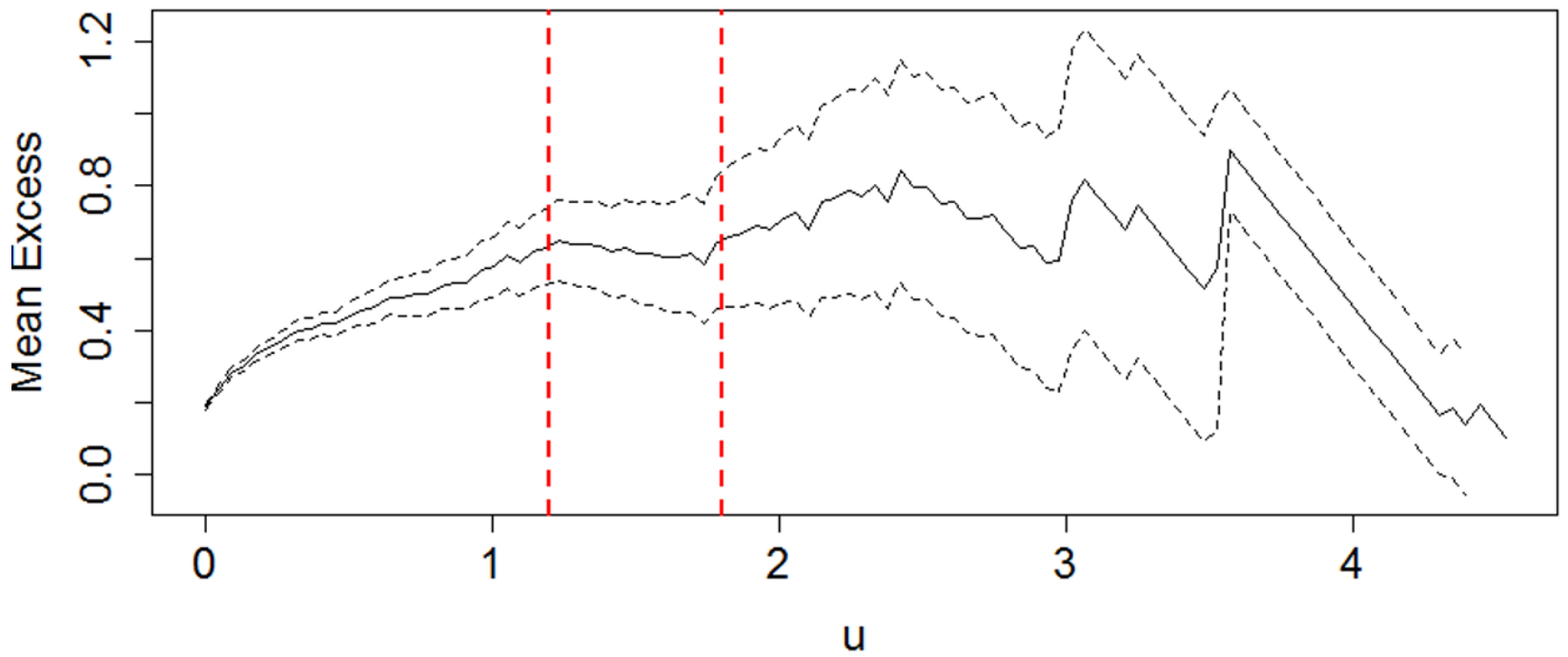
- We compare three different choices of threshold below ($u = 0.5, 1.0, 1.5$) to show the importance of getting the choice right.



u	$\% > u$	$n > u$
0.5	2.08	759
1.0	0.58	213
1.5	0.25	91

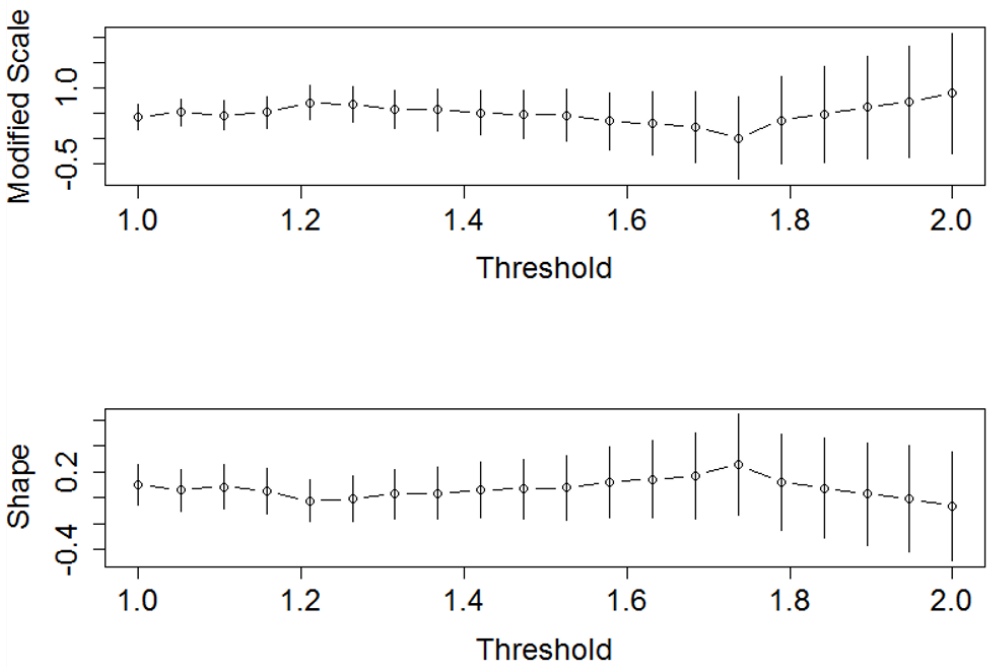
Mean Residual Life Plot

- We can fit a mean residual life plot to identify a sensible choice of threshold.
- It appears that a value of u somewhere between 1.2 and 1.8 would be an appropriate choice here - this is where the plot appears to be linear.



Sensitivity Analysis

- We can also carry out a sensitivity analysis to see the effect of choosing different threshold values on the estimated model parameters.
- The plot below shows the parameter estimates at different thresholds - they seem fairly robust.



POT models in R

- We can use the `extRemes` package to fit a Generalised Pareto distribution in R.
- The function `fevd` allows several extreme value distributions (including GEV and GPD) to be fitted, and can also provide return levels.

```
fitGP <- fevd(Fort, threshold=1.5, type="GP",
                    time.units="365/year")

return.level(fitGP, c(10,20,100), type="GP")

[1] "Return Levels for period units in years"
 10-year level  20-year level 100-year level
      2.857184       3.340219       4.581339
```

Correlation

- Threshold exceedances are not always independent due to temporal correlation.
- If we have high temperatures today, it's likely we might also have high temperatures tomorrow.
- We have to account for this dependence within our model, for example by using the ARIMA approaches outlined in the time series section.
- Alternatively, we could use a 'declustering' approach which identifies these temporal clusters and simply uses the cluster maxima.

Summary

- Estimating extremes is challenging and the results can be unreliable even with large datasets.
- We have identified two main approaches for dealing with these data - block maxima and point over threshold.
- The block maxima approach uses the Generalised Extreme Value (GEV) distribution to model the highest value(s) within a specific block of time.
- The point over threshold approach uses the Generalised Pareto distribution (GPD) to model all obsevations which exceed a certain value.