Hydroinformatics for Hydrology: Extreme Value Modelling

Hugo Winter

EDF Energy UK R&D Centre

9th April 2018

Outline

Table of Contents

Aims of this session

1. Provide an introduction to extreme value theory.
2. Look at some hot topics in the academic field of extreme value theory and try to understand how these approaches can be applied.
3. To outline some common pitfalls when undertaking an extreme value analysis.
4. To provide you with a set of packages and functions that can be used to undertake and extreme value analysis.

My background

- I'm the Natural Hazards & Environment manager at the EDF Energy R&D UK Centre.
- Previously worked from 2015-2017 as a researcher on extreme weather and coastal flooding.
- My PhD was titled Extreme value modelling of heatwaves and was jointly supervised by Jonathan Tawn at Lancaster University and Simon Brown at the Met Office.
- Main skills used during PhD were extreme value statistics applied to environmental problems.

Table of Contents

What is extreme value analysis?

- A statistical approach for analysing extreme data values of a variable of interest.
- First mentioned in 1928 by Fisher and Tippett.
- Formalised into statistical methods by Gumbel in paper in 1958.
- Use for environmental problems introduced in the 1950's.

Left: Roland Fisher; Centre: Leonard Tippett; Right: Emil Gumbel

Why use extreme value analysis?

- Provides a mathematically rigorous framework for modelling extreme values.
- Data are by definition sparse.
- Empirical approaches based upon the observed data an only provide accurate results within the range of the observed data $\rightarrow$ we often wish to extrapolate to higher levels.
- Different statistical models can lead to different tail behaviours $\rightarrow$ can often be too light-tailed and underestimate the probability of extreme events.
- Many statistical models are driven by average values as opposed to extreme values.

When to use EVA?

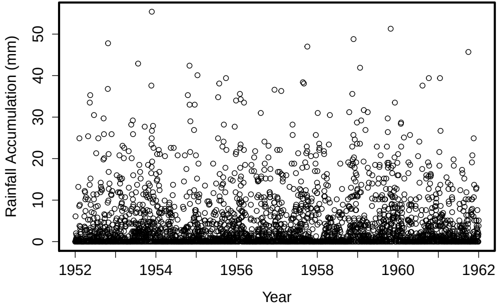
- When the variable of interest is stochastic (as opposed to deterministic) $\rightarrow$ e.g. storm surge ✓, tide ✗.
- When physical models are unavailable or unrealistic.
- When interested in obtaining estimates for extreme quantities that lie outside the range of the observed data.
- When there are at least 20-30 years of observations.

Modelling univariate extreme values

- Two main approaches exist for modelling univariate (one-dimensional) extreme values:
  - Block maxima
  - Threshold exceedances
- Block maxima methods were first to be developed.
- Threshold exceedance methods are most commonly used now.
- Modelling strategies for both assume observations are independent and identically distributed (iid).

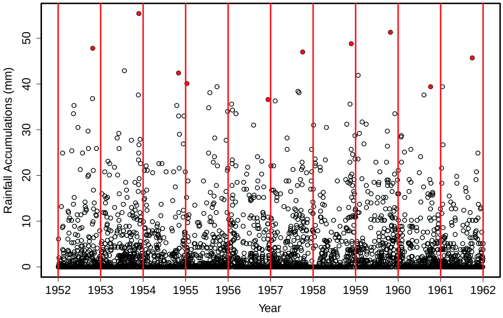
My recurring rainfall data example

- Daily rainfall accumulations at a location in south-west England recorded over the period 1914-1962.
- Data taken from Coles (2001) and freely available in `ismev` package in `R`.



Block maxima

- Model the maxima of time periods of a certain length.
- Annual maxima often taken to remove the effect of seasonality.



Generalized extreme value distribution

Let $M_1, \ldots, M_n$ be random variables for the cluster maxima from n time blocks. The generalized extreme value (GEV) distribution can be used to model these maxima such that

$$G(x) = \mathrm{P}(M \leq x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{1/\xi}\right\},$$

for $1 + \xi(x - \mu/\sigma) > 0$ where

- $\mu \in (-\infty, \infty)$ is the **location** parameter
- $\sigma \in [0, \infty)$ is the **scale** parameter
- $\xi \in (-\infty, \infty)$ is the **shape** parameter

More detail on parameters

- The shape parameter ξ is a very important parameter in EVA.
- Controls the heaviness of the tail $\Rightarrow$ directly affects the extremes.
- The shape parameter of the GEV covers three different types of tail behaviour:
  - $\xi > 0$ - Fréchet distribution $\rightarrow$ Heavy upper tail
  - $\xi < 0$ - Negative Weibull distribution $\rightarrow$ Bounded upper tail
  - $\xi = 0$ - Gumbel distribution $\rightarrow$ Exponential upper tail

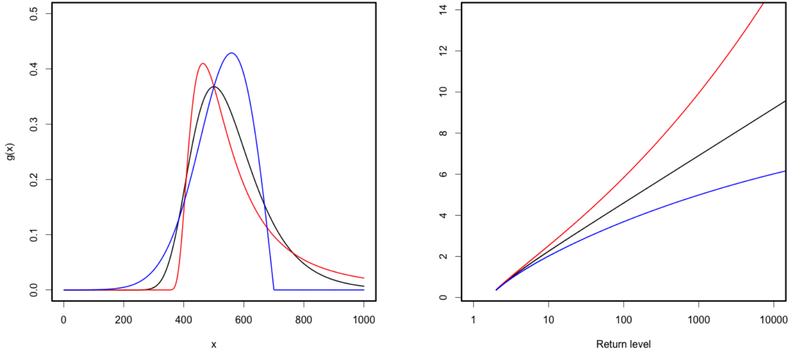
Return levels

- We are most interested in estimating the severity of extreme events.
- One way to summarise this is in terms of the T**-year return level** z_T.
- This is the event that happens once in every T years (i.e. has annual exceedance probability $1/T$).

For the GEV distribution fitted to annual maxima

$$z_T = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1 - 1/T)\}^{-\xi}\right] & \text{if} \quad \xi \neq 0 \\ \mu - \sigma \log\{-\log(1 - 1/T)\} & \text{if} \quad \xi = 0. \end{cases}$$

Effect of different shape parameters

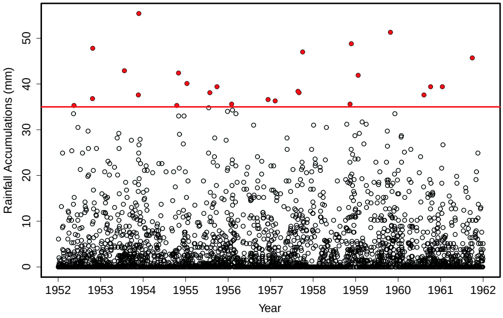


Left: GEV distribution function **Right:** Return level curves

Black: $\xi = 0$ **Red:** $\xi > 0$ **Blue:** $\xi < 0$

Threshold exceedances

- Model exceedances above a fixed high threshold.
- More efficient as more data are available but not necessarily independent.



Generalized Pareto distribution

Let $X_1, \ldots, X_n$ be a sequence of random variables. The distribution G of the exceedances above a high threshold u is a generalized Pareto distribution (Davison & Smith 1990) of the form

$$G(x) = \mathrm{P}(X \leq x \mid X > u) = 1 - \left(1 + \xi \frac{x - u}{\sigma_u}\right)_+^{-1/\xi},$$

for $x > u$ where

- $\sigma_u \in [0, \infty)$ is the **scale** parameter
- $\xi \in (-\infty, \infty)$ is the **shape** parameter

More detail on parameters

- The scale parameter for the GPD is threshold-dependent. For any higher threshold $v > u$

$$\sigma_v = \sigma_u + \xi(v - u)$$

- The scale parameter can be modified to make it threshold invariant.
- The shape parameter of the GPD is equal to the shape parameter of the corresponding GEV distribution.
- Threshold choice is important.
- Equivalent formulation is available in terms of Poisson process (REF)

Choosing the threshold

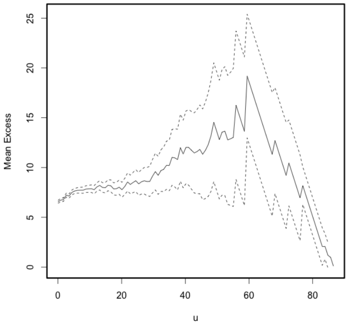
- A bias variance trade-off exists when choosing the threshold.
- We wish to set the threshold low to use as many data points as possible in our analysis.
- Need the threshold set high enough for underlying limit assumptions of EV model to hold.

- **Threshold too high** ⇒ not enough data, high uncertainty.
- **Threshold too low** ⇒ non extreme data modelled, model not suitable.

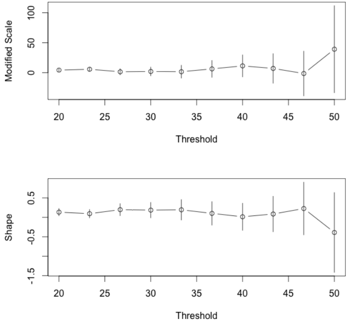
Choosing the threshold

- Two standard diagnostics exist for threshold choice:
  - Mean residual life (MRL) plot.
  - Parameter stability plots.

MRL plot



Parameter stability plots



- Other approaches exists (e.g. Northrop et al. (2016))

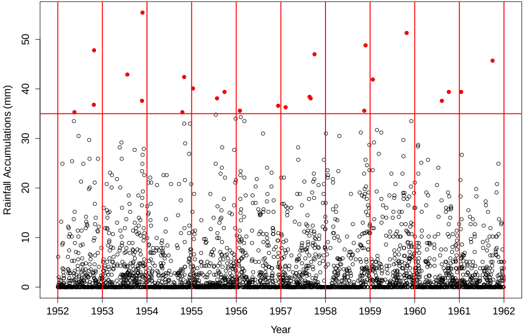
Return levels

- Calculated in a similar way as for the GEV distribution.
- Since data are conditional upon having exceeded a high threshold we need to undo this conditioning by multiplying by $\lambda_u = \mathrm{P}(X > u)$.
- The m-observation return level is given below

$$z_m = \begin{cases} u + \sigma_u/\xi \left[(m\lambda_u)^\xi - 1\right] & \text{if} \quad \xi \neq 0 \\ u + \sigma_u \log\left(m\lambda_u\right) & \text{if} \quad \xi = 0, \end{cases}$$

where m must be sufficiently large to ensure that $x_m > u$. If n_T is defined as the number of observations in a year then $T = m/n_T$.

Comparison of approaches



- Threshold exceedance approach allows more data to be used in an analysis.
- Clustering may occur so independence assumption may not always hold.

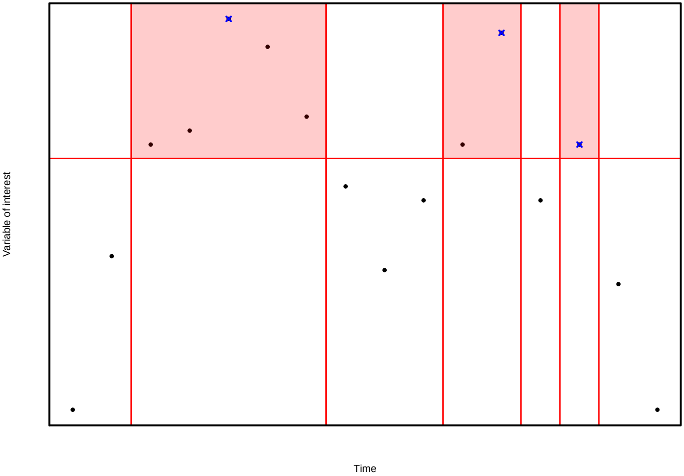
Declustering

- When fitting an extreme value distribution to data an important assumption made is that data are independent and identically distributed (IID).
- When using block maxima (for a sufficient block length) this is satisfied.
- This could be an issue for threshold exceedances tend to occur in clusters.
- If we model using all the exceedances it is likely that we will be overconfident and our confidence intervals will be too narrow.
- To solve this we usually undertake declustering to extract the peaks over the threshold (POTs).

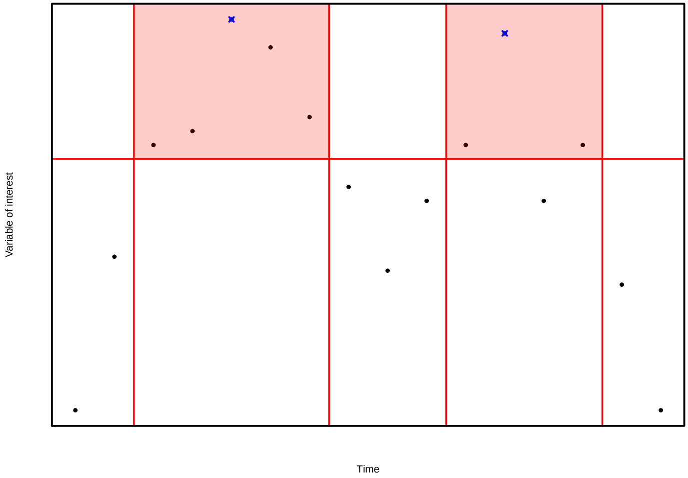
Runs method declustering

- The most well used declustering approach is called the runs method.
- A cluster is commenced by an exceedance of a threshold u.
- A cluster lasts until we have observed R consecutive observations below u.
- R is known as the run length.
- The peak of each cluster is extracted and a statistical model is fitted to the POTs.

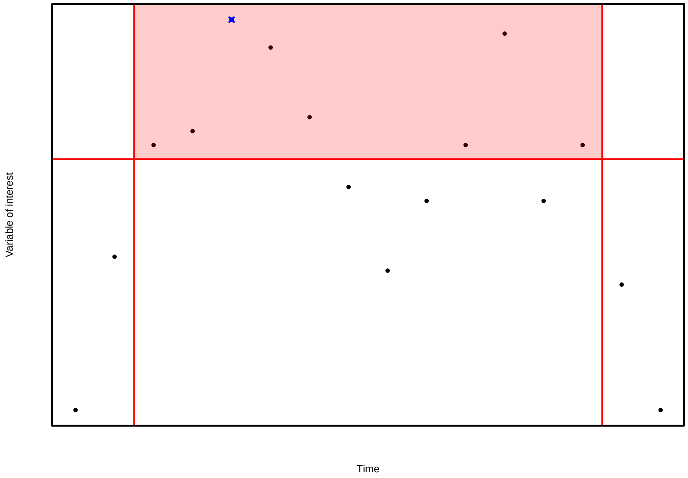
Runs method - run length 1



Runs method - run length 2



Runs method - run length 4



Declustering

- The number of clusters we define is dependent on the run length.
- In some situations the run length can be defined by the environmental context - i.e. if we expect a certain type of event to last around 6 observations this may suggest $R = 6$.
- Other methods are available including an automatic approach called intervals declustering (Ferro & Segers 2003).

Inference for univariate extreme value models

- Many different approaches exist for fitting both types of extreme value model:
  - Maximum likelihood → most commonly used.
  - L-moments → faster in certain situations.
  - Bayesian methods → modern approach, more computationally expensive.
- Many packages exist in `R` to fit extreme value models:
  - `evd` - Basic functions for an EVA
  - `extRemes` - Slightly more advanced set of functions
  - `ismev` - Companion package to Coles (2001)
  - `POT` - Peaks over threshold modelling
  - `evir` - More basic functions for an EVA

Confidence intervals

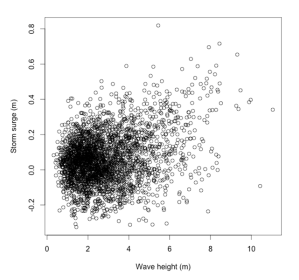
- Confidence intervals can be obtained in several ways:
  - Delta method
  - Profile likelihood
  - Bootstrapping (parametric and non-parametric)
- When looking at extreme quantities these intervals can get quite wide - this motivates methods for pooling data to obtain narrower intervals.
- Often profile likelihood or bootstrap intervals are preferred.

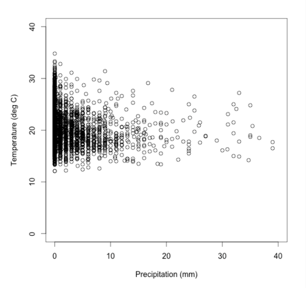
Table of Contents

Multivariate EVA - motivation

- Univariate EVA is a useful tool when considering the extremes of a single variable.
- Using univariate EVA we can provide estimates of return levels for one variable at a single site.
- In many situations we are interested in the probability of two or more hazards occurring at the same time or extremes at maore than one location.
- Can also use multivariate EVA to analyse:
  - Extremes occurring at multiple locations.
  - Extremes that persist through time.

Examples of combinations





Left Wave and surge height (m) measurements at a single location off the south-west coast of England.

Right Daily maximum temperatures (^{o}C) and rainfall (mm) at Waddington in east of UK - only looking at summer days (June-August) 1949-2015.

Summarising dependence

- We need measures to understand whether variables are dependent.
- The most common measure is the **correlation** coefficient ρ.
  - Positive - both variables increase together
  - Negative - as one variable increases, the other decreases.
- But estimation of the correlation is driven by non-extreme data.
- We need to define a different measure of **extremal dependence**.

Summarising extremal dependence

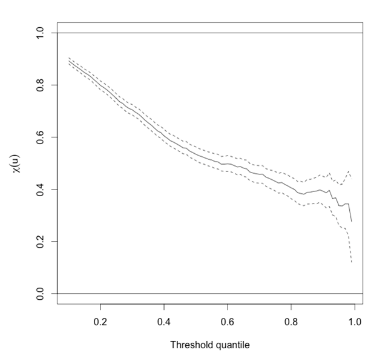
- A commonly used measure in multivariate EVA is the **extremal dependence measure** $\chi(u)$ (Coles et al. 1999).
- For two variables (X, Y) with a sufficiently high threshold u

$$\chi(u) = \mathrm{P}(Y > u \mid X > u)$$

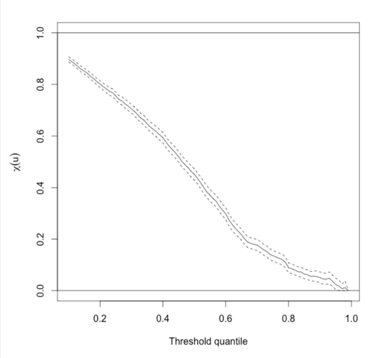
- Different values of $\chi(u)$ are associated with different levels of dependence
  - $\chi(u) = 1 \Rightarrow$ perfect dependence
  - $\chi(u) = \mathrm{P}(Y > u) \xrightarrow{u \text{ large}} 0 \Rightarrow$ independence
- $\chi(u)$ allows us to understand about the relationship between variables at high levels.

For the data sets introduced previously

Wave and surge



Rainfall and temperature

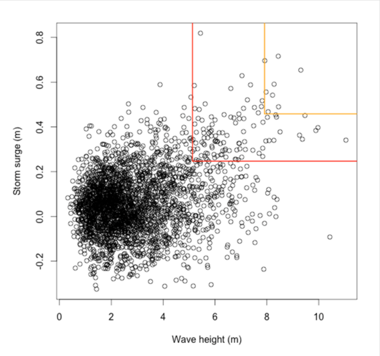


MVE models

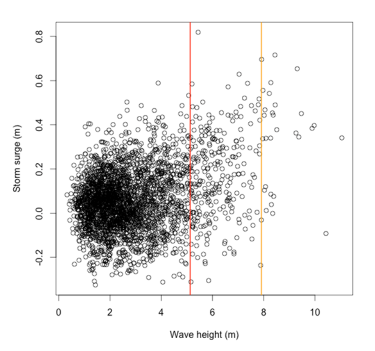
- Previous analysis gives estimates of joint probability within the scope of the data.
- Need to reliably estimate extremal dependence at higher levels.
- Different options for multivariate EVA
  - (i) **Copulas** - large class of models, scale to higher dimensions poorly (Nelson 2007).
  - (ii) **Joint tail model** - more flexible than copulas, scales to higher dimensions poorly (Ledford & Tawn 1997).
  - (iii) **Conditional extremes approach** - flexible, scales well (Heffernan & Tawn 2004).

Pictorial representation of MVE models

(i) and (ii)



(iii)



Red: Modelling threshold **Orange:** Critical level

Available packages in R

- Copulas:
  - `copula` - Fit a wide selection of different copulas. Contains functions to fit extreme value copulas.
  - `evd` - Fit bivariate extreme value distributions with different dependence structures.
- Joint tail model: No R packages explicitly for this.
- Conditional extremes:
  - `texmex` - Fit the conditional extremes approach in multivariate setting.

Results for data examples

- Set v as the critical level associated with the 10^{-4} annual exceedance probability
  - $\chi(v) = 0.01$ for wave and surge data
  - $\chi(v) = 0$ for rainfall and temperature data
- With conditional extremes model can derive conditional quantiles, i.e. Surge|Wave$> v$ or Temp|Rain$> v$
  - 0.759m (0.522, 1.040) for wave and surge data
  - 18.2°C (15.1, 21.2) for rainfall and temperature data
- Many other measures can be obtained using the conditional extremes approach.

Multivariate EVA - summary

- Multivariate extreme value models can be used to model hazard combinations - usage depends on factors such as data availability.
- Can be seen that data can have different behaviour in the extremes than at lower levels.
- Different choices are available conditional extremes models provide the most flexible choice.
- These approaches can be used to model dependence over space and time.

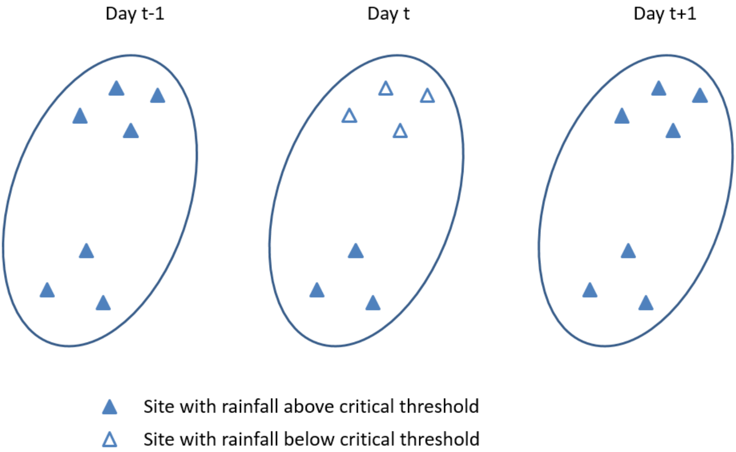
Spatial pooling - Motivation

- Univariate EVA methods are well used by academics and industry practitioners.
- As just presented, the extension to multivariate EVA is now well established in academic literature and starting to be used more in applied contexts.
- But how can we improve our methods for estimating return levels?
- Here, we shall briefly focus on two areas:
  - Regional frequency analysis - Hosking & Wallis (2005), Weiss et al. (2014)
  - Spatial Bayesian extreme value models - Reich et al. (2014).

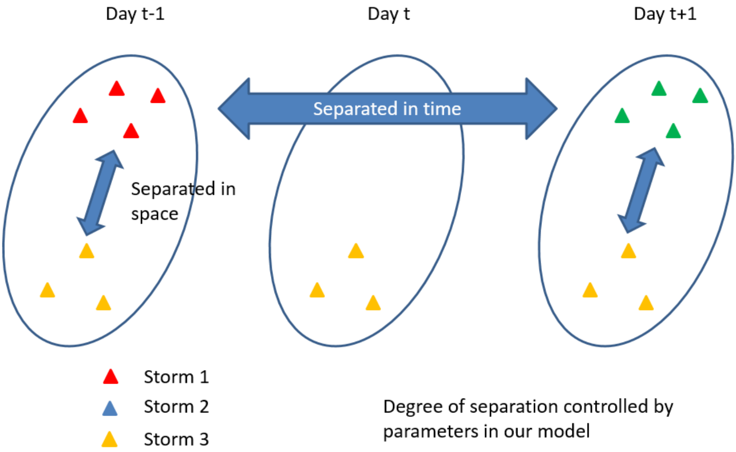
RFA Motivation

- The estimation of return levels is usually still undertaken using univariate EVA methods, i.e. we are only using data from a single site.
- In practice, we will have a lot more data available to us. Any chance to use more data should also be encouraged.
- There may also be situations where localised extreme events striking a region by chance miss a measurement gauge $\Rightarrow$ our single-site model may not take account of such an event!
- Is there a better way to use the available information to improve return level estimates?

RFA methodology - storm selection



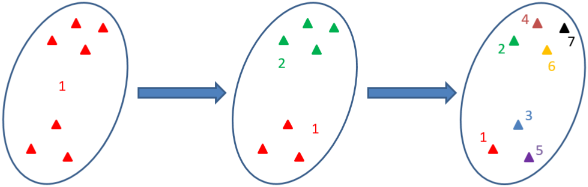
RFA methodology - storm selection



Degree of separation controlled by parameters in our model

RFA methodology - homogeneous region selection

- We now have a set of storms defined, need to convert this to a set of homogeneous regions.
- To do this we estimate the probability of two sites being affected by the same storm and repeat for all pairs of sites.
- This set of probabilities can be fed into a Hierarchical Agglomerative Clustering (HAC) algorithm to suggest a partition of the sites.



RFA methodology - statistical technique

- Within each region, the most extreme storms are extracted by applying a second thresholding.
- This threshold u_i often chosen to model on average 1 exceedance at each site per year, needs to be chosen high enough to ensure dependency of storms.
- Within each region, the chosen exceedances are normalised (i.e. $Y_i = X_i/u_i$) and we fit a Generalised Pareto Distribution to all the exceedances within the region, i.e.

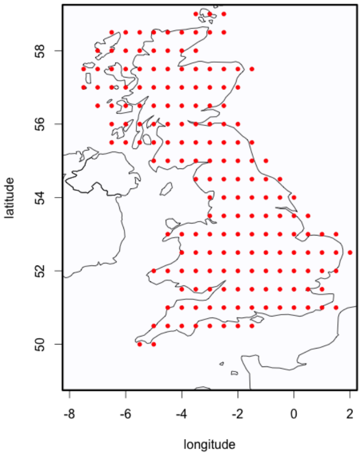
$$Y_i \mid Y_i > 1 \sim \text{GPD}(\gamma, k)$$

- The distribution at the local site scale can be recovered and used to estimate extreme return levels, i.e.

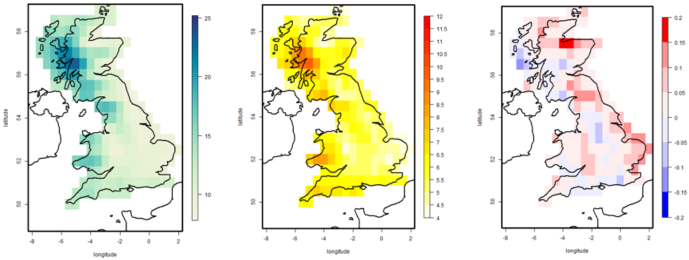
$$X_i \mid X_i > u_i \sim \text{GPD}(u_i\gamma, k)$$

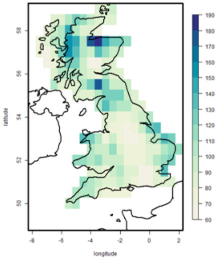
Rainfall example

- Data are taken from the Climate Forecast Services Reanalysis (CFSR) which are available from 1979-2016.
- Hourly data are available for a number of different variables - here hourly rainfall is accumulated to the daily scale.
- For this study the 0.5° grid was chosen, to provide sufficient coverage across the UK.



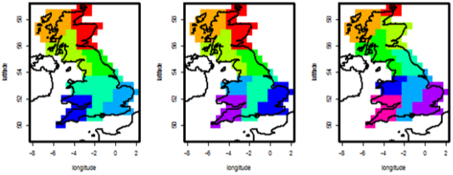
Rainfall example - Univariate EVA results

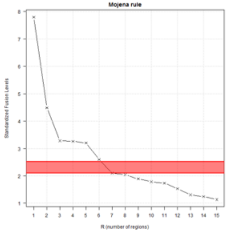




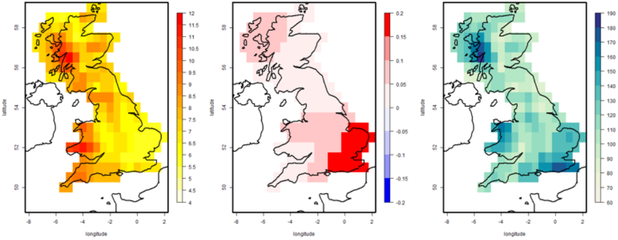
Top left: Threshold fixed at 95th quantile
Top centre: Estimates of scale parameter
Top right: Estimates of shape parameter
Bottom left: Estimates of 10,000-year return level

Rainfall example - RFA results





Different potential region choices (above) with Mojena rule for region selection (right)



Estimates of:

- Scale parameter (far left)
- Shape parameter (middle)
- 10,000-year return level (far right)

Discussion and Conclusion on RFA

- The range of return level estimates obtained using both approaches is reasonably consistent.
- We observe an improvement in the spatial consistency of the estimates when using RFA.
- However, There are also additional subjective steps in the methodology which can induce uncertainty.
- There are also some issues near the boundaries of homogeneous regions.
- Is there a way to create a smoother model across space?

Bayesian models motivation

- The use of Bayesian extreme value theory can help us to create a more physically consistent model.
- Bayesian statistics is a branch of statistics named after the Reverend Thomas Bayes (1701-1761).
- This area of statistics revolves around Bayes' Theorem:

$$\mathrm{P}(\theta \mid x) = \frac{\mathrm{P}(\theta, x)}{\mathrm{P}(x)} = \frac{\mathrm{P}(x \mid \theta)\mathrm{P}(\theta)}{\mathrm{P}(x)}$$

- $\mathrm{P}(\theta \mid x)$ is called the *posterior distribution*, $\mathrm{P}(x \mid \theta)$ is the *likelihood* and $\mathrm{P}(x)$ and $\mathrm{P}(\theta)$ are *prior distributions*.

What does Bayesian statistics mean in practice?

- Previously, we assumed that our data were one realisation of a sequence generated from the random variable X with distribution F.
- This realisation was then used to estimate a 'best set' of model parameters with uncertainty bounds to allow for the sampling uncertainty.
- Now, we have a belief in our parameter θ expressed through prior $\mathrm{P}(\theta)$ and we use the observed data x to update this belief to generate a posterior $\mathrm{P}(\theta \mid x)$.
- In our context, this posterior distribution can then be used to generate a whole distribution of return level estimates.

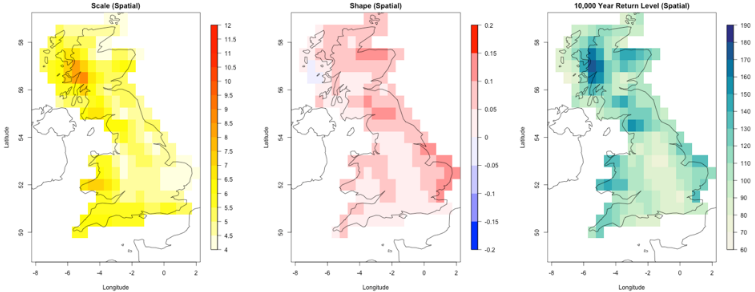
How can we use Bayesian statistics to improve our RFA results?

- Now we not only model site-by-site using a GPD, but include a spatial hierarchical model to borrow strength across locations.
- This model assumes an underlying spatial process in the means of the GPD parameters $\boldsymbol{\theta}_j = (\log(\sigma_u), \xi)$ and an underlying spatial process ϕ_j, i.e.

$$\boldsymbol{\theta}_j \sim \mathrm{N}(\mathbf{X}_j\beta + \phi_j, T_{\boldsymbol{\theta}}^{-1})$$

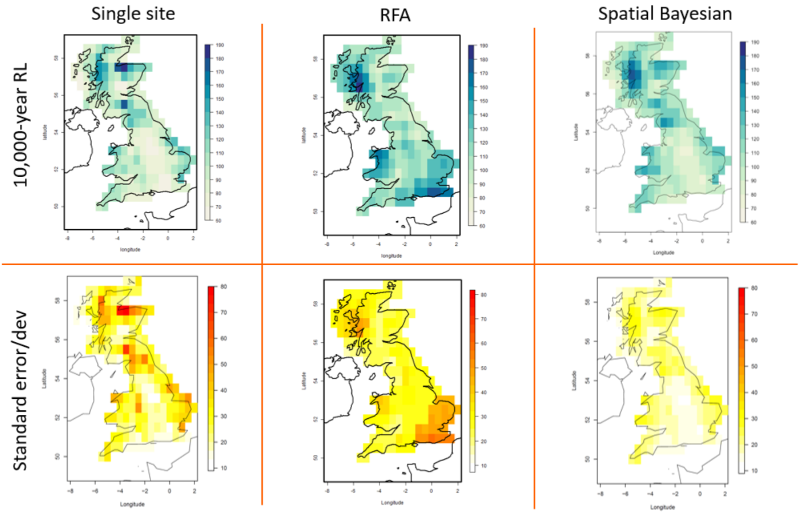
- This type of model can be fitted using Markov Chain Monte Carlo (MCMC).
- But how does this type of model perform in practice?

Rainfall example - Bayesian model results



Left: Estimates of scale parameter
Centre: Estimates of shape parameter
Right: Estimates of 10,000-year return level

Rainfall example - Comparison of all approaches



Discussion and Conclusion on Bayesian models

- The Bayesian model developed leads to the most spatially consistent return level estimates.
- The uncertainty estimates are greatly reduced when compared to single site EVA and RFA.
- However, this comes at a greater computational cost.
- The methods are currently still in the developmental phase and are yet to be tested on sparse observation data with missing values.
- The Bayesian paradigm also presents the opportunity to incorporate information from experts through the specification of prior distributions.

Spatial pooling - Summary

- The aim of this part has been to introduce a couple of novel approaches for improving extreme value characterisation.
- They have the potential to reduce the uncertainty associated to return level estimation.
- However, these benefits come at an additional computational cost and require more complicated modelling approaches.

Clustering of extremes

- Environmental variables have a tendency to form **clusters**, e.g. high temperatures caused by anticyclonic conditions.
- When fitting extreme value distributions we make the assumption that observations are independent.
- If they are not, we risk assuming we have more information than we actually do and therefore producing uncertainty bounds that are too narrow.
- Need methods to ensure that modelled data are independent.

How to account for clustering?

- Partition the data into independent time blocks and take the maximum of each.
  - Leads to the same results as for block maxima approach.
- For each cluster, model only the peak over the threshold as opposed to all values above the threshold.
  - Most commonly used approach.
  - Can use more information than for block maxima but can lead to underestimation.
- Explicitly model the within cluster dependence.
  - Permits the use of all values above the threshold.
  - Requires multivariate EVA - similar models can be used as have previously been shown.

The extremal index

- One commonly used measure of the amount of clustering is given by the extremal index $\theta \in (0, 1]$ (Leadbetter 1983).
  - $\theta = 1 \Rightarrow$ independence.
  - $\theta \to 0 \Rightarrow$ perfect dependence.
- The average number of exceedances is given by the reciprocal of the measure, i.e. θ^{-1}.
- Can be estimated empirically as

$$\hat{\theta} = \frac{n_c}{n_u} = \frac{\#\{\text{Clusters}\}}{\#\{\text{Exceedances of } u\}}$$

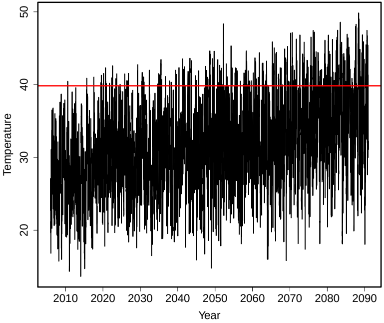
- Number of clusters can be estimated via the runs method or intervals method.

Explicitly modelling temporal dependence

- Some initial work has been undertaken to explicitly model the extremal dependence over time.
- Multivariate EVA models are used but instead the aim is to model $X_{t+1}|X_t > u$ and so on.
- Winter & Tawn (2016) used the conditional extremes approach to model the dependence between consecutive values to estimate the probability of heatwaves.
- This has been extended for higher-order processes (Winter & Tawn 2017).

Non-stationarity in extremes

- In many situations data are not stationary:
  - Heavier rainfall in summer than winter.
  - Climate change leading to hotter temperatures.
- Standard univariate methods may perform poorly.



How to account for non-stationarity?

- Directly in model parameters. For a covariate y and GEV would have

$$\mu(y) = \mu_0 + \mu_1 y$$
$$\sigma(y) = \exp(\sigma_0 + \sigma_1 y)$$
$$\xi(y) = \xi_0 + \xi_1 y$$

- By pre-processing the data first to remove non-stationarity and then applying stationary EVA (Eastoe & Tawn 2009).
- Use variable threshold (Northrop & Jonathan 2011).

Table of Contents

Concluding remarks

- I hope that this course has given a basic overview of EVA and more advanced techniques that are currently being developed.
- This area of statistics is seen as a vital skill across industry to estimate the risks posed by natural hazards.
- The go-to introductory book is still Coles (2001).
- Many packages are now out there (especially within R) to do this type of analysis.
- If you have any questions off the back of the course, feel free to get in touch.

References I

Coles, S. G. (2001), *An Introduction to Statistical Modeling of Extreme Values*, Springer Verlag.

Coles, S. G., Heffernan, J. E. & Tawn, J. A. (1999), 'Dependence measures for extreme value analyses', *Extremes* **2**(4), 339–365.

Davison, A. C. & Smith, R. L. (1990), 'Models for exceedances over high thresholds (with discussion)', *Journal of the Royal Statistical Society: Series B* **52**(3), 393–442.

Eastoe, E. F. & Tawn, J. A. (2009), 'Modelling non-stationary extremes with application to surface level ozone', *Journal of the Royal Statistical Society: Series C* **58**(1), 25–45.

Ferro, C. A. T. & Segers, J. (2003), 'Inference for clusters of extreme values', *Journal of the Royal Statistical Society: Series B* **65**(2), 545–556.

References II

Heffernan, J. E. & Tawn, J. A. (2004), 'A conditional approach for multivariate extreme values (with discussion)', *Journal of the Royal Statistical Society: Series B* **66**(3), 497–546.

Hosking, J. & Wallis, J. (2005), *Regional Frequency Analysis: An Approach Based on L-Moments*, Cambridge University Press.

Leadbetter, M. R. (1983), 'Extremes and local dependence in stationary sequences', *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete* **65**(2), 291–306.

Ledford, A. W. & Tawn, J. A. (1997), 'Modelling dependence within joint tail regions', *Journal of the Royal Statistical Society: Series B* **59**(2), 475–499.

Nelson, R. B. (2007), *An Introduction to Copulas*, 2nd edn, Springer.

References III

Northrop, P. J., Attalides, N. & Jonathan, P. (2016), 'Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity', *Journal of the Royal Statistical Society. Series C: Applied Statistics* pp. 93–120.

Northrop, P. J. & Jonathan, P. (2011), 'Threshold modelling of spatially dependent nonstationary extremes with application to hurricane-induced wave heights', *Environmetrics* **22**, 799–809.

Reich, B. J., Shaby, B. A. & Cooley, D. (2014), 'A hierarchical model for serially-dependent extremes: A study of heat waves in the western US', *Journal of Agricultural, Biological, and Environmental Statistics* **19**, 119–135.

Weiss, J., Bernardara, P. & Benoit, M. (2014), 'Formation of homogeneous regions for regional frequency analysis of extreme significant wave heights', *Journal of Geophysical Research: Oceans* **119**.

References IV

Winter, H. C. & Tawn, J. A. (2016), 'Modelling heatwaves in central France: a case study in extremal dependence', *Journal of the Royal Statistical Society: Series C* **65**(3), 345–365.

Winter, H. C. & Tawn, J. A. (2017), 'kth-order Markov extremal models for assessing heatwave risks', *Extremes* .