DSNE Autumn School - Extremes

Dan Clarkson[1] and Callum Murphy-Barltrop[2]
[1]Data Science of the Natural Environment, Department of Mathematics and Statistics, Lancaster University LA1 4YF, United Kingdom
[2]STOR-i Centre for Doctoral Training, Lancaster University LA1 4YR, United Kingdom

1st November 2022

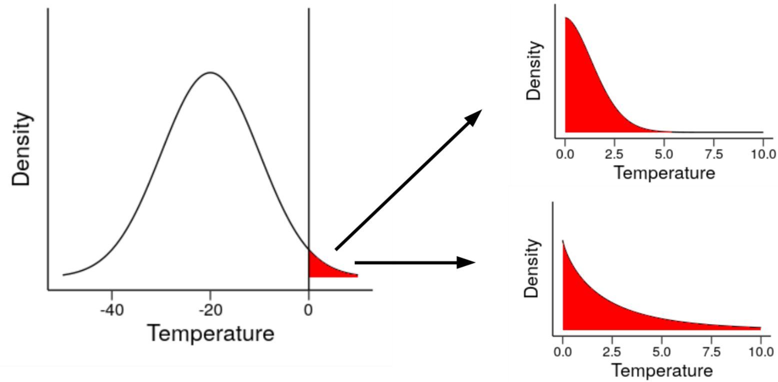
Introduction

1. Extreme value analysis (EVA)
   - Motivation
   - Annual maxima
   - Peaks over threshold
2. Additional topics and extensions
   - Non-stationarity
   - Temporal dependence
   - Multivariate extremes

Motivation

- Understanding extreme events central to the study of natural hazards
  - Droughts
  - Floods
  - Heatwaves
- Questions of interest around increased frequency, severity, intensity of extreme events under climate change
- Statistical models are essential because historical events are rare

Motivation

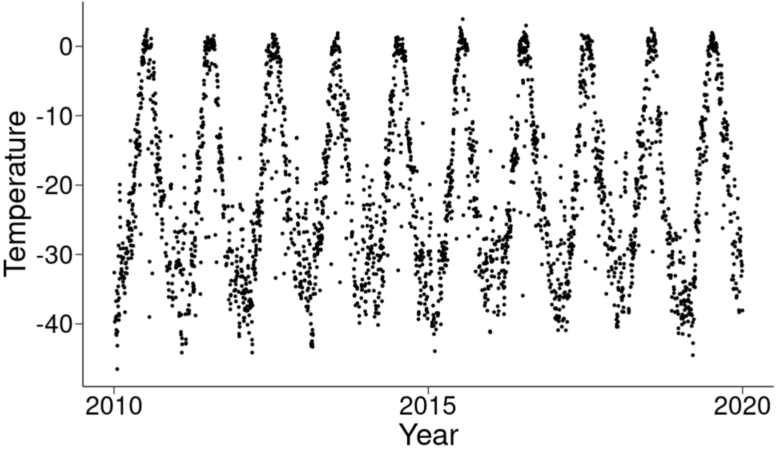


Aim: accurately model the tail of a distribution in order to extrapolate/make inference on very rare events

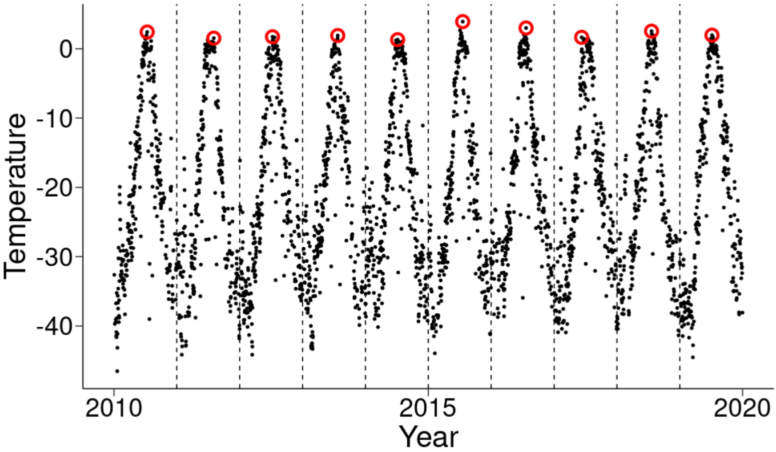
Approach

1. Identify extreme values
   - Block maxima
   - Threshold exceedances
2. Model with an extreme value distribution
3. Extrapolate/make inference on rare events

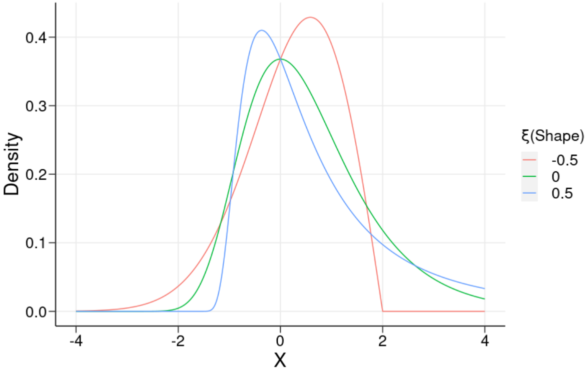
Method 1: Block maxima



Method 1: Block maxima



Method 1: Block maxima



GEV(μ, σ, ξ):

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

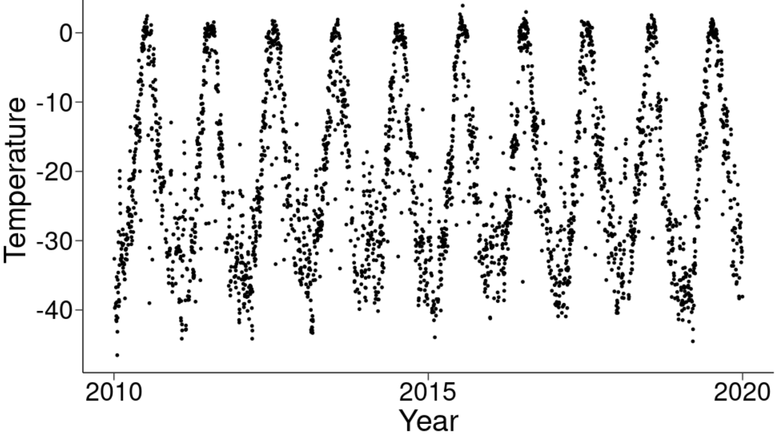
where $x_{+} = max(x, 0)$ and $\sigma > 0$

Method 1: Block maxima

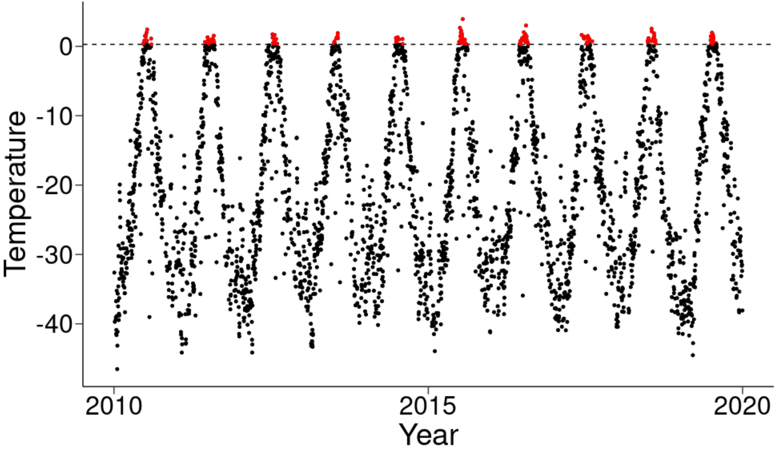
Considerations

- More appropriate for longer time series/natural blocks
- Can be inefficient use of data depending on block size
- Distribution can have upper or lower endpoints

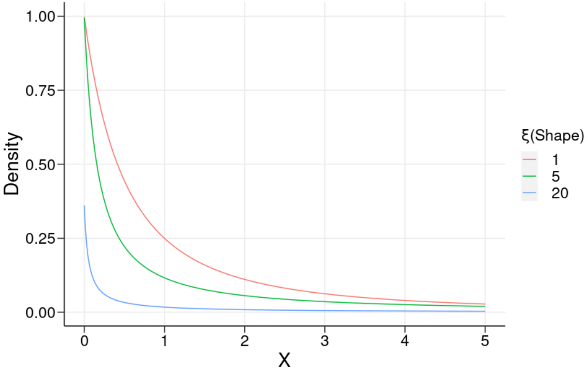
Method 2: Peaks over threshold



Method 2: Peaks over threshold



Method 2: Peaks over threshold



GPD(σ, ξ):

$$\Pr\left(Y_u < y \mid Y_u > 0\right) = 1 - \left\{1 + \xi\left(\frac{y}{\sigma_u}\right)\right\}_{+}^{-1/\xi} \qquad y > 0$$

Method 2: Peaks over threshold

Threshold selection

- Often this is reasonably simplistic
  - $95, 97.5, 99\%$ quantiles
- Lower threshold $=$ more data
- Higher threshold $=$ more asymptotically justified
- Measures to assess most appropriate threshold

Additional topics and extensions

Like anything, EVA is more complex than it first appears...

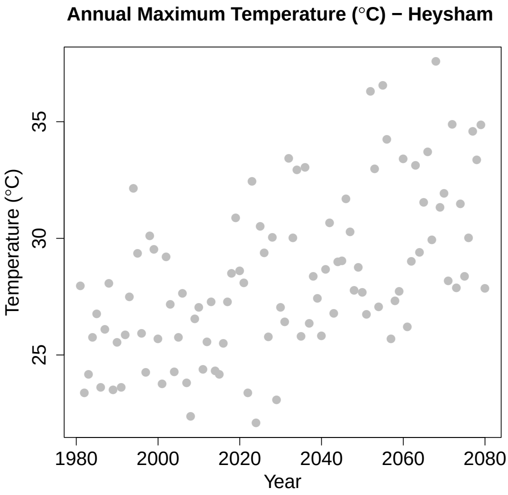
1. Non-stationarity.
2. Temporal dependence.
3. Multivariate extremes.

Non-stationarity

- Typically, we assume data are **independent and identically distributed.**
- When the second assumption is violated, we say data are **non-stationary**.
- **Very common** in environmental data.
- Example: temperature increase over time due to climate change.

Non-stationarity

Example: UKCP18 projections for Heysham, UK. Clear (linear?) trend in the data.



Non-stationarity

- **Many** methods for capturing these trends in a data series.
- For example, we can **add trends** to **parameters**.

Non-stationarity

- Rather than assuming

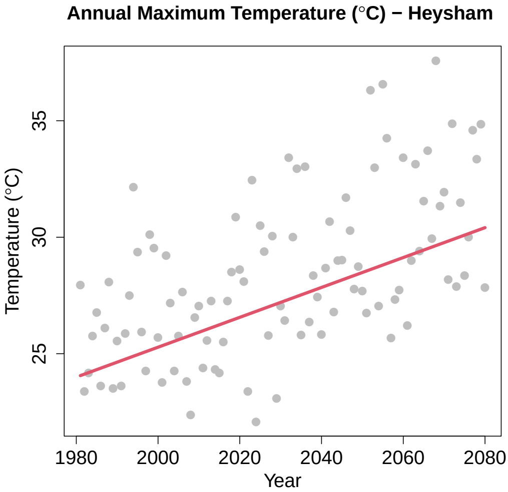
$$X_t \sim \text{GEV}(\mu, \sigma, \xi),$$

we could assume

$$X_t \sim \text{GEV}(\mu_0 + \mu_1 t, \sigma, \xi).$$

- This corresponds to a linear trend in the **location** parameter.

Non-stationarity



Non-stationarity

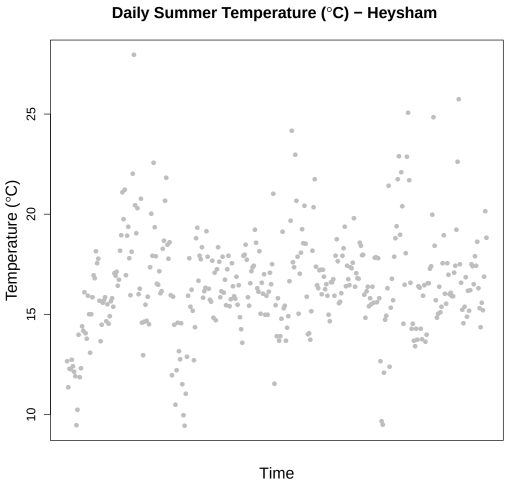
- Non-stationary modelling is **never** ‘one size fits all’.
- Requires careful assessment of data features.
- Trends typically **non-linear**.

Temporal dependence

- Typically assume data are **independent and identically distributed.**
- When first assumption is violated, we say data are **dependent**.
- Again, **very common** in environmental scenarios.
- Example: if it rains today, it is more likely to rain tomorrow.

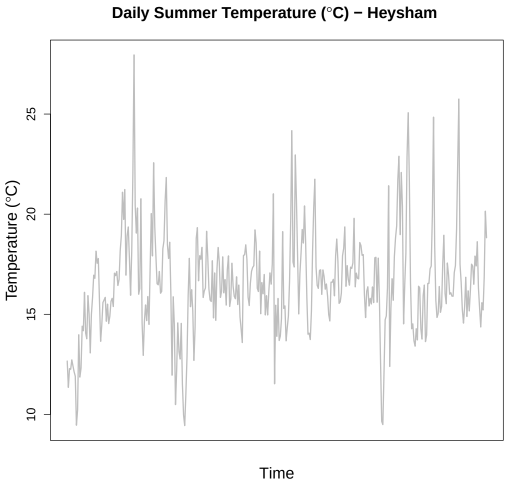
Temporal dependence

Example: UKCP18 projections for Heysham, UK - just **summer** months. Extreme observations tend to ‘cluster’.



Temporal dependence

Example: UKCP18 projections for Heysham, UK - just **summer** months.



Temporal dependence

- The issue: since nearby datapoints are closely linked, we have **less information**.
- Can't consider each data point as an **individual event**.
- Less information = **more uncertainty**.
- For example, a sample of 5 people from the **same** social group/family/company will provide **less information** for statistical analysis compared to 5 **randomly** chosen people.

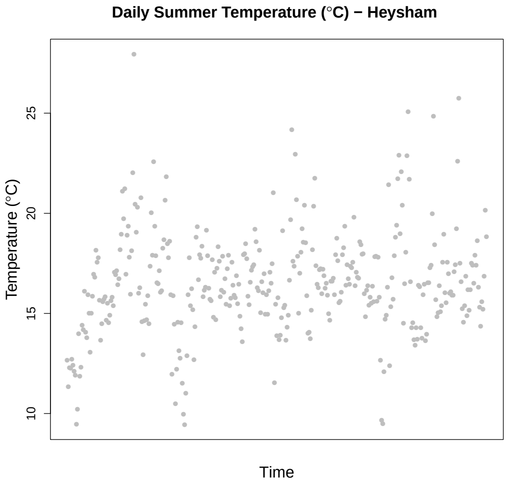
Temporal dependence

- How do we account for temporal dependence?
- **Good news:** can still fit same distributions.
- For block maxima data, **fit GEV** as normal.
- Dependence less of an issue.

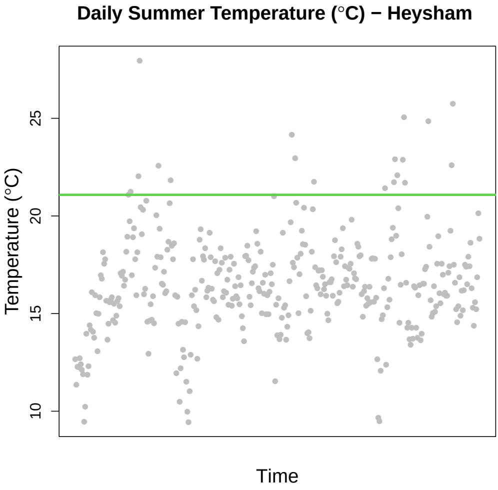
Temporal dependence

- For peaks over threshold, we typically
  1. Define clusters.
  2. Take cluster maxima.
  3. Assume cluster maxima to be independent.
  4. Fit GPD to cluster maxima.
- Removing lots of data - hence **more uncertainty**.

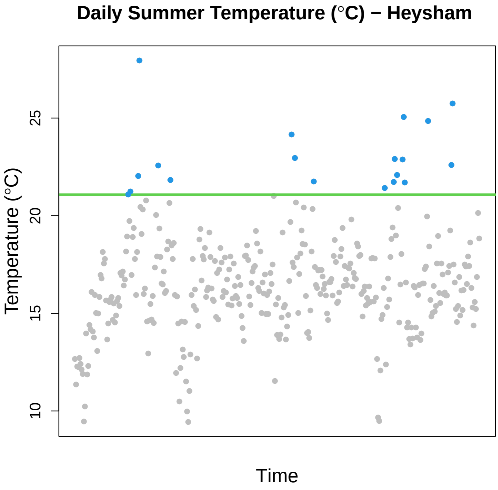
Temporal dependence



Temporal dependence

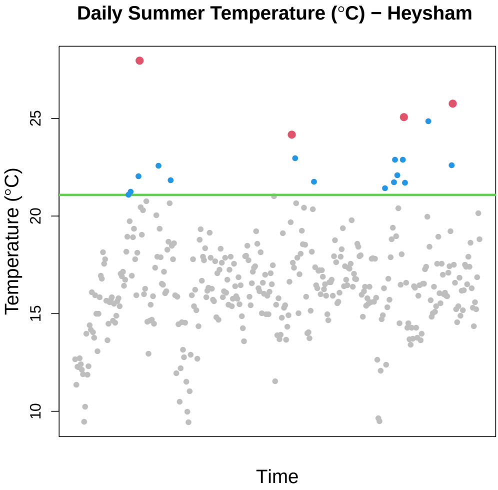


Temporal dependence



Temporal dependence

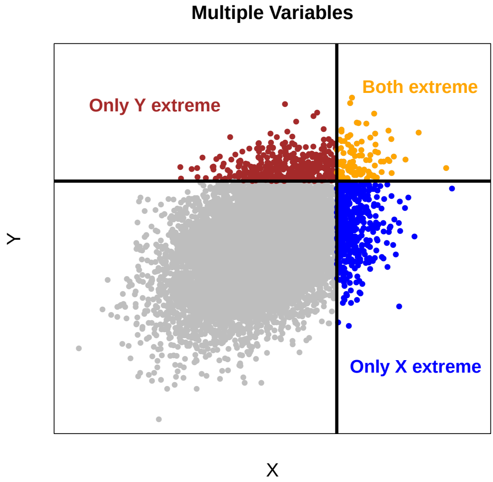
Less data = more uncertainty.



Multivariate extremes

- Multivariate = multiple variables.
- Multivariate extremes more **ambiguous**.
- No **natural ordering**.
- No **unique definition** of extremes.

Multivariate extremes

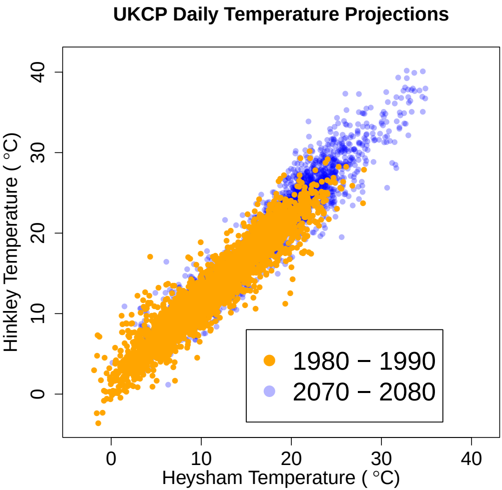


Multivariate extremes

- Definition of multivariate extremes varies between applications.
- **Many models** proposed in literature.
- Still a very active area of research (hence mine + Dan's PhDs).

Multivariate extremes

We still have non-stationarity and temporal dependence to contend with...

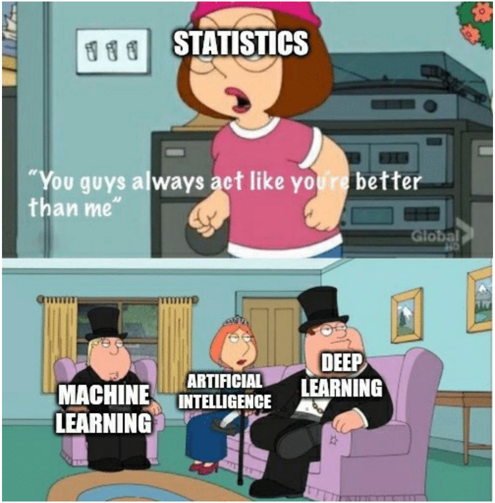


Further reading

- Chapters 5, 6, and 8 of Coles (2001) provide a great summary of all additional topics.
- See Beirlant et al. (2004) for a more theoretical outlook.
- Google Scholar - depending on what you need.

Thank you all for listening!

Does anyone have any questions?



References I

Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J., De Waal, D., and Ferro, C. (2004). *Statistics of extremes: Theory and applications*. John Wiley & Sons, Inc.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer London, London.