Extreme Climatic Events: Impacts for Ireland

John O'Sullivan, Conor Sweeney, and Andrew Parnell
andrew.parnell@ucd.ie

UCD School of Mathematics and Statistics

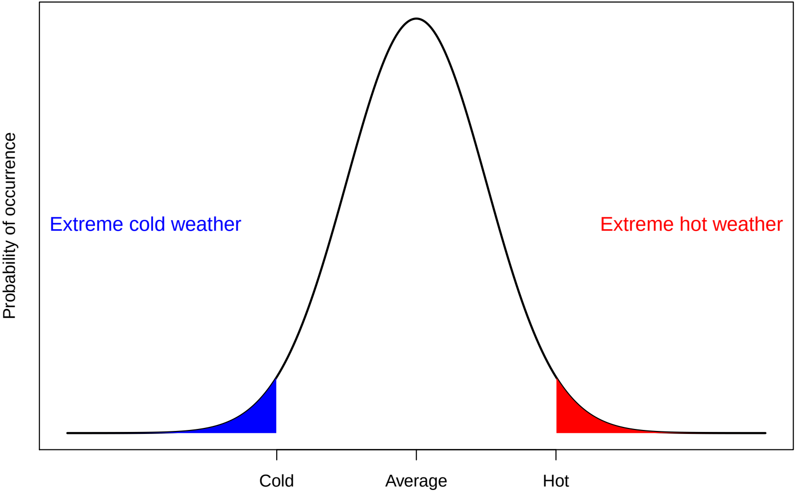
Environmental Protection Agency
An Ghníomhaireacht um Chaomhnú Comhshaoil

UCD Earth Institute
Better understand today's world.
Inform solutions for tomorrow.

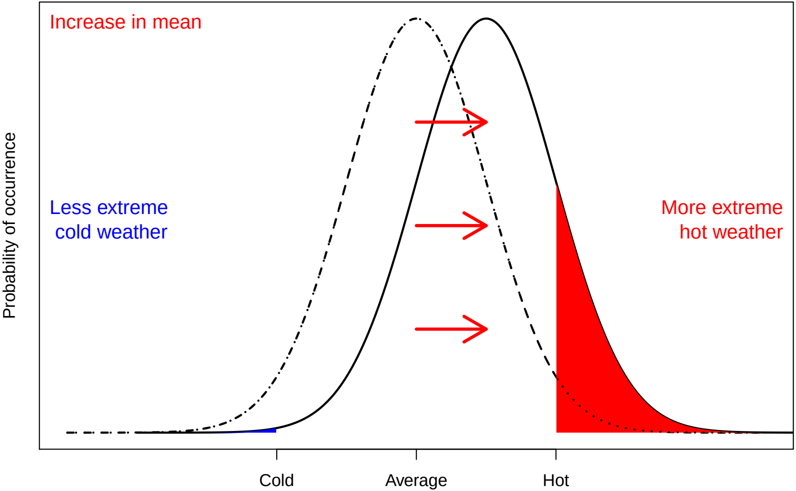
What is this talk about?

- We want to predict extreme temperatures, wind speeds, and rainfall for Ireland up to 2100
- We have a large set of data covering 50 years of weather in Ireland
- We have some down-scaled climate model runs of weather/climate up to 2100
- We want to use all this information together to estimate past, present and future extremes whilst accounting for uncertainty
- In this talk I'm only going to cover changes in past extremes

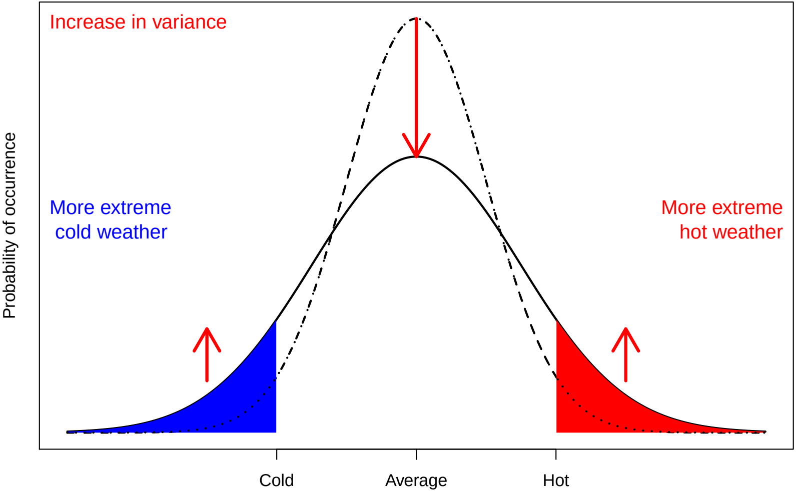
What is climate and what are extremes?



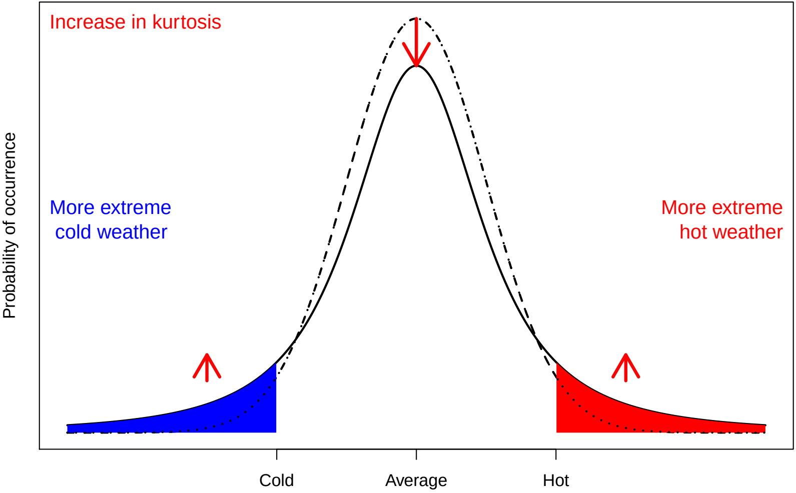
What is climate and what are extremes?



What is climate and what are extremes?



What is climate and what are extremes?

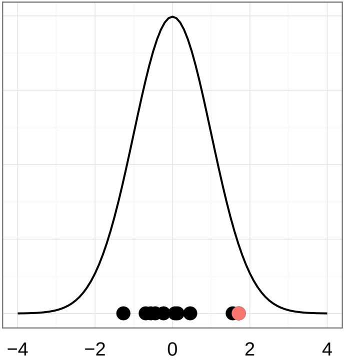


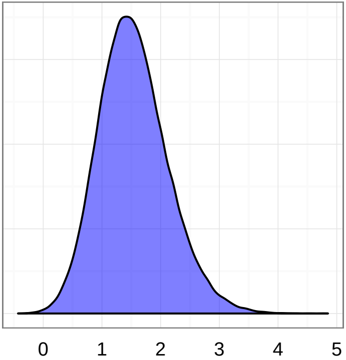
Some comments

- These plots are still an **over-simplification** (e.g. skew/multi-modal/multivariate distributions)
- Thinking about climate changes only in terms of the mean is not sufficient in determining the frequency or size of extreme events
- If we are interested in extremes we need to focus on the **tail behaviour** of the probability distribution
- The discipline of **Extreme Value Theory** was invented to determine changes in maxima (or minima) of series. It is not widely taught

The amazing Fisher-Tippett-Gnedenko theorem

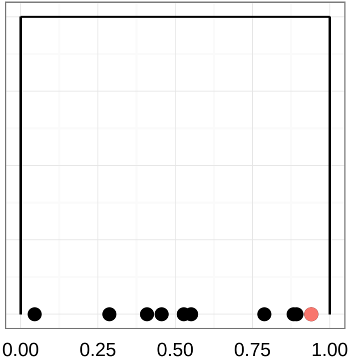
- A remarkable theorem: repeatedly take the maxima of any distribution and you will end up with a **Generalised Extreme Value** (GEV) distribution

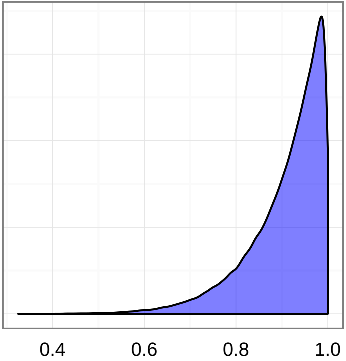




The amazing Fisher-Tippett-Gnedenko theorem

- A remarkable theorem: repeatedly take the maxima of any distribution and you will end up with a **Generalised Extreme Value** (GEV) distribution





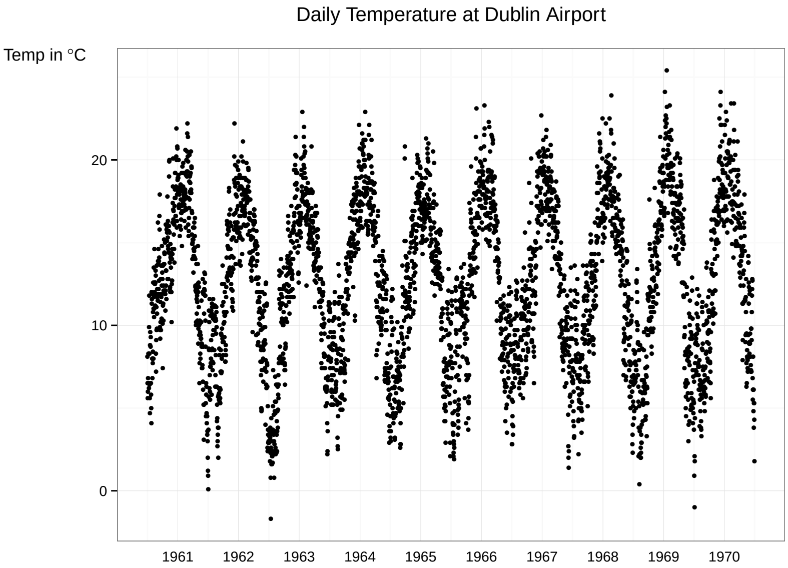
Extreme Value Theory

- We can use the GEV distribution to model the **tail behaviour** of the distribution
- There are two key parameters, the **scale**, which measures how spread out the distribution is, and the **shape**, which measures how long-tailed the distribution is
- If the shape is **negative**, then there is a maximum value in the data. If the shape is **positive**, then the data are long-tailed
- Positive shape parameters for climate are **dangerous** - more extreme extremes!

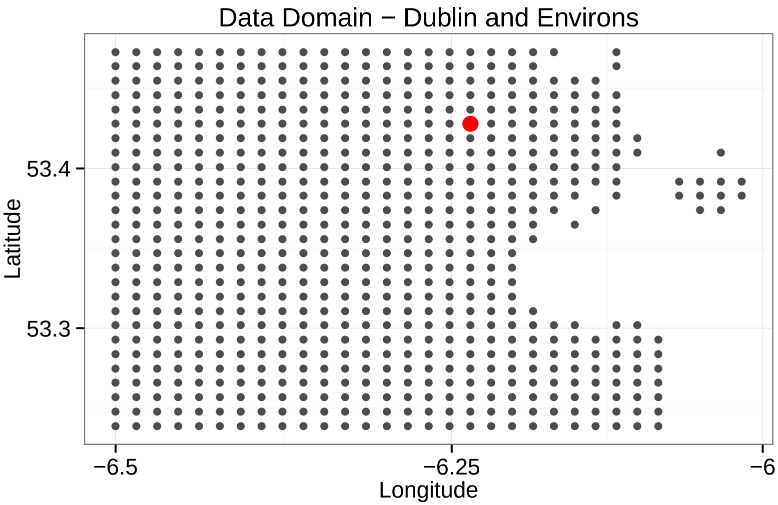
Where does this data come from?

- The raw data have numerous **missing values** from different stations at differing points in time
- Instead we use **re-analysis** data created by Seamus Walsh at Met Éireann. These re-analysis data are available at every grid point and every time point
- They are provided on a grid of 1km by 1km, totalling 72,000 grid points with over 50 years of daily data
- Over **1.3 billion** data points altogether!

Our data - temporal aspect



Our data - spatial aspect



Estimating extremes at just a single site

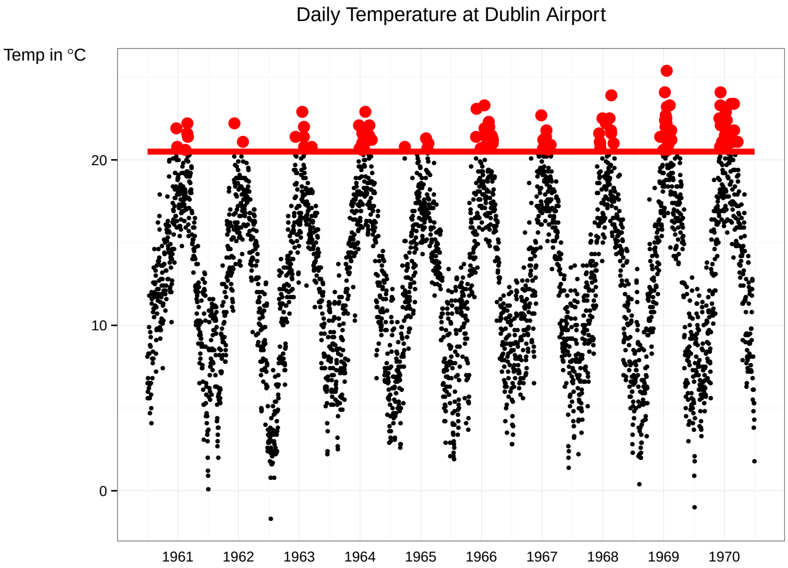
- Suppose we took just the daily maximum temperature at one particular site, Dublin airport
- We can fit the GEV distribution to this which will estimate the shape of the distribution curve
- From the output we can create a **return level plot** showing how likely different extreme temperatures are at different **return periods**
- The return level plot is just a function of the parameters

The output: return level plots

A slight alternative; the Generalised Pareto distribution

- The GEV has a slight disadvantage in that it only uses **maxima**
- An alternative, the **Generalised Pareto Distribution** (GPD) uses all the data above a threshold
- This means we get a better estimate of extremal behaviour, but we have to set a **threshold**, and can occasionally end up with more data than we can deal with
- The model fitting and output is exactly the same as the GEV

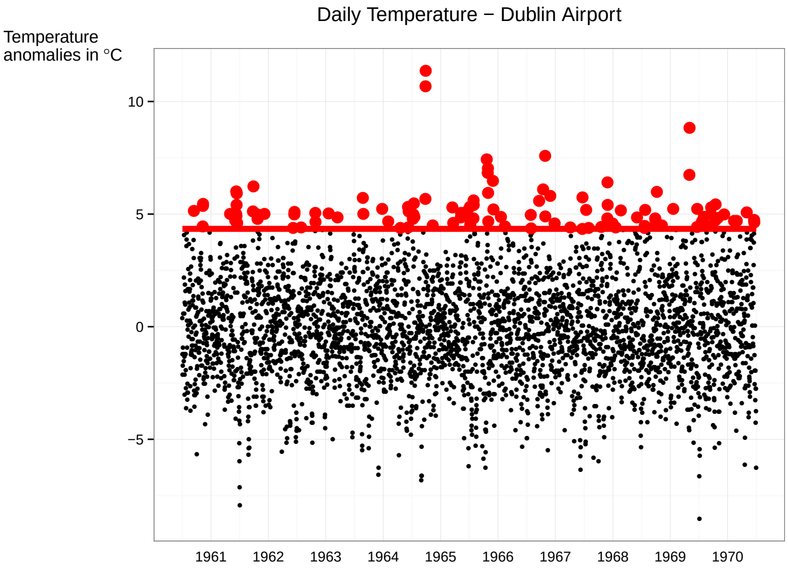
The GPD approach



Choosing a threshold

- A tricky part of using the GPD is choosing a threshold above which you keep the data and use it to estimate extremes
- If you set the threshold too low, the assumptions of the GPD aren't valid. If you set it to high you get a really poor noisy fit
- There isn't enough information in the data usually to estimate the threshold as another parameter
- Instead it is common to choose a **high percentile** (e.g. 99%)
- With data that vary spatially or over time it might be necessary to **vary the threshold**

Variable thresholds via anomalies

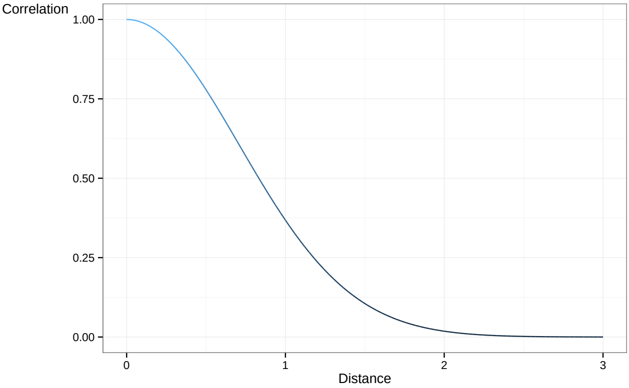


Spatial modelling

- We have spatially indexed data so we want to take account of the fact that extremes at neighbouring sites are likely to be similar
- Spatial statistical modelling is a very mature field, however many people still run extreme value models **separately** for each site
- We use **Gaussian Processes** to force neighbouring sites to have similar extreme values

A short introduction to Gaussian processes

- Gaussian processes work by fitting a multivariate normal distribution to the data at hand, with a restricted correlation matrix that requires high correlation of nearby sites



Combining spatial and EVT models

We use a **hierarchical model**:

$$y_i(s) \sim GPD(\sigma(s), \xi(s)),\ i = 1, \ldots, n$$

$$\sigma(s) \sim MVN(\mu_\sigma, \Sigma_\sigma)$$

$$\xi(s) \sim MVN(\mu_\xi, \Sigma_\xi)$$

with

$$\mu_\sigma = \alpha_\sigma,\ \mu_\xi = \alpha_\xi,$$

and

$$\Sigma_\sigma^{i,j} = \tau_\sigma^2 \rho_\sigma^{ij} + \omega_\sigma^2,\ \Sigma_\xi^{i,j} = \tau_\xi^2 \rho_\xi^{ij} + \omega_\xi^2$$

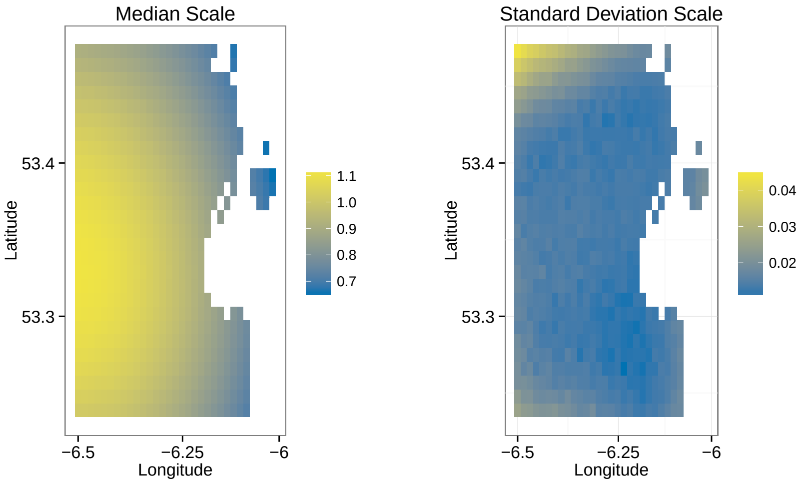
The advantages of a model-based Bayesian approach

- Put everything into **one model**
- Easy to produce predictions with **uncertainty**
- Easy to combine lots of different data sets
- Possible to include informative **external information**

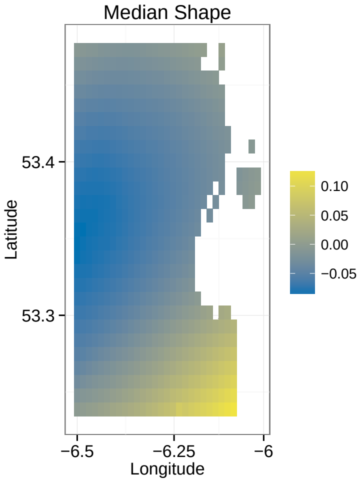
Computational issues

- We fit the model using **Markov chain Monte Carlo**
- This is a trial-and-error algorithm which produces probability distributions of the unknown parameters (e.g. $\sigma(s)$ and $\xi(s)$) given the data
- This probability distribution is **joint** in the sense that we have all the uncertainties and correlations between all the parameters together
- A disadvantage is that it's **slow to run** for big data sets, especially with nested Gaussian Processes because you have to solve a matrix at each step
- From the output we can use some Gaussian Processes tricks to produce predictions with uncertainty for any set of grid points we like

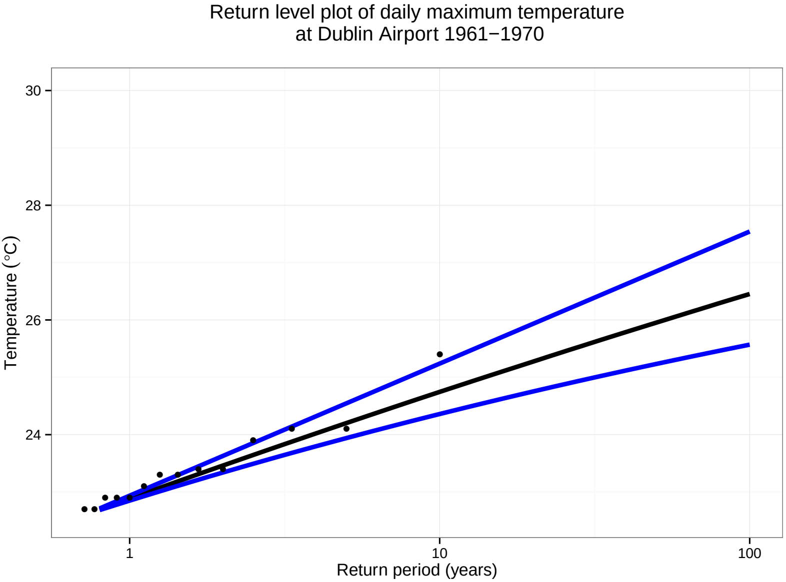
Output - Dublin region for the 1960s



Output - Dublin region for the 1960s

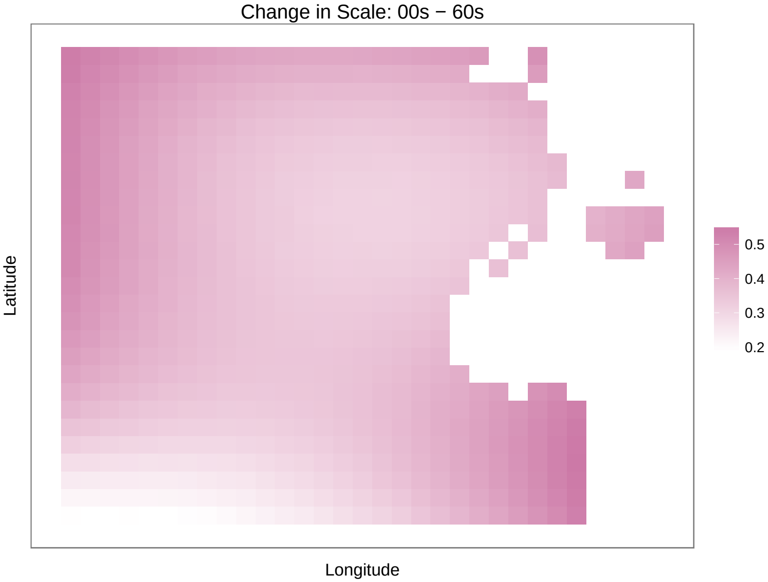


Output - return levels for a site

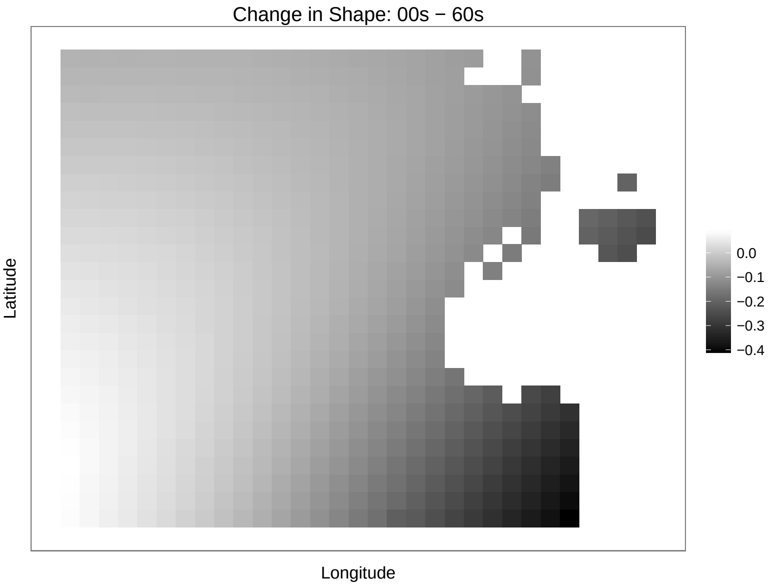


Output - return levels over the decades

Output - changes in scale and shape



Output - changes in scale and shape



Spatio-temporal models

Conceptually simple to extend to **spatio-temporal** models:

$$y_i(s,t) \sim GPD(\sigma(s,t), \xi(s,t)),\ i = 1, \ldots, n$$

$$\sigma(s,t) \sim MVN(\mu_\sigma, \Sigma_\sigma)$$

$$\xi(s,t) \sim MVN(\mu_\xi, \Sigma_\xi)$$

We then have the option of including the time series component in the mean or the covariance, or both

Challenges:

Statistical:

- Richer ways of **specifying spatio-temporal models**
- Richer(but more complicated) **likelihoods for extremes**
- **Faster** computational methods (SPDE-INLA, predictive processes)

Climatological:

- Incorporating climate model data, **grid matching issues**
- **Projecting** into the future

References and funding details

- The best book on statistical extreme value theory: Coles, S. An Introduction to the Statistical Modeling of Extreme Values, 2001, Springer
- A good paper on Hierarchical Bayesian modelling of extremes: Cooley, D., et al, Bayesian Spatial Modeling of Extreme Precipitation Return Levels, Journal of the American Statistical Association, 2005
- For the really advanced material, see, e.g. R. Huser and A. C. Davison. Space-time modelling of extreme events, in Journal of the royal statistical society: series B, 2014 and references therein

Funded by EPA Strive PhD fellowship 2012-CCRP-PhD.3,
Presentation available at: andrewcparnell.github.io