University of Exeter

MMTH044 Project in Statistics

Downscaling Extreme Precipitation in South West England

Author
Daniel Williams

Supervisor
Dr. Ben Youngman

Abstract

The low resolution of gridded model output can cause inaccuracies in estimates of key climate variables, but some of this bias can be corrected by using a downscaling model. Over the small region of South West England there are significant variations in orography that cause larger amounts of precipitation, which is not captured accurately across the scale of a model grid cell, exemplifying the problem. This research presents a way of downscaling extreme precipitation, firstly by parameterising the extremes using a generalised extreme value (GEV) distribution, for both the gridded model and the observations. A generalised additive model (GAM) is used for downscaling the estimated return levels calculated from these GEV models. Incorporation of spatial covariates into the GEV models explain the distribution of precipitation extremes over the South West more accurately than what was achieved with a stationary model. Spatial relationships are also used in the downscaling model, and are found to be important in creating accurate predictions. These predictions maintain the spatial composition of the estimated return levels calculated from the observations, and are of a similar size. This shows that it is possible to predict return levels for areas with sparse amounts of observations, using a pre-existing relationship between observations and model output.

Contents

1 Introduction

Numerical solutions of complex dynamical equations that define the climate system can provide information about long-term states of the world, and can give forecasts for expected levels of precipitation over the next few hundred years. Observations over a long historical period of time can also give information about how much precipitation will occur. However, due to climate change, extremes of precipitation have increased over the last century [Groisman et al., 1999], and so there is necessity for a climate model with some form of forcing applied to account for these effects.

Whilst the numerical climate models are needed, their estimates of various climate variables are often innacurate due to having a low-resolution gridded scale. Reanalysis models objectively combine observations and a numerical model to derive estimates of different weather variables. These types of models are beneficial where observations are sparse, as they provide a fixed set of data over space and time, but are still subject to a low resolution.

Over a smaller region such as South West England, finer scale variations in precipitation are not detected by gridded models. Observations are point-level data, and can provide more information on a finer scale. By downscaling, a statistical model can use information at the point-level and combine it with gridded model output to model the relationship between them both.

This project aims to provide methodologies to parameterise the distribution of the extreme values of precipitation for both gridded model output and observations, and derive a downscaling model to describe the connection between the two. A downscaling model would allow the interpolation of gridded data onto a finer scale, enabling predictions of point-level data through low resolution gridded model output. These methodologies would be tested over South West England, which would assess the capability of downscaling in a small region.

1.1 Extreme Precipitation in South West England

Extreme precipitation events classify as a natural hazard, as they can be the cause of serious flooding. Understanding the distribution of the extreme events of precipitation will have widespread applications. For example, in an ideal scenario, knowledge of the maximum amount of rainfall that will occur over the next 100 years will help in flood defense for villages near large rivers.

Over the South West, there is large variation in orography due to Dartmoor and Exmoor, which has a large effect on the amount of precipitation [M. L. Wigley et al., 1984]. These smaller regions of high precipitation are not captured accurately by gridded model output, as the grid is too sparse to provide the information. Note that the formal region of the South West is slightly larger than the domain being considered in this project, and references to the South West from here on refer to the south-western part of the official South West region. The domain considered throughout this project can be seen in Figure 1.

Figure 1 also shows the elevation levels across the South West, as well as the elevation points on a gridded scale. Since the gridded data samples every 0.25° in longitude and

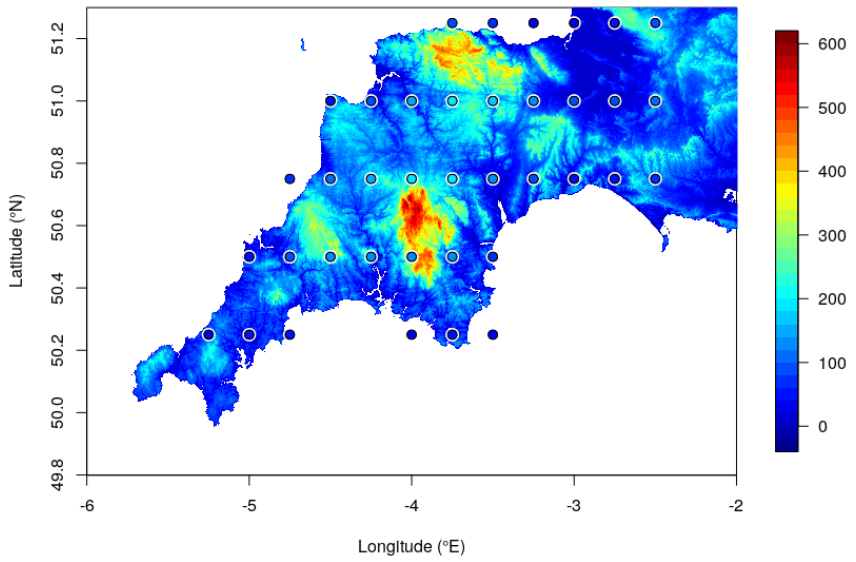


Figure 1: Elevation (m) shown in a colour plot for the South West, with a grid that samples every 0.25° in longitude and latitude overlaid, which contains elevations from a gridded model with the same colour scheme.

latitude, the high elevation levels of Dartmoor and Exmoor are not sampled. This means that the distribution of precipitation over the South West given by gridded models will likely be different to what is observed, most specifically for the extremes, at these high elevation regions. This presents an interesting challenge in the ability to use information from the gridded model data across this region to predict future levels of precipitation at positions within the grid cells. This will be a key focus for this project, as being able to predict at finer positions than grid cells, using only information from a gridded model, would allow estimates of extremes of precipitation at locations that contain few or no observations.

1.2 Extreme Value Theory and Downscaling

Normal modelling of precipitation involves modelling the bulk of its distribution, but modelling extremes is concerned with modelling the tails of the distribution, as this is where the extreme events would occur. Extreme events are rare by definition; for example, if we wanted to model likely floods across a certain region, we would not be interested in the average rainfall, but instead be interested in the most extreme rainfall that is possible. Extreme value theory is used to achieve this goal, and methods to do so include modelling the precipitation maxima across time, or modelling the excesses of a high quantile of rainfall. Since the early twentieth century, extreme value theory has been developing as a branch of statistics [Fisher and Tippett, 1928], and has had popular use in the modelling of extremes of climate and weather variables, such as precipitation.

Downscaling methods have been a popular field of research as a way of bridging the gap between numerical models and observations. Maraun et al. [2009] give a relevant summary of current techniques for precipitation downscaling under climate change, and cite that downscaling adds value to projections from global climate models.

Section 2 begins by introducing current methods for both modelling extreme values and for downscaling extremes, and how they are used in application. The mathematical framework for these methods is discussed in Section 4.

2 Literature Review

This section will explore the most relevant literature concerning modelling extreme values, and downscaling of the respective variables. A lot of research exists on parameterising extreme values, and it has been a topic of interest over the last century. Downscaling extremes is a developing field, and less literature exists on this topic, but the available research will be discussed.

2.1 Modelling Extreme Values

Many methods exist on modelling extreme values of a particular weather variable, and there are a range of distributions that can be used for this purpose. The generalised extreme value (GEV) distribution is amongst one of the most common, but other methods exist that deal with modelling excesses of a certain threshold, such as the generalised Pareto distribution (GPD).

The GEV distribution provides a parametric framework to model extreme values, and can easily be interpreted by the return values, which are the extreme quantiles of the distribution. This distributional result was first developed by Fisher and Tippett [1928], and later developed by Jenkinson [1955]. Gumbel [1958] introduced one of the first applications of the GEV distribution in engineering, and provided accessible methodology which is still relevant today. The derivation and methodology of the GEV distribution is explained in Section 4.1.1, but succinctly, it models the maxima of a particular variable, and contains parameters that describe the location, scale and shape of the distribution [Coles, 2001].

Many studies have found success in using the GEV distribution for modelling precipitation maxima across a large time period. Boudrissa et al. [2017] modelled the annual maxima of daily precipitation in Northern Algeria using a GEV distribution. In a review of statistical procedures for flood frequency estimation in the UK, Kjeldsen et al. [2008] found the GEV distribution to be amongst one of the best methods for accurately modelling various precipitation variables.

The GEV distribution can be extended to be non-stationary through its parameter estimates. Described in depth in Section 4.1.2, this allows the inclusion of spatial dependence in its estimates. This is important due to the spatial and orographical dependence of precipitation [M. L. Wigley et al., 1984]. Cooley et al. [2007] also implement covariates that describe climatological properties as well as spatial ones, which could be important to further explain variations in precipitation.

Some criticism of the GEV distribution involves its wasteful approach to excluding data, as most research uses annual maxima [Coles, 2001; Buishand, 1991] which would involve 'throwing away' on average 364 measurements per year. Davison and Huser [2015] instead use a monthly maxima approach instead of an annual one, which involves introducing parameters to deal with seasonality. If annual maxima are considered, seasonality does not need to be dealt with, as the data are considered stationary during that time period [Coles, 2001]. Other criticisms of the GEV distribution involve its inability to be used effectively for limited time periods [Cook, 1982].

An alternative method to modelling maxima is looking at the exceedance of a particularly high threshold. This is a less wasteful approach, and is most commonly done with a generalised Pareto distribution (GPD), originally implemented by Pickands [1975]. However, in a study of maximum daily precipitation by season in smaller regions of the Brazilian Amazon, Santos et al. [2015] found the GEV distribution to have a better fit to the data than the GPD. Since this research also focuses on a smaller region, the GEV distribution is likely more appropriate for measuring the extremes of precipitation in this case.

2.2 Downscaling Extremes

In a study of extreme precipitation in North America, Mannshardt-Shamseldin et al. [2010] make direct comparisons between the return levels of both gridded model output and observations, after fitting a GEV distribution to both, and derive a linear relationship between the two. This regression based approach parameterises the differences in return levels through easily interpretable terms, and inference can be made in how different the two variables are by inspection of the parameters. This approach will be a focus of this research, as it enables a simple parameterisation through easily accessible methods.

The n-year return level represents the value of precipitation that is expected to exceed over the next n years. Denoting $z_{n,\text{obs}}$ as the n-year return value estimated by the observations, and $z_{n,\text{mod}}$ as the same output given by the gridded model, a simple version of the method implemented by Mannshardt-Shamseldin et al. [2010] takes the form:

$$\begin{aligned} z_{n,\text{obs}} &\sim N(\nu, \sigma^2), \\ \nu &= \beta_0 + \beta_1 z_{n,\text{mod}}. \end{aligned}$$

In this model, β_0 and β_1 are estimated using standard linear modelling techniques; by maximising the log-likelihood of the Normal distribution with respect to these parameters. Mannshardt-Shamseldin et al. use this regression method with 'station averaged' observations for each grid cell, and the grid cell itself, using spatial covariates of longitude, latitude and elevation, of which longitude and latitude are cubic polynomials. Mannshardt-Shamseldin et al. note that spatial covariates were important to include in their downscaling model, and through doing so, their approach provided similar results as a kriging approach.

A more complex approach involves modelling the differences in the cumulative distribution functions (CDFs) of the data themselves. Déqué [2007] presented a 'Quantiles-matching' method, which was later improved by Michelangeli et al. [2009]. The basis of this idea is

that there exists a transformation $T : [0, 1] \to [0, 1]$ that would describe the relationship between the CDFs of the model output and the observations, such that

$$T\left(F_X(x)\right) = F_Y(x),$$

for a generic weather variable x. Here, F_X is the CDF of the model output, and F_Y is the CDF of the local observation data. The transformation T can be estimated non-parameterically from the data, based on a historical data set. The same transformation can then be applied for future periods, and statistical characteristics such as return levels can be inferred for this time period.

This CDF method presents a way of modelling the differences between the distribution without the need for parameterisation. However, this non-parameteric approach would require a large data set, and would be difficult to implement in the South West due to the smaller quantity of data. Non-parameteric methods such as quantile matching would not give spatially continuous estimates for this proposed problem in the research, and hence has been avoided. Kallache et al. [2011] expand the method to a semi-parameteric one, which combines the CDF matching approach and the use of a threshold exceedance extreme value distribution (a GPD). This would be suitable for an area of land such as the South West, but for this project, a strict parameteric approach is preferred, due to simplicity and the ability to quantify effects of covariates in how they would improve the model.

3 Data

The observation data are from the Met Office integrated data archive system (MIDAS) [Met-Office, 2018], and come in the form of daily precipitation accumulations for 863 stations, with a total of 4,976,704 measurements from the period of 1988 to 2017. Each station does not contain measurements for the entire duration of 30 years, nor do they all contain a measurement for every day in a particular year. It is assumed that the missing values in the data are missing at random, so that the remaining observations are expected to give unbiased estimates [Rubin, 1976].

The gridded model output is from ERA5 [Copernicus Climate Change Service, 2018] - the newest version of the ERA-interim reanalysis, where estimates of precipitation come from a reanalysis of temperature, humidity and wind observations [Dee et al., 2011]. These come in the form of hourly precipitation accumulations, and contains 105 grid cells over the region of the South West, for the period of 1979 to 2018.

3.1 Quality Control

The first form of quality control implemented was based on quality control flags set in place by the Met Office, recorded for each individual observation. Different flags described different points at which internal quality control failed. Observations were removed if the measurement was obtained via an estimation method which did not contain sufficient

information about the actual precipitation value at that time. Other measurements were excluded when the internal climate quality control program was not run [Met-Office, 2018].

A simple removal of those points reduced the amount of qualifying stations to 854, with a total of 2,102,960 measurements. The next form of quality control was determining whether extremely high values of precipitation were accurate depending on their surrounding measurements. Observation points that were above a certain value, which was arbitrarily taken to be 50 mm, were removed if their measurements up to two days before or after were more than four times lower than them. This reduced the total number of measurements to 2,093,793.

After this, stations were only considered reliable if they contained at least 10 years worth of data, of which each year contained at least 200 observations. Additionally, since this project is concerned with downscaling extreme values between model output and observations, observation stations that did not correspond to a grid cell for the available data set were omitted. This ensured no disparity in location between the gridded output and the observation data set. This left a final total of 491 stations, and 1,540,921 measurements.

Whilst the observations would be more accurate if the quality control were stricter, this proved sufficient to remove any large and obviously unreliable measurements and stations. Due to the aim of the project primarily being downscaling, the precision of the estimates is not as important as the estimates themselves. Provided that the observations and the model output are describing the same variable, inference on the difference between the two can still be made.

3.2 Inspection of Maxima

Since this project is concerned with modelling the extremes of precipitation, it is important to inspect the distribution of precipitation maxima that the observations and the ERA5 output provide, as that is what is being modelled. Figure 2 shows the distribution of maxima across the observations and the ERA5 output after quality control. The interesting aspect is the difference in size between the two: the observation maxima are generally larger than the maxima from the ERA reanalysis.

Since there is a large size difference at this preliminary stage of data analysis, it is expected that this systematic difference in the magnitude of the precipitation maxima will exist throughout further analysis of the data. The reason for this large difference is likely explained by the varying nature of precipitation over the South West, and since the observations are on a finer scale, they capture higher precipitation values at areas of higher elevation. This is demonstrated in Figure 1, and will be important when making comparisons between observations and the gridded model later on.

The observations are considered as an unbiased representation of the actual precipitation that occurred, and the precipitation maxima from ERA5 are considered to have a bias which can be corrected by downscaling. However, it is clear from the observations that there is a much larger upper tail of the maxima distribution than there is in the ERA5 reanalysis. This could be due to the quality control missing incorrect measurements, or it could simply be due to a feature present in the observation data. Whilst this looks

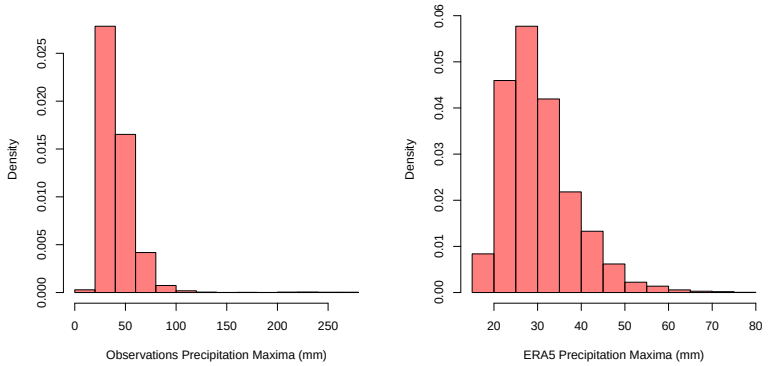


Figure 2: Histogram of precipitation maxima (mm) for observations and ERA5 reanalysis, across the entire time period and region.

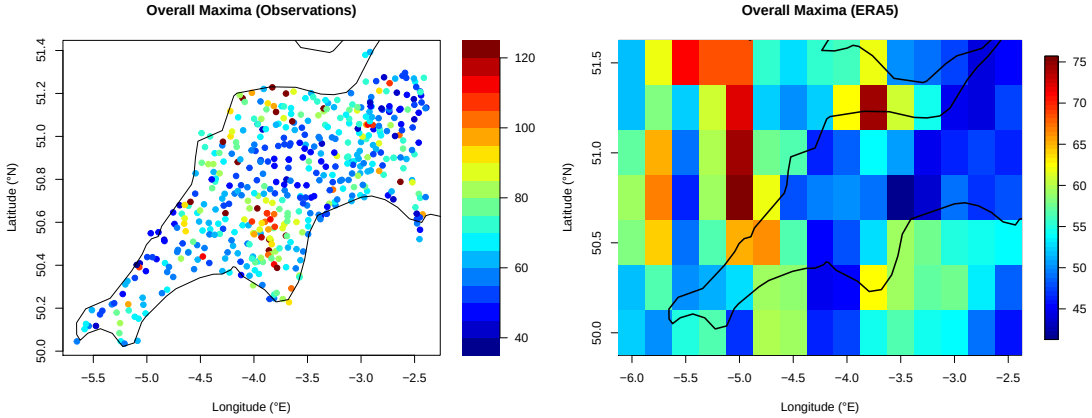


Figure 3: Overall precipitation maxima (mm) for observations (left) and ERA5 model output (right). *Note: overall maxima in observations is restricted to a maximum of 120 mm, to avoid spurious estimates in the spatial pattern, and a different scale used for clarity.*

troubling upon inspection, these values do not cause any discrepancies later on in the model fitting, and keeping a larger data set allows more reliable inferences to be made when inspecting the spatial variation (which would not be possible if the quality control was stricter).

We can inspect the overall maxima of the observations and the model output across space, seen in Figure 3. As expected, the observations show larger overall maxima at regions of high elevation, but there is no clear pattern seen in the ERA5 output. This is the type of spatial variation that would be expected to be seen throughout further analyses, as we will be modelling the yearly maxima with the GEV distribution. This is what will be used for comparisons later on.

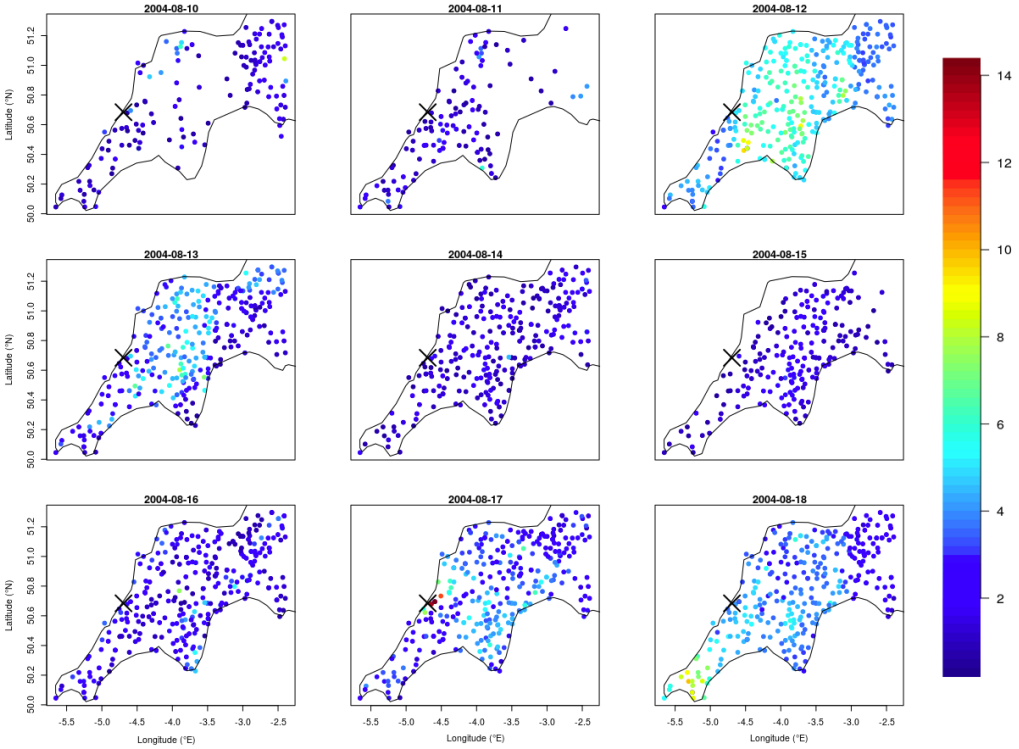


Figure 4: Square root of daily precipitation accumulations (mm) from the observations for the South West for the days leading up to the flood in Boscastle, and two days afterwards. The location of Boscastle is indicated with a black cross. A square root scale has been used for clarity in precipitation values. The real highest precipitation value here (at Boscastle) was 200.4 mm.

3.3 Motivating Example: Boscastle Floods

To demonstrate why modelling extremes of precipitation is important, and to see the disparity between the observations and the gridded ERA5 data, take as an example the floods in Boscastle that occurred on the 16th of August 2004. Boscastle is a small village and fishing port on the north coast of Cornwall in the South West.

Boscastle was the centre of an unfortunate flooding event that destroyed approximately 100 homes. This event was caused by heavy rain across Cornwall for several hours, and was intensified by the local topography of the area surrounding Boscastle [Met-Office, 2019]. This serious flooding caused long lasting damage, and rebuilding took several years to accomplish.

Figures 4 and 5 show the observations around the time of the flooding, as well as the gridded model output from ERA5 for the same time. Note that because precipitation is observed from accumulations from 9 a.m. - 9 a.m. every 24 hours, observed records for flooding that occurred may be shown a day later than when it happened. For the ERA5 output, predictions are made from midnight to midnight for a particular day, so these are also 'out of sync'.

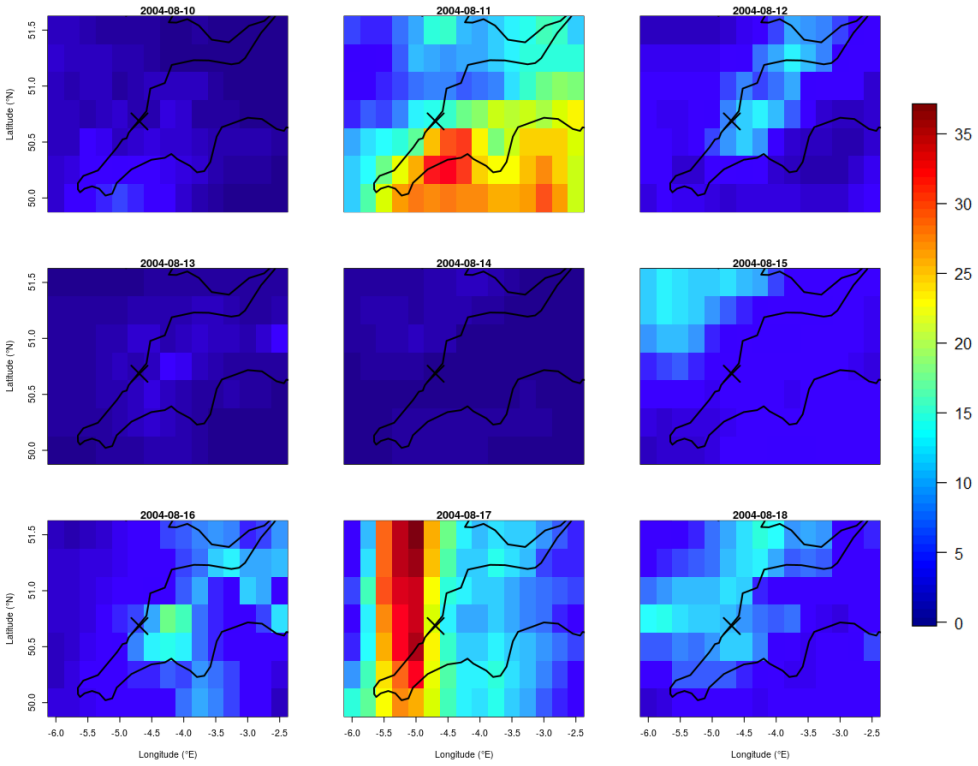


Figure 5: Daily precipitation accumulations (mm) from the ERA5 output for the South West for the days leading up to the flood in Boscastle, and two days afterwards. The location of Boscastle is indicated with a black cross.

The observations show an extremely large precipitation increase that occurred at the time of the floods, as expected. However, whilst the ERA5 output predicted that there would be a large amount of rainfall at this time (around 35 mm of rain), this is a large underestimation of the actual event. If we had downscaling information about ERA5, we would know that it would consistently underestimate extreme rainfall events. Whilst this project does not focus on day-to-day events such as this one, this example makes it clear that the relationship between gridded model output and the observations needs to be taken into account. It would be impossible to have used either data source on its own to predict this event, as the observations are not able to be used for predictions, and the ERA5 data did not predict such a large event.

4 Methodology

Whilst this research primarily focuses on modelling precipitation, this methodology is generally applicable to any form of extreme value modelling. Firstly, a way of parameterising the extreme values is presented through the generalised extreme value (GEV) distribution. This parameterisation allows estimates of return levels, which can be compared between point-level data such as observations and the large scale data from a gridded model output. The discrepancies can be modelled by downscaling, which requires further methodology to

quantify the differences between the two, which is achieved through the use of an additive model. This model captures the discrepancies in space which causes the return levels to be different across the two data sets.

4.1 Extreme Value Theory

4.1.1 Generalised Extreme Value Distribution

The GEV distribution can parameterise extreme values in the form of block maxima, and provide information about the distribution of these maxima. Parameterising extreme values means we can calculate diagnostics and output from a known GEV distribution, based on parameters that describe the location, scale and shape of the distribution.

These block maxima take the form of $M_n = \max\{X_1, ..., X_n\}$, for independent and identically distributed random variables $X_1, ..., X_n$. For stationarity in their realisations, sufficient block size is chosen. Most research opts for a block size of $n = 365$, so that M_n would represent the annual maxima of precipitation, which is what will be modelled in this project. Whilst most environmental variables are expected to have interannual variability, meaning that they are non-stationary within each year, it is reasonable to assume stationarity in the distribution of annual maxima due to its large block size.

Fisher and Tippett [1928] identified three limiting distributions for M_n depending on the value of the shape parameter. Jenkinson [1955] parameterised these three limiting distributions into a singular family of distributions, given in Equation (1).

Theorem 1 *If there exist a sequence of constants* $\{a_n\} > 0$ *and* $\{b_n\}$*, then*

$$Pr\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z)$$

as $n \to \infty$*, then* $G(z)$ *must be of the form*

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

defined on $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$*, where* $-\infty < \mu < \infty$ *is the location parameter,* $\sigma > 0$ *is the scale parameter and* $-\infty < \xi < \infty$ *is the shape parameter [Fisher and Tippett, 1928; Jenkinson, 1955]. See Leadbetter et al. [1983] for proof.*

Theorem 1 means that $G(z)$ approximates the common distribution function of the maxima M_n, provided $G(z)$ is non-degenerate [Fisher and Tippett, 1928]. In practice, the GEV distribution can approximate the distribution of the block maxima of some data, which we take to be annual maxima of precipitation. The GEV distribution has log-likelihood of the

form

$$\ell(\mu,\sigma,\xi) = \begin{cases} -m\log\sigma - (1+1/\xi)\sum_{i=1}^{m}\log\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right] \\ \qquad\qquad -\sum_{i=1}^{m}\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{-1/\xi}, & \text{for } \xi \neq 0, \\ -m\log\sigma - \sum_{i=1}^{m}\left(\frac{z_i-\mu}{\sigma}\right) - \sum_{i=1}^{m}\exp\left\{-\frac{z_i-\mu}{\sigma}\right\}, & \text{for } \xi = 0, \end{cases} \tag{2}$$

for block maxima z_i. This can be maximised for estimates of the GEV parameters $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$.

The return values of the GEV distribution are a high quantile of the data, depending on how large the period is. For a return period p, the corresponding return value is calculated by setting $G(z) = 1 - 1/p$ from Equation (1). These calculations can be found in the Appendix A, so the return levels for the GEV distribution are

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - y_p^{-\xi}\right], & \text{for } \xi \neq 0, \\ \mu - \sigma\log\left(y_p\right), & \text{for } \xi = 0, \end{cases} \tag{3}$$

with $y_p = -\log(1-1/p)$. Return level estimates can be determined by substitution of parameters for their estimates $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$. Since the GEV distribution will model the annual maxima, in our case these return level estimates would represent the expected exceedance of annual precipitation maxima over the next p years. Return values are often the most intuitive output of extreme value analysis, as they are most relevant in application. For example, in engineering, return values are required for the construction of safety barriers that are able to withstand the next 1000 years, which would correspond to the 1000 year return value output from the GEV distribution [Jonathan et al., 2011].

4.1.2 Non-stationarity

The GEV distribution can be extended to allow for non-stationarity, which allows the assumption of variables being identically distributed to be relaxed. This would allow spatial variation into the model, an important consideration for modelling precipitation maxima across space.

The approach to non-stationarity in this project is similar to the one originally described by Smith [1989], which is through the allowance of covariates into the estimates of the parameters. Smith described a basic form of the GEV parameters to be

$$\mu_i = \mu_0 + \mu_1 x_i, \;\; \sigma = \sigma_0, \;\; \xi = \xi_0 \tag{4}$$

so that only the location parameter would depend on a covariate x_i through the parameter μ_1, and intercept parameter μ_0. So instead of estimating μ directly, both μ_0 and μ_1 form a basis which makes up μ.

If s denotes a spatial position, a general spatial GEV model that describes the distribution of Z_t could be of the form

$$Z_t(s) \sim GEV(\mu(s), \sigma(s), \xi(s))$$

where each of $\mu(s)$, $\sigma(s)$ and $\xi(s)$ have an expression in terms of spatial covariates as in Equation (4). Link functions can also be introduced to constrain parameters in a certain way, which relate a parameter to its covariates through a predetermined function. Most research uses an exponential link on the scale parameter [Coles, 2001], so that it is constrained to be above zero, which is one of the criteria for the GEV distribution, from Theorem 1. For the location parameter μ and shape parameter ξ, an identity link function is used, so that there are no constraints.

Spatial covariates will be important in modelling precipitation, and are likely to have the largest effect on describing the distribution of rainfall over the South West [M. L. Wigley et al., 1984]. The log-likelihood in this extended model is the same as that given in (2), with the scalar parameters replaced with vectors that have the same length as the data, for each combination of covariates.

Other forms of non-stationarity include using generalised additive models. This is a methodology that was considered for modelling non-stationarity in space, but over a smaller region such as South West England, a more complex model is not needed to explain the spatial trend. In practice, a simpler parameterisation works just as well and is easier to interpret through simpler terms in the model. It is worth noting however, that if the region were to be larger, it might be necessary to use a more complex model to capture the spatial trend. This is a key assumption: we will not need to use a complex spatial trend, such as a Gaussian process, as our region of interest is small enough that it would not be necessary.

4.1.3 GEV Model

The final GEV model uses spatial dependence in a GEV distribution, across a range of longitudes, $\text{lat}(s)$, latitudes, $\text{lon}(s)$ and elevations, $\text{elev}(s)$. Let $X_{\text{obs},t}(s)$ be the annual maxima of the observations, and $X_{\text{mod},t}(s)$ be the annual maxima of the ERA5 gridded model output, for location s and time t. Then the models take the form

$$\begin{aligned} X_{obs,t}(s) &\sim \text{GEV}(\mu(s), \sigma(s), \xi), \\ \mu(s) &= \mu_0 + \mu_1 \text{lon}(s) + \mu_2 \text{lon}^2(s) + \mu_3 \text{lat}(s) + \mu_4 \text{lat}^2(s) \\ &\quad + \mu_5 \text{lon}(s)\text{lat}(s) + \mu_6 \text{elev}(s), \\ \log(\sigma(s)) &= \sigma_0 + \sigma_1 \text{lon}(s) + \sigma_2 \text{lon}^2(s) + \sigma_3 \text{lat}(s) + \sigma_4 \text{lat}^2(s) \\ &\quad + \sigma_5 \text{lon}(s)\text{lat}(s) + \sigma_6 \text{elev}(s), \\ \xi &= \xi_0, \end{aligned} \tag{5}$$

for the observations, and

$$\begin{aligned} X_{mod,t}(s) &\sim \text{GEV}(\mu(s), \sigma(s), \xi), \\ \mu(s) &= \mu_0 + \mu_1 \text{lon}(s) + \mu_2 \text{lon}^2(s) + \mu_3 \text{lat}(s) + \mu_4 \text{lat}^2(s) \\ &\quad + \mu_5 \text{lon}(s)\text{lat}(s) + \mu_6 \text{elev}(s), \\ \log(\sigma(s)) &= \sigma_0 + \sigma_1 \text{lon}(s) + \sigma_2 \text{lat}(s) + \sigma_3 \text{elev}(s), \\ \xi &= \xi_0, \end{aligned} \tag{6}$$

for the ERA5 gridded model output. This model assumes conditional independence over space and time. The covariates were selected through various techniques. AIC [Akaike, 1975] was used initially, but since interest lies in capturing the spatial trend of return values relating to location and elevation, the main model selection was done by inspection: ensuring that the spatial modelling of return values looked similar to the overall maxima shown in Figure 3.

Starting with the most complex form of the model (one with all covariates and quadratic terms), the model was reduced in the amount of covariates until the spatial trend was still recognised and provided a 'smoother' fit than the stationary GEV distribution. In this case, the best spatial trend appeared in the most complex form of the model for the observations, detailed in Equation (5). For the gridded model output, a simpler form of the scale parameter σ better captured the trend, detailed in Equation (6). The shape parameter ξ is kept constant, because whilst the shape parameter changes significantly with location [Ragulina and Reitan, 2017], the area of land over the South West is not large enough to warrant a varying shape parameter, and the model was kept in a simpler form.

This model is used to calculate estimates of return levels for the observations and gridded model, which are $\hat{RL}_{\text{obs}}(s)$ and $\hat{RL}_{\text{mod}}(s)$ respectively. Return levels are accurate in a stationary climate, but due to the effects of climate change, return levels are expected to be slightly larger than what is calculated by these models. However, the effects of climate change are a low source of uncertainty for extremes, and sampling variability is a much greater source of uncertainty in this case. This is another assumption of this method: climate change will have no significant effect on the return levels.

4.1.4 GEV Model Fit

We can assess the fit of the GEV model through the use of a probability plot [Coles, 2001]. This compares the empirical distribution of the ordered block maxima against its fitted GEV distribution. If the ordered block maxima $z_{(i)}$, and the respective parameter estimates $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$, are substituted into the equation for the GEV family in Equation (1), we get

$$\hat{G}(z_{(i)}) = \begin{cases} \exp\left\{-\left[1 + \hat{\xi}\left(\frac{z_{(i)} - \hat{\mu}}{\hat{\sigma}}\right)\right]^{-1/\hat{\xi}}\right\}, & \text{for } \hat{\xi} \neq 0, \\ \exp\left\{-\exp\left\{-\frac{z_{(i)} - \hat{\mu}}{\hat{\sigma}}\right\}\right\}, & \text{for } \hat{\xi} = 0. \end{cases} \tag{7}$$

These values calculated should be approximately uniform, if the GEV distribution is a good fit to the data. The empirical distribution of $z_{(i)}$ can be calculated as

$$\tilde{G}(z_{(i)}) = \frac{i}{m+1},$$

where i is the index at position i, for data of length m. If the GEV distribution is a good fit, then the empirical distribution will be approximately equal to the fitted distribution. We can compare this in a probability plot, which consists of points

$$\left\{\left(\tilde{G}(z_{(i)}), \hat{G}(z_{(i)})\right), i = 1, ..., m\right\}, \tag{8}$$

so that the x-axis will consist of the empirical distribution of z, and the y-axis will consist of the fitted GEV distribution [Coles, 2001]. If these are approximately equal, the points on this plot should lie close to the unit diagonal.

4.1.5 Calculations and R Code

To estimate the parameters of the GEV distribution, the log-likelihood given in Equation (2) is maximised with respect to $\theta = (\mu, \sigma, \xi)$, to obtain estimates $\hat{\theta} = (\hat{\mu}, \hat{\sigma}, \hat{\xi})$. Non-stationary estimates of parameters were calculated by formulation of model matrices. For example, for n covariates and data of length m, the basis for the location parameter $\boldsymbol{\beta_\mu}$ and model matrix $\boldsymbol{X}$ are composed of

$$\boldsymbol{X} = \begin{pmatrix} 1 & \dots & x_{1,n} \\ 1 & \dots & x_{2,n} \\ \vdots & & \vdots \\ 1 & \dots & x_{m,n} \end{pmatrix}, \qquad \boldsymbol{\beta_\mu} = \begin{pmatrix} \mu_0 \\ \vdots \\ \mu_n \end{pmatrix},$$

and the values that the location parameter would take for each of the data would be

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta_\mu} = \begin{pmatrix} 1 & \dots & x_{1,n} \\ 1 & \dots & x_{2,n} \\ \vdots & & \vdots \\ 1 & \dots & x_{m,n} \end{pmatrix} \begin{pmatrix} \mu_0 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} \mu_0 + \ldots + \mu_n x_{1,n} \\ \mu_0 + \ldots + \mu_n x_{2,n} \\ \vdots \\ \mu_0 + \ldots + \mu_n x_{m,n} \end{pmatrix}. \tag{9}$$

This would be the same for the other parameters, σ and ξ, if they both had covariates. These parameter estimates would maximise the log-likelihood. As explained in Section 4.1.2, link functions are used to constrain parameters in a certain way. For the scale parameter, using an exponential link would result in the inverse link, a logarithmic link on σ, so that the first element of the scale parameter would be

$$\log \sigma_1 = \sigma_0 + \ldots + \sigma_n x_{1,n}.$$

Using identity functions for the location parameter and a constant shape parameter would result in the general GEV distribution of

$$Z_t \sim GEV(\boldsymbol{\mu}, \log \boldsymbol{\sigma}, \xi),$$

where Z_t is some general form of block maxima. Return values can be estimated using the formulas as given in Equation (3). The parameters estimates $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$ were calculated in the same way as described in Equation (9), through dependence on a series of covariates that describe the spatial distribution of the parameters.

Currently, no R package exists that allows accessible fitting of a GEV distribution with this non-stationary method and separate link functions. For this reason, I have developed a range of functions that can

- maximise the GEV likelihood with non-stationary parameters $\mu(s)$, $\sigma(s)$ and $\xi(s)$
- use different link functions that explain the relationship between the parameters and their covariates
- calculate return levels based on these parameters, for each unique set of covariates

The parameter estimates are given an initial condition, and optimised using a BFGS method [Shanno, 1970] using the optimisation function in R. An analytical gradient function was supplied to speed up and provide a more robust convergence, which decreases fitting times by about three-fold compared to a numerical gradient method. The gradients used can be found in Appendix A. The functions that are used to do everything described in this section this are given in Appendix B, and an example of the code is given below.

4.1.5.1 R Code Example

The main function that is used for fitting is `fit.GEV`, which has the general form:

```
fit.GEV(formulas, data, response, rl.n, links)
```

The inputs `formulas` and `links` are lists of three elements, each containing formulae and link functions for the location μ, scale σ, and shape parameter ξ. The input `data` is the data frame that corresponds to the covariates in `formulas`, and `response` is the response variable, which will be in the form of the block maxima of interest. The `rl.n` input represents the n-year return value that will be calculated after the fitting is completed, and is input into this function for ease of calculations, so the return values do not have to be calculated separately.

To start, we ensure the data are set up correctly to go into the fitting function.

```
str(obs)
```

```
## 'data.frame':    4731 obs. of  5 variables:
##  $ lon : num  -3.74 -3.47 -3.35 -3.35 -3.08 ...
##  $ lat : num  51.1 51.2 51.1 51.1 51.1 ...
##  $ elev: int  381 8 271 96 25 20 40 183 21 3 ...
##  $ rain: num  51.8 21.4 33.4 33.1 29 34.4 30 33.3 38.7 33.2 ...
##  $ id  : chr  "1286" "1288" "1290" "1291" ...
```

In this example, `rain` is the response variable we want to use, and `lon`, `lat` and `elev` are the covariates that we want to include in the parameters. We can choose which parameters take which covariates through specification into `formulas`.

```
obs.formula = list(~ lon + lat + elev + lon:lat + I(lon^2) + I(lat^2),
                   ~ lon + lat + elev,
                   ~ 1)
obs.links = list("identity", "exponential", "identity")
obs.fit = fit.GEV(formula = obs.formula, data = obs, response = obs$rain,
                  rl.n = 100, links = obs.links)
```

This function inputs a set of initial conditions for all the parameters, and optimises them to maximise the likelihood given in Equation (2), using the in-built R function `optim`, and the BFGS method [Shanno, 1970]. This has fit a GEV distribution of the form

$$\begin{aligned} M_n &\sim GEV(\mu(s), \sigma(s), \xi), \\ \mu(s) &= \mu_0 + \mu_1 \text{lon}(s) + \mu_2 \text{lat}(s) + \mu_3 \text{elev}(s) + \mu_4 \text{lon}(s)\text{lat}(s) \\ &\quad + \mu_5 \text{lon}^2(s) + \mu_6 \text{lat}^2(s), \\ \sigma(s) &= \sigma_0 + \sigma_1 \text{lon}(s) + \sigma_2 \text{lat}(s) + \sigma_3 \text{elev}(s), \\ \xi &= \xi_0, \end{aligned}$$

where the variables are defined in Equation (5), with the same notation. The output from the fit has two elements, the GEV model itself and the data frame including the calculated return values. The GEV model contains the parameter estimates for all covariates, in the order of μ, σ then ξ, as well as other outputs from the fitting procedure that can be used in other functions (such as model checking).

4.2 Downscaling Extremes

This section explores the methods used to downscale extreme values of precipitation, using output from the GEV distributions fit to gridded model output and observations, explained in Section 4.1. Firstly, the methods used to identify what needs to go into the downscaling model are explained, and secondly, the methods used to model these discrepancies are presented.

The downscaling method will focus on modelling the differences in return levels, similar to an approach by Mannshardt-Shamseldin et al. [2010]. As described in Section 4.1.1, return levels are most important for application, which is the reason they are the focus of downscaling.

4.2.1 Identifying Discrepancies in Return Levels

To identify the forms of systematic discrepancies between gridded model output and observations, two approaches are used. The first approach is based on theory of precipitation variation, and the second approach is by inspection of the return levels of both outputs.

Theory of precipitation variation tells us that the main components are due to variations in elevation and spatial co-ordinates [M. L. Wigley et al., 1984]. Since we know that gridded model output samples at more sparse locations than the observations, it will miss out on key orographical and spatial variation within grid cells, that the observations will capture.

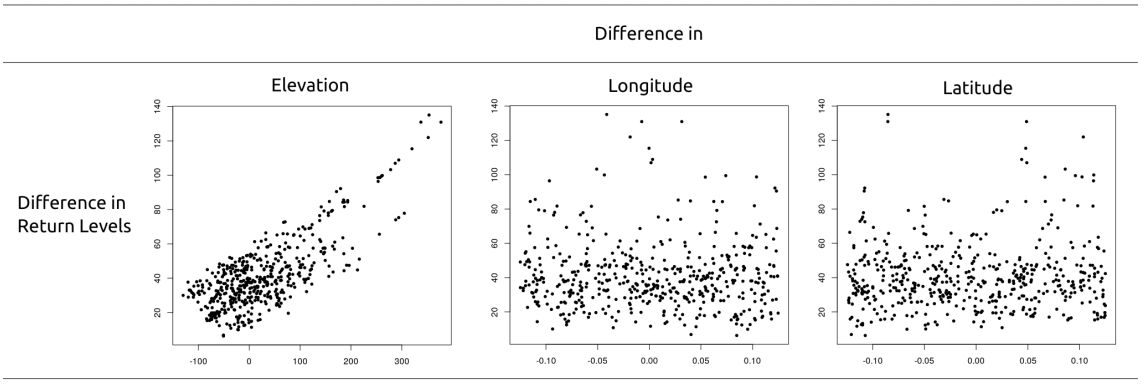


Figure 6: Differences in return levels (mm) for each observation and its corresponding grid cell, plotted against elevation (m), longitude (°E) and latitude (°N), respectively.

For this reason, we know that one of the key differences between the return level estimates will be due to the sampling method. The longitude and latitude of the gridded model are evenly spread, but due to the nature of elevation, it won't be evenly distributed across the sampling points of the gridded model. This leads to the conclusion that elevation could contribute to the largest discrepancy in return levels.

After fitting the spatial GEV model to both model output and observations, we can inspect the differences in elevation, longitude and latitude between the observation locations and the position at which the ERA5 output samples. How these differences relate to differences in return levels is considered. This would be inspected for each observation, and its difference taken from the corresponding grid cell for which the location of the observation is closest to.

Figure 6 shows the relationship between the differences in return levels and differences in elevation, longitude and latitude. The most obvious linear trend can be seen between return level differences and elevation differences, indicating this will be the primary variable of interest in the downscaling model. The trend between longitude and latitude appears to be non-existent, and so further exploration and analysis will be needed to decide whether these variables should be included into the model. This is described in Section 5.2.1.

4.2.2 Downscaling Model

From the relationship shown in Figure 6, we can see an approximately linear trend with return level differences and elevation. A simple parameterisation of these differences can be achieved with a linear model. The linear model can be extended to a generalised additive model (GAM) [Hastie and Tibshirani, 1986], for added flexibility and allowance of non-linear relationships between the return value and the covariates.

A GAM allows the linear predictor of a generalised linear model to depend on unknown smooth functions of some covariates. This semi-parametric approach increases the flexibility of fitting the data in the downscaling model, which is appropriate in this case as there is large variation in space between the observed return values and the return values given by

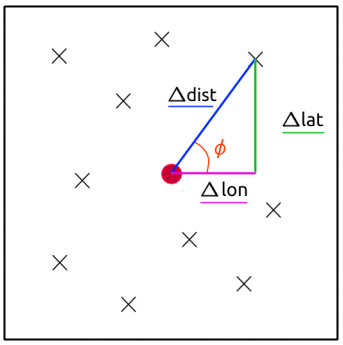


Figure 7: Schematic of the derivation of covariates for the downscaling model. The cross marks represent observations that are located within a particular grid cell. The red dot in the centre represents the mid-point of the grid cell, where the ERA5 model output samples.

ERA5's gridded output.

The general form of the downscaling model is

$$\begin{aligned} \hat{\text{RL}}_{\text{obs}}(s) &\sim N(\nu(s), \sigma^2), \\ \nu(s) &= \beta_0 + f_1(\hat{\text{RL}}_{\text{mod}}(s)) + f_2(\Delta\text{elev}(s)) + f_3(\Delta\text{lon}(s), \Delta\text{lat}(s)) \\ &\quad + f_4(\text{elev}(s)) + f_5(\text{lon}(s), \text{lat}(s)) + \beta_1 \Delta\text{dist}(s) + \beta_2 \phi(s), \end{aligned} \tag{10}$$

where s denotes spatial position, the variables superseded with Δ represent the difference between the variable from the ERA5 output grid point and the corresponding location of each observation which is closest to a particular grid cell mid-point. The covariate $\Delta\text{dist}(s)$ is the distance from the mid-point to the observation, and $\phi(s)$ is the angle from the horizontal to the same location. The derivation of these covariates (with exception to $\Delta\text{elev}(s)$) can be seen in Figure 7. Here, $\hat{\text{RL}}_{\text{obs}}(s)$ represents the estimated return level from the observations, and $\hat{\text{RL}}_{\text{mod}}(s)$ is the estimated return level from the ERA5 output. The functions $f_1, ..., f_5$ are unknown functions estimated from the data, and β_1 and β_2 are parameters that need to be estimated that describe the effect of $\Delta\text{dist}(s)$ and $\phi(s)$ respectively.

The model described here in Equation (10) is chosen based on a combination of choosing covariates that result in the smallest AIC [Akaike, 1975] and root mean squared error (RMSE). This enables model selection that would give the best model fit according to AIC, and provide predictions that are close to the actual observation return values. This is detailed further in Section 5.2.1.

To fit this GAM model, the `mgcv` R package will be used [Wood, 2017]. This allows easy fitting of a generalised additive model, through specification of a formula, which includes specification of covariates to use and how the unknown functions $f_1, ..., f_5$ are estimated. In this case, thin plate regression splines are used for this estimation [Wood, 2003], but similar results can be obtained through different splines. See Wood [2017] for an overview of different fitting procedures for GAMs.

5 Results

5.1 Extreme Values

This section will discuss the parameterisation of the extreme values of precipitation in the South West, using the methods described in Section 4.

5.1.1 GEV Distribution Fit

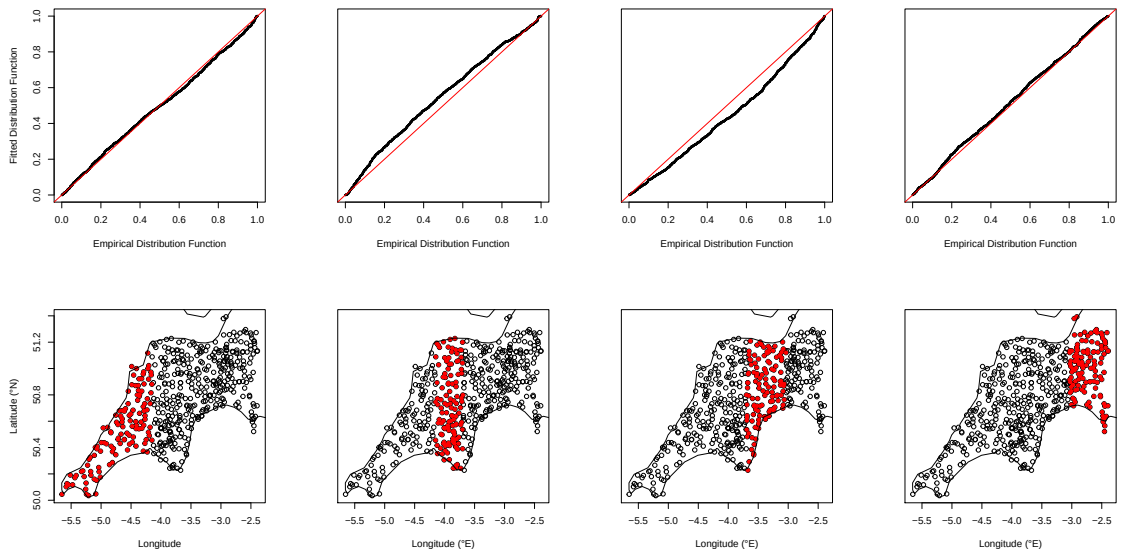
The form of the GEV model that has been fit is described in Equations (5) and (6), in Section 4.1.3. Table 1 shows the parameter estimates obtained by maximising the log-likelihood of the GEV distribution with respect to each covariate. The parameter estimates give an indication of the differences between the observations and ERA5. For example, there are different elevation effects for the estimates of the location and scale parameter; the parameter estimates for the observations are positive, but those of the ERA5 output are negative, showing the differing effects of elevation between the observations and the gridded-scale data. ERA5 also has a much larger parameter estimate for the effect of longitude than the observations do. The other parameter estimates have similar values, or are around the same scale.

		Covariate Estimates						
		Intercept	lon(s)	lat(s)	elev(s)	lon^2(s)	lat^2(s)	lon(s)lat(s)
Observations	$\hat{\boldsymbol{\mu}}$	34.250	−0.766	9.859	0.0440	−0.202	−0.176	−1.435
	$\hat{\boldsymbol{\sigma}}$	2.895	−2.179	0.0294	0.001 91	−0.001 42	0.0291	−0.0946
	$\hat{\boldsymbol{\xi}}$	0.123	-	-	-	-	-	-
ERA5	$\hat{\boldsymbol{\mu}}$	26.847	38.635	1.742	−0.0247	−0.0394	−0.893	−0.741
	$\hat{\boldsymbol{\sigma}}$	−2.589	−0.0443	0.0825	−0.001 01	-	-	-
	$\hat{\boldsymbol{\xi}}$	0.0525	-	-	-	-	-	-

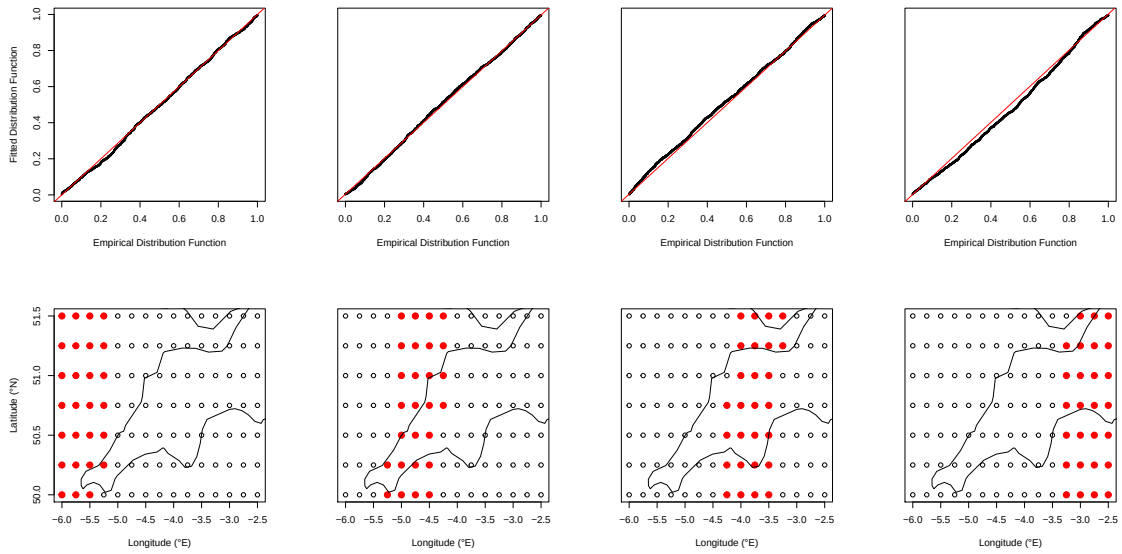
Table 1: Table of parameter estimates given by the GEV models as described by the models from Equations (5) and (6). This shows the values of parameter estimates that were assigned to each covariate for each parameter, and a dash means that that particular parameter did not contain that covariate in the model.

Another aspect of this model to note is the positive shape parameter for both observations and ERA5. From the definition of the GEV distribution, maximum likelihood methods give a confident fit provided that $\xi > -0.5$ [Coles, 2001]. This condition has been met, so we are satisfied the methods used to obtain these parameter estimates are justified. A positive shape parameter implies that there is no upper bound on rainfall estimates, so that for high return periods, the estimated return value will tend towards infinity. This is not a problem in our case however, as we consider 100 year return periods, which are not large enough for the return value to increase beyond expected levels.

The model fit of the GEV distribution can also be considered, as described in Section 4.1.4. Figure 8 shows the calculated probability plots for the South West, where each of the four plots are for a separate section of the South West, so that the model fit can be



(a) Probability plots for the observations.



(b) Probability plots for the ERA5 gridded model output.

Figure 8: Probability plots as defined by Equation (8) for the GEV distribution fit to the observations (top) and ERA5 model output (bottom) over the South West, split into four sections. The sections which each plot refers to is shown below the corresponding probability plot.

considered in more than one place. For some sections the points do not lie on the diagonal line exactly, suggesting that there may be some small failures of the GEV model in these areas. However, these do not seem to show systematic bias, as there is no region dominated by a particular physical attributes that would cause this failure. The majority of these probability plots give no reason to doubt the validity of the fitted model, as the points mostly lie on the diagonal line, showing that the empirical distribution is approximately equal to the fitted distribution. This leads to an increase in confidence in the fitted GEV model for both the observations and the gridded model output.

5.1.2 Estimated Return Levels

To justify the use of a spatial GEV model, as opposed to a non-stationary one, consider a stationary GEV model that has no covariates, so that there is a separate GEV distribution fit to each station and each grid cell. The return levels calculated from this stationary GEV fit are shown in Figure 9.

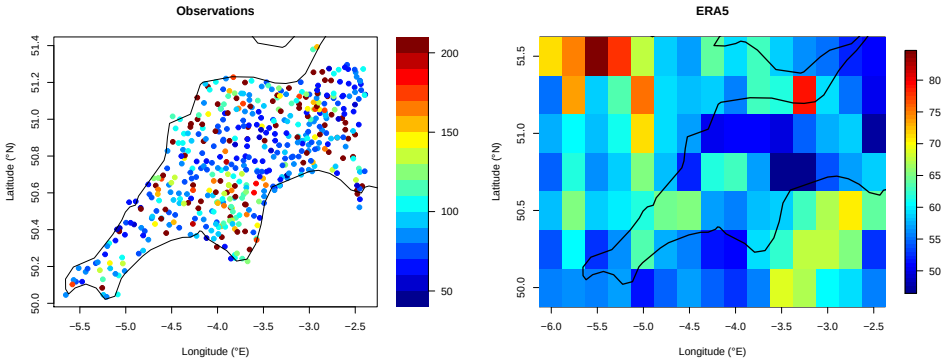


Figure 9: Estimated precipitation return levels (mm) calculated from fitting a stationary GEV model to each station/grid cell, for observations (left) and ERA5 model output (right).

There is no significant spatial variation shown in either of these return level plots in Figure 9, and they do not look similar to the overall maxima shown in Figure 3. For this reason, a spatial non-stationary GEV model as described in Equations (5) and (6) is necessary to calculate more accurate return levels. The return levels calculated from this spatial GEV model are shown in Figure 10.

The spatial variation in Figure 10 is more similar to that in the overall maxima in Figure 3, and so an initial inspection favours this model over its stationary counterpart. Since rainfall varies significantly with space and elevation, there is also grounds for the non-stationary GEV model instead of the stationary one on a theoretical basis. One of the interesting aspects of the return levels in Figure 10 is the large difference in size between the two. The observations take on much larger values of precipitation than the model output. This is expected due to our pre-existing knowledge that the measured precipitation in the observations are on average much higher than the derived precipitation from the ERA5 output. This highlights the necessity for downscaling, further demonstrated in Figure 11.

This disparity is likely due to the differences in elevation between the observations and

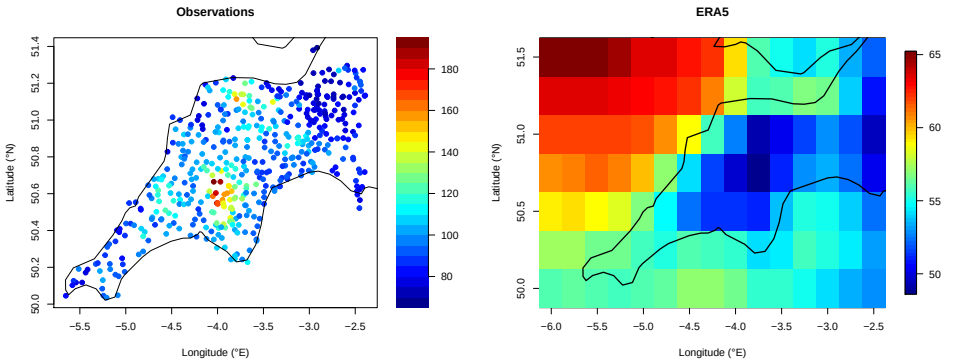


Figure 10: Estimated precipitation return levels (mm) calculated from fitting a non-stationary GEV model across all locations, described in Equations (5) and (6), with parameter estimates from Table 1. For observations (left) and ERA5 model output (right).

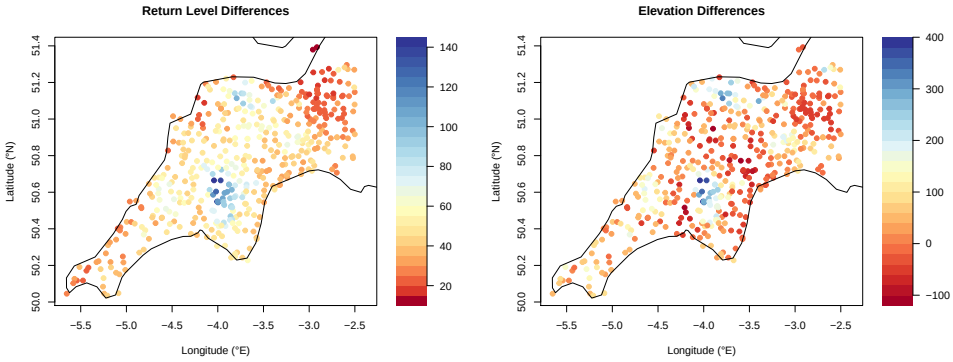


Figure 11: Differences in estimated precipitation return levels (mm) between observations and ERA5 interpolated to observation locations (left). Differences in elevations (m) between observations and ERA5 interpolated to observation locations (right).

the model output, caused by the sampling pattern of the ERA5 reanalysis. At points of lower elevation, or at positions where the ERA5 model output samples, the return values are much more similar. This is shown in the differences plot in Figure 11. The differences in the return levels show the same spatial pattern to the differences in the elevations, exemplifiying the problem of elevation being the primary source of differences between model output and observations.

5.2 Downscaling Extremes

5.2.1 Downscaling Model Selection

Figure 12 shows the estimated return levels from the observations and from the ERA5 output. The vertical lines of points in this plot show the large variation of observation return levels that occur within each grid cell. To explain this within grid cell variation, spatial covariates are introduced.

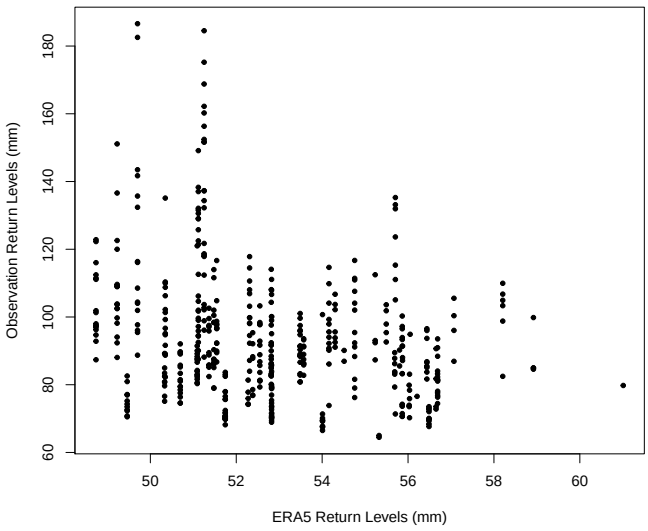


Figure 12: Plot of observation return levels against ERA5 return levels, where each ERA5 return level corresponds to a set of observation return levels of which the location is closest to that grid cell's midpoint. Return levels are calculated from the GEV model detailed in Sections 4.1.3 and 5.1.2.

	Covariates							
	elev(s)	lon(s), lat(s), elev(s)	Δelev(s)	Δlon(s), Δlat(s), Δelev(s)	lon(s), lat(s), elev(s), Δlon(s), Δlat(s), Δelev(s)	lon(s), lat(s), elev(s), Δlon(s), Δlat(s), Δelev(s), $\phi(s)$	lon(s), lat(s), elev(s), Δlon(s), Δlat(s), Δelev(s), Δdist(s)	lon(s), lat(s), elev(s), Δlon(s), Δlat(s), Δelev(s), Δdist(s), $\phi(s)$
RMSE	1.77478	0.264422	2.00227	1.69405	0.251372	0.248692	0.248382	**0.247813**
AIC	2293.92	654.526	2485.65	2402.02	589.516	589.923	**589.272**	590.189

Table 2: RMSE and AIC values for the downscaling GAM model $\hat{RL}_{\text{obs}}(s) \sim \beta_0 + f_1(\hat{\text{RL}}_{\text{mod}}(s)) + \ldots$, with additional covariates also being unknown functions, labelled in the columns. The lowest (best) scores are highlighted in bold.

To justify the covariate choices in the downscaling model, Table 2 shows different combinations of covariates that the downscaling model could include, and the resulting RMSE and AIC that each respective model gives. RMSE is defined as the absolute difference between the predicted and actual observation return levels. The best scoring

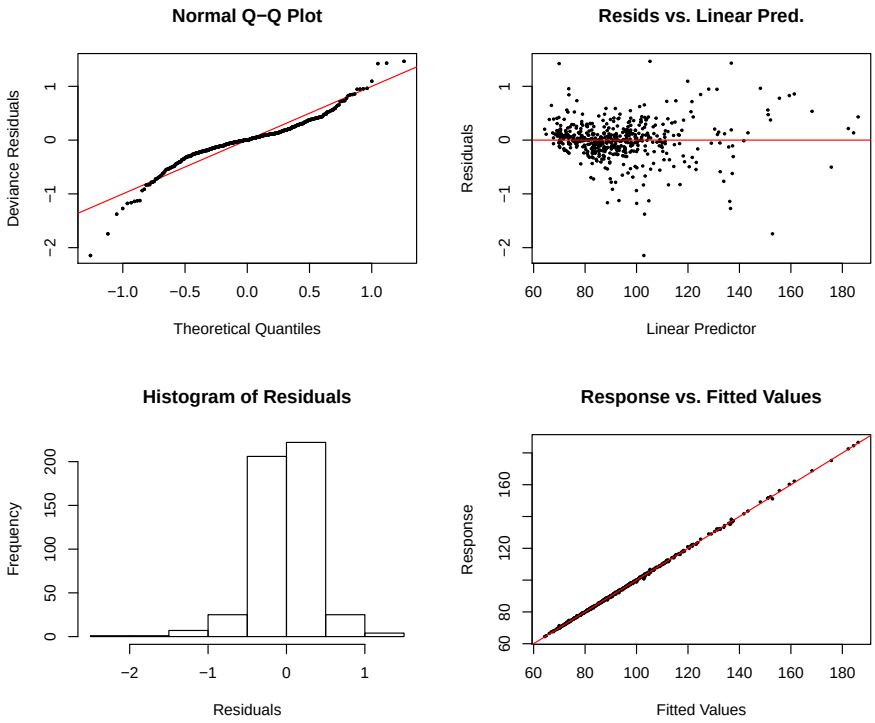


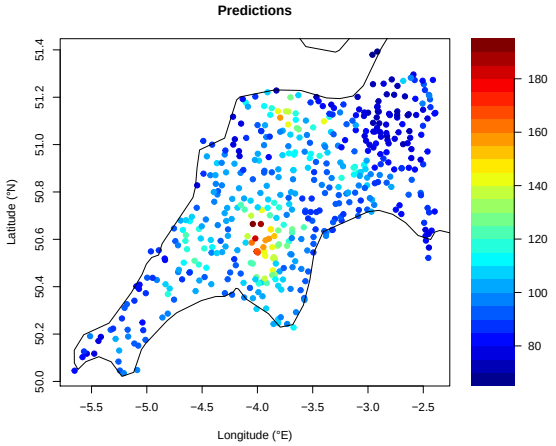
Figure 13: Diagnostic plots for the downscaling GAM as described in Equation (10).

model in terms of AIC was different to the best model in terms of RMSE. However, the difference in AIC between these models is less than 1, and since this difference is almost negligible, the model with a better predictive score is used. For this reason, the model in the final column, which takes all possible covariates is used for downscaling.

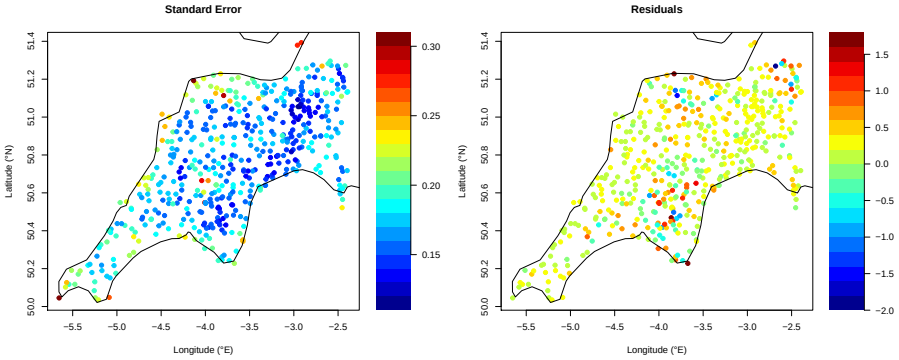
5.2.2 Downscaling Model Fit

The model fit of the GAM can be identified by inspection of the residuals. A Normal Q-Q plot can be used to compare the distribution of the quantiles of the residuals against the theoretical quantiles. For a good fit, the Q-Q plot will have points which lie upon a diagonal line. The residuals can be also checked for their Normality through a residuals versus linear predictor plot, and a histogram. For a good model fit, the residuals will be approximately Normally distributed, and so the histogram should appear as a probability density function of a $N(0, 1)$ distribution, provided the model fit is good. The final way to judge the model fit of a GAM is by its predictions. If the predictions of the GAM are close to the response variable, the plot of the response against the fitted values should be approximately diagonal, so that the predictions map to the response exactly. These diagnostic plots can be seen in Figure 13.

Figure 13 meets these criteria within an acceptable degree of accuracy. The Normal Q-Q plot has points that lie on the diagonal line for the most part, although there is deviation at the tails, specifically on the lower tail. The residuals seem to be Normally distributed, as



(a) Predicted observation return levels.



(b) Standard errors. **(c)** Residuals from the downscaling model.

Figure 14: Predictions and diagnostics from the downscaling model from Equation (10). The RMSE of these predictions in (a) is 0.247813.

the histogram of residuals seem to follow approximately a $N(0, 1)$ distribution, but there is more mass concentrated at zero than at the tails. There are no systematic deviations from the mean line at zero in the residuals versus linear predictor plot, which gives confidence in the model. The response versus fitted values plot gives credibility to the predictions from the model, as the points lie almost exactly on the diagonal line, and so the predictions almost map exactly to the response variable. Further analysis of the predictions are given in Section 5.2.3. Whilst there appears to be a shortcoming of the GAM at the lower tail of the distribution, it is not significant enough to discredit the model. Therefore, from these model checking plots it is clear that the GAM explains the differences between the different return levels well through incorporation of spatial covariates, although there is room for improvement. This is discussed further in Section 6.3.

5.2.3 Downscaled Return Levels

This downscaling model can be used to provide predictions at the same locations of the observations. By using the longitudes, latitudes, elevations, and the corresponding differences of these from their respective ERA5 grid point, the predicted differences in return levels can be obtained by using the parameter esimates. The predictions are shown in Figure 14a.

These predictions are obtained using spatial characteristics of the locations, and the ERA5 return levels only, using the downscaling model as information from the pre-existing relationship from observation return levels. The map showing the predicted observation return levels in Figures 14a are extremely similar to the actual observation return levels in Figure 10, showing the skill of the predictions in the downscaling model. The RMSE seems low upon inspection, but this is expected as the covariate of estimated return level from the ERA5 output is included, which itself is an estimate of observation return levels.

The standard errors in Figure 14b show higher uncertainty at the coast line, and some areas of uncertainty at high elevation points. This is similar in the residuals plot in Figure 14c, where the residuals from the GAM model are larger at higher elevation regions. This is likely due to there being a larger variability of precipitation in these areas, and the downscaling model captures this variability correctly. The residuals are not large enough to cause concern, as they do not leave the 95% confidence interval of a Normal distribution (do not go above 2 nor below -2).

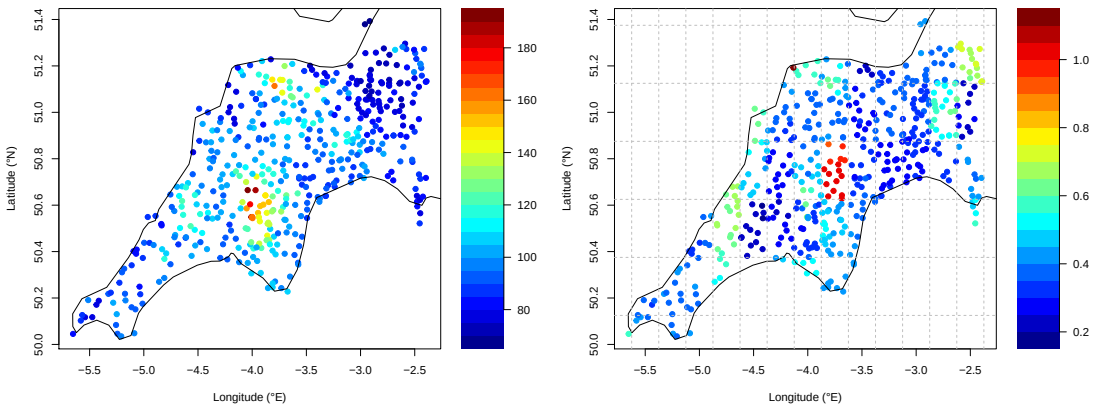
6 Discussion and Conclusions

6.1 Discussion

6.1.1 Sensitivity Analyses

This project is concerned with downscaling extreme precipitation in South West England, but the methods used could be applicable to other regions. To simulate the predictive skill of another region, or for an area outside of the downscaling region, a form of cross validation can be used. This approach will predict a smaller area of observation return levels using information from outside the area of prediction. This is achieved by the process:

1. Fit the GEV model to the ERA5 data for all grid cells and calculate the return levels.
2. Remove all the observations that correspond to a singular grid cell.
3. Fit the GEV model to the remaining observations and calculate the return levels.
4. Fit the downscaling model to these return levels and the corresponding ERA5 return levels, including all covariates.
5. Use the estimates from the downscaling model to predict the return values of the removed observations.
6. Repeat for all grid cells and hence for all observations.



(a) Fitted return levels.

(b) Standard error.

Figure 15: (a) Predicted observation return levels (mm) and (b) the standard error, as calculated by the cross-validation method for the sensitivity analyses. The grid structure in (b) represents the gridded structure of the ERA5 output.

This can be used to estimate return levels for the region using only predictions that were not used to derive the model itself. For this reason, this method simulates predicting for an area that is outside the domain of interest, and would produce comparable results to if the downscaling model were used for a different region.

Using this method, we get a RMSE of 0.555, which is slightly higher than the RMSE for predictions within the downscaling model, shown earlier as 0.247. The map of return levels calculated by this method can be seen in Figure 15a, and the standard errors can be seen in Figure 15b. The fitted return levels look similar to those predicted by normal methods, seen in Figure 14a. The spatial structure of extreme precipitation is still intact, but the uncertainty has increased by a significant margin. This increased uncertainty is expected, and is different for each grid cell, shown by the gridded pattern of uncertainty in the figure.

The accuracy of predictions in this cross-validation prediction method is still high compared to the actual observation return levels, showing the integrity of the downscaling model in predicting for areas outside of the domain of the model. The next step in this regard would be to test the same model fit to the South West on a different region of a similar size.

6.1.2 Comparison to Other Literature

The methodology presented by Mannshardt-Shamseldin et al. [2010] is similar to the methodology provided in this research. Mannshardt-Shamseldin et al. use a GEV distribution with a threshold exceedance approach to model extremes of precipitation, and so use the same parameters of the GEV distribution. The region in this case is the continent of North America, so it is over a much larger area than the South West. However, the disparity in return levels between the gridded model output and observations is similar to those presented here.

Kallache et al. [2011] model the empirical CDF of point-level and gridded-level precipitation data across five different stations, and show that their forms differ a lot. Whilst Kallache et al. do not calculate the return levels in their work, the disparity between the empirical CDFs of the extremes imply that the return levels would also be significantly different. The extreme value results presented in this research concurs with results presented by Mannshardt-Shamseldin et al. [2010] and Kallache et al. [2011].

Mannshardt-Shamseldin et al. use a linear model with covariates of elevation and cubic polynomials of longitude and latitude for downscaling. The results are similar to the research presented here: the fitted values of observation return levels show the same spatial pattern that the actual return levels do. Using only spatial covariates in the downscaling model succeeded in modelling the differences while keeping the spatial variation of the observations intact. Mannshardt-Shamseldin et al. [2010] is the only existing research that compares downscaled return levels between gridded output and observations. Since these results agree with those presented by Mannshardt-Shamseldin et al. [2010], it inspires confidence that they are indeed valuable.

This work also expands on the methodologies presented by Mannshardt-Shamseldin et al. [2010], by not considering averaged observation return levels for each grid cell. Instead, we use all observation locations and include the differences in spatial position from the grid cell mid-point. Further extension of the work of Mannshardt-Shamseldin et al. includes the use of a generalised additive model instead of a regular linear model, improving the flexibility of the downscaling.

6.2 Conclusions

This research has presented a relatively parsimonious way to parameterise extreme values of precipitation over South West England. Using a spatially varying GEV distribution, extreme precipitation variation over the space can be characterised by the distribution of their maxima. The GEV model was found to be stable over the small region, and captured the rainfall maxima well. Only quadratic terms in longitude and latitude were required to capture the spatial trend over the smaller area, and a more complex model was not necessary in this case.

The 100-year return levels calculated from the GEV models vary most significantly with elevation in the observations, but due to the sampling pattern of the gridded output, elevation played the opposite role in informing about return levels for the ERA5 data. The significance of the effects from the longitude and latitude covariates is also different between the observations and the gridded model, all of which contributes to a large difference in calculated return levels. Return levels estimated from ERA5 are significantly lower than those from the observations, meaning that ERA5 does not predict the extremes of precipitation accurately in the South West.

The disparity between these 100-year return levels are modelled successfully through a generalised additive model framework, using covariates of grid-level return levels and spatial characteristics. The predictions from this model are highly accurate: for return levels up to the scale of 180 mm, the mean error of the predictions was only around 0.25 mm. This shows the predictive power of the downscaling GAM.

The downscaling model used only spatial covariates to explain the variation of return levels for each model grid cell, and it provided accurate predictions that were indistinguishable from the actual observation return levels. This shows that the return levels can be modelled as a function of their spatial properties only, and still provide accurate predictions. This means that spatial covariates are the most important variable to consider when modelling extreme values of precipitation.

6.3 Future work

Climatological properties would also help to explain the variation in precipitation [Cooley et al., 2007], but are not included in the interest of choosing a simple model over a small region that would likely not benefit from the added complexity. Other physical aspects of precipitation over land, such as precipitation occuring on the windward side of mountains, were not considered during this project. Only a simple model of high elevation causing high precipitation was included into the statistical model describing extreme rainfall patterns. Since most wind in the UK blows in a westerly direction, a west-to-east effect of elevation could improve the spatial characteristics of the GEV model. This could be implemented by adding an additional covariate to each station (grid cell) which is the elevation of the station (grid cell) closest East.

The choice of covariates for the GEV model could also be improved similar to a method implemented by Kallache et al. [2011], who choose covariates dependent on each station. A different set of covariates could be chosen depending on the characteristics of each station or grid cell, which could improve the spatial distribution of return levels.

Seasonality has also not been considered in this project. Since the methods include the use of a GEV distribution to model annual precipitation maxima, if the maxima tend to consistently be within a particular season, then that season will be preferred over the others. A threshold exceedance method would be appropriate for reducing the bias introduced by lack of including seasonality into the model, but is not as simple to implement as modelling maxima. Other methods to deal with seasonality include performing separate analyses by season, similar to the approach by Mannshardt-Shamseldin et al. [2010], or introducing seasonally varying parameters into the GEV distribution, and using monthly maxima instead of annual maxima, described by Davison and Huser [2015].

Inspection of Figure 14 reveals information about how the downscaling model could be further improved. The standard errors in Figure 14b shows larger uncertainty at the coast lines. The likely source of this uncertainty is due to there being fewer observations at these points that correspond to a single grid cell, but additional covariates could be included to explain this uncertainty and provide more accurate estimates. In this research, covariates of distance to coast and proportion of grid box that contains land were experimented with to reduce the uncertainty, but neither were successful. With more time for experimentation, a more accurate downscaling model could be achieved.

The uncertainty in the fitted values of the GAM, presented in Figure 14, is not entirely accurate. The uncertainties here are only from the downscaling model itself, whilst there is also uncertainty involved in the GEV model (not described). The GEV model estimates are

treated as fixed for the purposes of downscaling, but their uncertainties could be included in addition to the uncertainties of the fitted return levels to better quantify them.

The model diagnostic plots from Figure 13 also demonstrate that the downscaling model is not performing to its full potential. The residuals do not follow a strict $N(0, 1)$ distribution, as most of the mass is concentrated near zero, instead of having an even distribution across the centre and the tails of the distribution. It seems upon inspection of the Q-Q plot that there is something missing that could explain the deviation from the diagonal line. Further experimentation could include additional covariates, or different fitting techniques, such as alternative splines instead of thin plate regression splines.

References

Akaike, H. (1975, 01). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on 19*, 716 – 723.

Boudrissa, N., H. Cheraitia, and L. Halimi (2017). Modelling maximum daily yearly rainfall in northern algeria using generalized extreme value distributions from 1936 to 2009. *Meteorological Applications 24*(1), 114–119.

Buishand, T. A. (1991). Extreme rainfall estimation by combining data from several sites. *Hydrological Sciences Journal 36*(4), 345–365.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. London: Springer-Verlag.

Cook, N. (1982). Towards better estimation of extreme winds. *Journal of Wind Engineering and Industrial Aerodynamics 9*(3), 295 – 323.

Cooley, D., D. Nychka, and P. Naveau (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association 102*(479), 824–840.

Copernicus Climate Change Service, C. (2017 (accessed November 2, 2018)). Era5: Fifth generation of ecmwf atmospheric reanalyses of the global climate.

Davison, A. and R. Huser (2015). Statistics of extremes. *Annual Review of Statistics and Its Application 2*(1), 203–235.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart (2011). The era-interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society 137*(656), 553–597.

Déqué, M. (2007). Frequency of precipitation and temperature extremes over france in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global and Planetary Change 57*(1), 16 – 26. Extreme Climatic Events.

Fisher and Tippett (1928). On the estimation of the frequency distributions of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society 24*, 180–190.

Groisman, P. Y., T. R. Karl, D. R. Easterling, R. W. Knight, P. F. Jamason, K. J. Hennessy, R. Suppiah, C. M. Page, J. Wibig, K. Fortuniak, V. N. Razuvaev, A. Douglas, E. Førland, and P.-M. Zhai (1999, May). Changes in the probability of heavy precipitation: Important indicators of climatic change. *Climatic Change 42*(1), 243–283.

Gumbel, E. (1958). *Statistics of Extremes*. Columbia University Press, New York.

Hastie, T. and R. Tibshirani (1986). Generalized additive models. *Statistical Science 1*(3), 297–310.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society 81*(348), 158–171.

Jonathan, P., K. Ewans, and J. Flynn (2011, 01). On the estimation of ocean engineering design contours. *Journal of Offshore Mechanics and Arctic Engineering 136*.

Kallache, M., M. Vrac, P. Naveau, and P.-A. Michelangeli (2011). Nonstationary probabilistic downscaling of extreme precipitation. *Journal of Geophysical Research: Atmospheres 116*(D5).

Kjeldsen, T., D. Jones, and A. Bayliss (2008). *Improving the FEH statistical procedures for flood frequency estimation: Science Report: SC050050*. Denmark: European Environment Agency (EEA).

Leadbetter, M., G. Lindgren, and H. Rootzén (1983). *Extremes and Related Properties of Random Sequences and Processes*.

M. L. Wigley, T., J. Lough, and P. Jones (1984, 01). Spatial patterns of precipitation in england and wales and revised england and wales precipitation time series. *Journal of Climatology 4*, 1 – 25.

Mannshardt-Shamseldin, E. C., R. L. Smith, S. R. Sain, L. O. Mearns, and D. Cooley (2010, 03). Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data. *Ann. Appl. Stat. 4*(1), 484–502.

Maraun, D., H. W. Rust, and T. J. Osborn (2009). The annual cycle of heavy precipitation across the united kingdom: a model based on extreme value statistics. *International Journal of Climatology 29*(12), 1731–1744.

Met-Office (2011 (accessed March 3, 2019)). Boscastle flood. `https://www.metoffice.gov.uk/binaries/content/assets/mohippo/pdf/6/boscastle_flood_-_16_august_2004.pdf`.

Met-Office (2012 (accessed November 2, 2018)). Met office integrated data archive system (midas) land and marine surface stations data.

Michelangeli, P.-A., M. Vrac, and H. Loukos (2009). Probabilistic downscaling approaches: Application to wind cumulative distribution functions. *Geophysical Research Letters 36*(11).

Pickands (1975, 01). Statistical inference using extreme order statistics. *Ann. Statist. 3*(1), 119–131.

Ragulina, G. and T. Reitan (2017). Generalized extreme value shape parameter and its nature for extreme precipitation using long time series and the bayesian approach. *Hydrological Sciences Journal 62*(6), 863–879.

Rubin, D. B. (1976, 12). Inference and missing data. *Biometrika 63*(3), 581–592.

Santos, E. B., P. S. Lucio, and C. M. Santos e Silva (2015). Seasonal analysis of return periods for maximum daily precipitation in the brazilian amazon. *Journal of Hydrometeorology 16*(3), 973–984.

Shanno, D. (1970). Conditioning of quasi-newton methods for function minimization. *Math. Comput. 24*(111), 647–656. cited By 6.

Smith, R. L. (1989, 11). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Statist. Sci. 4*(4), 367–377.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R* (2 ed.). Chapman and Hall/CRC.

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B) 65*(1), 95–114.

Appendix

A GEV distribution calculations

We can calculate the return levels for the GEV distribution by setting $G(z) = 1 - 1/p$ from Equation (1). For the $\xi \neq 0$ case, we have

$$\begin{aligned} G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} &= 1 - \frac{1}{p} \\ \implies \left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi} &= -\log\left(1-\frac{1}{p}\right) \\ \implies -\frac{1}{\xi}\log\left[1+\xi\frac{z-\mu}{\sigma}\right] &= \log\left[-\log\left(1-\frac{1}{p}\right)\right] \\ \implies 1+\xi\left(\frac{z-\mu}{\sigma}\right) &= \left(-\log\left(1-\frac{1}{p}\right)\right)^{-\xi} \\ \implies z_p &= \mu - \frac{\sigma}{\xi}\left[1-\left(-\log\left(1-\frac{1}{p}\right)\right)^{-\xi}\right], \end{aligned}$$

and for the $\xi = 0$ case, we have

$$\begin{aligned} G(z) = \exp\left\{-\exp\left\{-\frac{z-\mu}{\sigma}\right\}\right\} &= 1 - \frac{1}{p} \\ \implies -\frac{z-\mu}{\sigma} &= \log\left(-\log\left(1-\frac{1}{p}\right)\right) \\ \implies z_p &= \mu - \sigma\log\left(-\log\left(1-\frac{1}{p}\right)\right). \end{aligned}$$

The gradients of the GEV log-likelihood are also used to speed up and provide a more robust convergence when finding parameters that maximise the log-likelihood. These are

$$\frac{\partial \ell}{\partial \mu} = -\sum_{i=1}^{N} \frac{\left(1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right)^{-1/\xi} - \xi\left(1+\frac{1}{\xi}\right)}{\sigma\left(1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right)},$$

$$\frac{\partial \ell}{\partial \sigma} = -\sum_{i=1}^{N} \frac{1}{\sigma}\left(\left(\frac{1}{\left(1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right)^{1/\xi}} - \xi\left(1+\frac{1}{\xi}\right)\right)\left(\frac{z_i-\mu}{\sigma\left(1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right)}\right) + 1\right)$$

$$\begin{aligned} \frac{\partial \ell}{\partial \xi} = -\sum_{i=1}^{N} \frac{1}{\xi}\Bigg(\left(1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right)^{-1/\xi} - 1\Bigg)\frac{1}{\xi}\log\left(1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right) - \frac{z_i-\mu}{\sigma\left(1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right)^{1+1/\xi}}, \\ + \left(1+\frac{1}{\xi}\right)\left(\frac{z_i-\mu}{\sigma\left(1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right)}\right). \end{aligned}$$

B R Code: GEV Functions

Below are the functions that have been used to fit the GEV distribution, of which their purpose is expained in Section 4.1.5. Some of these functions are specific to the data set

that was used for this project, but can be easily adapted to be more general if need be.

Firstly, the following functions are the 'smaller' functions that tend to be called in the 'larger' functions. These include the gradient function, a function to get initial conditions, and various others.

```
GEVingrad <- function(mu, sig, xi,z){

  muterm1 = (1 + xi*((z-mu)/sig))^(-1/xi)
  muterm2 = xi*(1+(1/xi))
  muterm3 = sig*(1+xi*((z-mu)/sig))

  sigterm1 = (1 + xi*((z-mu)/sig))^(-1/xi)
  sigterm2 = xi*(1+1/xi)
  sigterm3 = (z-mu)/(sig*(1+xi*((z-mu)/sig)))

  xiterm1 = (1/(1+xi*((z-mu)/sig))^(1/xi)) - 1
  xiterm2 = log1p(xi*((z-mu)/sig))/xi
  xiterm3 = (z-mu)/(sig*(1+xi*((z-mu)/sig))^(1+(1/xi)))
  xiterm4 = (1 + (1/xi))*((z-mu)/(sig*(1+xi*((z-mu)/sig))))

  list(mu = -((muterm1 - muterm2)/muterm3),
       sig = -(1/sig)*((sigterm1 - sigterm2) * sigterm3 + 1),
       xi = - ( (xiterm1*xiterm2-xiterm3 )/xi + xiterm4))
}

getLinkFunction = function(link="identity"){
  # used in fitting GEV, returns R functions for multipling model matrices
  if(link == "identity") outf = function(B,X) X %*% B
  if(link == "exponential") outf = function(B,X) exp(X%*%B)
  if(link == "none") outf = function(B,X) X[,1] * B[1]
  return(outf)
}

getBetaIndex = function(mulink,siglink,xilink,p){
  # used in fitting GEV, gets indices of where parameters appear
  # depending on link function
  mup = 1; sigp = 1; xip = 1
  if(mulink!="none") mup = p
  if(siglink!="none") sigp = p
  if(xilink!="none") xip = p
  bindex = c(rep(1,mup),rep(2,sigp),rep(3,xip))
  return(bindex)
}

getInit = function(z,mup,sigp,xip,siglink="exponential"){
```

```
  # Gets initial conditions for fitting in optim (taken from gev.fit)

  init = numeric(sum(mup,sigp,xip))
  init[mup+1] = sqrt(6*var(z))/pi

  init[mup+sigp+1] = 0.1
  init[1] = mean(z) - 0.57722*init[mup+1]
  if(mup>1) if(siglink=="exponential" && init[mup+1]!=0) {
    init[mup+1] = log(init[mup+1])
  }

  return(init)
}

GEV_getParSums = function(gevmodel,data){
  # need model matrices for each cov so that know which covariates to sum
  # needs very specific inputs, that is output from GEVfitX_diffX
  # covariates is data frame of all variables

  mu  = (gevmodel$par[gevmodel$bindex==1])
  sig = (gevmodel$par[gevmodel$bindex==2])
  xi  = (gevmodel$par[gevmodel$bindex==3])

  munames  = colnames(gevmodel$modX[[1]])
  signames = colnames(gevmodel$modX[[2]])
  xinames  = colnames(gevmodel$modX[[3]])

  xinames = xinames[!is.na(xinames)]

  mucovariates  = data[,munames]
  sigcovariates = data[,signames]
  xicovariates  = data[,xinames]

  # Get link function to transform
  fmu = fsig = fxi = identity
  fmu = getLinkFunction(gevmodel$mulink)
  fsig = getLinkFunction(gevmodel$siglink)
  fxi = getLinkFunction(gevmodel$xilink)

  musum = fmu(as.matrix(mu), as.matrix(mucovariates))
  sigsum = fsig(as.matrix(sig), as.matrix(sigcovariates))
  xisum = fxi(as.matrix(xi), as.matrix(xicovariates))

  return(list("musum"=musum,"sigsum"=sigsum,"xisum"=xisum))
}
```

The next set of functions are used for fitting the GEV distribution. The `X` at the end of most of these functions refers to the final iteration that the functions went through, in this case it means that model matrices were used to fit. These functions include calculating gradients, log-likelihoods, optimising the log-likelihood and a wrapper that combines this all together with easy inputs.

```
GEVlikX_diffx = function(theta, z, muX, sigX, xiX,
                         mulink="identity", siglink="exponential",
                         xilink="identity"){
  # compute likelihood for GEV distribtuion with link functions for parameters
  # allows different choices of covariates in different parameters

  # X is the model matrix (supplied in input for each parameter)
  # z is the response
  mup = dim(muX)[2]
  sigp = dim(sigX)[2]
  xip = dim(xiX)[2]
  if(mulink=="none") mup=1
  if(siglink=="none") sigp=1
  if(xilink=="none") xip=1

  # different indices for different parameters
  muindex  <- getBetaIndex(mulink,siglink,xilink,mup)
  muindex = muindex[muindex==1]
  sigindex <- getBetaIndex(mulink,siglink,xilink,sigp)
  sigindex = sigindex[sigindex==2]
  xiindex  <- getBetaIndex(mulink,siglink,xilink,xip)
  xiindex = xiindex[xiindex==3]
  bindex = c(muindex,sigindex,xiindex)

  # Set parameters based on beta-index
  mubeta  <- theta[bindex==1]
  sigbeta <- theta[bindex==2]
  xibeta  <- theta[bindex==3]

  # Set up link functions
  mulinkF = getLinkFunction(mulink)
  siglinkF = getLinkFunction(siglink)
  xilinkF = getLinkFunction(xilink)

  # Create design matrices based on link functions
  mumat <- mulinkF(mubeta, muX)
  sigmat <- siglinkF(sigbeta, sigX)
  ximat <- xilinkF(xibeta, xiX)
```

```
  # Set up a recurring expression in advance
  z0 <- 1 + ximat*((z-mumat)/sigmat)

  # Catch errors in parameters
  if(any(z0<0)) return(-1e20)
  if(any(sigmat<0)) return(-1e20)
  if(any(ximat<0)) return(-1e20)

  loglike = -sum(log(sigmat) + (1 + 1/ximat)*log(z0) + (z0^(-1/ximat) ) )

  return(loglike)
}


GEVgradX_diffx <- function(theta, muX, sigX, xiX, z,
                           mulink="identity", siglink="exponential",
                           xilink="identity"){
  # Function to calculate gradient of parameters, accounting for
  # non-stationarity. Takes same arguments as GEVlikX_diffx so it
  # can go into the optim function

  # Set up parameters
  mup = dim(muX)[2]
  sigp = dim(sigX)[2]
  xip = dim(xiX)[2]

  muindex  <- getBetaIndex(mulink,siglink,xilink,mup)
  muindex = muindex[muindex==1]
  sigindex <- getBetaIndex(mulink,siglink,xilink,sigp)
  sigindex = sigindex[sigindex==2]
  xiindex  <- getBetaIndex(mulink,siglink,xilink,xip)
  xiindex = xiindex[xiindex==3]
  bindex = c(muindex,sigindex,xiindex)

  # Set parameters
  mubeta  <- theta[bindex==1]
  sigbeta <- theta[bindex==2]
  xibeta  <- theta[bindex==3]

  # Set up link function
  mulinkF = getLinkFunction(mulink)
  siglinkF = getLinkFunction(siglink)
  xilinkF = getLinkFunction(xilink)

  # get matrices for parameters with link functions
```

```
mumat <- as.vector(mulinkF(mubeta, muX))
sigmat <- as.vector(siglinkF(sigbeta, sigX))
ximat <- as.vector(xilinkF(xibeta, xiX))
gradients = list("mu"=mumat, "sig"=sigmat, "xi"=ximat)

# Criteria for shape parameter to be zero (or close enough)
xizeros = round(ximat,6)==0

# hard code if siglink is exponential (which is likely)
# and separating likelihood functions for when shape==0
# these functions are similar to GEVingrad, for different situations
if(siglink!="exponential") {
  gradientsno0s <- GEVingrad(mumat[!xizeros], sigmat[!xizeros],
                             ximat[!xizeros], z[!xizeros])
} else if(siglink=="exponential") {
  gradientsno0s <- GEVingradLog(mumat[!xizeros], log(sigmat[!xizeros]),
                                ximat[!xizeros], z[!xizeros])
}

if(siglink!="exponential") {
  gradients0s <- GEVingradxi0(mumat[xizeros], sigmat[xizeros],
                              ximat[xizeros], z[xizeros])
} else if(siglink=="exponential") {
  gradients0s <- GEVingradLogxi0(mumat[xizeros], log(sigmat[xizeros]),
                                 ximat[xizeros], z[xizeros])
}

# Separate gradients for different parameters
gradients$mu[xizeros] = gradients0s$mu
gradients$mu[!xizeros] = gradientsno0s$mu
gradients$sig[xizeros] = gradients0s$sig
gradients$sig[!xizeros] = gradientsno0s$sig
gradients$xi[xizeros] = gradients0s$xi
gradients$xi[!xizeros] = gradientsno0s$xi

# Multiply by model matrices
gradients$mu <- muX * gradients$mu
gradients$sig <- sigX * gradients$sig
gradients$xi <- xiX * gradients$xi

# Sum relevant columns
gradients$mu = colSums(as.matrix(gradients$mu[,1:mup]))
gradients$sig = colSums(as.matrix(gradients$sig[,1:sigp]))
gradients$xi = colSums(as.matrix(gradients$xi[,1:xip]))
```

```
  # Return gradients as a vector
  return(unlist(gradients))
}

GEVfitX_diffx = function(z,muX,sigX,xiX, gr = TRUE, se = TRUE,
                         mulink="identity", siglink="exponential",
                         xilink="identity", init=NULL){
  # Fit the GEV distribution with allowances of different model matrices
  # these are specified in advance
  # fit.GEV will create model matrices based on links and the data frame,
  # and is a wrapper for this function

  # Get initial conditions
  mup = dim(muX)[2]; sigp = dim(sigX)[2]; xip = dim(xiX)[2]
  if(is.null(init)) init = getInit(z,mup,sigp,xip)
  bindex = c(rep(1,mup),rep(2,sigp),rep(3,xip))

  # Set up gradient function (optional)
  if(gr)  gr1 = GEVgradX_diffx
  if(!gr) gr1 = NULL

  # Use optim with fnscale=-1 to maximise the likelihood
  maxl = optim(par = init, fn = GEVlikX_diffx, gr=gr1, hessian = TRUE,
               method = "BFGS", control=list(fnscale=-1,maxit=1000),
               z=z, muX=muX, sigX = sigX, xiX=xiX, mulink=mulink,
               siglink=siglink,xilink=xilink)
  val = maxl$value

  if(!se) se = NULL
  if(se){
    cov = solve(maxl$hessian)
    se = sqrt(diag(abs(cov)))
  }

  # Output normal optim output but also details of the model so that when
  # the model is input to other functions, other information can be used
  return(list("value"=maxl$value, "par"=maxl$par, "se"=se, "bindex"=bindex,
              "mulink"=mulink, "siglink"=siglink, "xilink"=xilink,
              "modX" = list(muX,sigX,xiX)))

}



fit.GEV = function(formulas, data, response, rl.n,
```

```
                    links = list("identity","exponential","identity")){
  # A wrapper for what goes into muX, sigX, xiX etc.
  # Will output return values as well, just for tidying scripts
  # formulas is a list containing mu, sig, xi formulas
  # links is a list of link functions for mu, sig, xi

  muX = model.matrix(formulas[[1]], data = data)
  sigX = model.matrix(formulas[[2]], data = data)
  xiX = model.matrix(formulas[[3]], data = data)
  mul = links[[1]]; sigl = links[[2]]; xil = links[[3]]

  GEVmod = GEVfitX_diffx(response, muX, sigX, xiX, gr = TRUE,
                         mulink = mul, siglink=sigl, xilink=xil)

  f1 = labels(terms(formulas[[1]]))
  f2 = labels(terms(formulas[[2]]))
  f3 = labels(terms(formulas[[3]]))

  fall = unique(c(f1,f2,f3))

  if(length(fall)>0)  big.formula = reformulate(fall)
  if(length(fall)==0) big.formula = ~1

  unique.data = model.matrix(big.formula, data = data)
  unique.data = unique(unique.data)

  rl = GEV_rl(rl.n, GEVmod, unique.data)
  rl = fix.colnames(rl)

  return(list("GEVmod"=GEVmod, "df.returns" = rl))
}
```

These last few functions are used to calculate return levels. `GEV_rl` calculates return levels given a GEV model output from the fit function, and a data frame that contains all forms of covariates included into the model. `GEV_rl_individual` will calculate return levels given a stationary GEV, fit for each station.

```
GEV_rl = function(n, gevmodel, data){
  # Calculate return levels, given a GEV model (output from fit.GEV)
  # Not needed to call this function directly, as fit.GEV will output RLs

  parsums = GEV_getParSums(gevmodel,data)
  xizeros = round(parsums$xisum,6)==0
```

```
  # Hard code if shape parameter has 0 elements
  if(length(parsums$xisum)!=1){
    z_n = numeric(length(parsums$xisum))
    z_n[!xizeros] = parsums$musum - (parsums$sigsum/parsums$xisum)*
                                    (1-(-log(1-1/n))^(-parsums$xisum))
    z_n[xizeros] = parsums$musum - parsums$sigsum * log(-log(1-(1/n)))
  } else {
    z_n = rep(NA, length(parsums$musum))
    z_n[!xizeros] = parsums$musum - (parsums$sigsum/parsums$xisum)*
                                    (1-(-log(1-1/n))^(-parsums$xisum))
    z_n[xizeros] = parsums$musum - parsums$sigsum * log(-log(1-(1/n)))

  }
  retlevels = data.frame(data, return=z_n)
  return(retlevels)
}


GEV_rl_individual = function(n, sw.data, type = "Obs"){
  # Calculate stationary return levels, given data in format of sw.obs
  # n-year return level, can be used for obs or model (different formats)

  if(type == "Obs") {
    # Split maxima by station
    sw.data.sub <- split(sw.data, sw.data$SRC_ID)
    list_fun = function(x){
      gevmod = fit.GEV(list( ~1, ~1, ~1), x, x$PRCP_AMT, 100,
                       links = list("identity", "identity","identity"))
      cbind(x[1,c("LON","LAT")],gevmod$df.returns[,2])
    }
  }
  if(type == "Model") {
    # Split by unique id
    sw.data.sub = split(sw.data, sw.data$id)
    list_fun = function(x){
      gevmod = fit.GEV(list( ~1, ~1, ~1), x, x$rain, 100,
                       links = list("identity", "identity","identity"))
      cbind(x[1,c("lon","lat")],gevmod$df.returns[,2])
    }
  }

  # Apply list_fun to the split data,
  # returning a return value at each unique lon/lat
  return = lapply(sw.data.sub, list_fun)
```

```
  return = do.call(rbind, return)
  colnames(return)[ncol(return)] = "return"
  return(return)
}
```