Statistics of Extremes

Dáire O'Kane, Gregoris Georgiou, Antonis Hadjiosif

March 2022

Contents

Abstract

Arguably the most pertinent use of contemporary statistical applications is the prevention of environmental disasters. In this report we will be employing extreme value techniques to consider changes in distribution of the spatial predicted 100-year annual monthly maximum burnt area return levels, i.e. the annual monthly maximum burnt area (BAx) value that is expected to be exceeded once every 100 years, as a result of wildfire occurrence in the United States, across the years of 2000, 2005, 2010 and 2015. We will be working with a large gridded dataset based on cell longitude and latitude containing BAx and values for other auxiliary variables related to land cover, weather, and altitude for the USA across the years of 1993 to 2015. Several generalised extreme value (GEV) models are considered, with options to allow for the variation of model parameters over space and the inclusion of additional variables to assess their potential impact on BAx prediction. We will be quantifying how the 100-year return level changes across the four years, as well as comparing the variation and uncertainty across the models employed. Our analysis indicates the 100-year return level range decreases from 2000 to 2015, with maximum values reduced and minimum values increased. Averaging over the spatial domain of the USA, we see from our analysis that the 100-year BAx return level corresponds to a 30-year return level in 2015.

1 Introduction

In a rapidly changing world where the earth increasingly suffers from the consequences of global warming, more and more extreme events are observed, including increasing heat, changing rain and snow patterns, shifts in plant communities, and other climate-related changes (Borunda, 2020). All of which are directly related to changes in the likelihood of wildfires in different regions across the world. Anthropogenic factors such as climate change have resulted in extremely hot and dry conditions (Harrison, 2021), making the occurrence of wildfires more probable, including wildfires resulting in extreme burnt area and devastation. This report will focus on the region of the continental USA where many of its regions were greatly affected by thousands of wildfires in the summer of 2021. A state which was massively affected was the state of California resulting from a historic low on the region's rainfall and reservoir levels (Gammon, 2021). More specifically, in 2021 alone the USA experienced 58,733 wildfires, with approximately 7% occurring in California alone, resulting in a total destroyed area of 7,139,713 acres (NCDC, 2021). A recent paper by (Virginia Iglesias, 2022) studies fire frequency across the USA and reports that all regions are affected by more fires since 2000, with the most extreme wildfires concurrent and widespread. Mitigating the devastation of a wildfire is of paramount importance to habitats of many wildlife, the homes of settlers and the overall preservation of the earths environment, due to the increased carbon monoxide that results from wildfire devastation, further contributing to the greenhouse effect and global warming. Several methods in supressing and detecting wildfires are included in (Alkhatib, 2014), involving controlled burning of dry areas under the supervision of firefighters, lightning detectors to detect strike coordinates and general physical factors such as an increased deployment of fire emergency services as well as educating and encouraging awareness around wildfire reporting protocol.

Figure 1: Firefighters battling with a wildfire in California, USA. (NBC, 2020)

The aim of this report is to conduct an extreme value analysis of the annual maximum monthly burnt area, which we denote as BAx, of wildfires occurring across the USA, with the intention of making statements and inference about the distribution of BAx 100-year return levels, which we denote as 100RLx, to help predict and mitigate the impact of future wildfires. We are working with a large, gridded dataset consisting of approximately 3500 unique locations USA, specified by longitude and latitude coordinates, with observations for 44 variables, ranging from 1993 to 2015. Since we are in the setting of modelling extreme wildfires, we appeal to extreme value theory, and in particular, consider a generalised extreme value model for the variable BAx. Additionally, as it is likely many land, environmental and climactic factors have an influence on the burnt area of a wildfire, we employ techniques of generalised additive modelling, in conjunction with extreme value theory to consider a model permitting for the non-stationary estimation of parameters i.e. allowing the model parameters to vary as a function of covariates. We are interested in studying the changes and distribution of 100-year return levels, in particular, how they vary across our spatial domain, as well as how they are changing across a range of years, as they are an informative way of encapsulating the model parameter estimates to explain the BAx value likely to be exceeded once every 100 years. This is pertinent to help predict and mitigate for the devastation caused by wildfires.

Several other modelling methods entered our discussion, such as a generalised pareto model, to which we consider a threshold and deem any wildfires resulting in burnt area value greater than this as extreme events, as discussed in (Coles, 2001). We opted for the GEV model to avoid any issues around setting a threshold and or the problem of choosing a separate threshold for individual regions of the USA. Another potential model in consideration was a Gaussian Markov Field Model (GRMF), due to the discretised nature of our data. This method takes account of the wildfire burnt area spatial dependency between neighbouring cells and is a potential inclusion, given the option to undertake further development of our analysis.

The paper is structure as follows. Section 2 introduces the statistical theory relevant to extreme value theory and generalised additive modelling, as well as the modelling techniques used in R. Section 3 applies our techniques and theory in practice under the subject of a case study involving modelling BAx across the USA, with an analysis of our results.

2 Statistical Theory

Extreme value theory will be the focal point of this report using the theory covered from the two books, *An Introduction to Statistical Modelling of Extremes* (Coles, 2001) and *Generalised Additive Models* (Wood, 2006). For the purposes of extreme value analysis in R, we will be mostly using the packages evgam (Youngman, 2003) and Optim (R-Documentation, 2022). The theory behind the techniques used to carry out extreme value analysis is provided below.

2.1 Introduction to extreme value theory

Let $X_1, ..., X_n$ be a sequence of independent random variables which have a common distribution function F. Then we define the maximum measurement of a process over a time period with n observations as

$$M_n = \max\{X_1, ..., X_n\}.$$

For convenience, in applications the X_i are values of a measurement which takes place on a regular time-scale, for example hourly, daily, monthly, yearly etc. This is done in order for M_n to represent the maximum over n time units, for example if n is the number of measurements or observations in a day, then M_n is the daily maximum.
Even though we can derive the distribution of M_n for all values of n, it is not convenient to use. There is a simple reason why. We know that the probability of M_n being less than z is

$$\begin{aligned} P\{M_n \leq z\} &= P\{X_1 \leq z, \ldots, X_n \leq z\} \\ &= P\{X_1 \leq z\} \times \ldots \times P\{X_n \leq z\} \\ &= \{F(z)\}^n. \end{aligned}$$

Even if we use the observed data and find an estimate of F, since it is raised in the power of n, a small error or difference in the value of F leads to great difference of F^n. In order to overcome this, we can use something similar as the central limit theorem. Instead of approximating distributions of sample means, using our extreme data only, we will find families of models for F^n. Thus we come up to the extremely important theorem, the extremal types theorem, which states that if there exist a sequence of normalising constants $\{a_n > 0\}$ and $\{b_n\}$, then the limiting distribution of $(M_n - b_n)/a_n$, as $n \to \infty$, has a non-degenerate limit distribution $G(z)$, which is one of three following extreme value distributions.

$$A : G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, -\infty < z < \infty$$

$$B : G(z) = \begin{cases} 0, & z \leq b, \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b; \end{cases}$$

$$C : G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, & z < b, \\ 1, & z \geq b; \end{cases}$$

the three extreme values distributions A, B and C are known as the Gumbel, Fréchet and Weibull families respectively. In order to combine these three families into a single one, it was derived a distribution function named, The Generalized Extreme Value family of distributions (GEV), which has the following form.

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

defined on the set $\{z : \xi(z-\mu)/\sigma > 0\}$, and the parameters satisfy $\infty < \mu < \infty$, $\sigma > 0$ and $\infty < \xi < \infty$. Here μ is the location parameter, σ is the scale parameter, and ξ is the shape parameter. When $\xi = 0$, $\xi > 0$ or $\xi < 0$, it corresponds to the Gumbel, Fréchet or Weibull family respectively.
Now, let's say we want to find a way to find the value that is expected to be equaled or exceeded on average once every a specific time period. We can do this by calculating the return level z_p. z_p is the return level associated with the return period $1/p$, and if we let $y_p = -log(1-p)$, is defined as:

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

where $G(z_p) = 1 - p$. For example, if we are looking into annual maxima, the return level z_p is expected to be equalled or exceeded by the annual maximum with probability p, in any particular year.
In order for us to use the formula for the return levels, we have to find estimates for the three parameters. Thus, it would be very useful to find the log-likelihood for the GEV parameters, and then use optimization methods to find the estimated parameters that maximize it. With the assumption that $Z_1, \ldots, Z_m$ have the GEV distribution and are independent, when $\xi \neq 0$, we have the following log-likelihood for the GEV parameters :

$$\ell(\mu, \sigma, \xi) = -m \log \sigma - (1 + 1/\xi) \sum_{i=1}^{m} \log\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi}$$

provided that

$$1 + \xi\left(\frac{z_i - \mu}{\sigma}\right) > 0, \text{for } i = 1, ..., m \tag{2}$$

If (2) is not satisfied, then the likelihood is zero and as a consequent the log-likelihood is ∞. In the case that $\xi = 0$, we have to handle it using the Gumbel limit of the GEV distribution, therefore leading to the log-likelihood for the GEV parameters:

$$\ell(\mu, \sigma) = -m \log \sigma - \sum_{i=1}^{m}\left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\}.$$

2.2 Generalised Regression Models

Generalised Linear Models:

A Generalised Linear Model (GLM) allows for response distributions other than normal, and for a degree of non-linearity in the model structure. It is generally assumed that the response variables, Z_i, are independent and follow an exponential family distribution. The basic structure of a GLM is

$$g(\mu_i) = \mathbf{X_i}\beta$$

where $\mu_i = \mathbf{E}(Z_i)$, g is a smooth, monotonic function, usually called a link, $\mathbf{X_i}$ is the ith row of a model matrix and β is a vector of unknown parameters (Wood, 2006).
Example:
Consider a model of the form $\mu_i = ab^{x_i}$,where a and b are unknown parameters. Using a log link function, we can transform this model into GLM form

$$\log(\mu_i) = log(a) + log(b)x_i = \beta_0 + \beta_1 x_i$$

with $\beta_0 = \log(a)$, $\beta_1 = \log(b)$, the model is now linear in its parameters. A suitable distribution for the response could be a gamma distribution, given that the model could represent exponential growth of a population.

Generalised Additive Models:

A Generalised Additive Model (GAM), is a GLM with linear predictor involving a sum of smooth functions of covariates. It has the general structure:

$$g(\mu_i) = \mathbf{X_i^*}\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots$$

where $\mu_i = \mathbf{E}(Z_i)$, g, Z_i are the same as for a GLM and $\mathbf{X_i^*}$ is the ith row of a model matrix for any strictly parametric model components. θ is the corresponding parameter vector and f_j are smooth functions of the covariates, x_k. We are interested in making inference about the smooth functions, f_j. To outline how GAM works, we will consider a simple model consisting of one univariate smooth function:

$$z_i = f(x_i) + \epsilon_i \tag{3}$$

for $\epsilon_i \sim N(0, \sigma^2)$. Our goal is to estimate f using linear methods, so we wish to represent f in such a way that (3) becomes a linear model. This can be achieved by representing f in terms of the basis functions from the space of which f belongs. We get that

$$f(x) = \sum_{i=1}^{q} b_j(x)\beta_j$$

for basis functions, $b_j(x)$ and unknown parameters, β_j. Subbing into (3), we get the following linear model:

$$z_i = b_1(x_i)\beta_1 + \dots + b_q(x_q)\beta_q + \epsilon_i.$$

Certain examples of bases for f include splines. A spline is a function defined piecewise by polynomials, satisfying certain conditions on continuity and differentiability. The locations at which the polynomials are joined are known as knots. Spline functions can be used in additive modelling as smoothing functions. A cubic spline would be a suitable base for the univariate situation of (3), (Wood, 2006).

2.3 Penalised methods

Penalised least squares approach:

Estimating an additive model involves estimating the unknown basis parameters, $\boldsymbol{\beta}$. An impact on the estimation of our model is model smoothness, which is controlled by the basis dimension, q, where q is the number of knots plus the number of unknown parameters. Methods of controlling smoothness involve keeping the basis dimension fixed at a degree assumed to be a little larger than necessary and adding a "wiggleness" penalty to the least squares fitting objective. Subsequently, the penalised regression spline problem of an additive model is then to minimise

$$\|\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta}$$

w.r.t. β, where $\mathbf{S}$ is a matrix of known coefficients and λ is the model smoothing parameter. It can be shown that the penalised least square estimator of $\boldsymbol{\beta}$, given λ, is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T X} + \lambda \mathbf{S})^{-1} \mathbf{X^T z}.$$

For a generalised additive model, the basis parameters, $\boldsymbol{\beta}$, must be estimated by a penalised maximum likelihood approach, which in practice, is achieved by penalised iterative least squares (Wood, 2006).

Penalised Log Likelihood:

Estimating model parameter values μ, σ and ξ corresponds to estimating basis coefficients, $\boldsymbol{\beta}$. Since this is a generalised additive model, this involves maximising a penalised log likelihood of the form:

$$\ell(\boldsymbol{\beta}_{\boldsymbol{\lambda}}, \lambda) = \ell(\boldsymbol{\beta}_{\boldsymbol{\lambda}}) - \frac{1}{2}\boldsymbol{\beta}^T S_\lambda \beta,$$

where $\lambda = (\lambda_1, \ldots, \lambda_k)$ is the model smoothing parameter and S_λ is a penalty matrix, with elements correpsonding to the choices for b_{kd} basis functions (Youngman, 2003).

2.4 Extreme Value Analysis in R

evgam package:

In this report, we will be using the package evgam in the statistical language R, to carry out generalised additive modelling techniques applied to extreme value theory. In our case, this pertains to the modelling of annual monthly maximum wildfire burnt area across the USA. The theory behind Extreme Value Generalised Additive Models (EVGAMs) follows directly from the theory of GAMs and extreme value theory. The package evgam allows for fitting of extreme value distributions with parameters varying flexibly in covariate, x. The function assumes GAM forms in parameters, which like in GAMs, rely on basis representations.
If we consider the covariate, x, and GEV parameters varying with x, $\mu(x), \sigma(x), \xi(x)$, evgam relates parameters to fixed link functions, via η^*, which has a basis representation. The link functions are as follows:

$$\mu(x) = \eta_\mu(x), \log(\sigma(x)) = \eta_\sigma(x), \gamma(x) = \eta_\gamma(x),$$

where

$$\eta_*(x) = \beta_0 + \sum_{k=1}^{K} \sum_{d=1}^{D_k} \beta_{kd} b_{kd}(x)$$

and β_{kd} and b_{kd} are the basis coefficients and functions, respectively. As this is an additive model, $\eta_*(x) = \mathbf{X^T}\boldsymbol{\beta}$, and since the support of σ is defined on $\mathbb{R}^+$, evgam uses a log link function for this parameter (Youngman, 2003).

Optim package:

Another part of our analysis on wildfire annual monthly maximum burnt area involves using the function Optim in R. Optim is used in conjunction with our defined GEV log likelihood function, which is an input argument as the function to be maximised in Optim. When using Optim, we also provide an inital parameter vector, as well as the optimisation method. In our analysis, we used the Nelder-Mead method. Optim will be used for the side-by-side extreme value analysis on the data with evgam, comparing their return levels for the annual maximum burnt area. Important to note that in our usage of Optim, we are not permitting for nonstationary fittings in model parameters, i.e. we are not considering the GEV parameters to be a function of covariates, like in the case of using evgam.

Assessing model fit and the AIC:

We want to use a measure of goodness of fit on the evgam model to remove any insignificant variables which are insignificant to the burnt area. This will help in reducing the data size and as a result the function will run more efficiently and the final prediction of the return levels will be more accurate. The measure of goodness of fit that will be used is the 'Akaike information

criterion', a measure of information loss of the model (Wikipedia, 2021). Hence, we will proceed with the model that has the lower value of AIC:

$$AIC = 2[-\ell(\hat{\theta}) + p],$$

the p term in the AIC penalises models with more parameters than necessary, counteracting the tendency of the likelihood to favour even larger models (Wood, 2006).

3 Case study: modelling wildfire burnt area across the USA

Wildfires are a great reason why nature and wildlife are destroyed but also the cause of people losing their lives. We will explore the recorded data of wildfires across the USA, and use them to find out what factors may affect the burnt area throughout the years. Using this information, we will carry out an extreme value analysis, to estimate the annual monthly maximum burnt area from a wildfire for future return periods and also approximate the location that this fire may be located in the future.

3.1 USA wildfires data

In this Mathematics project we use the data from the Data challenge "A competition in the field of spatio-temporal regression modelling of extremes.", organised by the University of Edinburgh through the event 'EVA 2021' (UoE, 2021).
The dataset contains some variables which are thought to be related to wildfires and will be utilized for the analysis of extremes. We should note here that some of these variables will be removed as a result of an initial refinement of the dataset. Additionally, some variables will turn out to be insignificant in the extreme value analysis of the burnt area , using a measure of goodness of fit and thus will be dropped from the dataset.
This dataset covers data relating to wildfires in the Continental USA, excluding the state of Alaska and islands of the USA territories (Hawaii, Virgin Islands, etc.). Moreover, this dataset spans from 1993 to 2015 and it only contains data from March to September of each year. While it is thought that the peak season of wildfires is in the summer months (Jack Beckwith and Wolf, 2018), this will still create some bias in the analysis of the data.

3.1.1 Initial refinement of data

Our initial dataset contains 37 variables and 563,983 observations which are significantly a lot more than what the package evgam can handle to predict coefficients. Therefore, it was decided that the extreme analysis is performed on the categories: meteorological variables & other variables. Consequently, the category related to land cover is dropped from the dataframe. Furthermore, we introduce some new variables that are calculated using some of our existing variables. We introduce Relative Humidity (RH), Wind Speed (WS) and Ground Net Thermal Radiation(gntr). Through prior knowledge the following couples of variables are combined into a single one:

- $clim_3$ & $clim_4$ to Relative Humidity
- $clim_1$ & $clim_2$ to Wind Speed
- $clim_6$ & $clim_7$ to Ground Thermal Net Radiation

Using additional information from the competition (UoE, 2021) the formulae for the above variables are calculated by:

- $RH = \frac{\exp(17.502*(clim_3-273.16)/(clim_3-32.19))}{\exp(17.502*(clim_4-273.16)/(clim_4-32.19))}$
- $WS = \sqrt{({clim_1}^2 + {clim_2}^2)}$
- $gntr = clim_6 - clim_7$

Variable Category	Variable	Description
Land cover (Proportion of each category within grid cells of longitude and latitude: $0.5^o \times 0.5^o$)	lc_1	cropland rainfed
	lc_2	cropland rainfed herbaceous cover
	lc_3	mosaic cropland
	lc_4	mosaic natural vegetation
	lc_5	tree broadleaved evergreen closed to open
	lc_6	tree broadleaved deciduous closed to open
	lc_7	tree needleleaved evergreen closed to open
	lc_8	tree needleleaved deciduous closed to open
	lc_9	tree mixed
	lc_{10}	mosaic tree and shrub
	lc_{11}	shrubland
	lc_{12}	grassland
	lc_{13}	sparse vegetation
	lc_{14}	tree cover flooded fresh or brakish water
	lc_{15}	shrub or herbaceous cover flooded
	lc_{16}	urban
	lc_{17}	bare areas
	lc_{18}	water
Meteorological variables	$clim_1$	wind speed in Eastern direction, m/s
	$clim_2$	wind speed in Northern direction, m/s
	$clim_3$	Dewpoint temperature, K
	$clim_4$	Temperature, K
	$clim_5$	Potential evaporation,m
	$clim_6$	Surface net solar radiation, J/m^2
	$clim_7$	Surface net thermal radiation, J/m^2
	$clim_8$	Surface pressure, Pa
	$clim_9$	Evaporation, m
	$clim_{10}$	Precipitation, m
Other Variables	CNT	number of wildfires
	BA	burnt area of wildfires, ac
	lon	longitude
	lat	latitude
	$area$	the proportion of a grid cell that overlaps the continental USA
	$month$	From March to September
	$year$	From 1993 to 2015
	$altiMean$	Mean of altitude related variables
	$altiSD$	Standard deviation of altitude related variables

Table 1: Table of the full list of variables from the Data challenge organized by the University of Edinburgh, EVA 2021

This reduces the dataset to 16 variables. Initially, we create a correlation matrix using R that indicates the pairwise correlation between these variables. We choose to exclude from our dataset variables with a correlation above the threshold of $r = |0.7|$, using the Pearson correlation coefficient, which indicates a high positive/negative correlation between the variables.

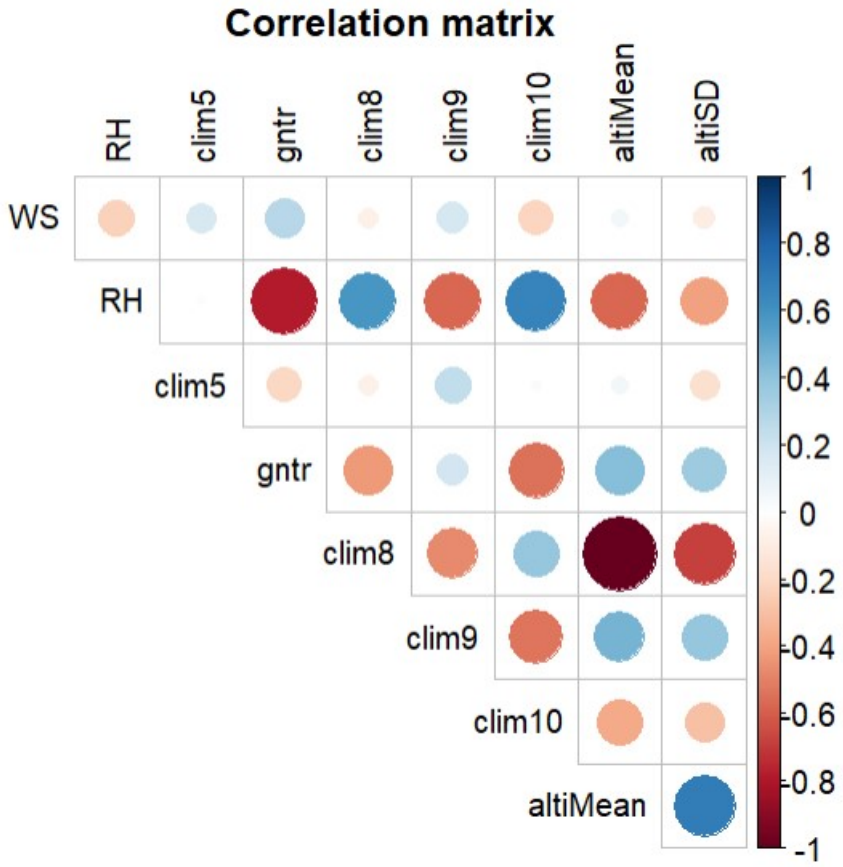

Figure 2: Correlation Matrix of the variables in our dataset using Pearson correlation coefficient

The following variables have high correlation between them:

- *gntr* (Net thermal radiation) & *RH* (Relative Humidity) have a negative correlation with a coefficient, r= -0.7877645. We exclude *gntr* as relative humidity *RH* is known to be highly associated with the likelihood of wildfires *BA*.
- *altiMean* (altitude) & *clim*8 (surface pressure) have a negative correlation with a coefficient, r= -0.9965072. Thus, we get rid of variables *altiMean* & *altiSD* since surface pressure can have similar effects on *BA*

Thus, after the initial refinement of the data this results in a dataframe of 13 variables which will be further reduced with a goodness of fit test before moving on to the extreme analysis.
As our aim is to conduct an extreme value analysis on the annual monthly maximum burnt area (BAx), our next task is to compute the monthly maximum burnt area for each unique longitude and latitude, for each different year. Additionally, since for certain years at specific locations there were no wildfires, we will have some entries of maximum burnt area of zero, so we will filter them out.

In addition, since the dataset is quite large, to simplify procedures and to allow for more in depth analysis of results, we split the data into 16 blocks, based on cell longitude and latitude. Calculating and using the quantiles of all latitudes, we draw lines on the map of the United States to divide it in 16 blocks, with block 1 being the lower left block and block 16 the top right block, as illustrated in Figure 3.

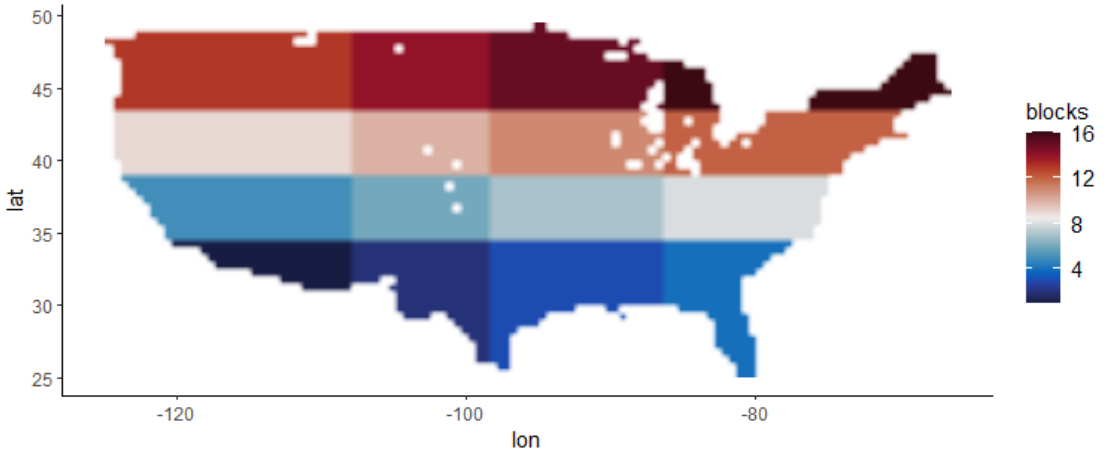

Figure 3: Map of the United states partitioned into 16 blocks

3.2 Statistical Models

The main statisical model used in our analysis is an extreme value generalised additive model, as described in Section 2, which permits for the estimation of GEV model parameters in a non-stationary manner.

3.2.1 Model AIC analysis

As previously mentioned there is only a certain amount of data that evgam can handle. As a result, for the goodness of fit test the model is sampled on the first four blocks cite map.
Starting from the variables that are already known to have an effect on burnt area, we are going to manually add and remove the rest of the variables from the model to see how AIC is affected and reach the final model. The initial variables are:

- longitude
- latitude
- months
- number of wildfires
- Relative Humidity

Note that 'longitude' & 'latitude' are modelled using a spline model and a separate spline model will be used for 'months'. The results can be seen on the table of Figure 4.

Variables already included in Model	*New variables Test*	*AIC*
✓ Longitude	WS	177913.5
✓ Latitude	Clim5	**177804.7**
✓ CNT	Clim8	177823.7
✓ Month	Clim9	178632.7
✓ RH	Clim10	178409.3
	None	177917.8
✓ Clim5	WS	177811.0
	Clim8	**177726.5**
	Clim9	178888.1
	Clim10	178542.1
✓ Clim8	WS	**177720.0**
	Clim9	177728.1
	Clim10	178580.5
✓ WS	Clim9	177933.5
	Clim10	178084.1

Figure 4: AIC values for different models used

As we can see the lowest AIC occurs when we include in our model the following variables:

- Potential Evaporation
- Surface Pressure
- Wind Speed

This results in the following final model, where GEV parameters for BAx are estimated as follows:

$$\mu_{i,j,t} = 1 + f_{11}(lon_i, lat_j) + f_{21}(month_t) + \boldsymbol{\beta_1 x},$$
$$\log \sigma_{i,j,t} = 1 + f_{12}(lon_i, lat_j) + f_{22}(month_t) + \boldsymbol{\beta_2 x},$$
$$\xi_{i,j,t} = 1 + \boldsymbol{\beta_3 x},$$

where i, j stand for the grid's location in longitude and latitude respectively, t is the time in months, f_{ij} are smooth functions which we wish to estimate and $\boldsymbol{\beta_k}$ is a column vector of the regression coefficients for $\boldsymbol{x} = [CNT, RH, clim5, clim8, WS]$, a row vector which includes the variables significant to the burnt area. As we can see from above, the final model involves assuming non-stationarity in the GEV model parameters, in the form of a generalised additive model for μ and σ, and a linear model for ξ. The GAM used for both parameters involves a mixture of smooth and parameteric components. Notice that *lon* and *lat* are covariates of the same smooth function as they are interacting and not additive. We also have a separate smooth function for the covariate, *month*. The parametric components selected are based on the results of the AIC model analysis carried out above.

We will use the final model described above, in conjunction with the package evgam in R to calculate parameter estimates for μ, σ and ξ, for each unique longitude and latitude cell location. Due to computational limitations, we will not be able to apply our model at all unique locations for all available years and have decided to perform analysis using the annual monthly maximum burnt area (BAx) for the years 2000, 2005, 2010 and 2015. We will then use the parameter estimates to calculate 100-year return levels at each cell, for the given year, which as defined generally here (1), are the BAx values likely to be exceeded once every 100 years. Furthermore, thanks to the 'predict()' function of the evgam package, we are able to extract the standard error for each cell's return level, allowing us to construct 95% confidence intervals for the return levels at each location in our spatial domain, for the given year.
Another approach in conducting our extreme value analysis on the variable BAx is a model using the log likelihood and return levels functions we have created in R, outlined at the end of this report (4.3), based on the extreme value theory in Section (2.1). Unlike the previous method, where it was not computationally feasible to model for all years, this method involves grouping the data by their unique longitude and latitude and then applying the function Optim in conjunction with our log likelihood function to estimate the GEV parameters μ, σ, ξ for each unique location, for each year, from 1993 to 2015. Based on our parameter estimates, we obtain 100-year return levels for the annual monthly maximum burnt area (100RLx) for each cell for each year and then take an average across all the years to obtain the 100RLx average for each cell. This method is slightly different as we are assuming stationarity in GEV parameters and are therefore not assuming they are functions of covariates.

3.3 Risk ratios and contour plots

To further quantify changes in the 100RLx across our spatial domain, we will consider the method of risk ratios (Graeme Auld, 2021). Let $Z_{it} \sim GEV(\mu_{it}, \sigma_{it}, \xi_{it})$, i.e. Z_{it} is a GEV random variable with parameters based on the obtained estimates at location i for year t. If we let $i = 23$ denote the year 2015 and $k =$ 8,13,18 denote the years 2000, 2005 and 2010 respectively, the risk ratio at location i is the value corresponding to how likely the 100-year return level (1), in the year k environment is to be exceeded in the 2015 environment.

$$\frac{\mathbb{P}\{Z_{i23} > z_{ik}(0.01)\}}{\mathbb{P}\{Z_{ik} > z_{ik}(0.01)\}} = \frac{\mathbb{P}\{Z_{i23} > z_{ik}(0.01)\}}{0.01}. \tag{4}$$

Since we are considering the non stationary fitting of the GEV location, and scale parameters, we can obtain contour plots from evgam to view the estimated smooth term for lon and lat, for each of these parameters, for the years 2000, 2005, 2010 and 2015. These plots are helpful to illustrate how our estimation changes over the spatial domain, for each of the four years.

3.4 Results

Considering a holistic view of Figure 5, it is clear that across the four years, the mid to south-western region of the USA has the largest 100-year BAx return levels (100RLx), with the eastern regions of the USA, specifically northeast, having the smallest 100RLx. This is indicated in Figure 5 by the large accumulation of orange to yellow points in the west and black to purple points in the east region. It is important to note that many cells appear white in the figure. This is a result of either no wildfire occurring at this specific location, or due to large lakes, reservoirs and rivers, making it impossible for a wildfire to occur there. This trend is most apparent across the years of 2000 and 2010, which we can see have the largest extremum range of 100RLx, due to the lightest yellow and darkest blacks appearing across the spatial domain. For the year 2000 specifically, the largest wildfire expected to occur in the next 100 years will result in the destruction of approximately 2.6×10^{10} acres of land, which is about 1 percent of the USA's total surface area. Similarly, a maximum value of around 1×10^{10} acres is predicted for the year 2010. Contrasting this with the years 2005 and 2015, there appears to be an overall reduced disparity in the extremum of the 100RLx across the majority of our spatial domain, with the 100RLx maximum generally decreasing but the 100RLx minimum in the eastern region tending to be larger, than in previous years.

Thus, generally indicating that for these years, certain regions in the east are forecasted to have their 100RLx exceeded at higher values than in previous years, or equivalently, that the 100RLx in the years 2000 and 2010 would be more likely exceeded than once every 100 years for certain regions in the east, for the years 2005 and 2015. This result is most significant for the year 2015, which appears to have a more evenly distributed 100RLx spread than of any of the previous years examined, with predictions for less extreme clusters of wildfires to occur, but predictions for harsher fires to happen in regions that were previously less affected by wildfires. This is seen by the more subdued purple and light red occurring across the spatial domain.

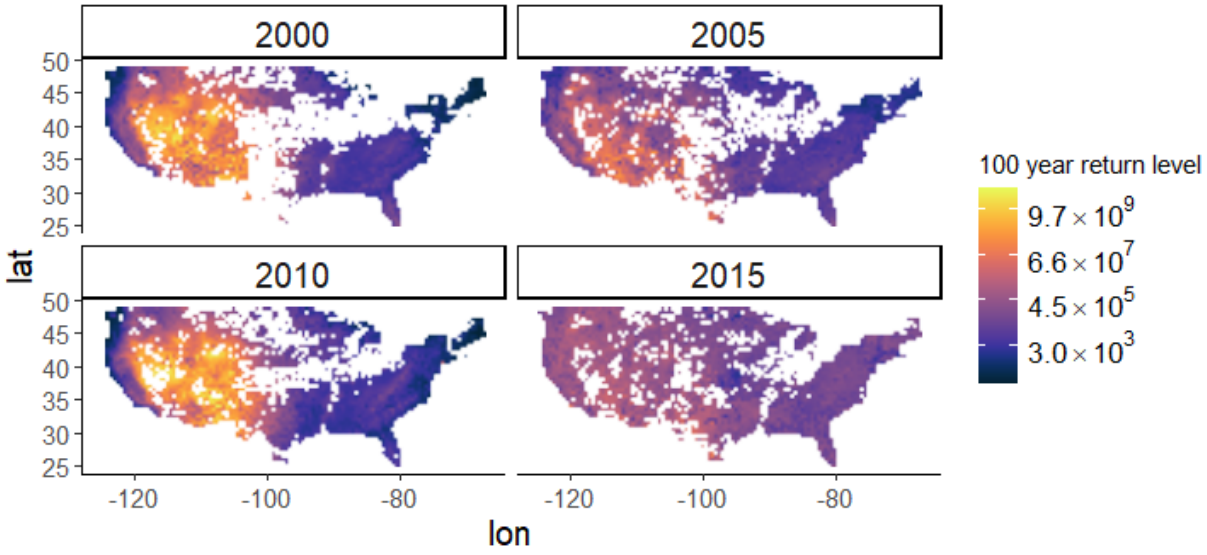

Figure 5: Spatial 100-year return levels (acres) for the years 2000,2005,2010 and 2015

We have quantified our uncertainty in the 100RLx estimates, across the USA, by computing the lower and upper bounds of a 95% confidence interval for each year, as displayed in Figure 6. We have obtained these estimates using the predict function in the package evgam, as described in Section 3.2. We must firstly offer a caveat to our approach, in that if the lower standard error estimates were calculated to be higher than the point estimates, the lower bound of the confidence interval for this cell would return a negative value. However, since the burnt area of a wildfire cannot be a negative number, the lower bound of the confidence interval was set to zero, causing the resulting output to appear grey. This is particularly pertinent in the years 2005 and 2010, with all of 2005 appearing grey and the majority of 2010, except for the region associated with intense 100RLx prediction. The impact of having high standard errors translates to a greater uncertainty in our results for the regions affected. Considering the results of Figure 6, and comparing with the results from Figure 5, the lower and upper interval estimates appear very similar to the 100RLx estimates as a result of the standard errors being relatively small for the 100RLx estimates, for the years 2000 and 2015, respectively. Additionally, it appears that the upper interval estimates are considerably larger for the years 2005 and 2010, which is seen by the lighter purple in previously darker regions and more intense orange colours observed in previously subdued regions from Figure 5. This is again a consequence of the high standard errors computed for spatial domain for these years and suggests the proneness of these regions to more extreme and unpredictable wildfires.

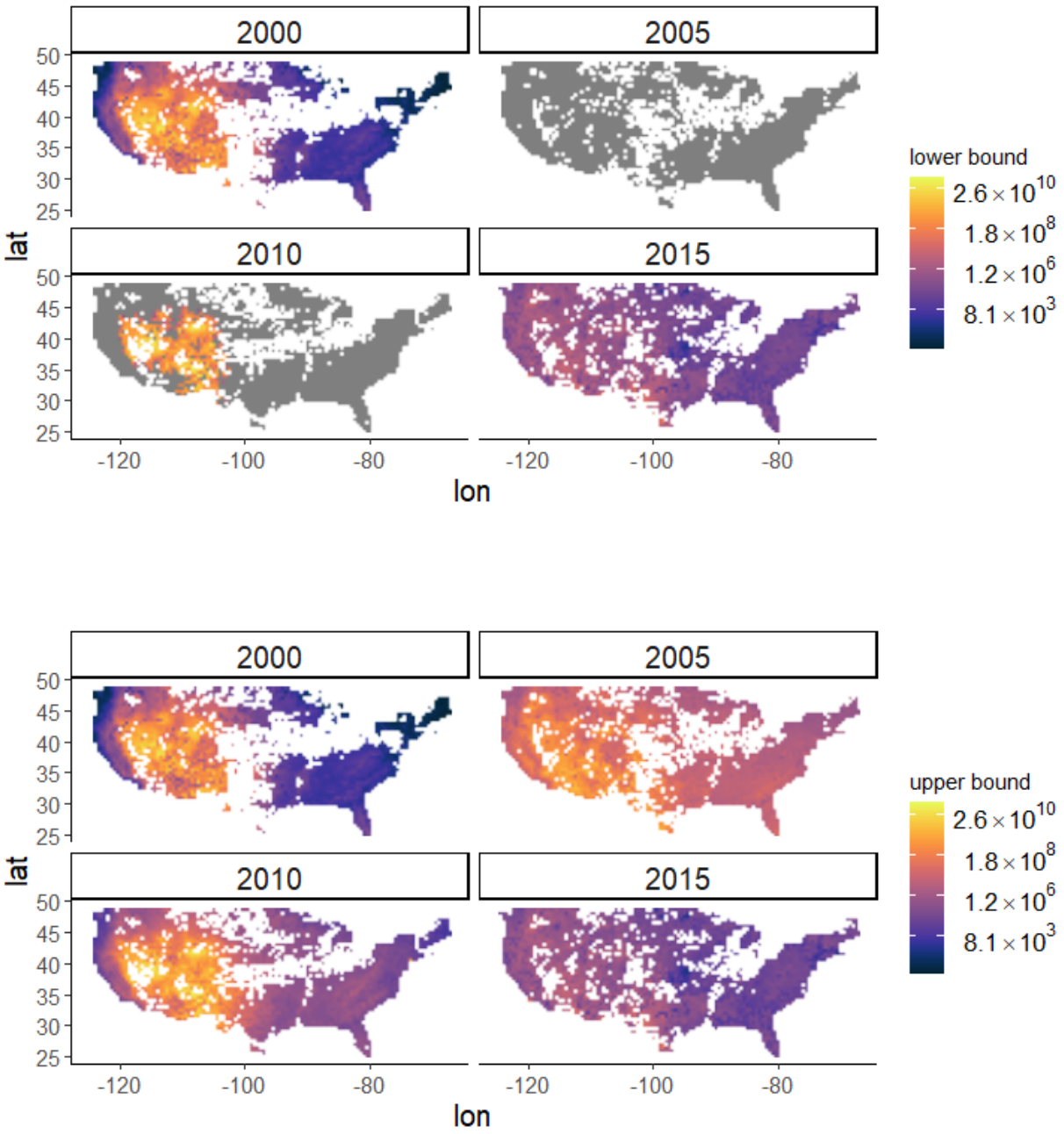

Figure 6: Lower and upper 100-year return level (acres) 0.95% confidence intervals for the years 2000, 2005, 2010 and 2015

The results from our spatial 100RLx average across the years 1993 to 2015 using the Optim model are displayed in Figure 7. Macroscopically, the general trend is similar to what was observed in Figure 5 and it is clear that the western region is experiencing the highest 100RLx values throughout. Interesting to note is that there is a protuberance of more extreme 100RLx values being observed in further easter regions than previously observed in Figure 5, and that based on the concentrated distribution of dark orange, we can see the 100RLx looks to be varying less than in the case of the evgam model, with there being a reduced disparity in the distribution of 100RLx across the spatial domain. We must however offer a caveat in comparing this with Figure 5, as we are not taking account for any covariates that could be in correlation with the burnt area, nor are we working over the same year range. We will carry out any further investigation of the 100RLx values using evgam but we thought it worthwhile to include the 100RLx values from the model based on the function that we created.

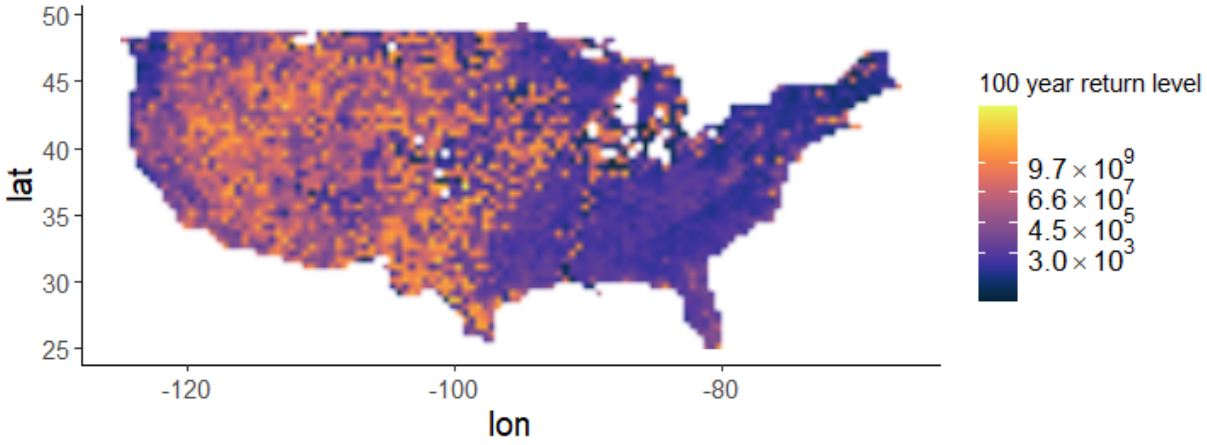

Figure 7: Spatial 100-year return levels (acres) using the functions built in R and Optim, without taking account of any covariates

Returning to the setting of evgam, we consider the question raised by the cell-by-cell 100RLx analysis of our spatial domain of how more likely the BAx values are to be exceeded in certain years, across certain regions of our spatial domain, than others. This motivated the investigation of risk ratios, as described in the Section 3.3, for the year 2015, with the years 2000, 2005 and 2010 respectively. This indicates how many times more likely, the 100RLx cell value based on the climate for each of the previous years are to be exceeded in the 2015 climate environment. In our domain, the calculation of risk ratios is based off the spatial locations that had a wildfire occurring in both 2015 and the year of comparison for that calculation, i.e., 2000, 2005 or 2010. Our results are illustrated in Figure 8, which displays for each cell in our spatial domain, the 100RLx risk ratios of 2015 with the three years. From observation of the graphs, it is clear that due to the lighter purple colour, the 100RLx risk ratios are highest in the north eastern region of the USA, with high values also occurring on the boundaries of the domain, such as north west and south east. It is also important to notice that the most extreme risk ratios are not occurring in the regions prone to extreme wildfires, such as that of the mid to south west, due to a more concentrated colour of black and dark purple observed here. This is indicative to the

100RLx varying less in these regions, consistent across the three years. We also observe a small number of extreme cells, which we can see due to the bright yellow colour, occurring especially in the north east and corresponding to very high risk ratio values. Upon investigation of this, our analysis revealed that the maximum value of risk ratio for each year was 63.11, 62.65 and 61.48 respectively. This is very extreme and approximately translates to a 1.5 year return level in 2015. We would recommend to authorities around, to be even more prepared for wildfires in this region, especially if there wasn't any extreme wildfire issue in the past there.

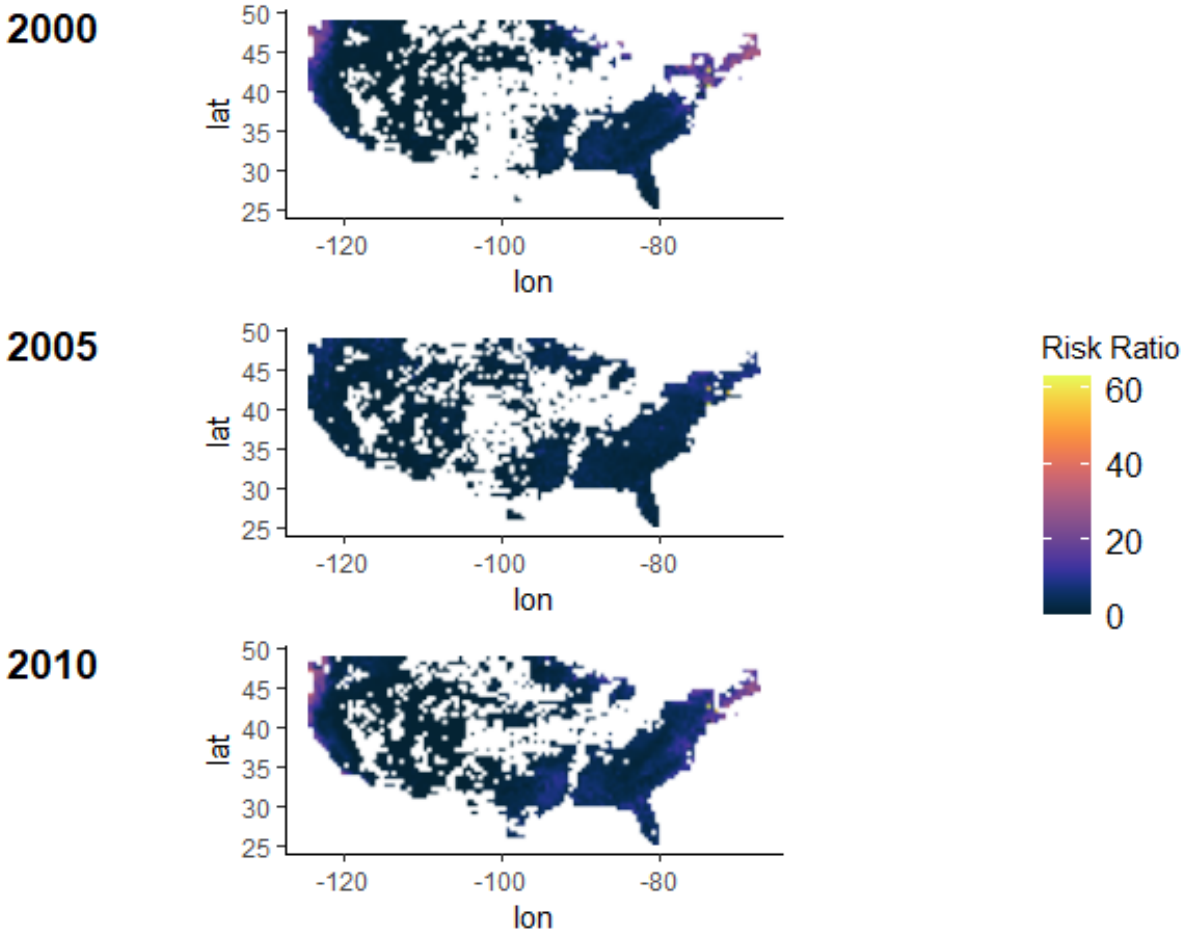

Figure 8: Map of USA, indicating the risk ratio of each Location, for 2000, 2005, and 2010

It is clear that the risk ratios are not constant across the entire spatial domain and so to aid for a more involved investigation of the distribution of risk ratios and to facilitate for computational ease, we have partitioned our spatial domain into 16 blocks, as shown in Figure 3. The risk ratio value at each block is calculated based on the spatial 100RLx risk ratio average for the locations in that block. Illustrated in Figure 9 and Table 2, we can see how the value changes across each block, for each of the three years. Having data for the three years allows us to make useful comparisons on how the risk ratio spatial average and block average is changing across the three years as Table 2 shows. We can see when compared to the 2000 climate, the 100RLx is 3.31 times likely to be exceeded in the 2015 climate, allowing for a fixed stationarity. This means that on average, a 100-year return level for 2000 loosely translates to a 30-year return level in 2015, with a similar result observed for the years 2005 and 2010. We can also see from Figure 9, how the 100RLx risk ratio value for 2015 varies across each of the 16 blocks for each of the three years. There is a clear trend, which matches our above analysis and illustrates how the prediction for the 100RLx in the eastern and north western regions in previous years is increasingly being

exceeded when considered in the 2015 environment. This is significant and on a whole, paints the picture of the potential for previously unaffected regions to become more prone to severe wildfires. Returning to Table 2, we can see that block 16, comprising the most north eastern region of the USA, is consistently having the maximum 100RLx exceedance for 2015, across the three years. We also observe that the overall block average 100RLx risk ratio maximum across the three years is occurring in block 16, at a value of 20.67, for 2015 when compared with 2000.

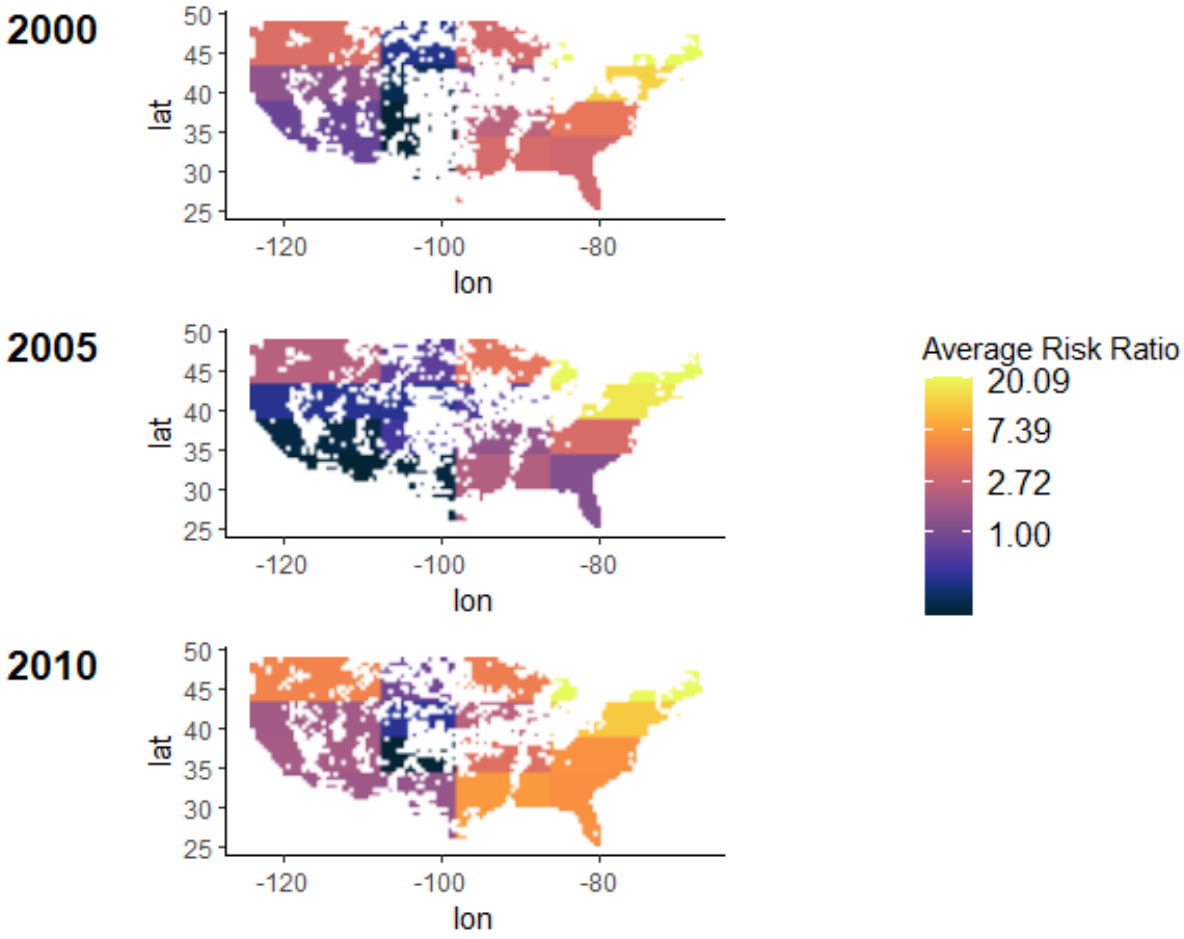

Figure 9: Map of USA, indicating the average risk ratio of each block, for 2000, 2005, and 2010

From a holistic perspective, we can see that the majority of the average block risk ratios for each year are large, with the regions prone to extreme wildfires, such as the south west, appearing to have risk ratios significantly smaller than those of the east, but with many still above one, especially in the year 2010, where majority of the block 100RLx risk ratios are greater than or close to one. This is quite significant as even the extreme blocks are being forecasted to have 100RLx exceeded at a rate greater than one in 2015.

Spatially Averaged Risk Ratios			
Region	**2000 - 2015**	**2005 - 2015**	**2010 - 2015**
Entire Domain	3.31	2.30	**3.58**
Block 1	0.79	0.88	1.50
Block 2	0.20	0.92	1.36
Block 3	3.23	2.37	5.13
Block 4	2.96	1.85	4.63
Block 5	0.81	0.93	1.64
Block 6	0.19	1.32	0.26
Block 7	2.35	2.02	2.97
Block 8	4.14	2.87	4.74
Block 9	1.33	1.20	1.55
Block 10	0.25	1.20	1.55
Block 11	1.21	1.41	2.00
Block 12	13.64	5.63	7.93
Block 13	3.50	2.43	3.94
Block 14	0.39	1.50	0.90
Block 15	3.21	3.12	3.66
Block 16	**20.67**	**6.12**	**12.45**

Table 2: Table of the spatially averaged risk ratios of each block and overall average across our domain, for 2015, with the years 2000, 2005 and 2010

Contour plots for the spline model

In this section we will investigate the contour plots of model estimates for the smooth function of longitude and latitude covariates, plotted obtained using evgam, for the location and logscale parameters, for each of the 4 years.

2000 :

Below we have the two contour plots for the predicted coefficients of location and logscale. Location's coefficient varies from -15 to 45 throughout the USA and this coefficient changes more rapidly towards the eastern USA. On the other hand, for the logscale's coefficient changes are smaller and have a range from -3 to 3.

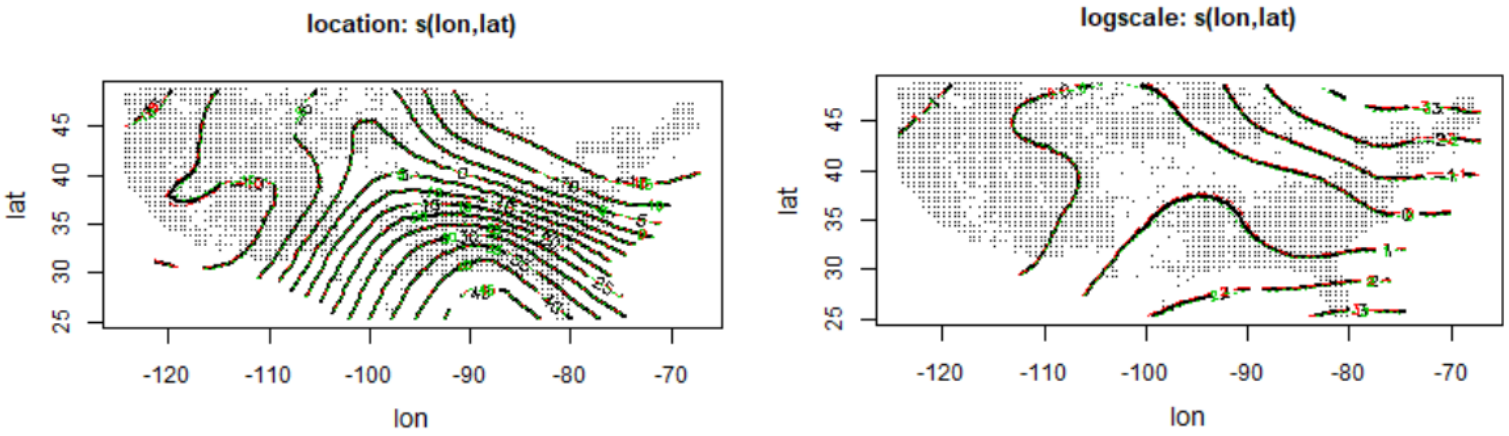

Figure 10: Contour plots for the predicted coefficients of the spline model using longitude and latitude, using the wildfire data of the year 2000

2005:

The Contour plots for this year are a lot smoother than year 2000 and predictions have a smaller range indicating that the coefficient values change slower. More precisely, location's coefficient varies from -8 to 10 and logscale's from -1.2 to 1.4.

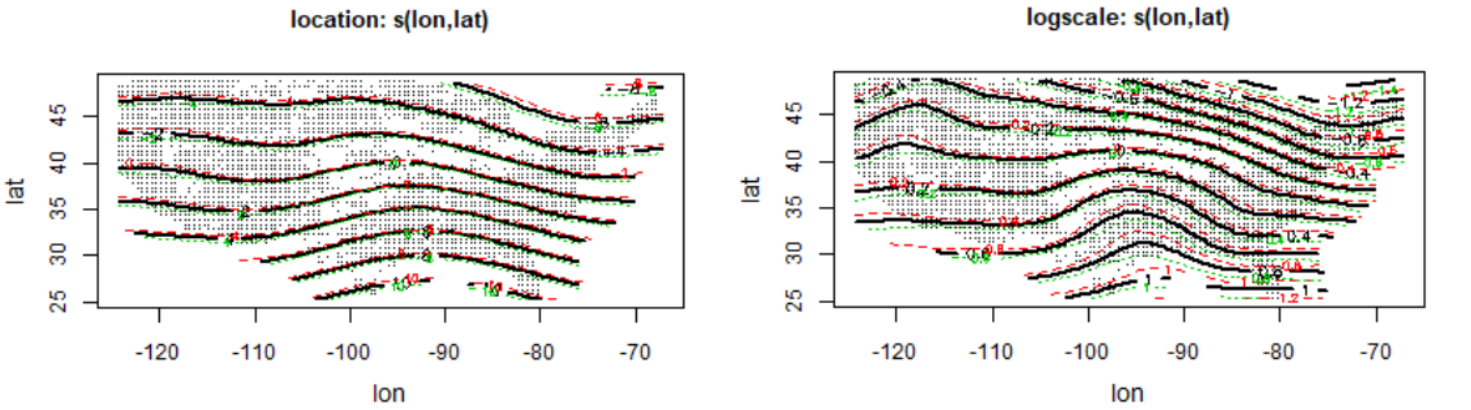

Figure 11: Contour plots for the predicted coefficients of the spline model using longitude and latitude, using the wildfire data of the year 2005

2010:
Contour plots for location coefficients vary from 20 to -5. We should note here that again these coefficients only vary in the eastern USA and remain constant for the western side. Logscale's coefficents vary from -1.5 to 1.5.

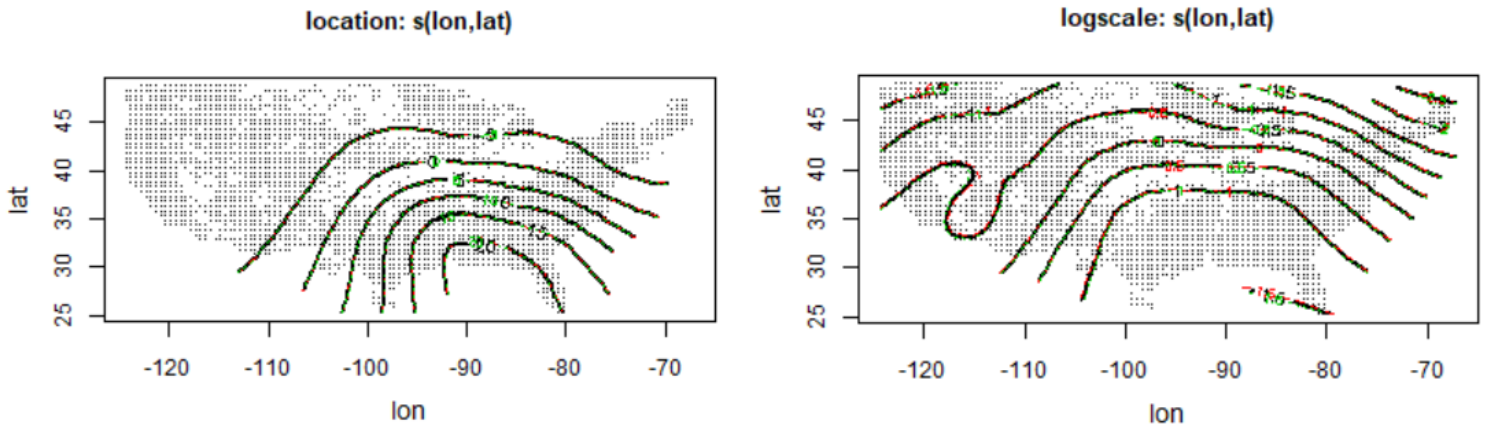

Figure 12: Contour plots for the predicted coefficients of the spline model using longitude and latitude, using the wildfire data of the year 2010

2015:
For year 2015 location's coefficients have a range from -2.5 to 2.5 and logscale's from -0.3 to 0.3. These ranges are the smallest from the previous four years and indicate small changes for the coefficients throughout the USA.

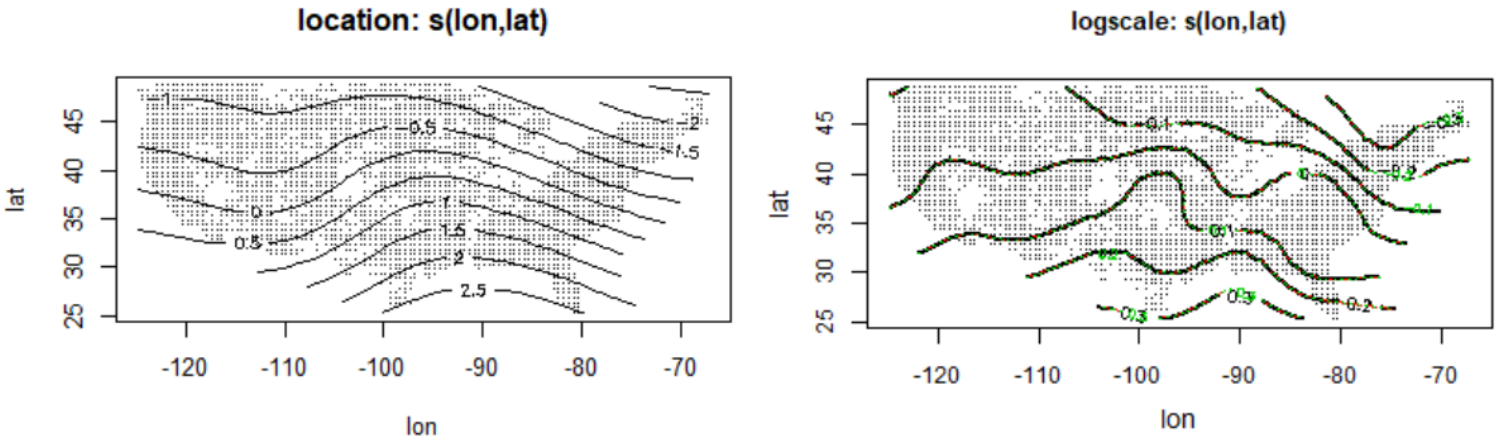

Figure 13: Contour plots for the predicted coefficients of the spline models using longitude and latitude, using the wildfire data of the year 2015

4 Appendix

4.1 Demonstrating convergence in distribution of block maxima to a GEV type distribution

In this section we are going to demonstrate the result of Theorem 3.11 (Coles, 2001), to show the convergence of any non-degenerate distribution of maxima to a GEV type distribution. For our demonstration, we are going to assume X_i are i.i.d. and $X_i \sim Beta(\alpha, \beta)$ for $i = 1 \dots n$. By our theorem, the distribution of block maximum is, $F^n(x_n) \to G(z)$ as $x_n \to 1$, where $x_n = a_n z + b_n$, for some non-negative sequence $\{a_n\}$ and a sequence $\{b_n\}$. $G(z)$ is the limit of the distribution of maxima, which we claim is one of three GEV types, i.e.

$$F^n(x_n) \to G(z)$$

Applying logs to each side and using that $F(x_n) = 1 - (1 - F(x_n))$, we have

$$n \log(1 - (1 - F(x_n))) \to \log(G(z)).$$

Then using the approximation $\log(1 + x) \approx x$ for small x, we get

$$-n(1 - F(x_n)) \to \log(G(z)),$$

and since the survival function is defined as, $\overline{F}(x_n) = 1 - F(x_n)$ we get that

$$n\overline{F}(x_n) \to -\log(G(z)) \tag{5}$$

Since the survival of a beta distribution,

$$\overline{F}(x_n) = \int_{x_n}^{1} \frac{1}{\beta e(\alpha, \beta)} z^{\alpha-1}(1-z)^{\beta-1} dz \tag{6}$$

cannot be evaluated analytically, we must approximate its value. This will be carried out by an iterative application of integration by parts. Upon our first application, calling (6) $= I_1$, we see that

$$I_1 = \frac{1}{\beta} \frac{1}{\beta e(\alpha, \beta)} \left[x_n^{\alpha-1}(1-x_n)^{\beta} - (\alpha - 1) \int_{x_n}^{1} \frac{1}{\beta e(\alpha, \beta)} z^{\alpha-2}(1-z)^{\beta} dz \right]$$

Now let $I_2 = \int_{x_n}^{1} \frac{1}{\beta e(\alpha,\beta)} z^{\alpha-2}(1-z)^{\beta} dz$ and then generally, $I_n = \int_{x_n}^{1} \frac{1}{\beta e(\alpha,\beta)} z^{\alpha-n}(1-z)^{\beta+(n-1)} dz$. Repeating the process iteratively on $I_2 \dots I_n$, and subbing into (6), we get that

$$\begin{aligned} I_1 =& \frac{1}{\beta} \frac{1}{\beta e(\alpha, \beta)} x_n^{\alpha-1}(1-x_n)^{\beta} \\ &\times \left[1 - \frac{(\alpha-1)}{x_n(\beta+1)}(1-x_n) - \frac{(\alpha-1)(\alpha-2)(1-x_n)^2}{(\beta+1)(\beta+2)x_n^2} - \dots - \frac{(\alpha-1)\dots(\alpha-(n-1))(1-x_n)^{n-1}}{(\beta+1)(\beta+2)\dots(\beta+(n+1))x_n^2} \dots \right] \end{aligned}$$

Since $x_n \to 1$ as $n \to \infty$, $(1 - x_n) \to 0$, so

$$\overline{F}(x_n) = I_1 \to \frac{1}{\beta} \frac{1}{\beta e(\alpha, \beta)} x_n^{\alpha-1}(1-x_n)^{\beta}.$$

Reintroducing this into (5), we have that

$$n \left(\frac{1}{\beta} \frac{1}{\beta e(\alpha, \beta)} x_n^{\alpha-1}(1-x_n)^{\beta} \right) \to -\log(G(z))$$

letting $K = \frac{1}{\beta} \frac{1}{\beta e(\alpha,\beta)}$ and observing that as $n \to \infty$, $x_n^{\alpha-1} \to 1$ and $(1 - x_n)^{\beta} \to 0$.

We want to find x_n which balances

$$(1 - x_n^\beta)n \simeq 1$$

we get that $(1 - x_n) \sim \frac{1}{n^{\frac{1}{\beta}}}$ and hence $x_n \sim 1 - \frac{1}{n^{\frac{1}{\beta}}}$.
As a result,

$$x_n = a_n z + b_n \sim 1 + \frac{(-z)}{n^{\frac{1}{\beta}}},$$

so we have that

$$Kn\left(1 - \left(1 + \frac{(-z)}{n^{\frac{1}{\beta}}}\right)\right) = Kn\left(\frac{(-z)}{n^{\frac{1}{\beta}}}\right)^\beta = -Kz^\beta.$$

Thus,

$$Kz^\beta \to \log(G(z))$$

Then, by exponentiating both sides, we arrive to

$$\exp(Kz^\beta) = \exp(K^{1/\beta}z)^\beta = \exp\left(\frac{z - 0}{K^{-1/\beta}}\right)^\beta \to G(z)$$

which is a reverse Weibull distribution with $\mu = 0$, $\sigma = K^{-1/\beta}$ and $\xi = \frac{-1}{\beta}$.

4.2 Demonstrating convergence in distribution of threshold maxima to a GPD distribution

We can apply the analogue of this for threshold maxima to show by Theorem 4.1 (Coles, 2001), the limiting distribution is a Generalised Pareto Distribution of the form $\tilde{H}(y) = (1 + \frac{\xi y}{\tilde{\sigma}})_+^{-1/\xi}$.
Like before, assume $X \sim Beta(\alpha, \beta)$, $x \in (0, 1)$ and define the survival function, $\overline{F}(x) = 1 - F(x)$, from our GEV derivation, $\overline{F}(x) \to \frac{1}{\beta}\frac{1}{Be(\alpha,\beta)}x^{\alpha-1}(1 - x)^\beta$.
Derivation for threshold exceedance:

$$\begin{aligned}
\mathbf{P}(X > u + h(u)y | X > u) &= \frac{\mathbf{P}(X > u + h(u))}{\mathbf{P}(X > u)} \\
&= \frac{\overline{F}(u + h(u)y)}{\overline{F}(u)} \\
&\approx \frac{\frac{1}{\beta}\frac{1}{Be(\alpha,\beta)}(u + h(u)y)^{\alpha-1}(1 - (u + h(u)y)^\beta}{\frac{1}{\beta}\frac{1}{Be(\alpha,\beta)}u^{\alpha-1}(1 - u)^\beta} \\
&= \frac{(u + h(u)y)^{\alpha-1}(1 - (u + h(u)y)^\beta}{u^{\alpha-1}(1 - u)^\beta} \qquad (7)
\end{aligned}$$

By the GEV exercise, we know:

$$\mathbf{P}\left(\frac{M_n - b_n}{a_n} \leq z\right) \to \exp\left(\frac{z - 0}{K^{-1/\beta}}\right)^\beta$$

which is a reverse weibull and corresponds to $\xi < 0$. We need to find $h(u)$ s.t. (7) converges to $\tilde{H}(y) = (1 + \frac{\xi y}{\tilde{\sigma}})_+^{-1/\xi}$.

Choosing $h(u) = \frac{\overline{F}(u)}{f(u)}$, the reciprocal of the hazard function,

$$h(u) \approx \frac{\frac{1}{\beta}\frac{1}{Be(\alpha,\beta)}(u)^{\alpha-1}(1-u)^{\beta}}{\frac{1}{\beta}\frac{1}{Be(\alpha,\beta)}u^{\alpha-1}(1-u)^{\beta-1}} = \frac{1/\beta}{(1-u)^{-1}} = \frac{1-u}{\beta}$$

hence, our guess is $h(u) = \frac{1-u}{\beta}$ and subbing this into (7) we get that

$$\begin{aligned}\frac{(u+\frac{1-u}{\beta}y))^{\alpha-1}(1-(u+\frac{1-u}{\beta}y))^{\beta}}{u^{\alpha-1}(1-u)^{\beta}} &= \frac{(1/\beta)^{\alpha+\beta-1}(u\beta+(1-u)y)^{\alpha-1}(1-u)^{\beta}(\beta-y)^{\beta}}{u^{\alpha-1}(1-u)^{\beta}} \\ &= \frac{(1/\beta)^{\alpha+\beta-1}(u\beta+(1-u)y)^{\alpha-1}(\beta-y)^{\beta}}{u^{\alpha-1}} \qquad (8)\end{aligned}$$

Taking the limit as $u \to 1$, $(1-u) \to 0$. Therefore, we get that

$$(8) \approx (\frac{\beta+y}{\beta})^{\beta} = (1+\frac{y}{\beta})^{\beta}$$

which corresponds to a GPD with $\xi = \frac{-1}{\beta}$ and $\tilde{\sigma} = -1$. Notice that for a GPD, the parameter ξ is the same as in the GEV case.

4.3 Functions created in R

```
# Function that calculates the log-likelihood of a GEV model.
# par is a vector of mu,sigma and xi, and Y is a vector containing
# all the readings of the input

gev_llik <- function(par,Y){

  m <- length(Y)
  sigma <- exp(par[2])
  mu <- par[1]
  xi <- par[3]

  for (val in Y) {
    if(1 + xi * (val - mu)/ sigma <= 0){
      return(-Inf)
    }
  }

  A <- -m * log(sigma)

  if(xi > -0.05 && xi < 0.05) {
    B <- -sum(Y - mu) / sigma
    C <- -sum(exp(-((Y - mu) / sigma)))
  } else {
    B <- -(1 + 1/xi) * sum(log(pmax(1 + xi * ((Y - mu)/sigma))))
    C <- -sum(pmax(1 + xi * ((Y - mu)/sigma))^(-1/xi))
  }
  return(A + B + C)
}

# Function that calculates the negative log-likelihood, so as to
# be used in the optim() function
neg_gev_llik <- function(par,Y){
```

```
  num <- gev_llik(par,Y)

  return(-num)
}

# Function that calculates the log-likelihood of a pareto model.
# par is a vector of sigma and xi, and Y is a vector containing
# all the readings of the input

pareto_llik <- function(par,Y){

  k <- length(Y)
  sigma <- exp(par[1])
  xi <- par[2]

  for (val in Y) {
    if(1 + xi * val / sigma <= 0){
      return(-Inf)
    }
  }

  A <- - k * log(sigma)

  if(xi > -0.05 && xi < 0.05) {
    B <- -sum(Y) / sigma
  } else {
    B <- -(1 + 1/xi) * sum(log(pmax(1 + xi * (Y / sigma))))
  }


  return(A + B)
}

# A function to find the 1/p return level.
# The parameters are a vector of the estimates of mu,sigma and xi
# and p is to determine the magnitude of the return level.

return_level <- function(par, p) {

  sigma <- exp(par[2])
  mu <- par[1]
  xi <- par[3]
  y_p <- -log(1-p)

  if(xi > -0.05 && xi < 0.05) {
    z_p <- mu - sigma * log(y_p)
  } else{
    z_p <- mu - (sigma/xi) * (1-y_p^(-xi))
  }



  return(z_p)
}
```

References

Alkhatib, Ahmad A. A. (2014). *A Review on Forest Fire Detection Techniques*. Sage journals. URL: https://doi.org/10.1155%2F2014%2F597368 (visited on 03/10/2021).

Borunda, Alejandra (2020). *The science connecting wildfires to climate change*. National Geographic. URL: https://www.nationalgeographic.com/science/article/climate-change-increases-risk-fires-western-us (visited on 01/22/2021).

Coles, Stuart (2001). *An Introduction to Statistical Modelling of Extreme Values*.

Gammon, Katharine (2021). *California's rainfall is at historic lows.* BBC. URL: https://www.theguardian.com/us-news/2021/feb/11/california-dry-weather-drought-wildfire-agriculture (visited on 02/11/2022).

Graeme Auld Gabriele C. Hegerl, Ioannis Papastathopoulos (2021). *Changes in the distribution of observed annual maximum temperatures in Europe*. arXiv. URL: https://arxiv.org/pdf/2112.15117.pdf (visited on 03/17/2021).

Harrison, Sandy P (2021). *Understanding and modelling wildfire regimes: an ecological perspective*. Environmental Research. URL: https://doi.org/10.1088/1748-9326/ac39be (visited on 03/10/2021).

Jack Beckwith, Michael Hester and Tyler Wolf (2018). *Wildfires occurence*. The DataFace. URL: https://thedataface.com/2018/11/public-health/wildfires-map (visited on 02/12/2021).

NBC (2020). *Wildfire in California*. NBC. URL: https://www.nbcnews.com/news/us-news/blistering-heat-continues-western-u-s-wildfire-threat-intensifies-n1237087 (visited on 02/21/2022).

NCDC (2021). *2021 US wildfires*. ncdc. URL: https://www.ncdc.noaa.gov/sotc/fire/202113 (visited on 02/21/2021).

R-Documentation (2022). *General-purpose Optimization*. DataCamp. URL: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/optim (visited on 02/22/2021).

UoE, School of Mathematics (2021). *Data Challenge*. The University of Edinburgh. URL: https://www.maths.ed.ac.uk/school-of-mathematics/eva-2021/competitions/data-challenge (visited on 02/11/2021).

Virginia Iglesias Jennifer K. Balch, William R. Travis (2022). *U.S. fires became larger, more frequent, and more widespread in the 2000s*. Science Advances. URL: https://doi.org/10.1126/sciadv.abc0020 (visited on 03/17/2021).

Wikipedia (2021). *AIC*. Wikipedia. URL: https://en.wikipedia.org/wiki/Akaike_information_criterion (visited on 02/22/2021).

Wood, Simon N. (2006). *Generalised Additive Models*.

Youngman, Benjamin D. (2003). *evgam: An R package for Generalized Additive Extreme Value Models*. Journal of Statistical Software. URL: https://arxiv.org/pdf/2003.04067.pdf (visited on 02/22/2021).