Extremes (2023) 26:381–398
https://doi.org/10.1007/s10687-023-00469-7

A marginal modelling approach for predicting wildfire extremes across the contiguous United States

Eleanor D'Arcy[1] · Callum J. R. Murphy-Barltrop[1] · Rob Shooter[2] · Emma S. Simpson[3]

Received: 31 December 2021 / Revised: 14 March 2023 / Accepted: 20 March 2023 /
Published online: 1 April 2023

Abstract

This paper details a methodology proposed for the EVA 2021 conference data challenge. The aim of this challenge was to predict the number and size of wildfires over the contiguous US between 1993 and 2015, with more importance placed on extreme events. In the data set provided, over 14% of both wildfire count and burnt area observations are missing; the objective of the data challenge was to estimate a range of marginal probabilities from the distribution functions of these missing observations. To enable this prediction, we make the assumption that the marginal distribution of a missing observation can be informed using non-missing data from neighbouring locations. In our method, we select spatial neighbourhoods for each missing observation and fit marginal models to non-missing observations in these regions. For the wildfire counts, we assume the compiled data sets follow a zero-inflated negative binomial distribution, while for burnt area values, we model the bulk and tail of each compiled data set using non-parametric and parametric techniques, respectively. Cross validation is used to select tuning parameters, and the resulting predictions are shown to significantly outperform the benchmark method proposed in the challenge outline. We conclude with a discussion of our modelling framework, and evaluate ways in which it could be extended.

Keywords Extreme value theory · Semi-parametric modelling · Wildfire prediction

Mathematics Subject Classification 62

Eleanor D'Arcy, Callum J. R. Murphy-Barltrop, Rob Shooter and Emma S. Simpson are authors who all contributed equally to this work.

✉ Callum J. R. Murphy-Barltrop
c.barltrop@lancaster.ac.uk

Extended author information available on the last page of the article

1 Introduction

1.1 Motivation and data description

This paper details an approach to the data challenge organised for the EVA 2021 conference. The subject of the challenge was wildfire modelling, and two important sub-challenges were proposed within this setting. In particular, teams were asked to develop methods for predicting the number of fires (i.e., individual fires that are separated in space), as well as the amount of burnt land resulting from these fires, over different months for gridded locations across the continental United States (US).

In the absence of mitigation, wildfires can have devastating consequences, including loss of life and damage to property. The northern California wildfire in October 2017 burned approximately 150,000 acres of land, resulting in 7,000 damaged structures and 100,000 evacuations [1]. Recent increases in both the number and severity of wildfires can be linked to climate change, and in particular to anthropogenic warming [2]. Focusing specifically on the western US, [3] demonstrate that a high proportion of the observed increases in weather events leading to wildfires may be attributed to this aspect of climate change. Extreme events in wildfire modelling are especially important; the more individual wildfires that occur, the greater the potential destruction, and the impact of large wildfires (in terms of the amount of land area burnt) can be particularly devastating. It is therefore of interest to develop models for wildfires, and in particular wildfire extremes.

The challenge data set consists of monthly wildfire count (CNT) and burnt area (BA) observations from 1993 to 2015 at 3,503 grid cell locations spanning the contiguous US. There are 35 auxiliary variables also recorded relating to land cover types, climate and altitude. Observation locations are arranged on a $0.5° \times 0.5°$ (approximately 55 km $\times$ 55 km) regular grid of longitude and latitude coordinates, with observations recorded from March to September; further details are provided by [4].

In order to compare the predictions produced by the teams participating in the data challenge, several observations were removed from the data to act as a validation set; this contained 80,000 observations for each of CNT and BA. The selection of these validation points was not done completely at random, so there is some spatio-temporal dependence between them. This will be discussed further in Section 3.4, with a pictorial example given in Fig. 4. Let CNT_i and BA_i, $i = 1, \ldots, N$, denote the i-th observation of the wildfire CNT or BA data, respectively, where $N = 563,983$ is the total number of observations across the training and validation sets for each variable over all sites, months and years. We denote the set of observation indices in the validation sets for CNT and BA by $CNT^{val}, BA^{val} \subset \{1, \ldots, N\}$, respectively, with $|CNT^{val}| = |BA^{val}| = 80,000$. An important feature is that the validation indices are not identical for the CNT and BA data, but there is a reasonable overlap, i.e., $CNT^{val} \neq BA^{val}$ but $CNT^{val} \cap BA^{val} \neq \emptyset$. We discuss ways to exploit this aspect in Section 2.2.

The objective of the challenge was to predict cumulative probability values for both CNT and BA at the times and locations in their respective validation sets.

The resulting estimates were then ranked using a score computed from the true observed values, with lower scores corresponding to more accurate probability predictions. These scores were weighted so that more importance is placed on the estimation of the extremes; see [4]. Statistical techniques that do not explicitly model the tail are therefore unlikely to produce the best scores.

1.2 Data exploration

In this section, we give an overview of features of the data set that motivate our modelling approach. We consider the relationship between CNT and BA, as well as the temporal non-stationarity of each variable separately; we also investigate how these features vary over the spatial domain.

We begin by exploring the dependence between CNT and BA; for the bulk of the data, we consider Kendall's τ measure of rank correlation, whilst for the extremes we consider the widely-used measures χ and $\bar{\chi}$. Consider a random vector (X, Y) with marginal distribution functions F_X and F_Y, respectively. Coles et al. [5] define $\chi = \lim_{u\to 1} \chi(u)$, where $\chi(u) = \Pr(F_Y(Y) > u \mid F_X(X) > u) \in [0, 1]$, as a measure of asymptotic dependence. If $\chi \in (0, 1]$, X and Y are said to be asymptotically dependent, with $\chi = 1$ corresponding to perfect dependence. Asymptotic independence between X and Y is present only when $\chi = 0$, meaning that χ fails to signify the level of asymptotic independence. To account for this, [5] define a further measure that provides additional detail in this case, namely $\bar{\chi} = \lim_{u\to 1} \bar{\chi}(u) \in (-1, 1]$ where

$$\bar{\chi}(u) = \frac{2 \log \Pr(F_Y(Y) > u)}{\log \mathbb{P}(F_Y(Y) > u, F_X(X) > u)} - 1.$$

Under asymptotic dependence, $\bar{\chi} = 1$, and for asymptotic independence, $\bar{\chi} < 1$; the further sub-cases $\bar{\chi} \in (0, 1)$ and $\bar{\chi} \in (-1, 0)$ correspond to positive and negative association, respectively, while $\bar{\chi} = 0$ indicates independence.

We estimate these measures separately for subsections of the US to investigate spatial variability in the dependence structure between CNT and BA. We start by splitting the spatial domain into quadrants corresponding to the north east (NE; > 37.5°N, < 100°W), south east (SE; ≤ 37.5°N, < 100°W), south west (SW; ≤ 37.5°N, ≥ 100°W) and north west (NW; > 37.5°N, ≥ 100°W). Kendall's τ measure suggests strong overall correlation between CNT and BA, with estimates of 0.926 (0.925, 0.927), 0.827 (0.825, 0.829), 0.858 (0.855, 0.860) and 0.868 (0.867, 0.870) for the NE, SE, SW and NW respectively, with the values in brackets denoting 95% confidence intervals obtained via bootstrapping. However, estimates of $\chi(u)$ and $\bar{\chi}(u)$ suggest this dependence diminishes in the extremes, leading to asymptotic independence. We obtain estimates (and 95% bootstrap confidence intervals) of $\chi(0.999) = 0.071(0.043, 0.126)$, $0.038(0.017, 0.072)$, $0.012(0,\ 0.024)$ and $0.043(0.024,\ 0.077)$, and $\bar{\chi}(0.999) = 0.438(0.343, 0.521)$, $0.282(0.191,\ 0.392)$, $0.092(-0.05, 0.179)$ and $0.317(0.253,\ 0.413)$, for the NE, SE, SW and NW regions respectively. The NE region exhibits the strongest dependence between CNT and BA in the bulk of the data, as well as the strongest extremal dependence. We extended this analysis to look at smaller spatial domains, but our conclusions did not change.

Figure 1 shows the spatial distribution of average CNT and BA values in two different time groupings: in the summer months (May, June, July and August; MJJA), when wildfires are more likely to occur, and in the remaining cooler months (March, April and September; MAS). The highest average CNT values are observed in the east for MAS and the west for MJJA. The highest average BA values typically occur in the west of the US during MJJA whilst the majority of the eastern US locations have relatively low average BA values in both time groups, with the exception of Florida. This demonstrates that there is both spatial and temporal variability in the wildfire observations.

To further demonstrate this spatio-temporal variability, Fig. 2 illustrates the months when the maximum CNT and BA observations occur for each grid cell. In the eastern US, the maxima of each variable tend to occur in July or August (shown by red points) whilst in the west, the maxima typically occur in March and April (illustrated by lighter yellow points). As global temperatures rise with anthropogenic climate change, the frequency and intensity of wildfires are generally expected to increase [6, 7]. To investigate this, we fit a linear model between year and annual mean CNT and BA separately, assuming independence across annual means. We find significant trends for both CNT and BA. Therefore, assuming stationarity across the entire spatial domain over the observation period would be unreasonable.

Due to the nature of wildfires, we expect to observe relationships between both CNT and BA observations and certain climate variables. For example, high temperature coupled with low rainfall and low wind speed are the ideal conditions for

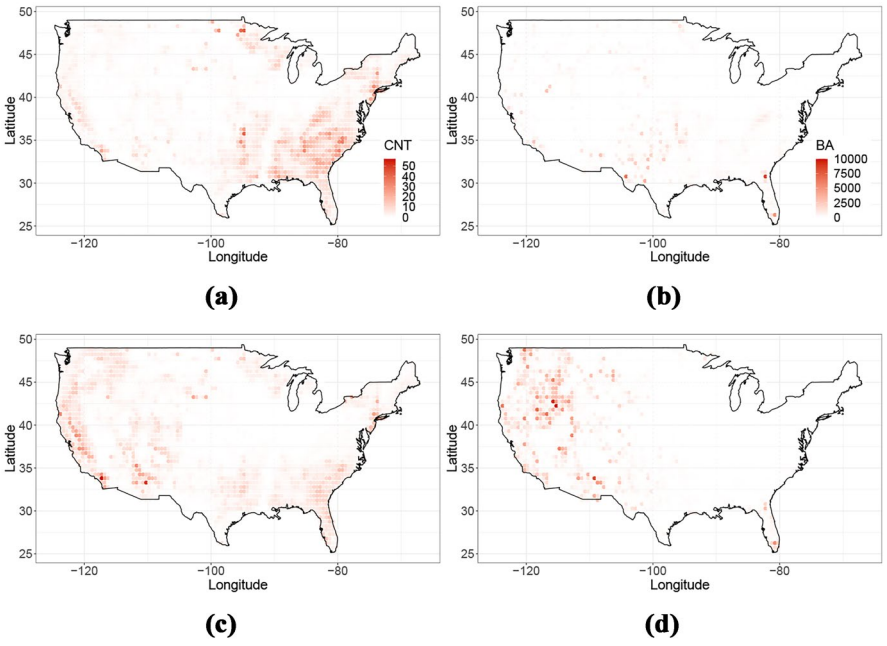

Fig. 1 Average CNT (**a** & **c**) and BA (**b** & **d**) across all years for each grid cell, for MAS (**a** & **b**) and for MJJA (**c** & **d**)

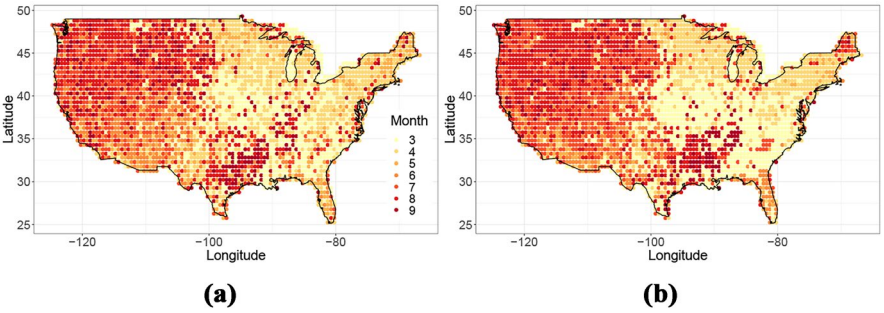

Fig. 2 Month where the maximum CNT (**a**) and BA (**b**) across all years occurs for each grid cell

wildfires to ignite and spread [8, 9]. No significant linear relationships exist for either wildfire variable with any of the climate covariates, suggesting such relationships are complex in nature. Figure 3a shows the average temperature for each grid cell; temperature is non-stationary across the US but there is some spatial dependence, with nearby locations exhibiting similar values. Some form of spatial dependence exists for all climate variables. Since these variables are given as monthly averages, it is difficult to associate these covariates directly with the wildfire observations, which are also given as monthly aggregates.

Another factor likely to alter wildfire behaviour across the US is the type of land cover. For example, locations with large proportions of water or urban areas are typically not conducive to wildfires, whilst those with forest areas probably are. Eighteen land cover variables, given as proportions of each grid cell, are provided in the challenge data set; these are denoted $lc(j)$ for $j = 1, \ldots, 18$ and defined in [4]. Figure 3b illustrates the maximum land cover variable for each location. Spatial heterogeneity can be observed over different regions. For example, a large portion of the western US is taken up by shrubland ($lc(11)$), whereas the eastern region is dominated by cropland ($lc(j)$ for $j = 1, 2, 3$) and tree-based land cover types ($lc(j)$ for $j = 5, 6, 7, 8$). Unsurprisingly, many coastal locations are predominantly covered by water ($lc(18)$), and regions containing national forests (such as Kootenai and Stanislaus) are easily identifiable, since they are mostly made up of tree categories.

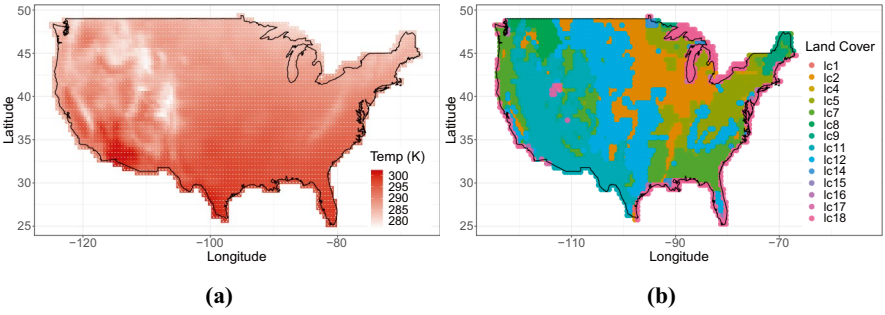

Fig. 3 Mean temperature in Kelvin (**a**) and the most common land cover variable (**b**) for each location

1.3 Existing methods

Various methods exist for modelling and predicting wildfire frequency and intensity. For example, generalised additive models (GAMs) with climatic, anthropogenic and/or spatial covariates are commonly used; see, e.g., [10] or [11]. The latter captures covariate information via a fire index; many such indices have been proposed within the literature [12]. Each index is typically developed with country-specific considerations in mind, such as land cover types and climate factors, and are often used by government bodies to assess risks and prioritise fire responses. In the US, the National Fire-Danger Rating System is the primary tool used for wildfire management [13]. There have been attempts in the literature to use fire indices as a means to model extreme wildfire events [14]. However, several approaches have found that certain fire indices are poor predictors of wildfires. For example, [15] show that the Forest Fire Danger Index, typically used in Australia, is inadequate for predicting the behaviour of moderate to high-intensity wildfires.

Machine learning techniques have also been adopted for wildfire modelling: [16] and [17] use deep learning techniques; [18] present a four-stage process including a random forest algorithm; and [19] develops a gradient boosting model trained with loss functions appropriate for predicting extreme values. We take a simpler, marginal-based approach.

The remainder of this paper is structured as follows. In Section 2, we illustrate how certain properties of the training data can be exploited to infer a subset of probability estimates for observations in the validation set. In Section 3, we introduce our marginal modelling techniques for both CNT and BA. We also discuss our technique for estimating spatial neighbourhoods and corresponding tuning parameters. We conclude with a discussion of our approach in Section 4.

2 Exploiting properties of the training data set

2.1 Re-scaling burnt area values

In this section, we discuss various properties of the wildfire data set, and how these can be exploited to improve the estimation of the predictive distributions for missing observations.

To begin, observe that BA is an absolute measurement; this results in varying measurement scales across different locations. To better understand this, consider that some grid cells in the data set do not lie completely inside the continental US; this feature is captured by the 'area' variable, denoted p_i, $i = 1, \ldots, N$, which describes the proportion of each grid cell that lies in the region of interest. BA observations depend upon this variable since for grid cells with smaller area values, there is less available land for wildfires to occur and hence lower BA values. For these reasons, the raw BA observations cannot be easily compared across locations.

To account for this, we propose re-scaling BA observations to ensure all observations are on a unified, relative scale. Recall that BA_i, $i = 1, \ldots, N$, with $N = 563,983$, denotes the i-th observation of the BA data, and that $BA^{val} \subset \{1, \ldots, N\}$ is the set of indices for missing BA observations. We consider here the i-th observation, with

corresponding grid cell area $p_i \in (0, 1]$. For each $i \in \{1, \ldots, N\}$, the total surface area of the grid cell is computed by taking the corresponding longitude and latitude coordinates and applying a formula derived from Archimedes' theorem [20]. We denote these surface area values by SA_i. The surface area contained within the continental US is then computed by multiplying the total surface area by the grid cell area variable, i.e., $SA_i \times p_i$. We denote these values by SA_i^*: such values will naturally vary between locations, especially for locations lying on a borderline. Moreover, SA_i^* values naturally decrease going from South to North of the continent, since grid cells defined using longitude and latitude suffer from unequal cell sizes [21]. We refer to this variable as the true surface area.

Using this variable, we derive a relative measure for BA, which we term burnt area proportion (BAP), i.e., for each i, define $BAP_i := BA_i/SA_i^* \in [0, 1]$. This value denotes the proportion of the true surface area that has been burnt for each observation. It is arguably a better indicator of the impact and/or severity of wildfire events compared to raw BA observations, since it puts the absolute magnitude in context for each location. Moreover, this proportion is a relative measure, meaning the data for all locations are on the same scale; this allows for a more straightforward comparison between neighbouring observations with different (true) surface areas.

We recall that the objective of the data challenge is to obtain probability estimates of the form $\Pr(BA_i \leq u)$ for all $u \in \mathcal{U}_{BA}$, where

$$\begin{aligned}\mathcal{U}_{BA} = \{&0, 1, 10, 20, 30, \ldots, 100, 150, 200, 250, 300, 400, 500, 1000,\\ &1500, 2000, 5000, 10000, 20000, 30000, 40000, 50000, 100000\},\end{aligned}$$

and $i \in BA^{val}$. This can be derived using the marginal distribution of BAP_i, since

$$\Pr(BA_i \leq u) = \Pr(BAP_i \times SA_i^* \leq u) = \Pr(BAP_i \leq u/SA_i^*).$$

Consequently, we evaluate the distribution function of BAP_i for all $u \in \mathcal{U}_{BAP}^i$, where $\mathcal{U}_{BAP}^i := \mathcal{U}_{BA}/SA_i^*$, to obtain the required predictive probabilities. We introduce our technique for estimating this distribution function in Section 3.3.

We can also use these proportional data to deduce information about the upper tail of the distribution for BAP_i at any $i \in BA^{val}$. Since it is impossible to observe a BA observation which exceeds the true surface area at any location, we can immediately deduce that $\Pr(BAP_i \leq u) = 1$ for any $u \in \mathcal{U}_{BAP}^i$ with $u \geq 1$. In practice, over 1% of missing BA observations satisfied the inequality $\max\{\mathcal{U}_{BAP}^i\} \geq 1$, meaning a non-negligible amount of information can be uncovered via this preliminary step.

We considered a similar re-scaling for CNT observations; however, there did not appear to be any obvious relationship between the true surface area and CNT values. Furthermore, unlike BA, no natural upper bound arises for CNT observations, so we cannot deduce properties of the upper tail distribution for missing observations.

2.2 Exploiting features of the missing data

Before introducing our marginal modelling procedures, we highlight how the training data can be used to provide information about the missing values we are required

to estimate. This is possible since the missing values in the CNT and BA variables do not always occur at the same space-time locations, although there is some overlap in their missingness. We show that if exactly one of the CNT or BA values is known at a particular index, we can deduce information about the other.

Recall that we are interested in estimating the predictive distribution of CNT_i for some $i \in CNT^{val}$, i.e., $\Pr(CNT_i \leq u)$ for $u \in \mathcal{U}_{CNT}$, where $u \in \mathcal{U}_{CNT} = \{0, 1, \ldots, 9, 10, 12, \ldots, 30, 40, \ldots, 100\}$. If $i \notin BA^{val}$ and $BAP_i = 0$, we can immediately deduce that $CNT_i = 0$ and $\Pr(CNT_i \leq u) = 1$ for all $u \in \mathcal{U}_{CNT}$. Moreover, if $i \notin BA^{val}$ and $BAP_i > 0$, we have that $CNT_i > 0$, implying $\Pr(CNT_i \leq 0) = 0$, though we are still required to estimate the predictive distribution for all $u \in \mathcal{U}_{CNT} \backslash \{0\}$. The values we can infer for BAP_i from CNT_i, with $i \in BA^{val}$, are analogous, so the detail is omitted here.

We find that $CNT_i = 0$ for approximately 23% of the points in the CNT validation set, and that $CNT_i > 0$ for an additional 15%. We can also deduce similar proportions for the BAP values we are required to predict. A reasonable amount of information can therefore be uncovered using this simple step.

We also found that for the non-missing CNT and BA observations, the probability of observing a zero observation exceeded 0.999 for both variables when $lc(18)_i > 0.94$, where $lc(18)_i$ denotes the proportion of each location covered by water. Therefore, for any $i \in CNT_{val}$ ($i \in BA_{val}$) with $lc(18)_i > 0.94$, we set $CNT_i = 0$ ($BAP_i = 0$), implying $\Pr(CNT_i \leq u) = 1, \forall\, u \in \mathcal{U}_{CNT}$ ($\Pr(BAP_i \leq u) = 1, \forall\, u \in \mathcal{U}_{BAP}$).

As well as improving estimates of the predictive distribution of some locations, the additional steps introduced in this section also increase the amount of information available. This aids the marginal estimation procedures detailed in Sections 3.2 and 3.3.

3 Marginal modelling of missing values

3.1 Neighbourhood selection

For our approach, we make the following assumption: for any observation with index $i \in \{1, \ldots, N\}$, there exists some spatial neighbourhood of indices, $\mathcal{N}_i$, where all corresponding observations come from the same marginal distribution. Through estimation of this distribution, we can obtain predictive probabilities for missing CNT and BAP observations. In this section, we introduce our approach for selecting these neighbourhoods for CNT observations; the approach for BAP is analogous.

Consider the observation with index $i \in \{1, \ldots, N\}$, and denote the corresponding spatial location, month and year by $\boldsymbol{s}_i \in \mathbb{R}^2$, $m_i \in \{3, \ldots, 9\}$ and $y_i \in \{1993, \ldots, 2015\}$, respectively. We define the spatial neighbourhood as

$$\mathcal{N}_i := \left\{ j \in \{1, \ldots, N\} \,:\, \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \leq k_1^{CNT}, m_j = m_i, y_j = y_i \right\}, \tag{1}$$

for some $k_1^{CNT} \geq 0$, i.e., the indices of all observations occurring in the same year and month as observation i with a spatial distance of at most k_1^{CNT} from $\boldsymbol{s}_i$. The spatial distance $\|\cdot\|$ is measured in kilometres (km) using the Haversine formula;

in practice, these are calculated via the `distm` function in the `R` package `geosphere` [22]. We treat k_1^{CNT} as a tuning parameter and introduce a cross validation technique to select it in Section 3.4. We denote the *CNT* values corresponding to neighbourhood $\mathcal{N}_i$ by $CNT^{\mathcal{N}_i} = \{CNT_j : j \in \mathcal{N}_i\}$. The definitions of k_1^{BAP} and $BAP^{\mathcal{N}_i}$ for $i \in \{1, \ldots, N\}$ are analogous.

More complex spatial neighbourhoods, which incorporated temporal and covariate-based information, were also considered but ultimately resulted in worse quality marginal estimates. This is discussed in the Appendix, where we present prediction scores for other neighbourhoods we considered.

3.2 A parametric approach for modelling CNT

Following [23], we assume all observations in the set $CNT^{\mathcal{N}_i}$ follow a zero-inflated negative binomial distribution for all $i \in CNT^{val}$, i.e., for any $CNT \in CNT^{\mathcal{N}_i}$, we have that

$$\Pr(CNT = j) = \begin{cases} \pi + (1-\pi)g(0) & \text{if } j = 0, \\ (1-\pi)g(j) & \text{if } j > 0, \end{cases} \tag{2}$$

where $\pi \in [0, 1]$ denotes a probability and $g(j)$, $j \geq 0$, is the probability mass function of the negative binomial distribution. We estimate the parameter π and those of the negative binomial distribution using likelihood inference. We then evaluate distribution (2) for all $u \in \mathcal{U}_{CNT}$ using the estimated parameters, resulting in the predictive distribution for the missing observation CNT_i. In practice, we use the same tuning parameter, k_1^{CNT}, for all $i \in CNT^{val}$; we discuss our approach to selecting this value in the Section 3.4.

3.3 A semi-parametric approach for modelling BAP

Given any $i \in BA^{val}$, we assume all observations in the set $BAP^{\mathcal{N}_i}$ follow the semi-parametric marginal distribution given in [24]. This distribution was proposed for modelling precipitation data, which are similar to wildfire data in the sense that they typically contain a large number of zero observations. These data structures are referred to as mixture distributions, since they are a mix of a discrete (zero observations) and a continuous (positive BAP observations) process. Values in the bulk of the data, including zeros, are modelled empirically, while values in the upper tail are modelled using a generalised Pareto distribution (GPD). This distribution is typically referred to in the context of the 'peaks over threshold' approach [25, 26], whereby a GPD is fitted to independent and identically distributed exceedances of a high threshold. This overall marginal model is given by

$$\Pr(BAP \leq x) = \begin{cases} z_i & \text{if } x = 0, \\ \frac{1-\lambda_i-z_i}{F_i^*(u_i)} F_i^*(x) + z_i & \text{if } 0 < x \leq u_i, \\ 1 - \lambda_i(1 - H_{u_i}(x)) & \text{if } x > u_i, \end{cases} \tag{3}$$

for all $BAP \in BAP^{\mathcal{N}_i}$, where z_i is the probability of observing a zero, u_i is some high threshold to be chosen, $\lambda_i = \Pr(BAP > u_i)$, F_i^* is the distribution function of strictly positive observations, and $H_{u_i}(x)$ denotes the cumulative distribution function of the GPD, i.e., $H_{u_i}(x) = 1 - \left[1 + (\xi_i(x - u_i))/(\sigma_i)\right]_+^{-1/\xi_i}$, with $x_+ = \max\{x, 0\}$ and $(\sigma_i, \xi_i) \in \mathbb{R}_+ \times \mathbb{R}$. We refer to σ_i and ξ_i as the scale and shape parameters, respectively. See [27] for a more detailed discussion of the peaks over threshold approach.

We set $k_2^{BAP} := 1 - \lambda_i$ for all $i \in BA^{val}$ and treat k_2^{BAP} as another tuning parameter, which we again estimate using cross validation; see Section 3.4. Both u_i and z_i can be estimated empirically, alongside F_i^*. Note that this marginal model is only valid when $z_i < 1 - \lambda_i$: in such cases, the GPD scale and shape parameters are estimated using likelihood inference. We then evaluate the distribution described in Eq. (3) at the fitted parameters for all $u \in \mathcal{U}_{BAP}^i$, resulting in the predictive distribution for the missing observation BAP_i.

In the cases when $z_i \geq 1 - \lambda_i$ (i.e., the marginal model is not valid), we use a fully empirical distribution. Such cases occur when the estimated threshold equals zero, corresponding to neighbourhood sets containing a significant proportion of zeros, indicating a low occurrence of wildfires.

3.4 Tuning parameter selection

We now consider how to select the tuning parameters k_1^{CNT}, k_1^{BAP} and k_2^{BAP} used in our marginal modelling approaches. One option is to use leave-one-out cross validation and select the tuning parameter values that minimise the score used for ranking in the data challenge: see [4] for more information. However, the locations for the validation data are not randomly distributed across the spatial domain, and are generally clustered in space and time. We demonstrate this in Fig. 4, where we show the locations of the CNT validation data for March 1994; the resulting plots have similar features for BAP, as well as for different months and years. In the case of BAP data, this implies that for a fixed value of k_1^{BAP}, there will be a larger number of missing values in the set $BAP^{\mathcal{N}_i}$ for $i \in BA^{val}$ than for $i \notin BA^{val}$, on average. The same holds when considering CNT data for a fixed value of k_1^{CNT}. This feature of the validation set means that using standard leave-one-out cross validation over all training locations could lead to selecting smaller neighbourhoods than are really appropriate.

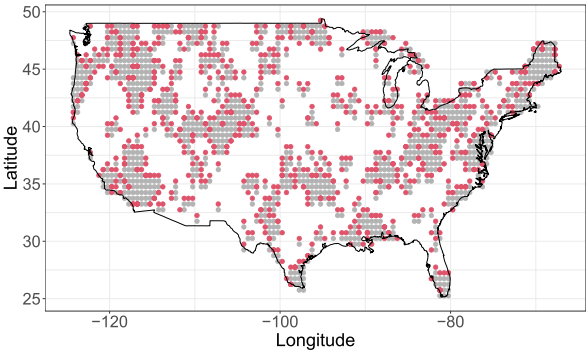

Fig. 4 Locations in the set CNT^{val} for March 1994 (grey) and the corresponding locations of observations for tuning parameter selection (red)

We instead propose to carry out the parameter selection procedure using only a subset of the observations in the training data. Focusing on BAP, for each $i \in BA^{val}$ and any combination of (k_1^{BAP}, k_2^{BAP}) values, we allow the observation indexed by

$$\underset{j: m_i = m_j, y_i = y_j, j \notin BA^{val}}{\arg\min} \|\boldsymbol{s}_i - \boldsymbol{s}_j\|,$$

to contribute to the score, i.e., giving the spatially-nearest non-missing observation that occurs in the same month and year as observation i. Ties may be broken at random, or using any rule that results in only one nearest neighbour per location. Since these locations can be the nearest neighbour of more than one validation location, some of the corresponding observations are included more than once in the score calculation. An alternative would have been to use each of these observations only once to avoid duplicates, but this means some validation locations would not be represented in the score calculation.

We consider the following candidate values for the tuning parameters: $k_1^{BAP} \in \{50, 75, \ldots, 400\}$, $k_2^{BAP} \in \{0.05, 0.10, \ldots, 0.95\}$. For each combination of candidate values, we recalculate the score function proposed in [4], summing over all values corresponding to our set of nearest neighbours, before finally selecting the parameter combination that minimises the score. The procedure in the CNT case is analogous, albeit without the GPD quantile parameter k_2^{BAP}. In Fig. 4, we demonstrate the locations of observations that contribute to the tuning parameter selection procedure for CNT in March 1994. This results in selected tuning parameter values of $k_1^{BAP} = 175$, $k_2^{BAP} = 0.5$, and $k_1^{CNT} = 125$.

We note that our final approach has similarities with the winning entry to the 2017 EVA data challenge [28], where the authors combine data across locations with sufficient observations, in order to fit a generalised extreme value distribution for predicting precipitation extremes. They advocate the use of cross validation for tuning parameter selection and to compare potential modelling approaches in data challenges such as this, where the aim is to optimise some pre-determined metric.

4 Discussion of limitations and possible extensions

In this paper, we have discussed a marginal modelling approach for predicting wildfire events across the contiguous US. This framework was applied to obtain estimates of the cumulative distribution function at locations with missing entries for either CNT or BA. The resulting estimates were then "ranked" using a score function weighted to give higher importance to extreme observations [4]. Our method produced scores of 4080.559 and 3640.92 for CNT and BA, respectively, resulting in an overall score of 7721.479; this is a significant improvement on the proposed benchmark technique.

Unlike all the techniques introduced in Section 1.3, our approach does not attempt to specify the relationships between the auxiliary and wildfire variables. Such relationships appear to be complex and non-linear in nature, which may be explained by

a variety of hypotheses. For example, the monthly aggregated format of the wildfire variables arguably makes it more difficult to associate them with any climate covariates, which are given as monthly means. Instead, our approach relies on the assumption that wildfire observations within spatial neighbourhoods arise from the same marginal distribution. We believe that this is realistic since neighbouring locations are likely to have similar auxiliary covariates, as demonstrated in Fig. 3, and a large wildfire event occurring at one location is likely to increase the probability of wildfires in neighbouring locations. Furthermore, since the values in neighbourhood sets vary over time for each missing CNT or BA observation, our approach accounts for the temporal non-stationarity discussed in Section 1.2. We also propose several preliminary steps in Section 2; these steps do not require expert knowledge of wildfires to implement. Furthermore, such steps lead to significant improvements in the predictive distributions obtained using our approach by increasing the amount of information available and bringing all BA observations onto a unified scale.

While developing our approach, we investigated the possibility of accounting for covariate influence on extreme CNT values via the use of GAMs (this option was also mentioned in Section 1.3), but found their predictive performance to be poor in this setting; see [29] for details on these types of models. In particular, we fitted a continuous GPD for each month, with the scale parameter having a GAM form [30] comprising spatial and climatological covariates; a continuous approximation was used due to having discrete CNT values. The advantage of this method is that covariate effects can be directly assessed by examining the smooth functions underlying the models. From the fitted GAMs, the general spatial behaviour of the CNT data was modelled fairly well in each month, but these models did appear to suffer from oversmoothing, even when using models with the lowest level of smoothness. On the other hand, we found that physically-interpretable covariate behaviour for the climatological variables was hard to capture, making model selection difficult; the precise reason for this is unclear. It is likely that the aforementioned oversmoothing combined with other issues, such as poor convergence of the underlying numerical optimisation routines and difficulties combining the fitted GAMs with models for the bulk of the CNT values, lead to poor model performance against the benchmark. Therefore, it appears that this type of approach may not be favourable in situations where the prediction of unknown values is required, and is more suited to analyses where the aim is to account for uncertainty whilst modelling complex covariate effects. Indeed, in addition to the GAM-based approach mentioned in Section 1.3, [11, 31] and [32] have also successfully applied GAM techniques for modelling wildfire data.

One possible extension of the modelling techniques proposed in Sections 3.2 and 3.3 would be to introduce weights into the marginal estimation procedures. In the current format, observations within spatial neighbourhoods are given equal weights, even though it is likely that locations with a closer proximity to a missing observation would provide more useful information than locations that are further away. Our current method could therefore be extended by introducing weights to the marginal estimation procedures, with closer observations contributing more to probabilistic estimates. We would expect different values of the tuning parameters k_1^{CNT} and k_1^{BAP}, defining the spatial range of the neighbourhoods, to be appropriate in this case, but our cross validation approach could be used analogously.

Changes to the definition of the neighbourhoods in equation (1) could lead to improvements with our approach. One drawback with our current implementation is that the values k_1^{CNT} and k_1^{BAP} are chosen to be the same across all validation locations. Although we selected these tuning parameters carefully via cross validation, it is possible that this is an over-simplification and allowing the values to depend on covariates such as location, month or year may have been more appropriate. An extension of our approach could allow for this possibility, e.g., by separating the spatial domain into smaller sections and implementing cross validation separately in each one. It may also be reasonable to apply clustering algorithms as a preliminary step, to inform the spatial regions where setting the tuning parameters (k_1^{CNT}, k_1^{BAP}) as constant is a reasonable assumption. Allowing these parameters to vary across space also has the potential to provide insight into the behaviour of wildfires across the spatial domain. Additionally, we considered allowing the neighbourhoods themselves to depend on covariate-based clusters or to cover larger time windows, but the results presented in the Appendix suggest the simpler spatial neighbourhood approach was more successful.

While the zero-inflated negative binomial distribution proposed for CNT neighbourhoods is not motivated by extreme value theory, our analysis indicated the fitted marginal distributions performed reasonably well, including in the upper tail in the majority of cases. Several other distributions were tested, including fully empirical and discrete GPD models [33]; however, in every case, these distributions resulted in poorer prediction quality when ranked by the objective function given in [4]. This is likely due to the difficulties that arise in trying to capture behaviour in the bulk and tail simultaneously, and perhaps due to an insufficient amount of data in each of our spatial neighbourhoods for fitting the discrete GPD. In addition, alternative marginal distributions for BAP observations have the potential to further improve the predictive ability of our modelling framework.

Appendix: Spatio-temporal neighbourhoods

As alternatives to the spatial neighbourhoods $\mathcal{N}_i$, $i \in \{1, \ldots, N\}$, defined in Eq. (1), we also considered neighbourhoods of the form

$$\mathcal{N}_i^t := \big\{ j \in \{1, \ldots, N\} : \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \leq k_1^{CNT}, m_j = m_i, \\ y_j \in \{y_i - k_y^{CNT}, y_i - k_y^{CNT} + 1, \ldots, y_i + k_y^{CNT}\} \big\},$$

and

$$\mathcal{N}_i^c := \big\{ j \in \{1, \ldots, N\} : \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \leq k_1^{CNT}, m_j = m_i, y_j = y_i, c_j = c_i \big\},$$

for some $k_1^{CNT} \geq 0$ and $k_y^{CNT} \in \mathbb{N}$, where c_j denotes a covariate-based cluster assignment for each observation $j \in \mathcal{N}_i$. Analogous neighbourhoods were also considered for BAP.

With the observation month fixed and the spatial range defined as in $\mathcal{N}_i$, the neighbourhood $\mathcal{N}_i^t$ incorporates additional observations from neighbouring years, thus increasing the amount of data available for marginal estimation and adding a temporal element to the modelling procedure. On the other hand, the month and year are fixed for $\mathcal{N}_i^c$ so that only those data points with the same cluster assignment as observation i are considered, thus reducing the amount of information available for a fixed k_1^{CNT} or k_1^{BAP} value. However, assuming we can define clusters such that observations in the same cluster have more similar marginal tail properties, this additional step has the potential to improve marginal estimation.

Cluster assignments used within the $\mathcal{N}_i^c$ neighbourhoods were computed using divisive hierarchical clustering [34] for two different covariates: temperature and precipitation. We select these variables since they have been shown to be positively and negatively associated, respectively, to wildfire events [35, 36], and therefore may allow us to group together locations with similar marginal properties for CNT and BAP.

We use hierarchical clustering since this technique has been used in practice to approximate spatial clusters with similar wildfire properties [37–39]. The clustering procedure is as follows:

1. Standardise the auxiliary variable data (temperature or precipitation) for every location in $\mathcal{N}_i$.
2. Apply hierarchical clustering, using the standardised covariate data, to obtain two clusters, i.e., for each $j \in \mathcal{N}_i$, $c_j = 1$ or 2.
3. Compute the subset of locations with the same cluster assignment as location i, i.e., $\{j \in \mathcal{N}_i \mid c_j = c_i\}$.

In our analysis, we found that both the neighbourhoods $\mathcal{N}_i^t$ and $\mathcal{N}_i^c$ resulted in worse prediction scores compared to the simpler spatial neighbourhood approach outlined in Section 3. This is illustrated by the results in Tables 1 and 2, where we present the overall prediction scores, as outlined in [4], for $\mathcal{N}_i^t$ and $\mathcal{N}_i^c$, respectively; recall that we aim to minimise this score.

Table 1 Total scores obtained using neighbourhood $\mathcal{N}_i^t$

Distance	$k_y = 1$	$k_y = 2$	$k_y = 3$	$k_y = 4$	$k_y = 5$	$k_y = 6$
$k_1 = 50$	8393.3	8089.3	8003.6	7991.2	7925.9	7909.3
$k_1 = 75$	8225.5	8134.2	8161.9	8187.4	8162.6	8159.6
$k_1 = 100$	8295.6	8222	8264.4	8293.9	8275.5	8274.1
$k_1 = 125$	8439.9	8401.4	8457	8487.2	8474.1	8474
$k_1 = 150$	8547	8519	8578.2	8608.6	8599.5	8600.6
$k_1 = 175$	8624.8	8604.8	8664.5	8694.1	8687.5	8689
$k_1 = 200$	8711.2	8696.7	8755.2	8784.2	8778.4	8781.5
$k_1 = 225$	8779.8	8768.9	8823.8	8852.3	8848	8851.2
$k_1 = 250$	8843.9	8838.3	8892	8919.1	8915.4	8919.4

For these scores, we let $k_1^{CNT} = k_1^{BAP} := k_1 \in \{50, 75, \ldots, 250\}$ to incorporate a variety of spatial distances and set $k_2^{BAP} = 0.5$ to match the existing selected tuning parameter from Section 3.4. For $\mathcal{N}_i^t$, we let $k_y^{CNT} = k_y^{BAP} := k_y \in \{1, 2, 3, 4, 5, 6\}$, resulting in time windows of up to 13 years. Note that these scores correspond to the final prediction scores, i.e., when the missing data are known, and in practice, one would need to select the tuning parameters using the cross validation procedure outlined in Section 3.4.

One can observe that all the prediction scores from Tables 1 and 2 exceed the final score obtained using the method described in Section 3, and in many cases, the scores obtained were significantly worse. The fact that these predictions were worse for a wide range of tuning parameter combinations gives further support to our main modelling approach.

In the case of temporal neighbourhoods, since the predictive scores do not tend to decrease with the parameter k_y, our results suggest that marginal wildfire behaviour can vary significantly over neighbouring years.

Therefore, even though incorporating information from neighbouring years increases the amount of data available for model fitting, it does not appear to improve the quality of marginal estimates.

In the case of the cluster-based neighbourhoods, our results indicate that observations with similar temperature and precipitation values may not be those with similar wildfire behaviour. We suspect this may occur due to the complex nature of the relationships between the wildfire and auxiliary variables described in Section 4. Such relationships are unlikely to be picked up by incorporating this additional clustering step. Clustering also reduces the amount of data available for model fitting, which also appears to reduce the quality of marginal estimates.

On the whole, these results suggest that incorporating additional information, both from temporal windows and covariate-based clusters, does not improve the quality of marginal estimates for either CNT or BAP under our modelling approach. Combined with the principle of parsimony, we do not consider these alternative neighbourhoods further.

Table 2 Total scores obtained using neighbourhood $\mathcal{N}_i^c$ with clusters computed using temperature and precipitation

Distance	Temperature clusters	Precipitation clusters
$k_1 = 50$	11149	11161
$k_1 = 75$	9608.8	9607.7
$k_1 = 100$	9245.2	9309.2
$k_1 = 125$	8651.9	8718.5
$k_1 = 150$	8522.9	8582.1
$k_1 = 175$	8431.5	8505.2
$k_1 = 200$	8460.3	8489.5
$k_1 = 225$	8470.2	8511.1
$k_1 = 250$	8493.6	8539.5

Acknowledgements We are grateful to the referees and guest editor for constructive comments that have greatly improved the paper.

Funding This paper is based on work completed while Callum Murphy-Barltrop and Eleanor D'Arcy were part of the EPSRC funded STOR-i centre for doctoral training (EP/L015692/1 and EP/S022252/1, respectively). Emma Simpson was partly supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. OSR-2017-CRG6-3434.02.

Data availibility The data set analysed during the current study is available from the corresponding author on reasonable request.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

1. Wong, S.D., Broader, J.C., Shaheen, S.A.: Review of California wildfire evacuations from 2017 to 2019. University of California Institute of Transportation Studies, Technical report, UC Office of the President (2020)
2. Jones, M.W., Smith, A., Betts, R., Canadell, J.G., Prentice, I.C., Le Quéré, C.: Climate change increases risk of wildfires. ScienceBrief Review (2020)
3. Zhuang, Y., Fu, R., Santer, B.D., Dickinson, R.E., Hall, A.: Quantifying contributions of natural variability and anthropogenic forcings on increased fire weather risk over the western United States. Proc. Natl. Acad. Sci. **118**(45), 2111875118 (2021). https://doi.org/10.1073/pnas.2111875118
4. Opitz, T.: Editorial: EVA 2021 Data Competition on spatio-temporal prediction of wildfire activity in the United States. Extremes (to appear) (2022)
5. Coles, S.G., Heffernan, J.E., Tawn, J.A.: Dependence measures for extreme value analyses. Extremes **2**(4), 339–365 (1999)
6. Preisler, H.K., Brillinger, D.R., Burgan, R.E., Benoit, J.: Probability based models for estimation of wildfire risk. Int. J. Wildland Fire **13**(2), 133–142 (2004)
7. Wuebbles, D.J., Fahey, D.W., Hibbard, K.A., Arnold, J.R., DeAngelo, B., Doherty, S., Easterling, D.R., Edmonds, J., Edmonds, T., Hall, T., et al.: Climate science special report: Fourth national climate assessment, volume I. Technical report, U.S. Global Change Research Program, Washington, DC, USA (2017). https://doi.org/10.7930/J0J964J6
8. Holden, Z.A., Swanson, A., Luce, C.H., Jolly, W.M., Maneta, M., Oyler, J.W., Warren, D.A., Parsons, R., Affleck, D.: Decreasing fire season precipitation increased recent western US forest wildfire activity. Proc. Natl. Acad. Sci. **115**(36), 8349–8357 (2018)
9. Son, R., Kim, H., Wang, S.-Y., Jeong, J.-H., Woo, S.-H., Jeong, J.-Y., Lee, B.-D., Kim, S.H., LaPlante, M., Kwon, C.-G., Yoon, J.-H.: Changes in fire weather climatology under 1.5 °C and 2.0 °C warming. Environ. Res. Lett. **16**(3), 034058 (2021)
10. Krawchuk, M.A., Moritz, M.A., Parisien, M.-A., Van Dorn, J., Hayhoe, K.: Global pyrogeography: the current and future distribution of wildfire. PloS one **4**(4), 5102 (2009)

11. Sá, A.C.L., Turkman, M.A.A., Pereira, J.M.C.: Exploring fire incidence in Portugal using generalized additive models for location, scale and shape (gamlss). Model. Earth Syst. Environ. **4**(1), 199–220 (2018)
12. Ziel, R.H., Bieniek, P.A., Bhatt, U.S., Strader, H., Rupp, T.S., York, A.: A comparison of fire weather indices with MODIS fire days for the natural regions of Alaska. Forests **11**(5), 516 (2020)
13. Cohen, J.D., Deeming, J.E.: The National Fire-Danger Rating System: basic Equations (1985). https://www.fs.usda.gov/research/treesearch/27298
14. Koh, J., Pimont, F., Dupuy, J.-L., Opitz, T.: Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. The Annals of Applied Statistics **17**(1), 560–582 (2023). https://doi.org/10.1214/22-AOAS1642
15. Sharples, J.J., McRae, R.H.D., Weber, R.O., Gill, A.M.: A simple index for assessing fire danger rating. Environ. Model. Softw. **24**(6), 764–774 (2009)
16. Richards, J., Huser, R.: A unifying partially-interpretable framework for neural network-based extreme quantile regression. arXiv preprint: 2208.07581 (2022)
17. Ivek, T., Vlah, D.: Reconstruction of incomplete wildfire data using deep generative models. Extremes (2023). https://doi.org/10.1007/s10687-022-00459-1
18. Cisneros, D., Gong, Y., Yadav, R., Hazra, A., Huser, R.: A combined statistical and machine learning approach for spatial prediction of extreme wildfire frequencies and sizes. Extremes (2023). https://doi.org/10.1007/s10687-022-00460-8
19. Koh, J.: Gradient boosting with extreme-value theory for wildfire prediction. Extremes (2023). https://doi.org/10.1007/s10687-022-00454-6
20. Kelly, K., Šavrič, B.: Area and volume computation of longitude-latitude grids and three-dimensional meshes. Trans. GIS **25**(1), 6–24 (2021)
21. Budic, L., Didenko, G., Dormann, C.F.: Squares of different sizes: effect of geographical projection on model parameter estimates in species distribution modeling. Ecol. Evol. **6**(1), 202–211 (2016)
22. Hijmans, R.J.: Geosphere: Spherical Trigonometry. (2019). R package version 1.5-10. https://CRAN.R-project.org/package=geosphere
23. Joseph, M.B., Rossi, M.W., Mietkiewicz, N.P., Mahood, A.L., Cattau, M.E., St. Denis, L.A., Nagy, R.C., Iglesias, V., Abatzoglou, J.T., Balch, J.K.: Spatiotemporal prediction of wildfire size extremes with Bayesian finite sample maxima. Ecol. Appl. **29**(6), 01898 (2019)
24. Richards, J., Tawn, J.A., Brown, S.: Modelling extremes of spatial aggregates of precipitation using conditional methods. Ann. Appl. Stat. **16**(4), 2693–2713 (2022)
25. Balkema, A.A., de Haan, L.: Residual life time at great age. Ann. Probab. **2**(5), 792–804 (1974). https://doi.org/10.1214/aop/1176996548
26. Pickands, J.: Statistical inference using extreme order statistics. Ann. Stat. **3**(1), 119–131 (1975)
27. Coles, S.G.: An Introduction to Statistical Modeling of Extreme Values. Springer, London (2001)
28. Stephenson, A.G., Saunders, K., Tafakori, L.: The MELBS team winning entry for the EVA2017 competition for spatiotemporal prediction of extreme rainfall using generalized extreme value quantiles. Extremes **21**, 477–484 (2018)
29. Wood, S.N.: Generalized Additive Models: An Introduction with R. CRC Press, Boca Raton (2017)
30. Youngman, B.D.: Generalized additive models for exceedances of high thresholds with an application to return level estimation for U.S. wind gusts. J. Am. Stat. Assoc. **114**(528), 1865–1879 (2019)
31. Zhang, Y., Lim, S., Sharples, J.J.: Wildfire occurrence patterns in ecoregions of New South Wales and Australian Capital Territory, Australia. Nat. Hazards **87**, 415–435 (2017)
32. Rodríguez-Pérez, J.R., Ordóñez, C., Roca-Pardiñas, J., Vecín-Arias, D., Castedo-Dorado, F.: Evaluating lightning-caused fire occurrence using spatial generalized additive models: a case study in central Spain. Risk Anal. **40**(7), 1418–1437 (2020)
33. Hitz, A.S., Davis, R.A., Samorodnitsky, G.: Discrete extremes. arXiv pre-print **1707**, 05033 (2017)
34. Rokach, L., Maimon, O.: Clustering methods. In: Data Mining and Knowledge Discovery Handbook, pp. 321–352. Springer, New York (2005)
35. Duane, A., Castellnou, M., Brotons, L.: Towards a comprehensive look at global drivers of novel extreme wildfire events. Climatic Change **165**, 43 (2021)
36. Crockett, J.L., Westerling, A.L.: Greater temperature and precipitation extremes intensify Western U.S. droughts, wildfire severity, and Sierra Nevada tree mortality. J. Clim. **31**(1), 341–354 (2018)
37. Rodrigues, M., Costafreda-Aumedes, S., Comas, C., Vega-García, C.: Spatial stratification of wildfire drivers towards enhanced definition of large-fire regime zoning and fire seasons. Sci. Total Environ. **689**, 634–644 (2019)

38. Rodrigues, M., González-Hidalgo, J.C., Peña-Angulo, D., Jiménez-Ruano, A.: Identifying wildfire-prone atmospheric circulation weather types on mainland Spain. Agric. For. Meteorol. **264**, 92–103 (2019)
39. Rahimi, S., Sharifi, Z., Mastrolonardo, G.: Comparative study of the effects of wildfire and cultivation on topsoil properties in the Zagros forest, Iran. Eurasian Soil Sci. **53**, 1655–1668 (2020)

Authors and Affiliations

Eleanor D'Arcy[1] · Callum J. R. Murphy-Barltrop[1] · Rob Shooter[2] · Emma S. Simpson[3]

Eleanor D'Arcy
e.darcy@lancaster.ac.uk

Rob Shooter
robert.shooter@metoffice.gov.uk

Emma S. Simpson
emma.simpson@ucl.ac.uk

[1] STOR-i Centre for Doctoral Training, Lancaster University, Fylde Avenue, Lancaster LA1 4YR, United Kingdom

[2] Department, Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, United Kingdom

[3] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, United Kingdom