Extremes (2023) 26:241–250
https://doi.org/10.1007/s10687-023-00465-x

Editorial: EVA 2021 data challenge on spatiotemporal prediction of wildfire extremes in the USA

Thomas Opitz[1]

Published online: 27 March 2023

1 Wildfires and their extremes: a challenging problem

The papers in this Special Issue describe the contributions of teams participating in a data competition that took place prior to the 12th Extreme-Value Analysis (EVA) conference. It started in December 2020 with final predictions of participants due in May 2021. During a dedicated session of the EVA conference held from June 28th to July 2nd of 2021 and taking place online because of the difficult and uncertain sanitary situation due to Covid, the results of the challenge were presented and prizes were awarded to the top teams.

Wildfires are defined as uncontrolled fires of combustible material composed of natural vegetation, such as forests or shrubland. They represent an environmental hazard with major impacts worldwide, and their frequency of occurrence and size is expected to further increase with global warming (Jones et al. 2020). Each year, wildfires cause many direct human casualties, and they are at the origin of extreme air pollution episodes and the loss of biodiversity and other ecosystem services. They further contribute an important fraction of global greenhouse gas emissions each year, such that they could further exacerbate climate change. Wildfire modeling has been tackled with a wide variety of approaches from statistics and machine learning [see, *e.g.*, Preisler et al. (2004), Pereira and Turkman (2019), and Xi et al. (2019)], some of them invoking extreme-value theory [see, *e.g.*, Pereira and Turkman (2019), and Koh et al. (2023)].

As to the origin of wildfire ignitions, lightning is the principal natural cause, but in the majority of cases human activities are responsible. They may be intentional (arson) or accidental (debris burning, agricultural activities, campfires, smoking). Typically, wildfires are the result of the concurrence of the presence of combustible material (*e.g.* forest), its easy flammability (*e.g.* resulting from extreme weather conditions such as droughts), and a trigger (*e.g.* lightning or human activity). In most wildfire-prone areas over the globe, wildfire activity shows seasonal cycles

✉ Thomas Opitz
thomas.opitz@inrae.fr

[1] Biostatistics and Spatial Processes, INRAE, 228 route de l'Aérodrome, Avignon 84914, France

due to the seasonality of favorable weather conditions modifying the vulnerability of forests.

To aid in wildfire management, it is crucial to understand and predict how various risk factors interact in contributing to wildfire behavior and its spatio-temporal variation. Wildfire management includes a multitude of tasks, including monitoring of forest ecosystems, deployment of preventive measures, firefighting logistics, as well as short-term forecasting and long-term projections of wildfire activity.

The data challenge focused on two important components of wildfire activity: wildfire occurrence, and wildfire size. Given a region of space and a period of time (*e.g.* a voxel in a regular spatiotemporal grid), we consider the number of spatially separated (*i.e.*, spatially and temporally non contiguous) wildfire events as an observation with respect to the first aspect (occurrence), and the aggregated burnt area of wildfires originating in the area of interest as an observation with respect to the second aspect (size). In light of the non-Gaussian nature and the relatively heavy tails in observations of both of these variables, and of the lack of strong autocorrelation of such variables across space and time, their accurate prediction is a challenging task. The most extreme impacts and the biggest difficulties in wildfire management are associated to large values of burnt area and/or of wildfire counts, for both of which prediction is highly challenging. Therefore, extreme-value theory is a promising conceptual and methodological framework for studying such data.

2 Description and preprocessing of original data

The dataset provided for the competition has been composed using several data sources. Here we provide some background. Data are provided for the continental United States, excluding the state of Alaska and islands such as Hawaiï. Spatial coordinates are given in the WGS84 system, that is, the usual longitude and latitude coordinates. All data are aggregated to a monthly $0.5^o \times 0.5^o$ grid of longitude and latitude coordinates (roughly 55 by 55 km) covering the study area, and we explain in the following how this has been achieved. Only the months between March and September are considered for wildfire data.

2.1 Original datasets

2.1.1 US wildfires

We use a comprehensive dataset of wildfires covering the period from 1993 to 2015 for the continental United States, which has been gathered from various wildfire inventories in Short (2017). It reports a set of unified attributes available for each wildfire. In this data competition, we specifically use the geographic position, the time of occurrence and the burnt area of individual wildfires. This information is aggregated towards a monthly longitude-latitude grid at 0.5-degree resolution (roughly 55 km) by counting the number of wildfires within each grid point (variable CNT), and by summing up the burnt areas of these wildfires (variable BA).

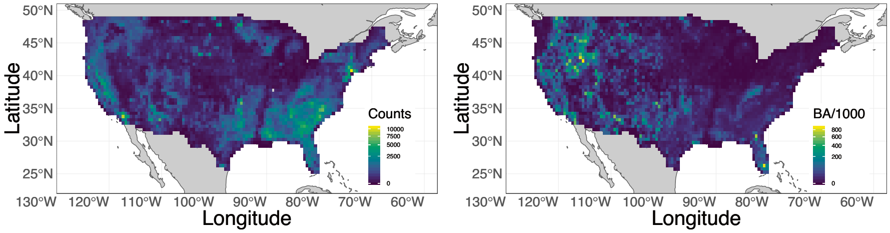

Fig. 1 Wildfire data for the US, aggregated over the study period for each pixel. Left: Counts. Right: Burnt areas

Figure 1 shows spatial maps of these two variables where they have been further aggregated over all months of the study period. The heavy-tailed nature of the BA variable becomes obvious from the histograms in Fig. 2 of positive BA values (left histogram) and logarithms of positive BA values (right histogram).

2.1.2 Land cover

The land monitoring service of the European Union's COPERNICUS service for remote sensing produces global land cover classification maps at 300 m spatial resolution and annual temporal resolution. Data are free of access and are provided online through the COPERNICUS Climate Data Store. The classification uses 38 classes whose definition is based on the United Nations Food and Agriculture Organization's (UN FAO) Land Cover Classification System (LCCS). Of these 38 categories, only 18 are observed with nonnegligible proportion in the study area defined by the contiguous US. For illustration, a map with classes of land cover in the US is shown in Fig. 3. For the data competition, data are aggregated to the $0.5^o \times 0.5^o$ grid of longitude and latitude by considering the proportion of each of

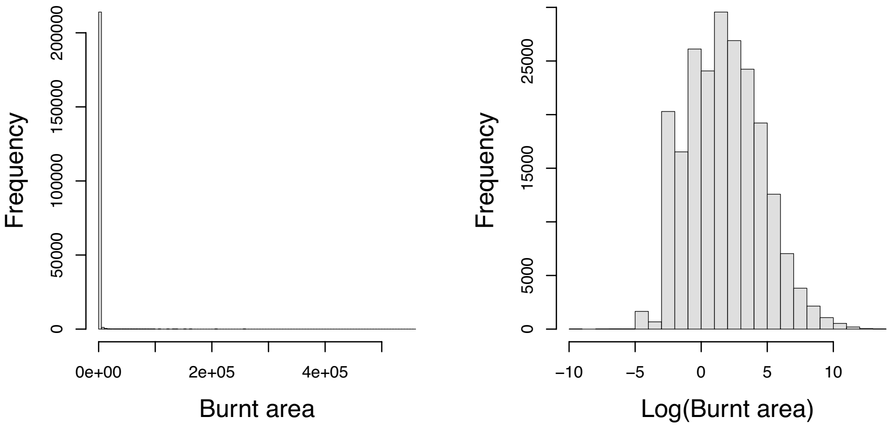

Fig. 2 Histogram of positive burnt areas (left) and of logarithms of positive burnt areas (right)

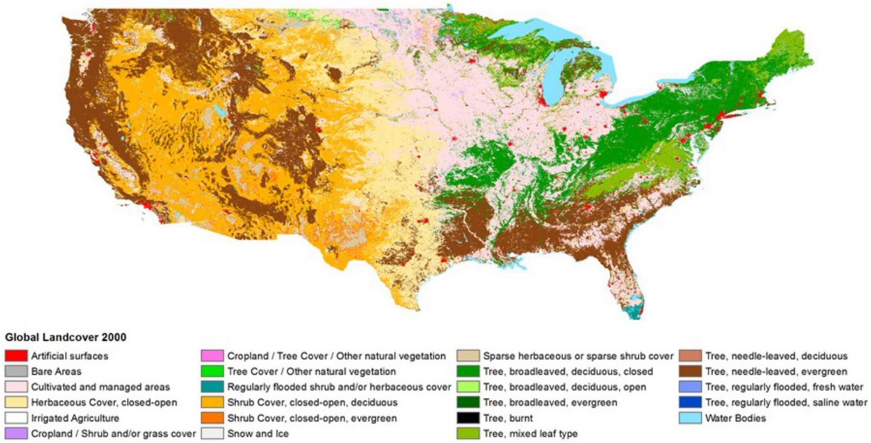
Global Landcover 2000
Artificial surfaces
Bare Areas
Cultivated and managed areas
Herbaceous Cover, closed-open
Irrigated Agriculture
Cropland / Shrub and/or grass cover
Cropland / Tree Cover / Other natural vegetation
Tree Cover / Other natural vegetation
Regularly flooded shrub and/or herbaceous cover
Shrub Cover, closed-open, deciduous
Shrub Cover, closed-open, evergreen
Snow and Ice
Sparse herbaceous or sparse shrub cover
Tree, broadleaved, deciduous, closed
Tree, broadleaved, deciduous, open
Tree, broadleaved, evergreen
Tree, burnt
Tree, mixed leaf type
Tree, needle-leaved, deciduous
Tree, needle-leaved, evergreen
Tree, regularly flooded, fresh water
Tree, regularly flooded, saline water
Water Bodies

Fig. 3 Map of the land cover classes for the US

these 18 categories within each grid cell. Proportions of the following categories are considered and are named lc_1 to lc_{18} in the dataset provided for the competition. Their denominations as given in the original dataset are as follows: 1) cropland rainfed; 2) cropland rainfed herbaceous cover; 3) mosaic cropland; 4) mosaic natural vegetation; 5) tree broadleaved deciduous closed to open; 6) tree broadleaved deciduous closed; 7) tree needleleave evergreen closed to open; 8) tree needleleave evergreen closed; 9) tree mixed; 10) mosaic tree and shrub; 11) shrubland; 12) grassland; 13) sparse vegetation; 14) tree cover flooded fresh or brakish water; 15) shrub or herbaceous cover flooded; 16) urban; 17) bare areas; 18) water.

The area proportions lc_1 to lc_{18} do not always sum to exactly 1 for each pixel and month since a few classes with quasi-0 proportion have been removed. These 18 predictors are therefore almost *collinear*, *i.e.*, $\mathrm{lc}_{i_0} \approx 1 - \sum_{i=1,\, i\neq i_0}^{18} \mathrm{lc}_i$ for $i_0 = 1, \dots, 18$.

2.1.3 Meteorological variables

The meteorological variables provided in the dataset of the competition are based on the gridded output of monthly means obtained within the ERA5-reanalysis on Land surface, available for a global grid of resolution $0.1^o \times 0.1^o$ and downloadable from the COPERNICUS Climate Data Store. The following 10 variables, called clim_1 to clim_{10}, respectively, in the provided data, are considered for this data competition, with their units given in parentheses: 1) 10 m U-component of wind (the wind speed in Eastern direction) (m/s); 2) 10 m V-component of wind (the wind speed in Northern direction) (m/s); 3) Dewpoint temperature (temperature at 2 m from ground to which air must be cooled to become saturated with water vapor, such that condensation ensues) (Kelvin); 4) Temperature (at 2 m from ground) (Kelvin); 5) Potential evaporation (the amount of evaporation of water that would take place if a sufficient source of water were available) (m) 6) Surface net solar radiation (net flux of shortwave radiation; mostly radiation coming from the sun) (J/m^2); 7) Surface net thermal radiation (net flux of longwave radiation; mostly radiation emitted by the surface) (J/m^2); 8) Surface pressure (Pa); 9) Evaporation (of water) (m); 10) Precipitation (m).

2.1.4 Variables related to altitude

Finally, two variables related to altitude are made available. The variable called `altiMean` provides the mean altitude for each cell of the longitude-latitude grid, and the variable called `altiSD` provides the corresponding standard deviation. Original gridded data were provided by the Shuttle Radar Topography Mission (SRTM) at 90 m spatial resolution.

2.2 Split into training and validation datasets

The dataset has been split into a training dataset and a validation dataset. Validation is carried out based on the predictions for the variables of aggregated burnt areas (BA) and counts (CNT). Only training data were provided to the teams participating in the data challenge. No data have been masked for uneven years (1993, 1995,...,2015). For even years (1994, 1996,..., 2014), overall 80,000 observations of each of the two variables are kept for validation; that is, they have been masked by setting them to a missing value flag in the dataset. The spatial and temporal positions of validation data are not completely random, but they tend to be clustered in space and time. Moreover, the validation locations for BA and CNT are not the same, but they are positively correlated. This has been achieved by defining masked locations as the exceedance locations of a bivariate space-time Gaussian process with positive spatiotemporal correlation and positive cross-correlation. Therefore, the probability of having to predict both BA and CNT for a given grid cell and month is higher than the product of the two probabilities of having to validate BA or CNT. The training data values and validation data locations for a given month of the study period are shown in Fig. 4.

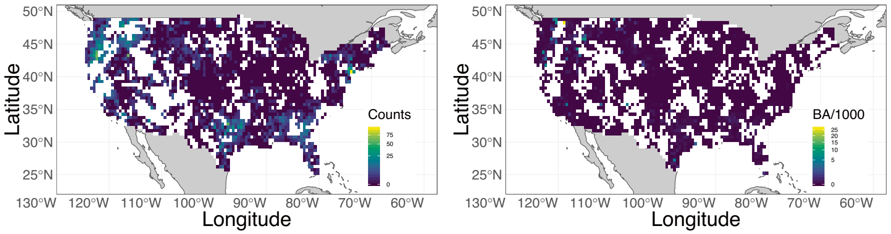

Fig. 4 Example of training data for a given month. Left: Counts. Right: Burnt areas

3 Prediction goal and benchmark model

3.1 Prediction goal

The aim of this challenge was to estimate predictive distributions for the BA and CNT validation data, *i.e.*, for the data masked in the training dataset. More precisely, the value of the predictive distribution function of each masked observation of BA and CNT must be estimated for a list of severity thresholds. The list of thresholds was chosen to put relatively strong focus on exceedances of extreme levels.

For CNT, 28 severity thresholds are fixed as follows:

$$\mathcal{U}_{CNT} = \{u_1, u_2, \dots, u_{28}\} = \{0, 1, 2, \dots, 9, 10, 12, 14, \dots, 30, 40, 50, \dots, 100\}.$$

For each CNT validation data point i and each $u \in \mathcal{U}_{CNT}$, the probability $p_{CNT,i}(u) = \mathbb{P}(\mathrm{CNT}_i \leq u)$ must be estimated. For BA, the following 28 severity thresholds are used:

$$\begin{aligned}\mathcal{U}_{BA} = \{u_1, u_2, \dots, u_{28}\} = \{&0, 1, 10, 20, 30, \dots, 100, 150, 200, 250, 300, 400, 500,\\ &1000, 1500, 2000, 5000, 10000, 20000, 30000, 40000, 50000, 100000\}.\end{aligned}$$

For each BA validation data point i and each $u \in \mathcal{U}_{BA}$, the probability $p_{BA,i}(u) = \mathbb{P}(\mathrm{BA}_i \leq u)$ must be estimated.

3.2 Prediction scores

The performance of predictions of the different teams was compared using a prediction score, and teams were ranked based on the prediction score, with lower scores resulting in better rank. Separate prediction scores were calculated for BA and CNT, resulting in separate rankings of teams participating in one or both of the sub-competitions of predicting BA and CNT, respectively. The overall prediction score, used to determine the overall ranking and the winning teams, resulted from adding up these two separate scores and ranking them. A relatively lower score is better and leads to a better ranking of the corresponding team.

The scores used for the competition are variants of weighted rank probability scores, which put relatively strong weight on good prediction in the extremes of the distribution of counts and burnt areas. We denote by $\hat{p}_{CNT,i}(u)$ the predicted probability for $\mathbb{P}(\mathrm{CNT}_i \leq u)$ and by $\hat{p}_{BA,i}(u)$ the predicted probability for $\mathbb{P}(\mathrm{BA}_i \leq u)$. There were $k_{CNT} = k_{BA} = 80,000$ values to be predicted for CNT and BA, respectively. For the counts in CNT to be predicted, the score

$$S_{CNT} = \sum_{i=1}^{k_{CNT}} \sum_{u \in \mathcal{U}_{CNT}} \omega_{CNT}(u)\big(\mathbb{I}(u \geq \mathrm{CNT}_i) - \hat{p}_{CNT,i}(u)\big)^2$$

was used, where the weight function is given as

$$\omega_{CNT}(u) = \frac{\tilde{\omega}_{CNT}(u)}{\tilde{\omega}_{CNT}(u_{28})}, \quad \tilde{\omega}_{CNT}(u) = 1 - (1 + (u+1)^2/1000)^{-1/4},$$

and $\mathbb{I}$ is the indicator function defined as

$$\mathbb{I}(u \leq x) = \begin{cases} 1, & \text{if } u \leq x \\ 0, & \text{if } u > x. \end{cases}$$

The division by $\tilde{\omega}_{CNT}(u_{28})$ above ensures that the largest weight $\omega_{CNT}(u_{28})$ is 1. By analogy, the score for burnt areas in BA is

$$S_{CNT} = \sum_{i=1}^{k_{BA}} \sum_{u \in \mathcal{U}_{BA}} \omega_{BA}(u)\big(\mathbb{I}((u \geq \mathrm{BA}_i) - \hat{p}_{BA,i}(u)\big)^2$$

where

$$\omega_{BA}(u) = \frac{\tilde{\omega}_{BA}(u)}{\tilde{\omega}_{BA}(u_{28})}, \quad \tilde{\omega}_{BA}(u) = 1 - (1 + (u+1)/1000)^{-1/4}.$$

The overall score was then given as $S_{TOTAL} = S_{CNT} + S_{BA}$. The weights were determined through preliminary analyses to ensure that the contributions of the two scores to the overall score were approximately of the same order. Each participating team had to provide two numerical matrices of predictions for counts and burnt areas, each of dimension $80,000 \times 28$, where the (i, j)-th entry corresponds to the prediction $\hat{p}_{BA,i}(u_j)$ or $\hat{p}_{CNT,i}(u_j)$, respectively.

3.3 Benchmark model

Moreover, a benchmark score was provided. For CNT, it was obtained by estimating a generalized linear model with Poisson response distribution and log-link using all available covariates, and the exceedance probabilities were calculated from the predictive distributions of the model. For positive values of BA, the benchmark score was obtained by estimating a generalized linear model with log-Gaussian response and log-link function using all available covariates. The probability predictions of BA are obtained by multiplying the predictions of exceedance probabilities for the log-Gaussian BA model (estimated using only the observations with BA > 0) with the predicted probability of CNT > 0 obtained from the Poisson model. Note that this benchmark model does not exploit the information that BA $= 0$ if the observed CNT $= 0$, but the teams were allowed to use this information.

4 Ranking of the teams participating in the competition

More than 60 participants from around 20 countries worked on the prediction task of the competition, and 13 teams submitted final predictions. These teams were invited to submit a paper for the present Special Issue of the Extremes journal to describe their approach, and finally seven papers were submitted and have undergone the usual peer-review process. These seven papers are presented in this issue of

Table 1 Final ranking of the participating teams that accepted publication of their results. The rightmost column indicates the author reference if there is a paper describing the team's approach in this Special Issue

Team name	S_{CNT}	rank$_{CNT}$	S_{BA}	rank$_{BA}$	S_{total}	rank$_{total}$	SI paper
BlackBox	2805	1	3316	1	6121	1	Ivek & Vlah
Kohrrelation	2990	3	3446	3	6436	2	Koh
Bedouins	3146	4	3408	2	6554	3	Hazra et al.
KUNGFUPANDA	3166	5	3513	5	6679	4	Makowski
RedSea	3419	7	3467	4	6886	5	Zhang et al.
NaiveTom	3403	6	3565	7	6968	6	–
THEFIRETASTICFOUR	2979	2	4133	11	7111	7	–
EdX	3520	8	3595	8	7115	8	–
SNUBRL	4074	9	3530	6	7604	9	Kim et al.
MayLaB	4418	11	3765	10	8182	10	–
LancasterDucks	4926	12	3719	9	8645	11	D'Arcy et al.
FUFighters	4329	10	4863	13	9191	12	–
BENCHMARK	5565	13	4244	12	9810	13	–

Extremes and bear witness of the large variety of statistical approaches devised and implemented by the teams.

Each team could submit up to two preliminary predictions to compare their approach to the predictions of other teams. The final ranking was based solely on the final prediction. Final rankings were published during the EVA 2021 conference and are reported in Table 1 for 12 teams. Note that teams could choose not to appear in the published rankings. All ranked teams were successful in substantially outperforming the benchmark model through their approaches. The best team BlackBox managed to improve on the benchmark score by almost 40%, and its performance is particularly impressive since it also achieved the highest absolute improvement over the next lower-ranking team among the 8 top-ranked teams.

5 Discussion of implemented approaches and results

For details on the various implemented approaches, we refer the reader to the corresponding papers in this Special Issue. The top-ranked team BlackBox (Ivek & Vlah) implemented a deep learning approach with adaptations to take into account the heavy-tailed distributions with singular mass at 0 and the incomplete spatiotemporal coverage of observations due to masked data. The second-ranked team Kohrrelation (Koh) developed a gradient-boosting algorithm with the choice of the loss function guided by the extreme-value setting. The third-ranked team Bedouins (Hazra et al.) developed a multi-stage modeling approach, including Bayesian hierarchical models and Random Forests as its components, and the team ranked fifth, RedSea (Zhang et al.), also resorted to hierarchical Bayesian modeling. The use of a classical Random Forest algorithm, without specific extreme-value features but appropriate for high-dimensional datasets, ensured the fourth place for team KUNGFUPANDA (Makowski). Finally, the SNUBRL team of Kim et al. used spatial quantile autoregressive models, and team LancasterDucks (D'Arcy et al.) first grouped similar locations using clustering algorithms before fitting distributions motivated by extreme-value theory with different parameter values for the different clusters.

In conclusion, some general observations can be made based on the analysis of the implemented approaches and the corresponding prediction performances:

1. Assessing and comparing predictions for heavy-tailed phenomena remains challenging. This is particularly true if data do not provide enough information to construct much lighter-tailed predictive distributions, such that predictions remain heavy-tailed. Good and simple tools to score predictions in this framework still seem to be lacking, but recent results in the literature hint at potential solutions invoking extreme-value theory [e.g., Taillardat et al. (2022)].
2. State-of-the-art techniques for machine learning and artificial intelligence can be successfully adapted to improve extreme-value predictions thanks to appropriate choices of cost functions to be minimized, i.e., by making them aware of the specific data structures induced by observations of extreme events (e.g. generalized Pareto distributions, censoring techniques).
3. Another important aspect of successful prediction of environmental extremes is efficient use of spatiotemporal dependencies, either explicitly by defining stochastic processes with spatiotemporal autocorrelation, or implicitly by appropriate choice of spatiotemporal data features used for prediction when training machine learning models.
4. The prediction setting of this competition was challenging due to heterogeneous data structures (e.g. discrete and continuous) and the requirement to predict over the whole range of values of the data, and not only in the tail. Therefore, direct use of standard extreme-value limit processes (max-stable processes, r-Pareto processes) was not implemented by any team, and such models would certainly have been too unwieldy to yield competitive predictions without major modeling extensions. Nevertheless, most approaches have incorporated concepts and tools from extreme-value theory to improve the prediction performance, especially in the tail.

Acknowledgements I would like to warmly acknowledge a series of preliminary discussions about the content of the data challenge with members of the scientific and organisation committees of EVA 2021. I am particularly grateful to Ioannis Papastathopoulos (University of Edinburgh) and Raphaël Huser (KAUST), who were co-organisers of this challenge and provided precious advice and logistic support for its organisation. I also address a big thank-you to the Editor-in-Chief of Extremes, Thomas Mikosch, for making this Special Issue possible and for all his guidance during the review process for the published papers.

Finally, I would like to point out that organizing this challenge was a real pleasure and a very rewarding experience for me, and I am already looking forward to the next one at EVA 2023, maybe going back to the role of a simple participant of one of the competing teams.

References

Jones, M.W., Smith, A., Betts, R., et al.: Climate change increases risk of wildfires. Sci. J. Rev. (2020)

Koh, J., Pimont, F., Dupuy, J.L., et al.: Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. Ann. Appl. Stat. **17**, 560–582 (2023)

Pereira, J., Turkman, K.: Statistical models of vegetation fires: spatial and temporal patterns. In: Handbook of Environmental and Ecological Statistics, pp. 401–420. Chapman and Hall/CRC (2019)

Preisler, H.K., Brillinger, D.R., Burgan, R.E., et al.: Probability based models for estimation of wildfire risk. Int. J. Wildland Fire **13**(2), 133–142 (2004)

Short, K.C.: Spatial wildfire occurrence data for the United States, 1992–2015. Tech. rep., Forest Service Research Data Archive, Fort Collins, CO. (2017). https://doi.org/10.2737/RDS-2013-0009.4

Taillardat, M., Fougères, A.L., Naveau, P., et al.: Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. Int. J. Forecast. (2022)

Xi, D.D.Z., Taylor, S.W., Woolford, D.G., et al.: Statistical models of key components of wildfire risk. Annu. Rev. Stat. Appl. **6**, 197–222 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.