Joint modeling and prediction of massive spatio-temporal wildfire count and burnt area data with the INLA-SPDE approach

Zhongwei Zhang1,2 · Elias Krainski2 · Peng Zhong2 · Harvard Rue2 · Raphaël Huser2

Received: 15 February 2022 / Revised: 10 October 2022 / Accepted: 16 January 2023 / Published online: 14 March 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
This paper describes the methodology used by the team RedSea in the data competition organized for EVA 2021 conference. We develop a novel two-part model to jointly describe the wildfire count data and burnt area data provided by the competition organizers with covariates. Our proposed methodology relies on the integrated nested Laplace approximation combined with the stochastic partial differential equation (INLA-SPDE) approach. In the first part, a binary non-stationary spatio-temporal model is used to describe the underlying process that determines whether or not there is wildfire at a specific time and location. In the second part, we consider a non-stationary hurdle log-Gaussian Cox process (hurdle-LGCP) for the positive wildfire count data, i.e., an LGCP is used to model the shifted positive count data, and a non-stationary log-Gaussian model for positive burnt area data. Dependence between the positive count data and positive burnt area data is captured by a shared spatio-temporal random effect. Our two-part modeling approach performs well in terms of the prediction score criterion chosen by the data competition organizers. Moreover, our model results show that surface pressure is the most influential driver for the occurrence of a wildfire, whilst surface net solar radiation and surface pressure are the key drivers for large numbers of wildfires, and temperature and evaporation are the key drivers of large burnt areas.

Keywords INLA-SPDE · Marked point process model · Multivariate processes · Non-stationarity · Spatio-temporal model · Wildfire modeling

AMS 2000 Subject Classifications 60G70 · 62M30 · 62P12

Zhongwei Zhang
zhongwei.zhang@unige.ch

Extended author information available on the last page of the article
1 Introduction

Wildfires have significant social and economic impact, and might pose significant threat to infrastructure, human safety, and natural resources (Rosenthal et al. 2021; Burke et al. 2021). Furthermore, wildfires are an important source of CO$_2$ emissions and contribute substantially to the global greenhouse effect (Liu et al. 2014). Over the past four decades, the wildfire burnt area has roughly quadrupled in the United States (US), which has led to substantial increases in US government expenditures on wildfire suppression in recent years (Burke et al. 2021). There is thus a pressing need to develop flexible statistical models for wildfire activity and to improve our understanding of wildfire risks so as to support fire management decision-making.

Wildfire risks consist of various components, such as fire occurrence, fire intensity and growth, fire duration, and fire size. Statistical science has played a key role in modeling and prediction of these components; see Taylor et al. (2013) and Xi et al. (2019) for an overview. In the data competition of the Extreme Value Analysis (EVA) 2021 conference, the main goal is modeling and prediction of aggregated monthly numbers of wildfire occurrences and their burnt areas in each cell of a regular grid covering the continental US.

Log-Gaussian Cox processes, which are Poisson point processes with log-intensity specified by a Gaussian random field, have been identified as useful models for wildfire occurrences (Serra et al. 2014; Opitz et al. 2020). As for the burnt area, which represents the size of wildfires, a variety of different models have been considered in the literature, including (truncated) power-law distributions (Cumming 2001; Butry et al. 2008), the Weibull distribution (Reed and McKelvey 2002), or the log-normal distribution (Hantson et al. 2016), for positive burnt area data, as well as the generalized Pareto distribution for large wildfires only (Holmes et al. 2008). The number of wildfires and their burnt areas are clearly linked since if one of them is zero, the other one must also be zero, and a large number of wildfires might correspond to large burnt areas. It is thus sensible to model these two wildfire characteristics jointly.

A natural approach is to consider a marked point process model with point process identifying the occurrences of wildfires and marks identifying the sizes. However, the marks might not be separable from the points (Schoenberg 2004), and thus one challenge is how to specify the dependence between them. Here, we propose a novel two-part model to jointly describe the wildfire count data and burnt area data with environmental covariates, by combining the integrated nested Laplace approximation fast inference with the stochastic partial differential equation (INLA-SPDE) approach. In the first part, we use a binary spatio-temporal model $Z(s, t), s \in D, t \in T$, for the underlying process that determines wildfire occurrences, i.e., whether or not there is wildfire at time $t$ and location $s$, where $D \in \mathbb{R}^2$, $T \in \mathbb{R}$ are the spatial and temporal domains, respectively. The first-part modeling is very useful since it accounts for the zero-inflated pattern in the data (more than 60% of the observed count and burnt area data are zeros); see Liu et al. (2019). In the second part, we consider a non-stationary hurdle log-Gaussian Cox process (hurdle-LGCP) model $X_{\text{CNT}}$ for the shifted positive wildfire count data (minus 1, specifically), i.e., the point pattern, and a non-stationary Gaussian model $X_{\text{BA}}$ for the logarithm of the positive burnt area data, i.e., the marks.
We capture the dependence between the point pattern and marks by a shared spatio-temporal random effect.

The prime goal of this data competition is to estimate the predictive distribution of wildfire occurrences and burnt areas at certain times and sites. In terms of the prediction score criterion chosen by the data competition organizers, our two-part modeling approach clearly outperforms the benchmark model, which is a generalized linear model with Poisson response for the wildfire count data and a generalized linear model with Gaussian response for the logarithm of positive burnt area data. Furthermore, we also aim to identify the key drivers for $Z, X_{\text{CNT}}$, and $X_{\text{BA}}$, respectively.

The paper is structured as follows: Section 2 introduces the data and presents some exploratory analysis. Section 3 details our modeling approach. Section 4 presents the results with interpretations. Section 5 concludes with a discussion.

## 2 Data

The dataset contains monthly wildfire information covering March to September from 1993 to 2015 in the continental United States. More specifically, the study area is partitioned into 3503 cells based on a $0.5^\circ \times 0.5^\circ$ grid of longitude and latitude coordinates. The number of wildfires and their burnt areas in each grid cell are then aggregated monthly and this yields the two main variables in the dataset, namely counts (CNT) and burnt area (BA). There are also 35 auxiliary variables, providing the spatial, temporal, meteorological and land cover information. More details can be found in Opitz (2021).

Table 1 shows the zero and missing value pattern in the variables CNT and BA of this dataset. One can observe that if one of them is zero, the other one must also be zero, and more than 60% of the observed values are zeros. For statistical modeling of zero-inflated nonnegative continuous data, two different approaches are generally adopted, i.e., a Tobit model or a two-part model (Liu et al. 2019). The two-part modeling approach is adopted here since it allows us to model the positive wildfire count data and positive burnt areas jointly; see Section 3 for more details on the proposed model.

Figure 1 depicts the monthly empirical mean of positive CNT and BA in all grid cells. It shows that large wildfire burnt areas often occur for two or three consecutive months, which might be due to the fact that large wildfires often persist for a long period. Moreover, very large wildfires seldom occur in two nonconsecutive months in the same year due to the reduction of wildland vegetation. This observation has

| BA | CNT | 0 | > 0 | NA | Sum |
|----|-----|---|-----|----|-----|
| 0  | 279762 | 0 | 18375 | 298497 |
| > 0| 0   | 173168 | 12318 | 185486 |
| NA | 18831 | 12222 | 48947 | 80000 |
| Sum| 298593 | 185390 | 80000 | 563983 |
motivated us to construct spatio-temporal models rather than pure spatial models, aiming to capture the temporal dependence between months.

Figure 2 shows the logarithm of the empirical mean and standard deviation of the positive CNTs and BAs over time at each grid cell. One can clearly observe spatial non-stationarity in the mean and standard deviation of the positive observations of both CNT and BA. Moreover, the pattern in the non-stationarity of CNT and BA appears to be rather different. Specifically, the western coast and southeastern part of the US appear to have larger numbers of wildfires (large CNTs), but most of the large wildfires (large BAs) appear to occur in the middle- and south-west of the US. Motivated by this observation, we propose to use non-stationary spatial models for CNT and BA; see Section 3 for details of our method to capture spatial non-stationarity in the mean and standard deviation of CNT and BA.

Fig. 1 Empirical mean of positive CNTs and BAs, i.e., excluding observations of zeros, in all the grid cells in each month

Fig. 2 Empirical mean and standard deviation (SD) of positive CNTs and BAs over all months at each grid cell. The grid cells where there are no positive observations are shown in white. The grid cells with only one positive observation, or more than one positive observations but they are equal (resulting in zero SD), are also shown in white in the lower panel
3 Modeling and inference

3.1 Two-part model

In this section we describe our two-part modeling approach. Although it is conceptually convenient to think of our constructed models over continuous time and space, we have to discretize the temporal and spatial domain in order to estimate our model in practice. Our main assumption is that the probability of wildfire occurrence in the binary process, the intensity function of the wildfire point pattern, and the density function of the marks do not vary within each temporal and spatial unit. Here, the temporal unit is chosen as one month and spatial unit is one grid cell.

The first part of our model is a binary logistic process \( Z(s, t) \), \( s \in D, t \in T \), which describes whether or not there is wildfire at site \( s \) and time \( t \). The sets \( D \in \mathbb{R}^2 \), \( T \in \mathbb{R} \) are the spatial and temporal domains, respectively. More specifically, our model has the following structure:

\[
Z(s, t) \mid p(s, t) \sim \text{Bernoulli}(p(s, t)),
\]

\[
\logit(p(s, t)) = \log \left( \frac{p(s, t)}{1 - p(s, t)} \right) = \beta_0^Z + \beta_{\text{time}}^Z t_{\text{year}} + \beta_{\text{time}}^Z t_{\text{month}} + \sum_{i=1}^{20} \beta_{i}^{Z,\text{land}} x_{i,\text{land}}(s) + \sum_{i=1}^{10} \beta_{i}^{Z,\text{clim}} x_{i,\text{clim}}(s, t) + W_Z^1(s) + W_Z^2(s, m(t)),
\]

where \( \beta_0^Z, \beta_{\text{time}}^Z, \beta_{i}^{Z,\text{land}}, \beta_{i}^{Z,\text{clim}} \) are the regression coefficients to estimate. There are 32 covariates in total in the fixed effects, namely 2 temporal covariates \( t_{\text{year}} \in \{1993, 1994, \ldots, 2015\} \) and \( t_{\text{month}} \in \{3, 4, \ldots, 9\} \), 20 spatial covariates \( x_{i,\text{land}}, i = 1, \ldots, 20 \) including 18 land cover covariates and 2 altitude-related covariates, and 10 meteorological covariates \( x_{i,\text{clim}}, i = 1, \ldots, 10 \).

The process \( W_Z^1(s) \) in (1) is a non-stationary spatial random effect aiming to capture the spatial dependence and non-stationarity in space. The non-stationarity in \( W_Z^1(s) \) is incorporated in a similar way as Ingebrigtsen et al. (2014) who impose a parametric function of explanatory variables in the parameters that define the Matérn SPDE model. More precisely, here we define \( W_Z^1(s) \) as the Gaussian Matérn SPDE model

\[
(k^2 - \Delta)^{v+1} \tau(s) W_Z^1(s) = \dot{M}(s), \quad s \in D,
\]

where \( v > 0 \) is a smoothness parameter and here set to 1, \( \kappa > 0 \) is a range parameter, \( \tau(s) \) controls the marginal variability at location \( s \), \( \Delta \) is the Laplacian, and \( \dot{M} \) is Gaussian white noise (Whittle 1963). The solution \( W_Z^1(s) \) is non-stationary because \( \tau(s) \) varies in space. As suggested in Lindgren and Rue (2015) and Ingebrigtsen et al. (2014), a more natural parametrization for this Gaussian Matérn field is to consider the practical range parameter \( \rho = \sqrt{8v}/\kappa \) and standard deviation parameter \( \sigma(s) \approx \Gamma(v)/(\Gamma(v + 1)2\pi^{1/2}k^v\tau(s)) \). Here we want to have a parametrization of the form

\[...\]
\[ \log(\rho) = \theta^Z, \]
\[ \log\{\sigma(s)\} \approx \theta^Z_2 + \theta^Z_3 \hat{\sigma}^Z(s), \]
where \( \hat{\sigma}^Z(s) \) is the empirical standard deviation of \( Z(s,t) \) at site \( s \) over all months, and \( \theta^Z_1, \theta^Z_2, \theta^Z_3 \) are hyperparameters that we need to estimate. Translating this into the SPDE model parameters \( \kappa \) and \( \tau(s) \) implies that we can parameterize them as
\[ \log(\kappa) = \log(\kappa_0) - \theta^Z_1, \]
\[ \log\{\tau(s)\} = \log(\tau_0) + \nu \theta^Z_1 - \theta^Z_2 - \theta^Z_3 \hat{\sigma}^Z(s), \]
where \( \tau_0 \) and \( \kappa_0 \) are fixed constants. For more details about the SPDE approach and the construction of non-stationary models, we refer to Lindgren et al. (2011), Ingebrigtsen et al. (2014) and Krainski et al. (2019).

The process \( W^Z_2(s,m(t)) \) in Eq. (1) is a stationary spatio-temporal random effect defined on the monthly level, i.e., \( m(t) \in \{3, \ldots, 9\} \) denotes the month corresponding to time \( t \) and the effects in different years are considered as replicates, aiming to capture the temporal dependence between months and the remaining spatial dependence that is left out by the non-stationary model \( W^Z_1(s) \). The temporal structure is defined in an autoregressive manner (AR(1), specifically), i.e.,
\[ W^Z_2(s,m(t)) = \rho W^Z_2(s,m(t) - 1) + \sqrt{1 - \rho^2} \epsilon_{m(t)}(s), \quad \rho \in (-1,1), \]
for \( m(t) \in \{4, 5, \ldots, 9\} \), where \( W^Z_2(s, 3) = \epsilon_3(s) \), and \( \epsilon_{m(t)}(s) \) are spatial Matérn SPDE innovation fields.

In the second part, we consider modeling the positive observations of CNT and BA jointly. In order to use a model based on marked point Poisson processes, we subtract the positive CNTs by 1 to transform the data from range \( \{1, 2, 3, \ldots\} \) to non-negative integers. Then the resulting point pattern is modeled by a Poisson process \( X_{\text{CNT}} \) with intensity \( \Lambda(s,t) \), and logarithm of the marks (log BA) are modeled by a non-stationary Gaussian process \( X_{\text{BA}} \). Specifically, \( X_{\text{BA}} \) and log \( \Lambda \) have the following additive structures:
\[ X_{\text{BA}}(s,t) = \rho^BA_0 + \rho^BA_{\text{time}} t_{\text{year}} + \rho^BA_{\text{time}} t_{\text{month}} + \]
\[ \sum_{i=1}^{20} \rho^BA_{\text{land},i} X^\text{land}_i(s) + \sum_{i=1}^{10} \rho^BA_{\text{clim},i} X^\text{clim}_i(s,t) + W^BA_1(s) + W^BA_2(s,m(t)) + \epsilon^BA(s,t), \]
\[ \log \Lambda(s,t) = \rho^CNT_0 + \rho^CNT_{\text{time}} t_{\text{year}} + \rho^CNT_{\text{time}} t_{\text{month}} + \]
\[ \sum_{i=1}^{20} \rho^CNT_{\text{land},i} X^\text{land}_i(s) + \sum_{i=1}^{10} \rho^CNT_{\text{clim},i} X^\text{clim}_i(s,t) + W^CNT_1(s) + W^CNT_2(s,m(t)) + \alpha W^BA_2(s,m(t)), \]
where \( \alpha, \rho^BA_0, \rho^CNT_0, \rho^BA_{\text{type}}, \rho^CNT_{\text{type}}, \) type \( \in \{\text{time, land, clim}\} \) are regression parameters to estimate, \( \epsilon^BA(s,t) \) can be thought of as a noise or measurement error process which
has independent and identical Gaussian distributions with zero mean and unknown precision parameter at any time and space, \( W_{\text{CNT}}^1(s) \), \( W_{\text{BA}}^1(s) \) are non-stationary spatial random effects with non-stationarity constructed in the same way as \( W_1^Z(s) \), and \( W_{\text{CNT}}^2(s, m(t)) \), \( W_{\text{BA}}^2(s, m(t)) \) are stationary spatio-temporal random effects with AR(1) temporal structures constructed as \( W_2^Z(s, m(t)) \). The dependence between \( X_{\text{BA}} \) and \( \log \Lambda(s, t) \) is specified by the shared random effect \( W_{\text{BA}}^2(s, m(t)) \), and controlled through the parameter \( \alpha > 0 \). Prior distributions and hyperparameters are discussed in the next section.

### 3.2 Bayesian inference using INLA

For each of the random effects \( W_j^i, i = 1, 2, j \in \{ Z, \text{BA}, \text{CNT} \} \), we use the SPDE approach to approximate the Gaussian random fields with Matérn covariance by Gaussian Markov random fields, thus enabling computationally efficient inference with INLA (Rue et al. 2009; Lindgren et al. 2011). The SPDE approach is based on a triangulation of the bounded spatial domain \( D \). Due to computationally considerations, we here choose a fine mesh with 3967 nodes (mesh 1 in Fig. 3) for the spatial random effects \( W_1^j, j \in \{ Z, \text{BA}, \text{CNT} \} \) and a coarse mesh with 743 nodes (mesh 2 in Fig. 3) for the spatio-temporal random effects \( W_2^j, j \in \{ Z, \text{BA}, \text{CNT} \} \).

For the model \( Z \), we have 6 hyperparameters, i.e., \( \theta_1^Z, \theta_2^Z, \theta_3^Z \) for the non-stationary spatial random effect \( W_1^Z \), and a range parameter, a variance parameter, and the temporal autoregression coefficient \( \rho \) for the spatio-temporal random effect \( W_2^Z \). Here we choose a penalized complexity (PC) prior (Simpson et al. 2017; Fuglstad et al. 2019) for the range parameter, variance parameter, and \( \rho \) of random effect \( W_2^Z \), and default vague priors in the R-INLA package (Lindgren and Rue 2015) for other parameters. Specifically, the PC prior distributions are fixed such that the prior probability of having a covariance range less than 55 km is 0.1, of having a variance larger than 0.25 is 0.1, and of having temporal autocorrelation parameter \( \rho \) below 0 is 0.05.

The joint model of \( X_{\text{CNT}} \) and \( X_{\text{BA}} \) has 14 hyperparameters, including one precision parameter for \( \Theta_{\text{BA}} \), three parameters for each of \( W_j^i, i = 1, 2, j \in \{ \text{BA}, \text{CNT} \} \), and parameter \( \alpha \) for the shared random effect. We again use PC priors for the temporal autoregression parameter, the range parameter, and variance parameter of the spatio-temporal random effect \( W_2^j \) and default vague priors for other parameters. The PC priors for \( W_2^j, j \in \{ \text{BA}, \text{CNT} \} \) are set the same as that for \( W_2^Z \).

Fig. 3 Triangulation over continental US. Green dots indicate the observational sites. Left panel: fine mesh with 3967 nodes for spatial random effects \( W_1^j, j \in \{ Z, \text{BA}, \text{CNT} \} \); right panel: coarse mesh with 743 nodes for spatio-temporal random effects \( W_2^j, j \in \{ Z, \text{BA}, \text{CNT} \} \)
INLA provides a useful tool for Bayesian modeling and inference for latent Gaussian models. This is the case for our two-part model as $Z$ is a binomial regression model with logit link, $X_{BA}$ is a Gaussian model, and $X_{CNT}$ is a Poisson regression model with log link, all of which are conditionally independent of the data level and include latent effects that are jointly Gaussian. The new package PARDISO (van Niekerk et al. 2021) has enabled parallel computation in INLA and further increased its scalability, which allows us to fit our complex model to this massive wildfire data. The computation time for fitting our first-part model is around 2 hours on a cluster with 48 cores and 2.9 TB memory, and around 42 hours for fitting our second-part model on the same cluster. The R code is available at https://github.com/zhongwei-zh/EVA2021-data-competition.

4 Results

4.1 Prediction performance

We first report the prediction performance of our model as this is the goal of this data competition. All participants of the competition are required to submit an estimation of the distribution of CNT and BA evaluated at a list of 28 severity values, at 80000 different time and locations. The prediction score of each participant is then calculated based on a modified version of weighted ranked probability score chosen by the organizers, where relatively strong weight is assigned to large values of CNT and BA; see Opitz (2021) for more details. For our model, the prediction score for CNT is 3498.74 and the one for BA is 3389.50, which clearly outperforms the benchmark model whose prediction scores for CNT and BA are 5565.15 and 4244.36, respectively. As a comparison, we have also run our model with stationary variances (i.e., with $\theta_j^Z = 0$, $j = Z, BA, CNT$), and the prediction scores for CNT and BA are respectively 3512.21 and 3389.79. This modest improvement in prediction performance further confirms the advantage of using non-stationary models over stationary ones for this dataset, especially for predicting wildfire counts.

4.2 Hyperparameter estimates

In this section we report estimates of the hyperparameters in our model. Table 2 presents the posterior mean estimates and 95% credible intervals of all the hyperparameters. The results show that non-stationarity in the spatial random effects $W_Z^1, W_{CNT}^1$, and $W_{BA}^1$ are all significant since the 95% credible intervals for $\theta_3^Z, \theta_3^{CNT}$ and $\theta_3^{BA}$ do not cover zero. Temporal dependence between months are very significant since estimates of the temporal autocorrelation parameters for $W_Z^2, W_{CNT}^2$, and $W_{BA}^2$ are all far from zero and their 95% credible intervals do not contain zero. Furthermore, spatial dependence for $Z$ and $X_{CNT}$ seems to be stronger than that for $X_{BA}$, as estimates of the range parameters of $W_Z^2$ and $W_{CNT}^2$ are much larger than that of $W_{BA}^2$ and estimates of $\theta_1^Z$ and $\theta_1^{CNT}$ are also much larger than that of $\theta_1^{BA}$. Finally, the dependence between $X_{CNT}$ and $X_{BA}$ is significant and
inclusion of the shared random effect is necessary since the estimate of the coefficient $\alpha$ is far from zero and its 95% credible interval does not contain zero.

### 4.3 Influence of covariates on $Z, X_{\text{CNT}}, X_{\text{BA}}$

In addition to accurate wildfire prediction, we also aim at identifying the most influential covariates on the three different responses $Z, X_{\text{CNT}},$ and $X_{\text{BA}}$. Following the recommendation of Gelman et al. (2008), we standardize all the covariates in a preliminary step to make them have mean 0 and standard deviation 1, so that the effects of different covariates are comparable and interpretable. Table 3 presents the three most positively and negatively influential covariates and their estimated coefficients for the models $Z, X_{\text{CNT}}, X_{\text{BA}}$. One interesting observation is that the most influential covariates for the three models $Z, X_{\text{CNT}},$ and $X_{\text{BA}}$ are in general not the same. More specifically, surface pressure is the most positively influential covariate to the occurrence of a wildfire, which might be the case because lightning is the principle natural cause of wildfire ignition and surface pressure is often quite high when lightning occurs. Moreover, while high surface pressure may lead to large numbers of wildfires (possibly because of lightning), high water evaporation, which often occurs when the temperature is high, the air is dry and the wind is strong, leads to large burnt areas. For the negatively influential covariates, shrubland is a
significant covariate for three models, especially for $Z$ and $X_{\text{CNT}}$. This might be due to the fact that large areas of shrubland means less human activity and this results in less human-caused wildfires. Furthermore, high dew point temperature leads to less wildfires and smaller burnt areas, which might be due to the fact that the higher the dew point temperature, the greater the amount of moisture in the air. Finally, altitude is a negatively influential covariate for $X_{\text{BA}}$, which might be explained by the fact that high altitude corresponds to lower temperature and often less human activity.

### 5 Discussion

In this paper we have proposed a novel two-part statistical model for jointly modeling zero-inflated wildfire count data and burnt area data. Our model clearly outperforms the benchmark model in terms of its prediction performance. Although it performs slightly worse than the model proposed by the best-performing team (named “BlackBox”) of this data competition, which uses algorithmic models based on deep learning methods, our model yields interpretable results and some understanding of the most important causative drivers that may trigger wildfires.

There are various interesting future research directions. For instance, one can explore better usage of the covariates information. Here we only considered a linear additive structure of the temporal, land cover, altitude-related, and meteorological covariates, but a non-linear relationship between some of them and the response...
variable might exist. One possibility is to include interaction effects of covariates, e.g., low precipitation and high temperatures defining very dry conditions. Alternatively, one can investigate how to better build the non-stationary random effects. Here we chose the empirical marginal variance as the explanatory variable for the variance parameter of the SPDE model, but one can include further variables as one wishes, as long as the added computational cost is acceptable. As pointed out by one referee, since the temporal resolution of the land cover variables is annual, which might be too coarse, restructuring them might bring extra flexibility and improve the model’s prediction ability. For example, one might fix the proportion of the land cover variables in a certain month as the one given by the competition organizers, and use interpolation to derive the proportions in other months. One could also consider constructing more complex space-varying regression models as in Opitz et al. (2022). All these efforts might be rewarded with a better prediction performance.

Finally, here we did not consider asymptotic models justified by extreme value theory for large values of CNT or BA. One reason is that current spatio-temporal extremes models are limited to problems of moderate dimensions and are not suitable for the massive wildfire dataset. Another practical reason is that although an approach similar to Opitz et al. (2018) or Castro-Camilo et al. (2019) may be taken to consider a generalized Pareto distribution for high threshold exceedances of log BA, a key feature in the burnt area data is that they are bounded from above, i.e., burnt area cannot exceed the area of their respective grid cells. This means that the generalized Pareto distribution might either have to be truncated at some point above the threshold, or restricted to negative shape parameters, which then brings new modeling challenges and might also involve substantial extra computational cost. Therefore, investigation of how to integrate extreme value theory in this data application is another interesting future research direction.

Acknowledgements We thank two anonymous referees and the guest editor for helpful comments that have improved the manuscript. We also thank Thomas Opitz for organizing the data competition, and Thomas Mikosch for welcoming a special issue on this topic.

Funding This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. OSR-CRG2020-4394.

Data Availability Not applicable.

Code Availability The code is available at GitHub and the link is https://github.com/zhongwei-zh/EVA2021-data-competition.

Declarations

Conflicts of interest Not applicable.

References

Burke, M., Driscoll, A., Heft-Neal, S., Xue, J., Burney, J., Wara, M.: The changing risk and burden of wildfire in the United States. PNAS 118(2) (2021)
Butry, D.T., Gumpertz, M., Genton, M.G.: The production of large and small wildfires. In: T.P. Holmes, J.P. Prestemon, K.L. Abt (eds.) The Economics of Forest Disturbances: Wildfires, Storms and Invasive Species, pp. 79–106 (2008)

Castro-Camilo, D., Huser, R., Rue, H.: A sliced Gamma-generalized Pareto model for short-term extreme wind speed probabilistic forecasting. J. Agric. Biol. Environ. Stat. 24, 517–534 (2019)

Cumming, S.G.: A parametric model of the fire-size distribution. Can. J. For. Res. 31, 1297–1303 (2001)

Fuglstad, G., Simpson, D., Lindgren, F., Rue, H.: Constructing priors that penalize the complexity of Gaussian random fields. J. Am. Stat. Assoc. 114(525), 445–452 (2019)

Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.S.: A weakly informative default prior distribution for logistic and regression models. Ann. Appl. Stat. 2(4), 1360–1383 (2008)

Hanson, S., Pueyo, S., Chuvieco, E.: Global fire size distribution: from power law to log-normal. Int. J. Wildland Fire 25(4), 403–412 (2016)

Holmes, T.P., Huggett, R.J.J., Westerling, A.L.: Statistical analysis of large wildfires. In: T.P. Holmes, J.P. Prestemon, K.L. Abt (eds.) The Economics of Forest Disturbances: Wildfires, Storms and Invasive Species, pp. 59–77 (2008)

Ingebritsens, R., Lindgren, F., Steinsland, I.: Spatial models with explanatory variables in the dependence structure. Spat. Stat. 8, 20–38 (2014)

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., Rue, H.: Advanced spatial modeling with stochastic partial differential equations using R and INLA. Chapman and Hall/CRC (2019)

Lindgren, F., Rue, H.: Bayesian spatial modelling with R-INLA. J. Stat. Softw. 63(19), 1–25 (2015)

Lindgren, F., Rue, H., Lindström, J.: An explicit link between the Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. (Series B) 73, 423–498 (2011)

Liu, L., Shih, Y.T., Strawderman, R.L., Zhang, D., Johnson, B.A., Chai, H.: Statistical analysis of zero-inflated nonnegative continuous data: a review. Stat. Sci. 34(2), 253–279 (2019)

Liu, Y., Goodrick, S., Heilman, W.: Wildland fire emissions, carbon, and climate: Wildfire-climate interactions. For. Ecol. Manage. 317, 80–96 (2014)

Opitz, T.: Editorial: EVA 2021 data competition on spatio-temporal prediction of wildfire activity in the United States. Extremes (2022)

Opitz, T., Bakka, H., Huser, R., Lombardo, L.: High-resolution Bayesian mapping of landslide hazard with unobserved trigger event. Ann. Appl. Stat. (2022)

Opitz, T., Bonneu, F., Gabriel, E.: Point-process based Bayesian modeling of space-time structures of forest fire occurrences in Mediterranean France. Spat. Stat. 40(100429) (2020)

Opitz, T., Huser, R., Bakka, H., Rue, H.: INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. Extremes 21, 441–462 (2018)

Reed, W.J., McKelvey, K.S.: Power-law behaviour and parametric models for the size-distribution of forest fires. Ecol. Model. 150, 239–254 (2002)

Rosenthal, A., Stover, E., Haar, R.J.: Health and social impacts of California wildfires and the deficiencies in current recovery resources: An exploratory qualitative study of systems-level issues. PLoS One 16(3), (2021)

Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society (Series B) 71(2), 319–392 (2009)

Schoenberg, F.P.: Testing separability in spatial-temporal marked point processes. Biometrics 60, 471–481 (2004)

Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Díaz-Ávalos, C., Rue, H.: Spatio-temporal log-Gaussian Cox processes for modelling wildfire occurrence: the case of Catalonia, 1994–2008. Environ. Ecol. Stat. 21, 531–563 (2014)

Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sorbye, S.H.: Penalising model component complexity: a principled, practical approach to constructing priors. Stat. Sci. 32(1), 1–28 (2017)

Taylor, S.W., Woolford, D.G., Dean, C.B., Martell, D.L.: Wildfire prediction to inform fire management: statistical science challenges. Stat. Sci. 28(4), 586–615 (2013)

van Niekerk, J., Bakka, H., Rue, H., Schenk, O.: New frontiers in Bayesian modeling using the INLA package in R. J. Stat. Softw. 100(2), 1–28 (2021)

Whittle, P.: Stochastic processes in several dimensions. Bull. Int. Stat. Inst. 40, 974–994 (1963)

Xi, D.D., Taylor, S.W., Woolford, D.G., Dean, C.B.: Statistical models of key components of wildfire risk. Ann. 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.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Zhongwei Zhang\textsuperscript{1,2} · Elias Krainski\textsuperscript{2} · Peng Zhong\textsuperscript{2} · Harvard Rue\textsuperscript{2} · Raphaël Huser\textsuperscript{2}\textsuperscript{3}

Elias Krainski
elias.krainski@kaust.edu.sa

Peng Zhong
peng.zhong@kaust.edu.sa

Harvard Rue
Haavard.Rue@kaust.edu.sa

Raphaël Huser
raphael.huser@kaust.edu.sa

\textsuperscript{1} Research Center for Statistics, University of Geneva, Geneva 1205, Switzerland

\textsuperscript{2} Statistics Program, CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia