Joint Modeling and Prediction of Massive Spatio-Temporal Wildfire Count and Burnt Area Data with the INLA-SPDE Approach

Zhongwei Zhang, Elias Krainski, Peng Zhong, Håvard Rue,

and Raphaël Huser

Statistics Program, CEMSE Division, King Abdullah University of

Science and Technology

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 model that is based on log-Gaussian Cox processes for positive wildfire 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.

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 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 & 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 $Z(\boldsymbol{s}, t), \boldsymbol{s} \in \mathcal{D}, t \in T$, for the underlying process that determines wildfire occurrences, i.e., whether or not there is wildfire at time t and location $\boldsymbol{s}$, where $\mathcal{D} \in \mathbb{R}^2, T \in \mathbb{R}$ 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 log-Gaussian Cox process model X_{CNT} for the shifted positive wildfire count data (minus 1, specifically), i.e., the point pattern, and a non-stationary Gaussian model X_{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

Table 1: Zero and missing value pattern in the variables CNT and BA of the wildfire dataset

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

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_{CNT}, and X_{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 $0.5° \times 0.5°$ 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 (2022).

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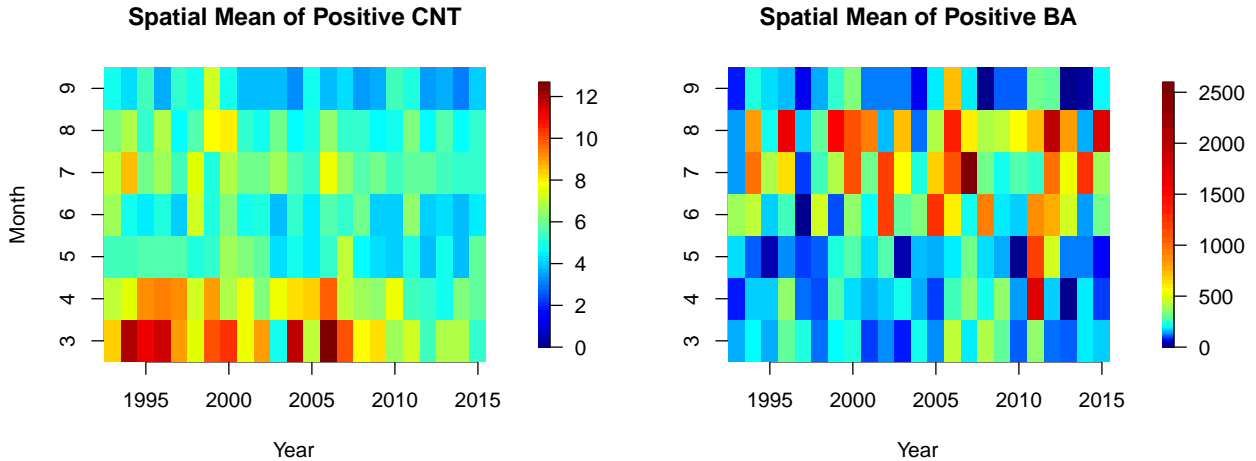
Figure 1: Empirical mean of positive CNTs and BAs, i.e., excluding observations of zeros, in all the grid cells in each month.

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 motivated us to construct spatio-temporal models rather than pure spatial models, aiming to capture the temporal dependence between months.

Figure 2 shows the 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 appear 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
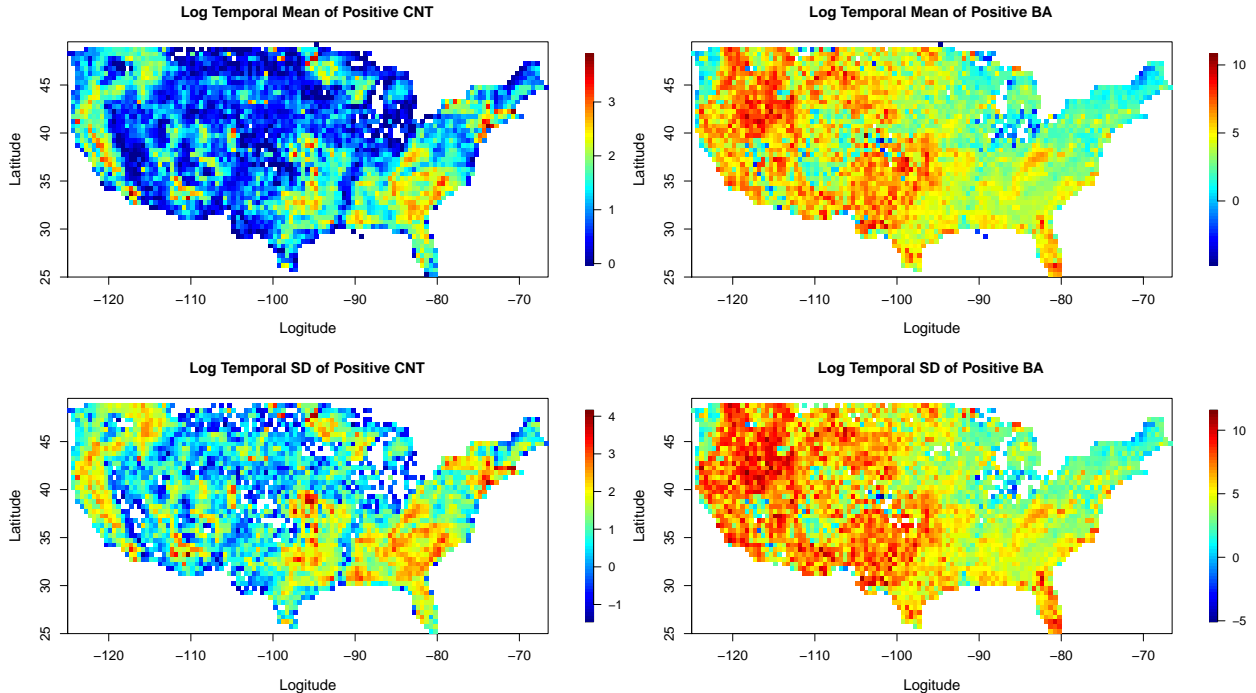
5

Figure 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.

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.

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

$$Z(\boldsymbol{s}, t) \mid p(\boldsymbol{s}, t) \sim \text{Bernoulli}(p(\boldsymbol{s}, t)),$$

$$\text{logit}(p(\boldsymbol{s}, t)) = \log \left\{ \frac{p(\boldsymbol{s}, t)}{1 - p(\boldsymbol{s}, t)} \right\} = \beta_0^Z + \beta_1^{Z,\text{time}} t_{\text{year}} + \beta_2^{Z,\text{time}} t_{\text{month}} +$$

$$\sum_{i=1}^{20} \beta_i^{Z,\text{land}} x_i^{\text{land}}(\boldsymbol{s}) + \sum_{i=1}^{10} \beta_i^{Z,\text{clim}} x_i^{\text{clim}}(\boldsymbol{s}, t) +$$

$$W_1^Z(\boldsymbol{s}) + W_2^Z(\boldsymbol{s}, a(t)), \tag{1}$$

where $\beta_0^Z, \beta_i^{Z,\text{time}}, \beta_i^{Z,\text{land}}, \beta_i^{Z,\text{clim}}$ 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 $W_1^Z(\boldsymbol{s})$ in (1) is a non-stationary spatial random effect aiming to capture the spatial dependence and non-stationarity in space. The non-stationarity in $W_1^Z(\boldsymbol{s})$ 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. Here we choose a linear function in terms of the empirical marginal variance as the explanatory variable. More precisely, we model $W_1^Z(\boldsymbol{s})$ through the Matérn SPDE model

$$\{\kappa^2 - \Delta\}^{\nu+1} \tau(\boldsymbol{s}) W_1^Z(\boldsymbol{s}) = \dot{\mathcal{M}}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D},$$

where $\nu > 0$ is a smoothness parameter and here set to 1, Δ is the Laplacian operator, $\dot{\mathcal{M}}$

is Gaussian white noise (Whittle, 1963), and where we set

$$\log\{\tau(\boldsymbol{s})\} = \log(\tau_0) + \theta_1^Z + \theta_2^Z - \hat{\sigma}^Z(\boldsymbol{s})\theta_3^Z,$$

$$\log\{\kappa\} = \log(\kappa_0) - \theta_1^Z + \theta_2^Z,$$

with τ_0, κ_0 constants, $\hat{\sigma}^Z(\boldsymbol{s})$ the empirical standard deviation of $Z(\boldsymbol{s}, t)$ at site $\boldsymbol{s}$ over all months, and $\theta_1^Z, \theta_2^Z, \theta_3^Z$ hyperparameters that we need to estimate. In this case the solution $W_1^Z(\boldsymbol{s})$ is a non-stationary Gaussian random field because τ varies with location. Furthermore, the resulting non-stationarity only lies in the marginal variances and we approximately have the marginal variance

$$\sigma(\boldsymbol{s})^2 \approx \frac{1}{4\pi\kappa^2\tau(\boldsymbol{s})^2}.$$

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 $W_2^Z(\boldsymbol{s}, a(t))$ in (1) is stationary spatio-temporal random effect defined on the monthly level, i.e., $a(t) \in \{3, \ldots, 9\}$ 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 $W_1^Z(\boldsymbol{s})$. The temporal structure is defined in an autoregressive manner (AR(1), specifically), i.e.,

$$W_2^Z(\boldsymbol{s}, a(t)) = \rho W_2^Z(\boldsymbol{s}, a(t) - 1) + \sqrt{1 - \rho^2}\epsilon_{a(t)}(\boldsymbol{s}), \quad \rho \in (-1, 1),$$

for $a(t) \in \{4, 5, \ldots, 9\}$, where $W_2^Z(\boldsymbol{s}, 3) = \epsilon_3(\boldsymbol{s})$, and $\epsilon_{a(t)}(\boldsymbol{s})$ 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 nonnegative integers. Then

the resulting point pattern is modeled by a Poisson process X_{CNT} with intensity $\Lambda(\boldsymbol{s}, t)$, and logarithm of the marks ($\log \mathrm{BA}$) are modeled by a non-stationary Gaussian process X_{BA}. Specifically, X_{BA} and $\log \Lambda$ have the following additive structures:

$$X_{\mathrm{BA}}(\boldsymbol{s}, t) = \beta_0^{\mathrm{BA}} + \beta_1^{\mathrm{BA,time}} t_{\mathrm{year}} + \beta_2^{\mathrm{BA,time}} t_{\mathrm{month}} +$$
$$\sum_{i=1}^{20} \beta_i^{\mathrm{BA,land}} x_i^{\mathrm{land}}(\boldsymbol{s}) + \sum_{i=1}^{10} \beta_i^{\mathrm{BA,clim}} x_i^{\mathrm{clim}}(\boldsymbol{s}, t) +$$
$$W_1^{\mathrm{BA}}(\boldsymbol{s}) + W_2^{\mathrm{BA}}(\boldsymbol{s}, a(t)) + \epsilon^{\mathrm{BA}}(\boldsymbol{s}, t),$$
$$\log \Lambda(\boldsymbol{s}, t) = \beta_0^{\mathrm{CNT}} + \beta_1^{\mathrm{CNT,time}} t_{\mathrm{year}} + \beta_2^{\mathrm{CNT,time}} t_{\mathrm{month}} +$$
$$\sum_{i=1}^{20} \beta_i^{\mathrm{CNT,land}} x_i^{\mathrm{land}}(\boldsymbol{s}) + \sum_{i=1}^{10} \beta_i^{\mathrm{CNT,clim}} x_i^{\mathrm{clim}}(\boldsymbol{s}, t) +$$
$$W_1^{\mathrm{CNT}}(\boldsymbol{s}) + W_2^{\mathrm{CNT}}(\boldsymbol{s}, a(t)) + \alpha W_2^{\mathrm{BA}}(\boldsymbol{s}, a(t)),$$

where $\alpha, \beta_0^{\mathrm{BA}}, \beta_0^{\mathrm{CNT}}, \beta_i^{\mathrm{BA,type}}, \beta_i^{\mathrm{CNT,type}}, \mathrm{type} \in \{\mathrm{time}, \mathrm{land}, \mathrm{clim}\}$ are regression parameters to estimate, $\epsilon^{\mathrm{BA}}(\boldsymbol{s}, t)$ can be thought of as a noise or measurement error process which has independent and identical Gaussian distribution with zero mean and unknown precision parameter at any time and space, $W_1^{\mathrm{CNT}}(\boldsymbol{s})$, $W_1^{\mathrm{BA}}(\boldsymbol{s})$ are non-stationary spatial random effects with non-stationarity constructed in the same way as $W_1^{\mathrm{Z}}(\boldsymbol{s})$, and $W_2^{\mathrm{CNT}}(\boldsymbol{s}, a(t))$, $W_2^{\mathrm{BA}}(\boldsymbol{s}, a(t))$ are stationary spatio-temporal random effects with AR(1) temporal structures constructed as $W_2^{\mathrm{Z}}(\boldsymbol{s}, a(t))$. The dependence between X_{BA} and $\log \Lambda(\boldsymbol{s}, t)$ is specified by the shared random effect $W_2^{\mathrm{BA}}(\boldsymbol{s}, a(t))$, 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 $W_i^j, i = 1, 2, j \in \{Z, \mathrm{BA}, \mathrm{CNT}\}$, 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;
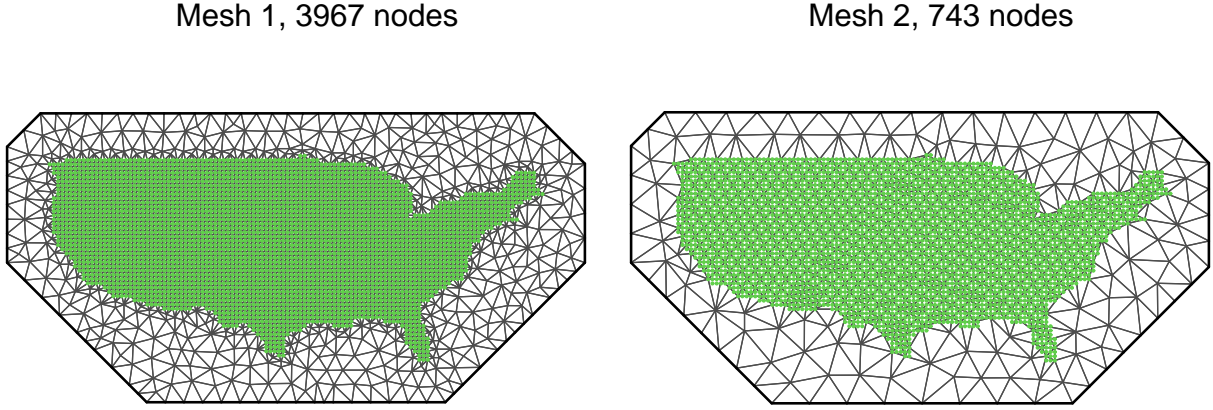
Figure 3: Triangulation over continental US. Green dots indicate the observational sites. Left panel: fine mesh with 3967 nodes for spatial random effects $W_1^j, j \in \{Z, \mathrm{BA}, \mathrm{CNT}\}$; right panel: coarse mesh with 743 nodes for spatio-temporal random effects $W_2^j, j \in \{Z, \mathrm{BA}, \mathrm{CNT}\}$.

Lindgren et al., 2011). The SPDE approach is based on a triangulation of the bounded spatial domain $\mathcal{D}$. Due to computationally considerations, we here choose a fine mesh with 3967 nodes (mesh 1 in Figure 3) for the spatial random effects $W_1^j, j \in \{Z, \mathrm{BA}, \mathrm{CNT}\}$ and a coarse mesh with 743 nodes (mesh 2 in Figure 3) for the spatio-temporal random effects $W_2^j, j \in \{Z, \mathrm{BA}, \mathrm{CNT}\}$.

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 ρ 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 ρ of random effect W_2^Z, and default vague priors in the R-INLA package (Lindgren & 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 ρ below 0 is 0.05. The joint model of X_{CNT} and X_{BA} has 14 hyperparameters, including one precision parameter for ϵ^{BA}, three parameters for each of

$W_i^j, i = 1, 2, j \in \{\text{BA}, \text{CNT}\}$, and parameter α 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.

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 (2022) for more details. For our model, the prediction score for CNT is 3498.73 and the one for BA is 3389.51, which clearly outperforms the benchmark model whose prediction scores for CNT and BA are 5565.15 and 4244.36, respectively.

4.2 Hyperparameter Estimates

Table 2: Posterior mean estimates and 95% credible intervals of hyperparamters in the models $Z, X_{\text{CNT}}, X_{\text{BA}}$.

Model	Random effect	Hyperparameter	Estimate	95% CI
Z	W_1^Z	θ_1^Z	2.07	[1.97, 2.17]
	W_1^Z	θ_2^Z	-1.73	[-1.84, -1.63]
	W_1^Z	θ_3^Z	0.073	[0.052, 0.093]
	W_2^Z	Spatial range (km)	429	[413, 447]
	W_2^Z	Standard deviation	1.99	[1.94, 2.03]
	W_2^Z	Temporal autocorrelation	0.851	[0.845, 0.856]
X_{CNT}	W_1^{CNT}	θ_1^{CNT}	2.38	[2.36, 2.39]
	W_1^{CNT}	θ_2^{CNT}	-0.83	[-0.85,- 0.82]
	W_1^{CNT}	θ_3^{CNT}	-0.33	[-0.34, -0.31]
	W_2^{CNT}	Spatial range (km)	389	[378, 398]
	W_2^{CNT}	Standard deviation	0.64	[0.61, 0.66]
	W_2^{CNT}	Temporal autocorrelation	0.812	[0.804, 0.818]
	Shared random effect	α	0.703	[0.696, 0.711]
X_{BA}	W_1^{BA}	θ_1^{BA}	2.05	[2.03, 2.06]
	W_1^{BA}	θ_2^{BA}	-0.52	[-0.55, -0.50]
	W_1^{BA}	θ_3^{BA}	0.089	[0.082, 0.096]
	W_2^{BA}	Spatial range (km)	175	[171, 179]
	W_2^{BA}	Standard deviation	1.38	[1.36, 1.39]
	W_2^{BA}	Temporal autocorrelation	0.482	[0.470, 0.494]
	ϵ^{BA}	Precision	0.268	[0.266, 0.270]

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_1^Z, W_1^{CNT}, and W_1^{BA} are all significant since the 95% credible intervals for θ_3^Z, θ_3^{CNT} and θ_3^{BA} do not cover zero. Temporal dependence between months are very significant since estimates of the temporal autocorrelation parameters for W_2^{Z}, W_2^{CNT}, and W_2^{BA} 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_2^Z and

W_2^{CNT} are much larger than that of W_2^{BA}. Finally, the dependence between X_{CNT} and X_{BA} is significant and inclusion of the shared random effect is necessary since estimate of the coefficient α is far from zero and its 95% credible interval does not contain zero.

4.3   Influence of Covariates on $Z, X_{\mathrm{CNT}}, X_{\mathrm{BA}}$

Table 3: Three most positively and negatively influential covariates (based on the values of their posterior means) for the models $Z, X_{\mathrm{CNT}}, X_{\mathrm{BA}}$. The type of the covariates, the posterior mean estimates of the corresponding coefficients, and their 95% credible intervals are presented.

Model	Covariate	Type	Estimate	95% CI
Z	Shrubland	Land cover	-2.67	[-4.44, -0.91]
	Cropland rainfed herbaceous cover	Land cover	-2.04	[-3.54, -0.55]
	Grassland	Land cover	-1.76	[-3.16, -0.37]
	Surface pressure	Climate	2.50	[1.88, 3.12]
	Temperature	Climate	0.62	[0.46, 0.77]
	Surface net solar radiation	Climate	0.56	[0.50, 0.63]
X_{CNT}	Shrubland	Land cover	-0.43	[-0.79, -0.07]
	Dewpoint temperature	Climate	-0.28	[-0.35, -0.22]
	Cropland rainfed herbaceous cover	Land cover	-0.25	[-0.55, 0.06]
	Surface net solar radiation	Climate	0.40	[0.36, 0.43]
	Surface pressure	Climate	0.32	[0.17, 0.46]
	Temperature	Climate	0.21	[0.14, 0.28]
X_{BA}	Dewpoint temperature	Climate	-0.58	[-0.71, -0.45]
	Altitude mean	Topography	-0.49	[-0.82, -0.17]
	Shrubland	Land cover	-0.40	[-1.54, 0.74]
	Temperature	Climate	0.60	[0.48, 0.73]
	Evaporation	Climate	0.38	[0.34, 0.42]
	Surface net solar radiation	Climate	0.19	[0.14, 0.25]

In addition to accurate wildfire prediction, we also aim at identifying the most influential covariates on the three different responses Z, X_{CNT}, and X_{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 $Z, X_{\mathrm{CNT}}, X_{\mathrm{BA}}$. One interesting observation is that the most influential covariates for the three models Z, X_{CNT}, and X_{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_{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_{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. 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. 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 should be truncated at some point above the threshold, 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.

References

Burke, M., Driscoll, A., Heft-Neal, S., Xue, J., Burney, J., & Wara, M. (2021). The changing risk and burden of wildfire in the United States. *PNAS*, *118*(2), e2011048118.

Butry, D. T., Gumpertz, M., & Genton, M. G. (2008). 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).

Castro-Camilo, D., Huser, R., & Rue, H. (2019). A sliced Gamma-generalized Pareto model for short-term extreme wind speed probabilistic forecasting. *Journal of Agricultural, Biological and Environmental Statistics*, *24*, 517–534.

Cumming, S. G. (2001). A parametric model of the fire-size distribution. *Canadian Journal of Forest Research*, *31*, 1297–1303.

Fuglstad, G., Simpson, D., Lindgren, F., & Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, *114*(525), 445–452.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Hantson, S., Pueyo, S., & Chuvieco, E. (2016). Global fire size distribution: from power law to log-normal. *International Journal of Wildland Fire*, *25*(4), 403–412.

Holmes, T. P., Huggett, R. J. J., & Westerling, A. L. (2008). 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).

Ingebrigtsen, R., Lindgren, F., & Steinsland, I. (2014). Spatial models with explanatory variables in the dependence structure. *Spatial Statistics*, *8*, 20–38.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., & Rue, H. (2019). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC.

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, *63*(19), 1–25.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between the Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society (Series B)*, *73*, 423–498.

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

Liu, Y., Goodrick, S., & Heilman, W. (2014). Wildland fire emissions, carbon, and climate: Wildfire-climate interactions. *Forest Ecology and Management*, *317*, 80–96.

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

Opitz, T., Bakka, H., Huser, R., & Lombardo, L. (2022). High-resolution Bayesian mapping of lanslide hazard with unobserved trigger event. *Annals of Applied Statistics*, to appear.

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

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

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

Rosenthal, A., Stover, E., & Haar, R. J. (2021). 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), e0248617.

Rue, H., Martino, S., & Chopin, N. (2009). 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.

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

Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Díaz-Ávalos, C., & Rue, H. (2014). Spatio-temporal log-Gaussian Cox processes for modelling wildfire occurence: the case of Catalonia, 1994-2008. *Environmental and Ecological Statistics*, *21*, 531–563.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: a principled, practical approach to constructing priors. *Statistical Science*, *32*(1), 1–28.

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

van Niekerk, J., Bakka, H., Rue, H., & Schenk, O. (2021). New frontiers in Bayesian modeling using the INLA package in R. *Journal of Statistical Software*, *100*(2), 1–28.

Whittle, P. (1963). Stochastic processes in several dimensions. *Bulletin of the International Statistical Institute*, *40*, 974–994.

Xi, D. D., Taylor, S. W., Woolford, D. G., & Dean, C. B. (2019). Statistical models of key components of wildfire risk. *Annual Review of Statistics and Its Application*, *6*, 197–222.