A combined statistical and machine learning approach for spatial prediction of extreme wildfire frequencies and sizes

Daniela Cisneros1 · Yan Gong1 · Rishikesh Yadav1 · Arnab Hazra1,2 · Raphaël Huser1

Received: 29 December 2021 / Revised: 23 November 2022 / Accepted: 16 December 2022 / Published online: 12 February 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Motivated by the Extreme Value Analysis 2021 (EVA 2021) data challenge, we propose a method based on statistics and machine learning for the spatial prediction of extreme wildfire frequencies and sizes. This method is tailored to handle large datasets, including missing observations. Our approach relies on a four-stage, bivariate, sparse spatial model for high-dimensional zero-inflated data that we develop using stochastic partial differential equations (SPDE), allowing sparse precision matrices for the latent processes. In Stage 1, the observations are separated in zero/nonzero categories and modeled using a two-layered hierarchical Bayesian sparse spatial model to estimate the probabilities of these two categories. In Stage 2, we first obtain empirical estimates of the spatially-varying mean and variance profiles across the spatial locations for the positive observations and smooth those estimates using fixed rank kriging. This approximate Bayesian inference method is employed to avoid the high computational burden of large spatial data modeling using spatially-varying coefficients. In Stage 3, we further model the standardized log-transformed positive observations from the second stage using a sparse bivariate spatial Gaussian process. The Gaussian distribution assumption for wildfire counts developed in the third stage is computationally effective but erroneous. Thus, in Stage 4, the predicted exceedance probabilities are post-processed using Random Forests. We draw posterior inference for Stages 1 and 3 using Markov chain Monte Carlo (MCMC) sampling. We then create a cross-validation scheme for the artificially generated gaps and compare the EVA 2021 prediction scores of the proposed model to those obtained using some competitors.

Keywords Approximate Bayesian inference · Extreme wildfire frequencies and sizes · Gaussian Markov random field · Random Forests · Stochastic partial differential equation

AMS 2000 Subject Classifications 62G32 · 62H11 · 62J05 · 62J12 · 62P12

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

Wildfires have become an important concern in recent years because they cause air pollution, extinction of flora and fauna, significant economic loss, irreparable damage to the environment and the atmosphere, and threats to people’s lives. Wildfires occur because of multiple causes, such as human intervention (e.g., agricultural activities, campfires, and smoking) (Pyne et al. 1996), lightning, volcanic eruption, debris burning, sparks from rock falls, and spontaneous combustion (Scott 2000). Typically, wildfires are accelerated by favorable conditions such as extreme weather events (e.g., a drought) and the presence of combustible materials (e.g., forest matter). Recent wildfires in the United States (US) have led to considerable economic losses and social stresses (Brown et al. 2021). Moreover, there is concern that climate change may increase the intensity, duration, and frequency of wildfires (Abatzoglou and Williams 2016; Wuebbles et al. 2017; Brown et al. 2021). Wildfire prediction is a critical component of wildfire management because it impacts resource distribution, mitigation of adverse effects, and recovery efforts. It is thus necessary to develop resilient statistical methods that can predict extreme wildfire events reliably over space and time.

From a probabilistic viewpoint, we can see wildfire occurrences and sizes as random spatiotemporal processes. Thus, it is crucial to understand their spatiotemporal distributions and the underlying risk factors. In particular, we can link wildfires to their spatial coordinates, e.g., the location of the fire origin, the center of a burnt area, their temporal instant, and other relevant covariates. Extreme wildfires are attracting considerable attention because they are increasingly hazardous and often considered a severe threat to ecosystems.

Multiple statistical approaches have been proposed in the literature to predict the counts and sizes of wildfires using univariate probability models (Cumming 2001; Preisler et al. 2004; Preisler and Westerling 2007; Preisler and Ager 2013; Pereira and Turkman 2019). Xi et al. (2019), in particular, made multiple important contributions to modeling fire risk components over recent decades, describing some key yet often overlooked fire characteristics. They highlighted various recent research areas that may enhance fire-risk assessment models. However, their model disregarded spatial/temporal dependence between wildfires, which is crucial for developing an effective disaster management policy. Several papers proposed using the so-called K and L functions for treating fire occurrences as spatial point-pattern datasets (Genton et al. 2006; Hering et al. 2009; Juan et al. 2012). Serra et al. (2012) and Bivand et al. (2015) treated such point-pattern datasets as gridded spatial datasets, in which the authors modeled the counts of fire occurrences (CNT) in each grid cell as spatially dependent count data. Furthermore, they modeled the spatial dependence in terms of a Gaussian process (GP), which is the most common tool for modeling spatial dependence because of its attractive theoretical and computational properties (Gelfand and Schliep 2016). A log-Gaussian Cox process (Møller et al. 1998), or LGCP in short, is a doubly stochastic construction consisting of a Poisson point process at the data level, with a random log-intensity modeled by a GP. For spatiotemporal point pattern data, LGCP models have been used in different applications; Diggle et al. (2013) reviewed the literature on LGCP. For fire occurrence modeling,
LGCP have been used by several authors, including Serra et al. (2012), Møller and Díaz-Avalos (2010), Gabriel et al. (2017). Abdelfatah et al. (2016) and Trucchia et al. (2018) used GPs for modeling burnt area data.

For large spatial datasets on discretized spatial domains, we often replace continuous-space GPs with discrete Gaussian Markov random fields (GMRFs, Rue and Held 2005). GMRFs allow the use of sparse precision matrices, affording faster computations. Many researchers used joint analyses to study fire occurrences and sizes together in large datasets. Rios-Pena et al. (2018) proposed a zero-inflated beta distribution where they used zero-inflation to model the absence of fires and modeled the fraction of the burnt area using a beta distribution. Furthermore, they assumed spatially-varying model parameters and used GMRF priors in a fully Bayesian analysis. Joseph et al. (2019) compared some common probability distributions for modeling frequencies and sizes of large wildfires to generate a posterior predictive distribution based on finite sample maxima for extreme events. A zero-inflated negative binomial model performed the best for CNT, and a log-normal model for burnt areas (BAs). A marked LGCP model was proposed by Pimont et al. (2021) in which the authors modeled the occurrences using a point process and treated the fire sizes as marks. However, their approach estimated the model components for occurrences and fire sizes separately, which limits the interaction between these two components. More recently, Koh et al. (2021) rectified this method by allowing joint estimation for all components using fully Bayesian inference.

Motivated by the 2021 Extreme Value Analysis conference (EVA 2021) data challenge, where we participated as the team The Bedouins, we propose a statistics and machine learning (ML)-based approach for the spatial prediction of BA and CNT at spatiotemporal locations with masked observations. We build a four-stage, high-dimensional, bivariate spatial model for zero-inflated data using stochastic partial differential equations (SPDEs). In Stage 1, we separate the data into zero/nonzero categories. Moreover, we fit a two-layered hierarchical Bayesian model where the first layer defines the zero/nonzero data as two categories of a real-valued latent process. In the second layer, we construct the latent process using an SPDE. In Stage 2, we first obtain empirical estimates of the spatially-varying mean and variance profiles across the spatial locations for the positive observations and smooth those estimates using fixed rank kriging (FRK, Cressie and Johannesson 2008). In Stage 3, we model the standardized log-transformed positive observations from the second stage using a bivariate spatial GMRF. Despite the computational advantages of GP-based modeling, the model in Stage 3 erroneously assumes that the marginal distribution of log-CNT is Gaussian. Thus, the predicted exceedance probabilities of CNT are post-processed in Stage 4 using Random Forests (RFs), where we treat BA as a covariate. We replace the missing BA values with the predicted BA values in Stage 3. In this study, CNT data are calibrated independently at each spatial location. Here, we use the simulation-based Markov Chain Monte Carlo (MCMC) method to draw a posterior inference in Stages 1 and 3. This method involves only moderate computational time despite the high spatial dimensions because of the sparse spatial precision matrix implied by the SPDE. Our modeling framework focuses on wildfire prediction in the US; however, it is worth noting that we can also adapt it to other data scenarios with large spatial dimensions and zero inflation.
The paper is structured as follows: In Section 2, we discuss an exploratory analysis of the US wildfire dataset. Section 3 shows the development of an approach combining statistics and RFs for the joint modeling of BA and CNT. Section 4 provides a brief overview of the computational details. We apply the proposed approach to the US wildfire dataset and then discuss the results in Section 6. Section 7 concludes and provides some perspectives for future research.

2 The US wildfire dataset and exploratory analysis

In this section, the primary features of the US wildfire dataset are described and an exploratory graphical support for our modeling choices is provided.

2.1 Data description and the missing data pattern

The wildfire dataset for the EVA 2021 data challenge comprises monthly observations at 3503 grid cells across the US Mainland using a spatial resolution of $0.5^\circ \times 0.5^\circ$. The original dataset (before masking) contains aggregated burnt areas (BA) and wildfire counts (CNT) in each pixel over a period of 23 years between (1993–2015) and for seven months per year (March to September). Information about 18 spatiotemporal land cover covariates (e.g., proportion of urban area, shrubland, grassland), 10 spatiotemporal meteorological covariates (e.g., temperature at 2 m above the ground, precipitation, evaporation of water), as well as certain purely spatial and temporal covariates (mean and standard deviation of the altitude, proportion of a pixel that is within the US Mainland, longitude and latitude for the center of the pixel, year, and month) was available. The organizers used the Shuttle Radar Topography Mission (SRTM) database available at a 90–m spatial resolution to calculate altitude-related covariates. A detailed description of the dataset is available at (Opitz 2022). The primary aim of this study was to estimate the predictive distribution function of BA and CNT at 28 severity thresholds as follows:

$$U_{BA} = \{0, 1, 10, 20, 30, \ldots, 100, 150, 200, 250, 300, 400, 500, 1000, 1500, 2000, 5000, 10000, 20000, 30000, 40000, 50000, 100000\},$$

$$U_{CNT} = \{0, 1, 2, \ldots, 9, 10, 12, 14, \ldots, 30, 40, 50, \ldots, 100\}.$$  

The final evaluation of all the registered teams was based on the weighted sum of the squared error between the empirical and predictive distribution functions, with weights given by $\hat{\omega}_{BA}(u) = \hat{\omega}_{CNT}(u) = 1 - (1 + (u + 1)^2/1000)^{-1/4}$, $u \in U_{BA}$ or $U_{CNT}$, then rescaled to add up to 1.

The original dataset does not contain any missing values. However, the organizers masked a total of 80,000 observations (14.18%) across all seven months of the even years (1994, 1996, 1998, 2000, 2002, 2004, 2006, 2008, 2010, 2012, and 2014) to compare the spatial prediction performances of models proposed by participating teams. We shall refer to locations where observations have been masked simply as “masked locations” in the following, for brevity. These masked spatiotemporal locations are not identical for BA and CNT; see Opitz (2022); furthermore, the data were
masked at pixels irrespective of being near low or high fire-prone regions. This demonstrates the requirement for modeling spatial dependence to borrow information from nearby pixels. Figure 1 shows the spatial maps of $\log(1+BA)$ and $\log(1+CNT)$ in March 1994, where the masked data locations are highlighted in white. The masked observations were from small and large clusters of pixels and from regions in the southeast of the US, where high CNT values are observed close to masked pixels. Table 1 shows the proportions of masked observations of BA and CNT. As reported, either BA or CNT data or both are missing in 14.18% of total cases. They are jointly missing in 8.68% of cases. Thus, BA and CNT information can be borrowed for the rest of 5.50% observations where one of variables is available, which leads to smaller standard errors of model parameters.

2.2 Zero inflation

The high proportion of zero observations in the US wildfire dataset is one of its important characteristics; BA being zero/nonzero at a spatiotemporal location is equivalent to CNT being zero/nonzero. Table 1 lists the proportions of zero and nonzero values for BA and CNT. Using this equivalence, some of the missing data can be retrieved. CNT is zero in 3.34% (out of the 14.18%) of the cases where BA has missing value; hence, the BA values are zeros in these instances. Similarly, BAs are zero in 3.32% (out of the 14.18%) of the cases where CNT has missing values; hence, CNT values in these cases are zero. Thus, by filling these masked locations with zeros, a dataset with a smaller number of missing values can be obtained (specifically, 48,947 cases instead of 80,000 cases). Here, out of the final available

| CNT   | BA   | Zero | Nonzero | Missing | Total |
|-------|------|------|--------|--------|-------|
| Zero  |      | 49.60% | 0.00%  | 3.34%  | 52.94%|
| Nonzero| 0.00% | 30.70% | 2.17%  |        | 32.87%|
| Missing| 3.32% | 2.18%  | 8.68%  |        | 14.18%|
| Total | 52.93%| 32.89% | 14.18% |        | 100.00%|

Table 1 The proportions of zero, nonzero, and missing values of BA and CNT in the total 563,983 observations
observations, 61.61% of the values are zeros. Therefore, the available information can be divided into two parts. In the first part, a spatiotemporal dataset of binary observations indicating whether BA/CNT is zero or not, is obtained; in the second part, only the positive values are maintained (treating zeros as missing data). For modeling positive BA and CNT, the logarithmic transformation can be used, and we call the resulting transformed dataset log-BA and log-CNT, respectively.

We now explore the covariate effects on the zero/nonzero indicators. The results of a probit regression model demonstrated significance for certain examined covariates at a significance level of 0.01. However, assuming a single regression coefficient for the entire spatiotemporal domain is not realistic. A cross-validation study does not demonstrate any significant improvement in the prediction performance after incorporating the available covariate information, in addition to increasing the computational burden. Thus, to simplify calculations we choose to ignore the covariate information at this stage. Similarly, the spatiotemporal covariates can be ignored in the modeling of log-BA and log-CNT, unless they do not increase significantly the computational burden.

We then explore the spatial and temporal correlation profiles of the zero/nonzero indicators, log-BA, and log-CNT, as well as the cross-correlation between log-BA and log-CNT. The left panel of Fig. 2 shows (smoothed) empirical spatial correlation profiles as a function of distance. All three spatial correlation profiles decrease towards zero with the increase in geographical distance between pixels, and they demonstrate similar spatial range and small-scale variability (nugget effect). Thus, a separable correlation structure can be reasonably assumed for multivariate spatial modeling of log-BA and log-CNT. For every spatial location, the lag-1 temporal autocorrelation was empirically calculated and was reported to be not significant at a significance level 0.01 for a large proportion of spatial locations (64.13% locations for the zero/nonzero indicators, 96.39% locations for log-BA, 89.46% locations for log-CNT). A cross-validation study does not demonstrate any significant improvement in the prediction performance by incorporating temporal dependence, where we assume a separable correlation structure across space and time, constructed from a Matérn SPDE spatial field and an AR(1) process in time. Thus, ignoring the temporal autocorrelation for all three spatiotemporal

![Spatial correlation and cross-correlation profiles](image)

**Fig. 2** Left: empirical spatial correlation (smoothed) profiles of the zero/nonzero indicators, log-BA, and log-CNT. Right: histogram of the empirical cross-correlation between log-BA and log-CNT across space.
processes is reasonable, particularly considering the additional computational burden in a high spatial dimension. The cross-correlation between log-BA and log-CNT is empirically calculated at every spatial location (based on the temporal replicates), and the right panel of Fig. 2 shows the histogram of empirical cross-correlation values. The cross-correlation is quite high for most spatial locations, which indicates the requirement for joint statistical modeling of log-BA and log-CNT.

Finally, we explore the requirement for spatially-varying marginal distribution parameters for log-BA and log-CNT. For modeling log-BA and log-CNT, we use a probability distribution from a location-scale family with a spatially-varying location profile, which is more common than allowing the scale to vary spatially for computational reasons. Figure 3 shows the empirical location-wise standard deviations of log-BA and log-CNT. For log-BA, standard deviations are generally lower in Eastern US than in Western US. For log-CNT, standard deviations are lower in the middle regions of US compared to the states closer to the Atlantic or the Pacific coasts. A similar spatial pattern is observed for the location-wise mean values of log-BA and log-CNT. Thus, a joint analysis of log-BA and log-CNT using a model with spatially-varying location and scale parameters is required.

3 Joint modeling of burnt area (BA) and counts of fire occurrences (CNT)

In this section, a four-stage model based on statistics and machine learning for the joint analysis of BA and CNT is described. As reported in Section 2, the available information is divided into two parts. In the first part, a spatiotemporal dataset of binary observations is obtained to determine whether BA/CNT values are zero; in the second part, only the positive values are maintained, and log-BA and log-CNT are modeled on the logarithmic scale. In Stage 1, a sparse latent GP model for binary spatial data is proposed. For the other three stages, the modeling of log-BA and log-CNT using a combination of an approximate Bayesian inference technique and Random Forests (RF) is described.
3.1 Stage 1: A sparse latent Gaussian process model for wildfire occurrence data

The wildfire occurrence data are assumed independent and identically distributed (IID) across months and years; furthermore the observations across the US mainland for each month are assumed to only be spatially dependent. Thus, by ignoring the month–year combinations, a generic notation is used for temporal replicates.

For a spatial location \( s_i \) and time \( t \), the BA and CNT values are denoted \( BA_i(s_i) \) and \( CNT_i(s_i) \), respectively, where \( i \in \{1, \ldots, N\} \) and \( t \in \{1, \ldots, T\} \). \( N \) is the total number of pixels (\( N = 3503 \)), and \( T \) is the total number of months (\( T = 161 \)). The wildfire occurrence indicator \( Z_t(s_i) \) at location \( s_i \) and time \( t \) is defined as

\[
Z_t(s_i) = \begin{cases} 
1, & \text{if} \, BA_i(s_i) > 0, \, CNT_i(s_i) > 0, \, \text{or} \, CNT_i(s_i) \text{ is missing, or} \, BA_i(s_i) \text{ is missing,} \\
0, & \text{if} \, BA_i(s_i) = 0, \, CNT_i(s_i) = 0, \, \text{or} \, CNT_i(s_i) \text{ is missing, or} \, BA_i(s_i) \text{ is missing,} \\
\text{NA, if} \, BA_i(s_i) \text{ and } CNT_i(s_i) \text{ are both missing.}
\end{cases}
\]

By assumption the replicated indicator processes, \( Z_i(\cdot), t = 1, \ldots, T \), are IID across time \( t \), and we model \( Z_i(\cdot) \) as

\[
Z_i(s_i) = \begin{cases} 
1, & \text{if } X_i(s_i) > 0, \quad \text{where } X_i(s_i) = \mu_i(s_i) + \epsilon_i(s_i), \\
0, & \text{if } X_i(s_i) < 0,
\end{cases}
\]

and \( \epsilon_i(\cdot), t = 1, \ldots, T \), are IID spatial GPs, in which \( \text{E}[\epsilon_i(s_i)] = 0 \) and \( \text{Var}[\epsilon_i(s_i)] = 1 \) for all \( i \in \{1, \ldots, N\} \). The process \( \epsilon_i(\cdot) \) is assumed to follow an isotropic Matérn spatial correlation (with nugget effect) given by

\[
\rho_\epsilon(s_i, s_j) = \frac{r_\epsilon}{\Gamma(\nu)2^{-\nu-1}} \left( \frac{d(s_i, s_j)}{\phi_\epsilon} \right)^\nu K_\nu \left( \frac{d(s_i, s_j)}{\phi_\epsilon} \right) + (1 - r_\epsilon) \delta(s_i = s_j),
\]

where \( d(s_i, s_j) \) is the Euclidean distance between \( s_i \) and \( s_j \), \( \phi_\epsilon > 0, \nu > 0 \) and \( r_\epsilon \in [0, 1] \) are the range, smoothness, and ratio of the spatial to total variation, respectively. In (3), \( K_\nu \) is the modified Bessel function of the degree \( \nu \), and \( \delta(s_i = s_j) = 1 \) if \( s_i = s_j \), and 0 otherwise. When \( r_\epsilon = 1 \), \( \epsilon_i(\cdot) \) is mean-square differentiable if \( \nu \) is an integer. For practical applications, identifying \( \nu \) is difficult, and thus, it is generally fixed a priori. Here, the process \( \epsilon_i(\cdot) \) is not observable and hence estimating \( \nu \) is more challenging. Therefore, we here set \( \nu \) to one. To overcome the high computational burden due to the large spatial dimension, \( \epsilon_i(\cdot) \) is defined as a Gaussian Markov random field (GMRF) that has a spatial covariance structure (approximately) equivalent to a dense spatial Matérn correlation \( \rho_\epsilon(\cdot, \cdot) \). As described in Lindgren et al. (2011) the equivalence is derived from the one-to-one link between dense isotropic Matérn GPs and GMRFs. We next briefly summarize this link.
Suppose that $\mathbf{e}(\cdot)$ is a dense GP with correlation structure (3) and $r_\varepsilon = 1$. Then, $\mathbf{e}(\cdot)$ is the solution to the SPDE $(8\phi_\varepsilon^{-2} - \Delta)\mathbf{e}(s) = \mathbf{W}(s)$, where $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is the Laplacian operator, $(8\phi_\varepsilon^{-2} - \Delta)$ is a differential operator, and $\mathbf{W}(s)$ is a standard Gaussian white noise. We can solve this SPDE using finite element methods (Lindgren et al. 2011) over a triangular mesh in $\mathbb{R}^2$, where the triangles are formed following a Delaunay triangulation. Let the set of mesh nodes be denoted by $S^* = \{s_1^*, \ldots, s_N^*\}$. We construct a finite element representation of the solution to $\mathbf{e}(s) = \sum_{j=1}^{N^*} \mathbf{w}_j(s) e_j^*$ for some chosen basis functions $\mathbf{w}_j(\cdot)$ and normally distributed weights $e_j^*$ defined at the mesh nodes $S^*$. We calculate the inner products $\langle \mathbf{w}_j(\cdot), 1 \rangle$ and $\langle \nabla \mathbf{w}_j(\cdot), \nabla \mathbf{w}_j(\cdot) \rangle$, where $\langle f, g \rangle = \int f(s) g(s) ds$, and obtain three $(N^* \times N^*)$-dimensional finite element matrices $\mathbf{C}$, $\mathbf{G}_1$, and $\mathbf{G}_2$. Here, $\mathbf{C}$ is a diagonal matrix, of which the $(j,j)^{th}$ entry $C_{jj} = \langle \mathbf{w}_j(\cdot), 1 \rangle$, $\mathbf{G}_1$ is a sparse matrix, of which the $(j_1,j_2)^{th}$ entry $G_{1_{j_1,j_2}} = \langle \nabla \mathbf{w}_{j_1}(\cdot), \nabla \mathbf{w}_{j_2}(\cdot) \rangle$, and $\mathbf{G}_2 = G_{1}^{-1} G_{1}$. Further theoretical details are discussed in Bakka et al. (2018). We have that $\mathbf{e}^* = [e_1^*, \ldots, e_{N^*}]' \sim \text{Normal}_{N^*}(\mathbf{0}, \mathbf{Q}_{\phi_\varepsilon})$, where the precision matrix is $\mathbf{Q}_{\phi_\varepsilon} = (4\pi)^{-1} \mathbf{\phi}_\varepsilon^{-4} \mathbf{C} + 2\phi_\varepsilon^{-2} \mathbf{G}_1 + \mathbf{G}_2$. To project $\mathbf{e}^*$ back to the data locations $S$, we evaluate $a_{ij} = \mathbf{w}_j(s_i)$ for each location $s_i$ and mesh node $s^*_i$. The $(N \times N^*)$-dimensional matrix $\mathbf{A}$, of which the $(i,j)^{th}$ entry is $a_{ij}$, is called the SPDE projection matrix, and $\mathbf{A}\mathbf{e}^* \sim \text{Normal}_{N}(\mathbf{0}, \mathbf{AQ}_{\phi_\varepsilon}^{-1} \mathbf{A}^*)$. The covariance matrix $\mathbf{AQ}_{\phi_\varepsilon}^{-1} \mathbf{A}^*$ approximates the Matérn correlation matrix obtained by evaluating (3) at $S$ (for $r_\varepsilon = 1$).

Suppose that for each $t \in \{1, \ldots, T\}$, $\mathbf{e}_t^*$ is an independent copy of $\mathbf{e}^*$. We construct a vector $\mathbf{e}_t = [e_t(s_1), \ldots, e_t(s_N)]'$ from $\mathbf{e}_t^*$ as

$$
\mathbf{e}_t = \sqrt{r_\varepsilon \mathbf{A}\mathbf{e}_t^*} + \sqrt{1 - r_\varepsilon} \tilde{\mathbf{e}}_t,
$$

where $\tilde{\mathbf{e}}_t = [\tilde{e}_t(s_1), \ldots, \tilde{e}_t(s_N)]'$ with $\tilde{e}_t(s_i) \overset{\text{iid}}{\sim} \text{Normal}(0,1)$. The final covariance matrix of $\mathbf{e}_t$ is $\mathbf{\Sigma}_t = r_\varepsilon \mathbf{AQ}_{\phi_\varepsilon}^{-1} \mathbf{A}^* + (1 - r_\varepsilon) \mathbf{I}_N$, and it approximates the Matérn correlation matrix obtained by evaluating (3) at $S$, for any $r_\varepsilon \in [0, 1]$. We then discuss the GMRF approximation accuracy in our data application, for the values of $\phi_\varepsilon$ and $r_\varepsilon$ set to 3 and 0.8 respectively. These values are similar to estimates obtained in our wildfire data analysis; here, there are $N^* = 1027$ mesh nodes. In this SPDE mesh, we calculate the covariance between every pair of spatial locations (the elements of the matrix $\mathbf{\Sigma}_t$) and the corresponding true Matérn correlation based on (3). The results of these calculations are presented as a function of distance in the right panel of Fig. 4. The true correlation structure of dense GP is well approximated by the corresponding GMRF. Moreover, the sparsity of $\mathbf{Q}_{\phi_\varepsilon}$ can be exploited to allow quick computations. Moreover, while categorizing the latent process given in (2), the conditional spatial independence structure $\mathbf{e}_t|\mathbf{e}_t^* \sim \text{Normal}_{N}(\sqrt{r_\varepsilon \mathbf{A}\mathbf{e}_t^*}, (1 - r_\varepsilon) \mathbf{I}_N)$ allows the univariate imputation of latent variables, which is exploited for spatial prediction.

After marginalization with respect to $X_t(s_i)$, we obtain that $\text{Pr}(Z_t(s_i) = 1) \approx \Phi(\mu_Z(s_i))$, where $\Phi(\cdot)$ is the standard normal distribution function. The conditional distribution of $Z_t(s_i)$ given $\mathbf{e}_t^*$ is $\text{Pr}(Z_t(s_i) = 1 | \mathbf{e}_t^*) = \Phi((\mu_Z(s_i) + \sqrt{r_\varepsilon} a_i e_t^*) / \sqrt{1 - r_\varepsilon})$, where $a_i$ is the $i^{th}$ row of $\mathbf{A}$, and $\text{Pr}(Z_t(s_i) = 0 | \mathbf{e}_t^*) = 1 - \text{Pr}(Z_t(s_i) = 1 | \mathbf{e}_t^*)$. The variables $Z_t(s_i)$ and
$Z_t(s_j)$ are conditionally independent given $\epsilon^*_t$. There is no closed form expression for the joint distribution of $Z_t(s_i)$ and $Z_t(s_j)$, after marginalizing with respect to $X_t(s_i)$.

### 3.2 Stage 2: Approximate Bayesian inference to smooth spatially-varying parameters

Despite the discreteness of CNTs, we proceed by modeling both log-BA and log-CNT using a GP with spatially-varying location and scale terms, because of the computational attractiveness of GPs, as follows

$$
\begin{align*}
\log -BA_t(s_i) &= \mu_1(s_i) + \sigma_1(s_i)W_{1t}(s_i), \\
\log -CNT_t(s_i) &= \mu_2(s_i) + \sigma_2(s_i)W_{2t}(s_i),
\end{align*}
$$

where $W_t(s_i) = [W_{1t}(s_i), W_{2t}(s_i)]^T$ is a bivariate standard GP (zero mean and unit variance for all marginal distributions). A fully Bayesian inference is computationally challenging, and thus, in Stage 2, we focus on estimating the parameter surfaces $\mu_1(\cdot)$, $\sigma_1(\cdot)$, $\mu_2(\cdot)$, and $\sigma_2(\cdot)$ only, while the parameters within the bivariate spatial GP $W_t(\cdot)$ are treated as nuisance parameters. Here, the procedure for $\mu_1(\cdot)$, which is the same procedure used to estimate the surfaces $\log[\sigma_1(\cdot)]$, $\mu_2(\cdot)$, and $\log[\sigma_2(\cdot)]$ is detailed; all surfaces evaluated at any $s_i$ can take values over the whole real line.

An approximate Bayesian inference scheme, similar to Max-and-Smooth (Hrafnkelsson et al. 2021; Johannesson et al. 2022; Hazra et al. 2021), is applied to obtain parameter surfaces in two steps. In the first step, we estimate $\mu_1(s_i)$, $\log[\sigma_1(s_i)]$, $\mu_2(s_i)$, and $\log[\sigma_2(s_i)]$ at each $s_i$ separately, using location-wise empirical means and standard deviations, ignoring any spatial/temporal trend or dependence. We denote the estimates by $\hat{\mu}_1(s_i)$, $\log[\hat{\sigma}_1(s_i)]$, $\hat{\mu}_2(s_i)$, and $\log[\hat{\sigma}_2(s_i)]$.

In the second step we smooth parameter surfaces by treating the preliminary estimates as noisy measurements of the true underlying parameters. Specifically, for the parameter surface $\mu_1(\cdot)$, we assume that $\hat{\mu}_1(s_i) = \mu_1(s_i) + e(s_i)$, where $\mu_1(s_i)$ is the true parameter value $e(s_i) \sim \text{Normal}(0, \sigma^2_\epsilon)$ is Gaussian white noise (measurement error term). Because of the large spatial dimension, the prior for $\mu_1(s_i)$ is assumed to follow a low-rank structure as
\[ \mu_1(s_i) = \beta_0 + \beta_1 \text{lon}(s_i) + \beta_2 \text{lat}(s_i) + \sum_{r=1}^{3} \sum_{k=1}^{K_r} h_{rk}(s_i) \omega_{rk}^* + \xi(s_i), \quad s_i \in S, \quad (6) \]

where \( h_{rk}(-) \) is the \( k \)th spatial Gaussian kernel at the \( r \)th resolution, the spatial random effects are \( \omega^* = [\omega_{r1}^*, \ldots, \omega_{rk}^*] \overset{\text{Indep}}{\sim} \text{Normal}(K_r, \Sigma(\theta_r)) \), and \( \xi(s) \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2_\xi) \). By selecting flat priors for the spatially-invariant hyperparameters \( \beta_0, \beta_1, \beta_2, \theta_r, \) and \( \sigma^2_\xi \), the estimation of \( \mu_1(s_i), s_i \in S \), lies in the setting of fixed rank kriging (FRK), which is readily implemented using the \( \mathbb{R} \) package FRK (Zammit-Mangion and Cressie 2021). For certain pixels, there was no available positive observation. In these instances \( \tilde{\mu}_1(s_i) \) is treated as missing and the corresponding values of \( \mu_1(s_i) \) are predicted based on the available first-step estimates. We repeat the same procedure for other parameter surfaces and obtain smoothed estimates, denoted by \( \tilde{\mu}_1(\cdot), \tilde{\sigma}_1(\cdot), \tilde{\mu}_2(\cdot), \) and \( \tilde{\sigma}_2(\cdot) \). Finally, we also obtain estimates of \( W_t(s_i), s_i \in S, t = 1, \ldots, T \), by plugging the smoothed estimates into (5).

### 3.3 Stage 3: Bivariate spatial modeling of standardized log-BA and log-CNT

In this stage, we model the standardized variables \( \hat{W}_{t_1}(s_i) = \{ \log \text{-BA}(s_i) - \bar{\mu}_1(s_i) \} / \tilde{\sigma}_1(s_i) \) and \( \hat{W}_{t_2}(s_i) = \{ \log \text{-CNT}(s_i) - \tilde{\mu}_2(s_i) \} / \tilde{\sigma}_2(s_i) \), obtained in Stage 2 using (5). Suppose that \( \hat{W}_{p} = [\hat{W}_{t_1}(s_1), \ldots, \hat{W}_{t_2}(s_N)]' \) for \( p = 1, 2 \), and \( \hat{W}_t = [\hat{W}_{t_1}, \hat{W}_{t_2}]' \). We model \( W_t \) as \( \hat{W}_t = [I_2 \otimes A] \eta_t^* + \tilde{\eta}_t \), where

\[ \eta_t^* \sim \text{Normal}_{2N}(0_{2N}, \rho_\eta \begin{pmatrix} 1 & \rho_\eta \\ \rho_\eta & 1 \end{pmatrix} \otimes Q_{\phi_\eta}^{-1}), \quad (7) \]

\[ \tilde{\eta}_t \sim \text{Normal}_{2N}(0_{2N}, (1 - \rho_\eta) \begin{pmatrix} 1 & \rho_\eta \\ \rho_\eta & 1 \end{pmatrix} \otimes I_N). \]

Similarly to Stage 1, we model the processes \( \hat{W}_{ip}(\cdot) \) using GMRFs. The same SPDE mesh and SPDE projection matrix as those used in Stage 1 are used. The spatial correlation of each component \( \hat{W}_{ip}(\cdot) \) is approximately equal to (3), with \( \phi_\eta \) and \( r_\eta \) replaced by \( \phi_\eta \) and \( r_\eta \), respectively. This is confirmed by similar empirical spatial correlation profiles to those of log-BA and log-CNT shown in Fig. 2. The marginal standard deviation of each \( \hat{W}_{ip}(s_i) \) is approximately equal to one (as described in Stage 1); thus, the cross-covariance between \( \hat{W}_{t_1}(s_i) \) and \( \hat{W}_{t_2}(s_i) \) is approximately equal to the cross-correlation \( \rho_\eta \in [-1, 1] \) for each \( s_i \) and \( t \). To summarize, a separable spatiotemporal bivariate GMRF, where the marginal distributions have zero mean and approximate unit variance, is defined.

At a spatiotemporal prediction location \( (s_i, t) \), we need to simulate the missing process \( \hat{W}_{ip}(s_i) \), possibly for both \( p = 1, 2 \). Let the elements of 2N-dimensional vector \([I_2 \otimes A] \eta_t^* \) be denoted by \([\tilde{\eta}_{t1}^*, \tilde{\eta}_{t2}^*]' \), where \( \tilde{\eta}_{t1}^* = [\tilde{\eta}_{t1}^*(s_1), \ldots, \tilde{\eta}_{t1}^*(s_N)]' \) and \( \tilde{\eta}_{t2}^* = [\tilde{\eta}_{t2}^*(s_1), \ldots, \tilde{\eta}_{t2}^*(s_N)]' \). Then,
\[ \hat{W}_t(s_i) \sim \text{Normal}_2 \left( \begin{pmatrix} \tilde{\eta}_{1t}^*(s_i) \\ \tilde{\eta}_{2t}^*(s_i) \end{pmatrix}, (1 - r_\eta) \begin{pmatrix} 1 & 0 \\ 0 & \rho_\eta \end{pmatrix} \right). \]

The predicted values are denoted by \( \hat{W}_{t\eta}(s_i), \ p = 1, 2 \). Then, the predicted values of log-BA and log-CNT are obtained by plugging \( \hat{\mu}_1(\cdot), \hat{\sigma}_1(\cdot), \hat{\mu}_2(\cdot), \) and \( \hat{\sigma}_2(\cdot) \) (obtained in Stage 2) and \( \hat{W}_{t\eta}(s_i) \) (obtained in Stage 3) into (5). The distribution functions of \( \text{BA} \) and \( \text{CNT} \) at \( (s_i, t) \) are approximately Lognormal(\( \mu_1(s_i), \sigma_1^2(s_i) \)) and Lognormal(\( \mu_2(s_i), \sigma_2^2(s_i) \)), denoted by \( F_{\text{LN}}(\cdot ; \mu_1(s_i), \sigma_1^2(s_i)) \) and \( F_{\text{LN}}(\cdot ; \mu_2(s_i), \sigma_2^2(s_i)) \), respectively. Furthermore, by incorporating zero inflation, \( F_{\text{BA}} \) and \( F_{\text{CNT}} \), the distribution functions of \( \text{BA} \) and \( \text{CNT} \) at \( (s_i, t) \), respectively, are (up to the SPDE approximation)

\[
F_{\text{BA}}(x; \mu_2(s_i), \mu_1(s_i), \sigma_1^2(s_i)) \approx 1 - \Phi(\mu_2(s_i)) + \Phi(\mu_2(s_i))F_{\text{LN}}(x; \mu_1(s_i), \sigma_1^2(s_i)),
\]

\[
F_{\text{CNT}}(x; \mu_2(s_i), \mu_2(s_i), \sigma_2^2(s_i)) \approx 1 - \Phi(\mu_2(s_i)) + \Phi(\mu_2(s_i))F_{\text{LN}}(x; \mu_2(s_i), \sigma_2^2(s_i)), x \geq 0.
\]

### 3.4 Stage 4: Random Forests (RF)

Until Stage 3, we considered a fairly simple statistical approach to jointly model \( \text{BA} \) and \( \text{CNT} \). In particular, in our joint modeling approach, the stage-specific models did not incorporate any covariate information (except in certain priors that are discussed later). As reported in Section 2, incorporating covariates in the statistical modeling framework and allowing for spatially-varying regression coefficients is computationally challenging and does not significantly improve the prediction performance. Moreover, assuming that the marginal distribution of log-CNT is Gaussian is computationally beneficial but crude due to the discrete nature of CNT. Thus, in Stage 4, we rectify the predicted values of CNT obtained from Stages 1 through 3 using Random Forests (RFs). Typically, machine learning (ML) approaches (such as RF) do not have any distributional assumptions and involve a moderate and controlled computational burden, while allowing a straightforward incorporation of covariate information.

For interpretation and prediction purposes, it is important to incorporate meaningful covariates, such as some landcover types and climate/weather conditions, because they play a key role in the occurrence of wildfires, as shown in several studies (Fusco et al. 2019; Nadeem et al. 2020). These covariates are interrelated and show seasonal dependence, but classical statistical methods, including generalized linear models, cannot account for the sophistication in this process. Jain et al. (2020) reviewed extensive literature (until 2019) on ML application in wildfire science and management, in which they reported RF to be a more common method for predicting fire occurrences since 2012, before models based on artificial neural networks (NNs) and support vector machines (SVMs) were widely reported in the literature. It is reasonable to use RF in the rectification of the US wildfire data analysis, because tree-based models are often suitable for classification problems, hence they are suitable for the
discrete CNT data. Moreover, RFs generate nonlinear regression models, which facilitate correct formulations of covariates.

The RF algorithm (Breiman 2001) involves an ensemble of many decision trees, in which individual trees are trained based on a random subset of the data, and are drawn with replacement. A random subset (resampled with replacement) of covariates is selected at each node of every decision tree. For classification problems, each individual tree forms a class; furthermore, the predicted class is selected by the highest votes. The success of RF is attributed to the low correlation among trees (each tree is trained independently from others); thus, it results in low prediction variance. Importantly, RF minimize the correlation between trees, hence they tend to provide higher accuracy than models based on individual trees, which explains the success of classification and regression trees (CART).

For the EVA 2021 data challenge, the prediction performance is evaluated based on score functions with high weights assigned to the data categorized into higher severity levels as described in Section 2. Considering the size of the dataset and underlying computational cost, this specific task (i.e., that of the challenge) was converted into a classification problem, in which CNT was classified into 29 categories, (based on the 28 severity thresholds provided for model evaluation in 1) and labeled at each spatiotemporal location. Then CNT is converted into categorical data, which inevitably decreased in resolution and essentially did not affect the ability to achieve high prediction accuracy. Moreover, BA is here included as a covariate, in which the missing values of BA are imputed by the joint statistical modeling of BA and CNT.

4 Computation

4.1 Computational details for Stages 1 and 3

Inferential statistical analysis is conducted on the model parameters mentioned in Sections 3.1 and 3.3 based on MCMC sampling. Conjugate priors are selected whenever possible. The full posterior distributions of the model parameters and hyperparameters are provided in the Supplementary Materials. We here briefly outline the MCMC steps. In Stage 2, a fixed rank kriging model was directly fitted using the R package FRK; hence, this is skipped here.

In Stage 1, the parameters and hyperparameters are $\Theta_1 = \{(X_t)_{t=1}^T, \mu_Z, \theta_\mu, \tau_\mu, (\varepsilon_t^*)_{t=1}^T, \phi_t, r_s\}$. By some abuse of notation, we reparametrize $\sqrt{\tau_s} \varepsilon_t^*$ by $\varepsilon_t^*$. The full posterior distribution of the latent variables $X_t(s_i)$ depends on $Z_t(s_i)$. If $Z_t(s_i)$ is missing, the posterior distribution of $X_t(s_i)$ is normal; otherwise, if $Z_t(s_i)$ is zero or one, the posterior of $X_t(s_i)$ is a truncated normal distribution, supported on the negative or positive side of the real line, respectively.

The prior distribution we choose for $\mu_Z$ is $\mu_Z \sim \text{Normal}_{N}(D\theta_\mu, \tau_\mu^{-1}I_N)$. Here, $D$ is a $(N \times 6)$-dimensional design matrix with its columns representing an intercept term, longitude, latitude, mean altitude, standard deviation of altitude, and the proportion of a pixel that is within the US Mainland. Because the aim of this study is to predict the underlying spatial process at a new set of locations, it is reasonable to assume an unstructured covariance for the prior of $\mu_Z$. For the hyperparameters $\theta_\mu$ and $\tau_\mu$, we choose weakly-informative
conjugate priors $\theta_\mu \sim \text{Normal}_d(0, 10^2I_d)$ and $\tau_\mu \sim \text{Gamma}(0.1, 0.1)$. The unconditional distribution of $\varepsilon^*_t$ is $\varepsilon^*_t \sim \text{Normal}_N(0, r_{\varepsilon}Q^{-1}_{\phi_\varepsilon})$ and the conditional distribution of $X_t$ given $\varepsilon^*_t$ is $X_t | \varepsilon^*_t \sim \text{Normal}_N(\mu_Z + AE^*_t(1 - r_{\varepsilon}Q_{\theta_\varepsilon}), \sigma^2)$. Thus, the full conditional posterior distribution of $\varepsilon^*_t$ is an $N^{\mu}$-variate normal distribution; furthermore, the calculation of the mean vector and covariance matrix is straightforward. Based on the remaining parameters and hyperparameters, $\varepsilon^*_t; t = 1, \ldots, T$, are conditionally independent and thus are updated in parallel. In case of the parameters $\phi_\varepsilon$ and $r_{\varepsilon}$, any existence of conjugate priors is not known and hence independent priors are selected as $\phi_\varepsilon \sim \text{Uniform}(0, 2\Delta_S)$ and $r_{\varepsilon} \sim \text{Uniform}(0, 1)$, where $\Delta_S$ is the largest Euclidean distance between two data locations. The posterior samples from $\phi_\varepsilon$ and $r_{\varepsilon}$ are drawn using the well-established Metropolis-Hastings algorithm.

In Stage 3, the parameters and latent variables are $\Theta_3 = \{(\eta^*_t)^T, \phi_\eta, r_\eta, \rho_\eta\}$. The unconditional distribution of $\eta^*_t$ is a $2N^{\eta}$-variate normal distribution; furthermore, the conditional distribution of $\eta^*_t$ is a $2N$-variate normal distribution, (the detailed expressions are presented in Section 3.3). The calculation of the full conditional posterior distribution of $\eta^*_t$ is straightforward and it is again a $2N^{\eta}$-variate normal distribution. When the remaining parameters and hyperparameters, $\eta^*_t; t = 1, \ldots, T$, are conditionally independent and thus are updated in parallel. For the parameters $\phi_\eta, r_\eta$, and $\rho_\eta$, any existence of conjugate priors is not known and we choose independent priors $\phi_\eta \sim \text{Uniform}(0, 2\Delta_S)$, $r_\eta \sim \text{Uniform}(0, 1)$, and $\rho_\eta \sim \text{Uniform}(0, 1)$. Posterior samples from $\phi_\eta$, $r_\eta$, and $\rho_\eta$ are drawn using a Metropolis-Hastings algorithm.

Each MCMC chain was run for 60,000 iterations with the first 10,000 iterations discarded as burn-in. The post-burn-in samples were then thinned keeping every fifth sample. Convergence of the chains was monitored via trace plots. The computation of Stage 1 and Stage 3 was undertaken on two desktop computers with Intel Xeon CPUs E5-2680, 2.40GHz processor and 128GB RAM, and the corresponding computational times for Stage 1 and Stage 3 were 447 minutes and 744 minutes, respectively. These two stages can be run in parallel.

### 4.2 Computational details for Stage 4

We used the function `randomForest` from the R package `randomForest`, which implements Breiman’s RF algorithm. The constructed RF models are trained and their performances tested using a cross-validation study detailed in Section 6.1. For classification problems, the optimal number of covariates used at each splitting node is $\sqrt{P}$, where $P$ is the total number of covariates. However, multiple models were fitted using different numbers of covariates, and the model with the best prediction performance was selected based on a cross-validation scheme. The running time on a computer with the same configuration described in Section 4.1, was approximately 20 minutes. After some trial-and-error, the tuning parameter configurations were finally chosen as follows: $mtry = 36$ (all available covariates), where $mtry$ is the number of covariates randomly sampled as candidates for each split; and $ntree = 200$, where $ntree$ denotes the number of trees to grow. Note that $ntree$ should not be extremely small to ensure that every input row is predicted a
few times. These two tuning-parameters were selected based on some exploratory experiments and a cross-validation study; see Table 2 for more details.

5 A competing model for CNT: Log-Gaussian Cox processes (LGCPs)

5.1 Modeling through LGCPs

Log-Gaussian Cox processes (LGCPs), or Cox process, are a commonly flexible approach for analyzing point pattern data and are obtained by assuming a hierarchical Bayesian structure. At the first level of this structure (data level), the response at every location is assumed to follow a Poisson process conditioned on its random intensity measure; at the second level (process level), the intensity measure is assumed to follow a log-Gaussian process defined over the spatial domain. Precisely, a spatial point process \( Y(\cdot) \) defined over \( S \subset \mathbb{R}^2 \) is called a Cox process if, for any subregion \( A \subset S \), counts \( Y(A) \) are Poisson distributed, i.e.,

\[
    Y(A) \sim \text{Poisson}(\int_A \exp[\Lambda(s)] \, ds),
\]

where the log-intensity process \( \{ \Lambda(s); s \in S \} \) is a GP, and with \( Y(A) \) and \( Y(B) \) being independent for any disjoint sets \( A, B \subset S \).

To fit the LGCP model in a tractable manner, a common approach is to divide the whole spatial region in a uniform grid and assume a constant log-intensity in each grid cell. Here, the US wildfire dataset is already gridded, and \( \{ s_i, i = 1, \ldots, N \} \) denote the spatial locations (the centroids of the grid cells) as defined in Section 3.1. The log-intensity vector of the \( t \)-th time point is denoted

\[
    \Lambda_t = [\Lambda_t(s_1), \ldots, \Lambda_t(s_N)]',
\]

where

\[
    \Lambda_t \sim \text{Normal}(X_t \beta, \sigma^2 \Lambda)
\]

is the vector of regression coefficients, \( \sigma^2 \Lambda \) is the common variance term, and \( C_\Lambda \) is the corresponding correlation matrix. For a dense correlation matrix \( C_\Lambda \), the estimation procedure involves a high computational burden and thus, similar to Section 3.1, we use an SPDE approximation-based construction of \( C_\Lambda \). We denote the CNT data at spatial location \( s_i \) and time \( t \) by \( \text{CNT}_t(s_i) \) and propose the following LGCP model:

\[
\begin{align*}
    \text{CNT}_t(s_i) \mid \Lambda_t(s_i) & \overset{\text{Indep}}{\sim} \text{Poisson}(\exp[\Lambda_t(s_i)]), \quad i = 1, \ldots, N, \ t = 1, \ldots, T, \\
    \Lambda_t(s_i) & = \mu_t(s_i) + \zeta_t(s_i),
\end{align*}
\]

where \( \zeta_t(\cdot), t = 1, \ldots, T, \) are IID copies from a GMRF that has an (approximately) equivalent isotropic Matérn spatial correlation given by (3) (see Section 3.1 for more details), with \( \Phi_\epsilon \) and \( r_\epsilon \) replaced by \( \Phi_\tau \) and \( r_\tau \), respectively, and \( \mu_t(s_i) \) denotes the spatiotemporal mean process at spatial location \( s_i \) and time \( t \) and is defined in terms of fixed covariates. Multiple combinations of covariates (i.e., all covariates, only spatial covariates, including/excluding BA) are compared and discussed in Section 6.

5.2 Computational details for fitting LGCPs

Posterior inference is obtained from the LGCP model (8) based on a stochastic gradient-based MCMC method (for more details, see Welling and Teh (2011) and Yadav et al. (2022)). In LGCP models of the form (8), the set of parameters, hyperparameters, and
latent variables is given by $\Theta_{\text{LGCP}} = \left\{ \{ \Lambda_t \}_{t=1}^T, \beta, \{ \xi_t^* \}_{t=1}^T, \phi_\xi, r_\xi \right\}$, where $\beta$ is the vector of regression coefficients associated with the known covariates, and $\left\{ \{ \xi_t^* \}_{t=1}^T, \phi_\xi, r_\xi \right\}$ has a similar interpretation as $\left\{ \{ e_t^* \}_{t=1}^T, \phi_e, r_e \right\}$ in Section 3.1. We use conjugate priors whenever possible, and these parameters are updated using Gibbs sampling. For $\beta$, we choose a weakly-informative conjugate Gaussian prior with mean zero and variance 100 when a small number of covariates are in the model, and for the case of a large number of covariates, we use an informative Gaussian prior with mean 0 and variance equal to 0.1 that allows appropriate penalization. For the Matérn correlation parameters $\phi_\xi$ and $r_\xi$, we use the same priors as for $\phi_e$ and $r_e$, respectively. We update them within MCMC using a Metropolis-Hastings algorithm, similar to updating $\phi_e$ and $r_e$. For the latent vectors $\{ \Lambda_t \}_{t=1}^T$, we do not have closed-from posteriors, and thus we update them jointly using the stochastic gradient Langevin dynamics, which is similar to Algorithm 1 in Yadav et al. (2022). For the latent vectors $\{ \xi_t^* \}_{t=1}^T$, we have closed-form full posteriors, and thus they are updated using Gibbs sampling. For the stochastic gradient MCMC algorithm, we set the batch size to be $b = 10$ (i.e., we update 10 out of the $T$ vectors $\Lambda_t$ at a time). Because the remaining parameters and hyperparameters, $\xi_t^*: t = 1, \ldots, T$, are conditionally independent, we update them in parallel.

The MCMC chains were run for a total of 250,000 iterations, and the first 200,000 samples were discarded as burn-in samples. The chains were thinned by keeping one for each 25 samples. Thus, all the summary statistics were calculated based on the final 2000 samples. The computation time is approximately 37 hours, when only spatial covariates are included in the models, and approximately 110 hours when we use all the covariates (both spatial and spatiotemporal).

6 Data application

6.1 Cross-validation schemes and model comparison

The prediction performance of the four-stage model proposed in Sections 3.1 to 3.4 is compared to a few sub-models, in addition to the alternative LGCP model described in Section 5.1. For comparison, a cross-validation study was performed by dividing the available data after masking into training and test sets. To ensure that the used cross-validation scheme is compatible to that used by the EVA 2021 data challenge organizers, we attempt to replicate the original missingness pattern to create new testing datasets for model selection, and we choose the same metric for model comparison as that used for the data challenge described in Section 2.1.

As reported in Section 2.1, the observations are available for 23 years (1993–2015), from March to September. Thus, data for a total of 161 months were available. Of the 23 years, complete observations for 12 years (84 months) are available, and a significant proportion of the data are missing (NA) for the remaining of 11 years (77 months). Here, two types of test sets were developed by replicating the spatial patterns of the missing observations for 77 months out of 84 months (chosen randomly) with complete data. The first type is called ‘Fixed Month’ scheme, where missingness is created for a month using the pattern from the same month index; e.g., for March 1993...
(which has no missing data), another month of March is randomly selected among the 11 years where certain data are missing, e.g., 1994, and then the missingness pattern for March 1993 is set to be the same as that for March 1994. The second type is called ‘Random Month’ scheme where missingness for a month is created using the pattern from any randomly selected month with certain observations missing. Figure 5 shows the process of generating the new test set for CNT using the principle of the ‘Fixed Month’ scheme. The original test set created by the organizers contains 80,000 spatiotemporal locations, and we masked data at additional 80,000 observations. These two types of cross-validation schemes are primarily selected to confirm the validity of the prediction performance of the proposed models while accounting for seasonality and clustering of masked data in space; these testing set choices have no connection to our temporal independence assumption in our proposed model.

For BA prediction, five competing models, closely linked to the final proposed model, are compared. These include the benchmark, which is a linear regression model for BA on a logarithmic scale, with all covariates including all filled CNT (achieved thanks to a Poisson regression). The spatial prediction performances, of the five models are compared using the same metric employed by the data challenge organizers. For the first competitor, a scenario where there is no data available for CNT, is assumed; thus, the zero/nonzero indicator created in Stage 1 is only based on the BA information. Consequently, the original 80,000 missing observations (rather than 48,947 cases, as mentioned in Section 2.1) in Stage 1 are obtained. Moreover, rather than a bivariate modeling of BA and CNT, a similar univariate SPDE-based spatial model is used for BA only. For other competing models, the

![Fig. 5 A graphical schematic of the cross-validation schemes. Top panel illustrates binary time series (across the X-axis) for the first 50 pixels (across the Y-axis), where data availability and missingness are presented in black and white, respectively. The pixels within two consecutive white dashed lines correspond to seven months within a year. The patterns of the first column of 1994 and the fifth column of 1996, which are highlighted in red and green respectively, are repeated in the latter columns highlighted in the same colors. The bottom panel, shows the binary time series for data availability and missingness, after masking additional 80,000 spatiotemporal cases under the ‘Fixed month’ scheme.](image)
CNT data is assumed to be available. For the second competitor, CNT data are used only for the zero/nonzero indicator part in Stage 1, for which data of only 48,947 spatiotemporal locations are missing. However, for the positive part of BA, the same univariate SPDE-based spatial model is fitted as in the case of the first competitor. For the third competitor, all the 35 spatiotemporal, purely spatial, or purely temporal covariates mentioned in Section 2.1 are used in Stage 1. However, the same univariate SPDE-based model in Stage 3 as for the first two competitors, is fitted. In Stage 1, we replace (2) as follows

\[
Z_t(s_i) = \begin{cases} 
1, & \text{if } X_t(s_i) > 0 \\
0, & \text{if } X_t(s_i) < 0,
\end{cases}
\]

where \(\alpha(\cdot)\) is a spatially-varying intercept term, \(D_{l,t}(s_i), l = 1, \ldots, 35\), are the 35 covariates, \(\gamma_t, l = 1, \ldots, 35\), are the corresponding spatially/temporally invariant regression coefficients, and \(\epsilon_t(s_i)\) are the same as in (2). Here, \(\Pr(Z_t(s_i) = 1) \approx \Phi(\alpha(s_i) + \sum_{l=1}^{35} \gamma_l D_{l,t}(s_i))\).

Table 2 lists the evaluation scores for both cross-validation schemes. Under both schemes, the final model described in Stages 1 to 3 exhibits better performance than the alternative univariate spatial models and the benchmark (Opitz 2022).

In CNT prediction, three types of competing models are selected with different settings for each type, including the benchmark, which is a Poisson regression model with all covariates except BA. The first type is a univariate LGCP model constructed using a latent SPDE, as in (8). In the first setting, it is assumed that the term \(\mu_t(s_i)\) in (8) is constant across space and time. In the second setting, \(\mu_t(s_i)\) is assumed to be only spatially-varying and it is written as a linear combination of purely spatial covariates, including an intercept term (the columns of the design matrix \(\mathbf{D}\) in Section 4.1). In the third setting, \(\mu_t(s_i)\) is assumed to be both spatially and temporally varying; it is written as a linear combination of the purely spatial covariates, as well as BA, where the missing values are filled using Stages 1 to 3. Finally, in the fourth setting, \(\mu_t(s_i)\) is written as a linear combination of all covariates and filled BA data. The second model type is the purely statistical model described in Stages 1 to Stage 3, without incorporating any rectification of the erroneous Gaussian assumption for CNT. The third model type is based on RF, in which all the spatiotemporal covariates, as well as filled BA, are used. Under this setting, we fix \(n_{tree} = 200\) as mentioned in Section 4.2, and set the number of covariates randomly sampled as candidates for each split (\(m_{try}\)) at 3, 6, and 36, under three different settings. The final evaluation scores for the two cross-validation schemes are presented in Table 2. Under both cross-validation schemes, the final model described in Stages 1 to 4 performs better than the alternative univariate LGCP and the bivariate model without the rectification using RFs, as well as the benchmark.

### 6.2 Results

In this section, the posterior means and posterior standard deviations of the model parameters and hyperparameters in Stages 1 and 3 are discussed. The fixed rank kriging estimates of parameter surfaces \(\mu_1(\cdot), \mu_2(\cdot), \sigma_1(\cdot), \) and \(\sigma_2(\cdot)\) in Stage 2, and
the prediction performance of the final model described in Stage 1 through Stage 4 are also reported and discussed.

The mixing and convergence of MCMC chains in Stages 1 and 3 were evaluated using trace plots. For scalar parameters, Fig. 6 shows the trace plots of the thinned MCMC chains. All plots exhibit good mixing and convergence diagnostics. For other parameter vectors and latent variables, the convergence and mixing are also confirmed, though not shown here. The computations in Stages 2 and 4 were conducted using the R packages FRK and randomForest, respectively.

Figure 7 shows the spatial maps of the posterior mean and posterior standard deviation of \( \mu_Z(\cdot) \) in Stage 1. The values of the posterior mean of \( \mu_Z(\cdot) \) are higher near the Southeastern (the state of Georgia) and Southwestern (the state of California) parts of the US and are generally lower in the middle parts of the US. Among available temporal replications, the posterior standard deviation is extremely high (more than 0.4) for 54 spatial locations, mostly in the northeastern parts of the US, where the zero/nonzero indicator is either always zero (for 53 locations) or always one (for 1 location).

Table 3 lists the posterior mean and standard deviation of the non-spatial parameters and hyperparameters in Stages 1 and 3. In a frequentist sense, all parameters \( \theta_{\mu,1} \) to \( \theta_{\mu,6} \) are significant (the absolute value of the ratio of the posterior mean to the
posterior standard deviation is larger than 2 for each of the six cases). The posterior means of \( \phi \) and \( r \) are 3.0491 and 0.5319, respectively, for the latent Gaussian process, which indicates a correlation of 0.050 at a spatial distance of 10 degrees. In Stage 3, the posterior means of \( \phi \) and \( r \) are 3.6408 and 0.3442, respectively, for each component of the bivariate spatial Gaussian process, which indicates a correlation of 0.052 at a spatial distance of 10 degrees. These results show the requirement for modeling the spatial dependence in fire occurrences and sizes. The posterior mean of \( \rho \) is 0.4575, which shows a strong positive correlation between BA and CNT.

Figure 8 shows the fixed rank kriging estimates of \( \mu_1(\cdot) \), \( \mu_2(\cdot) \), \( \sigma_1(\cdot) \), and \( \sigma_2(\cdot) \), in Stage 2. For \( \mu_1(\cdot) \), the FRK estimates are generally higher in a large portion of the southeastern US, while the values are lower in the mid-west, north-west and the north-east. For \( \mu_2(\cdot) \), the values are higher in two small regions in the southeast and southwest of the US. In some regions (mid-north), the estimated profile is not highly smooth indicating high local nonstationarity. For \( \sigma_1(\cdot) \), large estimates are visible in a small portion of the mid-US, while moderately large values are observed in a large region in the western US. For \( \sigma_2(\cdot) \), large values are observed in two large regions of the eastern and western US. In the mid-US, the
Table 3  Posterior mean and standard deviation (SD) of the non-spatial parameters and hyperparameters in Stage 1 and Stage 3. Here, $\theta_{\mu,1}, \ldots, \theta_{\mu,6}$ denote the six components of $\theta_{\mu}$ that represent an intercept term and the regression coefficients of longitude, latitude, mean altitude, standard deviation of altitude, and the proportion of a pixel that is within the mainland US, respectively.

| Parameter | Stage 1 Posterior mean | Stage 1 Posterior SD |
|-----------|------------------------|----------------------|
| $\theta_{\mu,1}$ | $-0.3805$ | $0.0262$ |
| $\theta_{\mu,2}$ | $0.0894$ | $0.0239$ |
| $\theta_{\mu,3}$ | $-0.2596$ | $0.0192$ |
| $\theta_{\mu,4}$ | $-0.2583$ | $0.0280$ |
| $\theta_{\mu,5}$ | $0.4540$ | $0.0289$ |
| $\theta_{\mu,6}$ | $0.1826$ | $0.0154$ |
| $r_{\mu}$ | $1.5398$ | $0.1276$ |
| $\phi_x$ | $3.0491$ | $0.9754$ |
| $r_x$ | $0.5319$ | $0.0194$ |

| Parameter | Stage 3 Posterior mean | Stage 3 Posterior SD |
|-----------|------------------------|----------------------|
| $\phi_\eta$ | $3.6408$ | $1.2179$ |
| $r_\eta$ | $0.3442$ | $0.0042$ |
| $\rho_\eta$ | $0.4575$ | $0.0022$ |

Estimated profile is not smooth in some regions, particularly, near the mid-north of the US, similar to the $\mu_2(\cdot)$ profile. The spatial maps of $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ illustrate the underlying spatial heteroscedasticity and sharp local variability.

Fig. 8  Approximate Bayesian estimates of $\mu_1(\cdot)$, $\mu_2(\cdot)$, $\sigma_1(\cdot)$, and $\sigma_2(\cdot)$ using the fixed rank kriging procedure described in Section 3.2.
After incorporating the Stage 2 estimates, we obtain the residuals \( \hat{W}_{ip}(s_i) \) (Section 3.3). Despite fitting a Gaussian model to log-BA and log-CNT in Stage 3 for computational suitability, we study the histograms of the standardized log-BA and log-CNT values to check the validity of our Gaussian assumption, and we present them in Fig. 9. For the standardized log-BA, the histogram is bell-shaped and symmetric around zero and the Gaussian assumption appears to be reasonable. However, for the standardized log-CNT, the histogram is bimodal and the shapes of the two modes are considerably different. Furthermore, the histogram appears to be right-skewed; thus, a Gaussian assumption for \( \hat{W}_{ip}(s_i) \) in Stage 3 for CNT is questionable and this justifies the rectification using RF in Stage 4.

In Stage 4, the RF algorithm, is used. In this algorithm, the final model based on cross-validation includes all the covariates and BA (missing values are imputed using Stages 1–3), as described in Section 6.1.

![Histograms of the standardizing log-BA and log-CNT based on the Stage 2 estimates](image)

**Fig. 9** Histograms of the standardizing log-BA and log-CNT based on the Stage 2 estimates

![Variable importance (or mean decrease accuracy plot) for the RF algorithm in Stage 4, in which all the available covariates and filled BA were used for classification](image)

**Fig. 10** Variable importance (or mean decrease accuracy plot) for the RF algorithm in Stage 4, in which all the available covariates and filled BA were used for classification
Figure 10 shows the variable importance plot (or mean decrease accuracy plot). The plot expresses how much accuracy the model loses by excluding each variable at a time during classification. The mean decrease in Gini coefficient (Menze et al. 2009) is a measure of how each variable contributes to the homogeneity of nodes and leaves in the resulting RF. The larger the mean reduction in the Gini coefficient, the more important the variable is in the model. Here, BA is reported to be the most important covariate, followed by clim5 (Evaporation) and clim1 (wind speed in Eastern direction), whereas lc6 (tree broadleaved deciduous closed) is the least important covariate. The covariates were included into the models in the order of their variable importance.

Finally, we compare the estimated predictive distribution functions for the 80,000 test observations, for each of BA and CNT, evaluated at 28 levels, with the empirical distribution functions calculated based on the test observations. Figure 11 shows boxplots of the absolute differences between the two cumulative distribution functions (CDFs) at 28 evaluation levels, specified in (1). At lower levels certain boxes and upper endpoints of error bars are significantly different from zero; however, the median of the absolute differences is close to zero in most cases, particularly for BA. For higher levels, which are assigned larger weights for model evaluation, the boxes are close to zero indicating that the proposed model performs well for predicting the distribution of BA and CNT at unobserved spatiotemporal locations.

![Boxplots of absolute differences between the estimated predictive distribution functions and empirical distribution functions evaluated at 28 levels given in (1)](image)
Discussions and conclusions

7.1 Summary

Motivated by the EVA 2021 data challenge, in which the first four authors of this paper participated as the The Bedouins team, a four-stage high-dimensional zero-inflated bivariate spatial model based on statistics and machine learning was proposed for the prediction of BA and CNT at masked spatiotemporal locations. Here, a spatial dependence structure was developed using SPDEs, which reduces the computational burden by allowing sparsity in the precision matrices. In Stage 1, the data were categorized into zero/nonzero categories and a two-layered Bayesian hierarchical model was fitted for estimating the probabilities of the two categories at unobserved sites. In Stage 2, the parameter surfaces were estimated using a two-step approximate Bayesian inference technique that avoids high computational burden.

In Stage 3, the standardized log-transformed positive observations from Stage 2 were modeled using a bivariate spatial GMRF. The log-Gaussian assumption for modeling positive wildfire frequencies was computationally helpful; however, it was erroneous because the observations are discrete-valued. Thus, in Stage 4, the predicted exceedance probabilities of wildfire frequencies were postprocessed using RF, in which BA was treated as a covariate, and the missing BA values were imputed by the predicted BA values in Stage 3. MCMC sampling was used to draw posterior inference in Stages 1 and 3. The computation in Stages 2 and 4 were done directly using the R packages FRK and randomForest, respectively. Our final model was shown to outperform some alternatives in a well-designed cross-validation study and to effectively predict low to high quantiles of BA and CNT at unobserved sites. Table 4 lists the final scores for the four best participating teams. While the teams BlackBox and Kohrrelation focused on purely ML-based approaches, our method only depends on ML algorithms for rectifying predictions given by statistical models. Moreover, it is interesting to note that our method outperformed the team Kung Fu Panda, who used only standard implementations of RFs. This shows that our relatively sophisticated statistical model in Stages 1–3 is worthwhile.

A cross-validation scheme was created to effectively compare the models. Although the univariate spatial model without covariates is nested inside the model with all covariates, we assume that the regression coefficients are constant across space and time. Based on some exploratory analyses (not shown), we think that this might be an oversimplification and that allowing spatially or spatiotemporally varying regression coefficients might improve the prediction performance. However, such a model would

|              | BlackBox | Kohrrelation | The Bedouins | Kung Fu Panda | Benchmark |
|--------------|----------|--------------|--------------|---------------|-----------|
| BA           | 3315.65  | 3446.02      | 3408.31      | 3513.44       | 4244.36   |
| CNT          | 2804.95  | 2989.85      | 3145.81      | 3165.68       | 5565.15   |
| Total        | 6120.60  | 6435.87      | 6554.12      | 6679.12       | 9809.51   |

Table 4 Final evaluation scores for the four best performing teams in the EVA 2021 data challenge and the benchmark. (A smaller score indicates better performance)
be highly computationally demanding. Hence, we have here decided to ignore covariates in the first three steps of the spatial model and use them only at the fourth stage of our model which is computationally very fast. In that way, by combining a statistical and a machine learning model, we take the advantage of both approaches while allowing a feasible computation time. For modeling discrete spatial data, an alternative approach using LGCPs was discussed. The fitted LGCP was constructed using a latent SPDE spatial effect; however, in case of spatial prediction of the US wildfire data, our proposed model outperforms the LGCP approach. While fully machine learning algorithms do not have any distributional assumption and are thus more robust, our proposed statistical method can better quantify the underlying uncertainty, specifically in Stages 1 and 3.

7.2 Drawbacks and possible solutions

Despite its elegant performance in predicting the distribution functions at masked sites, our approach also has some limitations.

First, we selected a multistage approach in which information was not borrowed from one stage to another. For example, in Stage 1, the data are treated as zero/nonzero indicators; however, in other stages, zeros are treated as missing data. Joint modeling of zero and nonzero data is a possible solution. Moreover, in Stage 2, we obtained smoothed parameter estimates in two steps; in the first step, we calculated empirical estimates, and in the second step, we treated such estimates as spatial data. Moving from Step 1 to Step 2, we ignored the uncertainty of the estimates in the first step. While the posterior coverage probabilities are only mildly affected by such a choice as shown in Hazra et al. (2019), a superior method for borrowing the uncertainty information from the first step to the second step has been recently discussed in Hrafnkelsson et al. (2021).

Second, because the study domain is in the Northern Hemisphere, the physical distance between two locations is not the same if they are one degree apart in longitude (about 84 km) or one degree apart in latitude (about 111 km). However, in our model construction in Stages 1 and 3, we have ignored this fact for simplicity and used the isotropic Matérn correlation function (3) with the Euclidean distance calculated from longitude-latitude coordinates. This implies that our model ends up being non-isotropic with slightly stronger dependence along the North–South direction. It would have been more natural to first project the domain onto a metric space where physical distances are preserved or to use the geodesic distance directly. Nevertheless, whether or not the true process is isotropic and stationary is also not clear with such a large domain. To properly account for this, it would be possible to make our model more complex by using an anisotropic and/or non-stationary Matérn correlation function (Hazra et al. 2018) or empirical orthogonal functions (Wikle 2010; Hazra and Huser 2021) instead, but this would also complicate inference substantially, so we have decided to leave it for future research. We also note that since the goal of the data competition was to predict high marginal quantiles of BA and CNT (with spatial dependence used here only to borrow strength across locations), we believe that our rather simplistic model choice approximates reality well enough that
it should not play a major role in the overall model predictive performance for this data competition.

Third, in Stage 4, we used the univariate random forest algorithm of Breiman (2001) independently at each site. However, recently, a random forest algorithm for spatially-dependent data has been proposed by Saha et al. (2021). Because we ignored covariate information in Stage 1 to Stage 3, the proposed model is not suitable to draw inference about the significance of a specific predictor in the context of fire modeling. A possible solution would be to fit a regression model for the mean surfaces.

7.3 Other applications

The model proposed here is a general tool for all bivariate zero-inflated spatiotemporal datasets, although our methodology has been motivated by the joint analysis of fire occurrences and sizes. The model supplements spatial marked point processes; thus, it is suitable for multiple scenarios. For example, Penttinen et al. (1992) discussed a marked point process approach for forest statistics, where ‘points’ are the tree positions, whereas the ‘marks’ are tree characteristics such as stem diameters and tree species. After summarizing the data over a grid, we can obtain a bivariate spatial dataset with two components representing the number of trees and average stem diameter within each grid cell. In such a scenario, our proposed inference approach would be suitable. Another example is the joint modeling of the number of rainy days within a month and the total monthly precipitation within a pixel/region. In addition to the marked point process scenario, our model can also be used to fit temperature and rainfall data over a spatiotemporal domain (Gelfand et al. 2005), for example. Temperature is a real-valued continuous variable for which a Gaussian assumption is common (and thus a log-transformation is not necessary), and rainfall data are non-negative and zeros for the dry periods; here, we can ignore Stage 4 if the log-transformed nonzero precipitation amounts follow a normal distribution.

7.4 Possible extensions

The method proposed in Sections 3.1 to 3.4 can be extended in multiple directions. The distribution function estimation problem for the EVA 2021 data challenge can be treated as a classification problem. Thus, by categorizing the data based on the 28 levels, a set of spatial indicators can be obtained, as described in Agarwal et al. (2021). Furthermore, a multivariate version of Stage 1 would be a reasonable model for drawing inference. The proposed model assumes temporal independence; thus, a natural extension would be to assume that the latent processes have temporal autocorrelation. By replacing the mean profiles as functions of the covariates, we can extend the proposed model to study the significance of a specific predictor in the context of fire modeling. The underlying dependence structures are assumed to be stationary for Stage 1 and Stage 3; however, some approaches available in the literature allow nonstationary spatial modeling for large datasets (Katzfuss 2013; Banerjee 2020), and we can extend the proposed model in that direction. Although the main focus in the data challenge is the accurate estimation of the upper tail of BA and CNT observations, we build our model
using Gaussian processes that have been criticized for modeling spatial extremes (Davison et al. 2013). While the extreme-value theory-justified models typically entail a huge computational burden even in low dimensions, a simpler alternative would be to consider scale mixture models (Huser et al. 2017; Huser and Wadsworth 2022; Hazra et al. 2021). Although we used an adaptive Metropolis-Hastings algorithm for the computations in Stages 1 and 3, we can extend it using some faster and more recently developed algorithms, like the stochastic gradient-based (Welling and Teh 2011) or deterministic transformation-based (Dutta and Bhattacharya 2014) MCMC algorithms.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10687-022-00460-8.

Acknowledgements The first three authors (Cisneros, Gong, Yadav) contributed equally to this work by implementing some of the methods and writing parts of the paper. The last two authors (Hazra, Huser) oversaw the whole project, with Hazra having a leading role throughout all practical aspects of the data competition (supervision of the Bedouins Team, methods’ implementation, results’ interpretation, writing). We would like to thank Thomas Opitz for organizing this very interesting data competition for the EVA 2021 Conference, as well as Thomas Mikosch for welcoming a Special Issue in Extremes about this topic. 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 The dataset analyzed during the current study is available from the corresponding author on reasonable request.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

Abatzoglou, J.T., Williams, A.P.: Impact of anthropogenic climate change on wildfire across western US forests. Proc. Natl. Acad. Sci. 113(42), 11770–11775 (2016)
Abdelfatah, K., Bao, J., Terejanu, G.: Environmental modeling framework using stacked Gaussian processes. Preprint at https://arxiv.org/abs/1612.02897 (2016)
Agarwal, G., Sun, Y., Wang, H.J.: Copula-based multiple indicator kriging for non-Gaussian random fields. Spat. Stat. 100524 (2021)
Bakka, H., Rue, H., Fuglstad, G.A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D., Lindgren, F.: Spatial modeling with R-INLA: a review. Wiley Interdiscip. Rev. Comput. Stat. 10(6), e1443 (2018)
Banerjee, S.: Modeling massive spatial datasets using a conjugate Bayesian linear modeling framework. Spat. Stat. 37, 100417 (2020)
Bivand, R., Gómez-Rubio, V., Rue, H.: Spatial data analysis with R-INLA with some extensions. J. Stat. Softw. 63(20), 1–31 (2015)
Breiman, L.: Random Forests. Mach. Learn. 45(1), 5–32 (2001)
Brown, E.K., Wang, J., Feng, Y.: US wildfire potential: a historical view and future projection using high-resolution climate data. Environ. Res. Lett. 16(3), 034060 (2021)
Cressie, N., Johannesson, G.: Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Series B Stat. Methodol. 70(1), 209–226 (2008)
Cumming, S.: A parametric model of the fire-size distribution. Can. J. For. Res. 31(8), 1297–1303 (2001)
Davison, A.C., Huser, R., Thibaud, E.: Geostatistics of dependent and asymptotically independent extremes. Math. Geosci. 45(5), 511–529 (2013)
Diggle, P.J., Moraga, P., Rowlingson, B., Taylor, B.M.: Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. Stat. Sci. 28(4), 542–563 (2013)
Dutta, S., Bhattacharya, S.: Markov chain Monte Carlo based on deterministic transformations. Stat. Methodol. 16, 100–116 (2014)

Fusco, E.J., Finn, J.T., Balch, J.K., Nagy, R.C., Bradley, B.A.: Invasive grasses increase fire occurrence and frequency across US ecoregions. Proc. Natl. Acad. Sci. 116(47), 23594–23599 (2019)

Gabriel, E., Opitz, T., Bonneu, F.: Detecting and modeling multi-scale space-time structures: the case of wildfire occurrences. J. Soc. Fr. Stat. 158(3), 86–105 (2017)

Gelfand, A.E., Banerjee, S., Gamerman, D.: Spatial process modelling for univariate and multivariate dynamic spatial data. Environmetrics 16(5), 465–479 (2005)

Gelfand, A.E., Schlief, E.M.: Spatial statistics and Gaussian processes: a beautiful marriage. Spat. Stat. 18, 86–104 (2016)

Genton, M.G., Butry, D.T., Gumpertz, M.L., Prestemon, J.P.: Spatio-temporal analysis of wildfire ignitions in the St. Johns River water management district, Florida. Int. J. Wildland Fire 15(1), 87–97 (2006)

Hazra, A., Huser, R.: Estimating high-resolution Red Sea surface temperature hotspots, using a low-rank semiparametric spatial model. Ann. Appl. Stat. 15(2), 572–596 (2021)

Hazra, A., Huser, R., Bolin, D.: A sparse Gaussian scale mixture process for short-range extremal dependence and long-range independence. Preprint at http://arxiv.org/abs/2112.10248 (2021)

Hazra, A., Huser, R., Jóhannesson, Á.V.: Latent Gaussian models for high-dimensional spatial extremes. Preprint at http://arxiv.org/abs/2110.02680 (2021)

Hazra, A., Reich, B.J., Reich, D.S., Shinohara, R.T., Staicu, A.M.: A spatio-temporal model for longitudinal image-on-image regression. Stat. Biosci. 11(1), 22–46 (2019)

Hazra, A., Reich, B.J., Shaby, B.A., Staicu, A.M.: A semiparametric spatiotemporal Bayesian model for the bulk and extremes of the Fosberg Fire Weather Index. Preprint at http://arxiv.org/abs/1812.11699 (2018)

Hering, A.S., Bell, C.L., Genton, M.G.: Modeling spatio-temporal wildfire ignition point patterns. Environ. Ecol. Stat. 16(2), 225–250 (2009)

Hrafneksson, B., Siegert, S., Huser, R., Bukka, H., Jóhannesson, Á.V.: Max-and-smooth: a two-step approach for approximate Bayesian inference in latent Gaussian models. Bayesian Anal. 16(2), 611–638 (2021)

Huser, R., Opitz, T., Thibaud, E.: Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixture models. Spat. Stat. 21, 166–186 (2017)

Huser, R., Wadsworth, J.L.: Advances in statistical modeling of spatial extremes. Wiley Interdiscip. Rev. Comput. Stat. 14, e1537 (2022)

Jain, P., Coogan, S.C., Subramanian, S.G., Crowley, M., Taylor, S., Flannigan, M.D.: A review of machine learning applications in wildfire science and management. Environ. Rev. 28(4), 478–505 (2020)

Jóhannesson, Á.V., Siegert, S., Huser, R., Bukka, H., Hrafneksson, B.: Approximate Bayesian inference for analysis of spatio-temporal flood frequency data. Ann. Appl. Stat. 16(2), 905–935 (2022)

Joseph, M.B., Rossi, M.W., Mietkiewicz, N.P., Mahood, A.L., Cattau, M.E., St, L.A., Denis, R.C., Nagy, V., Iglesias, J.T. Abatzoglou., Balch, J.K.: Spatiotemporal prediction of wildfire size extremes with Bayesian finite sample maxima. Ecol. Appl. 29(6), e01898 (2019)

Juan, P., Mateu, J., Saez, M.: Pinpointing spatio-temporal interactions in wildfire patterns. Stoch. Env. Res. Risk Assess. 26(8), 1131–1150 (2012)

Katzfuss, M.: Bayesian nonstationary spatial modeling for very large datasets. Environmetrics 24(3), 189–200 (2013)

Koh, J., Pimont, F., Dupuy, J.L., Opitz, T.: Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. Preprint at https://arxiv.org/abs/2105.08004 (2021)

Lindgren, F., Rue, H., Lindström, J.: An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Series B Stat. Methodol. 73(4), 423–498 (2011)

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F.A.: A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinf. 10(1), 1–16 (2009)

Møller, J., Díaz-Avalos, C.: Structured spatio-temporal shot-noise Cox point process models, with a view to modelling forest fires. Scand. J. Stat. 37(1), 2–25 (2010)

Møller, J., Syversveen, A.R., Waagepetersen, R.P.: Log-Gaussian Cox processes. Scand. J. Stat. 25(3), 451–482 (1998)

Nadeem, K., Taylor, S., Woolford, D.G., Dean, C.: Mesoscale spatiotemporal predictive models of daily human-and lightning-caused wildland fire occurrence in British Columbia. Int. J. Wildland Fire 29(1), 11–27 (2020)

Opitz, T.: Editorial: EVA 2021 Data Competition on spatio-temporal prediction of wildfire activity in the United States. Extremes (to appear). (2022)
Penttinen, A., Stoyan, D., Henttonen, H.M.: Marked point processes in forest statistics. Forest Sci. 38(4), 806–824 (1992)
Pereira, J., Turkman, K.: Statistical models of vegetation fires: Spatial and temporal patterns. Handbook of Environmental and Ecological Statistics, pp. 401–420. Taylor & Francis: Chapman and Hall/CRC (2019)
Pimont, F., Fargeon, H., Optiz, T., Ruffault, J., Barbero, R., Martin-StPaul, N., Rigolot, E., Rivière, M., Dupuy, J.L.: Prediction of regional wildfire activity in the probabilistic Bayesian framework of Firelihood. Ecol. Appl. 31(5), e02316 (2021)
Preisler, H., Ager, A.: Forest-fire models. Environ. Encycl. 3, 2181–2185 (2013)
Preisler, H.K., Brillinger, D.R., Burnag, R.E., Benoit, J.: Probability based models for estimation of wildfire risk. Int. J. Wildland Fire 13(2), 133–142 (2004)
Preisler, H.K., Westerling, A.L.: Statistical model for forecasting monthly large wildfire events in western United States. J. Appl. Meteorol. Climatol. 46(7), 1020–1030 (2007)
Pyne, S., Andrew, P., Laven, R.: Introduction to Wildland and Rural Fire. Princeton University Press, Princeton, NJ (1996)
Rios-Pena, L., Kneib, T., Cadarso-Suarez, C., Klein, N., Marey-Perez, M.: Studying the occurrence and burnt area of wildfires using zero-one-inflated structured additive beta regression. Environ. Model. Software 110, 107–118 (2018)
Rue, H., Held, L.: Gaussian Markov Random Fields: Theory and Applications. Taylor & Francis: Chapman and Hall/CRC (2005)
Saha, A., Basu, S., Datta, A.: Random forests for spatially dependent data. J. Am. Stat. Assoc. 1–19 (2021). https://doi.org/10.1080/01621459.2021.1950003
Scott, A.C.: The pre-quaternary history of fire. Palaeogeogr. Palaeoclimatol. Palaeoecol. 164(1–4), 281–329 (2000)
Serra, L., Saez, M., Varga, D., Tobías, A., Juan, P., Mateu, J.: Spatio-temporal modelling of wildfires in Catalonia, Spain, 1994–2008, through log-Gaussian Cox processes. Modelling, Monitoring and Management of Forest Fires III, pp. 11139. (2012)
Trucchia, A., Egorova, V., Pagnini, G., Rochou, M.C.: Surrogate-based global sensitivity analysis for turbulence and fire-spotting effects in regional-scale wildland fire modeling. Preprint at https://arxiv.org/abs/1809.05430 (2018)
Welling, M., Teh, Y.W.: Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 681–688. (2011)
Wikle, C.K.: Low-rank representations for spatial processes. Handbook of Spatial Statistics, pp. 114–125. Taylor & Francis: CRC Press (2010)
Wuebbles, D.J., Fahey, D.W., Hibbard, K.A., Arnold, J.R., DeAngelo, B., Doherty, S., Easterling, D.R., Edmonds, J., Edmonds, T., Hall, T. et al.: Climate science special report: Fourth national climate assessment (NCA4), vol. I. (2017)
Xi, D.D., Taylor, S.W., Woolford, D.G., Dean, C.: Statistical models of key components of wildfire risk. Annu. Rev. Stat. Appl. 6, 197–222 (2019)
Yadav, R., Huser, R., Optiz, T.: A flexible Bayesian hierarchical modeling framework for spatially dependent peaks over-threshold-data. Spat. Stat. 51, 100672 (2022)
Zammit-Mangion, A., Cressie, N.: FRK: an R package for spatial and spatio-temporal prediction with large datasets. J. Stat. Softw. 98(4), 1–48 (2021)

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

Daniela Cisneros¹ · Yan Gong¹ · Rishikesh Yadav¹ · Arnab Hazra¹,² ★ · Raphaël Huser¹

★ Arnab Hazra
ahazra@iitk.ac.in

1 Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia

2 Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Kanpur 208016, India