Spatio-Temporal Wildfire Prediction Using Multi-Modal Data

Chen Xu, Yao Xie, *Member, IEEE*, Daniel A. Zuniga Vazquez,
Rui Yao, *Senior Member, IEEE*, and Feng Qiu

***Abstract*—Due to severe societal and environmental impacts, wildfire prediction using multi-modal sensing data has become a highly sought-after data-analytical tool by various stakeholders (such as state governments and power utility companies) to achieve a more informed understanding of wildfire activities and plan preventive measures. A desirable algorithm should precisely predict fire risk and magnitude for a location in real time. In this paper, we develop a flexible spatio-temporal wildfire prediction framework using multi-modal time series data. We first predict the wildfire risk (the chance of a wildfire event) in real-time, considering the historical events using discrete mutually exciting point process models. Then we further develop a wildfire magnitude prediction set method based on the flexible distribution-free time-series conformal prediction (CP) approach. Theoretically, we prove a risk model parameter recovery guarantee, as well as coverage and set size guarantees for the CP sets. Through extensive real-data experiments with wildfire data in California, we demonstrate the effectiveness of our methods, as well as their flexibility and scalability in large regions.**

***Index Terms*—Spatial-temporal point process, conformal prediction, multi-sensor network, fire safety.**

I. Introduction

In recent years, widespread large-scale wildfire cause severe consequences, including direct property damage and economic losses, community evacuation, and fatalities, as well as impacts on nature such as higher CO_2 emissions [1]. To monitor and prevent severe consequence caused by large-scale wildfire, an imperative challenge was brought up: how to utilize multi-modal data collected through various sensing technologies, so as to precisely predict wildfire risk and magnitude for a local region and monitor the predictions in real-time.

Wild fire risk prediction is particularly important for power utility companies to enhance their capability in making precise location-wise wildfire risk predictions. To prevent damage and economic losses, the utility companies also perform schedule utility shutdown for high wild-fire risk regions [2]. Despite such urgent and essential need, utility companies often only leverage simple models/metrics for risks assessment, such as the burning index (BI) [3] and the fire load index [4], which are static metrics that do not take into account the contribution from historical wildfire incidents and auxiliary environmental information. Imprecise wildfire risk prediction is causing sub-optimal power operator actions (such as unnecessary shutdown) that significantly disrupt reliable power delivery to customers.

Meanwhile, thanks to the development of sensing technology, there have been abundant multi-modal data collected through a variety of sensing mechanisms to gather wildfire information [5], which provides the unique opportunity for using sensing to perform precise location-wise real-time wildfire prediction. Common approaches to identify wildfire incidents include reports from human observers, wireless sensing [6], and infrared technology. Additional environmental information (e.g., weather and environmental conditions) has been integrated with each record, thus providing excellent opportunities for subsequent statistical analyses. As a result, each wildfire record is multi-modal: we know not only when and where it occurred but also its magnitude, the condition of the surrounding (e.g., infrastructure type), current weather information, and so on. Nevertheless, most existing wildfire modeling approaches [7], [8], [9], [10] have not been designed to utilize such abundant multi-modal data.

In this paper, we present a framework for predicting wildfire risk and magnitude using multi-modal sensing data, based on a mutually exciting point process model and time series conformal prediction sets. Our model can capture the complex spatial-temporal dependence of the multi-modal data through mutually exciting point processes, which is a natural framework for real-time prediction, since the conditional probability can be used to capture fire risk given the past observations. In addition, we present a fire magnitude prediction algorithm through time-series CP sets. Theoretically, we first prove model parameter recovery guarantees of the point process model for risk prediction. We then present coverage guarantees of fire magnitude prediction sets. Through extensive real-data experiments, we verify our models' competitive performances against other baseline methods regarding the precision of wildfire risk prediction.

Our prediction framework has the following features: (i) Predicting the wildfire risk — the chance of *binary* fire

Manuscript received 10 October 2022; revised 7 March 2023 and 20 April 2023; accepted 9 May 2023. Date of publication 22 May 2023; date of current version 21 August 2023. This work was supported in part by NSF CAREER under Grant CCF-1650913; in part by NSF under Grant DMS-2134037, Grant CMMI-2015787, Grant CMMI-2112533, Grant DMS-1938106, and Grant DMS-1830210; and in part by the U.S. Department of Energy Advanced Grid Modeling Program under Grant DE-OE0000875. *(Corresponding author: Yao Xie.)*

Chen Xu and Yao Xie are with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: yao.xie@isye.gatech.edu).

Daniel A. Zuniga Vazquez, Rui Yao, and Feng Qiu are with the Energy Systems Division, Argonne National Laboratory, Lemont, IL 60439 USA.

This article has supplementary downloadable material available at https://doi.org/10.1109/JSAIT.2023.3276054, provided by the authors.

Digital Identifier 10.1109/JSAIT.2023.3276054

event (no fire versus fire) at a given locations and times, given historical observations and available multi-modal data (which can be treated as marks of the point processes), using a flexible marked spatio-temporal Hawkes process model [11]. Specifically, we model the *mutual exciting property* in that historical and neighboring occurrences likely affect the occurrence likelihood, where certain occurrences may increase the chance while others inhibit the chance. The model parameters are efficiently estimated using an alternating optimization approach, in contrast to the more expensive expectation-maximization method [12]. (ii) Exploiting interdependence among different geographic regions and the mutually exciting point process model is highly interpretable. (iii) Predicting fire magnitude using time-series CP set, which can guarantee to contain true fire magnitude with a user specified high probability.

The rest of the paper is organized as follows. Section II describes background on sensing and the wildfire dataset. Section III contains our proposed methods. In particular, Section III-A introduces proposed spatio-temporal Hawkes process models, which either linearly (i.e., `LinearSTHawkes`) or nonlinearly (i.e., `NonLinearSTHawkes`) quantify feature contributions to fire hazards. Section III-B describes the objective function, the estimation procedure, and how to yield binary predictions based on predicted risks. Section III-C describes the CP sets for wildfire magnitude prediction. Section IV has two parts. We first present the theoretical analyses regarding the accuracy of fire risk prediction as a result of model recovery guarantee in Section IV-A. Section IV-B then verifies coverage guarantee of the prediction sets, whose size also converge to the true fire sizes asymptotically. Section V first validates the proposed model on a small-scale real-data experiment, where Section V-B compares `LinearSTHawkes` with baseline methods and Section V-C demonstrates the further advantage of `NonLinearSTHawkes`. Section VI then shows the scalability of our methods on a significantly larger region, where Section VI-B further examines the empirical coverage of prediction sets by the CP method. Finally, Section VII concludes the work with discussion on future steps. The Appendix contains additional derivations and algorithms.

A. Related Work

Wildfire prediction and modeling is an essential procedure for analyzing the occurrence of wildfire events. There have many indices, such as the BI [13] and the fire danger index [14] for general awareness of fire risks. Despite their popularity, these indices often fail to account for events' interactions. Meanwhile, regression-based approaches [9], [15], [16] are more flexible and often yield satisfactory predictions. However, their performance can be sensitive to the number of available observations per location and thus not applicable under arbitrary spatial granularity with a fixed amount of training data. Lastly, stochastic point-process models [17], [18], [19] have been leveraged to examine the conditional fire risk given past data and allow a deeper understanding of the underlying stochastic mechanism. However, most current works focus on model evaluation through the akaike information criterion (AIC) rather than predicting the binary occurrence of wildfire events using one-class data. In practice, making a binary prediction is essential for forestry managers and utility owners to understand the fire risk.

Since our proposed fire occurrence model is based on the Hawkes process, we briefly survey existing methods in a wider context. Initially proposed in [11], the Hawkes process is a stochastic temporal point-process model for rates of events conditioning on historical ones. There have been many extensions that take into account spatial interactions [20], [21], [22] and influences by auxiliary features (i.e., marks) [23], [24], [25]. Neural-network-based Hawkes process models [26], [27], [28] have also been proposed for greater model expressiveness. These models have shown great promise in fields such as financial markets [29], social networks [30], disease modeling [31], and neurophysiological studies [32]. Despite their emerging popularity and flexibility, how to make a prediction based on rate estimates and comparisons against predictive models has been less well studied.

We briefly surveyed CP, the primary tool used for constructing prediction sets that quantify uncertainty in fire magnitude prediction. Originated in the seminal work [33], CP has gained wide popularity for uncertainty quantification [34]. It is particularly appealing as the methods are distribution-free, model-agnostic, and easily implementable. The only assumption is that observations are exchangeable (e.g., i.i.d.). On a high level, CP methods assign non-conformity scores to potential outcomes of the response variable. The outcomes that have small non-conformity scores are included in the prediction set. Many methods follow this logic with promising results [35], [36], [37], [38], [39]. More recently, works have also relaxed the exchangeability assumption [40], [41], [42], [43], [44], [45], but time-series CP methods are still limited, and their applications to wildfire predictions remain largely unexplored.

II. Sensing for Wildfire and Real-Data Illustration

The latest technology provides multi-modal data for wildfire risk prediction and monitoring. Below, we briefly describe a few common sensing and data collection techniques [5], [46].

- Air patrols: Patrollers typically consist of a pilot and a trained aerial observer. To identify and report observed wildfire phenomena, the plane flies over predetermined areas during periods associated with elevated fire danger. Wildfire activities are also commonly reported by commercial or recreational pilots.
- Infrared technology: Thermal imaging technology is commonly used to detect fire risks hot spots. It is also used to detect wildfire progression, contour the fire impact, and identify residual fire during extinguishment.
- Computer technology: Various management systems are used to obtain well-rounded multi-modal information. Such systems obtain up-to-date weather information, predict the fire probability and spread rate, and reports moisture levels in the natural surrounding.

A feature of our work is that we validate our model on a large-scale multi-modal dataset, 2014–2019 fire incident data collected by the California public utilities commission [46]. The wildfire occurrence dataset is publicly available and associated with three large utility companies: PG&E, SCE, and SDG&E. A total of 3191 fire incidents are recorded, where the latitude-longitude coordinates of each incident are enclosed within the coordinate rectangle $[32.24, -124, 38] \times [41.28, -114.67]$.

The wildfire data is multi-modal and collecting using various sensing mechanism. Each incident is multi-modal with additional information, which we call *marks* in our model. Marks can be categorized as being discrete/continuous and dynamic/static. Static marks do not change at a given location, and all discrete marks are one-hot encoded to be utilized in the model. Static and discrete marks include existing vegetation type and physiology (EVT_PHYS) [47], such as the road condition and agricultural condition, the name of the three utility companies, and the fire threat zone, which is classified into three levels indicating increasing levels of static fire danger [46]. Dynamic and discrete marks include seasonal information (e.g., spring, summer, autumn, and winter). Dynamic and continuous marks include relative humidity in % of the surrounding [48] temperature in celsius [48] large fire probability (LFP) [49], and fire potential index (FPI) [49]. In particular, LFP and FPI are forecasted by the United States geological survey (USGS) to indicate the risks associated with a region.

To pre-process the multi-modal data, we interpolate missing entries of each continuous mark using the spline function with degree 5. Each feature is also standardized to have unit variance and zero mean and further scaled to lie within the interval [0, 1] so that estimated parameters for different marks are on the same scale. The unit for risk prediction is in days, while we allow fractional time values during training where the exact hour and minutes are recorded along each incident.

III. Wildfire Prediction Framework

A. Wildfire Risk Prediction: Mutually Exciting Spatio-Temporal Point Processes

We observe a sequence of n fire incidents over a time horizon $[0, T]$, where each observation consists of time t_i, location u_i, and a mark $m_i \in \mathbb{R}^p$ (where p is the number of features):

$$x_i = (t_i, u_i, m_i), \quad i = 1, \dots, n. \tag{1}$$

Note that we specify $u_i \in \{1, \dots, K\}$ for K locations under space discretization.

We model these event data using a marked spatio-temporal Hawkes process. Given the σ-algebra $\mathcal{H}_t$ that denotes all historical fire occurrence before time t, the conditional intensity function is the probability of an event occurring at time t and location k, with current mark m:

$$\lambda(t, k, m|\mathcal{H}_t) = \lim_{\Delta t, \Delta u \to 0} \frac{\mathbb{E}[\mathbb{N}([t, t + \Delta t) \times B(k, \Delta k) \times B(m, \Delta m)) \mid \mathcal{H}_t]}{\Delta t \times B(k, \Delta k) \times B(m, \Delta m)}, \tag{2}$$

where $B(a, r)$ is a ball centered at a with radius r and $\mathbb{N}$ is the counting measure. For notation simplicity, we drop $\mathcal{H}_t$ in (2) from now on.

We can use the conditional intensity function above (2) to quantify the fire risk. For mutually exciting point processes, the conditional intensity function depend on the past events and they typically increase the chance of a future event in the neighborhood. This *mutual excitation* can be modeled by representing the conditional intensity function (2) as (see, e.g., [12]):

$$\begin{aligned} \lambda(t, k, m) &= \lambda_g(t, k) f(m|t, k) \\ &= \left(\mu(k) + \sum_{j: t_j < t} \mathcal{K}(u_j, k, t_j, t) \right) f(m|t, k), \end{aligned} \tag{3}$$

which factors the conditional intensity into product of ground process $\lambda_g(t, k)$ and conditional density $f(m|t, k)$. In (3), $\mu(k)$ is the scalar baseline intensity and $\mathcal{K}(u_j, k, t_j, t)$ measures spatial and temporal influence from event happening at t_j in u_j till current time t through a kernel function.

In general, functions $\mu(k)$, $\mathcal{K}(u_j, k, t_j, t)$, and $f(m|t, k)$ can take many possible forms. Such choices often depend on the application of interest. For computation simplicity and model interpretability, here we parametrize the model in (3) as

$$\mu(k) = \mu_k, \quad \mathcal{K}\big(u_j, k, t_j, t\big) = \alpha_{u_j,k} \beta e^{-\beta(t - t_j)}. \tag{4}$$

In equation (4), the parameters μ_k represent the baseline rate of fire risk at location k. The parameters $\alpha_{u_j,k}$ capture the spatial influence of fire incidents that occurred at location u_j and time t_j on the fire risk at location k and time t. To simplify the design of $\mathcal{K}(u_j, k, t_j, t)$ in (4), we use a negative exponential model. This choice is motivated by two key factors. Firstly, it results in an optimization problem whose parameters can be efficiently estimated with a performance guarantee (refer to Section IV). Secondly, domain experts have observed that past fire incidents can affect the risk of future fire incidents, but the impact of past events diminishes quickly over time.

Furthermore, we assume the distribution of the mark is either in linear form or, more generally, through a non-linear function g

$$f(m|t, k) = \gamma^T m, \qquad (\texttt{LinearSTHawkes}) \tag{5}$$

$$f(m|t, k) = g(m|t, k) \qquad (\texttt{NonLinearSTHawkes}) \tag{6}$$

Even though (5) is linear, it implicitly incorporates the spatial-temporal information through the mark m, which is collected in location k at time t. Meanwhile, $g(m|t, k)$ in (6) can be any feature extractor (e.g., neural networks) that outputs the score of m. Regarding the formulation differences of (5) and (6), note that `LinearSTHawkes` based on (5) is more interpretable, and also leads to more computationally efficient sequential convex optimization scheme with guarantees (see Section IV-A). On the other hand, `NonLinearSTHawkes` can be more expressive in terms of capturing the dependency of fire risks on marks through the feature extractor $g(m|t, k)$ in (6).

B. Point Process Parameter Estimation and Real-Time Prediction

We estimate the parameters in the model through maximum likelihood. For LinearSTHawkes, denote all parameters using $\theta = \{\mu, A, \beta, \gamma\}$, where $\mu = \{\mu_k\}_{k=1}^K$ and $A = [\alpha_{i,j}]_{i,j=1}^K$. We can derive and simplify the log-likelihood of $x_1, \ldots, x_n$ as follows similar to [12] (the full derivation can be found in Appendix G):

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^n \log\big(\lambda_g(t_i, u_i)\big) \\ &\quad + \sum_{i=1}^n \log(f(m_i|t_i, u_i)) - \sum_{k=1}^K \int_0^T \lambda_g(\tau, k) d\tau \\ &= \sum_{i=1}^n \log\left(\mu(u_i) + \sum_{j:t_j<t_i} \alpha_{u_j,u_i} \beta e^{-\beta(t_i-t_j)}\right) \\ &\quad + \sum_{i=1}^n \log(f(m_i|t_i, u_i)) - \sum_{k=1}^K T\mu(k) \\ &\quad - \sum_{i=1}^n \left(\sum_{k=1}^K \alpha_{u_i,k}\right)\left(1 - e^{-\beta(T-t_i)}\right). \qquad (7)\end{aligned}$$

Note that the likelihood term of the marks decouples from the rest. Thus, when using NonLinearSTHawkes based on (6), we first fit a feature extractor on the marks and then employ maximum likelihood estimation to estimate the rest parameters. To achieve better model estimation stability (since we believe few features should be effective in the model), we further add ℓ_1 regularization on γ:

$$\begin{aligned}\min_{\theta=\{\mu,A,\beta,\gamma\}} \quad & -\sum_{i=1}^n \log\left(\mu(u_i) + \sum_{j:t_j<t_i} \alpha_{u_j,u_i} \beta e^{-\beta(t_i-t_j)}\right) \\ & -\sum_{i=1}^n \log\big(\gamma^T m_i\big) + \sum_{k=1}^K T\mu(k) \\ & + \sum_{i=1}^n \left(\sum_{k=1}^K \alpha_{u_i,k}\right)\left(1 - e^{-\beta(T-t_i)}\right) + \|\gamma\|_1 \qquad (8) \\ \text{subject to} \quad & \alpha_{i,j} = 0 \text{ if } |i-j| \geq \tau, \qquad (9) \\ & \|\mu\|_2 \leq 1, \|A\|_2 \leq 1, \|\gamma\|_2 \leq 1, \qquad (10) \\ & \beta \geq 0, \mu(u_i) \geq 0 \ \forall u_i. \qquad (11)\end{aligned}$$

The purpose of constraints (9)-(11) can be explained as follows: (9) introduces sparsity in the interaction matrix and reduces the total number of parameters in the model for computational efficiency; (10) ensures the objective (8) is bounded and is reasonable since the rate $\lambda(t, k, m)$ is typically very small; (11) is introduced since baseline rates (i.e., $\mu(u_i)$) and interaction propagation over time (i.e., β) are non-negative. Note that the constraints define a convex feasible region.

In addition, we can show that $\ell(\theta)$ is concave in all other parameters with a fixed scalar β. Thus, we can device a method to solve (8) to global optimal solution: for a grid of β values, solve the corresponding convex optimization problem using solvers such as [50] to high numerical accuracy, and then choose the optimal β that gives the best overall objective value. The description of the algorithm, as well its computational efficiency, is in Algorithm 2 of Appendix H. In our experiments, we observe that the algorithm usually terminates in a small number of iterations (e.g., three), and each iteration only takes a few seconds to minutes, depending on the problem size. Hence, it is computationally friendly.

C. Fire Magnitude Prediction: Conformal Prediction Set

Besides predicting when and where fire occurs, fire magnitude prediction is also desirable—knowing the possible fire magnitude can better inform decision-makers of potential losses by such disasters and plan accordingly. The dataset described in Section II treats fire magnitude as discrete categories in its catalog. In principle, this can thus be achieved by variants of LinearSTHawkes and NonLinearSTHawkes for categorical data. However, making categorical prediction based on the estimated risks requires us to construct multi-class thresholds, which can greatly increase model design complexity. In addition, it is unclear how to quantify uncertainty in the resulting categorical estimates.

Thus, we treat fire magnitude prediction as a classification problem: given multi-modal features $X_i \in \mathbb{R}^p$ as in (1), we would like to build a multi-class classifier that outputs $\hat{Y}_i \in \{1, \ldots, C\}$ as the fire magnitude prediction (assuming C magnitude levels). Denote $\pi_i := P_{Y_i|X_i}$ as the true conditional distribution of $Y_i|X_i$, whose properties are unknown. In a typical classification setting, we assume the first N data are known to us as training data and the goal is to construct an estimator $\hat{\pi} := \mathcal{A}(\{(X_i, Y_i)\}_{i=1}^N)$, which satisfies $\sum_{c=1}^C \hat{\pi}_{X_i}(c) = 1, \hat{\pi}_{X_i}(c) \geq 0$ for any $i \geq 1$. Here, $\mathcal{A}$ is any classification algorithm, from the simplest multinomial logistic regression to a complex deep neural networks. Then, the point prediction $\hat{Y}_i := \arg\max_{c\in[C]} \hat{\pi}_{X_i}(c)$ is obtained for any test index $i > N$.

However, point predictions are often insufficient in such settings—there are inherent uncertainties in these predictions, which arise due to randomness in data generation, during the collection of multi-modal data, and when fitting the multi-class classifier. Therefore, a *confident* fire magnitude prediction is essential, which quantifies uncertainties in the point predictions and contains all the possible high-probability outcomes. One way for uncertainty quantification in classification is the construction of *prediction sets* around $\hat{Y}_i$ that contain actual observations Y_i with high probability before its realization. Formally, given a significance level $\alpha \in (0, 1)$, we construct a *prediction set* $\widehat{C}(X_i, \alpha) \subset \{1, \ldots, C\}$ such that

$$\mathbb{P}\big(Y_i \in \widehat{C}(X_i, \alpha)\big) \geq 1 - \alpha. \qquad (12)$$

We note that the significance level α in conformal prediction should be distinguished from the interaction parameters α_{ij} in the point-process model, the latter of which has double subscripts as in (4). A set satisfying (12) thus confidently predicts the actual fire magnitude Y_i with high probability. Note that a trivial construction that always satisfies (12) is $\widehat{C}(X_i, \alpha) = \{1, \ldots, C\}$, so we also want the prediction set to be as small as possible. This is a challenging question because fire incidents are highly correlated and non-stationary, and classifiers can be very complex (e.g., neural network classifiers).

To build prediction sets that satisfy (12) in practice, we produce uncertainty sets using recent advances in CP [36], [42], [51]. CP methods requires two ingredients. First, they define *non-conformity scores*, which quantify the dissimilarity of a potential fire magnitude. Second, they specify the prediction set based on non-conformity scores. As a result, CP methods assign non-conformity scores to each possible fire magnitude and the prediction set contains fire magnitude whose non-conformity scores are small compared to past ones.

We first specify a particular form of non-conformity score recently developed in [36] using any estimator $\hat{\pi}$. The notations are very similar and we include the descriptions for a self-contained exposition. Given the estimator $\hat{\pi}$, for each possible label c at test feature $X_i, i > N$, we make two other definitions:

$$m_{X_i}(c) := \sum_{c'=1}^{C} \hat{\pi}_{X_i}(c') \cdot \mathbb{I}\big(\hat{\pi}_{X_i}(c') > \hat{\pi}_{X_i}(c)\big). \tag{13}$$

$$r_{X_i}(c) := \left| \sum_{c'=1}^{C} \mathbb{I}\big(\hat{\pi}_{X_i}(c') > \hat{\pi}_{X_i}(c)\big) \right| + 1. \tag{14}$$

where $\mathbb{I}$ is the indicator function. In other words, (13) calculates the total probability mass of labels deemed more likely than c by $\hat{\pi}$. It strictly increases as c becomes less probable. Meanwhile, (14) calculates the rank of c within the order statistics. It is also larger for less probable c. Given a random variable $U_i \sim \text{Unif}[0, 1]$ and pre-specified regularization parameters $\{\lambda, k_{reg}\}$, we define the non-conformity score as

$$\hat{\tau}_i(c) := m_{X_i}(c) + \underbrace{\hat{\pi}_{X_i}(c) \cdot U_i}_{(i)} + \underbrace{\lambda\big(r_{X_i}(c) - k_{reg}\big)^+}_{(ii)}. \tag{15}$$

We interpret terms (i) and (ii) in (15) as follows. Term (i) randomizes the uncertainty set, accounts for discrete probability jumps when new labels are considered. A similar randomization factor is used in [35, eq. (5)]. In term (ii), $(z)^+ := \max(z, 0)$. Meanwhile, the regularization parameters $\{\lambda, k_{reg}\}$ force the non-conformity score to increase when λ increases and/or k_{reg} decreases. In words, λ denotes the additional penalty when the label is less probable by one rank and k_{reg} denotes when this penalty takes place. This term ensures that the sets are *adaptive*, by returning smaller sets for easier cases and larger ones for harder cases.

Then, the prediction set based on (15) is

$$\widehat{C}(X_i, \alpha) := \left\{ c \in [C] : \sum_{j=i-N}^{i-1} \mathbb{I}\big(\hat{\tau}_j \le \hat{\tau}_i(c)\big)/N < 1 - \alpha \right\}, \tag{16}$$

where $\hat{\tau}_j := \hat{\tau}_j(Y_j)$. The set in (16) includes all the labels whose non-conformity scores are no greater than $(1 - \alpha)$ fraction of previous N non-conformity scores. Following (15) and (16), we thus propose *ensemble regularized adaptive prediction set* (`ERAPS`) in Algorithm 1. In particular, `ERAPS` aggregates probability predictions from bootstrap multi-class classifiers to yield more accurate point prediction and leverage new feedback of Y_i to ensure adaptiveness in the prediction sets.

Algorithm 1 Ensemble Regularized Adaptive Prediction Set

```
Require: Training data$\{(X_i, Y_i)\}_{i=1}^{N}$, classification algorithm $\mathcal{A}$, $\alpha$, regularization parameters $\{\lambda, k_{reg}\}$, aggregation function $\phi$ (e.g., mean), number of bootstrap models $B$, the batch size $s$, and test data $\{(X_i, Y_i)\}_{i=N+1}^{N+N_1}$, with $Y_i$ revealed only after the batch of $s$ prediction intervals with $i$ in the batch are constructed.
Ensure: Ensemble uncertainty sets $\{\widehat{C}(X_i, \alpha)\}_{i=N+1}^{N+N_1}$
for $b = 1, \dots, B$ do    ▷ Train Bootstrap Estimators
    Sample with replacement an index set $S_b = (b_1, \dots, b_N)$ from indices $(1, \dots, N)$.
    Compute $\hat{\pi}^b = \mathcal{A}(\{(X_i, Y_i) \mid i \in S_b\})$.
end for
Initialize $\boldsymbol{\tau} = \{\}$ and sample $\{U_i\}_{i=1}^{N+N_1} \overset{i.i.d.}{\sim} \text{Unif}[0, 1]$.
for $i = 1, \dots, N$ do    ▷ LOO Ensemble Estimators and Scores
    Compute $\hat{\pi}^{\phi}_{-i} := \phi(\{\hat{\pi}^b : i \notin S_b\})$ such that for each $c \in \{1, \dots, C\}$
        $\hat{\pi}^{\phi}_{-i,X_i}(c) = \phi(\{\hat{\pi}^b_{X_i}(c) : i \notin S_b\})$.
    Compute $\hat{\tau}^{\phi}_i := \hat{\tau}_{X_i}(Y_i)$ using (15) and $\hat{\pi}^{\phi}_{-i}$.
    $\boldsymbol{\tau} = \boldsymbol{\tau} \cup \{\hat{\tau}^{\phi}_i\}$
end for
for $i = N + 1, \dots, N + N_1$ do    ▷ Build Uncertainty Sets
    Compute $\hat{\tau}^{\phi}_{i,cal} := q_{\boldsymbol{\tau},1-\alpha}(\boldsymbol{\tau})$ as the $(1 - \alpha)$-empirical quantile of $\boldsymbol{\tau}$.
    Compute $\hat{\pi}^{\phi}_{-i} := \phi(\{\hat{\pi}^{\phi}_{-i}\}_{i=1}^{N})$ so that for each $c \in \{1, \dots, C\}$
        $\hat{\pi}^{\phi}_{-i,X_i}(c) := \phi(\{\hat{\pi}^{\phi}_{-i,X_i}(c)\}_{i=1}^{N})$.
    Compute $\widehat{C}(X_i, \alpha)$ in (16) using $\hat{\pi}^{\phi}_{-i}$ and $\hat{\tau}^{\phi}_{i,cal}$.
    if $t - T = 0 \mod s$ then    ▷ Slide Scores Forward
        for $j = i - s, \dots, i - 1$ do
            Compute $\hat{\tau}^{\phi}_j := \hat{\tau}_{X_j}(Y_j)$ using (15) and $\hat{\pi}^{\phi}_{-j}$.
            $\boldsymbol{\tau} = (\boldsymbol{\tau} - \{\hat{\tau}^{\phi}_1\}) \cup \{\hat{\tau}^{\phi}_j\}$ and reset index of $\boldsymbol{\tau}$.
        end for
    end if
end for
```

IV. Theoretical Guarantee

In this section, we establish some theoretical performance guarantees for the proposed algorithms. Section (IV-A) provides parameter recovery guarantee for the point-process model defined in (3). Section (IV-B) provides coverage guarantee (see Eq. (12)) and the tightness of the fire magnitude prediction set by `ERAPS`.

A. Parameter Recovery for Point Process Model

Note that for fixed β, the problem for estimating the rest of the parameters in θ via (7) for `LinearSTHawkes` is convex (it can be shown that the objective function is concave in θ other than β, and constraints induce convex feasible domain). We can establish the following bound using a similar technique as in [52], [53]. We do not consider the bound for `NonLinearSTHawkes` in (6) because it is impossible to verify convexity for a generic feature extractor g.

We first obtain parameter recovery bound for minimizing a generic continuously differentiable strictly convex function $f(\theta) : \Theta \to \mathbb{R}$, where $\Theta \subset \mathbb{R}^p$ is a convex set. Let $F(\theta) := \nabla f(\theta)$ be the gradient of f on Θ. We know that $F(\theta)$ is *monotone* [52]:

$$\big[F(\theta) - F(\theta')\big]^T\big[\theta - \theta'\big] \geq 0 \ \forall \theta, \theta' \in \Theta.$$

Let $\theta^* \in \Theta$ be the unique global minimizer of f, which exists as f is strictly convex. To estimate θ^*, we use the projected gradient descent procedure, starting at an arbitrary $\theta_0 \in \Theta$:

$$\theta_k := \mathrm{Proj}_\Theta(\theta_{k-1} - t_k F(\theta_{k-1})), \tag{17}$$

where $t_k > 0$ determines the step size and $\mathrm{Proj}_\Theta(\hat{\theta}) := \arg\min_{\theta\in\Theta} \|\hat{\theta} - \theta\|_2$. To analyze the error $\|\theta_k - \theta^*\|_2$ after k iterations, we need the following conditions:

Assumption 1: Assume that there exist $D, \kappa, M > 0$ where

$$(i) \quad \|\theta - \theta'\|_2 \leq D \ \forall \theta, \theta' \in \Theta, \tag{18}$$

$$(ii) \quad \big[F(\theta) - F(\theta')\big]^T\big[\theta - \theta'\big] \geq \kappa\|\theta - \theta'\|_2^2 \ \forall \theta, \theta' \in \Theta, \tag{19}$$

$$(iii) \quad \|F(\theta)\|_2 \leq M \ \forall \theta \in \Theta. \tag{20}$$

We now have the following lemma that yields the error bound in (22). The proof is contained in Appendix A.

Lemma 1: Under Assumptions 1:(18)—(20) and with the step sizes

$$t_k := [\kappa(k+1)]^{-1}, \tag{21}$$

Estimates θ_k obtained through (17) obey the error bound

$$\|\theta_k - \theta^*\|_2^2 \leq \frac{M^2}{\kappa^2(k+1)}. \tag{22}$$

We can now use Lemma 1 to obtain the parameter recovery guarantee for minimizing $\ell(\theta)$ via solving (7). For a fixed $\beta > 0$, let

$$\theta[\beta] := \theta - \{\beta\} \tag{23}$$

contain all the model parameters except β when solving (7). We thus know that under Lemma 1, the estimate $\hat{\theta}[\beta]$ converges to the global minimum $\theta^*[\beta]$ at rate $1/k$. Meanwhile, since the optimal parameter β^* is non-negative scalar, we can estimate it up to arbitrary precision using one one-dimensional grid search. In particular, assume $\beta^* \in [\beta_0, \beta_1]$ with known values of β_0, β_1. For a fixed integer $J \geq 1$, divide the region $[\beta_0, \beta_1]$ into $J+1$ points $\beta_0, \ldots, \beta_J$, where

$$\beta_j := \beta_0 + \frac{j}{J}(\beta_1 - \beta_0), j = 0, \ldots, J. \tag{24}$$

Then, we can obtain estimates $\hat{\theta}[\beta_j]$ via solving (7) using the projected gradient descent procedure (17) at the fixed β_j. Given J pairs of estimates $(\beta_j, \hat{\theta}[\beta_j])$, we define

$$\hat{\theta} := \Big(\beta_{j^*}, \hat{\theta}\big[\beta_{j^*}\big]\Big) \tag{25}$$

$$j^* := \mathop{\arg\min}_{j=0,\ldots,J} \ell\Big(\Big[\beta_j, \hat{\theta}\big[\beta_j\big]\Big]\Big), \tag{26}$$

which denotes the estimate that reaches the smallest log-likelihood out of these M estimates. We then bound in the following theorem the parameter estimation error of $\hat{\theta}$ in (25). The proof is contained in Appendix B.

Theorem 1 (`LinearSTHawkes` Parameter Recovery Guarantee): Let θ^* be a minimizer of $\ell(\theta)$ in (7) under `LinearSTHawkes` in (5). Under Assumption 1:(18)—20, the estimate $\hat{\theta}$ in (25) obeys the bound

$$\|\hat{\theta} - \theta^*\|_2^2 = \mathcal{O}\bigg(\frac{1}{J^2} + \frac{1}{k+1}\bigg). \tag{27}$$

In (27), J is the number of grid searches for β^* in $[\beta_0, \beta_1]$ and k is the number of projected gradient descent step (17) of $\theta[\beta_j]$ in (23) at each search point β_j.

The implication of Theorem 1 is that we can recover the *true model* of $\lambda(t, k, m)$ in (3) for `LinearSTHawkes` in (5). This is because `LinearSTHawkes` reaches the smallest negative log-likelihood under θ^* and log likelihood is also the highest under the true model. Thus, when estimates $\hat{\theta}$ approach true parameters θ^* in ℓ_2 norm, the corresponding model estimate also recover the true model.

B. Conformal Prediction Set Guarantee

Note that in existing CP literature, it is typically assumed that observations (X_i, Y_i) are exchangeable. This assumption is unrealistic in our setting when strong correlation exists within data. Instead, we impose assumptions on the quality of estimating the non-conformity scores and on the dependency of non-conformity scores in order to bound coverage gap of (12). Most of the assumptions and proof techniques extends our earlier work [42], but we extend it to the classification setting under arbitrary definitions of non-conformity scores. In particular, we allow arbitrary dependency to exist within features X_i or responses Y_i.

Given any feature X, a possible label c, and a probability mapping p such that $\sum_{c=1}^C p_X(c) = 1, p_X(c) \geq 0$, we denote $G : (X, c, p) \to \mathbb{R}$ as an arbitrary non-conformity mapping and $\tau_X^p(c) := G(X, c, p)$ as the non-conformity score at label c. For instance, we may consider

$$G(X, c, p) = \sum_{c'=1}^C p_X\big(c'\big) \cdot \mathbb{I}\big\{p_X\big(c'\big) > p_{X_i}(c)\big\}, \tag{28}$$

which computes the total probability mass of labels that are deemed more likely than c by p. The less likely c is, the greater $\tau_i^p(c)$ is, indicating the non-conformity of label c. For notation simplicity, the oracle (resp. estimated) non-conformity score of each training datum $(X_i, Y_i), i = 1, \ldots, N$ under the true conditional distribution $\pi := P_{Y|X}$ (resp. any estimator $\hat{\pi}$) is abbreviated as $\tau_i = \tau_{X_i}^\pi(Y_i)$ (resp. $\hat{\tau}_i$).

We now impose these two assumptions that are sufficient for bounding coverage gap of (12). First, we make assumptions about the quality of estimation by the chosen classifier:

Assumption 2 (Error Bound on Estimation): Assume there is a real sequence $\{\vartheta_i\}$ where $\frac{1}{N}\sum_{j=i-N}^{i-1}(\hat{\tau}_j - \tau_j)^2 \leq \vartheta_N^2$.

Then we make assumptions about to the property of true non-conformity scores:

Assumption 3 (Regularity of Non-Conformity Scores): Assume $\{\tau_j\}_{j=i-N}^i$ are independent and identically distributed (i.i.d.) according to a common cumulative density function (CDF) F with Lipschitz continuity constant $L > 0$.

We brief remark on implications of the Assumptions above. Note that Assumption 2 essentially reduces to the point-wise estimation quality of π by $\hat{\pi}$, which may fail under data overfitting—all N training data are used to train the estimator. In this case, $\hat{\pi}$ tends to over-concentrate on the empirical conditional distribution under $(X_i, Y_i), i = 1, \dots, N$, which may not be representative of the true conditional distribution $P_{Y|X}$. A common way to avoid this in the CP literature is through data-splitting—train the estimator on a subset of training data and compute the estimated non-conformity scores $\hat{\tau}$ only on the rest training data (i.e., calibration data). However, doing so likely results in a poor estimate of π and as we will see, the theoretical guarantee heavily depends on the size of estimated non-conformity scores. On the other hand, Assumption 3 can be relaxed as stated in [42]. For instance, the oracle non-conformity scores can either follow linear processes with additional regularity conditions [42, Corollary 1] or be strongly mixing with bounded sum of mixing coefficients [42, Corollary 2]. The proof techniques directly carry over, except for slower convergence rates.

Lastly, define the empirical distributions using oracle and estimated non-conformity scores:

$$\tilde{F}(x) := \frac{1}{N} \sum_{j=i-N}^{i-1} \mathbb{I}\big(\tau_j \le x\big), \qquad \text{[Oracle]}$$

$$\hat{F}(x) := \frac{1}{N} \sum_{j=i-N}^{i-1} \mathbb{I}\big(\hat{\tau}_j \le x\big). \qquad \text{[Estimated]}$$

We then have the following coverage results at the prediction index $t > T$.

Lemma 2 [42, Lemma 2]: Suppose Assumptions 2 and 3 hold. Then,

$$\sup_x |\tilde{F}(x) - \hat{F}(x)| \le (L+1)\vartheta_N^{2/3} + 2 \sup_x |\tilde{F}(x) - F(x)|.$$

The proof of Lemma 2 appears in Appendix C.

Lemma 3 [42, Lemma 1]: Suppose Assumption 3 holds. Then, for any training size N, there is an event A within the probability space of non-conformity scores $\{\tau_j\}_{j=1}^N$, such that when A occurs,

$$\sup_x |\tilde{F}(x) - F(x)| \le \sqrt{\log(16N)/N}.$$

In addition, the complement of event A occurs with probability $\mathbb{P}(A^C) \le \sqrt{\log(16N)/N}$.

The proof of Lemma 3 appears in Appendix D.

As a consequence of Lemmas 2 and 3, the following bound of coverage gap of (12) holds:

Theorem 2 (Coverage Guarantee, [42, Th. 1]): Suppose Assumptions 2 and 3 hold. For any training size N and significance level $\alpha \in (0, 1)$, we have

$$|\mathbb{P}\big(Y_i \notin \widehat{C}(X_i, \alpha)\big) - \alpha| \le 24\sqrt{\log(16N)/N} + 4(L+1)\vartheta_N^{2/3}. \tag{29}$$

The proof of Theorem 2 appears in Appendix E. Note that Theorem 2 holds uniformly over all $\alpha \in [0, 1]$ because Lemmas 2 and 3 bound the sup-norm of differences of distributions. Hence, users in practice can select desired parameters α *after* constructing the non-conformity scores. Such a bound is also useful when building multiple prediction intervals simultaneously, under which α is corrected to reach nearly valid coverage [54].

In addition to coverage guarantee, we can analyze the convergence of $\widehat{C}(X_i, \alpha)$ to the oracle prediction set $C^*(X_i, \alpha)$ under further assumptions. Given the true conditional distribution function $\pi := P_{Y|X}$, we first order the labels so that $\pi_{X_i}(i) \ge \pi_{X_i}(j)$ if $i \le j$. Then, we have

$$C^*(X_i, \alpha) = \big\{1, \dots, c^*\big\},$$

where $c^* := \min_{c \in [C]} \sum_{k=1}^c \pi_{X_i}(k) \ge 1 - \alpha$.

Theorem 3 (Set Size Convergence Guarantee): Suppose Lemmas 2 and 3 hold and denote F^{-1} as the inverse CDF of $\{\tau_j\}_{j=i-N}^{i}$. Further assume that

(1) $c_1^* = c_2^*$ *where*

$$c_1^* := \arg\min_c \big\{\textstyle\sum_{k=1}^c \pi_{X_i}(k) \ge 1 - \alpha\big\},$$
$$c_2^* := \arg\max_c \big\{\tau_i(c) < F^{-1}(1-\alpha)\big\}.$$

(2) There exists a sequence ϑ_i' converging to zero with respect to N such that $\|\tau_i - \hat{\tau}_i\|_\infty \le \vartheta_i'$, where the ∞-norm is taken over class labels.

Then, there exists N large enough such that for all $i > N$,

$$\widehat{C}(X_i, \alpha) \Delta C^*(X_i, \alpha) \le 1, \tag{30}$$

where Δ in (30) denotes set difference.

The proof of Theorem 3 appears in Appendix F. Note that if the non-conformity score at any label c is defined in (28), which is the total probability mass of labels $c' \neq c$ that are more likely than c based on a conditional probability mapping p, then the first additional assumption (i,e., $c_1^* = c_2^*$) in Theorem 3 can be verified to hold. In general, whether this assumption is satisfied depends on the particular form of the non-conformity score.

V. Model Validation by Real-Data

We apply the proposed models on the 2014-2019 California wildfire data described in Section II. The experiment is organized as follows. Section V-A describes the setup details, including the dataset and evaluation metrics. Section V-B compares `LinearSTHawkes` with competing baselines on data from a small region. Section V-C compares `LinearSTHawkes` and `NonLinearSTHawkes` on the same region to highlight their performance differences.

A. Evaluation Metrics

We use the F_1 score for performance assessment, which is a standard metric for classification when data are *imbalanced*—note that the number of no occurrence of fire incidents (denoted as 0) significantly outweighs the other (denoted as 1). The goal is to predict as many fire occurrences as possible without making too many false positives. In our case, false positives measured at each location refers to be a prediction of fire incidents at a specific date t when there is no fire incident. Quantitatively, we define the set of fire occurrences as

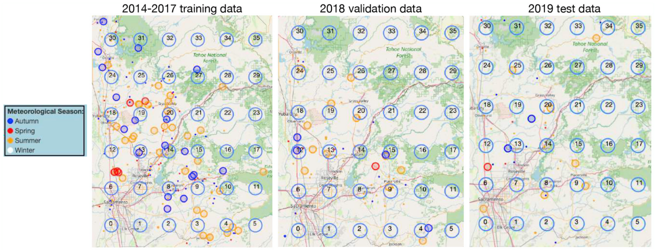


Fig. 1. Visualize data and grid discretization on data from different years. There are grid-wise shifts in data distribution—for instance, fire incidents cluster more closely around grid 12 in 2018 (validation) than in 2014—2017 (training) or in 2019 (test).

U and our predicted set as V. Then the *precision* P and *recall* R are defined as

$$P = |U \cap V|/|V|, \;\; R = |U \cap V|/|U|, \tag{31}$$

where the notation $|\cdot|$ denotes the size of the set. In the definition (31), we write P and/or R to be 1 if the ratio is 0/0 (i.e., there is no fire incident at a specific location and the model correct predicts none). The F_1 score is thus a combination: $F_1 = 2/(P^{-1} + R^{-1}) = 2PR/(P+R)$, where a high F_1 score indicates both a large of true detection and a small number of false positives. In general, when one of P and R is more important, one can consider a weighted F_1 that assigns imbalanced weights to precision and recall. We use non-weighted F_1 scores in all our experiments.

We construct dynamic thresholds to make binary prediction based on estimated fire risk $\hat{\lambda}(t, k, m)$ defined in Eq. (3). The detailed Algorithm 3 is provided in Appendix I. In particular, we observe that rate estimates $\hat{\lambda}(t, k, m)$ have clear seasonality (e.g., a sharp drop from summer to fall and a sharp rise from spring to summer). At the same time, fire incidents often occur when rate estimates suddenly increase on certain days. For instance, Figure 4 illustrates the performance of our model based on the observations above.

B. *LinearSTHawkes vs. Baselines*

We first focus on a small region because the distribution of fire incidents within the region and the performance of our model can be visualized clearly. The model is trained with incidents between 2014 and 2017 and examined on validation data in 2018. There were 238 fire occurrences in 2014-2017 and 70 in 2018. Upon consulting domain experts, we set the sides of discretized cells to be 0.24-degree in both longitude and latitude directions so that 36 non-overlapping cells cover the region. Figure 1 visualizes both the training and validation data, from which it is clear that the validation data have a much less number of actual fires; only a few grids have fires that occurred near them.

Estimated parameters. In practice, our feature m_i includes both temporal dynamic features m_d (e.g., weather information) and location-specific information m_l (e.g., road condition), so that we re-write $\gamma^T m$ as

$$\gamma^T m = \gamma_d^T m_d + \gamma_l^T m_l, \tag{32}$$

TABLE I
ESTIMATED PARAMETERS OF STATIC MARKS γ_l AND DYNAMIC MARKS γ_d DEFINED IN (32). "PHYS=" INDICATES ROAD TYPE OR EXISTING VEGETATION TYPE. A LARGER PARAMETER ESTIMATE INDICATES MORE CONTRIBUTION OF THE FEATURE TO FIRE HAZARDS. NOTE THAT *Temperature* AND *Relative Humidity* IN γ_d ALSO DEFINE THE WIDELY-USED FIRE DANGER INDEX SO THAT LinearSTHawkes SELECTS PHYSICALLY MEANINGFUL FEATURES

	Three Largest Estimates			Three Smallest Estimates		
γ_l estimate	0.301	0.231	0.184	0.046	0.024	0.008
γ_l feature name	Fire Tier1	Fire Tier2	Fire Tier3	PHYS=Developed-Roads	PHYS=Conifer	PHYS=Developed
γ_d estimate	0.57	0.472	0.46	0.217	0.117	0.02
γ_d feature name	Summer	Temperature	Relative Humidity	LFP	Spring	Winter

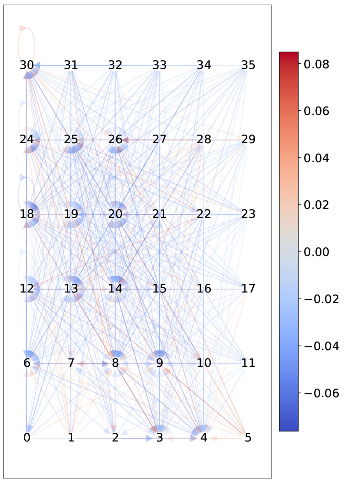


Fig. 2. The distribution of α_{ij} closely follows the data distribution in Figure 1.

which decompose the contribution of m into the sum of both terms.

Based on (32), we interpret the feature and interaction parameters of LinearSTHawkes, estimated via Algorithm 2. First, Table I shows the estimated parameters for features (i.e., marks), whose magnitude indicates feature importance. Higher magnitude of estimates contribute more significantly to the growth of fire risk. Noticeably, the top two features in γ_d (excluding summer, the seasonality parameter) are also factors in defining the *Fire Danger Index*, which is a most commonly used index for fire hazard monitoring [55]. Therefore, the model estimates of feature parameters are physically meaningful. Next, Figure 2 examines the location-to-location interaction parameters α_{ij}, which is forced to be zero if centroids of two cells exceeds 4×0.24 degrees. Values of α_{ij} above or below zero indicate excitatory or inhibitory effects from nearby and past events. The distribution of interaction effects closely aligns with the 2014—2017 training data in Figure 1. For instance, we see clusters of fire incidents in 2014-2017 training data in Figure 1 around location 20 and as a result, location 20 in Figure 2 also interacts intensively with its nearby neighbors. Quantitatively, if we use α_{ij} to roughly measure the amount of influence of location i on location j:

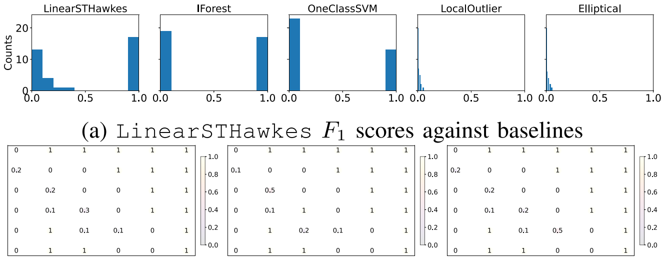
- The amount of positive influence into location 20 (i.e., $\sum_{j:\alpha_{j,20}>0} \alpha_{j,20}$) is 0.40.
- The amount of negative influence into location 20 (i.e., $\sum_{j:\alpha_{j,20}<0} \alpha_{j,20}$) is -0.30.
- The amount of positive influence from location 20 (i.e., $\sum_{j:\alpha_{20,j}>0} \alpha_{20,j}$) is 0.29.

- The amount of negative influence from location 20 (i.e., $\sum_{j:\alpha_{20,j}<0} \alpha_{20,j}$) is -1.44.

In addition, we can perform *counterfactual analyses* using the estimated parameters: suppose a decision-maker wants to know the increase in risk when an external condition changes from A to B (e.g., Fire tier zone shift, changes in vegetation types, etc.). Then, the change in risk at a certain location and time is $\Delta(A, B) := \lambda(t, k, B) - \lambda(t, k, A)$. Similar analyses can be performed for a change in location from k to k_1. Such analyses can help one better study the effect of different factors on fire risks, making risk management more effective.

Prediction results. We first compare LinearSTHawkes with several one-class classification baselines. We choose isolation forest [56], one-class SVM [57], local outlier factor [58], and elliptic envelope [59] due to their popularity and generality. These classifiers, including static and dynamic marks, use the same data as LinearSTHawkes. Figure 3(a) visualizes the histograms of F_1 scores by each method, which show that LinearSTHawkes outperforms competing methods by yielding less zero F_1 scores and more one F_1 scores. Note that zero (resp. one) F_1 scores appear at locations that are the easiest (resp. hardest) to predict discussed earlier. In addition, LinearSTHawkes can yield non-trivial fractional F_1 scores at other locations by capturing a decent number of true positives. Nevertheless, our model also yields many zero F_1 scores because the task is inherently challenging: it makes 365 daily predictions at each of 36 locations, in a total of 13140 predictions, when there are only 70 actual fire occurrences across all 36 locations.

We now illustrate the location-wise prediction results of LinearSTHawkes. Figure 3(b)—3(d) visualizes F_1 score, recall, and precision on each of the 36 location. The result helps us assess the prediction difficulty at various locations, where we suspect the difficulty arises partially due to the distribution shift of data in 2018 comparing to data in 2014-17 (cf. Figure 1). To better illustrate how LinearSTHawkes makes a prediction, we further visualize in Figure 4 the trajectory of rate prediction on top of actual incidents. Dynamic thresholds are obtained by using Algorithm 3. The figure shows that sharp increases in predicted fire risks tend to occur near true fire events, which helps us make correct predictions. In the future, to reduce the number of false positives, we may refit the model parameters during validation using newly observed incidents.



(a) LinearSTHawkes F_1 scores against baselines

(b) On map: F_1 score (c) Recall (d) Precision

Fig. 3. Comparison across methods (top) and LinearSTHawkes performance per location (bottom). Histograms of F_1 scores over all locations on the top row show that our LinearSTHawkes outperforms other methods by yielding fewer zero F_1 scores, a moderate number of fractional F_1 scores, and more one F_1 scores. The bottom row visualizes the F_1 score, recall, and precision of LinearSTHawkes at each location.

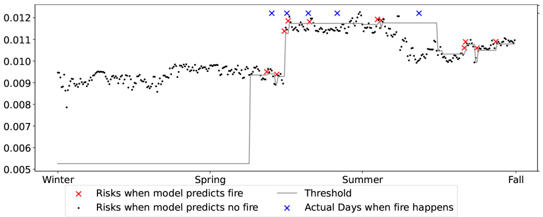
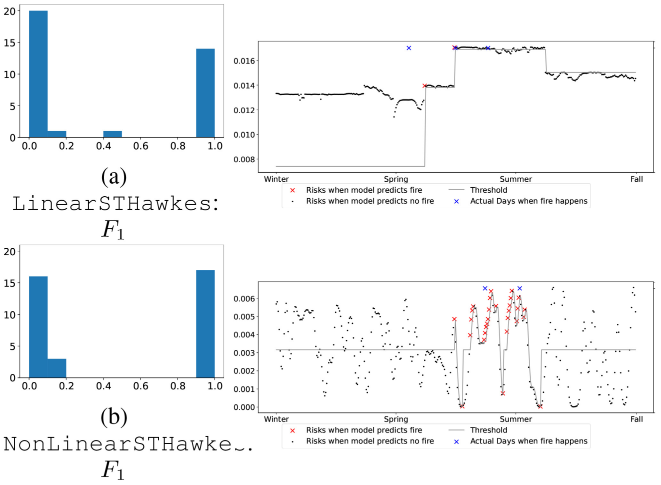


Fig. 4. Real-time prediction of fire risks and incidents on top of actual incidents and dynamic thresholds. The prediction by LinearSTHawkes can closely match the actual data.



(a) LinearSTHawkes: F_1

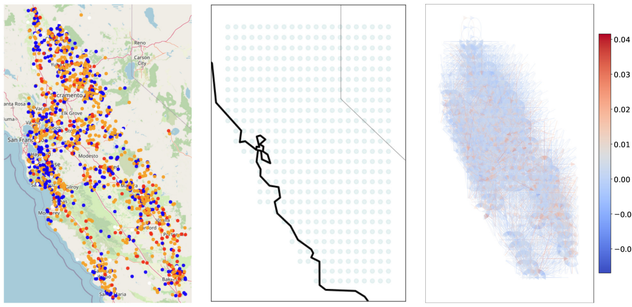
(b) NonLinearSTHawke.. F_1

Fig. 5. Compare LinearSTHawkes with NonLinearSTHawkes on 2019 test data. Both models are trained on 2014-2018 data. The top row shows results under LinearSTHawkes, and the bottom row shows those under NonLinearSTHawkes. In comparison, NonLinearSTHawkes shows improved performance because of a more flexible feature extractor and the ability to yield less zero F_1 scores.

C. Compare LinearSTHawkes vs. NonLinearSTHawkes

We now compare LinearSTHawkes and NonLinearSTHawkes on 2019 test data (cf. Figure 1 right), where we train the feature extractor $g(m|t, k)$ in (6) using the one-class SVM. In principle, one can use any feature extractor, but we choose SVM due to the flexibility of the kernel function. Based on earlier results, we only include seasonal and weather information, LFP, and FPI in the dynamic marks.

Figure 5 compares the performance of both methods and there are several observations. First, the histograms of F_1 scores (cf. Figure 5(a) & 5(b)) show that NonLinearSTHawkes performs better than LinearSTHawkes, as the former yields more non-zero F_1 scores. To explain the improvement, we found the empirical distribution of estimates $g(m|t, k)$ by NonLinearSTHawkes to closely match the Frechet distribution, a classic example from *extreme value theory* [60]. Although the Frechet distribution is not used to aid modeling, the connection allows NonLinearSTHawkes to make a more accurate prediction because many rare events (e.g., fire incidents)



(a) Raw data (b) Grid discretization (c) Interaction α_{ij}

Fig. 6. Data visualization. (a) shows fire events colored by season as in Figure 1, (b) shows the grid discretization, and (c) visualizes the location-location interaction matrix parameters α_{ij}.

follow the Frechet distribution. Further discussions appear in Appendix J. Second, the trajectory of predicted fire risks by NonLinearSTHawkes (cf. Figure 5, lower right) fluctuates much more than LinearSTHawkes (cf. Figure 5, top right). For this prediction task, such fluctuation enables better detection because actual fire incidents are often associated with sudden risk increases.

Remark 1 History-Dependent Mark in NonLinearSTHawkes): Accumulated weather conditions can often induce fire events (e.g., several dry days earlier can lead to elevated fire risks). Thus, it seems natural to include in each m_i additional spatio-temporal marks to account for accumulation effects. However, doing so has two drawbacks:

1) Data acquisition and storage are much more expensive. One must collect a complete record of historical marks at each grid to fit the models. The issue mainly arises when the number of grids is large (e.g., hundreds) and marks frequently arrive (e.g., hourly).
2) The curse of dimensionality rises when each mark contains longer historical values. Note that the total number of fire incidents is fixed and typically small (e.g., hundreds over multiple years). Therefore, parameter estimation can be more difficult as the feature dimension increases. How to choose historical values appropriately to reduce the effect of this issue would increase difficulty in training.

VI. Large-Scale Data Validation

We now show that our LinearSTHawkes and NonLinearSTHawkes are scalable to a large region with much more fire incidents and locations. There are a total of 2011 fire occurrences in this region, comprising 63% of total wildfire incidents in California from 2014 to 2019. Figure 6(a) visualizes fire incidents within the region on the map, and Figure 6(b) illustrates the resulting 453 grids after discretization into squares with side lengths equal to 0.24 degrees; we remove regions that lie inside the ocean. Most grids have no fire in the 5-year horizon since fire incidents seem to cluster near the coastal line with large populations. We remark that the setup and hyperparameter choices are the same as those in Section V-B. The distribution of estimated interaction parameters α_{ij} (cf. Figure 6(c)) still closely align with that of the actual data. For instance, Figure 6(a) shows there are clusters of true fire incidents around the coastal line on the west side and few incidents in the mid-south side. As a result, estimates in Figure 6(c) are much denser in distribution around the west side than around the mid-south side. As a concrete example, location 140 is on the west side along the coastal line, where there are clusters of fire incidents. Quantitatively, if we use α_{ij} to roughly measure the amount of influence of location i on location j:

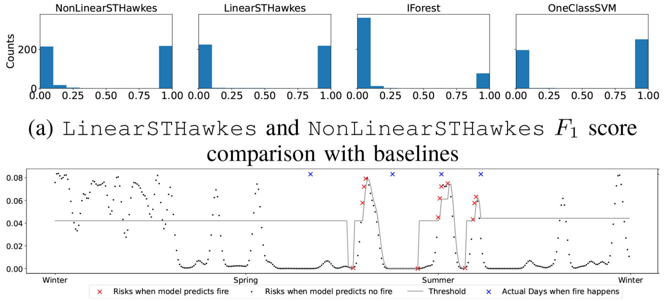
- The amount of positive influence into location 140 (i.e., $\sum_{j:\alpha_{j,140}>0} \alpha_{j,140}$) is 0.17.
- The amount of negative influence into location 140 (i.e., $\sum_{j:\alpha_{140}<0} \alpha_{j,140}$) is -0.30.
- The amount of positive influence from location 140 (i.e., $\sum_{j:\alpha_{140,j}>0} \alpha_{140,j}$) is 0.23.
- The amount of negative influence from location 140 (i.e., $\sum_{j:\alpha_{140,j}<0} \alpha_{140,j}$) is -0.47.

In comparison, location 20 is in the mid-south region of few clusters of fire incidents. Quantitatively, if we use α_{ij} to roughly measure the total influence of location i on location j:

- The amount of positive influence into location 20 (i.e., $\sum_{j:\alpha_{j,20}>0} \alpha_{j,20}$) is 0.00.
- The amount of negative influence into location 20 (i.e., $\sum_{j:\alpha_{j,20}<0} \alpha_{j,20}$) is -0.09.
- The amount of positive influence from location 20 (i.e., $\sum_{j:\alpha_{20,j}>0} \alpha_{20,j}$) is 0.00.
- The amount of negative influence from location 20 (i.e., $\sum_{j:\alpha_{20,j}<0} \alpha_{20,j}$) is 0.00.

A. Real-Time Fire Risk Prediction

Figure 7(a) compares the prediction performances of NonLinearSTHawkes, LinearSTHawkes, IForest, and OneClassSVM. We see that NonLinearSTHawkes performs better than both the LinearSTHawkes and the isolation forest by yielding more non-zero F_1 scores and a large number of F_1 scores being one. Due to its flexible feature extractor, the NonLinearSTHawkes is also competitive against the one-class SVM; importantly, it yields more F_1 scores between zero and one, making it



(a) LinearSTHawkes and NonLinearSTHawkes F_1 score comparison with baselines

(b) NonLinearSTHawkes real-time prediction

Fig. 7. On 2019 test data: The top row compares the histograms of F_1 score under various methods. The leftmost NonLinearSTHawkes has the most number of non-zero F_1 scores, with many being 1. The bottom row visualizes the temporal predicted risks by NonLinearSTHawkes at one grid. Overall, NonLinearSTHawkes yields the best performance among all models.

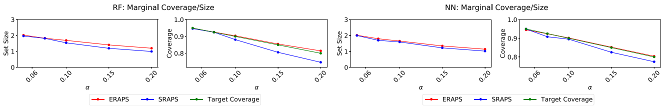


Fig. 8. Marginal coverage (12) and size of prediction sets by ERAPS and SRAPS under the random forest classifier and the neural network classifier. ERAPS always maintains desired coverage, whereas competing methods can fail to do so.

more informative than the one-class SVM on certain locations. Hence, NonLinearSTHawkes maintains improved performance than other models even if the number of grids significantly increases. Figure 7(b) further visualizes the real-time prediction behavior of NonLinearSTHawkes, where the peaks identified as fire incidents closely align with the actual incidents.

B. Fire Magnitude Conformal Prediction Sets

We show that prediction sets by ERAPS maintain desired coverage defined in (12). Data in 2014-2018 are training data, and data in 2019 are test data, where there are a total of five possible fire magnitude. Both the random forest classifier (RF) and the neural network classifier (NN) are used as prediction algorithms; their setup is the same as those in [51]. We let regularization parameters $(\lambda, k_{\text{reg}}) = (1, 2)$ as suggested in [51]. Figure 8 shows marginal coverage under both classifiers, where we also compare ERAPS against a competing method titled *split regularized adaptive prediction set* (SRAPS) [36]. The details of SRAPS are described in [51, Algorithm 1]. We have two findings. First, ERAPS performs very similarly under both classifiers and always maintains $1-\alpha$ coverage, whereas SRAPS tends to lose coverage at different values of α. Thus, ERAPS is more robust and consistent in terms of coverage. Second, both methods return prediction sets with almost the same sizes, but ERAPS is preferable due to its ability to maintain near $1-\alpha$ coverage.

VII. Conclusion and Discussions

We have developed a predictive framework for wildfire risk and magnitude using multi-modal sensing data, based on a mutually exciting spatio-temporal point process model as well as time series CP set. We established performance guarantees of the proposed methods, and demonstrate the good performance on large-scale real data experiments. Overall, our method is efficient in model parameter, enjoys interpretability, accurate prediction against existing methods. There are several future works. Regarding the point process model, we can consider beyond the parametric forms in (4) and (5), such as the more general neural network-based formulations. The development of dynamic marks in Algorithm 3 can also be refined. Regarding conformal uncertainty quantification, remaining questions include how to better utilize the existing time-series method when data have an additional spatial dimension.

From our numerical results, we observe that distribution shifts may exist sometime for wildfire prediction. Although our LinearSTHawkes and NonLinearSTHawkes are not designed to explicitly consider distribution shift, they still yield improved performance against baseline models on real data. In particular, as shown in Fig. 3(a) on small-scale data and Fig. 7 on large-scale data, our proposed models always outperform the baseline one-class classifiers. As a result, although the performance of our proposed framework may vary from year to year, it is still preferable in terms of predictive ability. We believe this is due to the model design to capture spatial-temporal information (e.g., past fire incidents around neighbors) and mark contribution (e.g., how multi-modal sensor information contributes to fire risks). To mitigate the adverse effects of distribution shifts, one approach is to introduce uncertainty into model parameters. For instance, instead of specifying the parameters in the optimization problem (8) as unknown constants in our models, one could allow them to vary within a pre-specified range (or even treat them as random variables). With accurate parameter estimation, the estimated model could better address model shifts that arise from distribution shifts in test data. However, we do not explore this model design in this work, as our goal is to propose simple yet effective models for capturing fire risks using multi-modal data and establishing theoretical guarantees based on the proposed models (see Theorem IV-A).

References

[1] M. O. Andreae and P. Merlet, "Emission of trace gases and aerosols from biomass burning," *Global Biogeochem. Cycles*, vol. 15, no. 4, pp. 955–966, 2001.

[2] "Wildfire and wildfire safety—Cpuc.ca.gov." Accessed: Oct. 7, 2022. [Online]. Available: https://www.cpuc.ca.gov/industries-and-topics/wildfires

[3] "Fire weather week 2 forecasts—Cpc.ncep.noaa.gov." Accessed: Oct. 7, 2022. [Online]. Available: https://www.cpc.ncep.noaa.gov/products/people/mchen/fireWeather/cpc_wk2fw_index.html

[4] "NFDRS system inputs and outputs | NWCG—Nwcg.gov." Accessed: Oct. 7, 2022. [Online]. Available: https://www.nwcg.gov/publications/pms437/fire-danger/nfdrs-system-inputs-outputs

[5] "Detecting wildfire | environment and natural resources—Enr.gov.nt.ca." Accessed: Oct. 7, 2022. [Online]. Available: https://www.enr.gov.nt.ca/en/services/wildfire-operations/detecting-wildfire

[6] A. Srinivasan and J. Wu, "A survey on secure localization in wireless sensor networks," in *Encyclopedia of Wireless and Mobile Communications*, vol. 126. Boca Raton, FL, USA: CRC Press, 2007.

[7] B. S. Lee, M. E. Alexander, B. C. Hawkes, T. J. Lynham, B. J. Stocks, and P. Englefield, "Information systems in support of wildland fire management decision making in Canada," *Comput. Electron. Agr.*, vol. 37, pp. 185–198, Dec. 2002.

[8] B. M. Wotton, "Interpreting and using outputs from the Canadian forest fire danger rating system in research applications," *Environ. Ecol. Stat.*, vol. 16, pp. 107–131, Mar. 2008.

[9] P. Jain, S. C. P. Coogan, S. G. Subramanian, M. Crowley, S. Taylor, and M. D. Flannigan, "A review of machine learning applications in wildfire science and management," *Environ. Rev.*, vol. 28, no. 4, pp. 478–505, 2020.

[10] A. Jaafari, E. K. Zenner, M. Panahi, and H. Shahabi, "Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability," *Agr. Forest Meteorol.*, vols. 266–267, pp. 198–207, Mar. 2019.

[11] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, pp. 83–90, Apr. 1971.

[12] A. Reinhart, "A review of self-exciting spatio-temporal point processes and their applications," *Stat. Sci.*, vol. 33, no. 3, pp. 299–318, Aug. 2018, doi: 10.1214/17-STS629.

[13] F. P. Schoenberg, H.-C. Chang, J. E. Keeley, J. Pompa, J. Woods, and H. Xu, "A critical assessment of the burning index in Los Angeles County, California," *Int. J. Wildland Fire*, vol. 16, no. 4, pp. 473–483, 2007.

[14] L. A. Sanabria, X. Qin, J. Li, R. P. Cechet, and C. Lucas, "Spatial interpolation of McArthur's forest fire danger index across Australia: Observational study," *Environ. Model. Softw.*, vol. 50, pp. 37–50, Dec. 2013.

[15] W. H. Frandsen, "Ignition probability of organic soils," *Can. J. Forest Res.*, vol. 27, pp. 1471–1477, Sep. 1997.

[16] M. P. Plucinski and W. R. Anderson, "Laboratory determination of factors influencing successful point ignition in the litter layer of Shrubland vegetation," *Int. J. Wildland Fire*, vol. 17, no. 5, pp. 628–637, 2008.

[17] A. A. Cunningham and D. L. Martell, "A stochastic model for the occurrence of man-caused forest fires," *Can. J. Forest Res.*, vol. 3, pp. 282–287, Jun. 1973.

[18] H. Xu and F. P. Schoenberg, "Point process modeling of wildfire hazard in Los Angeles County, California," *Ann. Appl. Stat.*, vol. 5, pp. 684–704, Jun. 2011.

[19] J. Koh, F. Pimont, J.-L. Dupuy, and T. Opitz, "Spatiotemporal wildfire modeling through point processes with moderate and extreme marks," *Ann. Appl. Stat.*, vol. 17, no. 1, pp. 560–582, 2023.

[20] E. Gabriel and P. J. Diggle, "Second-order analysis of inhomogeneous spatio-temporal point process data," *Statistica Neerlandica*, vol. 63, no. 1, pp. 43–51, 2009.

[21] P. J. Diggle, *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Boca Raton, FL, USA: CRC Press, 2013.

[22] A. C. Miller, L. Bornn, R. P. Adams, and K. Goldsberry, "Factorized point process intensities: A spatial analysis of professional," in *Proc. ICML*, 2014, pp. 235–243.

[23] J. D. Scargle, "An introduction to the theory of point processes, vol. I: Elementary theory and methods," *Technometrics*, vol. 46, no. 2, p. 257, 2004.

[24] S. Zhu and Y. Xie, "Spatiotemporal-textual point processes for crime linkage detection," *Ann. Appl. Stat.*, vol. 16, no. 2, pp. 1151–1170, 2022. [Online]. Available: https://doi.org/10.1214/21-AOAS1538

[25] L. Holden, S. Sannan, and H. Bungum, "A stochastic marked point process model for earthquakes," *Nat. Hazards Earth Syst. Sci.*, vol. 3, nos. 1–2, pp. 95–101, 2003.

[26] H. Mei and J. Eisner, "The neural Hawkes process: A neurally self-modulating multivariate point process," in *Proc. NIPS*, 2017, pp. 6757–6767.

[27] S. Li, S. Xiao, S. Zhu, N. Du, Y. Xie, and L. Song, "Learning temporal point processes via reinforcement learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 10804–10814.

[28] S. Zuo, H. Jiang, Z. Li, T. Zhao, and H. Zha, "Transformer Hawkes process," in *Proc. ICML*, 2020, pp. 11692–11702.

[29] S. J. Hardiman, N. Bercot, and J.-P. Bouchaud, "Critical reflexivity in financial markets: A Hawkes process analysis," *Eur. Phys. J. B*, vol. 86, pp. 1–9, Oct. 2013.

[30] R. Kobayashi and R. Lambiotte, "TiDeH: Time-dependent Hawkes process for predicting Retweet dynamics," in *Proc. ICWSM*, 2016, pp. 191–200.

[31] E. Choi, N. Du, R. Chen, L. Song, and J. Sun, "Constructing disease network and temporal progression model via context-sensitive Hawkes process," in *Proc. IEEE Int. Conf. Data Min.*, 2015, pp. 721–726.

[32] F. Gerhard, M. Deger, and W. A. Truccolo, "On the stability and dynamics of stochastic spiking neuron models: Nonlinear Hawkes process and point process GLMs," *PLoS Comput. Biol.*, vol. 13, no. 2, 2017, Art. no. e1005390.

[33] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *J. Mach. Learn. Res.*, vol. 9, pp. 371–421, Jun. 2008.

[34] M. Fontana, G. Zeni, and S. Vantini, "Conformal prediction: A unified review of theory and new challenges," *Bernoulli*, vol. 29, no. 1, pp. 1–23, 2023.

[35] Y. Romano, M. Sesia, and E. Candès, "Classification with valid and adaptive coverage," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3581–3591.

[36] A. N. Angelopoulos, S. Bates, M. Jordan, and J. Malik, "Uncertainty sets for image classifiers using conformal prediction," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=eNdiU_DbM9

[37] M. Eklund, U. Norinder, S. Boyer, and L. Carlsson, "The application of conformal prediction to the drug discovery process," *Ann. Math. Artif. Intell.*, vol. 74, pp. 117–132, Sep. 2013.

[38] N. Bosc, F. Atkinson, E. Felix, A. Gaulton, A. Hersey, and A. R. Leach, "Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery," *J. Cheminformatics*, vol. 11, p. 4, Jan. 2019.

[39] J. Smith, I. Nouretdinov, R. Craddock, C. R. Offer, and A. Gammerman, "Anomaly detection of trajectories with kernel density estimation by conformal prediction," in *Proc. AIAI Workshops*, 2014, pp. 271–280.

[40] R. J. Tibshirani, R. F. Barber, E. Candès, and A. Ramdas, "Conformal prediction under covariate shift," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, pp. 2530–2540.

[41] S. Park, E. Dobriban, I. Lee, and O. Bastani, "PAC prediction sets under covariate shift," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–33. [Online]. Available: https://openreview.net/forum?id=DhP9L8vIyLc

[42] C. Xu and Y. Xie, "Conformal prediction for time series," 2020, *arXiv:2010.09107*.

[43] C. Xu and Y. Xie, "Conformal anomaly detection on spatio-temporal observations with missing data," 2021, *arXiv:2105.11886*.

[44] K. Stankevivciūtė, A. M. Alaa, and M. van der Schaar, "Conformal time-series forecasting," in *Proc. NeurIPS*, 2021, pp. 1–13.

[45] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," 2022, *arXiv:2202.13415*.

[46] "California public utilities commission (CPUC)." Accessed: Oct. 7, 2022. [Online]. Available: https://www.cpuc.ca.gov/wildfires

[47] "LaNDFiRE program: Home—-Landfire.gov." Accessed: Oct. 7, 2022. [Online]. Available: https://www.landfire.gov/

[48] "NLDaS: North American land data assimilation system | NCaR—Climate data guide—Climatedataguide.ucar.edu." Accessed: Oct. 7, 2022. [Online]. Available: https://climatedataguide.ucar.edu/climate-data/nldas-north-american-land-data-assimilation-system

[49] "Fire danger forecast | U.S. geological survey—Usgs.gov." Accessed: Oct. 7, 2022. [Online]. Available: https://www.usgs.gov/fire-danger-forecast

[50] S. Diamond and S. Boyd, "CVXPY: A python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 83, pp. 2909–2913, 2016.

[51] C. Xu and Y. Xie, "Conformal prediction set for time-series," 2022, *arXiv:2206.07851*.

[52] A. B. Juditsky and A. S. Nemirovski, "Signal recovery by stochastic optimization," *Autom. Remote Control*, vol. 80, no. 10, pp. 1878–1893, 2019.

[53] M. Zhang, C. Xu, A. Sun, F. Qiu, and Y. Xie, "Solar radiation ramping events modeling using spatio-temporal point processes," 2021, *arXiv:2101.11179*.

[54] A. Farcomeni, "A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion," *Stat. Methods Med. Res.*, vol. 17, pp. 347–388, Aug. 2008.

[55] "Wildland fire danger index (FDI) / links and information / fire weather / wildland fire / forest &amp; wildfire / home—Florida department of agriculture &amp; consumer services—Fdacs.gov." Accessed: Oct. 7, 2022. [Online]. Available: https://www.fdacs.gov/Forest-Wildfire/Wildland-Fire/Fire-Weather/Links-and-Information/Wildland-Fire-Danger-Index-FDI

[56] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Min.*, 2008, pp. 413–422.

[57] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[58] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. SIGMOD*, 2000, pp. 93–104.

[59] P. J. Rousseeuw and K. van Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[60] L. De Haan and A. Ferreira, *Extreme Value Theory: An Introduction*, vol. 3. New York, NY, USA: Springer, 2006.

[61] M. R. Kosorok, *Introduction to Empirical Processes and Semiparametric Inference*, vol. 61. New York, NY, USA: Springer, 2008.

[62] M. C. Grant and S. P. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*. London, U.K.: Springer 2008, pp. 95–110.

[63] M. Raginsky, R. M. Willett, C. Horn, J. Silva, and R. F. Marcia, "Sequential anomaly detection in the presence of noise and limited feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5544–5562, Aug. 2012.

[64] D. E. A. Sanders, "The modelling of extreme events," *Brit. Actuarial J.*, vol. 11, no. 3, pp. 519–557, 2005.