Impact of Geophysical and Anthropogenic Factors on Wildfire Size: A Spatiotemporal Data-Driven Risk Assessment Approach Using Statistical Learning

Nima Masoudvaziri

University at Buffalo

Prasangsha Ganguly

University at Buffalo

Sayanti Mukherjee (✉ sayantim@buffalo.edu)

University at Buffalo School of Engineering and Applied Sciences <https://orcid.org/0000-0002-5833-1617>

Kang Sun

University at Buffalo

Research Article

Keywords: Wildfire size, Spatiotemporal analysis, Risk assessment, Statistical learning, Predictive analytics, Geophysical and anthropogenic factors

Posted Date: June 2nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-539684/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Stochastic Environmental Research and Risk Assessment on October 30th, 2021. See the published version at <https://doi.org/10.1007/s00477-021-02087-w>.

Impact of geophysical and anthropogenic factors on wildfire size: A spatiotemporal data-driven risk assessment approach using statistical learning

Nima Masoudvaziri · Prasangsha Ganguly · Sayanti Mukherjee · Kang Sun

Received: date / Accepted: date

Abstract Wildfire spread is a stochastic phenomenon driven by a multitude of geophysical and anthropogenic factors. In this study, we propose a spatiotemporal data-driven risk assessment framework to understand the effect of various geophysical/anthropogenic factors on wildfire size, leveraging a systematic machine learning approach. We apply this framework in the state of California—the most vulnerable US state to wildfires. Using county-level annual wildfire data from 2001-2015, and various geophysical (e.g., landcover, wind, surface temperature) and anthropogenic features (e.g., population density, housing type), we trained, tested, and validated a suite of ensemble tree-based learning algorithms to identify and evaluate the key factors associated with wildfire size. The extreme gradient boosting (XGBoost) algorithm outperformed all the other models in terms of generalization performance, categorization of important features, and risk performance. We found that standard deviations of meteorological variables with long-tailed distributions play a key role in predicting wildfire size. Specifically, the top ten factors associated with high risk of larger wildfires include larger standard deviations of surface temperature and vapor pressure deficit, higher wind gust, more grassy and barren land covers, lower night-time boundary layer height and higher population density.

Nima Masoudvaziri
Department of Civil, Structural, and Environmental Engineering
University at Buffalo, The State University of New York
E-mail: nimamaso@buffalo.edu

Prasangsha Ganguly
Department of Industrial and Systems Engineering
University at Buffalo, The State University of New York
E-mail: prasangs@buffalo.edu

Sayanti Mukherjee (corresponding author)
Department of Industrial and Systems Engineering
University at Buffalo, The State University of New York
E-mail: sayantim@buffalo.edu

Kang Sun
Department of Civil, Structural, and Environmental Engineering
University at Buffalo, The State University of New York
E-mail: kangsun@buffalo.edu

Our proposed risk assessment framework will help federal/state decision-makers to adequately plan for wildfire risk mitigation and resource allocation strategies.

Keywords Wildfire size · Spatiotemporal analysis · Risk assessment · Statistical learning · Predictive analytics · Geophysical and anthropogenic factors

1 Introduction

Wildfire is a highly dynamic phenomenon comprising of various physical interactions at different spatiotemporal scales. Just after the initial ignition, various environmental conditions (e.g., wind, fuel type and moisture) may impact the fire spread. Moreover, there are several, often coupled, factors such as updraft plume and released moisture that influence these conditions. Therefore, computational cost of simulating a wildfire spread highly depends on the level of complexity considered in the analysis as well as the spatial scale of the simulation. Nonetheless, there exist sufficiently-detailed, accurate simulation tools that leverage numerical models like the Coupled Atmosphere-Wildland Fire Environment (*CAWFE*) model (Coen, 2013; Coen et al., 2020) and *WRF – SFIRE* or simply *SFIRE* (Spread FIRE) model (Mandel et al., 2011, 2014) to simulate the spread of a single wildfire incident. Although these models are useful in understanding behavior of specific wildfire events, they cannot be used in studying wildfire spread patterns in a large area (e.g., a state, a country, or a continent) due to prohibitive computational cost. To address this issue, models like the large-fire simulator system (Finney et al., 2011) are used to investigate wildfire spread in larger areas. Such models comprise of simplified equations and assume fire-driving variables spatially or temporally invariant to keep the computational costs manageable. These models run comparatively faster; thus, stochastic studies can be conducted and probabilistic maps can be produced leveraging such models (Thompson et al., 2011).

Modeling a specific wildfire incident at different stages is subject to a range of uncertainties, mostly associated with the representation of surrounding environment and weather conditions along with anthropogenic factors (Thompson and Calkin, 2011; Boulanger et al., 2018; Daniel et al., 2017; Amatulli et al., 2007). Given these uncertainties and stochastic nature of wildfires, looking at a single fire incident that occurs at a certain time and location usually is not instructive enough to represent the general fire activities and their patterns in a region. Thus, to better understand the patterns of wildfire activities in a region, a macro-level, regional, probabilistic model needs to be developed to help informing decision-makers about the risk of wildfire spread in a region, given its environmental conditions.

Recently, there is an expansion in data-driven research studying different aspects of wildfires including modeling the uncertainties and understanding the patterns of wildfire activities in a region. This has been facilitated by the availability of informative, robust and reliable datasets along with cheaper computational costs (Jain et al., 2020). Some studies focus on understanding and predicting the occurrence of wildfire incidents (Sayad et al., 2019). Other studies investigate the number of incidents or final size of fires that have occurred within a specific region or during a particular time-period (Yang et al., 2015; McCandless et al., 2020; Bendick and Hoylman, 2020). From a methodological perspective, different statistical methods such as regression models, cluster analyses, and time-series models have

been implemented (Cortez and Morais, 2007; Rodrigues and de la Riva, 2014). Given the complexity of factors influencing a wildfire spread, diverse features have been explored. Anthropogenic factors like population, roads, power lines, and developments and geophysical variables such as meteorological and climatic factors (e.g., temperature, precipitation, drought, wind), vegetation type and density, and topography are considered in most of the existing studies (Amraoui et al., 2015; Costafreda-Aumedes et al., 2018; Rodrigues et al., 2018; Shen et al., 2019).

Although data-driven studies are gaining popularity, they are still beset with a number of challenges. One of the foremost challenges is the significant variation in the spatial and temporal scales of wildfire characteristics. For example, there are large wildfires that burn hundreds of thousands of acres over weeks or months (e.g., Mendocino Complex Fire occurred in 2018, California, U.S.), and there are small fires that burn a few acres in the matter of hours and then put out naturally. Moreover, environmental features that are known to be influential in wildfire studies are highly heterogeneous both spatially and temporally. This heterogeneity signifies the importance of scales in which the features are reported in. As an example, topography is known to affect the wildfire behavior by influencing wind flow (Nelson, 2002; Viegas, 2004; Linn et al., 2007). A high-resolution digital elevation model (DEM) is necessary to capture the complex terrain effects on fire spread. In contrast to the static DEM, high temporal resolution is essential for meteorological drivers of fire spread. Hence, the effect of spatiotemporal scales renders it to be a labyrinthine process to capture a wildfire and its surrounding setting's information, adequately.

In addition, when data-driven studies are conducted at a scale that is large enough for managerial purposes (e.g., provincial, nationwide), they usually analyze spatially and temporally aggregated clusters of wildfires. The results from such analyses are proven to be sensitive to the extent of aggregation (Juan et al., 2012; Costafreda-Aumedes et al., 2016; Koutsias et al., 2010; Gralewicz et al., 2012). Correspondingly, features are also clustered, and given their inherent heterogeneity, the levels of spatially and temporally averaging will impact the explainability of the features. The temporal variations across feature clusters (e.g., seasonality) are demonstrated to be especially important and their consequential effects are usually referred to as 'preconditioning' (Urbietal et al., 2015; Trigo et al., 2016; Littell et al., 2009; Gedalof et al., 2005; Crimmins and Comrie, 2005).

To address the aforementioned challenges, our proposed framework introduces a generalized data-driven approach that is flexible to accommodate features with different spatial and temporal resolutions. More specifically, this wildfire risk assessment framework will: i) capture the stochastic behavior of the wildfire activities; ii) identify the key geophysical and anthropogenic factors that drive wildfire spread; and iii) understand their impacts on the size of the wildfire. Assembling a comprehensive dataset and leveraging a suite of non-parametric ensemble tree-based machine learning algorithms, we perform a county-level analysis with an objective to model the nexus between extent of acres burnt annually in each county and the corresponding geophysical and anthropogenic factors (e.g., land cover, surface temperature, wind, population density, etc.). The proposed framework goes beyond the existing deterministic models with linear architecture, and uses the state-of-the-art statistical learning techniques to identify and assess the key factors associated with the wildfire spread within various counties in the state of California, while accounting for the uncertainties for the future. The models'

performance is evaluated based on both goodness-of-fit and out-of-sample predictive accuracy to ensure high generalization performance as well as its ability to explain the variance in the historical data. Then, the model with the best performance in terms of generalizability and goodness-of-fit is used to characterize the relationships between the geophysical and anthropogenic factors and the extent of wildfire spread in California. We selected the state of California as a case study to establish our framework because:

- California is the third largest state by area and the most populated state in the U.S. (U.S. Census Bureau, 2012) and experiences numerous wildfire incidents of varied sizes annually (Short, 2017; California Department of Forestry and Fire Protection, 2020a). This enables us to capture a wide variation of information in both space and time.
- Wildfires in California are becoming more destructive and some of the costliest wildfire incidents in the U.S. has been recorded in this state (California Department of Forestry and Fire Protection, 2020b; Insurance Information Institute, 2020).

However, although we established our framework for the state of California, it is generalizable to other states (or any region/scale of interest), given the availability of the data.

The remaining of the paper is structured in the following way: Section 2 presents description of the data used, their format and the sources from which data is collected. Section 3 describes the methodology and research framework. This includes illustration of different steps involved in data processing, implementing algorithms, choosing a set of candidate models, selection of the final model, and methods used for statistical inference. In this section, the steps corresponding to the case study are also discussed. Section 4 presents the results of our case study. Finally, limitations and future work of the study along with the concluding remarks are summarized in sections 5 and 6.

2 Data Description

In this section, we describe the data (response and input/predictor variables) used in this study. We also provide an overview of the various data pre-processing steps leveraged to obtain the final dataset used for consecutive analysis, i.e., predicting the wildfire size and evaluating the key anthropogenic and geophysical factors that increase the risk of wildfire burnt acres.

2.1 Geographic boundaries

California Geographic Boundaries dataset in Shapefile format (California Department of Technology, 2019) provides the perimeter of counties which is used for different geospatial calculations such as assignment of wildfire incidents, calculating the land area in each county, the percent of the land covered by a specific type of vegetation, etc. Details of data pre-processing are provided in the section 3.1.3.

2.2 Response variable: Wildfire-induced burnt acres

We calculate the total number of acres burned annually in each county using Fire Program Analysis fire-occurrence database (FPA-FOD). This database provides consistent information on historical wildfire events and their characteristics such as ignition date and coordinate, size, confinement date, etc., which are reported by different agencies across the United States (Short, 2017).

2.3 Input/predictor variables: Anthropogenic and geophysical factors

Estimates of population and number of houses in a county are used in this study as indicators of anthropogenic activities. The annual estimates of these features are provided by the State Department of Finance. Three types of housing accommodations are reported in this dataset—single-family, multi-family and mobile houses, along with the number of occupied houses (Department of Finance, California, 2020a, 2012a). In addition, Rural Urban Continuum Codes (RUCC) (United States Department of Agriculture, 2013) is used as another indicator of human activity and development. RUCC categorizes counties into nine different types—three levels of metropolitan and six levels of non-metropolitan divisions.

Data on fuel types across the state are acquired from the Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Type Product “MCD12Q1” (Friedl and Sulla-Menashe, 2019). This product is based on the International Geosphere-Biosphere Programme (IGBP) land cover classification. This annual-level data is available in rasters with a resolution of 500 meters. The IGBP classification partitions the Earth cover into seventeen classes, and accounts for urban areas, barren lands, water bodies, croplands, permanent icy lands, and different types of vegetation including the various kinds and their densities. More details can be found in Loveland and Belward (1997) and Belward et al. (1999). For topography, Global Multi-resolution Terrain Elevation Data (GMTED2010) at ~ 1 km spatial resolution (Danielson and Gesch, 2011) is used.

Drought information is obtained from the United States Drought Monitor (USDM) (USDM, 2020). This dataset is available on a weekly basis at various spatial scales and in different formats. There are five categories of droughts describing different levels— $D0$, $D1$, $D2$, $D3$ and $D4$, representing abnormally dry, moderate drought, severe drought, extreme drought and exceptional drought, respectively. Also, a category exists for no drought condition (i.e., none). More information can be found in Svoboda et al. (2002). For this study, weekly categorical data are extracted for each county. Values in the dataset represent percent of the area in each county that has experienced the corresponding level of drought. Moreover, the USDM introduces an index called Drought Severity and Coverage Index ($DSCI$) (Akyuz, 2017), which is a score derived from weighted accumulation of categorical drought levels, and can be calculated by Equation (1). Therefore, seven variables (no drought, five levels of drought, and $DSCI$) are used in this study to characterize the drought condition of each county.

$$(1 \times D0) + (2 \times D1) + (3 \times D2) + (4 \times D3) + (5 \times D4) = DSCI \quad (1)$$

Other than drought, which is considered as an indicator of long-term climatic effects, a range of meteorological variables are obtained from ERA5 dataset. ERA5

Fig. 1 Temporal availability of data sources. Time interval of the case study is bound by the years 2001 and 2015.

Table 1 Summary of raw data used for this study

Data	Temporal availability	Spatial availability	Source
FPA-FOD	annual	coordinates of incidents	Short (2017)
Population and Housing	annual	county	Department of Finance, California (2020b), Department of Finance, California (2012b)
RUCC	every 10 years	county	United States Department of Agriculture (2013)
Land Cover	annual	raster (resolution: 500m)	Friedl and Sulla-Menashe (2019)
Drought	weekly	county	USDM (2020)
ERA5	hourly	raster (resolution: 0.25°)	Service (2017)
topography	static	raster (resolution: 0.01°)	Danielson and Gesch (2011)

is the most recent climate reanalysis model developed by the European Centre for Medium-Range Weather Forecasts (ECMWF), and provides hourly data on regular raster grids at 0.25° resolution. Variables used for this study include boundary layer height, eastward and northward components of wind at the height of 10m above surface, maximum wind gust at the height of 10m above surface, air temperature and dew point temperature at the height of 2m, surface temperature, rate of precipitation at the surface, and volume of water in first soil layer (0–7cm). Detailed information on ERA5 data set can be found in Hersbach et al. (2020).

It is noteworthy that all the datasets used in this study are publicly available. Time period of analysis considered in the case study is decided by the availability of certain data types. For example, the land cover data is available since year 2001, while the FPA-FOD data is available until year 2015. The final time period of our study spans from 2001 – 2015, limited by the start time of the land cover data and the end time of FPA-FOD. Figure 1 provides a schematic of the time span that each data source covers, and Table 1 summarizes the features collected, time frequency and spatial resolution of the raw data used in our analysis.

3 Methodology

In this section, we describe our proposed research framework in three general steps.

- *Data collection and pre-processing*: Data are collected and processed to the spatiotemporal resolution of interest. Different subsets of the variables are tested as predictors as described in sub-section 3.1.4.

Fig. 2 Schematics of the proposed framework. Three general stages are 1) data collection and processing; 2) creating subsets of variables, and selecting candidate models; 3) selecting final model and statistical inference.

- *Implementing statistical learning models*: Algorithms are implemented on different subsets of the variables, and candidates for the “best” model are selected based on the defined criteria as described in sub-section 3.2.
- *Final model selection and inference*: More in-depth comparisons are conducted to select the final model, followed by hyper-tuning of the selected model, which is then used for statistical inference as described in sub-section 3.3.

Figure 2 summarizes the proposed framework, and the details are discussed in the subsequent sub-sections.

3.1 Data collection and pre-processing

Data collection and pre-processing is a vital step of any data-driven study. Since different features are collected from different sources with diverse formats, pre-processing of the data is conducted to have all the variables in compatible resolutions.

3.1.1 Determining spatiotemporal resolution of the study

The first step in conducting data-driven spatiotemporal research is determining the spatial and temporal scales of the response variable. From the perspective of risk analysis, depending on the assets at risk and/or interest of the stakeholder, one can focus on investigating the risk associated with just a single incident and limit the study area to the regions affected by that fire, or can analyze the risk patterns associated with multiple fire incidents occurring across a large region (e.g., a state or a country). In fact, selecting the optimal spatiotemporal scale that is compatible with the decision-makers' objectives and help adequately addressing the research gaps is the key for any data-driven risk assessment approach (Fontecha et al., 2021). Hence, there is a wide range of spatial scales from which the relevant one needs to be chosen for the analysis to address the specific research gaps.

In this study, we conduct the analysis on the wildfire-induced acres burned (i.e., extent of the wildfire spread) where the spatial unit is county and the temporal unit is year. It means fires that have occurred in a specific county in a particular year are aggregated to represent one observation of our training data set. County-level analysis is suitable for identifying regional wildfire-induced vulnerabilities, and thus help with risk-informed decision making.

3.1.2 Collection of data

As we already explained in details in section 2, various data types were collected from multiple sources. They provide a comprehensive representative set of different influencing factors on wildfire occurrence and spread including meteorological factors, topography, land cover, and anthropogenic features. It should be noted that selecting a particular data source/format for a feature, from the available ones, depends mainly on the spatiotemporal resolution of study. It is ideal to use features with resolutions finer than (or at most equal to) targeted spatial and temporal resolutions of the study. Up-sampling the features leaves us with more flexibility compared to down-sampling.

3.1.3 Pre-processing of collected data

Because the raw data are in diverse formats and resolutions, features are processed to develop a set of variables with homogenized spatiotemporal resolutions. The objective of this step is to condense each feature to a single value assigned to each spatiotemporal unit. If the scale of a feature is coarser than or equal to that of the target spatiotemporal unit, then its value is interpolated to the centroids of the study's units. If a feature is available in a temporal resolution finer than the target unit, first we conduct the temporal aggregation and extract different

Fig. 3 Calculation procedures on different data in raster format. Black lines in the bottom row illustrate the shape of a spatial unit, i.e., a county.

statistics measures (e.g. mean, standard deviation) of our interest, which are then used to represent their corresponding feature. Next, if the feature is also available in a spatial resolution finer than the study's spatial unit, a weighted average is conducted for the corresponding variable units that constitutes the target spatial unit, where the weight is the fractional overlap between each grid cell and the county boundary shape. Figure 3 depicts a schematic of the procedure for this spatiotemporal aggregation.

In this research, the state of California was selected as a case study, and a county in a particular year was considered as the smallest spatiotemporal unit, which was assigned a value for each predictor variable. Fire incidents were also clustered within each county on an annual basis. California is partitioned into 58 counties. This study compiled diverse data sets with wide range of temporal resolutions (from hourly to annually). Annual-level variables (anthropogenic and land cover features) were used directly, while the drought levels (available on a weekly basis) were aggregated to annual averages. On the other hand, variables with higher frequencies (hourly meteorological features) were aggregated to quarters of a year. More details of this procedure and the processing of different features are explained later in this sub-section.

From FPA-FOD, subset of fires labeled for the state of California were filtered, and based on the reported ignition date and coordinate, a county and a year was

assigned to each incident. Acres burned by wildfires associated to the same county in a year were summed together. Herein, the term “acres burned” corresponds to the total number of reported acres that have been burned in each county of California over a year. Fires are aggregated to annual values because there is a huge variation in the time, resources and efforts required to contain different wildfires. Moreover, it also retains the practicality of the study for decision-making. Figure 4 demonstrates how the county-level acres burned varies across the state (top left) and over the years (top right). For optimal decision-making, it is desirable that the risk models capture the long tails of the distribution (i.e., the extreme incidents), since they are the most resource-demanding incidents that significantly influence the fire management decisions and logistics. However, since these incidents are high-impact, low-probability events, it is difficult for the learning algorithms to be well-trained on such extreme cases. Therefore, to address this serious issue of the right skewed data, logarithmic transformation has been conducted on the acres burned variable leveraging Equation (2). The resultant transformed variable was then used as the response variable for our analysis. The distribution of the actual acres burned and the log-transformed distribution is also illustrated in Figure 4 (bottom). Four cases (i.e., county in a particular year) had no fire reported and were removed from the study.

$$\text{response variable} = \log(\text{acres burned}) \quad (2)$$

Data on population, housing, RUCC, and drought had the same spatial unit as that of the county. They are also available annually, except for drought, so no temporal or spatial aggregation was required. Data on drought was available on a weekly basis, and its annual average was calculated such that it matched the study’s target temporal resolution (i.e., annual). Population and housing densities were calculated by normalizing the population and housing count per 10,000 acres of each county. Housing data contain the percentage of five different categories of housing types in a county. The categories include single-family, multi-family, mobile houses, total houses, and occupied houses. RUCC data is provided every ten years. We used the latest RUCC published in 2013 for all the counties during the period 2008 — 2015. The previous version published in 2003 was used for the years 2001 to 2007.

Annual land cover data from MODIS is available in raster format with regular grids of 500 meters. Each pixel of the raster has a value from 1 to 17 indicating a specific type of cover (missing values are labeled as 255). Hence, there are seventeen variables which characterize the land coverage of the counties. In this study, the area within a county covered by each land cover type was calculated for each year and then normalized by area of the county.

Availability of meteorological data on hourly basis, which is significantly finer than the target temporal unit of this study (i.e., year), provides a great opportunity to aggregate and extract information at different levels. For each of the downloaded parameters from ERA5, hourly rasters were compiled over each quarter of the year, and different statistics measures (mean, standard deviation, minimum, 1st, 5th, 10th, 25th, median, 75th, 90th, 95th, 99th-percentiles, and maximum) were evaluated at pixel-level, resulting in a raster for each statistics measure. Given the accumulative nature of the precipitation, an additional raster of total precipitation over each quarter was also calculated. Finally, across each county, weighted spa-

Fig. 4 Processed information on wildfires. Top left: Total acres burned in each county over the duration of the study. Top right: Total annual acres burned across California. Bottom left: Distribution of acres burned. Bottom right: Transformed distribution of acres burned. This distribution is used in training algorithms.

tial average of the pixel values overlapping with county was calculated. It should be noted that meteorological features are not aggregated annually. This process was conducted to capture the effect of seasonality and preconditioning (Urbieto et al., 2015; Trigo et al., 2016; Littell et al., 2009; Gedalof et al., 2005; Crimmins and Comrie, 2005). Moreover, (Masoudvaziri et al., 2020) also provides evidence that aggregating this data to a finer temporal resolution (e.g., monthly) does not provide any significant advantage. Details of these calculations are presented in Figure 3.

Over the time span of this study (15 years), topography of the region can be considered to be unchanged naturally. Since, we have varying annual values of acres burned, a constant feature would not be useful in training the algorithms. Thus, in this study, we combined topography with wind for two reasons. First, the combined effect of wind and topography on wildfire spread has been established in wildland fires literature (Weise and Biging, 1994, 1996; Viegas, 2004; Silvani et al., 2012). Studies indicate that up-slope and wind facilitate the fire spread. Moreover, leveraging this technique, we could incorporate the effect of topography on the fire spread in a dynamic way. From the elevation raster, slope and aspects were calculated for each pixel (Burrough et al., 2015). We introduced the variable “wind-

topo” in the study, which was calculated using Equation (3) on the coinciding pixels of the four rasters (slope, aspect, eastward and northward 10m wind rasters).

$$wind-topo = s(u \cos(\frac{\pi(90 + a)}{180}) - v \sin(\frac{\pi(90 - a)}{180})) \quad (3)$$

In this Equation 3, s and a represent pixel values of slope and aspect rasters, u and v represent eastward and northward wind, in [m/s], respectively. Equation 3 is the negative of dot product of two vectors (horizontal wind and topography gradient). Topography vector is represented by aspect (determining direction) and slope (determining magnitude). It should be noted that aspect is measured clockwise starting from North, and is in the unit of degrees. Furthermore, resolution of wind raster was coarser than elevation raster by a factor of 25. Hence, wind rasters were linearly interpolated to render four rasters with the same resolution.

In addition to surface temperature and volume of water in the first soil layer, which can be a good indicator of vegetation conditions, the feature vapor pressure deficit (VPD) was also calculated at 2m above ground level, using air and dew point temperature at the height of 2m. Detailed procedure can be found in Prenger and Ling (2001).

3.1.4 Creating predictor subsets

In this study, our focus was to use a wide range of predictors that could explain the variations in the acres burned and let the model choose the key predictors best explaining the response variable—thus, not imposing any biases by considering only a handful of specific features. To satisfy this objective, we extracted as many statistics measures as we could during data pre-processing step. As a result, a large number of features were generated, some being highly correlated with others. Although presence of highly-correlated variables in a statistical learning algorithm does not affect the predictive performance of the models, it might affect the statistical inferencing process as presence of a highly correlated variable often masks the effect of another key variable. This has been established in previous studies as well (Mukherjee and Nateghi, 2017, 2019; Mukherjee et al., 2019; Nateghi and Mukherjee, 2017; Alipour et al., 2019; Obringer et al., 2020). Thus, keeping all the highly correlated variables in the model might affect one of our main objectives that is to identify the key factors and understand their associations with the wildfire spread for the specific spatiotemporal resolution of interest. Therefore, it is necessary to select features subsets comprising of variables that were not significantly correlated with each other. To conduct this dimensionality reduction, three methods were incorporated in this study as follows:

- *Principal component analysis (PCA)*: This method was applied on clusters of predictors corresponding to a set of features (e.g., population, housing, drought, land cover, each meteorological feature). To bring more depth to the analysis, since the number of meteorological predictors was comparatively larger than other predictors and as expected they were strongly correlated, these variables were further categorized in three different ways: (a) including all the meteorological features at quarterly-level, i.e., all the meteorological features and their associated statistics recorded during a particular quarter of a year; (b) including all the variables associated with a meteorological feature during a

year, i.e., all the statistics associated with a meteorological feature recorded during a particular year; (c) including only a particular meteorological feature and its associated statistics measured during a particular quarter of a year (unlike (a) where all the meteorological variables are used at a time). PCA was applied on each predictor clusters (a, b, and c) with different thresholds on the percentage of variation explained by principal components (10%, 20%, 40%, 70%) and different thresholds on the total rotation—proportion of the principal components explained by variables—(10%, 20%, 40%, 70%). Thresholds affect the number of variables selected for each subset.

- *Descriptive statistics analysis*: We estimated different statistics measures of meteorological features, obtained during data pre-processing, and four groups of statistics were created to be used during each iteration of the analysis: (a) mean and standard deviation; (b) different percentiles (minimum, 25th, median, 75th, maximum); (c) combination of the previous two; (d) extremes (5th- and 95th-percentiles). These groups added with non-meteorological variables yielded four variable subsets which were analyzed separately (in four different iterations) in our study.
- *Correlation screening*: Correlation screening has been established as an efficient process of variable subset selection as well (Mukherjee and Nateghi, 2017). This process is performed to select representatives of highly correlated variables. The selected features were uncorrelated with each other, but had a meaningful correlation with the response variable.

Algorithm 1: Variable Selection

Input: The high dimensional data set D with multiple categories ($C = \{C_1, C_2, \dots, C_k\}$) of variables with several quarters ($Q_c = \{Q_{c1}, Q_{c2}, \dots, Q_{cl}\}$) for each category of the variable and several variables ($V_{cq} = \{V_{cq1}, V_{cq2}, \dots, V_{cqm}\}$) in each quarter of a particular category $q \in Q; c \in C$

Output: Data set with reduced dimension

begin

```

1  Set threshold  $t$  of variance explained
2  Final_Variables =  $\{\phi\}$ 
3  All_Variables =  $\{\phi\}$ 
4  for  $c$  in  $C$  do
5      | All_Variables  $\leftarrow$  All_Variables  $\cup$   $c$ ;
6  for  $q$  in  $Q$  do
7      | All_Variables  $\leftarrow$  All_Variables  $\cup$   $q$ ;
8  for  $v$  in  $V$  do
9      | All_Variables  $\leftarrow$  All_Variables  $\cup$   $v$ ;
10   $PC \leftarrow$  PCA(All_Variables);
11   $PC_{top} \leftarrow$  Top principal components with total variance  $\geq t$ ;
12  for  $p$  in  $PC_{top}$  do
13      |  $\Gamma \leftarrow$  the top variables of  $p$  based on threshold of PCA rotation;
14      | Final_Variables  $\leftarrow$  Final_Variables  $\cup \Gamma$ 

```

In this paper, we have 4 quarters, and 8 meteorological parameters and each meteorological parameter includes 13 variables in each quarter (precipitation includes 14 variables, see Table 2). It is noteworthy that the proposed spatiotemporal

risk assessment framework is insensitive of the methods and threshold values implemented for dimensionality reduction. However, feature selection is an important step of this framework.

3.2 Statistical learning for risk analysis

In this section, we present a brief background on the statistical supervised learning and the various modeling techniques that we adopted in our research.

3.2.1 Overview of Statistical Supervised Learning

The supervised statistical learning algorithms establish a relation between a given response variable (say Y) and p different predictor or independent variables (Hastie et al., 2009; James et al., 2013), $X = X_1, X_2, \dots, X_p$ which can be represented as,

$$Y = f(X) + \epsilon \quad (4)$$

Here f is some unknown function, and ϵ is the irreducible error term that arises from unobserved heterogeneity and thus can't be measured. ϵ is independent of X and assumed to be normally distributed $N(\mu, \sigma^2)$ where, μ = mean and σ^2 = variance. To predict the response values Y corresponding to the unknown X values, and to generate inference from the data set, the statistical learning algorithm tries to estimate the unknown function f (Hastie et al., 2009; James et al., 2013). A model is trained on a subset of data, known as training data. Then the model performance is evaluated on another subset of data, known as test data.

The statistical learning models can be broadly classified into three classes, namely: i) parametric models, ii) semi-parametric models, and iii) non-parametric models (Hastie et al., 2009; James et al., 2013). For parametric models, the unknown function f is estimated using a set of parameters characterizing the function by assuming a pre-defined structural form of the function. On the other hand, the non-parametric methods make no assumptions about the functional form and try to fit the closest function that matches f . A semi-parametric model is a hybrid of parametric and non-parametric models.

In this research, several non parametric, tree-based algorithms viz. Random Forest (*RF*) (Breiman, 2001), Gradient Boosting Method (*GBM*) (Friedman, 2001), Extreme Gradient Boosting (*XGBoost*), and Bayesian Additive Regression Trees (*BART*) (Chipman et al., 2010; Kapelner and Bleich, 2016), are investigated. In our preliminary work (Masoudvaziri et al., 2020), it is demonstrated that for such a problem, tree-based algorithms outperform linear models.

3.3 Model selection and statistical inference

In this study, we employ a hierarchical process of selecting the best model as described below. For our case study, 275 models are developed and assessed.

3.3.1 Model selection criteria:

The generalization performance of a predictive model is measured by how accurately the model is able to predict the response values for some unseen data (*low bias*) while having minimum overfitting of the learning algorithm in the training data (*low variance*). The *bias* of a predictive model is the deviation of the predicted functional form $\hat{f}$ from the actual function f . A model is termed to have high *variance* if a small change in the training data can result in a significant change in the predicted function (Hastie et al., 2009; James et al., 2013).

One of the major criteria for selecting the best model are the in-sample and out-of-sample root mean squared errors, herein named as RMSEin and RMSEout respectively. Since evaluating the wildfire spread risk at county-level spatial scale is prioritized over its prediction in this study, we selected the best model based on the goodness of fit of the models, i.e., lower RMSEin. However, we also evaluated the RMSEout to ensure that the models do not overfit the data. Mathematically, the in-sample mean squared error is given by (Alipour et al., 2019):

$$MSE_{in\ sample} = \frac{1}{k} \left[\sum_{j=1}^k \frac{1}{m} \left(\sum_{i=1}^m (y_{i,j} - \hat{y}_{i,j})^2 \right) \right] \quad (5)$$

k = number of times cross validation performed; m = number of observations in the training data during each cross validation $y_{i,j} = i^{th}$ actual observation that was randomly selected for training during the j^{th} cross-validation, $\hat{y}_{i,j}$ = predicted i^{th} observation for the training set data during the j^{th} cross validation. On the other hand, the out-of-sample MSE can be defined as follows (Alipour et al., 2019):

$$MSE_{out\ of\ sample} = \frac{1}{k} \left[\sum_{j=1}^k \frac{1}{m} \left(\sum_{i=1}^m (y_{i,j} - \hat{y}_{i,j})^2 \right) \right] \quad (6)$$

k = number of times cross validation performed; m = number of observations in the holdout data during each cross validation $y_{i,j} = i^{th}$ actual observation that was randomly holdout during the j^{th} cross-validation, $\hat{y}_{i,j}$ = predicted i^{th} observation during the j^{th} cross validation using the model developed using the training set data during the j^{th} cross validation.

Once the models are implemented and the candidate models with best generalization performance are selected, their performances are further evaluated. Three additional criteria are used: (a) a detailed comparison of errors; (b) variable importance plots (VIP) and partial dependence plots (PDP); and, (c) performance on the validation set (dataset not used for training or testing of the algorithms). Details of each of these criteria are described as follows:

- First, the model errors are compared. Observational data points are divided into three groups: small fire incidents falling in the first quartile ($size \leq 25th$ percentile), large fire incidents falling in the third quartile ($size \geq 75th$ percentile), and medium fire incidents ($size \geq 25th$ percentile through $size \leq 75th$ percentile), based on the fire size (acres burned). In each division, estimations by models are evaluated in four ways: (1) total error; (2) percentage of the total errors; (3) number of cases that are underestimated and overestimated; (4) percentage of errors due to underestimation and overestimation. The ideal

result should provide a closer match between every observation and the corresponding estimated acres burnt by the model. Note, in the context of risk assessment, when errors are close to each other, overestimation is preferred to underestimation.

- Second, the characteristic of the models to distinctly differentiate the predictors in terms of their importance is considered.
- Third, since data points in validation set have not been incorporated in any stage of the training and testing of the algorithms, it can be used to evaluate the candidate model’s performances. By implementing the models on the validation set, the corresponding predicted values are compared to the actual observations from the validation set. This validation set data is also used for validation of the final model.

Further, if the model selection result obtained after applying these criteria are not unanimous, then the criterion are prioritized as follows: 1) detailed model error analysis; 2) distinct stratification in variable importance; and 3) performance on validation set. The rationale is that the validation set will be different as time passes, and new data points become available. Also, investigating the model errors and variable importance are directly aligned with the objectives of the framework. All the models are compared with the “mean-only” or the “null” model. The “mean-only” model uses mean of the response variable instead of a statistical model, which is a common benchmark used in statistics to identify the power of statistical models in explaining the variance of the response (Mukherjee and Nateghi, 2019; Mukherjee et al., 2019; Ganguly and Mukherjee, 2021).

After the final model is chosen, hyperparameter tuning is conducted and the performance of the final model is validated leveraging the validation set. Finally, important geophysical and anthropological factors are obtained from the variable importance plots, and their relationship with log of acres burned are investigated using the partial dependence plots.

In a nutshell, besides the generalization performance of the learning models, we considered model complexity and ability of the models to distinctly classify the risk factors as the other important criteria for model selection. We performed a step-by-step model filtering strategy described in Figure 5 for identifying the best model. Essentially, the best model has the best generalization performance compared to the other models, as well as a lower complexity while distinctly classifying the risk factors in terms of their importance.

3.3.2 Algorithm description of Extreme Gradient Boosting (the final model)

Extreme gradient boosting (*XGBoost*) was found to outperform all the other models based on the three different criteria described in the previous section 3.3. In this section, a brief description of the gradient boosting method is provided (details of other models are provided in the Appendix in section 6). The Gradient Boosting method is a non-parametric statistical learning algorithm that combines several weak learners to create a predictive model (Hastie et al., 2009; James et al., 2013). The algorithm develops on the empirical observation that takes lots of small steps towards the right direction and makes better prediction on an unseen data.

Fig. 5 A framework for selecting the best model

In a regression setting, the model starts with a constant value (e.g. mean of the response variable) as a predicted value of the response. A shallow regression tree is then added based on the residual of the previously predicted value and the tree's contribution is scaled by a learning rate to reduce the variance. Following this step, new residuals are calculated, another shallow regression tree is fitted to predict these new residuals, and the contribution of this tree is scaled as previous. This process is repeated several times. The formal algorithm proposed by Friedman (2001) is described below. XGBoost is an extension of the normal gradient boosting algorithm to enhance the performance.

Algorithm 2: Gradient Boosting Algorithm

Input: The training data set $\{(x_i, y_i)\}_{i=1}^n$, a differentiable loss function $L(y, F(x))$, the number of iterations M and a learning rate ν
Output: The optimal model

begin

- 1 Initialize model with a constant value (e.g. mean of the response):

$$F_0(x) = \gamma \sum_{i=1}^n L(y_i, \gamma)$$
- 2 **for** $m = 1$ **to** M **do**
- 3 Compute pseudo-residuals as $r_{im} = -[\frac{\delta L(y_i, F(x_i))}{\delta F(x_i)}]_{F(x)=F_{m-1}(x)}$

$$\forall i = 1, \dots, n$$
- 4 Fit a weak learner (shallow tree) to predict the residuals r_{im} and create the terminal regions $R_{jm} \forall j = 1, \dots, J_m$
- 5 **for** $j = 1, \dots, J_m$ **do**

$$\quad \text{Compute } \gamma_{jm} = \gamma \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$$
- 6 Update the model: $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$
- 7 **return** $F_M(x)$

Table 2 Summary of variables processed for this study. Minus signs indicate number of items removed due to empty values.

Data	Period	Number of variables
Acres burned (FPA-FOD)	year	1
Population density	year	1
Housing types density	year	5
Rural-Urban Continuum Codes (RUCC)	year	1
Drought	year	7
Land cover percentage	year	17 (−1)
Meteorological variables:		
- Rate of precipitation at the surface	quarter	4×14 (−19)
- Boundary layer height	quarter	4×13
- Maximum wind gust at the height of - 10 meters above surface	quarter	4×13
- Surface temperature	quarter	4×13
- Volume of water in first soil layer (0–7cm)	quarter	4×13
- Wind-topo	quarter	4×13
- Vapor pressure deficit at 2m above surface	quarter	4×13

4 Results

Leveraging the methodology described in the previous section, we implemented the research framework for the state of California. The results are described in the following subsections.

4.1 Variable subset selection

Our analysis for the case study spans over fifteen years (2001 to 2015) and for the 58 counties in the state of California. For counties where no fire incident is observed in a year, that particular observation is removed. Four of such instances, indicating no fire incident, were observed in the dataset and thus, were not included in the study. Therefore, the study is conducted on a total of 866 observations ($15 \times 58 - 4$). The data pre-processing step yields a dataset of 376 variables. Table 2 summarizes the number of variables that correspond to each of the collected features and the time span that they cover. It should be noted that one land cover and 19 precipitation variables were removed from the data set because all of their values were zero. The rationale behind removing these variables is that a constant feature vector with no variation would not be useful for learning algorithms.

After conducting the subset selection procedure, there were 55 datasets to investigate, one of which was the complete dataset, 48 subsets were outputs of principal component analysis, 4 were outputs of descriptive statistics analysis, and 2 are outputs of correlation screening. Number of variables in each subset is provided in the Appendix (section 6).

4.2 Implementing candidate models

The library of ensemble tree-based algorithms were implemented using all the variable subsets and their performances were compared leveraging the three criteria

proposed in our study (see section 3.3). Figure 6 display the relative improvement in models' performances compared to the null model for each data subset.

First, it is observed that RF and XGBoost outperform other algorithms at in-sample RMSE criterion. The rationale for not selecting XGBoost as the best algorithm at this step based on its superior in-sample RMSE is that, it performs poorer in terms of out-of-sample RMSE (compared to the RF which has the second best in-sample RMSE results), implying a possible overfitting.

Second, it is observed that PCA does not cause significant changes to the models' performances, advocating that not all the variables assembled can provide significant information for training the models, which aligns with one of our objectives that is to let the algorithms decide the useful input features. Moreover, between the three categories of input features for PCA, *category c* yields lower errors (which is expected since this category includes more variables). In addition, the gradual increasing trend in the performance with higher thresholds of PCA (i.e., when number of variables are increased) is noteworthy and it is also well-expected.

Third, among all the subsets, it is observed that subsets based on different statistics yielded better results compared to the PCA subsets. We observe that the subset *correlation 1* yields a top performing model, while being remarkably simpler (fifty-eight variables). It is noteworthy that the simplest subset *correlation 2*, which is the outcome of a more restricted correlation screening (with only fifteen variables), causes a significant loss in performance. Thus, following Occam's Razor rule, "the simpler one is the better one", we select the simplest subset that yielded comparable results with the top performing complex models. Overall, the two models, XGBoost and RF, yielding comparable results on the variable subset *Correlation 1* are selected as candidates.

4.3 Final model

Following the criteria described in section 3.3, model errors are calculated and compared. The analysis is provided in Table 3. It is observed that XGBoost produces less error, a larger portion of which originates from overestimation. Moreover, when both models are applied on the validation set, their estimations are found to be close to each other, but XGBoost follows the real values more closely than the RF (see Figure 7). A significant difference between candidate models are observed when comparing variable importance plots, where XGBoost clearly differentiates the variables in terms of their importance, which substantially varies. In contrast, many features in RF have similar importance measures, indicating that the model has high variance and adding new observations to the model might change the rankings significantly, which is not aligned with our objective of data-driven discovery of the influential variables. The RF's variable importance plot also explains the unique behavior and insensitivity of the algorithm towards the different variable subsets (Figure 6), which can be attributed to its inability to effectively isolate the key features. Since both models are applied on the same dataset, this specifically becomes a useful criterion. Based on these observations, we selected XGBoost model developed using the variable subset *Correlation 1* as the final one for our case study.

Fig. 6 Performance of different algorithms on different input data sets. plots a and b present the performance of models incorporating the subsets made by correlation screening and statistics measures; plots c and d present the performance of models incorporating the subsets made by PCA. In the plots, x-axes represent the features subsets and y-axes represent the relative reduction in errors by different models, compared to the null model.

Table 3 Performance of the candidate models based on estimated values

	Small Fires		Medium Fires		Large Fires	
	RF	XGBoost	RF	XGBoost	RF	XGBoost
squared errors (SE)	155.87	56.6	69.5	61.29	173.48	127.64
percent of total SE	39	23	17	25	43	52
number of underestimated cases	25	21	175	164	180	177
number of overestimated cases	177	181	229	240	22	25
percent of SE underestimated	2	1.2	41.8	37.8	99.2	98.9
percent of SE overestimated	98	98.8	58.2	62.2	0.8	1.1

Grid search is leveraged for conducting hyperparameter tuning of the final XGBoost model. A combination of parameters — $\eta = (0.01, 0.05, 0.1, 0.3, 0.5)$, number of trees (500, 1000, 2000, 4000), maximum depth of trees (4, 6, 8, 10), and $\gamma = (0.001, 0.005, 0.01, 0.05)$ are investigated. The results are demonstrated in Figure 8. Considering the errors as well as simplicity of the models, the combination of $\eta = 0.01$, number of trees = 2000, maximum depth = 6, and $\gamma = 0.001$ is selected as the final parameters of the XGBoost model and then used for statistical

Fig. 7 In-depth comparison of the performance of XGBoost and RF on the data set “correlation 1”: a) Performance of the models on validation set; b) Comparing the estimations made on training data by XGBoost and RF (i.e., years 2001 to 2014); c) Variable Importance Plot of XGBoost; d) Variable Importance Plot of RF.

inference. It should be noted that effect of γ was insignificant. The final model is validated leveraging the validation data set and compared to the observed values in Figure 8.

4.4 Statistical inferencing

Figure 9 depicts the variable importance of the top ten most important features obtained from our final model. We observe that various geophysical features and an anthropogenic factor are included in this set of variables. This is an indicator of the diverse factors contributing to the acres burned.

4.4.1 Key geophysical factors

The effect of the key geophysical factors are described below.

1. *Surface/skin temperature:* The most influencing variables associated with the wildfire-induced acres burned (i.e., the wildfire spread) are the standard devia-

Fig. 8 Final model: XGBoost implemented on the data set "correlation 1". Top: Tuning of XGBoost hyperparameters. Bottom: Performance of the final model compared to the validation set (year 2015) and average of the training set (years 2001 to 2014).

tion in surface temperature for months April to September (i.e., quarters 2 and 3) as shown in Figure 10. We observe that larger variations are associated with widespread wildfires (i.e., larger acres burned). Our results, thus emphasize that the variations in surface temperatures are the more important than their absolute values. This is because it is an important indicator of the "preconditioning" that is discussed in the literature (Urbiet al., 2015; Trigo et al., 2016; Littell et al., 2009; Gedalof et al., 2005; Crimmins and Comrie, 2005).

2. *Wind gust*: Wind gust is another important feature known to affect acres burned during a wildfire event, mostly attributed to its direct effect on the rate of wildfire spread (Fovell and Gallagher, 2018; Mitchell, 2013). Its importance is apparent since three variables representing wind gust is among the top 10 variables (see Figure 9). It can be observed from Figure 11 that stronger gust is associated with larger acres burned in general. One interesting observation is that wind gust is an important feature almost throughout the entire year as we have variables representing wind gust from three different quarters and all ranked among the top 10 important variables.
3. *Land cover*: Our results show that counties with higher coverage of barren lands usually experience larger acres burned. Although it may seem counter-intuitive at first as barren lands generally contain less fuels, it should be noted

Fig. 9 Variable importance plot of the final model; Top 10 variables.

Fig. 10 PDPs of important skin temperature variables.

that barren land also coincides with climatic conditions with high temperature and dry air, which facilitates the spread of fire.

The Woody Savannas land cover (see Figure 12) is also found to be an important feature. However, a careful observation reveals that only for certain counties this land cover type has an association with the acres burned.

Fig. 11 PDPs of important wind gust variables.

4. *Vapor pressure deficit*: Similar to the skin temperature, a higher variation in the vapor pressure deficit during the second quarter, and not necessarily the absolute value, is associated with a larger wildfire spread (see Figure 13).
5. *Boundary layer height*: As observed from Figure 14, there is a clear inverse relationship between acres burned and the height of atmospheric boundary layer (in m). Many parameters and features affect the boundary layer height, making it difficult to conclude about its causal relationships with the acres burned. The large burnt area at low boundary layer height is likely due to the associated dominance of large scale high pressure system that is fire-prone.

4.4.2 Key anthropological factors:

Overall, as expected, anthropological factors are found to be relatively less important to the extent of wildfire spread when compared to the geophysical factors. Only population density is in the top-10 list of the key risk factors. The association of population density with the acres burned is given below.

Fig. 12 PDPs of important land cover variables.

Fig. 13 PDP of important vapor pressure deficit variable.

1. *Population density*: As depicted in Figure 15, increase in population density is positively associated with the wildfire-induced acres burned. This can be attributed to the fact that large proportion of wildfire occurrences are started by humans (Balch et al., 2017), and anthropogenic factors such as land-use modifications (causing change in fuel structure and load) and suppressive activities directly affect the acres burned (Syphard et al., 2007; Bowman et al., 2011; Fusco et al., 2016).

5 Limitations and Future Work

The proposed framework presented in this study focuses on augmenting new variables, especially for geophysical features. The rationale behind this is not to just rely on common statistics (like mean, median, maximum, etc.) of the variables,

Fig. 14 PDP of important boundary layer height variable.

Fig. 15 PDP of important population variable.

but also to consider their entire distribution and then identify the important features using statistical learning algorithms. One of the challenges in this method is the two types of correlation amongst the variables: 1) the correlation amongst variables that are drawn from the same distribution (e.g. different percentiles); and 2) the inherent correlation between meteorological features. Although this does not present any challenge for predicting the wildfire spread, it needs to be addressed if statistical inferencing is one of the major objectives. However, as explained in section 3.2, various variable reduction techniques can be considered to overcome this issue.

High level of spatiotemporal randomness inherited to the wildfire spread phenomenon (see Figure 4) imposes challenge to understanding the causal effect of the key features on the wildfire-induced acres burned. Although predicting the acres burned is not the prime objective of this study, models with less error and higher accuracy would be ideal, as then inference of features can be done with higher confidence. Consider our case study's validation set (i.e., year 2015) as an example. It is not the utmost year during the interval of the study overall (ranked 4th based on the total acres burned). As observed in Figure 8, although the model captures the average of the acres burned in each county over the years 2001 to 2014 (by a dashed blue line), we see some serious differences between model's estimations and the real values for validation set (by red solid line). A closer look into the top nine counties with the largest burnt areas (presented in Figure 16) reveals that for most of these cases, acres burned in 2015 is unprecedented in the corresponding county. If data for recent years become available, extreme incidents

Fig. 16 Variation of the annual burned area in some of the counties.

in 2017, 2018, and 2020 are expected to change the distribution of burned areas significantly, hence change the behavior of models.

The criteria and procedure used in the framework filter the candidate models adequately. However, it should be noted that these criteria help comparing the models, and quantitative thresholds and measures for selecting the final model are not introduced. The reason is that by changing the spatiotemporal scales, algorithms, and input features, different results might be obtained and further insights gained. Future studies can examine these effects. Also, VIP and PDP used in inference are based on variation of single variables while keeping the rest unchanged. Although this is a conventional way to investigate the sensitivity of models, we know that there exists a complicated dynamic and inter-connection between most of the variables.

Additionally, given the ultimate objective of the work that is to provide the decision and policy makers with beneficial information, the annual time scale and the county boundaries are selected as they conform well with administrative work-frames. However, a large fire sometimes burns across several counties, in which case the incident is called “complex”. Given the flexibility of the pre-processing stage, future research may leverage the perimeters of such regions affected by the fire, instead of specific county borders.

Moreover, considering more features to represent long-term effects in the future and patterns in the past may also improve our understanding of the problem. Since a complex dynamic phenomenon is associated with a wildfire spread, assessing synoptic climate activities, time-series analysis, and spatial dependencies can be conducted in future. Finally, data on response strategies can be included in the future to understand whether the emergency response time and the amount of resources allocated can significantly affect the wildfire spread risk.

6 Conclusion

A data-driven spatiotemporal wildfire risk assessment approach is introduced in this study to help policy makers capturing county-level wildfire spread behavior, and identifying the associated key contributing factors. The proposed framework is composed of three steps: i) data collection and pre-processing; ii) predictor subset selection; iii) final model selection and statistical inference.

To showcase the applicability of the framework, a case study is conducted at county-level for the state of California, spanning a 15-year period (2001 to 2015). In this study, wildfire-induced acres burnt per year are aggregated at county-level and different geophysical and anthropogenic features associated with the extent of wildfire spread have been investigated. The eXtreme Gradient Boosting (XGBoost) algorithm is found to outperform the other models in terms of generalization performance and simplicity (i.e., having a fewer number of features). The top ten factors associated with the wildfire-induced acres burned include standard deviations of surface temperature and vapor pressure deficit, wind gust, grassy and barren land covers, night-time boundary layer height and population density in counties.

One of the important characteristics of this framework is the proposed data pre-processing approach can be conducted over a range of spatiotemporal scales. The framework's flexibility facilitates incorporation of different additional data sources and its adaptation at various spatial and temporal scales. The combination of input data in raster format and vector format of perimeters (e.g., boundary of the counties in the case study) create the opportunity to investigate different features from various sources as needed for the specific research study and/or pertaining to the interest of the stakeholder/decision-maker. For example, a fine-scale application of this framework may use the perimeters of the wildfires directly as the boundaries of interest. This framework will help the decision-makers to adequately plan for wildfire risk mitigation and resource allocation strategies.

Declaration

Funding

Funding for this project was provided by 2019-20 SUNY Research Seed Grant Program RFP: 20-02-RSG, Proposal ID: RSG201039.2.

Conflicts of interest/Competing interests

The authors do not have any conflicts of interest.

Appendix

6.1 Background on statistical learning methods

In this section, the methodological backgrounds of the other algorithms used in the study besides eXtreme Gradient Boosting-XGBoost, i.e., Random forest (RF) and Bayesian Additive Regression Trees (BART) are discussed.

Random Forest (RF):

Random Forest is a non-parametric ensemble tree based method. The method ensembles B bootstrapped regression trees (T_b) where B is selected based on cross-validation. The final estimate is made by averaging the predictions across all trees as shown in the equation below.

$$f^B(X) = \frac{1}{B} \sum_{b=1}^B T_b(X)$$

The random forest are low bias techniques, i.e they can capture the pattern of the data very well. However, these models have high variance and sensitive to outliers.

Bayesian Additive Regression Trees (BART):

Bayesian Additive Regression Tree is a sum-of-trees model where the outputs from m ‘small’ decision trees are aggregated with an underlying Bayesian probability model to generate the response function Chipman et al. (2010); Kapelner and Bleich (2016). Mathematically, BART can be expressed as,

$$Y = [\sum_{j=1}^m g(X; T_j, M_j)] + \epsilon \quad \epsilon \sim N(0, \sigma^2) \quad (7)$$

There are m distinct regression trees T_j with their terminal node parameters M_j . The function $g(X; T_j, M_j)$ assigns the leaf node parameters M of tree T to the independent variables X for all m trees. The main difference of BART compared to other tree ensemble methods is that, BART develops on an underlying Bayesian probability model and consists of a prior, likelihood and posterior probability space. The prior terms are responsible for the tree structure, model complexity, regularization and incorporating expert knowledge in the model. Generally, the Metropolis-Hastings algorithm is used to generate draws from the posterior probability space.

Table 4 Naming of Subsets and Number of Variables in Subsets based on PCA (including the response variable)

Meteorological variable categories	Variance explained	Rotation			
		10%	20%	40%	70%
Both	10%	PCA_01, 56	PCA_04, 85	PCA_07, 146	PCA_10, 253
	20%	PCA_13, 56	PCA_16, 85	PCA_19, 146	PCA_22, 253
	40%	PCA_25, 57	PCA_28, 86	PCA_31, 147	PCA_34, 254
	70%	PCA_37, 58	PCA_40, 88	PCA_43, 150	PCA_46, 260
Parameter	10%	PCA_02, 38	PCA_05, 68	PCA_07, 146	PCA_10, 253
	20%	PCA_14, 38	PCA_17, 68	PCA_20, 131	PCA_23, 234
	40%	PCA_26, 39	PCA_29, 69	PCA_32, 132	PCA_35, 235
	70%	PCA_38, 39	PCA_41, 73	PCA_44, 143	PCA_47, 250
Parameter	10%	PCA_02, 38	PCA_05, 68	PCA_07, 146	PCA_10, 253
	20%	PCA_14, 38	PCA_17, 68	PCA_20, 131	PCA_23, 234
	40%	PCA_26, 39	PCA_29, 69	PCA_32, 132	PCA_35, 235
	70%	PCA_38, 39	PCA_41, 73	PCA_44, 143	PCA_47, 250
Quarter	10%	PCA_03, 28	PCA_06, 54	PCA_09, 108	PCA_12, 204
	20%	PCA_15, 28	PCA_18, 54	PCA_21, 108	PCA_24, 204
	40%	PCA_27, 33	PCA_30, 59	PCA_33, 118	PCA_36, 219
	70%	PCA_39, 37	PCA_42, 68	PCA_45, 136	PCA_48, 244

Table 5 Naming of Subsets and Number of Variables in Subsets based on PCA (including the response variable)

Name	Variables included	Number of Variables
mean + sd	Variables representing mean and standard deviation of meteorological features in every quarter + all variables representing population and housing, drought, land cover + response variable	87
percentiles	Variables representing minimum, maximum, median, first and third quartiles of meteorological features in every quarter + all variables representing population and housing, drought, land cover + response variable	164
extremes	Variables representing median, 5th and 95th percentiles of meteorological features in every quarter + all variables representing population and housing, drought, land cover + response variable	111
mean + sd + percentiles	Union of "mean+sd" and "percentiles" variables	220

6.2 Number of Variables

Here, number of variables available in each variable subset is provided. Also, list of the variables in the final model is delivered.

6.3 List of variables comprised in the final subset (Correlation 1):

[1] Response variable, [2] Population density, [3] D0, [4] DSCI, [5] LC4, [6] LC8, [7] LC10, [8] LC16, [9] LC17, [10] Precipitation - Median - Q1, [11] Precipitation - 99th percentile - Q1, [12] Precipitation - Standard Deviation - Q2, [13] Precipitation - Standard Deviation - Q3, [14] Precipitation - 75th percentile - Q3, [15] Precipitation - Median - Q4, [16] Precipitation - Maximum - Q4, [17] Boundary layer height - 5th percentile - Q1, [18] Boundary layer height - Standard Deviation - Q1, [19] Boundary layer height - 5th percentile - Q2, [20] Boundary layer height - Mean - Q2, [21] Boundary layer height - 10th percentile

- Q3, [22] Boundary layer height - Mean - Q3, [23] Boundary layer height - 25th percentile - Q4, [24] Boundary layer height - 90th percentile - Q4, [25] Skin Temperature - 25th percentile - Q1, [26] Skin Temperature - Standard Deviation - Q1, [27] Skin Temperature - Minimum - Q2, [28] Skin Temperature - Standard Deviation - Q2, [29] Skin Temperature - Minimum - Q3, [30] Skin Temperature - Standard Deviation - Q3, [31] Skin Temperature - Median - Q4, [32] Skin Temperature - Standard Deviation - Q4, [33] Volume of water in first soil layer - Mean - Q1, [34] Volume of water in first soil layer - 75th percentile - Q2, [35] Volume of water in first soil layer, 99th percentile - Q3, [36] Volume of water in first soil layer - Maximum - Q4, [37] Maximum wind gust at the height of 10 meters above surface - 1st percentile - Q1, [38] Maximum wind gust at the height of 10 meters above surface - Minimum - Q2, [39] Maximum wind gust at the height of 10 meters above surface - Median - Q2, [40] Maximum wind gust at the height of 10 meters above surface - Minimum - Q3, [41] Maximum wind gust at the height of 10 meters above surface - Median - Q3, [42] Maximum wind gust at the height of 10 meters above surface - Minimum - Q4, [43] Maximum wind gust at the height of 10 meters above surface - Maximum - Q4, [44] Vapor pressure deficit at 2m above surface - 25th percentile - Q1, [45] Vapor pressure deficit at 2m above surface - Minimum - Q2, [46] Vapor pressure deficit at 2m above surface - Standard deviation - Q2, [47] Vapor pressure deficit at 2m above surface - Minimum - Q3, [48] Vapor pressure deficit at 2m above surface - Standard deviation - Q3, [49] Vapor pressure deficit at 2m above surface - 25th percentile - Q4, [50] Vapor pressure deficit at 2m above surface - 95th percentile - Q4, [51] Wind-topo - 25th percentile - Q1, [52] Wind-topo - Mean - Q1, [53] Wind-topo - 25th percentile - Q2, [54] Wind-topo - Mean - Q2, [55] Wind-topo - 25th percentile - Q3, [56] Wind-topo - Mean - Q3, [57] Wind-topo - 25th percentile - Q4, [58] Wind-topo - Mean - Q4.

Fig. 17 Variable Importance Plot of the final model.

References

- United States Department of Agriculture ERS (2013) <https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx>
- Akyuz A (2017) Drought severity and coverage index. united states drought monitor. URL <https://droughtmonitor.unl.edu/About/AbouttheData/DSCI.aspx>
- Alipour P, Mukherjee S, Nateghi R (2019) Assessing climate sensitivity of peak electricity load for resilient power systems planning and operation: A study

- applied to the Texas region. *Energy* 185:1143–1153
- Amatulli G, Pérez-Cabello F, de la Riva J (2007) Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty. *Ecological modelling* 200(3-4):321–333
- Amraoui M, Pereira MG, DaCamara CC, Calado TJ (2015) Atmospheric conditions associated with extreme fire activity in the Western Mediterranean region. *Science of the total environment* 524:32–39
- Balch JK, Bradley BA, Abatzoglou JT, Nagy RC, Fusco EJ, Mahood AL (2017) Human-started wildfires expand the fire niche across the United States. *Proceedings of the National Academy of Sciences* 114(11):2946–2951
- Belward AS, Estes JE, Kline KD (1999) The IGBP-DIS global 1-km land-cover data set DISCover: A project overview. *Photogrammetric Engineering and Remote Sensing* 65(9):1013–1020
- Bendick R, Hoylman ZH (2020) Topological data analysis reveals parameters with prognostic skill for extreme wildfire size. *Environmental Research Letters*
- Boulanger Y, Parisien MA, Wang X (2018) Model-specification uncertainty in future area burned by wildfires in Canada. *International journal of wildland fire* 27(3):164–175
- Bowman DM, Balch J, Artaxo P, Bond WJ, Cochrane MA, D’antonio CM, DeFries R, Johnston FH, Keeley JE, Krawchuk MA, et al. (2011) The human dimension of fire regimes on Earth. *Journal of biogeography* 38(12):2223–2236
- Breiman L (2001) Random forests. *Machine Learning* 45:5–32
- Burrough PA, McDonnell R, McDonnell RA, Lloyd CD (2015) Principles of geographical information systems. Oxford university press
- California Department of Forestry and Fire Protection (2020a) Top 20 Largest California Wildfires. https://www.fire.ca.gov/media/11416/top20_acres.pdf, Accessed: September 3, 2020
- California Department of Forestry and Fire Protection (2020b) Top 20 Most Destructive California Wildfires. https://www.fire.ca.gov/media/t1rdhizr/top20_destruction.pdf, accessed: September 3, 2020
- California Department of Technology (2019) CA Geographic Boundaries. URL https://data.ca.gov/metadata_download/ca-geographic-boundaries
- Chipman A Hugh, George EI, McCulloch RE (2010) Bart: Bayesian additive regression trees. *The Annals of Applied Statistics* 4(1):266–298
- Coen J (2013) Modeling wildland fires: A description of the coupled atmosphere-wildland fire environment model (CAWFE)
- Coen JL, Schroeder W, Conway S, Tarnay L (2020) Computational modeling of extreme wildland fire events: a synthesis of scientific understanding with applications to forecasting, land management, and firefighter safety. *Journal of Computational Science* p 101152
- Cortez P, Morais AdJR (2007) A data mining approach to predict forest fires using meteorological data
- Costafreda-Aumedes S, Comas C, Vega-Garcia C (2016) Spatio-temporal configurations of human-caused fires in Spain through point patterns. *Forests* 7(9):185
- Costafreda-Aumedes S, Comas C, Vega-Garcia C (2018) Human-caused fire occurrence modelling in perspective: a review. *International Journal of Wildland Fire* 26(12):983–998
- Crimmins MA, Comrie AC (2005) Interactions between antecedent climate and wildfire variability across south-eastern Arizona. *International Journal of Wild-*

- land Fire 13(4):455–466
- Daniel CJ, Ter-Mikaelian MT, Wotton BM, Rayfield B, Fortin MJ (2017) Incorporating uncertainty into forest management planning: Timber harvest, wild-fire and climate change in the boreal forest. *Forest Ecology and Management* 400:542–554
- Danielson JJ, Gesch DB (2011) Global multi-resolution terrain elevation data 2010 (GMTED2010). US Department of the Interior, US Geological Survey
- Department of Finance, California (2012a) California, E-8 Historical Population and Housing Estimates for Cities, Counties, and the State, 2000-2010. <https://www.dof.ca.gov/Forecasting/Demographics/Estimates/E-8/2000-10/>
- Department of Finance, California (2012b) California, E-8 Historical Population and Housing Estimates for Cities, Counties, and the State, 2000-2010. URL <http://www.dof.ca.gov/Forecasting/Demographics/Estimates/E-8/2000-10/>
- Department of Finance, California (2020a) California, E-5 Population and Housing Estimates for Cities, Counties and the State — January 1, 2011-2020. <https://www.dof.ca.gov/Forecasting/Demographics/Estimates/e-5>
- Department of Finance, California (2020b) California, E-5 Population and Housing Estimates for Cities, Counties and the State — January 1, 2011-2020. URL <http://www.dof.ca.gov/Forecasting/Demographics/Estimates/e-5/>
- Finney MA, McHugh CW, Grenfell IC, Riley KL, Short KC (2011) A simulation of probabilistic wildfire risk components for the continental United States. *Stochastic Environmental Research and Risk Assessment* 25(7):973–1000
- Fontecha JE, Agarwal P, Torres MN, Mukherjee S, Walteros JL, Rodriguez JP (2021) A two-stage data-driven spatiotemporal analysis to predict failure-risk of urban sewer systems leveraging machine learning algorithms. *Risk Analysis*
- Fovell RG, Gallagher A (2018) Winds and gusts during the Thomas fire. *Fire* 1(3):47
- Friedl M, Sulla-Menashe D (2019) MCD12Q1 MODIS/Terra+ Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006. 2019, distributed by NASA EOSDIS Land Processes DAAC
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics* pp 1189–1232
- Fusco EJ, Abatzoglou JT, Balch JK, Finn JT, Bradley BA (2016) Quantifying the human influence on fire ignition across the western USA. *Ecological applications* 26(8):2390–2401
- Ganguly P, Mukherjee S (2021) A multifaceted risk assessment approach using statistical learning to evaluate socio-environmental factors associated with regional felony and misdemeanor rates. *Physica A: Statistical Mechanics and its Applications*
- Gedalof Z, Peterson DL, Mantua NJ (2005) Atmospheric, climatic, and ecological controls on extreme wildfire years in the northwestern United States. *Ecological Applications* 15(1):154–174
- Gralewicz NJ, Nelson TA, Wulder MA (2012) Factors influencing national scale wildfire susceptibility in Canada. *Forest Ecology and Management* 265:20–29
- Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media
- Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, et al. (2020) The ERA5 global reanalysis.

- Quarterly Journal of the Royal Meteorological Society
- Insurance Information Institute (2020) Facts + Statistics: Wildfires. <https://www.iii.org/fact-statistic/facts-statistics-wildfires>, Accessed: September 3, 2020
- Jain P, Coogan SC, Subramanian SG, Crowley M, Taylor S, Flannigan MD (2020) A review of machine learning applications in wildfire science and management. arXiv preprint arXiv:200300646
- James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol 112. Springer
- Juan P, Mateu J, Saez M (2012) Pinpointing spatio-temporal interactions in wildfire patterns. *Stochastic Environmental Research and Risk Assessment* 26(8):1131–1150
- Kapelner A, Bleich J (2016) Bartmachine: Machine learning with bayesian additive regression trees. *Journal of Statistical Software, Articles* 70(4):1–40
- Koutsias N, Martínez-Fernández J, Allgöwer B (2010) Do factors causing wildfires vary in space? evidence from geographically weighted regression. *GIScience & Remote Sensing* 47(2):221–240
- Linn R, Winterkamp J, Edminster C, Colman JJ, Smith WS (2007) Coupled influences of topography and wind on wildland fire behaviour. *International Journal of Wildland Fire* 16:183–195
- Littell JS, McKenzie D, Peterson DL, Westerling AL (2009) Climate and wildfire area burned in western US ecoprovinces, 1916–2003. *Ecological Applications* 19(4):1003–1021
- Loveland TR, Belward A (1997) The international geosphere biosphere programme data and information system global land cover data set (DISCover). *Acta Astronautica* 41(4-10):681–689
- Mandel J, Beezley JD, Kochanski AK (2011) Coupled atmosphere-wildland fire modeling with WRF 3.3 and SFIRE 2011. *Geoscientific Model Development* 4(3):591–610
- Mandel J, Amram S, Beezley J, Kelman G, Kochanski A, Kondratenko V, Lynn B, Regev B, Vejmelka M (2014) Recent advances and applications of WRF-SFIRE. *Natural Hazards and Earth System Sciences* 14(10):2829
- Masoudvaziri N, Ganguly P, Mukherjee S, Sun K (2020) Integrated risk-informed decision framework to minimize wildfire-induced power outage risks: A county-level spatiotemporal analysis. In: *Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference, Venice Italy*
- McCandless T, Kosovic B, Petzke W (2020) Enhancing wildfire spread modelling by building a gridded fuel moisture content product with machine learning. *Machine Learning: Science and Technology*
- Mitchell JW (2013) Power line failures and catastrophic wildfires under extreme weather conditions. *Engineering Failure Analysis* 35:726–735
- Mukherjee S, Nateghi R (2017) Climate sensitivity of end-use electricity consumption in the built environment: an application to the state of Florida, United States. *Energy* 128:688–700
- Mukherjee S, Nateghi R (2019) A data-driven approach to assessing supply inadequacy risks due to climate-induced shifts in electricity demand. *Risk Analysis* 39(3):673–694

- Mukherjee S, Vineeth C, Nateghi R (2019) Evaluating regional climate-electricity demand nexus: a composite bayesian predictive framework. *Applied Energy* 235:1561–1582
- Nateghi R, Mukherjee S (2017) A multi-paradigm framework to assess the impacts of climate change on end-use energy demand. *PloS One* 12(11):e0188033
- Nelson RM (2002) An effective wind speed for models of fire spread. *International Journal of Wildland Fire* 11:153–161
- Obringer R, Mukherjee S, Nateghi R (2020) Evaluating the climate sensitivity of coupled electricity-natural gas demand using a multivariate framework. *Applied Energy* 262:114419
- Prenger JJ, Ling PP (2001) Greenhouse condensation control understanding and using vapor pressure deficit (VPD)
- Rodrigues M, de la Riva J (2014) An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software* 57:192–201
- Rodrigues M, Jiménez-Ruano A, Peña-Angulo D, de la Riva J (2018) A comprehensive spatial-temporal analysis of driving factors of human-caused wildfires in spain using geographically weighted logistic regression. *Journal of environmental management* 225:177–192
- Sayad YO, Mousannif H, Al Moatassime H (2019) Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire safety journal* 104:130–146
- Service CCC (2017) ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate
- Shen H, Tao S, Chen Y, Odman MT, Zou Y, Huang Y, Chen H, Zhong Q, Zhang Y, Chen Y, et al. (2019) Global fire forecasts using both large-scale climate indices and local meteorological parameters. *Global Biogeochemical Cycles* 33(8):1129–1145
- Short KC (2017) Spatial wildfire occurrence data for the United States, 1992-2015 [FPA_FOD_20170508]
- Silvani X, Morandini F, Dupuy JL (2012) Effects of slope on fire spread observed through video images and multiple-point thermal measurements. *Experimental Thermal and Fluid Science* 41:99–111
- Svoboda M, LeCompte D, Hayes M, Heim R, Gleason K, Angel J, Rippey B, Tinker R, Palecki M, Stooksbury D, et al. (2002) The drought monitor. *Bulletin of the American Meteorological Society* 83(8):1181–1190
- Syphard AD, Radeloff VC, Keeley JE, Hawbaker TJ, Clayton MK, Stewart SI, Hammer RB (2007) Human influence on California fire regimes. *Ecological applications* 17(5):1388–1402
- Thompson MP, Calkin DE (2011) Uncertainty and risk in wildland fire management: a review. *Journal of environmental management* 92(8):1895–1909
- Thompson MP, Calkin DE, Finney MA, Ager AA, Gilbertson-Day JW (2011) Integrated national-scale assessment of wildfire risk to human and ecological values. *Stochastic Environmental Research and Risk Assessment* 25(6):761–780
- Trigo RM, Sousa PM, Pereira MG, Rasilla D, Gouveia CM (2016) Modelling wildfire activity in Iberia with different atmospheric circulation weather types. *International Journal of Climatology* 36(7):2761–2778
- Urbieta IR, Zavala G, Bedia J, Gutiérrez JM, San Miguel-Ayán J, Camia A, Keeley JE, Moreno JM (2015) Fire activity as a function of fire-weather seasonal severity and antecedent climate across spatial scales in southern Europe and

- Pacific western USA. *Environmental Research Letters* 10(11):114013
- US Census Bureau (2012) 2010 Census of Population and Housing, CPH-2-1, United States Summary. Tech. rep., U.S. Government Printing Office, Washington DC
- USDM (2020) United states drought monitor, the national drought mitigation center, university of nebraska-lincoln. <https://droughtmonitor.unl.edu/>
- Viegas DX (2004) Slope and wind effects on fire propagation. *International Journal of Wildland Fire* 13(2):143–156
- Weise DR, Biging GS (1994) Effects of wind velocity and slope on fire behavior. In: Takashi Kashiwagi, ed. 1994. *Fire Safety Science Proceedings of the Fourth International Symposium*. Intl. Assoc. for Fire Safety Science, Boston, MA: pp: 1041-1051, vol 4, pp 1041–1051
- Weise DR, Biging GS (1996) Effects of wind velocity and slope on flame properties. *Canadian Journal of Forest Research* 26(10):1849–1858
- Yang J, Weisberg PJ, Dilts TE, Loudermilk EL, Scheller RM, Stanton A, Skinner C (2015) Predicting wildfire occurrence distribution with spatial point process models and its uncertainty assessment: a case study in the Lake Tahoe Basin, USA. *International Journal of Wildland Fire* 24(3):380–390

Figures

6

Masoudvaziri, Ganguly, Mukherjee, Sun

Figure 1

Temporal availability of data sources. Time interval of the case study is bound by the years 2001 and 2015.

Figure 2

Schematics of the proposed framework. Three general stages are 1) data collection and processing; 2) creating subsets of variables, and selecting candidate models; 3) selecting nal model and statistical inference.

Figure 3

Calculation procedures on different data in raster format. Black lines in the bottom row illustrate the shape of a spatial unit, i.e., a county.

Figure 4

Processed information on wildres. Top left: Total acres burned in each county over the duration of the study. Top right: Total annual acres burned across California. Bottom left: Distribution of acres burned. Bottom right: Transformed distribution of acres burned. This distribution is used in training algorithms.

Figure 5

A framework for selecting the best model

Figure 6

Performance of different algorithms on different input data sets. plots a and b present the performance of models incorporating the subsets made by correlation screening and statistics measures; plots c and d present the performance of models incorporating the subsets made by PCA. In the plots, x-axes represent the features subsets and y-axes represent the relative reduction in errors by different models, compared to the null model.

Figure 7

In-depth comparison of the performance of XGBoost and RF on the data set \correla- tion 1": a) Performance of the models on validation set; b) Comparing the estimations made on training data by XGBoost and RF (i.e., years 2001 to 2014); c) Variable Importance Plot of XGBoost; d) Variable Importance Plot of RF.

Figure 8

Final model: XGBoost implemented on the data set "correlation 1". Top: Tuning of XGBoost hyperparameters. Bottom: Performance of the final model compared to the validation set (year 2015) and average of the training set (years 2001 to 2014).

Figure 9

Variable importance plot of the final model; Top 10 variables.

Figure 10

PDPs of important skin temperature variables.

Figure 11

PDPs of important wind gust variables.

Figure 12

PDPs of important land cover variables.

Figure 13

PDP of important vapor pressure deficit variable.

Figure 14

PDP of important boundary layer height variable.

Figure 15

PDP of important population variable.

Figure 16

Variation of the annual burned area in some of the counties.

Figure 17

Variable Importance Plot of the final model.