# Convolutional neural networks and the wildfire risk of California residential real estate*

Paulo Issler† Richard Stanton‡ Nancy Wallace§ Yao Zhao¶

December 2, 2024

**Abstract**

This paper uses spatiotemporal Convolutional Neural Networks (CNNs) to forecast wildfires across the state of California. CNNs capture both spatial and temporal dependencies and can identify correlations between neighboring data points in a time series. We find that CNN significantly outperforms logistic regression in estimating the likelihood of wildfire. Using our fire-likelihood estimates, we estimate expected annual fire-related property losses for each area, finding wide variation across areas and a total estimated expected loss for 2021 that closely matches the observed cost of fires that year. Finally, we discuss the implications of our results for the future prospects of the property and casualty insurance industry in California.

JEL classification: G21, G54.

Keywords: Housing, mortgages, climate risk, household finance, moral hazard.

*We are grateful for financial support from the Fisher Center for Real Estate and Urban Economics.

†Haas School of Business, U.C. Berkeley, 545 Student Services Building, Berkeley, CA 94720-1900. Email: pauloissler@berkeley.edu.

‡Haas School of Business, U.C. Berkeley, 545 Student Services Building, Berkeley, CA 94720-1900. Email: rhstanton@berkeley.edu.

§Haas School of Business, U.C. Berkeley, 545 Student Services Building, Berkeley, CA 94720-1900 Email: newallace@berkeley.edu.

¶Haas School of Business, U.C. Berkeley, 545 Student Services Building, Berkeley, CA 94720-1900 Email: yaozhao@haas.berkeley.edu.

# 1 Introduction

Wildfire risk threatens the California economy through increased greenhouse-gas emissions, loss of human life (both directly due to the fires and indirectly due to increased air pollution), and losses to real estate and infrastructure.[1] Paci et al. (2023) estimate that wildfires caused economic losses to the state averaging $117.4 billion per year between 2012 and 2021, of which $1.2 billion was the cost of extra greenhouse-gas emissions. Wildfires led to the destruction of more than 60,000 structures and to 302 civilian and firefighter fatalities in California between 2002 and 2021 (Safford et al., 2022). By itself, the 2018 Camp Fire, which burned Paradise, California, caused $27.7 billion in capital losses, $32.2 billion in health costs, $88.6 billion in indirect losses (Wang et al., 2021), 85 deaths, and the destruction of 18,804 structures.[2]

In this paper, we measure vegetative wildfires using data from the California Department of Forestry and Fire Protection's (CAL FIRE) Fire and Resource Assessment Program (FRAP).[3] Figure 1 presents annual wildfire counts and area burned from 2000 to 2021. The average annual number of wildfires was 447 in 2020 to 2021, compared with 314 from 2000 to 2019 (Panel a), an increase of 42.4%.[4] Similarly, the annual average of 3,347,473 acres burned in 2020 and 2021 (Panel c) was more than 5 times the annual average during the prior two decades.[5] Finally, the distribution of fire burn areas from 2000 to 2021 (Panel c) shows a right-tailed skew, similar to that found by Diaz (2022) for the entire western U.S .

To further underscore the distributional characteristics of the fire sizes, Figure 2 shows a QQ-plot of the quantiles of the burned areas against i) a Pareto distribution with shape parameter 0.8 (Panel a); and ii) an exponential distribution with rate parameter 1 (Panel b) . The figure shows that the wildfire-size quantiles almost perfectly match those of the Pareto distribution with shape parameter of 0.8 — a heavy-tailed distribution. In contrast, the plot of the same data against the (thin-tailed) exponential distribution lies far from the 45-degree line. Given the usual caveats of interpreting the graphical results of QQ-plots on finite samples, the plots of Figure 2 suggest that the California historical wildfire size distribution is likely to be characterized by a heavy tailed distribution such as the Pareto.

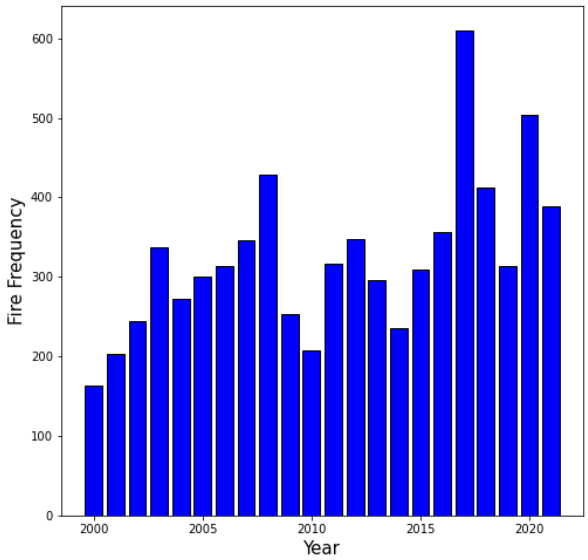
[1] See MacDonald et al. (2023); Safford et al. (2022).

[2] See https://www.fire.ca.gov/media/4jandlhh/top20$_$acres.pdf.

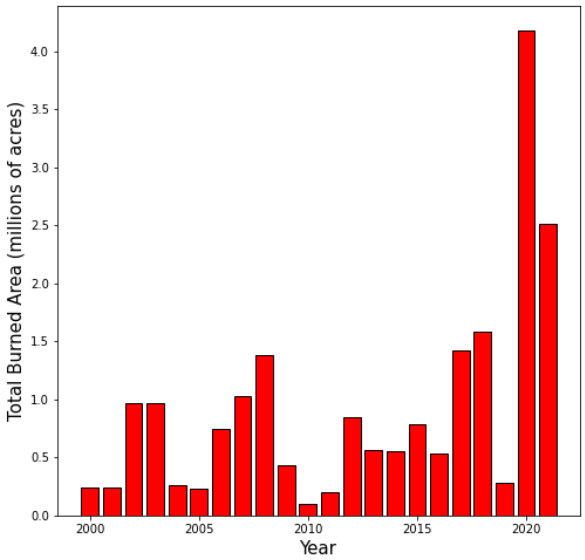
[3] The Fire and Resource Assessment Program (FRAP) distributes annual wildfire perimeter data sets for all public and private lands in California. CAL FIRE defines vegetative wildfires as burning a minimum area of 10 acres for timber fires, 30 acres for brush fires, and 300 acres for grass fires. The GIS data is developed with the cooperation of the United States Forest Service Region 5, the Bureau of Land Management, the National Park Service and the Fish and Wildlife Service (see https://www.fire.ca.gov/what-we-do/fire-resource-assessment-program/fire-perimeters).

[4] Looking further back, Buechi et al. (2021) found that the number of fires over the decade from 2009 to 2018 was 1.4 times the per-decade average between 1979 and 2009.

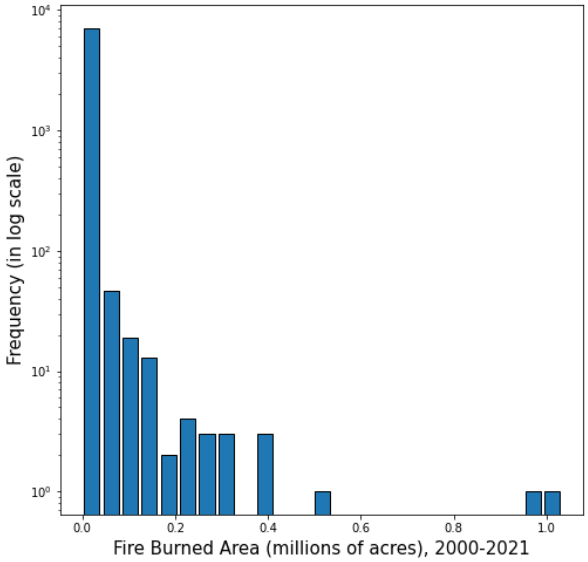
[5] Buechi et al. (2021) also found that the 7.08 million acres burned in the decade from 2009 to 2018 was 1.6 times larger than the per-decade average since 1979, and more than twice that from 1979 to 1988.




(a) Annual wildfire counts



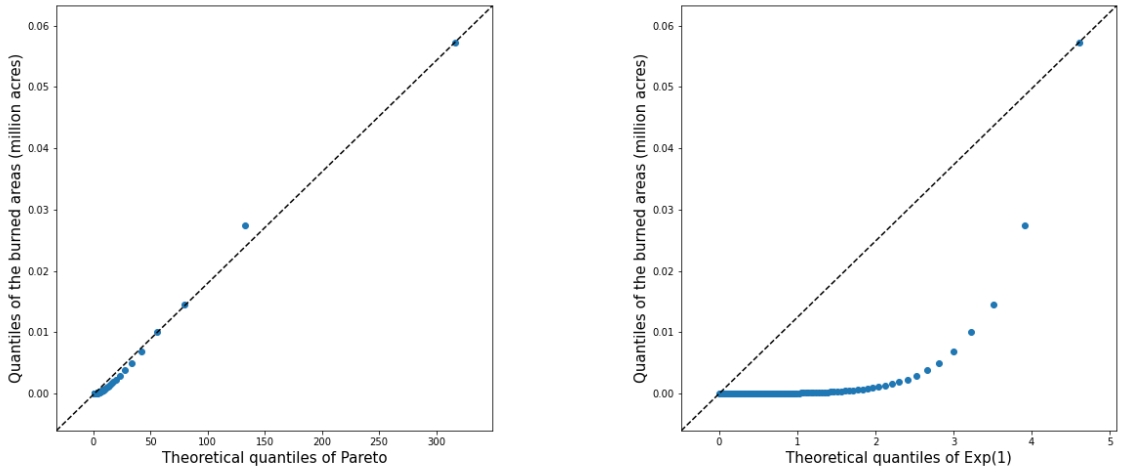

(b) Annual area burned (millions of acres)




(c) Distribution of burn areas

Figure 1: **Frequency and size of California wildfires, 2000–2021**. The wildfire incidence and burn area data are sourced from the California Department of Forestry and Fire Protection (CAL FIRE) (see https://www.frontlinewildfire.com/wildfire-news-and-resources/california-wildfires-history-statistics/).

Thus, following Cooke et al. (2014), reliance on the historical average of wildfire sizes, which has been the policy of the California Department of Insurance under Proposition 103 since 1989, would not be sufficiently informative for future predictions.




(a) Pareto Distribution

(b) Exponential(1) Distribution

Figure 2: **Historical fire burned area shows a heavy-tailed distribution.** We construct two QQ plots of all CAL FIRE identified fire burned areas between 2000 and 2021. The left figure is plotted against the theoretical quantiles of a Pareto distribution with shape parameter 0.8, which is a heavy-tailed distribution with infinite mean and variance. The right one is plotted against an exponential distribution with rate parameter 1, which is a thin-tailed distribution.

These results also suggest potential challenges to accurate statistical forecasting of California wildfires, for methods to diversify and securitize wildfire risks, for reserve strategies under Value-at-Risk management regimes, and for the design of risk management strategies due to spatial dependencies that affect many people, properties, and insurance lines simultaneously (see Abatzoglou and Williams, 2016; Joseph et al., 2019; Kousky, 2019; Kousky and Cooke, 2009; Li and Banerjee, 2021). As Kousky and Cooke (2009) and Cooke et al. (2014) point out, climate-change-related risk distributions such as these often have very small—even undetectable—correlations between variables,[6] they have heavier tails in which the probabilities of ever more serious damage decrease slowly relative to the extent of the damage, and tail dependencies in which bad outcomes are more likely to occur together. Some research

[6]The existence of global correlations can arise when every variable is correlated with some latent variable such as temperature or the effects of El Niño on the Pacific Ocean (see Flannigan et al., 2016; Voosen, 2024).

suggests that climate change may be directly fattening the tails of the distributions of many extreme events (Cardil et al., 2021; Koh et al., 2023).[7]

High temperatures and low precipitation play an important role in enhancing the flammability of vegetative fuels. Figure 3 shows the maximum annual temperature for the West Climate Region in the U.S. from 1895 to 2023, where a pronounced long-run increase can be seen. The current consensus view among climate scientists is that increasing global and regional temperatures increases of about 1.1°C (2°F) since 1980 are largely attributable to anthropogenic greenhouse gas emissions (Burke et al., 2021; MacDonald et al., 2023; Safford et al., 2022; Williams and Abatzoglu, 2016).

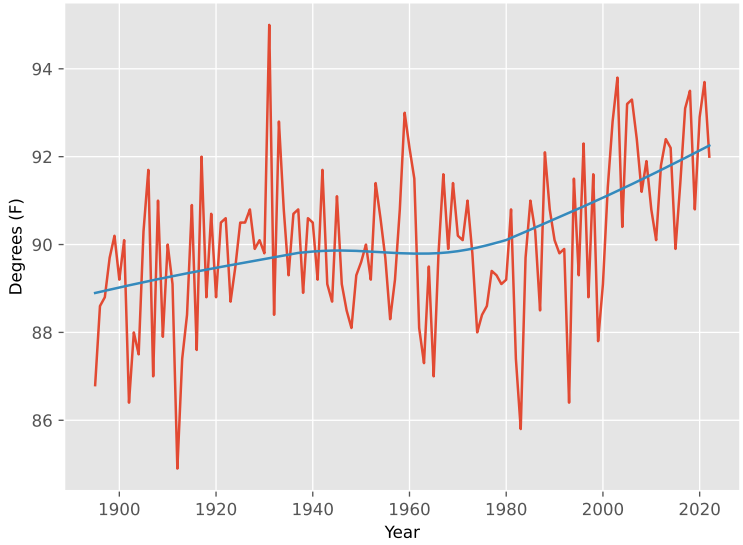



Figure 3: **Maximum annual temperature**. The red line shows the annual maximum temperature for the West Climate Region between 1895 and 2023 from the National Oceanographic and Atmospheric Administration (see https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/regional/time-series), with LOWESS trend line (Cleveland, 1979) in blue.

Although there is a large statistical literature investigating these and other factors affecting wildfire risk, much of this literature uses logistic regression or related extensions (Xi et al., 2019). While machine-learning models have been used, their application is mainly limited to cross-sectional and short-run time-series forecasting applications (see Casolaro et al., 2023; Chen et al., 2023; Makridakis et al., 2023). In this paper we estimate wildfire risk using spatiotemporal Convolutional Neural Networks (CNNs), a technique that has been used in

[7]See also the research goals of *The U.S. Global Change Research Program 2022–2031 Strategic Plan* (https://downloads.globalchange.gov/strategic-plan/2022/USGCRP_2022-2031_Decadal_Strategic_Plan.pdf).

other fields such as image recognition and traffic flow forecasting, but has not previously been applied to climate modeling. CNNs are particularly useful in our setting because they capture both spatial patterns and temporal dependencies and effectively identify correlations between neighboring data points in a time series. We find that CNN significantly outperforms logistic regression in estimating the likelihood of wildfire at a given point in space and time. Combining our fire-likelihood estimates with measures of the value of houses in each area, we next estimate expected annual fire-related property losses for each area. We find enormous variation across different areas, and a total estimated expected loss for 2021 that closely matches the actual costs of fires in that year.

The paper is organized as follows: Section 2 presents a discussion of the wildfire-related risks currently facing property and casualty insurance companies in California. Section 3 discusses the current state of wildfire modeling and presents our spatiotemporal CNN model. Section 4 discusses the data used in our analysis. Section 5 compares estimation results using CNN with those from logistic regression and calculates expected property losses in the state. Section 6 discusses recent problems faced by the insurance company State Farm in light of our estimation results, and Section 7 concludes.

# 2 Property and casualty insurance and wildfire risk

The California property and casualty insurance industry is facing significant wildfire-related challenges. For an insurance company to stay solvent, it must have access to enough capital to pay losses even in catastrophic years. For non-disaster lines of insurance, such as automobile insurance, the premiums in any given year are usually enough to cover claims from that year. For fire insurance, however, claims may greatly exceed annual revenue, due to the heavily skewed burn area distribution shown in Figure 1.[8] Property and casualty insurance firms underwriting wildfire risk in California must solve an intertemporal smoothing problem to cover their catastrophic loss years (see Jaffee and Russell, 2013). The likelihood of very significant losses associated with California wildfires, especially in the last five year, has required property and casualty insurance companies to build up reserves, purchase reinsurance, and use other insurance-linked securities to be able to pay their claims in high damage years (see Goss et al., 2020; Opitz, 2023). As shown in Figure 4, loss ratios for fire peril were significantly impacted by the heavy wildfire-loss years of 2017 and 2018. Only at the end of 2019, after two straight years of insurers paying out $1.85 in losses for every $1 of premium

[8]For example, in 2018 the California insurance industry collected $939,112,586 in property insurance premia and paid out $1,534,083,985 in incurred losses, a 164.22% loss ratio (https://www.insurance.ca.gov/01-consumers/120-company/04-mrktshare/2021/upload/PrmLssChartHistorical2021wa.pdf).

earned, did the California Department of Insurance approve 71 rate-increase requests from 50 different companies.[9]

There are also other internal capital market challenges for property casualty insurance companies that hinder solutions to their intertemporal smoothing problem. U.S. accounting requirements preclude earmarking capital surplus to a specific risk and current tax provision require that retained earnings are taxed as corporate income at set aside Jaffee and Russell (2013). Additionally the accumulation of capital holdings to preclude losses in particularly catastrophic future years, make these companies very susceptible to takeover risks. Another challenge has been the refusal of the California Department of Insurance to allow the inclusion of reinsurance costs in the calculation of premia for wildfire insurance.[10] As a result of these on-going challenges, between 2012 and 2022 California homeowners insurance companies performed significantly worse than the national average on key risk metrics such as the direct incurred loss ratio which was 73.9% compared to the U.S. average of 59.7% and the average direct underwriting profit of $-13.1\%$ as compared to the U.S. average of 3.6%.[11]

Rate setting limitations are another major challenge for the long-term viability of homeowner fire casualty insurance in California. In 1988, California voters passed Proposition 103, which required insurance companies to receive "prior approval" from the California Department of Insurance (CDI) before implementing property and casualty insurance rates. As shown by Oh et al. (2024), casualty insurance rates in states like California with high regulatory frictions have not adequately adjusted in response to the growth in losses. In addition, California state insurance regulations require wildfire insurers to set rates for future annual catastrophic coverage as the fraction of damages accrued from the 20-year historical mean rather than statistical, or actuarial, models. Additionally, the CDI does not allow for the costs, or changes in the cost, of reinsurance risk to be included in insurer rate requests. As a result, California's annual rates now rank next to the lowest in the U.S. (see Oh et al., 2024), perhaps threatening the future ability of California homeowners to successfully rebuild and continue to make their mortgage payments after large and destructive wildfires.[12]

On September 21, 2023 California Governor Gavin Newsom issued an Executive Order to authorize the State Insurance Commissioner, Ricardo Lara, to exercise his authority to

[9]See https://www.insurance.ca.gov/0250-insurers/0800-rate-filings/0100-rate-filing-lists/rate-filing-approvals/rate-filing-approvals-yearly.cfm.

[10]See California Code of Regulations. Title 10. § 2644.25. Reinsurance (https://govt.westlaw.com/calregs/Document/).

[11]http://insurance.ca.gov/0400-news/0100-press-releases/2023/upload/California-s-Sustainable-Insurance-Strategy-slides.pdf

[12]Since 2022, AIG and Chubb have left the high-value home insurance market. State Farm, Farmers, Allstate, USAA, Travelers, and Nationwide have all either limited or paused writing new policies (see https://www.insurance.ca.gov/01-consumers/180-climate-change/SustainableInsuranceStrategy.cfm).

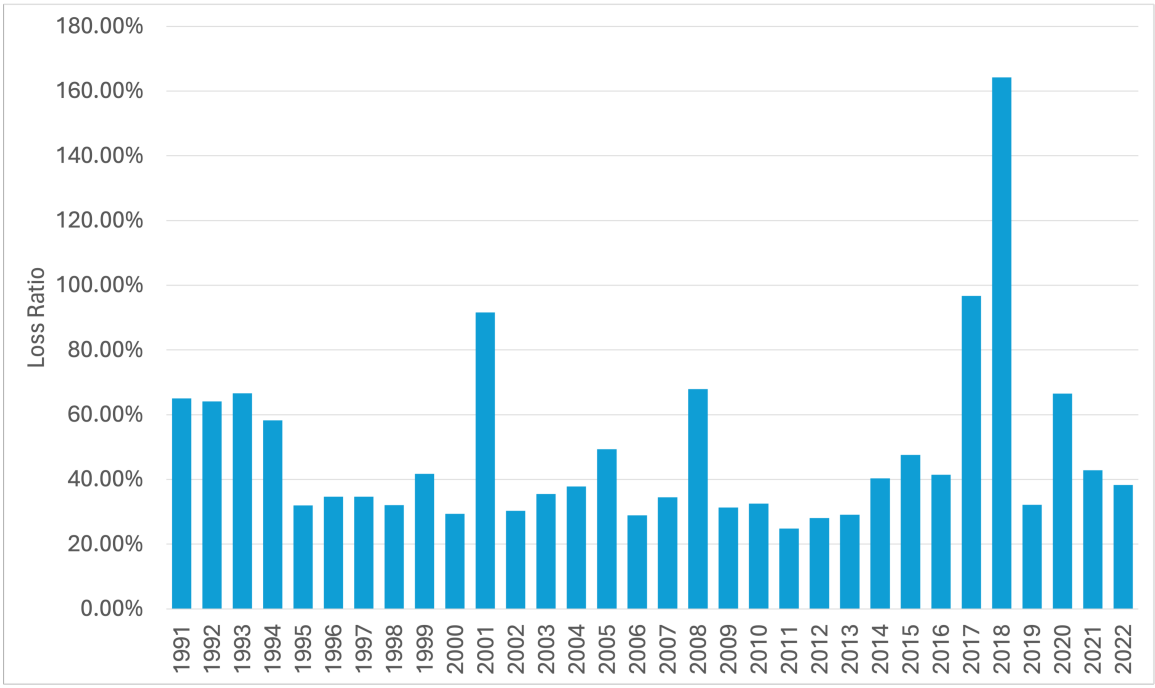



Figure 4: Realized loss rates (fire peril) for California Property and Casualty insurance companies. Source: https://www.insurance.ca.gov/01-consumers/120-company/04-mrktshare/2022/upload/PrmLssChartHistorical2022.pdf

stabilize California's property insurance markets.[13] At the same time, the CDI introduced the California Sustainable Insurance Strategy, which will allow insurance carriers in the future to apply forward-looking catastrophe models to more accurately assess and price climate-related risks in exchange for expanded property insurance coverage in risky areas.[14] The insurance commissioner's office is currently in the process of developing regulations for how exactly the new models can be used for rate setting in the future and who will vet these models.[15]

Boomhower et al. (2024) provide a rough summary measure of the relative granularity or complexity of the pricing algorithms used by California Homeowners Insurance companies. Given the available data on the models, they build a measure of model complexity, by counting the number of risk-rating variables that each insurer uses to assess the likelihood of wildfire damages for specific home locations. They find that some California insurers price wildfire risk using zip-code-level territory factors and others use parcel-level categorical

[13] https://www.gov.ca.gov/wp-content/uploads/2023/02/Feb-13-2023-Executive-Order.pdf.

[14] https://www.insurance.ca.gov/01-consumers/180-climate-change/SustainableInsuranceStrategy.cfm.

[15] https://www.politico.com/news/2023/09/21/newsom-orders-action-on-wildfire-insurance-00117488.

wildfire risk scores based on qualitative factors such as slope, vegetation, fuel load, and road access. The larger insurers use more granular measures generated by using probabilistic catastrophe models. Overall, it is the firms with the largest market share in high-hazard zip codes that use the most granular risk segmentation. They argue that the observed heterogeneity in California insurance price/risk modeling primarily reflects. the direct costs of licensing or developing state-of-the-art wildfire models. These costs together with the indirect costs of adoption include adapting the firm's internal systems and employing professional staff can run into annual costs of millions of dollars (see Jergler, 2021).

The effects of Proposition 103 and the California Department of Insurance's historical reluctance to allow probabilistic models to justify firm-level requests for rate increases, have led to the significant heterogeneity in the current state of statistical rate-setting technology found by Boomhower et al. (2024). The new policies of the California Department of Insurance should allow more active competition among firms to develop state-of-the-art wildfire modeling technology for wildfire rate pricing at the property level. The current challenge is that firm-level development of these newly allowed probabilistic wildfire risk models and the regulatory vetting of the models is unlikely to be fully completed by December 2024.

# 3 Wildfire modeling

The study of wildfire occurrence has led to a large statistical literature (for an overview see Oliveira et al., 2021; Prestemon et al., 2013; Xi et al., 2019) and a growing machine-learning literature (see Cruciata et al., 2024; First Street Foundation, 2022; Koh et al., 2023; Opitz, 2023) focused on identifying risk factors and producing risk maps or indices.[16] These modeling strategies are best suited to the investigation of general trends across wildfires as a function of features that are predictive of when and where wildfire ignitions occur.

The biophysical variables found in occurrence models are intended to causally explain why wildfire ignitions vary across space and time as the result of temporal and spatial variations in weather, climate, vegetative land coverage, and topography (see Brinkmann et al., 2022; First Street Foundation, 2022; Kearns et al., 2022; Prestemon et al., 2013). Other common variables used in statistical occurrence modeling include anthropogenic features that affect the rising risks of wildfire due to the interaction of human activity and climate (see Abatzoglou et al., 2018; Abatzoglou and Williams, 2016; Apt et al., 2023; Williams and Abatzoglou, 2020; Williams and Abatzoglu, 2016) as well as the effects of the increasing

[16] An even more recent literature, focused on physics-informed machine learning, has provided accurate and efficient ways of recognizing complex patterns and predicting spatiotemporal weather and climate processes that obey fundamental laws governing physical systems (Kashinath et al., 2020; Seydi et al., 2024).

encroachment of urban development into the wildland urban interface (see Alexandre et al., 2016; Kestelman, 2024; Price and Bradstock, 2014; Radeloff et al., 2018).

Historically, the most commonly used form of wildfire occurrence modeling has been logistic regression models or related extensions such as logistic generalized additive models (Xi et al., 2019). More recently, machine learning models such as gradient boosted random forest, neural networks, deep neural network, multi-layer perceptrons, and (for image analysis) convolutional neural networks (see Alkhatib et al., 2023; Ismail and Amarasoma, 2023; Jain et al., 2020; Tong and Gernay, 2023) have been used for wildfire occurrence modeling. However, other than the physics-informed machine learning models, the current generation of machine learning methodologies are limited to cross-sectional and short-run time-series forecasting applications (see Casolaro et al., 2023; Chen et al., 2023; Makridakis et al., 2023).

## 3.1 Standardization of the wildfire event space

Among climate scientists, the presence or absence of wildfire occurrence is measured over discrete space-time cells, called voxels, that are projected onto the earth's surface. The cell centroid is precisely defined at the latitude and longitude of the centroid location and the raw weather and climate data are usually measured as satellite-data projections to the centroids. Other measures such as vegetative land cover, urban density, or infrastructure measurement are also standardized to cell-level representations so as to harmonize the different spatial-temporal scales of wildfire and predictor data such as weather conditions, land cover and land use (Abatzoglou, 2013; Koh et al., 2023). Currently available climate space-time cells are sized at 4 km $\times$ 4 km or less for the dynamic hourly or daily climate and weather measures. The static predictors related to vegetative land coverage, topography, and housing density and electricity transmission lines among other measure are available at raster, or pixel levels, so that they can be easily merged to the cell data.

Given the ready availability of economic and demographic indicators at the census tract or zip-code level, several recent papers in the economics literature also use measures of wildfire occurrence at the zip-code or census-tract level (see Biswas et al., 2023; Kahn et al., 2024). A first concern with the use of census tracts or zip codes for wildfire occurrence measurement is that zip codes and census tracts are based on population not geography. The use of zip codes is especially a problem in California, where the average zip code is 60.12 square miles, the largest zip code is 1,773 square miles, and the smallest is 0.01 square miles. Additionally, zip codes are not spatially or temporally consistent with climatological, topographical, or vegetative measurement (see Abatzoglou, 2013). Secondly, any county-level or zip-code-level predictions may suffer from inaccuracy, because they will either incorrectly assume the

entire area is burned, or rely on historical fire sizes that as shown in Section 1 are likely to be heavy-tailed. This again supports our choice of conducting cell-level predictions, because predicting cell by cell can avoid using direct historical estimates of the fire size.

## 3.2 Convolutional Neural Networks

Spatiotemporal CNNs (Convolutional Neural Networks) are well suited to short-term forecasting problems due to their ability to automatically extract important spatial and temporal features from data without relying on hand-crafted features (Casolaro et al., 2023). Common applications of these models include still-image recognition and action recognition in videos (see Tran et al., 2018). The models have also been successfully applied to address the spatial correlations, temporal correlations, and heterogeneity of the traffic flow data used to forecast urban traffic congestion (Guo et al., 2019) and to the satellite detection of canopy-scale tree mortality and survival from California wildfires (see Dixon et al., 2023). However, to our knowledge they have not previously been applied to modeling the occurrence of wildfires.

Spatiotemporal CNN models are designed to capture spatial patterns. They are also effective in modeling temporal dependencies and identifying correlations between neighboring data points in a time series, just as they can recognize objects in images by analyzing patterns in pixel values. By applying 3-dimensional convolutional filters across both space and time, spatiotemporal CNNs can learn the motion patterns in time series data and fully use those patterns to account for how past values influence future predictions. Factorizing 3-dimensional convolutions into spatial and temporal components can further enhance accuracy and efficiency leading to lower training and testing errors (Sra, 2019).

An additional advantage of 3-dimensional CNNs is that they can introduce nonlinearities into the network, thus allowing for the complex functions that are needed to accurately model the joint spatial correlations and temporal dynamics of wildfire prediction. Another strength of the models for wildfire occurrence modeling is that they easily handle the cell adjacency correlation structure of wildfire — since if there is a wildfire in one cell location nearby cell locations are likely to also have wildfires. These CNNs also allow for the temporal aggregation of some wildfire features by accounting for the cumulative effects of phenomena such as maximum temperature and vegetative dryness on the cell-by-cell occurrence of wildfires. Another important strength includes the benefit of regularization. The fitted values from spatiotemporal CNNs are by nature correlated in space and time, which helps to prevent overfitting even with a high-dimensional nonlinear parameter space. They are also parsimonious because the weight parameters are shared across space and across time, thus significantly reducing model complexity. Finally, a weighted loss function can be applied to treat the

severe data imbalances that exist in most wildfire data arrays over time and space.

For our application of spatiotemporal CNNs, we are forecasting one year ahead to 2021, since the typical maturity of an homeowners insurance contract is annual. However, longer-run out-of-sample forecasting remains a limitation for applications of spatiotemporal CNN to longer term contracting. Another potential drawbacks of using spatiotemporal CNN is that the model needs to be slightly customized to prohibit the use of future temporal features to forecast future wildfire occurrences. However, as will be discussed below, overall spatiotemporal CNNs offer a powerful approach to time-series forecasting, particularly for tasks where capturing both spatial and temporal dependencies is crucial. Their ability to automatically learn relevant features from raw data, combined with the flexibility offered by different spatiotemporal convolutional designs, makes them a valuable tool for predicting future values in various time-dependent domains.

## 3.3 A spatiotemporal CNN model

Suppose we are given a panel dataset $\{(X_{it}, y_{it})\}_{i\in I, t\in T}$, where $y_{it}$ is the dependent variable observed for location $i$ at time $t$ and $X_{it} = (x_{it}^1, x_{it}^2, \cdots, x_{it}^k)'$ is the corresponding $k$-dimensional vector of explanatory variables. $I$ and $T$ denote the sets of location and timestamps, respectively. Assume we have $N$ locations in total, i.e., $|I| = N$. This is a common data structure in many location-based studies, such as real estate, climate, transportation, etc. For generalization purposes, we do not make specific assumptions about the data types of $y_{it}$. It can be either binary if it is a classification problem or continuous if it is a regression problem. Generally, we are trying to model the target variable $y_{it}$ as a function of $\{X_{is}\}_{i\in I, s\leq t}$, which contains all historical information across space up to time $t$.

**Notation** Stack the data across locations and define

- $\boldsymbol{y}_t = (y_{1t}, \cdots, y_{Nt})'$, an $N \times 1$ vector whose $i$-th entry is $y_{it}$ for location $i$,
- $\boldsymbol{X}_t = (X_{1t}, X_{2t}, \cdots, X_{Nt})'$, an $N \times K$ matrix, whose $i$-th row is a $k$-dimensional vector of characteristics for location $i$.

Now further stack the data across time and define

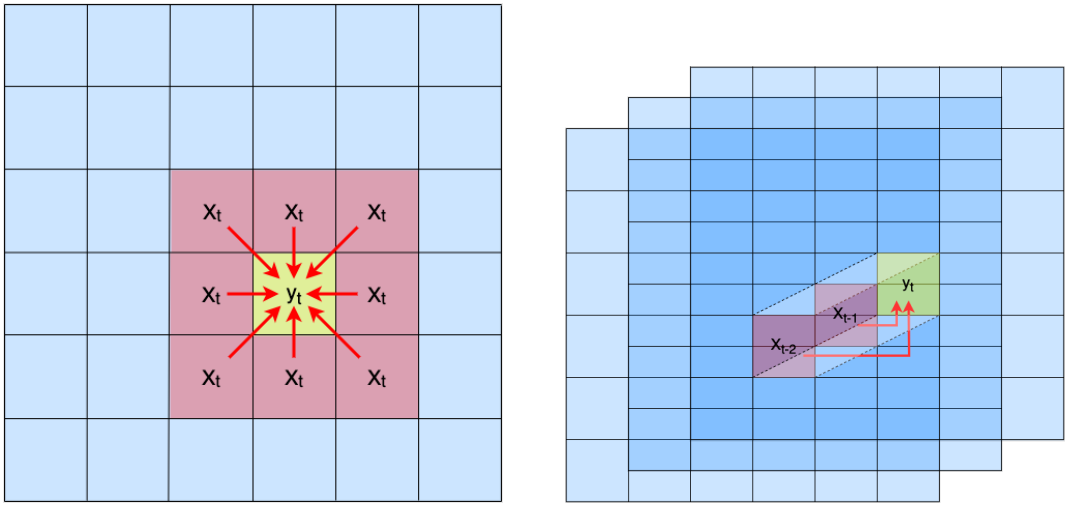
- $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_T)'$, a $T \times N$ matrix,
- $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_T)$, a tensor with dimension $T \times N \times K$.

## 3.4 Spatial and temporal dependence

Intuitively, the simplest model to start with is OLS for regression or logistic model for binary classification, where we assume all observations across space and time are independent and simply use the cross-sectional variations to fit $\boldsymbol{y}_t$ for any $t \in T$.

$$\boldsymbol{y}_t = \begin{cases} \sigma(\beta_0 \mathbf{1} + \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t) & \text{if classification,} \\ \beta_0 \mathbf{1} + \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t & \text{if regression,} \end{cases}$$

where $\sigma$ is the sigmoid function applied element-wise to the vector input, $\mathbf{1}$ is a vector of ones, $\boldsymbol{\beta}$ is the vector of coefficients, and $\boldsymbol{\varepsilon}_t$ is a vector of independent errors.




(a) Spatial dependence (b) Temporal dependence

Figure 5: **Visualizing the potential dependence structure in a spatiotemporal dataset.** The yellow cell refers to a target location-time, whose $y_{it}$ is our modeling goal. In addition to the yellow cell's own observations $X_{it}$, the characteristics of its nearby red cells are assumed to have an effect on the modeling goal as well, which will lead to spatial and temporal dependence. The red cells in Figure 5a are called the "spatial lags," while those in Figure 5b are the usual lags in time series. The remaining blue cells are assumed to have no impact on the yellow cells. In Figure 5b, each big square represents a slice of the panel data at a particular time.

Figure 5 demonstrates the potential dependence structure in space and time for a spatiotemporal panel dataset. Taking the wildfire as an example, at each time $t$ the maximum temperatures in location $i$'s nearby cells could also contribute to the wildfire occurrence in location $i$ through the flow of air, which is the spatial dependence plotted in Figure 5a. In spatial econometrics, the nearby influencing cells all called "spatial lags." In comparison, Figure 5b plots the usual "lags" to show temporal dependence, which means in location $i$ the

maximum temperatures of the past few days or weeks could impact the wildfire occurrence at time $t$ as well, because of the accumulated heat over time. One difference between the spatial and temporal dependence is that the spatial dependence does not have any specific direction, while the temporal dependence must be one-directional, meaning that only past values can affect the future, but not the other way round. However, in space any direction is allowed. There could exist the interaction of spatial and temporal lags too, as demonstrated in Figure 6. In short, this means the spatial lags of the temporal lags, or equivalently, the temporal lags of the spatial lags, might also affect our target location-time.

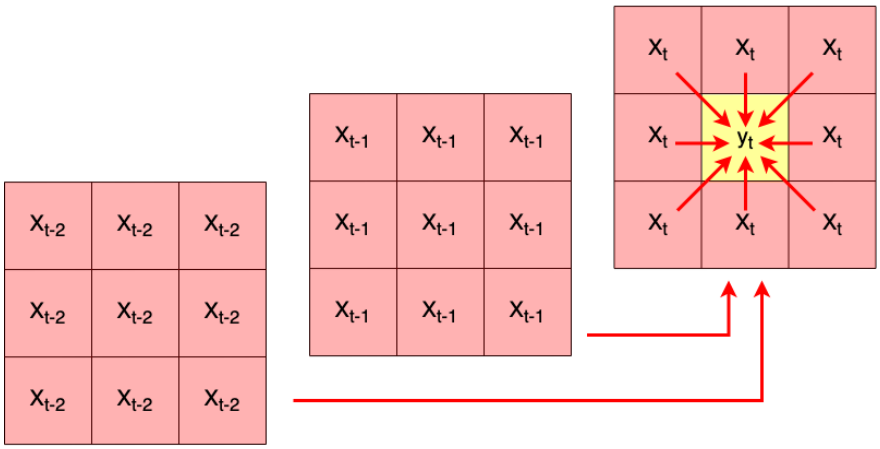



Figure 6: **Interactive spatial and temporal dependence.** Again, the yellow cell represents our target location, and the characteristics of red cells affect the target location's $y_{it}$ as well. For convenience, we plot only the cells which matter for the target location. Obviously, they are either spatial lags (as in time $t$) or the spatial lags of the temporal lags (as in time $t-2$ and $t-1$).

To sum up, the analysis above suggests that when dealing with spatiotemporal datasets, it is essential for the model to capture the dependence across both space and time. In the next two subsections, we will review several classical econometric models that are designed to deal with these potential dependence, and then explain how CNN corresponds to better versions of these models.

## 3.5 Finite distributed lag model and 1-d convolution

In time series econometrics, the finite distributed lag (FDL) model is designed to include lagged explanatory variables to fully account for delays in the explanatory variables.

$$\boldsymbol{y}_t = \begin{cases} \sigma(\beta_0 \mathbf{1} + \boldsymbol{X}_t\boldsymbol{\beta}_t + \boldsymbol{X}_{t-1}\boldsymbol{\beta}_{t-1} + \cdots + \boldsymbol{X}_{t-q}\boldsymbol{\beta}_{t-q} + \boldsymbol{\varepsilon}_t) & \text{if classification,} \\ \beta_0 \mathbf{1} + \boldsymbol{X}_t\boldsymbol{\beta}_t + \boldsymbol{X}_{t-1}\boldsymbol{\beta}_{t-1} + \cdots + \boldsymbol{X}_{t-q}\boldsymbol{\beta}_{t-q} + \boldsymbol{\varepsilon}_t & \text{if regression,} \end{cases}$$

where $q < \infty$ is the order of lags, $\boldsymbol{X}_{t-1}, \cdots, \boldsymbol{X}_{t-q}$ are lagged explanatory variables that are added to capture the temporal dependence, $\boldsymbol{\beta}_{t-1}, \cdots, \boldsymbol{\beta}_{t-q}$ are the corresponding coefficient vectors for different lags, and $\boldsymbol{\varepsilon}_t$ is the independent error term vector.

The idea of including lagged explanatory variables, as in the FDL model, can be perfectly replicated by an 1-d convolutional neural network (CNN). Figure 7 graphically shows how a filter in the 1-d convolutional layer works. We fix an arbitrary location cell $i$, and describe its observed $K$ features as a $K$-dimensional time series. For each feature, or equivalently, "channel" in CNN's terminology, applying the 1-d convolution can be considered as calculating weighted moving averages over a constant window. This moving window is called a "kernel", whose weights are the CNN parameters that will be learned by training. The kernel size is a hyper-parameter that can be tuned via validation. Different kernel sizes are equivalent to different order of lags $q$ in the FDL model. For convenience, we usually choose an odd number as the kernel size, so that for every moving average we will have a unique median cell as the target output cell, and the number of cells around the median cell will be the same on both sides.

The 1-d convolution is conducted channel by channel for all features. The convolutional outputs will then be linearly combined. The only difference between the 1-d convolutional filter and the FDL model is that a non-linear activation function relu will be applied to the final output to enhance model complexity. Mathematically, a filter in the 1-d convolutional layer can be described as below. Assuming that the kernel size is $p$, then for each cell $i$,

$$y_{it} = \text{relu}\left(\beta_0 + \sum_{k=1}^{K} \beta_k \sum_{l=1}^{p} w_l^k x_{it-L+l}^k\right),$$

where $L = (p+1)/2$ is a constant given $p$, $x_{it}^k$ represents the $k$-th explanatory feature in location $i$ at time $t$, and $\{w_l^k\}_{l=1,\ldots,p}$ are the kernel weights for the $k$-th channel.

With the default convolutional setting shown above, one might be concerned about a *looking-ahead problem* in the convolution, meaning that future values of the explanatory variables are involved in the convolution. To make sure that the temporal dependence is one-directional, we shift the filter output backward, as Figure 8 demonstrates. If we shift the output backward by $L-1$ cells, we can guarantee that for each moving window, the last time cell in the convolution happens to be the target cell. However, note that Figure 8 is just a conceptual visualization. In practice, one needs to shift the time series of the dependent

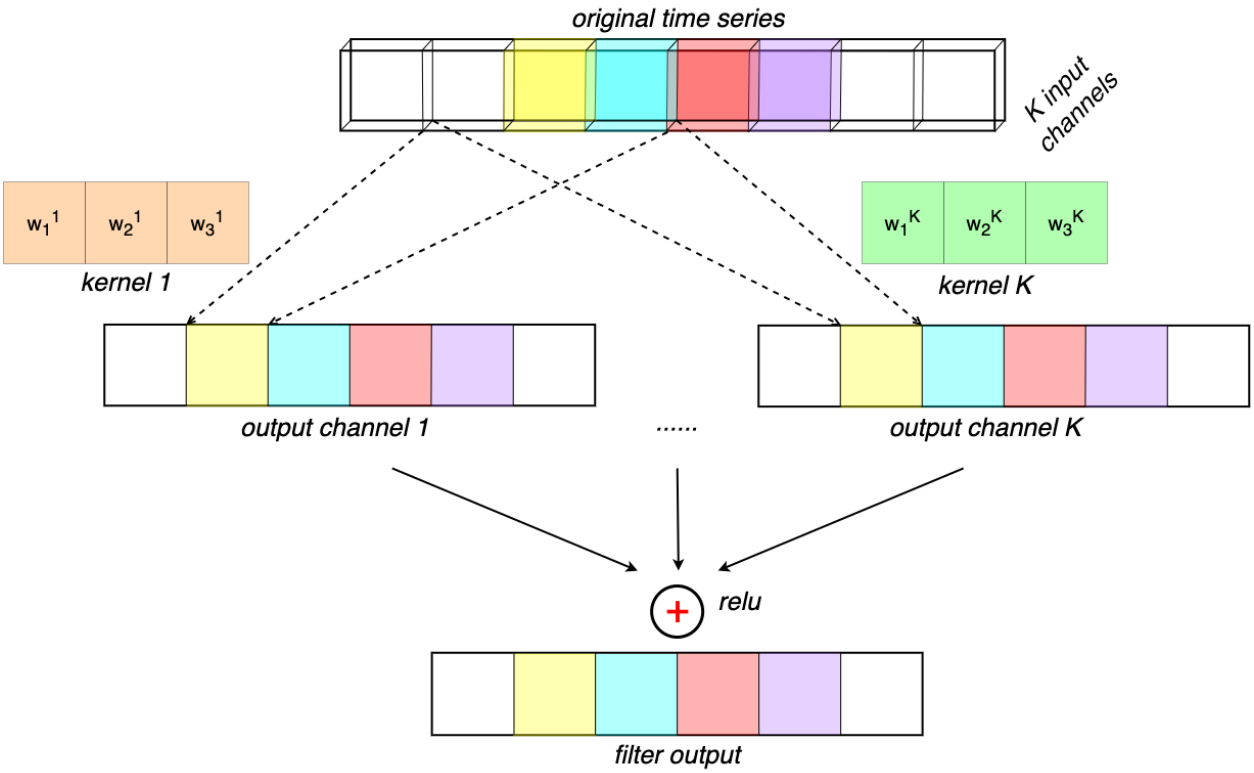



Figure 7: **An example of a filter in the 1-d convolutional layer.** In this figure, we use kernel size 3, which is arbitrarily chosen, as an example to show how a filter works in the 1-d convolutional layer. Note that each cubic or square cell represents a timestamp $t$. At the top we plot the original $k$-dimensional time series, where each dimension corresponds to a feature or an input "channel" in the CNN. Each channel is associated with a specific kernel, which can be understood as the convolutional weights applied to that channel. Specifically, the 1-d convolution is conducted by taking the inner product of the $3 \times 1$ kernel and the channel value of three consecutive cells, on a rolling basis across the entire channel, from left to right. We use the same color to keep track of the same timestamp $t$. For example, the value of the yellow cell in output channel 1 is the inner product of the kernel 1 vector and the closest 3 cells to the yellow cell, including itself, in the input channel 1 (marked by dashed lines). In other words, the central cell in the convolution is always the target cell. For simplicity, this figure plots only 1 filter, so the filter output has only 1 dimension. But in practice there could be a number of filters. The output of all filters will be stacked to form a multi-dimensional time series, which then becomes the input of the next layer in the neural network.

variable to match the time horizon. With the filter output being shifted, we can then rewrite the 1-d convolution formula as

$$y_{it} = \text{relu}\left(\beta_0 + \sum_{k=1}^{K} \beta_k \sum_{l=1}^{p} w_l^k x_{it-p+l}^k\right).$$

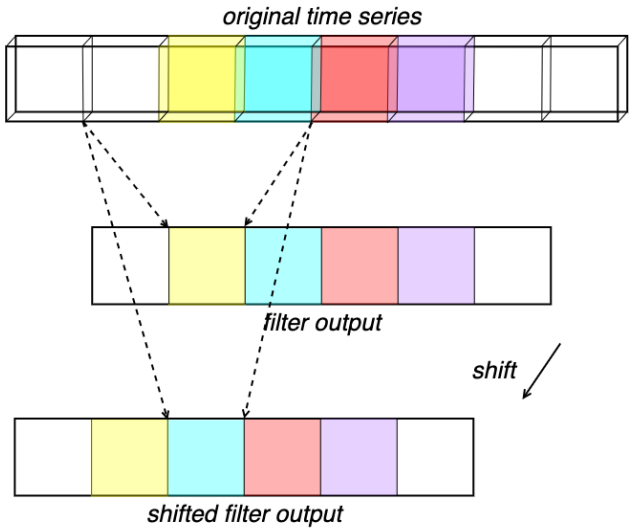



Figure 8: **Customize the 1-d convolution to avoid looking ahead.** In this figure, we conceptually demonstrate how to shift the convolution output so that we can avoid using future values of explanatory variables for estimation. Each cubic or square denotes a timestamp, and we use the same color to keep track of the same timestamp. Take the yellow cell as an example. Before shifting the output, the convolution will involve the blue cell, which is a future time. But after we shift the output backward, the target cell/timestamp of the same convolution changes to the blue one. The number of periods to shift backward will depend on the kernel size.

The 1-d convolution applies to the time dimension. Because it keeps different locations independent, we can simply stack the outputs across different locations.

$$\boldsymbol{y}_t = \text{relu}\left(\beta_0 \mathbf{1} + \boldsymbol{X}_t \boldsymbol{W}_p \boldsymbol{\beta} + \boldsymbol{X}_{t-1} \boldsymbol{W}_{p-1} \boldsymbol{\beta} + \cdots + \boldsymbol{X}_{t-p+1} \boldsymbol{W}_1 \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t\right),$$

where $\boldsymbol{W}_l = \text{diag}(w_l^1, w_l^2, \cdots, w_l^K)$ is a diagonal weight matrix, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_K)'$ is the coefficient vector. Note that the above formula is exactly in the same form as the FDL model in time series econometrics, except for two small differences. First, the order of lags in the FDL model $q$ differs from the kernel size $p$ by 1, but this is simply because of different definitions. Second, in the 1-d CNN model, an additional non-linear activation function relu is added to the output.

## 3.6 Spatial econometric models and 2-d convolution

Several spatial econometric models are proposed to deal with spatial dependence concerns, such as the spatial cross-regressive model (SLX), spatial lag model (SLM) and spatial error model (SEM). In this subsection we will briefly review these classical models, and we will explain how the CNN model could capture these models.

For all spatial econometrics, we will start with a $N \times N$ weight matrix $\boldsymbol{W} = \{w_{ij}\}_{i,j \in I}$, where $w_{ij}$ represents the weight that we impose on location $j$ when target location is $i$. Hence, the $i$-th row of $\boldsymbol{W}\boldsymbol{X}_t$ corresponds to the weighted characteristics of the nearby cells for location $i$, i.e., the spatial lags of location $i$. Note that here $\boldsymbol{W}$ needs to be determined ex-ante.

- Spatial cross-regressive model (SLX)
  In SLX, we introduce the spatially lagged exogenous regressors to the model, assuming that the spatial dependence can be captured by the spatial lags in the explanatory variables, namely $\boldsymbol{W}\boldsymbol{X}_t$.

$$\boldsymbol{y}_t = \begin{cases} \sigma(\beta_0 \mathbf{1} + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}_t\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_t) & \text{if classification,} \\ \beta_0 \mathbf{1} + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}_t\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_t & \text{if regression,} \end{cases}$$

  where the additional term $\boldsymbol{W}\boldsymbol{X}_t\boldsymbol{\gamma}$ will control for the effects of spatial lags, and $\boldsymbol{\gamma}$ are the corresponding coefficients. When $\boldsymbol{\gamma} = \mathbf{0}$, SLX will degenerate to an OLS model. Obviously, this method has drawbacks because the weight matrix $\boldsymbol{W}$ needs to be determined ex-ante. Although in most cases we can assume closer places should have higher weights, we are uncertain about the rate of spatial decay. On the other hand, even after controlling for the observed characteristics of spatial lags, we still have to conduct additional test to examine whether the spatial dependence is fully captured by $\boldsymbol{W}$. The advantage of SLX lies in that it is both conceptually and computationally easy, because the spatial lag variables can simply be treated as additional regressors. Hence, at least for the dependent variables with continuous values (i.e., for regression problems), this model can easily be estimated by OLS.
- Spatial lag model (SLM)
  Similar to SLX, we also use spatial lags to capture the spatial dependence in SLM. However, we switch from $\boldsymbol{W}\boldsymbol{X}_t$ to $\boldsymbol{W}\boldsymbol{y}_t$, assuming that the dependence manifests directly in the dependent variable. Geometrically, adding an autoregressive covariate corresponds to a ripple effect. In contrary to SLX where we usually assume only nearby cells matter, in SLM even if start with a few nearby cells, the spatial dependence could end up spreading much more widely in space, through cascades of lag effects. The rate

of spatial decay depends on the parameter $\rho$.

$$\boldsymbol{y}_t = \begin{cases} \sigma(\beta_0 \mathbf{1} + \boldsymbol{X}_t\boldsymbol{\beta} + \rho \boldsymbol{W}\boldsymbol{y}_t + \boldsymbol{\varepsilon}_t) & \text{if classification,} \\ \beta_0 \mathbf{1} + \boldsymbol{X}_t\boldsymbol{\beta} + \rho \boldsymbol{W}\boldsymbol{y}_t + \boldsymbol{\varepsilon}_t & \text{if regression.} \end{cases}$$

- Spatial error model (SEM)

$$\boldsymbol{y}_t = \begin{cases} \sigma(\beta_0 \mathbf{1} + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{u}_t) & \text{if classification,} \\ \beta_0 \mathbf{1} + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{u}_t & \text{if regression,} \end{cases}$$
$$\boldsymbol{u}_t = \lambda \boldsymbol{W}\boldsymbol{u}_t + \boldsymbol{\varepsilon}_t$$

  If we assume the spatial dependence is mainly due to some spatially correlated omitted variables, we can model it through the error terms. Specifically, we can apply a spatial version of autoregressive model on $u_t$, by assuming the error terms depend on their spatial lags $\boldsymbol{W}\boldsymbol{u}_t$. Hence, the parameter $\lambda$ controls the spatial decay rate, and $\boldsymbol{\varepsilon}_t$ is still the $i.i.d$ noise.

In both SLM and SEM, there exists a spatial autoregressive component, which will cause fairly high computational cost. For example, when $\boldsymbol{y}_t$ has continuous values, then to obtain the closed form solution for estimators we will have to invert a large matrix $(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}$ in SLM or $(\boldsymbol{I} - \lambda\boldsymbol{W})^{-1}$ in SEM, both of which have an $N \times N$ dimension. When $\boldsymbol{y}_t$ is a binary variable, this would be even more complicated because of the additional sigmoid function. In this sense, SLX seems computationally much cheaper.

There is a clear tie between SLX and the 2-d CNN model, in the sense that CNN is one instance of SLX with 1) a particular form of weight matrix and 2) additional nonlinearity. Figure 9 demonstrates a specific example of a filter in the 2-d convolutional layer. First, a filter consists of $K$ "kernels," where $K$ equals the number of covariates, or "channels," in the original data. For each kernel of size $k$, it represents a specific weight matrix, assuming that only the $k \times k$ nearest cells matter and farther away cells will have zero weights. Clearly, the size of spatial lags is determined by the kernel size $k$, which is a hyperparameter in this model. Second, the weights are shared across space. For each target location, we compute the convolution by taking the inner product of the characteristics observed in its $k^2$ closest cells and the kernel, which is equivalently the spatial lags in SLX. Since we use the same kernel for all locations, this will remarkably decrease the number of parameters from $|L|^2$ to $k^2$. Third, different kernels work independently, and the convolutional results are combined linearly, with a nonlinear activation function relu. Last, Figure 9 demonstrates only one filter. Actually one convolutional layers could have several independent filters, and all of the

convolutional results will be passed to the next layer, which works jointly for the estimation.

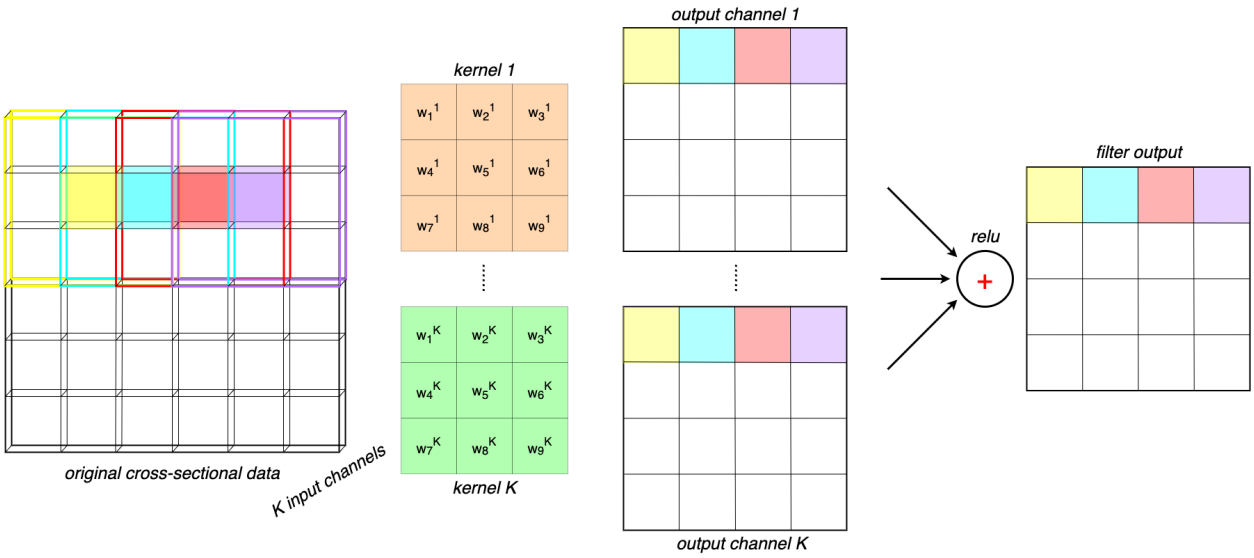



Figure 9: **An example of a filter in the 2-d convolutional layer.** This figure shows an example of a filter in the 2-d convolutional layer, consisting of $K$ kernels of size $3 \times 3$. The original cross-sectional data can be viewed as a map, where each cell represents a specific location. The data has K features, which are called K *input channels*. From left to right, we plot the original map, the kernels, the resulting maps (i.e., the *output channels*) and the final output of this filter. A kernel is associated with a specific weight matrix and its corresponding channel. The colors in the initial map and the channels match with each other. For example, the value in yellow cells are calculate by taking the inner product of the $3 \times 3$ yellow area in the initial map and the weight matrix. Note that the central cell of the $3 \times 3$ yellow area is our target location.

Mathematically, a 2-d convolutional filter in NN can be written as

$$y_{it} = \text{relu}\left(\beta_0 + \sum_{k=1}^{K} \beta_k \sum_{j=1}^{N} w_{ij}^k x_{it}^k\right),$$

where relu is a function that is widely used in neural networks to allow for nonlinearity. Besides, $x^k$ represents the $k$-th covariate/channel, and $w_{ij}^k$ represents the kernel weight on location $j$ when the target location is $i$. Note that $w_{ij}^k = 0$ when cell $j$ is out of the kernel coverage when we position the kernel to be centered at cell $i$. In other words, cell $j$ will receive a non-zero weight only when it is close enough to the target cell $i$ so that it could be covered within the kernel.

To see the connection between SLX and the CNN more clearly, we can re-write the

equation above as a spatial cross-regressive model with some nonlinearities. That is,

$$\begin{aligned}\boldsymbol{y}_t &= \text{relu}\left(\beta_0\mathbf{1} + \boldsymbol{W}^{\boldsymbol{f}} \circ \boldsymbol{X}_t\boldsymbol{\beta}\right) \\ &= \text{relu}\left(\beta_0\mathbf{1} + \boldsymbol{X}_t\boldsymbol{\beta} + (\boldsymbol{W}^{\boldsymbol{f}} - \boldsymbol{I}) \circ \boldsymbol{X}_t\boldsymbol{\beta}\right),\end{aligned}$$

where $\boldsymbol{W}^{\boldsymbol{f}} = [\boldsymbol{W}^1, \cdots, \boldsymbol{W}^K]$, which is a 3-d tensor of size $K \times N \times N$ that represents a collection of $K$ kernels for filter $f$, one for each covariate/channel. relu is applied element-wise to inputs of size $N$, and $(\boldsymbol{W}^{\boldsymbol{f}} - \boldsymbol{I}) \circ \boldsymbol{X}_t$ is defined as

$$(\boldsymbol{W}^{\boldsymbol{f}} - \boldsymbol{I}) \circ \boldsymbol{X}_t = \left[(\boldsymbol{W}^1 - \boldsymbol{I})\boldsymbol{X}_t\boldsymbol{e}_1, (\boldsymbol{W}^2 - \boldsymbol{I})\boldsymbol{X}_t\boldsymbol{e}_2, \ldots, (\boldsymbol{W}^K - \boldsymbol{I})\boldsymbol{X}_t\boldsymbol{e}_K\right],$$

where $\boldsymbol{e}_k$ is a one-hot vector that has value 1 in its $k$-th entry and 0 anywhere else. Therefore, $\boldsymbol{X}_t\boldsymbol{e}_k$ simply extracts the $k$-th column of $\boldsymbol{X}_t$, which is the collection of the $k$-th covariate for all $N$ locations. As demonstrated in Figure 9, we use the $k$-th kernel $\boldsymbol{W}^k$ to compute the spatial lags for the $k$-th explanatory variable. We deduct the identity matrix to separate out the original covariates and the spatial lags. For each weight matrix $\boldsymbol{W}^k$, although its size is $N \times N$ in this representation, its degree of freedom is in fact the kernel size $k \times k$, because the weights are shared across space.

Clearly, the convolutional filter is in essence an advanced version of SLX. First, compared with a fixed weight matrix in SLX, a convolutional filter allows for different spatial weights for different covariates. Second, weights are shared across space in the filter, which greatly reduces the degree of freedom from $N \times N$ to $k \times k$, and hence saves a lot of computational costs. Third, in SLX the weights are determined ex ante, while in the convolutional filter the weights are learned from the data, jointly with other parameters. Additionally, a convolutional layer is much better than SLX in terms of the model expressivity. On one hand, a nonlinear component, namely the relu, is added to the filter. On the other hand, a convolutional layer could have several independently working filters, which could capture different characteristics of the original data and jointly work for the estimation.

## 3.7 Wildfire forecast model

In the previous two subsections, we explained how 1-d and 2-d CNN models can capture the temporal and spatial dependence respectively. Intuitively, for a spatiotemporal prediction task, we will need a 3-d model to perform convolution on the spatiotemporal lags. In this subsection, we will describe the setup details of our spatiotemporal wildfire forecast model.

One main challenge in this forecast problem is how to make annual predictions based on daily meteorological data. Because most homeowner insurance contracts are annual, to

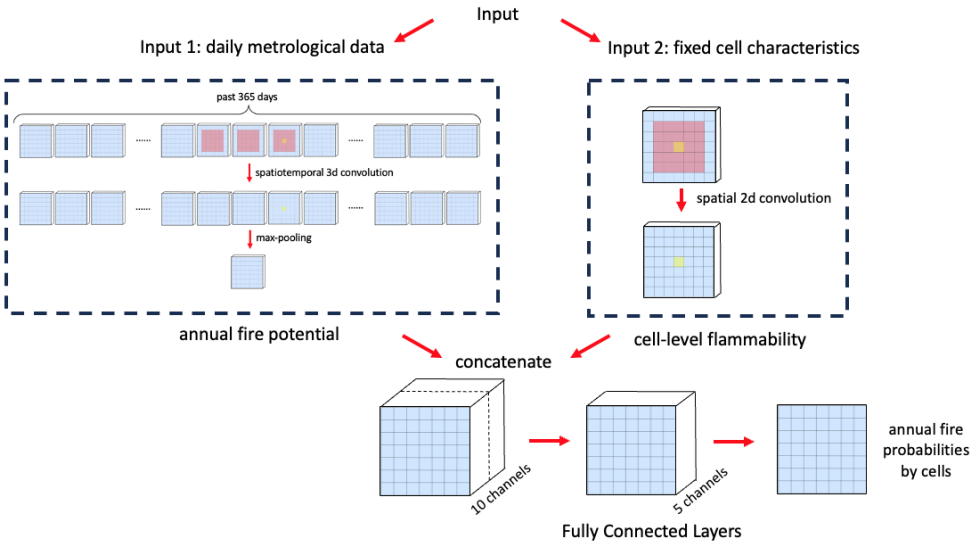



Figure 10: **Spatiotemporal CNN: Branching structure for dynamic and static measures:** Input 1: includes the daily meteorological measures at each cell; Input 2 includes the fixed cell features such as topography and vegetative coverage

price them on a daily basis, on each day we have to dynamically forecast the fire probability of the next 365 days. To solve this problem, we aggregate the fire occurrence over the next 365 days, and choose the daily meteorological data of the previous 365 days $x_{it-364:t}$ as the predictors for location $i$. We define our prediction target as below.

$$y_{it} = \begin{cases} 1 & \text{if a wildfire occurs in location } j \text{ between day } t+1 \text{ and } t+365, \\ 0 & \text{otherwise.} \end{cases}$$

In terms of the predictors, as Figure 10 shows, we divide them into a time-varying group and a time-invariant group. The time-varying group includes all daily meteorological features, such as the max temperature, relative humidity and wind events, based on which we could compute the fire potential dynamically for each location. On the other hand, the time-invariant group consists of fixed effects that are associated with the general flammability of each location, including its topological features, vegetation conditions, electrical lines, utility providers, etc. We will discuss these predictors in detail in Section 4 below. Generally, this branching structure reflects the fact that the wildfire occurrence is the joint effects of dynamic weather conditions and static location flammability.

We apply CNN of different dimensions to the two predictor groups. For each day $t$, the daily meteorological features from the past 365 days are 3-d panel data of dimension $365 \times N \times K_1$, where $N$ is the number of locations and $K_1$ is the number of time-varying predictors.

In comparison, the static features are simple cross-sectional data of dimension $N \times K_2$, where $K_2$ is the number of time-invariant predictors. To account for both spatial and temporal dependence, a spatiotemporal 3-d CNN is designed for the dynamic daily meteorological data. Figure 11 plots an example of the 3-d CNN structure. For each location-day, the figure describes how we use the weather data of the past few days from its surrounding cells, including itself, to forecast the fire potential. For one filter in the 3-d convolutional layer, we estimate

$$z_{it} = \text{relu}\left(\beta_0 + \sum_{k=1}^{K_1} \beta_k \sum_{l=1}^{p} \sum_{j=1}^{N} w_{lij}^k x_{it-p+l}^k\right),$$

where $z_{it}$ is the fire potential of cell $i$ at day $t$, $p$ is the size of temporal convolution, and $w_{lij}^k$ is location $j$'s kernel weight on location $i$, at day $t-p+l$, for the $k$-th channel. Note that $w_{lij} \neq 0$ only when location $j$ is within the spatial convolution range for location $i$. For time-invariant features, we use the common 2-d CNN to measure the general flammability of cell $i$,

$$m_i = \text{relu}\left(\alpha_0 + \sum_{k=K_1+1}^{K_2} \beta_k \sum_{j=1}^{N} w_{ij}^k x_{it}^k\right).$$

We use the max pooling structure to obtain annual fire potential measures. The idea of max pooling is simply taking the maximum value within a given time or spatial window. In our model, we apply a time window version of max-pooling to each location. As Figure 12 shows, after we perform the 3-d CNN and obtain daily measures of fire potential, we use the max pooling algorithm to get the maximum value from the previous 365 days, which we believe should be predictive for the wildfire occurrence of the next 365 days. One major advantage of this method is that we do not need to compute any statistics to aggregate those daily meteorological data into annual measures. Specifically, if we were to use a logistic regression, we would have to first calculate the mean, quantile or extreme values of the daily temperature, humidity, wind, etc., and then perform the annual forecast. This might lead to over-estimation because the warmest day may not happen to be the day with strong wind events. In comparison, with this CNN structure we work directly on the daily data, compute daily fire potentials and then take the maximum value, which perfectly avoid the mismatch problem as in the logistic regression.

The max pooling helps to reduce the time dimension of time-varying features from 365 to 1. This means we could now easily concatenate measures of the annual fire potential and the time-invariant features, as we show in Figure 10. Then we add several fully connected layers, which are simply linear functions with relu, to allow different features to interact with each other. We use the sigmoid function in the last layer to produce the annual fire probability for each location. In sum, we could summarize our model setup as follows: for

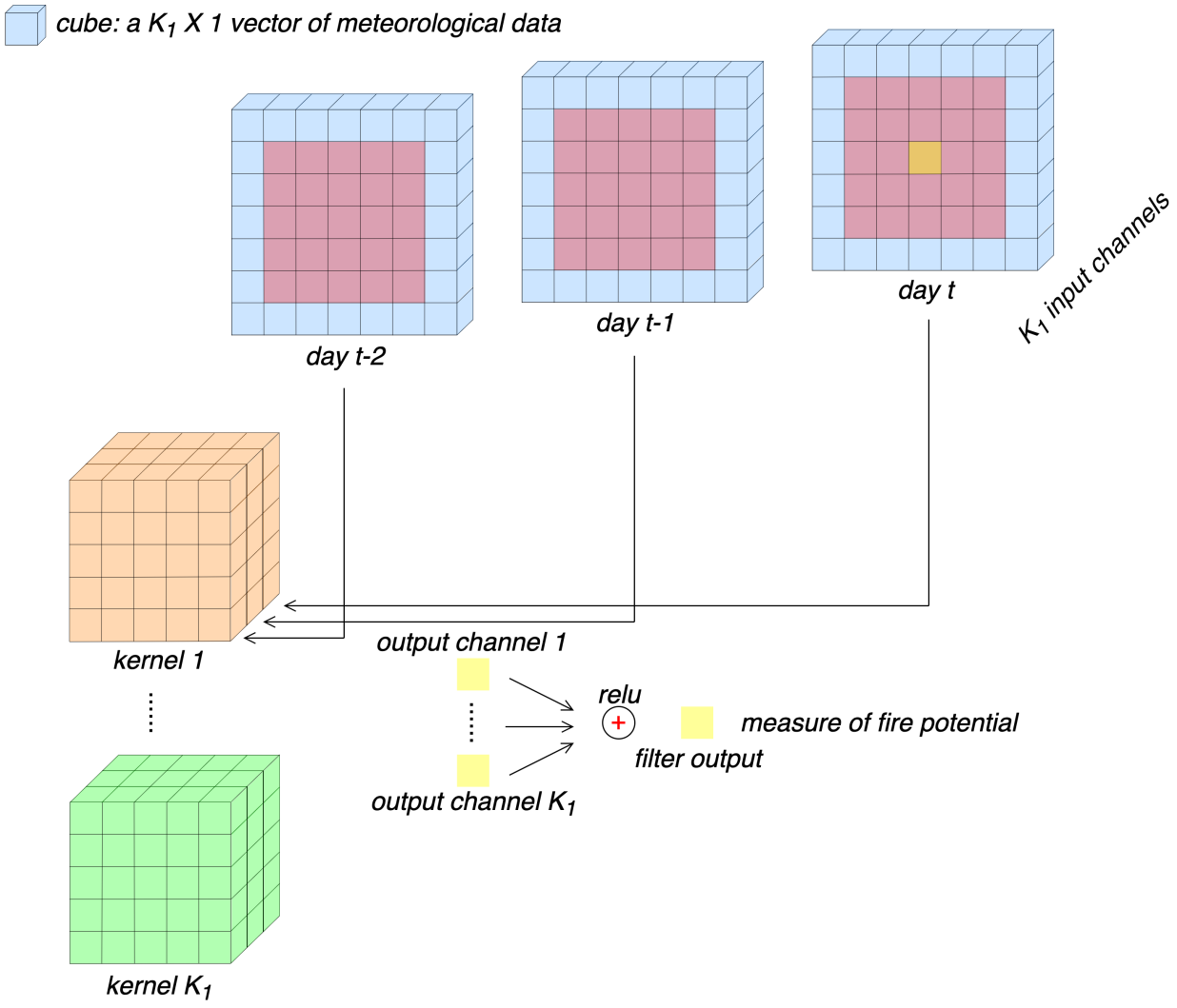



Figure 11: **Example of a filter in a 3d convolutional layer with kernel size (3,5)**: 3 is the temporal convolution size (i.e. the last three 3 days) and 5 is the spatial convolution size (i.e. the surrounding 5×5 cells). This kernel size is arbitrarily chosen as an example. Each cube represents a $K_1 \times 1$ vector of meteorological data, corresponding to $K_1$ input channels. Each channel is associated with a particular 3d kernel of dimension $(3, 5, 5)$. The yellow cube is our target location-time. Its surrounding red cubes plus the yellow one itself are involved in the 3d convolution, while the blue ones are out of the convolution range and irrelevant. We only show three days of data because other days are irrelevant. Just like the lower-dimensional cases, the convolution is performed by taking the inner products between the original data and the kernel, and the convolutional outputs of all channels are combined to produce the final output.

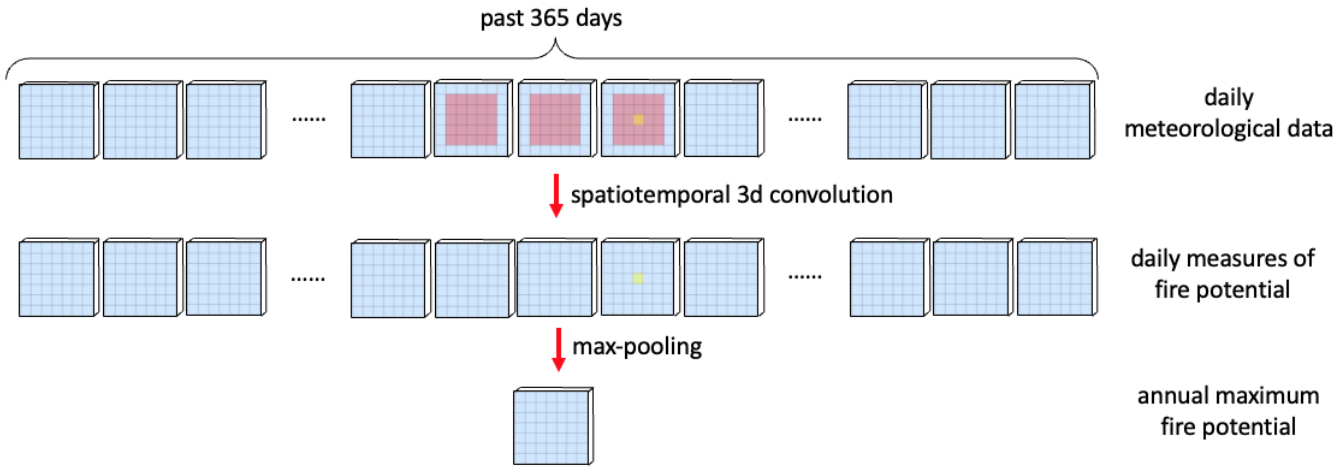



Figure 12: **Example of obtaining annual forecast by max-pooling** In this graph, we show how we obtain the annual fire potential for each location $i$ on day $t$. Each big blue cube represents a $K_1$-dimensional cross-sectional map on a specific day, while each small cube stands for a location. At the top layer, we collect the daily meteorological data from day $t-364$ to day $t$, and then apply the 3-d CNN to to obtain daily measures of fire potential. Then we apply max pooling, which takes the maximum value from the past 365 days for each location. The 365 maps of fire potential are aggregated into one map after max pooling.

$s = t - 365 + L, t - 364 + L, \cdots, t + 1 - L,$

$$
\begin{aligned}
z_{is} &= \text{relu}\left(\beta_0 + \sum_{k=1}^{K_1} \beta_k \sum_{l=1}^{p} \sum_{j=1}^{N} w_{lij}^k x_{is-p+l}^k\right), \\
z_{it}^a &= \max\left(z_{it-365+L}, z_{it-364+L}, \cdots, z_{it+1-L}\right), \\
m_i &= \text{relu}\left(\alpha_0 + \sum_{k=K_1+1}^{K_2} \beta_k \sum_{j=1}^{N} w_{ij}^k x_{it}^k\right),
\end{aligned}
$$

where $L = (p+1)/2$ and $p$ is the size of temporal convolution. For simplicity, here we assume we only use 1 filter in each layer. In practice, we have many filters in each layer, and the outputs are concatenated to be vectors with different channels. Then,

$$
\begin{aligned}
\boldsymbol{u}_{it} &= \text{relu}(\beta_0^u \mathbf{1} + \boldsymbol{\beta}_1^u \boldsymbol{z}_{it}^a + \boldsymbol{\beta}_2^u \boldsymbol{m}_i), \\
\boldsymbol{v}_{it} &= \text{relu}(\beta_0^v \mathbf{1} + \boldsymbol{\beta}^v \boldsymbol{u}_{it}), \\
\hat{y}_{it} &= \text{sigmoid}(\beta_0^y + (\boldsymbol{\beta}^y)' \boldsymbol{v}_{it}),
\end{aligned}
$$

where $\boldsymbol{z}_{it}^a$ and $\boldsymbol{m}_{it}$ are concatenated features that reflect annual fire potential and location flammability, $\boldsymbol{u}_{it}$ and $\boldsymbol{v}_{it}$ are results from the intermediate fully connected layers, and $\hat{y}_{it}$

is a scalar that predicts the wildfire probability. Note that $\boldsymbol{\beta}_1^u$, $\boldsymbol{\beta}_2^u$ and $\boldsymbol{\beta}^v$ are matrix that collects the coefficients for different channels, while $\boldsymbol{\beta}^y$ in the output layer is a vector.

Another challenge in this forecast problem is the severe data imbalance. Because wildfires are rare events, the majority of places in California did not burn. This means only a very tiny proportion of sample are *ones*, while we have massive number of *zeros*. To prevent the CNN from predicting no fire for all locations, we adopt a weighted cross entropy function as the loss function for training. Concretely,

$$\mathcal{L} = -\frac{1}{|I||T|}\sum_{i\in I}\sum_{t\in T}\gamma_0(1-y_i)\log(1-\hat{y}_i) + \gamma_1 y_i \log(\hat{y}_i),$$

where

$$\gamma_0 = \frac{\text{Total}_{\text{train}}}{2\cdot \text{Neg}_{\text{train}}},$$
$$\gamma_1 = \frac{\text{Total}_{\text{train}}}{2\cdot \text{Pos}_{\text{train}}},$$

and where $\text{Total}_{\text{train}} = |I||T_{\text{train}}|$ is the total number of observations in the training sample, $\text{Neg}_{\text{train}}$ is the number of negative cases (0, no fire), and $\text{Pos}_{\text{train}}$ is the number of positive cases (+1, fire). In our case, $\gamma_1$ will be much larger than $\gamma_0$, so that the classifier can heavily weight the few fires in our sample.

# 4 Data

The three key environmental conditions that determine vegetative wildfire behavior after ignition events are topography, the nature of fuel availability, and climatic conditions. CAL FIRE data includes information on the alarm date, the precise location of the wildfire, when available the ignition cause, and the wildfire perimeter for each incident in California from 2000 through 2022. We follow Abatzoglou (2013) and carry out all of our wildfire prediction modeling using 2 $\times$ 2 kilometer grid-cells projected over the entire state of California. We pre-process the dataset by correcting several typos in the fire dates and dropping any fire whose alarm date is missing or after its containment date. We then select all fires whose alarm dates are between 01/01/2000 and 12/31/2021, and merge them with our geocoded 2 $\times$ 2 kilometer cell network. Overall in our data, we have 7,163 wildfires in total and 42,659 fire grid-cells.

| Measures for each grid cell | Specific features |
|---|---|
| **1. Topography** | |
| http://apps.nationalmap.gov/downloader/ | |
| | Elevation |
| | Aspect |
| | Slope |
| **2. Large utility districts** | |
| https://hub.arcgis.com/datasets/CalEMA::california-electric-utility-service-territory/about | |
| | Gas & Electric |
| | Southern California Edison |
| | Pacific Corporation |
| **3. Transmission lines** | |
| https://www.arcgis.com/home/item.html?id=d4090758322c4d32a4cd002ffaa0aa12 | |
| | Count of transmission lines |
| **4. Daily meteorology** | |
| https://www.climatologylab.org/gridmet.html. | |
| | Maximum air temperature |
| | Specific humidity |
| **5. Hourly meteorology** | |
| https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=overview | |
| | Wind speed |
| | Wind direction |
| | Humidity |
| **6. Dry lightning** | |
| www.ncei.noaa.gov/pub/data/swdi/database-csv/v2/ | |
| | Dry lightning counts |
| **7. Vegetative types** | |
| https://gis.data.ca.gov/maps/CALFIRE-Forestry::california-vegetation-whrtype/about | |
| | Indicators for thirteen vegetative types |
| **8. Vegetative canopy** | |
| https://lpdaac.usgs.gov/products/mod44bv061/ | |
| | Percent covered by tree canopy |
| | Percent covered by non-tree canopy |

Table 1: **Table of grid-cell features and their data sources:** The table reports the eight classes of grid-cell measurement for topography, utility district service provider, transmission line counts, hourly meteorologic data, daily meteorologic data, vegetative types and tree/nontree canopy grid-cell percentage coverage. For each measurement class we also report the download information for the data sources as well as the specific features that we are using from each source.

## 4.1 Topography

Topography is a key factor in vegetative wildfire behavior. It influences the spatial variability of fuels and the biophysical conditions that determine wildfire ignition, the direction of spread, the intensity and the duration of wildfire. In California, topography and air pressure systems play a key role in the direction and speed of the hot dry northwesterly flowing Diablo winds of Northern California and the Santa Ana winds of Southern California. The steepness and southwest aspect of the slopes of these ranges and the high elevation of their ridge tops all induce relatively drier vegetative fuel conditions especially in the high ambient temperature months of July through October. Topography also has an impact on a variety of other features of fire behavior such as fire-line width, flame length, and the direction of spread. Another important aspect of fire behavior that is affected by topography is the rate of spread since many fires accelerate dramatically uphill, thus placing fire fighters, reservoir access, and utility infrastructure at risk (McClung and Mass, 2007).

As shown in Table 1, our slope, elevation and aspect measure are computed using topographical raster data from the U.S. Geological Services and geoprocessing this information using QGIS software to compute slope and aspect. We conduct the cosine transformation on the aspect, to measure how close the direction of the is to a southwestern orientation (i.e. the 225 degrees direction) (Kumar et al., 1997).

## 4.2 Utility districts and transmission lines

California utilities have struggled with wildfire related liability associated with inadequate vegetation management around their transmission lines, deferred maintenance of their transmission power pylons, and catastrophic wildfire ignition events associated with damage to lines and pylons within their service areas. California Assembly Bill 1054, passed in July, 2019, funded a $5 billion fund for utility wildfire safety investments that required utilities to file Wildfire Mitigation Plans with the California Public Utilities Commission in exchange for access to the mitigation plan funds for investment reimbursement. There are currently three utilities that participate in the California Wildfire Fund – San Diego Gas & Electric Company, Southern California Edison, and Pacific Gas & Electric Company. However, there remains considerable controversy concerning the most cost effective utility mitigation strategies: undergrounding of lines or covered conductors.[17]

Again, following Table 1, we focus on the three largest utilities: Pacific Gas & Electric, Southern California Edison, and Pacific Corporation. Two of these utilities currently par-

[17]See "This utility's undergrounding plan is causing sticker shock" by Wes Venteicher and Blanca Begert, *Politico*, October 5, 2023. https://www.politico.com/newsletters/california-climate/2023/10/05/pg-es-undergrounding-plan-is-causing-sticker-shock-00120290.

ticipate in the California Wildfire Fund under AB1054 and the other, Pacific Corporation, does not. We download the map of California Electric Utility Service Territory[18] and assign the utility provider to each of our 2km × 2km cells. In addition, we obtain the map of U.S. Electric Power Transmission Lines.[19] For each cell, we use ArcGIS to calculate the distance from its center to the nearest electrical lines.

- **Pacific Gas & Electric (PG&E)** is the largest investor-owned California utility and serves approximately 16 million people throughout a 70,000-square-mile service area in northern and central California.[20] Five of the 10 most destructive fires in California since 2015 have been linked to PG&E's electrical network. Regulators have found that in many fires, PG&E violated state law or could have done more to make its equipment safer.[21]
- **Southern California Edison (SCE)** is one of the nation's largest electric utilities and provide electric service to approximately 15 million people through 5 million customer accounts. SCE's service area includes portions of 15 counties and hundreds of cities and communities in a 50,000-square-mile service area within Central, Coastal and Southern California.[22] Public Utilities Commission investigators found SCE liable for damages from SCE power lines associated that ignited the 2017 Thomas fire burning more than 280,000 acres, damaging more than 1,300 structures and causing two deaths in Santa Barbara and Ventura counties. Another Public Utilities Commission again found liability related to Southern California Edison equipment that was likely "associated" with 2018's deadly Woolsey fire, which burned more than 1,600 structures in Los Angeles and Ventura counties and killed three people.[23]
- **Pacific Corporation** is the largest transmission-line grid operator in the West, with a service area of 141,500 square miles including parts of Oregon, Washington, California, Utah, Idaho and Wyoming, and has 2.1 million customers.[24] S&P Global reported in August 2024 that Pacific Corporation faces at least $46 billion in claims related to Western US wildfires following recent lawsuits in Oregon for fires in Oregon and California.[25]

[18] https://hub.arcgis.com/datasets/CalEMA::california-electric-utility-service-territory/explore

[19] https://www.arcgis.com/home/item.html?id=d4090758322c4d32a4cd002ffaa0aa12

[20] https://www.pge.com/en/about/company-information/company-profile.html.

[21] https://www.nytimes.com/interactive/2019/03/18/business/pge-california-wildfires.html.

[22] https://download.newsroom.edison.com/create_memory_file/?f_id=5cc32d492cfac24d21aecf4c&content_verified=True.

[23] https://www.latimes.com/california/story/2021-12-17/southern-california-edison-faces-550m-penalty-for-wildfires.

[24] https://www.pacificorp.com/about.html.

[25] https://www.spglobal.com/commodityinsights/en/market-insights/latest-news/electric-

Given the risks associated with density of California transmission lines and transmission power pylons, we also measure the underlying vegetative ground cover due to their differing susceptibility to combustion events when temperatures, wind, and relative humidity sufficiently lower moisture levels. To measure these risks, as shown in Table 1, we count the number of transmission lines that cross each grid-cell and then take the log of that count as a measure to risks of failures of the lines and pylons themselves.

## 4.3 Meteorology

### 4.3.1 Daily measures

Following Table 1, our daily measure of climate data are obtained from gridMET, which is a publicly available dataset of high-spatial resolution (4km×4km) surface meteorological data covering the contiguous US from 1979 till yesterday (see Chegini et al., 2021). We then use interpolation to map from the (4km×4km) grid cells to a denser 2km×2km spatial coverage of the state. gridMET data includes 15 meteorological variables in total, and is updated daily. As shown in the Appendix, Figure 8, given the high correlations between maximum air temperature (tmmx) and the gridMET measures for fuel moisture over 100 hours (fm100), fuel moisture over 1000 hours (fm100), reference evapotranspiration (etr), minimum air temperature (tmmn), vapor pressure deficit (vpd), and surface radiation (srad), we focus on two key climate measures found in gridMET: maximum air temperature and specific humidity.

### 4.3.2 Hourly measures

In California, there are two types of fire-associated wind: the Diablo Winds of northern California and the Santa Ana winds of southern California. Santa Ana winds have been the driving force behind many of southern California's most devastating fires (see Billmire et al., 2014; Jin et al., 2013; Kochanski et al., 2013), more recently the Diablo winds of northern California have become more dangerously associate with wildfire occurrence with their similarly low relative humidity, high temperatures, and very high wind speeds (see Bowers, 2018; Diaz, 2022; Keeley and Syphard, 2018; Linn et al., 2020; Liu, 2022; Liu et al., 2021). Although both of these winds are shaped by atmospheric conditions and are driven by the topography of region, they also have important differences.

The Diablo Winds (see Abatzoglou et al., 2018, 2021; Diaz, 2022; MacDonald et al., 2023) are strong northeasterly winds that flow over the western slopes of the Sierra Nevada

power/080524-wildfire-claims-against-pacificorp-surge-to-46b-on-oregon-mass-complaints.

range in eastern California where they heat up and lose humidity before passing through the California central valley. High pressure systems in the central valley then drive the Diablo winds over the California coastal range where the compression and the loss of moisture produces intense, dry, downslope winds. In contrast, the Santa Ana winds of southern California are gravity driven winds that occur when high pressure builds over the Great Basin to the West of the Sierra Nevada mountains — the Great Basin includes most of Nevada, half of Utah, and sections of Idaho, Wyoming, Oregon, and California — and low pressure systems develop over the California coast (see Cardil et al., 2021; Gershunov et al., 2021; Guzman-Morales, 2018; Keeley et al., 2021). The cold air from the Great Basin then sinks and the dry air from the desert is pushed toward the low lying coastal areas through the Sierra and coastal mountain canyons and where compression causes the winds to warm by tens of degrees Fahrenheit per mile as it travels.

Hourly climate measurement is required to measure the occurance of Diablo and Santa Ana winds. Diablo winds are characterized by their intensity, long duration, low moisture content, and northeasterly direction. Following Bowers (2018) and Diaz (2022), we identify the occurance of Diablo winds if the winds are northeasterly in direction, have speeds exceeding eight meters per second, have relative humidity that is below 25%, and have a duration of at least six hours. Following Guzman-Morales (2018), the Santa Ana winds are defined as northeasterly winds, with wind speeds of over 30 miles per hour (mph), relative humidity of below 10% (or dropping to single digits), and average durations of at least 12 hours of continuous wind speed. Following Table 1, our hourly climate measurement data are obtained from ERA5-Land, published by ECMWF (The European Centre for Medium-Range Weather Forecasts).[26] We identify wind events using hourly two meter dewpoint temperature, two meter temperature, ten meter directional components (u-component and v-components).

## 4.4 Dry lightning

Dry lightning, occurring when there has been less than 2.5 mm of rainfall, is a major source of wildfire ignition in central and northern California (see Kalashnikov et al., 2022). While human-caused wildfire ignitions predominate in southern California, lightning-caused fires are more common in the northern half of the state, particularly over mountainous terrain (see Balch et al., 2017; Brey et al., 2018; Chen and Jin, 2022; Keeley and Syphard, 2018). Summertime lightning outbreaks in northern California from July through August, unlike the predominantly human-caused fires that originate in a single location, can strike multiple

[26] See https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=overview.

locations and start numerous simultaneous wildfires (Miller et al., 2012). Widespread thunderstorms with dry lightning have produced some of the largest and longest-lasting wildfires in recent decades including the 1987 wildfire season (see Duclos et al., 1990), the 2008 wildfire season (Wallmann et al., 2010), and the disasterous 2020 wildfire season (Keeley and Syphard, 2021). Based on our CAL FIRE ignition-cause data we find that 20% of wildfires in northern California between 2003 and 2022 were caused by dry lightning.

## 4.5 Vegetative types

Vegetation has significant effects on wildfire behavior and is thus an important focus of wildfire prediction modeling (see Kearns et al., 2022; Price and Bradstock, 2014). Topographic features such as elevation, aspect, latitude, and slope also influence microclimatic conditions, such as temperature, precipitation, direct solar radiation, wind exposure, among others, which together influence the moisture content of fuel (Flannigan et al., 2016, 2009; Westerling, 2014). As noted above, topography can also affect ignition probabilities because steep slopes, ridge tops, and southwest facing slopes are all characterized by drier fuel conditions. Other important vegetative features that effect wildfire ignition probabilities and behavior include the spread of invasive non-native or non-conifer species (Brooks and Matchett, 2006; Calhoun et al., 2022; Holmes et al., 2008) and the degree of nearby urbanization (see Alexandre et al., 2016; Kestelman, 2024; Price and Bradstock, 2014).

As shown in Table 1 we focus on thirteen types of vegetative exposure for each grid cell: Agriculture, barren/other, conifer forest, conifer woodland, desert, desert shrub, hardwood forest, hardwood woodland, herbaceous, shrub, urban, water, and wetland following the California Department of Agriculture vegetative indexes discussed in Table 1. Sub Figure 13a presents the geographic locations of the vegetative ground coverage types. As shown, conifer woodlands/forests dominate the Sierra Nevada mountain range and the northern section of the coastal range. The hardwood forest and hardwood woodlands are found in the foothills of the Sierra Nevada and along the coastal range north of San Francisco. Herbaceous, shrub, and woodlands are the dominant vegetative ground covers in the coastal range areas around San Francisco and extend south to Santa Barbara County. The Los Angeles Basin including San Diego is dominated by herbaceous, desert shrub, and desert woodland with small areas of conifer forest at higher elevations. The central valley of California is dominated by irrigated agricultural vegetation. As shown in Subfigure 13c of Figure 13, the historical locations of wildfires in California are found in the western facing slopes of the Sierra Nevada and along coastal range in both northern and southern California. The northern California wildfires are dominated by hardwood/conifer woodland and forest vegetation. The southern

California wildfires have primarily occurred in areas dominated by herbaceous, desert/conifer woodland, and shrub vegetation. The irrigated areas of the central valley have the lowest historical wildfire incidence.

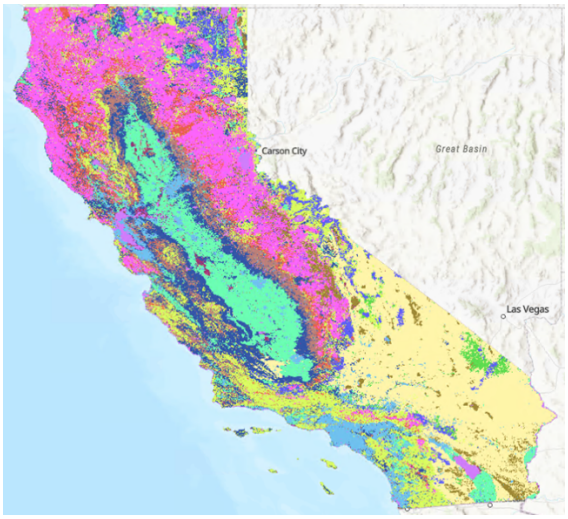
### 4.6 Vegetative canopy density

Finally, as shown in Table 1, in addition to our measures for the vegetative types found within grid cells, we also account for the canopy coverage of the vegetation within the cells. Dense canopy coverage is importantly associated with the effects of incoming solar energy and the live fuel content of the vegetation. The live fuel moisture content is a measure of the water content of live fresh foliage relative to its dry mass (Yebra et al., 2018) and it is an important determinant of the potential for fire ignitions to propagation. California has significant heterogeneity in its canopy coverage and the types of vegetation associated with the coverage. Northern California has dense canopies of conifer and hardwood in the foothills of the Sierra Nevada and the northern coastal range that are significantly prone to wildfire (see Chen et al., 2021). The shrubland canopies of the southern California chaparral areas from Monterey County south to the Los Angeles Basin and San Diego pose equally severe threat of wildfire occurrence and rapid propagation (see Dennison and Moritz, 2009; Dennison et al., 2008).

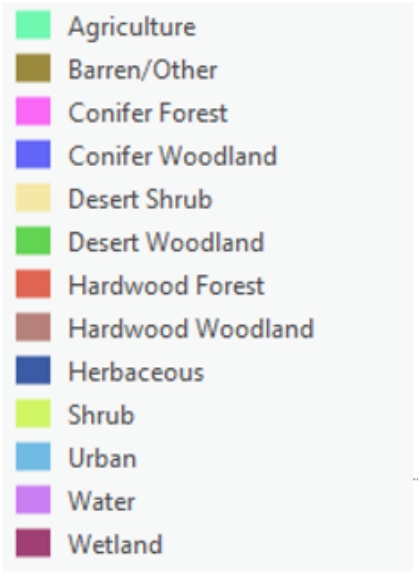
We obtain two measures of canopy from the Moderate Resolution Imaging Spectroradiometer's (MODIS) Vegetation Continuous Fields database from the Geological Survey Land Process Distributed Active Archive Center. The first measure is the "Percent covered by tree canopy," defined as the canopy coverage by woody plants that are greater than or equal to five meters tall (see Chen et al., 2021; DiMiceli et al., 2021). The second measure is the "Percent covered by non-tree canopy," defined as the canopy coverage associated with small trees (less than 2.5 meters), grass, or shrubs (see Lai et al., 2022; Mallinis et al., 2019).

## 5 Wildfire forecasting: logistic regression vs. CNN

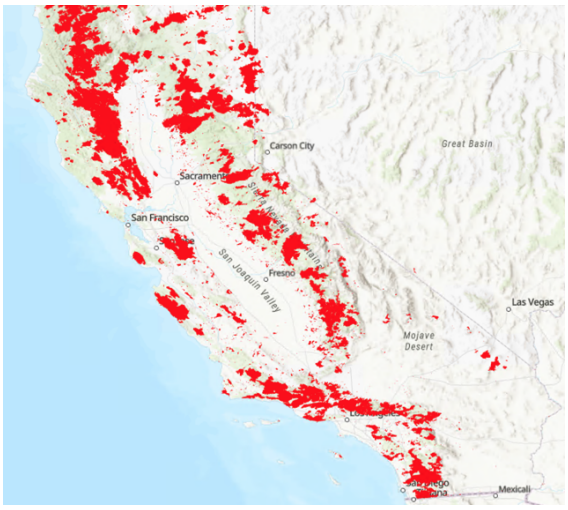
As discussed in Section 3, logistic generalized additive models been a common approach to the modeling of wildfire occurrence in the climate literature (Xi et al., 2019) and logistic regression is currently the common methodology that is applied in recent analyses of wildfire incidence in the economics literature (see Biswas et al., 2023; Kahn et al., 2024).




(a) Geographic location of vegetative types in California, 2021




(b) Vegetative types 2021




(c) Historical location of wildfire incidents in California (2000–2021)

Figure 13: **Relationship between vegetative distributions and wildfire incidents** The vegetative coverage data are sourced from CAL FIRE Forestry (https://gis.data.ca.gov/maps/CALFIRE-Forestry::california-vegetation-whrtype/about). The wildfire incidence and burn area data are sourced from California Department of Forestry and Fire Protection (CAL FIRE) (https://www.fire.ca.gov/what-we-do/fire-resource-assessment-program/fire-perimeters).

| | coef | std err | z | P> \|z\| |
|---|---|---|---|---|
| Constant | -5.7238 | 0.011 | -533.151 | 0.000 |
| Specific humidity (q5: May–Oct) | -0.0423 | 0.002 | -18.298 | 0.000 |
| Maximum air temperature (q95: May–Oct) | 0.2281 | 0.003 | 78.685 | 0.000 |
| Indicator: Diablo or Santa Ana wind events | 0.1527 | 0.010 | 15.682 | 0.000 |
| Indicator: dry lightning | 0.0667 | 0.006 | 11.578 | 0.000 |
| Transmision line count | 0.0592 | 0.006 | 10.332 | 0.000 |
| Transmision line count × PG & E Indicator | 0.0444 | 0.006 | 7.108 | 0.000 |
| Transmision line count × Southern California Edison Indicator | 0.1602 | 0.007 | 23.675 | 0.000 |
| Transmision line count × Pacific Corporation Indicator | -0.3480 | 0.012 | -28.193 | 0.000 |
| Grid-cell percentage tree canopy | 0.1782 | 0.004 | 50.321 | 0.000 |
| Grid-cell percentage non-tree canopy | 0.3837 | 0.004 | 97.561 | 0.000 |
| Slope | 0.4568 | 0.004 | 113.477 | 0.000 |
| Aspect | 0.0047 | 0.002 | 2.496 | 0.013 |
| Barren/other | 0.8927 | 0.021 | 41.851 | 0.000 |
| Conifer forest | 1.6291 | 0.013 | 125.272 | 0.000 |
| Conifer woodland | 1.1832 | 0.016 | 72.456 | 0.000 |
| Desert shrub | 0.0236 | 0.017 | 1.395 | 0.163 |
| Desert woodland | 0.6652 | 0.034 | 19.407 | 0.000 |
| Hardwood forest | 1.7103 | 0.013 | 129.134 | 0.000 |
| Harwood woodland | 1.3460 | 0.012 | 107.779 | 0.000 |
| Herbaceous | 1.5020 | 0.011 | 131.827 | 0.000 |
| Shrub | 2.0482 | 0.011 | 181.352 | 0.000 |
| Urban | 1.2230 | 0.014 | 87.214 | 0.000 |
| Water | 1.5557 | 0.021 | 74.361 | 0.000 |
| Wetland | 0.8069 | 0.028 | 28.361 | 0.000 |
| No. Observations: | 20,483,064 | | | |
| Pseudo R-squ.: | 0.07144 | | | |
| Log-Likelihood: | -1.5974e+06 | | | |
| LL-Null: | -1.7203e+06 | | | |
| LLR p-value: | 0.000 | | | |

Table 2: **Logistic regression results.** The table reports a logit regression of annual grid-cell incidence of wildfire using data from 2000 to 2021. incidence at a grid-cell at month t

## 5.1 Logistic regression

Because our goal is the annual prediction of wildfire occurrence and given the heavy tailed data distributions discussed in Section 1, we use annual aggregates measured at the $5^{th}$ quantile for specific humidity and at the $5^{th}$ quantile for maximum air temperature over the fire season months of May through October. For each grid cell, we construct a daily indicator variable for Diablo or Santa Ana wind exposure using grid-cell hourly measurements of the maximum annual wind speed, wind direction, and humidity that define these winds.

As shown in Table 2, all of the coefficient estimates are of the expected sign and nearly all are statistically significant at better than the .0001 level. The key meteorological features measured at the $5^{th}$ quantile for specific humidity and at the $5^{th}$ quantile for maximum air

temperature over the wildfire season of May through October are shown to have a statistically significant statistically negative coefficient for specific humidity and a statistically significant positive coefficient for maximum air temperature. Annual grid-cell wildfire occurrence is also shown to be statistically significantly positively associated with the indicator features for dry lightening and Diablo/Santa Ana winds. Additionally, consistent with recent wildfire events, higher counts of transmission lines within grid-cells and the interaction of higher counts with indicator variables for the location of those lines in either the PG&E and Southern California Edison service provision districts are also positively associated with higher levels of wildfire occurrence, whereas higher transmission line counts in the Pacific Corporation service area is negatively associated with wildfire incidence.

Consistent with the climate literature, the logistic regression results also show that the grid-cell percentage of both tree and small-tree/shrubland canopy coverage is positively associated with the occurrence of wildfire. Similarly the fixed-effects controls for other twelve vegetative types, other than desert shrub, are shown to be statistically significantly and positively associated with wildfire occurrence where the hold-out vegetative type is the irrigated agriculture. The two topographical measures for aspect and slope are also shown to be statistically significant and positively associated with the occurrence of wildfire consistent with the discussion found in Section 4.1.

Overall, the logistic regression results provide feature associations with the occurrence of wildfire that comport well with the climate, topographical, and vegetative coverage literature surveyed in Section 4 and are consistent with the cross-sectional econometric techniques that mostly characterize this literature as discussed in Section 3. Of course, these results also reflect the potential shortcomings of logistic regression methodologies including lack of controls for spatial and temporal correlations among the cell features, lack of controls for non-linearities in the association between features and the occurrence of wildfire, the likely need for a more saturated specification that accounts for all possible interactions between indicator variables and other continuous measured features, and, finally, the fact that logistic regressions are cross-sectional models.

## 5.2 Spatiotemporal CNN

Following the presentation of our spatiotemporal CNN for out-of-sample annual wildfire forecasts in Section 3, we re-analyze wildfire occurrence in California using spatiotemporal CNNs again using the meteorological, topographical, utility exposure, and vegetative features used in the logit regression reported in Table 2.

Since the California wildfire data is significantly imbalanced, where the unconditional

probability of wildfire at a given grid-cell is essentially zero and the number of grid-cell with wildfire occurrence is far outweighed by the number of grid-cells without an occurrence of wildfire, we first focus on the F1 scores, the harmonic mean of precision and recall. We divide the entire spatiotemporal dataset into the training, validation and the test sets, in the order of time. The training set covers the period between Dec 31, 2000 and Dec 31, 2018. For computational simplicity, we sample the last day of every month in the training set to train the 3-d CNN model, which includes altogether 217 days with 94,829 cells on each day. We use the day Dec 31, 2019 as the validation, and Dec 31, 2020 as the test. Note that for each day we are predicting one-year ahead, so the validation set corresponds to the wildfires in 2020, while the test set corresponds to 2021. In addition, we carefully divide the dataset to avoid any looking-ahead problem. The last day in the training set is Dec 31, 2018, which is one year before the validation day. We also leave an one year gap between validation and test days.

Table 3 reports the F1s for the training, validation, and out-of-sample test data for our annual spatiotemporal CNNs estimates of annual wildfire occurrence with differing hyperparameter structures. The rows and columns of Table 3 represent differing kernel sizes for the 3d convolutional layer, where, for example, row p=3 corresponds to the temporal convolution size of the last 3 days and the column k=5 corresponds to the spatial convolution size of the surrounding 5×5 cells. Thus, for the p=3 row and the k=5 column in Table 3, each of the F1s reported for the training, validation, and out-of-sample CNN are associated with a channel structure comprised of 3d kernels of dimension $(3, 5)$. Table 3 presents the F1 results for a range of hyperparameter structures from p=0 to p=30 for the temporal convolution sizes and k=1 to k=9 for the spatial convolution sizes.

As expected, given the challenges of imbalanced classification and the underrepresentation of the wildfire occurrence in the training data, the F1s for the training estimates are uniformly low, ranging from 0.0523 for the 3d kernels of dimension $(30, 7)$ to a high of 0.0622 for the 3d kernels of dimension $(3, 7)$. The highest F1 scores reported in Table 3 are obtained for both the validation data and the out-of-sample test data for the 3d kernels of dimension $(3, 5)$. As shown, the F1 is 0.2084 for the validation data and the F1 is 0.1485 test data, indicating that the spatiotemporal CNN performs only modestly well in forecasting the annual occurrence of wildfire while minimizing false positive forecasts.

Table 4 allows for a more nuanced interpretation of the differing hyperparameter structures and the F1 statistics by comparing a range of performance statistics for both the validation and out-of-sample test data and comparing the added nonlinearity of the spatiotemporal CNNS to the benchmark logistic regression forecasting model. As shown in Table 4, the CNNs uniformly outperform the logistic regression in terms of F1, Brier scores,

| Training | k=1 | k=3 | k=5 | k=7 | k=9 |
|---|---|---|---|---|---|
| p=1 | 0.0556 | 0.056 | 0.0581 | 0.0594 | 0.0549 |
| p=3 | 0.0568 | 0.0609 | 0.0573 | 0.0622 | 0.0618 |
| p=7 | 0.0560 | 0.0628 | 0.0620 | 0.0719 | 0.0643 |
| p=30 | 0.0556 | 0.0687 | 0.0690 | 0.0523 | 0.0605 |
| Validation | k=1 | k=3 | k=5 | k=7 | k=9 |
| p=1 | 0.1962 | 0.1994 | 0.2002 | 0.2061 | 0.1934 |
| p=3 | 0.2021 | 0.2170 | 0.2084 | 0.2084 | 0.2008 |
| p=7 | 0.1968 | 0.1917 | 0.1999 | 0.2269 | 0.2017 |
| p=30 | 0.1996 | 0.2002 | 0.2163 | 0.1814 | 0.1997 |
| Test | k=1 | k=3 | k=5 | k=7 | k=9 |
| p=1 | 0.1262 | 0.1243 | 0.1283 | 0.1279 | 0.1326 |
| p=3 | 0.126 | 0.1354 | 0.1485 | 0.1322 | 0.1501 |
| p=7 | 0.1301 | 0.1289 | 0.1383 | 0.1455 | 0.1265 |
| p=30 | 0.1195 | 0.1295 | 0.1456 | 0.1322 | 0.1355 |

Table 3: **F1 performance** The F1 scores on training, validation and test samples with different CNN hyper-parameters

true positives, precision, recall, and the 0.5 auc value shows the logistic regression to be no better than randomly classifying annual wildfire occurrence for prediction purposes. Given the inherent randomness of wildfire occurrence, the relative performance of the preferred 3d kernels of dimension $(3, 5)$ for the spatiotemporal CNN can be interpreted as providing a moderately successful positive classification of annual wildfire with an auc of 0.7499, a precision of .0821 due to the preponderance of false positive predictions on the part of the CNN, and a relatively high recall of 0.7752. Of course, from the vantage point of insurance companies the worst case classification is false negatives, when wildfire and presumably losses occur, suggesting that the high recall statistic would be informative. Whereas for homeowners, the worst case scenarios is false positives, when the model predicts wildfire occurrence thus falsely increasing the rate of incorrect non-renewals of wildfire insurance policies, suggesting that precision may be the more informative statistic for homeowners.

Table 5 presents an analysis of the relative effects of shuffling model features on the performance metrics of the preferred spatiotemporal CNN with 3d kernels of dimension $(7, 7)$. For each metric, we show its percentage change when a feature is shuffled, relative to its original value. We recognize that the shuffling-based importance measures is somewhat limited in that it is comparable only for features with similar distribution, such as normally-distributed features. However, in our study and other climate studies, many features would

| hyperparameters (p,k) | validation | | | | test | | | |
|---|---|---|---|---|---|---|---|---|
| metric | logistic | (1,1) | (3,5) | (7,7) | logistic | (1,1) | (3,5) | (7,7) |
| cross entropy | 0.9968 | 0.6874 | 0.6719 | 0.5742 | 0.6505 | 0.6907 | 0.5922 | 0.6404 |
| Brier score | 0.0646 | 0.2525 | 0.2461 | 0.2088 | 0.0422 | 0.2548 | 0.2112 | 0.2359 |
| true positive | 0 | 5524 | 5613 | 5165 | 0 | 3363 | 3100 | 3344 |
| false positive | 0 | 44658 | 42127 | 34233 | 0 | 45922 | 34651 | 38638 |
| true negative | 88701 | 44043 | 46574 | 54468 | 90830 | 44908 | 56179 | 52192 |
| false negative | 6128 | 604 | 515 | 963 | 3999 | 636 | 899 | 655 |
| accuracy | 0.9354 | 0.5227 | 0.5503 | 0.6288 | 0.9578 | 0.5090 | 0.6251 | 0.5856 |
| precision | 0 | 0.1101 | 0.1176 | 0.1311 | 0 | 0.0682 | 0.0821 | 0.0797 |
| recall | 0 | 0.9014 | 0.916 | 0.8429 | 0 | 0.8410 | 0.7752 | 0.8362 |
| auc | 0.5 | 0.7473 | 0.7858 | 0.7861 | 0.5 | 0.6944 | 0.7499 | 0.7638 |
| prc | 0.0646 | 0.1261 | 0.1452 | 0.1540 | 0.0422 | 0.0660 | 0.0816 | 0.0991 |
| f1 score | 0 | 0.1962 | 0.2084 | 0.2269 | 0 | 0.1262 | 0.1485 | 0.1455 |

Table 4: **Comparing the logistic regression with CNN** Note that when (p,k) = (1,1), the model will be equivalent to a logistic regression that has additional nonlinearity and is trained with weighted loss function.

have highly skewed or imbalanced distributions. For example, the majority of the locations in California won't suffer from Diablo or Santa Ana events, implying the wind event indicator will mostly be zeros. Similarly, the dry lightening flashes will be zero for the majority of areas too. It is concerning to us that the classical random shuffling method will under-estimate the importance of these features with skewed distribution, because the shuffled values will remain the same for the majority area. To take this into account, we customize the shuffling a bit for those variables. Specifically, we flipped the values for dummies with imbalanced distribution. For skewed distribution with massive zeros, we assign zero to original non-zero values, and randomly assign non-zero values to original zeros. We understand that the feature importance may still not be comparable across all features, so we suggest compare the importance measures only within features with similar original distributions. In Table 5, we report the feature importance separately for the two different groups of features.

Similar to the logistic regression results, the slope and the indicators for dry lightening, and Diablo or Santa Ana winds, and as well as the vegetative type, all have large impacts on the CNN model performance metrics both in the validation and the test sample. Interestingly, maximum temperature has a greater effect on the CNN metrics for the test data than for the validation data. The features with lower overall impact on the spatiotemporal CNN include the utility provider, the number of transmission lines, aspect and the percentage of tree and non-tree canopy.

| **Validation** | | | | |
|---|---|---|---|---|
| feature | f1_score | precision | recall | loss |
| features with normal distribution | | | | |
| Vegetative type indicators | -19.56 | -22.35 | 4.59 | -4.03 |
| Daily specific humidity | -11.29 | -6.29 | -33.96 | 44.01 |
| Slope | -11.02 | -8.36 | -25.01 | 29.29 |
| Utility provider | -5.45 | -6.77 | 4.05 | -3.17 |
| Grid-cell percentage tree canopy | -2.21 | -2.47 | -0.56 | 2.85 |
| Daily maximum air temperature | -2.03 | -0.3 | -11.89 | 17.55 |
| Aspect | 0.02 | -0.04 | 0.39 | 0.29 |
| Grid-cell percentage non-tree canopy | 0.58 | 1.17 | -3.06 | 2.63 |
| features with skewed/imbalanced distribution | | | | |
| Daily dry lightening flashes | -48.64 | -29.19 | -81.43 | 79.06 |
| Daily indicator: Diablo or Santa Ana wind | -44.34 | -48.49 | 15.66 | 38.51 |
| Number of transmission lines | -3.02 | -4.04 | 4.1 | -13.14 |
| **Test** | | | | |
| feature | f1_score | precision | recall | loss |
| features with normal distribution | | | | |
| Vegetative type indicators | -14.94 | -16.52 | 6.19 | -0.14 |
| slope | -12.42 | -11.07 | -24.46 | 22.65 |
| Daily maximum air temperature | -9.75 | -8.49 | -21.11 | 26.85 |
| Number of transmission lines | -7.41 | -7.84 | -2.72 | 2.29 |
| Daily specific humidity | -7.26 | -5.03 | -25.63 | 31.13 |
| Grid-cell percentage tree canopy | -4.17 | -4.3 | -2.78 | 6.13 |
| Utility provider | -2.65 | -3.41 | 6.13 | -3.73 |
| Aspect | 0.06 | 0.04 | 0.3 | -0.51 |
| Grid-cell percentage non-tree canopy | 4.34 | 4.78 | -0.03 | -1.16 |
| features with skewed/imbalanced distribution | | | | |
| Daily indicator: Diablo or Santa Ana wind | -42.95 | -45.65 | 19.08 | 132.15 |
| Daily dry lightening flashes | -26.22 | -20.1 | -59.09 | 50.57 |
| Number of transmission lines | -3.02 | -4.04 | 4.1 | -13.14 |

Table 5: **Feature importance analysis** CNN feature importance: Percentage change in metrics with shuffled features. Note that the feature importance are calculated based on random shuffling, so it will be affected by the data characteristics and distribution of each feature. This means we should compare the importance separately for dummies and non-dummies. For dummy variables, we flipped its value to compute the importance, while for variables with continuous values, we shuffled the data across time and space.

## 5.3 Out-of-sample wildfire predictions

Figure 14, presents the results of an out-of-sample forecasting exercise for annual wildfire occurrence in 2021, a very high actual wildfire incidence year. For comparison, we extends our prediction to 2022, a very low actual wildfire incidence year, by simply applying the trained model to the day Dec 31, 2021. Based upon our max-pooling strategy to achieve annual out-of-sample forecasts, we present our estimates for the annual wildfire occurrence for the 2021 fire season in the left-hand side of Figure 14. As shown, we accurately forecast the grid-cell pre-conditions for the location and occurrence of the Mcfarland Fire, the McCash-River Complex-Monument Fires, the Antelope Fire, the Windy-KNP Complex Fires, and the devastating Dixie and Caldor wildfires. As examples of the randomness of wildfire ignition events, the Dixie Wildfire ignition was caused by 65-foot Douglas fir tree that fell on a Pacific Gas and Electric (PG&E) transmission line and the Caldor Fire ignition event was arson. The accuracy of the spatiotemporal CNN forecasts of the locational preconditions of these fires is, however, an important justification for its use in annual wildfire prediction. Notably, the CNN accurately accounted for the cumulative effects of the late August heat, and low humidity as well as the effect of these conditions on the differing vegetative types and canopy coverage of grid-cells. The severity of these conditions then escalated in September due to seasonal offshore winds over challenging topography and wide-spread dry lightening strikes that occur with dry thunderstorms when the air so dry that rain evaporates before it hits the ground. According to the National Weather Service in San Francisco there were approximately 1,100 cloud to ground lightening strikes, many of them dry lightening, recorded in the state on the evening of September 9, 2021 and too many lightning strikes to count that were captured by NOAA Satellite sensors.[27]

The right-hand side of Figure 14, present our annual out-of-sample wildfire occurrence forecasts for the 2022 wildfire season. Compared with 2021, the fire probability is on average much lower in 2022, especially for Northern California. This is mainly due to the trend of the Diablo events being weaker compared with the previous years. Note that the wind events stand out in our feature importance table, which means it's one of the most informative predictors that could affect the prediction significantly.

## 5.4 Expected residential losses from wildfire

To obtain, the percentage of value loss for residential structures post wildfires, we construct a one year pre-wildfire and one-year post-wildfire panel data set of the assessed value of the improvement, using ATTOM Assessor files, for all residential single family properties found

[27] See https://x.com/NWSBayArea/status/1436327235661668357, September 10, 2021, 6:54am.

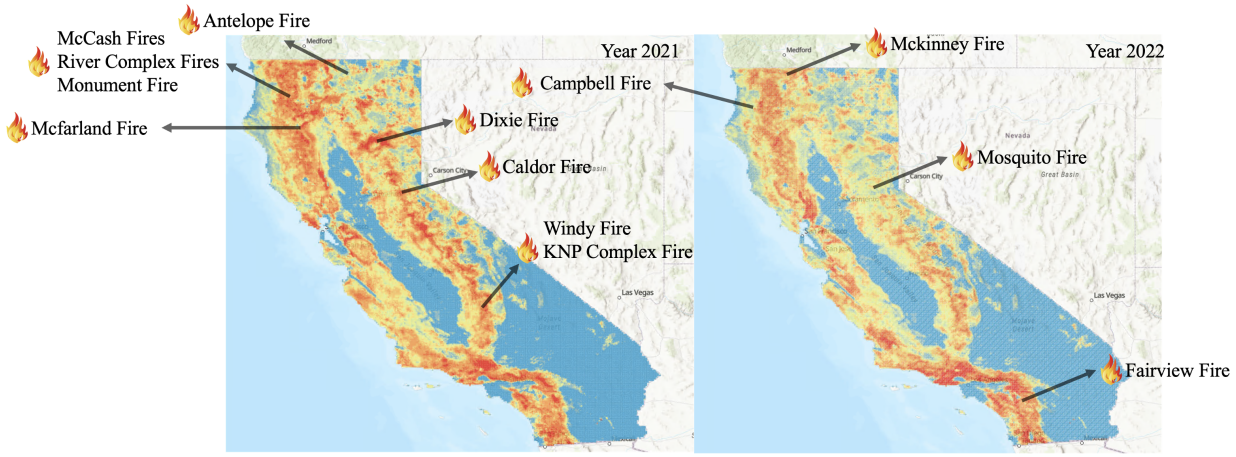



Figure 14: CNN one year ahead out-of-sample wildfire prediction

within the CAL FIRE wildfire burn perimeters. The residential single family house data are obtained from the ATTOM Assessor Files that have been merged to the CAL FIRE burn area shape files. The WUI intermix and interface data are obtained from the CAL FIRE Damage Inspection (DINS) data (see https://gis.data.cnra.ca.gov/datasets/CALFIRE-Forestry::cal-fire-damage-inspection-dins-data/explore). The housing density is computed by the authors. The elevation of the house and its aspect are obtained are obtained from the U.S. Geological Survey (see http://apps.nationalmap.gov/downloader/) given on the latitude and longitude of the property.

We report the results of the regression of percentage loss in pre-wildfire assessed value of the improvement on standard normal transformations for all of the features. As shown in Table 6, we find that the average loss is 28.3% of the assessed value of the improvement for wildfire exposed residential single family properties. The loss percentage increases to 32.5% for properties built before California strengthened its wildfire related building codes in 2008. Percentage losses also increase with the standardized year of the building's age. Single family residential properties at higher elevations and southwesterly aspects also experience higher losses. Standardized housing density is also associated with higher losses as are locations in either the WUI interface or intermix. Overall, these results confirm the merits of mandate building codes in reducing residential single family property losses from wildfire (see Baylis and Boomhower, 2021).

Figure 15 presents the 2021 spatiotemporal CNN out-of-sample predictions for annual wildfire occurrence by zip code aggregates of the grid-cell estimates. As shown, the highest annual wildfire prediction for zip aggregates is 1 for the zip codes just north of San Franciso in Marin County, zip code in the suburban area of Contra Costa County, a zip code in the foothills of the Sierra Nevada, and a zip code aggregate in suburban Los Angeles County.

| | coef | std err | z | P> \|z\| |
|---|---|---|---|---|
| constant | 0.2831 | 0.020 | 14.472 | 0.000 |
| Indicator: Built before 2008 codes | 0.0422 | 0.018 | 2.386 | 0.017 |
| Building Age | 0.0808 | 0.003 | 28.266 | 0.000 |
| Elevation | 0.1448 | 0.003 | 46.784 | 0.000 |
| Aspect | 0.0699 | 0.003 | 24.624 | 0.000 |
| Housing density | 0.1571 | 0.004 | 44.652 | 0.000 |
| Indicator: WUI Intermix | 0.2009 | 0.010 | 19.859 | 0.000 |
| Indicator: WUI Interface | 0.0732 | 0.011 | 6.950 | 0.000 |
| No. Observations: | 23,629 | | | |
| Adjusted R-squared | 0.275 | | | |
| F-Statistic | 1283. | | | |
| Prob (F-statistic): | 0.00 | | | |
| Log-Likelihood: | -12814 | | | |

Table 6: **Regression of the percentage of the pre-wildfire residential single family value (the assessed value of the improvement) that is lost due to wildfire exposure** We estimate the percentage loss of the pre-wildfire assessed value of the improvement one year after a wildfire for all residential single family structures located within CAL FIRE defined burn area perimeters from 2014 through 2020. The residential single family house data are obtained from the ATTOM Assessor Files that have been merged to the CAL FIRE burn area shape files. The WUI intermix and interface data are obtained from the CAL FIRE Damage Inspection (DINS) data (see https://gis.data.cnra.ca.gov/datasets/CALFIRE-Forestry::cal-fire-damage-inspection-dins-data/explore. The housing density is computed by the authors. The elevation of the house and its aspect are obtained are obtained from the U.S. Geological Survey (see http://apps.nationalmap.gov/downloader/) given on the latitude and longitude of the property.

Annual zip code aggregate grid-cell predictions of point 0.9 to 0.7 are found in the Coastal range north of San Francisco and in the foothill ares of the northern Sierra Nevada mountains, where as seen in Figure 14, very severe wild fires actually occurred in 2021. Other areas of annual wildfire occurrence predictions include the eastern suburbs of the Bay Area and southern California that are also coastal range shrubland and woodland forest vegetative areas. The primary difference between the severity of the prediction in the eastern Bay Area and that of southern California is the nature of the winds. The Bay Area is primarily exposed to Diablo Winds that were most severe in the north, whereas southern California is continually exposed to the hazards of the Santa Ana winds and its drying effects on the shrubland and chapparal vegetation and on dense non-tree canopy. The very severe wildfire conditions did lead to one devastating wildfire complex, called the Windy Fire KNP Complex Fire, and other smaller fires in the southern Sierra and again in the coastal range around Los Angeles and San Diego as shown again in Figure 14. In 2021 overall, there was a 167 per cent increase in wildfire acres burned in northern California compared to the ten-year average while a 40 per cent increase was observed in the Northern Rockies. California observed 200,000 fires through the course of the year with an estimated cost of $11.4 Billion.[28]

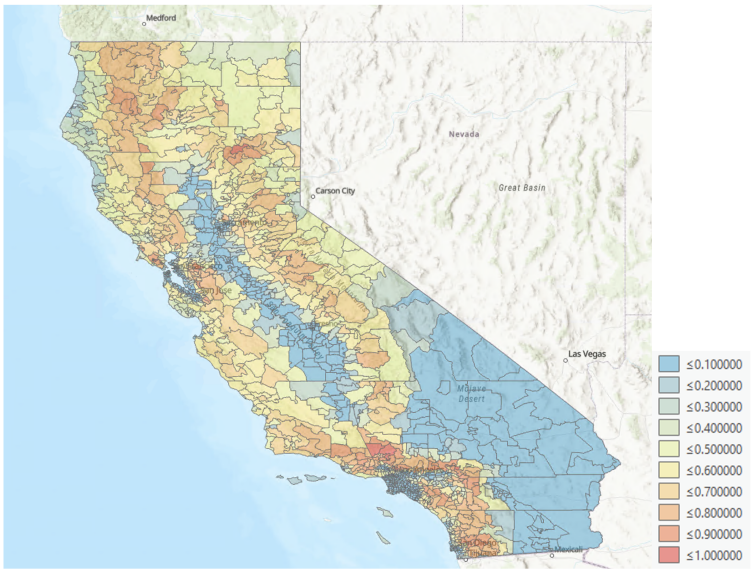



Figure 15: CNN one year ahead (2021) out-of-sample wildfire prediction at the zip code level

Figure 16 reports the grid-cell expected out-of-sample forecast for the per house structural losses from wildfire for 2021. The key to the figure report the dollar amount of loss per house using quintile cutoffs for visualization. The losses were calculated as the grid-cell expected

[28]see https://blueskyhq.io/blog/the-true-cost-of-2022-californian-wildfires

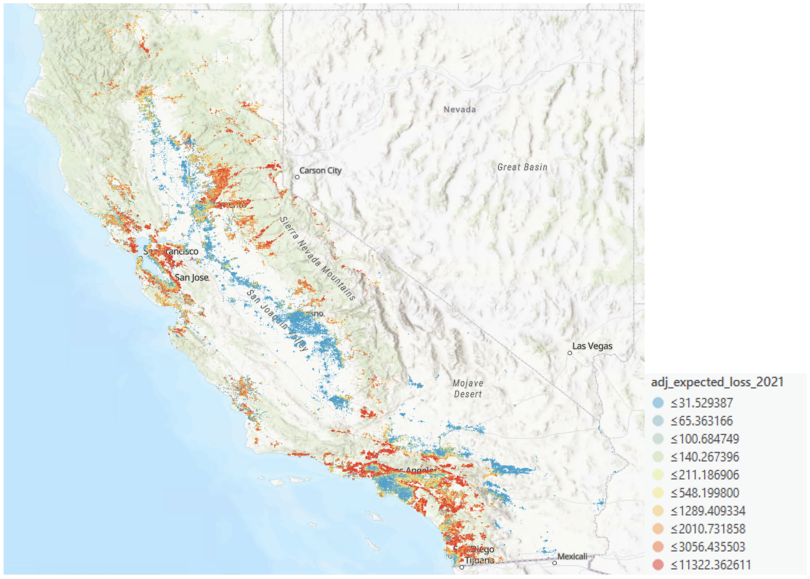



Figure 16: CNN one year ahead (2021) out-of-sample expected loss for all single family houses in California

probability times the percentage of structural loss given by the house by house characteristics of the grid-cell times the 2020 assessed value of the improvement house by house given training precision and recall.[29] As shown, the areas with the highest expected structural wildfire losses are found in the densely populated Bay Area and the Los Angeles Basin from Santa Barbara south to San Diego as well as the grid-cells found in the increasingly populated areas of Sacramento, Placer and El Dorado counties in the foothills of the Sierra Nevada. Overall, the expected residential single family structural losses from our spatiotemporal CNN wildfire predictions and our empirical estimates of the expected losses to structures from these fire is $10.8 billion for 2021. This out-of-sample forecast is remarkably close to the most recent tallies of the structural costs of the 2021 wildfires.

# 6 Spatiotemporal CNN and insurance risk

Given the Section 2 discussion concerning the recent intertemporal loss-smoothing problems of the U.S. fire peril insurance industry, the 2021 California wildfire season was especially damaging to the State Farm Group, the largest P&C insurance carrier in the state. As reported by the National Association of Insurance Commissioners (NAIC), State Farm ex-

[29]Since precision adjusts for the over-estimation, while recall adjusts for the under-estimation we adjust our classifications accordingly. Thus for example, if precision = 10%, then only one out of ten predicted fires will actually happen. So we multiply by 10% to adjust. If recall = 80%, then we only predicted four out of five fires, i.e. we missed one fire. So we divide by 80% to adjust.

perienced a 144.8% fire-peril loss ratio in 2021[30] and AM Best downgraded it's Financial Strength Rating (FSR) to B (Fair) from A (Excellent) and the Long-Term Issuer Credit Rating (Long-Term ICR) to "bb+" (Fair) from "a" (Excellent).[31] In September of 2024, the California Subsidiary of the State Farm group requested a 30% rate increase for homeowners insurance in California, citing concerns about financial solvency and the need to protect itself from potential insolvency due to rising costs and risks in the state. As part of State Farm's 30% rate hike request to the California Department of Insurance, State Farm's Exhibit 1 documented its prior insurance rate increases and its use of vendor wildfire risk models.[32]

The models that State Farm identified for its 2021 eligibility filings included CoreLogic Brushfire; the CoreLogic RQE; AIR Touchstone; and the GRID Fire Model. Of course, all of these models are proprietary and State Farm argued "...Public disclosure to competitors of eligibility criteria that constitute confidential trade secret information is bad public policy and impairs competition."[33] State Farm also provided a table of the impacted zip codes for the 23,480 P&C policy non-renewals the firm had scheduled for 282 zip codes in California.

Table 7 provides summary statistics of our CNN expected 2021 out-of-sample wildfire probabilities for the 272 zip codes where State Farm did not renew at least one P&C policy and for the zip codes where State Farm renewed all of its policies. At least at the empirical mean, our CNN expected wildfire probabilities support the State Farm zip code classifications. As shown, the non-renewal zip codes have a higher wildfire probability mean of 0.5466, standard deviation of 0.2526, whereas the 100% renewal zip codes have a lower wildfire probability mean of 0.2989, standard deviation of 0.2776. Interestingly, the non-renewed classification also includes grid-cell aggregates with quite low CNN wildfire probabilities suggesting the possible mis-classification of zip codes as risky when our CNN models finds that they are not. Similarly, the 100% renewed zip code aggregates include an upper tail of very high expected out-of-sample wildfire occurrence probabilities again suggesting mis-classification of zip codes as riskless where our CNN model finds that they are not.

Figure 17 presents a histogram of zip code renewal classifications organized by the deciles of our spatiotemporal CNN estimates of each zip-code aggregate's expected wildfire occurrence probability. Each of the 1,705 zip-codes aggregates of grid-cells are classified as either non-renewal zip codes, if at least one P&C policy is not renewed, or 100% renewal zip codes, if all of the State Farm policies are renewed. As shown in Figure 17, for the grid-cell aggre-

[30] https://www.insurance.ca.gov/01-consumers/120-company/04-mrktshare/2021/upload/Top25grps2021wa_Revised.pdf

[31] See https://news.ambest.com/pr/PressContent.aspx?refnum=34559&altsrc=2.

[32] https://srp-prod-public-pdfs.s3-us-west-2.amazonaws.com/33f3f200-083c-560f-8b03-f1cc1f939577

[33] https://srp-prod-public-pdfs.s3-us-west-2.amazonaws.com/33f3f200-083c-560f-8b03-f1cc1f939577

| | Zip codes with at least one non-renewed policy | Zip codes 100% renewed policies |
|---|---|---|
| | CNN wildfire expected probability in 2021 | CNN wildfire expected probability in 2021 |
| count | 272 | 1433 |
| mean | 0.5466 | 0.2989 |
| std | 0.2526 | 0.2776 |
| min | 0.0031 | 0 |
| 25% | 0.3703 | 0.0255 |
| 50% | 0.5956 | 0.2165 |
| 75% | 0.7443 | 0.5545 |
| max | 0.9564 | 0.9405 |

Table 7: **Summary statistics for the 2021 spatiotemporal CNN expected out-of-sample zip code aggregate grid-cell wildfire probabilities in non-renewed vs. renewed zip codes by State Farm.** We calculate the fire probabilities at the zip codes level by aggregating the cell-level results.

gates where our CNN expected wildfire probabilities were 0%, State Farm renewed nearly all of the policies in those locations. However, for our CNN expected probabilities of wildfire between 50% and 70%, despite the risks of wildfire occurrence in those locations, State Farm mostly renewed their policies. Surprisingly, for the zip codes aggregates with an 80% probability of occurrence, State Farm's eligibility assignment appears nearly random with 50% of the zip-codes with these high risk probabilities being renewed and 50% of the policies not being renewed. Of course, we are comparing two modeled results, State Farm's to ours. Nevertheless, given State Farm's dire fire loss-rate performance in 2021, it does appear that the modeling technology that they relied upon for both fire-peril pricing and eligibility criteria served them very poorly. At a minimum, this evidence suggests that these likely modeling anomalies should be an important public-policy focus for improving our understanding and vetting procedures for wildfire occurrence modeling by the U.S. P&C industry.

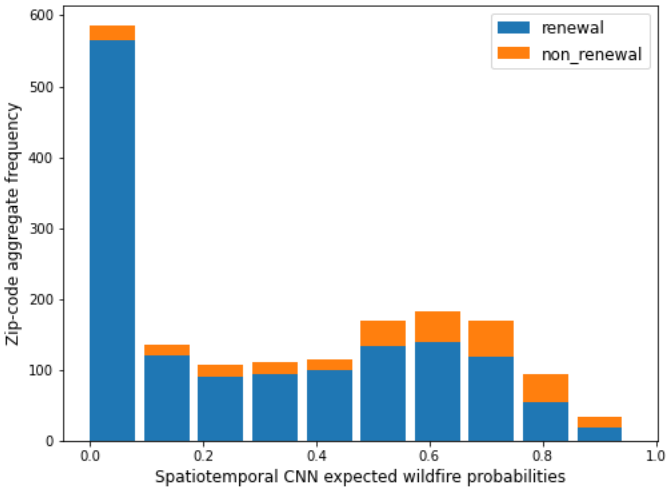



Figure 17: Histogram of State Farms P&C classifications of policy renewals and non-renewals from Spatiotemporal CNN expected 2021 out-of-sample estimate of wildfire occurrence probabilities

# 7 Conclusions

Wildfire risks are escalating at a very rapid pace in California in part driven by climatological factors such rising maximum temperatures from May to October in the Western States, increasing urbanization in risk prone locations, and global forces like El Niño and La Niña that affect drought cycles and snow fall during winter months. Given the very large economic costs associated with wildfires, property and casualty (P&C) insurance companies in California are increasingly finding themselves at the edge of financial survivability. In part, their financial problems have arisen from inherent moral hazard problems associated with insurance as well as legal and institutional frictions that make it difficult for insurance companies to spread risk over time.

More uniquely, California has a self-inflicted regulatory problem that has required wildfire insurers to set rates for future annual catastrophic coverage as the fraction of damages accrued from the 20-year historical mean rather than based on forward looking statistical, or actuarial, models. Additionally, the California Department of Insurance does not allow for the costs, or changes in the cost, of reinsurance risk to be included in insurer rate requests. As a result, California's annual rates now rank next to the lowest in the U.S. leading to insurance company fragility that threatens the future ability of California homeowners to successfully rebuild after fires or even to have access to the mortgage markets.

Since the prohibitions on actuarial models is scheduled to be lifted in December of 2024, this paper proposes a new class of actuarial modelling for wildfire occurrence risk based on spatiotemporal Convolutional Neural Networks (CNNs). We propose that these models are uniquely suited to forecast wildfires across the state of California based on highly imbalanced data and numerous important causal features that are characterized by heavily right-skewed distributions. CNNs capture both spatial and temporal dependencies and can identify correlations between neighboring data points in a time series. We find that spatiotemoral CNNs significantly outperforms logistic regression in estimating the likelihood of wildfire. Using our fire-likelihood estimates, we estimate expected annual fire-related property losses for thousands of grid-cells across the state. We find wide variation in the pre-conditions of wildfire and the estimated probability of wildfire occurrence across the northern and southern areas of the state. Overall we find a total estimated out-of-sample expected loss for 2021 that closely matches the observed cost of wildfires that year. Finally, we discuss the implications of our results for the future financial well-being of the U.S. and California P&C insurance industry and the likely harms to homeowners from modeled-based eligibility classifications that are being applied by the largest carrier in the State. With the partial lifting of actuarial model prohibitions in California, future work should be able to apply our CNN wildfire es-

timates to develop economically justifiable premium reductions in exchange for homeowner mitigation investments. These investments should significantly reduce the risk of property losses from wildfire despite the growing external pre-conditions of these risks and property-level mitigation should better align the incentives of P&C insurers, banks, and homeowners, thus reducing moral hazard.

# Appendix

## A Correlation matrix for gridMET meterology data

| | pr | sph | rmin | rmax | fm100 | fm1000 | etr | pet | vpd | tmmn | tmmx | srad | vs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pr | 1 | 0.04 | 0.39 | 0.23 | 0.38 | 0.3 | -0.25 | -0.25 | -0.22 | -0.14 | -0.25 | -0.29 | 0.33 |
| sph | 0.04 | 1 | 0.25 | 0.33 | 0.07 | -0.06 | 0.27 | 0.34 | 0.22 | 0.61 | 0.48 | 0.33 | 0 |
| rmin | 0.39 | 0.25 | 1 | 0.8 | 0.87 | 0.74 | -0.67 | -0.63 | -0.69 | -0.4 | -0.62 | -0.48 | 0.24 |
| rmax | 0.23 | 0.33 | 0.8 | 1 | 0.83 | 0.67 | -0.6 | -0.56 | -0.7 | -0.46 | -0.53 | -0.35 | 0.19 |
| fm100 | 0.38 | 0.07 | 0.87 | 0.83 | 1 | 0.88 | -0.73 | -0.72 | -0.73 | -0.56 | -0.69 | -0.57 | 0.21 |
| fm1000 | 0.3 | -0.06 | 0.74 | 0.67 | 0.88 | 1 | -0.67 | -0.65 | -0.69 | -0.57 | -0.67 | -0.49 | 0.18 |
| etr | -0.25 | 0.27 | -0.67 | -0.6 | -0.73 | -0.67 | 1 | 0.99 | 0.86 | 0.77 | 0.86 | 0.79 | 0.09 |
| pet | -0.25 | 0.34 | -0.63 | -0.56 | -0.72 | -0.65 | 0.99 | 1 | 0.85 | 0.78 | 0.87 | 0.85 | 0.05 |
| vpd | -0.22 | 0.22 | -0.69 | -0.7 | -0.73 | -0.69 | 0.86 | 0.85 | 1 | 0.83 | 0.9 | 0.59 | -0.17 |
| tmmn | -0.14 | 0.61 | -0.4 | -0.46 | -0.56 | -0.57 | 0.77 | 0.78 | 0.83 | 1 | 0.91 | 0.57 | -0.12 |
| tmmx | -0.25 | 0.48 | -0.62 | -0.53 | -0.69 | -0.67 | 0.86 | 0.87 | 0.9 | 0.91 | 1 | 0.69 | -0.2 |
| srad | -0.29 | 0.33 | -0.48 | -0.35 | -0.57 | -0.49 | 0.79 | 0.85 | 0.59 | 0.57 | 0.69 | 1 | -0.05 |
| vs | 0.33 | 0 | 0.24 | 0.19 | 0.21 | 0.18 | 0.09 | 0.05 | -0.17 | -0.12 | -0.2 | -0.05 | 1 |

Table 8: **Correlations between gridMET climate variables.** Label key: precipitation (pr), specific humidity (sph), minimum relative humidity (rmin), maximum relative humidity (rmax), fuel moisture over 100 hours (fm100), fuel moisture over 1000 hours (fm1000), reference evapotranspiration — Alfalfa (etr), reference evaportranspiration — Grass (pet), vapor pressure deficit (vpd), minimum air temperature (tmmn), maximum air temperature (tmmx), surface radiation (srad), wind speed at 10m (vs)

## B Variable definitions for the loss-given-fire regression

| variable | count | mean | std | min | max |
|---|---|---|---|---|---|
| Indicator: Built before 2008 codes | 23629 | 0.9738 | 0.1598 | 0 | 1 |
| Building Age | 23629 | 42.9303 | 20.419 | 1 | 120 |
| Elevation | 23629 | 463.0631 | 263.351 | 1.6841 | 2255.453 |
| Aspect | 23629 | 0.3262 | 0.6417 | -1 | 1 |
| Housing density | 23629 | 229.9863 | 159.6621 | 1.2346 | 1370.3704 |
| Indicator: WUI Intermix | 23629 | 0.5722 | 0.4948 | 0 | 1 |
| Indicator: WUI Interface | 23629 | 0.3394 | 0.4735 | 0 | 1 |

Table 9: **Variable definitions and summary statistics.**

# References


Abatzoglou, John T., 2013, Development of gridded surface meteorological data for ecological applications and modelling, *International Journal of Climatology* 33, 121–131.

Abatzoglou, John T., Jennifer K. Balch, Bethany A. Bradley, and Crystal A. Kolden, 2018, Human-related ignitions concurrent with high winds promote large wildfires across the USA, *International Journal of Wildland Fire* 27, 277–386.

Abatzoglou, John T., Benjamin J. Hatchett, Paul Fox-Hughes, Alexander Gershunov, and Nicholas J. Nausla, 2021, Global climatology of synoptically-forced downslope winds, *International Journal of Wildland Fire* 41, 31–50.

Abatzoglou, John T., and A. Park Williams, 2016, Impact of anthropogenic climate change on wildfire across western US forests, *Proceedings of the National Academy of Sciences* 113, 505–546.

Alexandre, Patricia M., Susan I. Stewart, Miranda H. Mockrin, Nicholas S. Keuler, Alexandra D. Syphard, Avi Bar-Massada, Murray K. Clayton, and Volker C. Radeloff, 2016, The relative impacts of vegetation, topography and spatial arrangement on building loss to wildfires in case studies of California and Colorado, *Landscape Ecology* 31, 415–430.

Alkhatib, Ramez, Wahib Sahwan, Anas Alkhatieb, and Brigitta Schütt, 2023, A brief review of machine learning algorithms in forest fires science, *Applied Sciences* 13, 2–15.

Apt, Jerome, Dennis Epple, and Fallaw Sowell, 2023, Forest fires: Why the large year-to-year variation in forests burned?, Working Paper 31738, NBER.

Balch, Jennifer K., Bethany A. Bradley, John T. Abatzoglou, R. Chelsea Nagy, Emily J. Fusco, and Adam L. Mahood, 2017, Human-started wildfires expand the fire niche across the United States, *Proceedings of the National Academy of Science* 114, 2946–2951.

Baylis, Patrick W., and Judson Boomhower, 2021, Mandated vs. voluntary adaptation to natural disasters: The case of U.S. wildfires, Working Paper 29621, NBER.

Billmire, Michael, Nancy H. F. French, Tatiana Loboda, R. Chris Owen, and Marlene Tyner, 2014, Santa Ana winds and predictors of wildfire progression in Southern California, *International Journal of Wildland Fire* 23, 1119–1129.

Biswas, Siddhartha, Mallick Hossain, and David Zink, 2023, California wildfires, property damage, and mortgage repayment, Working Paper 23-05, Federal Reserve Bank of Philadelphia.

Boomhower, Judson, Meredith Fowlie, Jacob Gellman, and Andrew Plantinga, 2024, How are insurance markets adapting to climate change? Risk selection and regulation in the market for homeowners insurance, Working Paper 32625, NBER.

Bowers, Carrie Lynn, 2018, *The Diablo Winds of Northern California: Climatology and Numerical Simulations*, Master's thesis, San Jose State University.

Brey, Steven J., Elizabeth A. Barnes, Jeffrey R. Pierce, Christine Wiedinmyer, and Emily V. Fischer, 2018, Environmental conditions, ignition type, and air quality impacts of wildfires in the southeastern and western United States, *Earth's Future* 6, 1442–1456.

Brinkmann, Peggy, Nancy Watkins, Cody Webb, Dave Evans, Gabriele Usan, Michael Glavan, Lillian Zhang, Carolyn Prescott, Tom Larsen, and Grace Lee, 2022, Catastrophe models for wildfire mitigation: Quantifying credits and benefits to homeowners and communities, Research Paper, Casualty Actuarial Society.

Brooks, Matthew L., and John R. Matchett, 2006, Spatial and temporal patterns of wildfires in the Mojave Desert, 1980–2004, *Journal of Arid Environments* 67, 148–164.

Buechi, Hanna, Paige Weber, Sarah Heard, Dick Cameron, and Andrew J. Plantinga, 2021, Long-term trends in wildfire damages in California, *International Journal of Wildland Fire* 30, 757–762.

Burke, Marshall, Anne Discoll, Sam Heft-Neal, Jiani Xue, Jennifer Burney, and Michael Wara, 2021, The changing risk and burden of wildfire in the United States, *PNAS* 118, 1–6.

Calhoun, Kendall L., Melissa Chapman, Carmen Tubbesing, Alex McInturff, Kaitlyn M. Gaynor, Amy Van Scoyoc, Christine E. Wilkinson, Phoebe Parker-Shames, David Kurz, and Justin Brashares, 2022, Spatial overlap of wildfire and biodiversity in California highlights gap in non-conifer fire research and management, *Diversity and Distributions* 28, 529–541.

Cardil, Adrián, Marcos Rodrigues, Joaquin Ramierz, Sergio de-Miguel, Carles A. Silva, Machela Mariani, and Davide Ascoli, 2021, Coupled effects of climate teleconnections on drought, Santa Ana winds and wildfires in southern California, *Science of the Total Environment* 756, 1–8.

Casolaro, Angelo, Vincenzo Capone, Gennaro Iannuzzo, and Francesco Camastra, 2023, Deep learning for time series forecasting: Advances and open problems, *Information* 14, 598.

Chegini, Taher, Hong-Yi Li, and L. Ruby Leung, 2021, HyRiver: Hydroclimate data retriever, *Journal of Open Source Software* 6, 1–3.

Chen, Bin, and Yufang Jin, 2022, Spatial patterns and drivers for wildfire ignitions in California, *Environmental Research Letters* 17, 055004.

Chen, Bin, Yufang Jin, Erica Scaduto, Max A. Moritz, Michael L. Goulden, and James T. Randerson, 2021, Climate, fuel, and land use shaped the spatial pattern of wildfire in California's Sierra Nevada, *Journal of Geophysical Research: Biogeosciences* 126, 1–18.

Chen, Liuyi, Bocheng Han, Xuesong Wang, Jiazhen Zhao, Wenke Yang, and Zhengyi Yang, 2023, Machine learning methods in weather and climate applications: A survey, *Applied Sciences* 13, 12019.

Cleveland, William S., 1979, Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* 74, 829–836.

Cooke, Roger M., Daan Nieboer, and Jolanta Misiewicz, 2014, *Fat-Tailed Distributions: Data, Diagnostics and Dependence*, volume 1 (John Wiley & Sons).

Cruciata, Giorgio, Liliana Lo Presti, Gabriele Ajello, Paolo Cicero, Giacomo Corvisieri, and Marco La Cascia, 2024, Wildfires classification: A comparative study, in *Image Analysis and Processing — ICIAP 2023 Workshops: Udine, Italy, September 11–15, 2023, Proceedings, Part I*, 62–73 (Springer-Verlag, Berlin, Heidelberg).

Dennison, Philip E., and Max A. Moritz, 2009, Critical live fuel moisture in chaparral ecosystems: A threshold for fire activity and its relationship to antecedent precipitation, *International Journal of Wildland Fire* 18, 1021–1027.

Dennison, Philip E., Max A. Moritz, and Robert S. Taylor, 2008, Evaluating predictive models of critical live fuel moisture in the Santa Monica mountains, California, *International Journal of Wildland Fire* 17, 18–27.

Diaz, Adam, 2022, *A Contribution to the Statistical Analysis of Climate-Wildfire Interaction in Northern California*, Ph.D. thesis, Clemson University.

DiMiceli, Charlene, John Townshend, Mark Carroll, and Robert Sohlberg, 2021, Evolution of the representation of global vegetation by vegetation continuous fields, *Remote Sensing of Environment* 254, 112271.

Dixon, Dan J., Yunzhe Zhu, Christopher F. Brown, and Yufang Jin, 2023, Satellite detection of canopy-scale tree mortality and survival from California wildfires with spatio-temporal deep learning, *Remote Sensing of Environment* 298, 113842.

Duclos, Philippe, Lee M. Sanderson, and Michael Lipsett, 1990, The 1987 forest fire disaster in California: Assessment of emergency room visits, *Archives of Environmental Health: An International Journal* 45, 53–58.

First Street Foundation, 2022, The First Street Foundation wildfire model.

Flannigan, M. D., B. M. Wotton, G. A. Marshall, W. J. DeGroot, J. Johnston, N. Jurko, and A. S. Cantin, 2016, Fuel moisture sensitivity to temperature and precipitation: Climate change implications, *Climatic Change* 134, 59–71.

Flannigan, Mike D., Meg A. Krawchuk, William J. de Groot, B. Mike Wotton, and Lynn M. Gowman, 2009, Implications of changing climate for global wildland fire, *International Journal of Wildland Fire* 18, 483–507.

Gershunov, Alexander, Janin Guzman Morales, Benjamin Hatchett, Kristen Guirguis, Rosana Aguilera, Tamara Shulgina, John T. Abatzoglou, Daniel Cayan, David Pierce, Park Williams, Ivory Small, Rachel Clemesha, Lara Schwarz, Tarik Benmarhnia, and Alex Tardy, 2021, Hot and cold flavors of Southern California's Santa Ana winds: Their causes, trends, and links with wildfire, *Climate Dynamics* 57, 2233–2248.

Goss, Michael, Daniel L. Swain, John T. Abatzoglou, Ali Sarhadi, Crystal A. Kolden, A. Park Williams, and Noah S. Diffenbaugh, 2020, Climate change is increasing the likelihood of extreme autumn wildfire conditions across California, *Environmental Research Letters* 15, 505–546.

Guo, Shengnan, Youfang Lin, Shijie Li, Zhaoming Chen, and Huaiyu Wan, 2019, Deep spatial-temporal 3D convolutional neural networks for traffic data forecasting, *IEEE Transactions on Intelligent Transportation Systems* 20, 3913–3926.

Guzman-Morales, Janin, 2018, *Santa Ana Winds of Southern California: Historical Variability and Future Climate Projections*, Ph.D. thesis, University of California San Diego.

Holmes, Thomas P., Jr. Huggett, Robert J., and Anthony L. Westerling, 2008, Statistical analysis of large wildfires, in T. P. Holmes, ed., *The Economics of Forest Disturbances: Wildfires, Storms, and Invasive Species*, 59–77 (Springer Science).

Ismail, Fathima Nuzla, and Shanika Amarasoma, 2023, One-class classification-based machine learning model for estimating the probability of wildfire risk, *Procedia Computer Science* 222, 341–352.

Jaffee, Dwight, and Thomas Russell, 2013, Catastrophe insurance, capital markets, and uninsurable risks, *Journal of Risk and Insurance* 64, 205–230.

Jain, Piyush, Sean C. P. Coogan, Sriram Ganapathi Subramanian, Mark Crowley, Steve Taylor, and Mike D. Flannigan, 2020, A review of machine learning applications in wildfire science, *Environmental Review* 28, 478–505.

Jergler, Don, 2021, Grim California wildfire outlook has insurers forking over big bucks for modeling, *The Insurance Journal* June 18.

Jin, Yufang, James T. Randerson, Nicolas Faivre, Scott Capps, Alex Hall, and Michael L. Goulden, 2013, Contrasting controls on wildland fires in Southern California during periods with and without Santa Ana winds, *Journal of Geophysical Research: Biogeosciences* 119, 432–450.

Joseph, Maxwell B., Matthew W. Rossi, Nathan P. Mietkiewicz, Adam L. Mahood, Megan E. Cattau, Lise Ann St. Denis, R. Chelsea Nagy, Virinia Iglesias, John T. Abatzoglou, and Jennifer K. Balch, 2019, Spatiotemporal prediction of wildfire size extremes with Bayesian finite sample maxima, *Ecological Applications* 29, 1266–1281.

Kahn, Mathew, Amine Ouzad, and Erkan Yönder, 2024, Adaptation using financial markets: Climate risk diversification through securitization, Working Paper 32244, NBER.

Kalashnikov, Dmitri A., John T. Abatzoblou, Nicholas J. Nauslar, Daniel L. Swain, Danielle Touma, and Deepti Singh, 2022, Meteorological and geographical factors associated with dry lightning in central and northern California, *Environmental Research: Climate* 1, 025001.

Kashinath, K., M. Mustafa, A. Albert, J-L. Wu, C. Jiang, S. Esmaeilzadeh, K. Asissadenesheli, R. Wang, A. Chattopadhyay, A. Singh, A. Manepalli, D. Chirila, R. Yu, R. Walters, B. White, H. Xiao, H. A. Tchelepi, P. Marcus, A. Anandkumar, P. Hassanzadeh, and Prabhat, 2020, Physics-informed machine learning: Case studies for weather and climate modelling, *Philosophical Transactions A, Royal Society* 379, 1–34.

Kearns, Edward J., David Saah, Carrie R. Levine, Chris Lautenberger, Owen M. Doherty, Jeremy R. Porter, Michael Amodeo, Carl Rudeen, Kyle D. Woodward, Gary W. Johnson,

Kel Markert, Evelyn Shu, Neil Freeman, Mark Bauer, Kelvin Lai, Ho Hsieh, Bradley Wilson, Beth McClenny, Andrea McMahon, and Farrukh Chishtie, 2022, The construction of probabilistic wildfire risk estimates for individual real estate parcels for the contiguous United States, *Fire* 5, 377–386.

Keeley, Jon E., Janin Guzman-Morales, Alexander Gershunov, Alexandra D. Syphard, Daniel Cayan, David W. Pierce, Michael Flannigan, and Tim J. Brown, 2021, Ignitions explain more than temperature or precipitation in driving Santa Ana wind fires, *Science Advances* 7, 1–9.

Keeley, Jon E., and Alexandra D. Syphard, 2018, Historical patterns of wildfire ignition sources in California ecosystems, *International Journal of Wildland Fire* 27, 781–799.

Keeley, Jon E., and Alexandra D. Syphard, 2021, Large California wildfires: 2020 fires in historical context, *Fire Ecology* .

Kestelman, Stephanie, 2024, Environmental externalities of urban growth: Evidence from the California wildfires, Working paper, Harvard University.

Kochanski, Adam K., Mary Ann Jenkins, Jan Mandel, Jonathan D. Beezley, and Steven K. Krueger, 2013, Real time simulation of 2007 Santa Ana fires, *Forest Ecology and Management* 294, 136–149.

Koh, Jonathan, François Pimont, Jean-Luc Dupuy, and Thomas Opitz, 2023, Sptatiotemporal wildfire modeling through point processes with moderate and extreme marks, *The Annals of Applied Statistics* 17, 560–582.

Kousky, Carolyn, 2019, The role of natural disaster insurance in recovery and risk reduction, *Annual Review of Resource Economics* 11, 399–418.

Kousky, Carolyn, and Roger M. Cooke, 2009, Climate change and risk management, Technical Report RFF DP 0903-REV, Resources for the Future.

Kumar, Lalit, Andrew K. Skidmore, and Edmund Knowles, 1997, Modelling topographic variation in solar radiation in a GIS environment, *International Journal of Geographical Information Science* 11, 475–497.

Lai, Gengke, Xingwen Quan, Marta Yebra, and Binbin He, 2022, Model-driven estimation of closed and open shrublands live fuel moisture content, *GIScience & Remote Sensing* 59, 1837–1856.

Li, Shu, and Tirtha Banerjee, 2021, Spatial and temporal pattern of wildfires in California from 2000 to 2019, *Scientific Reports* 11, 8779.

Linn, Rodman, Judith Winterkamp, Carleton Edminster, Jonah J. Colman, and William S. Smith, 2020, The strong, dry winds of central and northern California: Climatology and synoptic evolution, *Weather and Forecasting* 316, 2163–2178.

Liu, Lu, 2022, The demand for long-term mortgage contracts and the role of collateral, Working Paper, Wharton.

Liu, Yi-Chin, Pingkuan Di, Shu-Hua Chen, ZueMeng Chen, Jiwen Fan, John DaMassa, and Jeremy Avise, 2021, Climatology of Diablo winds in Northern California and their relationships with large-scale climate variables, *Climate Dynamics* 56, 1335–1356.

MacDonald, Glen, Tamara Wall, Carolyn A. F. Enquist, Sarah R. LeRoy, John B. Bradford, David D. Breshears, Timothy Brown, Daniel Cayan, Chunyu Dong, Donald A. Falk, Erica Fleishman, Alexander Gershunov, Molly Hunter, Rachel A. Loehman, Phillip J. van Mantgem, Beth Rose Middleton, Hugh D. Safford, Mark W. Schwartz, and Valerie Trouet, 2023, Drivers of California's changing wildfires: A state-of-the-knowledge synthesis, *International Journal of Wildland Fire* 32, 1039–1058.

Makridakis, Spyros, Evangelos Spiliotis, Vassilios Assimakopoulos, Artemios-Anargyros Semenoglou, Gary Mulder, and Konstantinos Nikolopoulos, 2023, Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward, *Journal of the Operational Research Society* 74, 840–859.

Mallinis, Giorgos, Marius Petrila, Ioannis Mitsopoulos, Adrien Lorent, Stefan Neagu, Bogdan Apostol, Vladimir Gancz, Popa Ionel, and Johann Georg Goldammer, 2019, Geospatial patterns and drivers of forest fire occurrence in Romania, *Applied Spatial Analysis and Policy* 12, 773–795.

McClung, Brandon, and Clifford F. Mass, 2007, Coupled influences of topography and wind on wildland fire behaviour, *International Journal of Wildland Fire* 16, 183–195.

Miller, J. D., C. N. Skinner, H. D. Safford, E. E. Knapp, and C. M. Ramirez, 2012, Trends and causes of severity, size, and number of fires in northwestern California, USA, *Ecological Applications* 22, 184–203.

Oh, Sangmin, Ishita Sen, and Ana-Maria Tenekedjieva, 2024, Pricing of climate risk insurance: Regulation and cross-subsidies, Working Paper, Columbia Business School.

Oliveira, Sandra, Jorge Rocha, and Ana Sá, 2021, Wildfire risk modeling, *Current Opinion in Environmental Science & Health* 23, 1–6.

Opitz, Thomas, 2023, Editorial: EVA 2021 data challenge on spatiotemporal prediction of wildfire extremes in the USA, *Extremes* 26, 241–250.

Paci, James, Matthew Newman, and Tim Gage, 2023, The economic, fiscal, and environmental costs of wildfire in California, Technical Report, Gordon and Betty Moore Foundation.

Prestemon, Jeffrey P., Todd J. Hawbaker, Michael Bowden, John Carpenter, Maureen T. Brooks, Karen L. Abt, Ronda Sutphen, and Samuel Scranton, 2013, Wildfire ignitions: A review of the science and recommendations for empirical modeling, General Technical Report SRS-171, Southern Research Station, Forest Service, United States Department of Agriculture.

Price, Owen, and Ross Bradstock, 2014, Countervailing effects of urbanization and vegetation extent on fire frequency on the wildland urban interface: Disentangling fuel and ignition effects, *Landscape and Urban Planning* 130, 81–88.

Radeloff, Volker C., David P. Helmers, H. Anu Kramer, Miranda H. Mockrin, Patricia M. Alexandre, Avi Bar-Massada, Van Butsic, Todd J. Hawbaker, Sebastián Martinuzzi, Alexandra D. Syphard, and Susan I. Stewart, 2018, Rapid growth of the US wildland-urban interface raises wildfire risk, *Proceedings of the National Academy of Sciences* 115, 3314–3319.

Safford, Hugh D., Alison K. Paulson, Zachary L. Steel, Derek J. N. Young, and Rebecca B. Wayman, 2022, The 2020 California fire season: A year like no other, a return to the past, or a harbinger of the future?, *Global Ecology and Biogeography* 31, 2005–2025.

Seydi, Seyd Teymoor, John T. Abatzoglou, Amir AghaKouchak, Yavar Pourmohamad, Ashok Mishra, and Mojtaba Sadegh, 2024, Predictive understanding of links between vegetative soil burn severities using physics-informed machine learning, *Earth's Future* 12, e2024EF004873.

Sra, Survit, 2019, CNN, Lecture note, 6.687 Fall 2019, Massachusetts Institute of Technology.

Tong, Qi, and Thomas Gernay, 2023, Mapping wildfire ignition probability and predictor sensitivity with ensemble-based machine learning, *Natural Hazards* 119, 1551–1582.

Tran, Du, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, 2018, A closer look at spatiotemporal convolutions for action recognition, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6450–6459.

Voosen, Pau, 2024, El Niño fingered as likely culprit in record 2023 temperatures, *Science* 386, 137.

Wallmann, James, Rhett Milne, Christopher Smallcomb, and Matthew Mehle, 2010, Using the 21 June 2008 California lightning outbreak to improve dry lightning forecast procedures, *Weather and Forecasting* 25, 1447–1462.

Wang, Daoping, Dabo Guan, Shupeng Zhu, Michael Mac Kinnon, Guannan Geng, Qiang Zhang, Heran Zheng, Tianyang Lei, Shuai Shao, Peng Gong, and Steven J. Davis, 2021, Economic footprint of California wildfires in 2018, *Nature Sustainability* 4, 252–260.

Westerling, Jaap F., 2014, "The global multi-asset market portfolio, 1959–2012": A comment, *Financial Analysts Journal* 70, 9.

Williams, A. Park, and John T. Abatzoglou, 2020, Warmer and drier fire seasons contribute to increases in area burned at high severity in Western US forests from 1985 to 2017, *Geophysical Research Letters* 47, e2020GL089858.

Williams, A. Park, and John T. Abatzoglu, 2016, Recent advances and remaining uncertainties in resolving past and future effects on global fire activity, *Current Climate Change Reports* 2, 1–14.

Xi, Dexen D. Z., Stephen W. Taylor, Douglas G. Woolford, and C. B. Dean, 2019, Statistical models of key components of wildfire risk, *Annual Review of Statistics and Its Applications* 6, 197–222.

Yebra, Marta, Xingwen Quan, David Riaño, Pablo Rozas Larraondo, Albert I. J. M. van Dijk, and Geoffrey J. Cary, 2018, A fuel moisture content and flammability monitoring methodology for continental Australia based on optical remote sensing, *Remote Sensing of Environment* 212, 260–272.