# Application of Machine Learning Methods to Predict the Air Quality Impact of Wildfires in Northern California

**Collin Frink[1], Eliot Kim[1], Brian Hu[1], Shreyans Saraogi[1], Jack Cai[1], Gautam Agarwal [1]**

[1] *University of Wisconsin-Madison, Wisconsin, United States of America*

*Github: https://github.com/GAInTheHouse/BadgerX*

*Dataset: https://drive.google.com/drive/folders/1MpH_IXB6W4YWAasxK1BlEM4xLf8DIm7U?usp=sharing*

November 30, 2020

## Abstract

Modern anthropogenic climate change and global warming have led to continuous growth in wildfire intensity and frequency, consequently resulting in increased emission of air pollutants. Timely and accurate predictions of air quality in the aftermath of wildfires is necessary to curtail growth of cardiovascular and respiratory diseases. Machine learning techniques have been explored in the literature as an effective tool to predict the impact of wildfires in general, but extensive work on air quality is less prevalent. This study aims to compare deep learning techniques in predicting air quality in Northern California from 2010 through 2019. Four categories of input features were selected: wildfire, air quality, meteorological, and land cover. The data gathered was transformed to conform with a custom $6^o$ by $6^o$ pixel grid covering Northern California consisting of pixels with dimensions $0.05^o$ by $0.05^o$. Machine learning and deep learning models were trained to predict air quality for the next day across a spatiotemporal domain. Although the accuracy of the models was insufficient for real-world applications, results from deep learning models indicated the potential for providing predictions at high spatiotemporal resolution.

## 1 Introduction

The smoke particulates released by wildfires can greatly harm human health. Accordingly, robust predictions of wildfire-induced air pollutant dispersion are necessary to minimize detrimental impacts. Health officials and the general populace both benefit from a better understanding of the air pollution risks caused by wildfires.

Wildfires are complex chemical processes; a single wildfire event generates several types of air pollutants. These include greenhouse gases (carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O)), photochemically reactive compounds (e.g., carbon monoxide (CO), non-methane volatile organic carbon (NMVOC), nitrogen oxides (NOx)), and fine and coarse particulate matter (PM) [40]. Even after the fire is considered extinguished, there is still a smoldering phase that may last for months, during which the majority of pollutants are produced [4]. The significant amount of harmful particulates generated causes an acute drop in air quality, followed by chronic negative health impacts on neighboring and distant communities. In the western United States, the increased spread of wildfires is expected to cause a 40% increase in organic carbonaceous aerosol concentrations by 2050 [38]. Air pollution from wildfires (such as smoke particles and ozone) has been linked to cardiovascular and respiratory diseases, thus increasing health risks for affected populations [30]. Spikes in air quality related hospital admissions one to two weeks after nearby wildfire events have been observed in the literature [26].

Therefore, the necessity for accurate air quality predictions will only intensify in the coming years, as there is significant evidence to suggest that the burned area of wildfires is increasing due to anthropogenic climate change. In particular, higher global temperatures due to greenhouse gas emissions are strongly linked to increases in the area burned by wildfires [11][15][42]. The increasing trend of fire activity observed across the western United States since the 1980s is strongly correlated with temperature which leads to higher plant transpiration rates [29]. Higher temperatures reduce atmospheric moisture saturation, causing increased transpiration rates from plants and reduced precipitation amounts. Thus, the vegetation which feeds wildfires becomes drier as temperatures rise, enabling fires to draw on more plentiful and flammable fuel [22]. This trend is particularly strong in areas

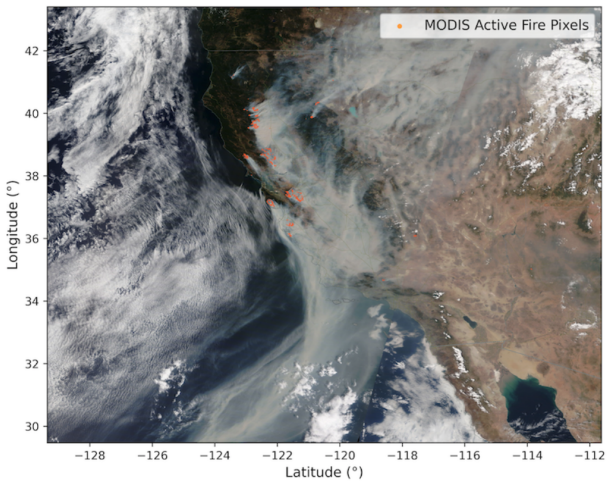


**Figure 1:** *MODIS Terra and Aqua Imagery, Western US, August 20, 2020 [43].*

with higher biomass density, such as Northern California. The four-fold increase in annual area burned by wildfires from 1972 and 2018 in California is largely attributable to lower atmospheric moisture levels in the North Coast and the Sierra Nevada regions [42]. Thus, global warming increases the severity of wildfires.

Northern California was chosen as the spatial area of study both because it is prevalent in the wildfire literature and has experienced severe wildfire seasons with major human impacts in recent years [7]. Wildfires in other regions are influenced by region-specific factors, such as Southern California's unique seasonal wind patterns [51]. Northern California, then, is more likely to have features generalizable across other regions. This study focuses on providing accurate and timely predictions of daily air pollutant concentrations as a result of wildfire incidents in Northern California.

Reliable prediction and timely communication of poor air quality events is a challenging process. One difficulty is the low spatial and temporal resolution of existing sources of air quality measurements. Satellites can only collect indirect air quality measurements, and ground-level monitors are too sparse to capture detailed air quality on a regional scale. Another difficulty for reliable predictions is that the current methods involved in air quality prediction (with and without wildfire effects) are largely deterministic. These models, however, can be significantly inaccurate as they fail to account for the complexity and the sheer number of relevant parameters [37].

Machine learning is capable of analyzing large, complex datasets and finding meaningful patterns that deterministic models cannot. There already is significant literature on applying machine learning to modeling wildfires as a whole, but studies with a specific focus on air quality are minimal. A recent field review found nearly 300 papers that applied machine learning algorithms to different sub-domains of wildfire science [20]. Of the 35 papers classified under the "Fire Effects" section, only seven dealt with modeling "Smoke and Particulate" levels, i.e. measures of air quality. While a couple of these did share some similarities in input parameters and output measures such as PM2.5 [49], in general, all the models were relatively disconnected in their main focus, location studied, and the model used [13][28][33][41][48][49].

This work aims to expand upon existing research by introducing different air quality measures and more recent data. Most prior research in this field has emphasized traditional machine learning models, such as random forests, generalized boosted regression, and multivariate linear regression [33][48][49][52]. However, deep learning models such as artificial neural networks and recurrent neural networks are particularly suited for modeling this subject due the complex nature of the data. By implementing machine learning and deep learning algorithms, this paper applies novel methods to the problem of modeling air pollution due to wildfires. The use of meteorological and land cover data from 2010 to 2019 ensures that our air quality forecasts are pertinent to present conditions.

# 2 Methodology

The domain of study spans Northern California from 2010 to 2019. All data was mapped onto a spatial region ranging from $36^oN$ to $42^oN$ and $118.5^oW$ to $124.5^oW$ (Figure 2). This ensures that relevant urban areas (San Francisco, Sacramento), agricultural regions, and coastal forests are studied. Each grid cell in the spatial region has dimensions of $0.05^o$ by $0.05^o$, with a total of 120 by 120 grid cells. All days from January 1st, 2010 to December 31st, 2019 were included in the study, for a total of 3652 days.

## 2.1 Data Sources

The selected predictor variables fall into four categories: fire, air quality, meteorological, and land cover. These categories were chosen based on their significant impact on the severity of wildfires and the dispersal of aerosol pollutants [2][6][17][34][42]. A full list of features and their units, sources, and spatiotemporal resolutions are listed in Table 1.

Fire data was obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Fire product [14]. This Level 2 dataset is derived from imagery captured by MODIS from the Terra satellite and has daily global retrievals with pixel dimensions of approximately 1km by 1 km. The data was obtained from annual country-based summaries in a CSV format, which were stored in a data archive maintained by Fire Information for Resource Management System (FIRMS).

Brightness (K) and radiative power (MW) are the primary fire-related parameters in this dataset, both

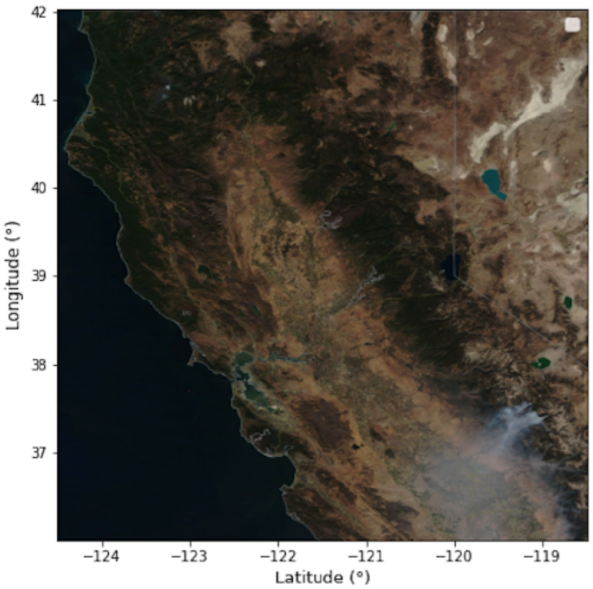


**Figure 2:** *Spatial domain of study with outlines of the custom grid [43].*

| Feature | Dataset |
|---|---|
| ODLO (AOD) | MODIS AOD |
| Fire Radiative Power | MODIS Fire |
| Fire Confidence | MODIS Fire |
| Inverse Distance to Nearest Fire | MODIS Fire |
| Maximum Temperature | NARR |
| Minimum Temperature | NARR |
| Humidity | NARR |
| Precipitation | NARR |
| u-wind velocity | NARR |
| v-wind velocity | NARR |
| Enhanced Vegetation Index | MODIS Vegetation |

**Table 1:** *Feature Description Table*

of which measure the intensity of a fire. These variables are provided for pixels in which a fire has been detected with reasonable confidence, which will be referred to as fire pixels. Radiative power is a linear function of brightness temperature, so fire radiative power was chosen as the measure of fire intensity in order to avoid collinearity within the dataset [12]. Fire confidence percentage for each fire pixel was also used as a measure of the locations of each fire.

In addition to the included fire radiative power and fire confidence percentage variables, a distance to nearest fire feature was derived from the MODIS fire data. This feature is used in several related studies [33][41][48], and provides additional explicit information for models to learn the association between wildfires and air quality. In this study, a regularized inverse distance was used to avoid infinite values on days without any wildfires. Inverse distance to the nearest fire pixel was calculated for each of the 14,400 grid cells, for each day in the domain of this study. The equation used to calculate this is given by equation 1:

$$d_j = \frac{1}{e_{ij} + 0.5} \tag{1}$$

where $e_{ij}$ is the Euclidean distance to the nearest fire pixel i, and $d_j$ is the resulting regularized inverse distance for the grid cell j. Clearly, grid cells with a $d_j$ value of 2 represent fire pixels.

Air quality data was provided by the MODIS Level 2 Aerosol Products, which uses MODIS Terra and Aqua satellite imagery to provide daily global measurements of aerosol optical depth (AOD) at a 3km spatial resolution. AOD is a unitless measure of the concentration of aerosols in the atmospheric columns captured by the MODIS instrument [23]. Due to the column-wise measurement of AOD, this variable is not a perfect measure of surface-level air quality, which is the most relevant factor for human health impacts. However, numerous studies in the literature show that AOD is an accurate and useful proxy variable for approximating PM2.5 and other surface-level aerosol measurements [9][18][24][46]. Furthermore, another advantage of AOD is that AOD is gridded and features high spatial resolution.

The specific MODIS AOD product dataset used in this project is from the Optical_Depth_Land_And_Ocean dataset, which provides measurements at the standard 550nm wavelength over both land and water in the area of study. However, there are many removed pixels in the public AOD data due to cloud cover, instrument configuration errors, and other factors. Among all grid cells in the ten years of this study, only around 27% of them had a valid MODIS AOD value. Thus, interpolation techniques were critical for ensuring the usability of the AOD data.

For meteorological features, the North American Regional Reanalysis (NARR) dataset was used. The NARR model uses the National Centers for Environmental Prediction (NCEP) Eta model and the Regional Data Assimilation System (RDAS) to produce highly accurate estimates of geographical and climatic variables. NARR outputs these variables 8 times daily at 29 atmospheric pressure levels, with each pixel spanning a 32 km by 32 km area. In order to best capture conditions at the surface, this study uses estimates from the lowest pressure level, which are the most relevant for human health impacts.

Temperature (K), precipitation (kg/m$^2$), humidity (kg water/kg air), vertical wind vector component (m/s), and horizontal wind vector component (m/s) were the meteorological variables used from the NARR dataset. Wind patterns were assimilated by taking the average wind speed daily for each component. Different wind patterns across different seasons affects the spread of air pollutants and wildfires. The temperature feature was separated into daily maximum and daily minimum temperatures in the input data. Together, they provide the models with more information than a simple daily average, and are thus commonly used in

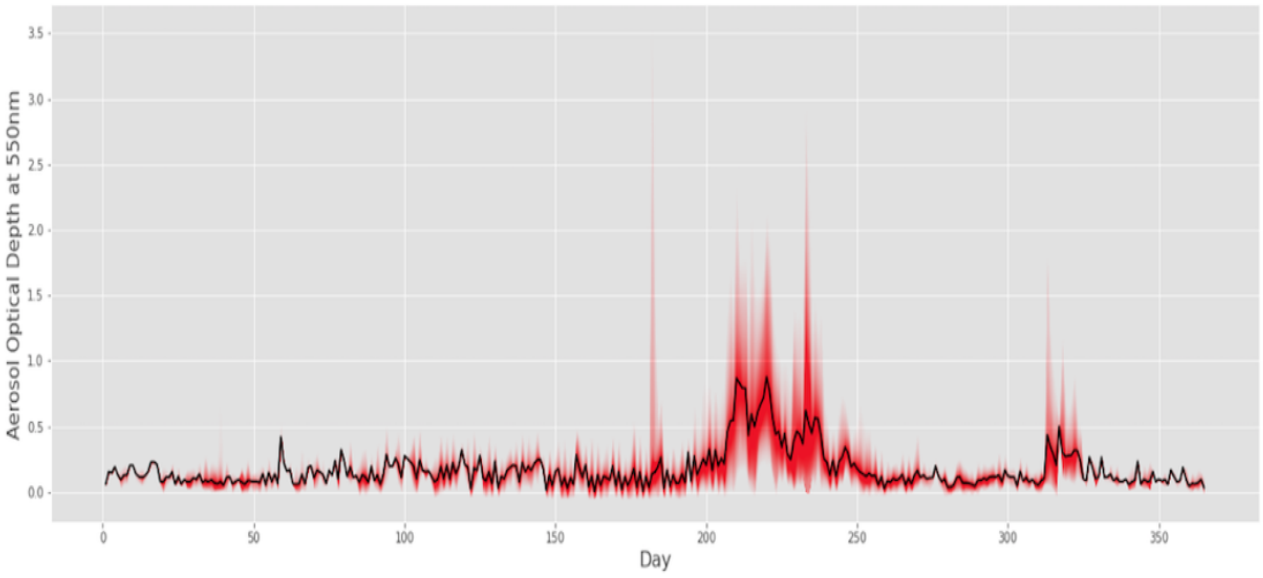


**Figure 3:** *Mean 2018 Daily MODIS AOD Measurements, Northern California.*

related work [27][32][42].

Vegetation data was obtained from the MODIS Vegetation Index product, which contains several datasets measuring vegetation levels at a $0.05^o$ by $0.05^o$ resolution every 16 days [21]. MODIS measurements of vegetation are obtained by estimating the amount of light reflected by leaves within each pixel. The dataset used for this study is the Enhanced Vegetation Index (EVI), which also accounts for the impact of the atmosphere on reflectance measurements. Due to the low temporal resolution of the vegetation data, the same vegetation values were assumed for the 16 day interval covered by each data file.

## 2.2 Regridding and Interpolation

All data was re-gridded onto the custom 120 by 120 grid defined for this study. For each input parameter, every given pixel's parameter value was assigned to the custom grid cell containing the center of the original pixel. If there were multiple pixel center points in a certain custom grid cell, the pixel feature values were averaged. MODIS AOD, NARR, and MODIS Fire were re-gridded using this method. MODIS Vegetation data has the same spatial alignment as the custom grid, thus no re-gridding was necessary.

After re-gridding, both NARR and MODIS AOD had a large number of missing values. For each day in the period of study, the custom grid has 345 valid NARR values out of the 14,400 custom grid cells. This is a result of the low spatial resolution of NARR compared to the grid. However, the NARR features (temperature, precipitation, humidity, and wind) can all be assumed to be relatively smooth across the domain, so spatial interpolation is appropriate. MODIS AOD has varying numbers of values per day, with a generally lower number of valid measurements in the winter months and more valid measurements in the summer. This difference is due to the greater cloud cover in Northern California during winter. MODIS Vegetation had no values over bodies of water as expected, but the default value for water grid cells was sufficient and included in the final input data. MODIS Fire had no missing data values in our region of study.

Spatial kriging was applied to both NARR and AOD datasets to obtain estimated values for the missing grid cells. Kriging interpolation determines missing values based on distance-dependent variance, which is modeled by a variogram [8][16]. A number of previous papers have successfully applied spatial kriging on MODIS AOD [47][51]. The openturns Python package, and specifically the OrdinaryKriging3D function, was applied to MODIS AOD and the six NARR variables for each day in the domain of this study [3]. Figure 5a shows uninterpolated MODIS AOD data for July 31st, 2018, and Figure 5b shows interpolated MODIS AOD data for the same day.

Kriging interpolation resulted in reasonable AOD estimates for almost all days. However, every year except 2011 and 2015 had at least one day with more than 100 values which far exceeds the typical range of AOD values [0,5]. Thus, for days with abnormal interpolated values, the AOD measurements from the neighboring days' were averaged to obtain an estimate for the current day. While this method does not account for daily AOD variations, it results in more accurate estimates compared to leaving the errant interpolated values or forgoing interpolation altogether.

Following re-gridding and interpolation, the 11 feature variables had daily values for each of the 14,400 grid cells in the region of study.

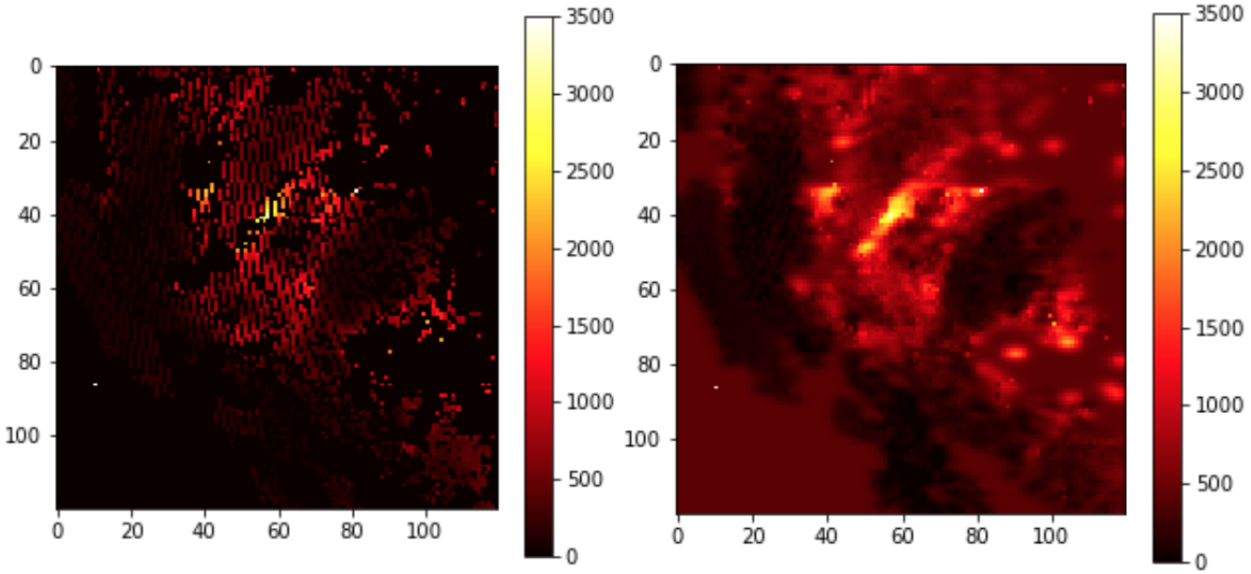


**Figure 4:** *Non-interpolated MODIS AOD for July 31st, 2018 (left); Interpolated MODIS AOD for July 31st, 2018 (right)*

## 2.3 Data Transformation

The Yeo-Johnson power transformation (1) was applied to all features and labels.

$$\phi(\lambda, y) = \begin{cases} \frac{((y+1)^{\lambda}-1)}{\lambda} & \text{if } \lambda \neq 0, y \geq 0 \\ log(y+1) & \text{if } \lambda \neq 0, y = 0 \\ \frac{-((-y+1)^{2-\lambda}-1)}{2-\lambda} & \text{if } \lambda \neq 2, y < 0 \\ -log(-y+1) & \text{if } \lambda = 2, y < 0 \end{cases} \quad (2)$$

$\phi$ is the transformation of the data, while $\lambda$ and y are the optimization parameter and dataset respectively. As an extension of the Box-Cox transformation for including non-positive data, the Yeo-Johnson method was used to normalize the data to reduce data skewness [50]. For each feature, the transformation was applied to the complete set of the data points ranging over the 2D spatial grid and the timeframe in order to ensure the consistent normalization of a 0 mean and standard deviation of 1.

Once the data was normalized, two techniques were applied and evaluated to reduce the dimensionality of the input dataset: Principal Component Analysis (PCA) and autoencoding. This was motivated by the large size of the dataset, which necessitates vast computational resources. PCA is a traditional method of dimensionality reduction that transforms the columns of the feature matrix into an orthonormal basis. Linear dependence ensures that the number of independent features is at most the smaller dimension of the feature matrix. For the training dataset, this implied an overall 3287 number of features, as there are 3287 training days. Using a PCA representation of the data to train, however, can eliminate the "time-series nature" of the data [25].

One method of better incorporating the time-series relationships of the data is to use an autoencoder to reduce the dimensionality. Further motivation for using autoencoders are their properties in noise reduction [19]. Because the MODIS AOD dataset initially contained lots of missing data points, some of which needed to be set to a constant zero for training, these noise-reduction properties could improve the overall performance of the model. Two autoencoders were implemented: one that reduces the training samples to 2500 features and one that reduces the training samples to 5000 features. In all, there are three implemented datasets, notated by PCA, AE2500, AE5000.

## 2.4 Model Structure

A baseline time series model was implemented to evaluate the performance of the learning models-persistence. Persistence is a naive predictor for time-series data that uses the air quality measurements from the previous day (t-1) as a prediction for the current day (t).

Machine and deep learning models have demonstrated success in modelling both wildfires and air quality [33][48][49][41][52]. For each of the feature datasets a number of models were implemented: support vector regression (SVR), boosted regression forest (BRF), linear regression, artificial neural networks (ANN), and recurrent neural networks (RNN). Ultimately, only the ANNs and RNNs developed demonstrated promising results, and are further expanded upon below.

For both the ANNs and RNNs developed, the number of layers, size of layers, optimizer, and other relevant hyperparameters were tuned. Two different network structures were finalized and are shown in Figure 6. The ANN models simply used fully connected layers as

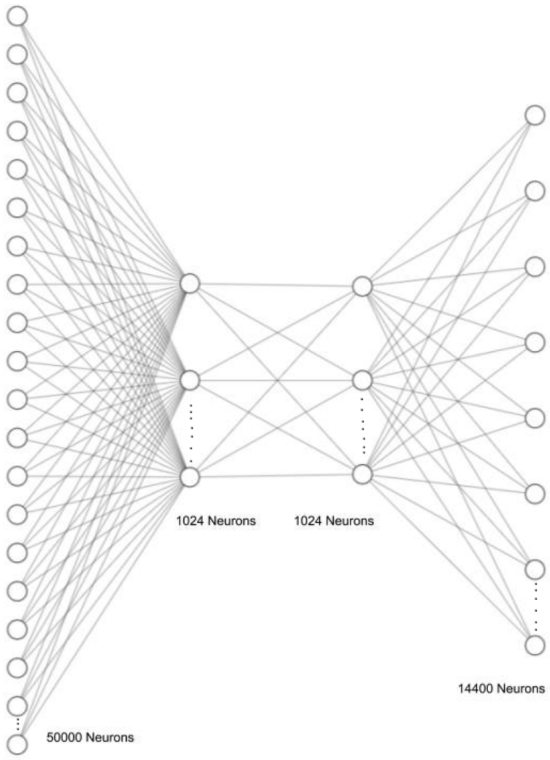


**Figure 5:** *Structure of artificial neural network model with two hidden layers.*

shown in Figure 6. Both a 2-layer ANN and a 4-layer ANN were trained for each dataset. The RNN used a similar structure to the ANN, except with LSTM cells replacing the nodes of the hidden layer. All the models implemented AdaGrad optimization, as opposed to the ADAM and stochastic gradient descent methods [10].

## 3 Experiments and Results

Figure 5 graphically depicts the RMSE values for each model when tested on completely unseen testing data. In every case where different datasets were tested, the majority of models performed better when trained on the auto encoded data compared to when trained

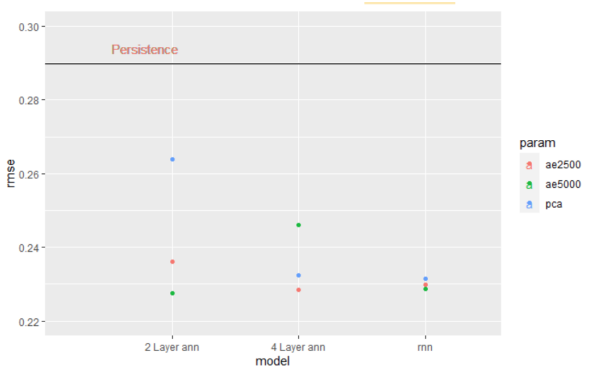


**Figure 6:** *RMSE values for trained models and persistence.*

| Model | PCA RMSE | AE1 RMSE | AE2 RMSE |
|---|---|---|---|
| Persistence | 0.28292 | 0.28292 | 0.28292 |
| ANN1 | 0.26389 | 0.23607 | 0.22755 |
| ANN2 | 0.23240 | 0.22840 | 0.24599 |
| RNN | 0.23155 | 0.22993 | 0.22882 |

**Table 2:** *RMSE Values*

on the PCA-reduced data. The best-performing model overall was the 2-layer artificial neural network trained on the auto encoded data (RMSE of 0.2278), however, many of the models came extremely close to it in terms of performance, namely, the recurrent neural networks trained on auto encoded data. See Table 2 for all RMSE results. All of the models outperformed the baseline persistence model, meaning that they are somewhat better than simply taking today's value as the next day's prediction.

However, further analysis shows that these models are far from being highly predictive, as initially hoped. Fig 8 displays the RMSE per grid cell for the RNN trained on auto encoded data (see below for why the best-performing ANN was not used) as well as the average AOD measurements for the test set. It is clear that the RNN failed to accurately predict AOD on the California mainland, which are the areas of interest. This occurred for a few reasons. First, the range of the RNN's predictions were scaled down by a tenfold relative to the range of the labels, making significant differences appear larger than they actually were. Second, while the RNN certainly learned some trends as indicated by its overall RMSE performance, it failed to learn the most important fire-related trends, which is an issue that arises from either model design or previously discussed data inadequacies. Overall, many of the models followed this pattern, thus showing that neural networks do have the capability to pick up on some trends, and their potential in this field must be further explored.

One notable model that did not follow this pattern was the 2-layer ANN trained on auto encoded data, which was also the best performing model. Figure 9 shows a comparison of one of the model's predictions, compared to the average AOD of the training labels. The model managed to derive an image fairly similar to the average AOD of the training labels to use as its predictions, which allowed it to score very well in terms of RMSE. However, it used this same image as its prediction for the vast majority of its predictions, making it pragmatically useless in terms of predictive capability. Furthermore, as with the other models, the predictions were also significantly scaled down compared to the labels.

Ultimately, the models we ran and tested are not ready for real-world applications and predictions. The conjunction of the 2-layer ANN's performance and unique predictions suggests that air quality over a large region largely remains stable over time. It may be more

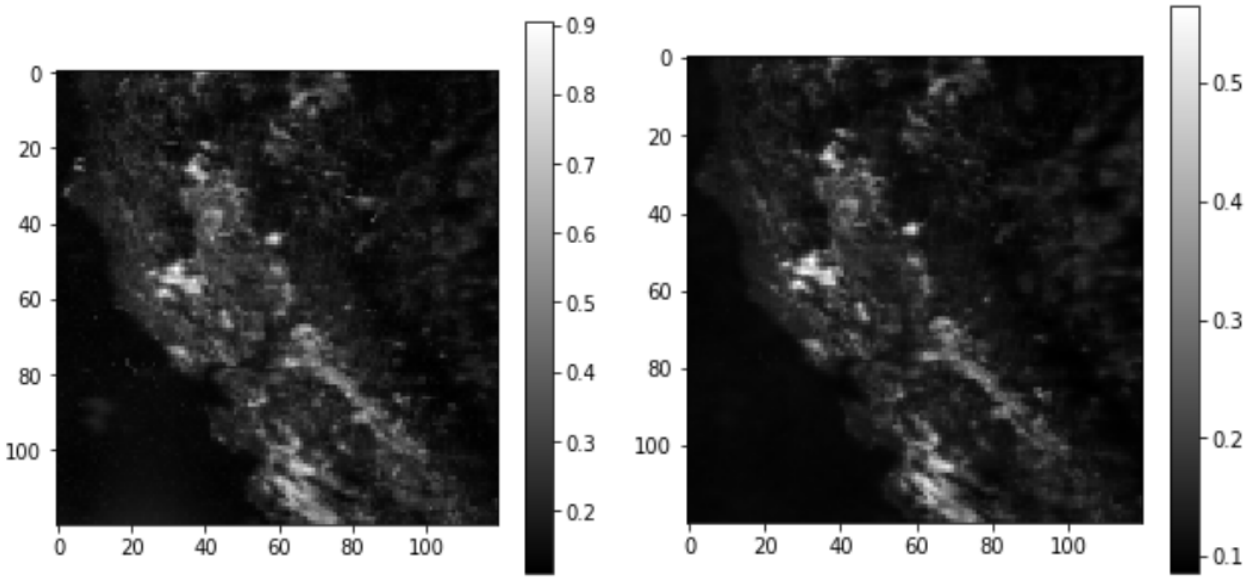


**Figure 7:** *Prediction Errors (left); Average AOD Test Labels (right)*

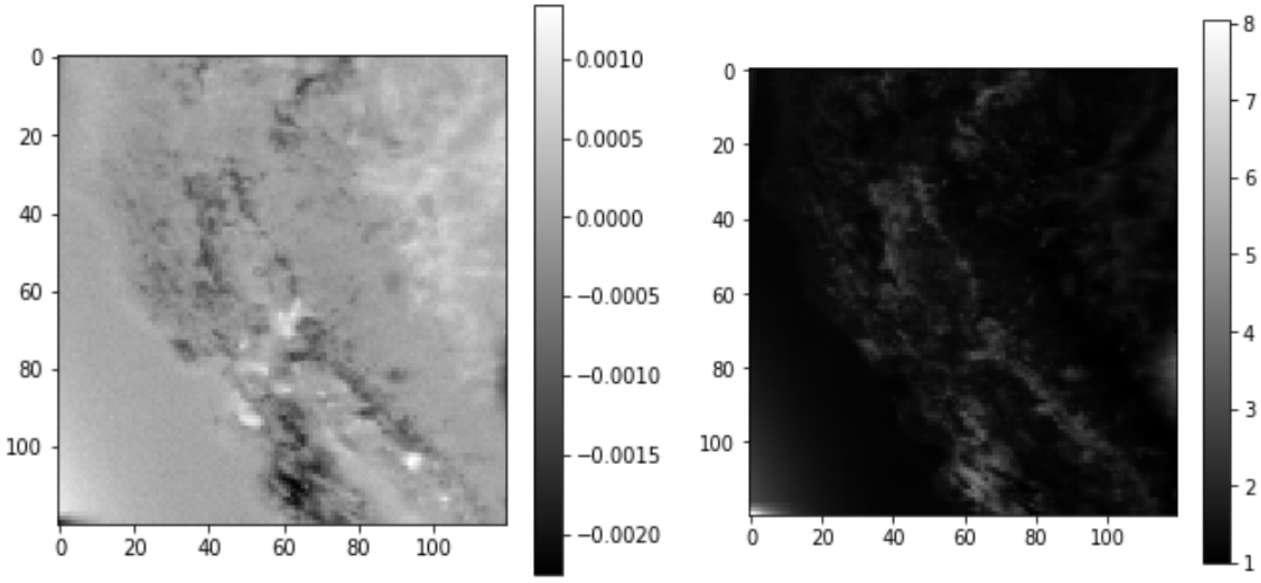


**Figure 8:** *Two hidden layer ANN prediction (left); Average AOD Train Labels (right)*

difficult to make nuanced predictions based on fire occurrences without more complete data, or a more specialized region. These models have shown some promise and should be further tested to determine if this problem is predictable over a large scale.

Despite the lack of overall model success, there was a clear distinction between the quality of results produced by the PCA dataset and the auto encoded datasets. Autoencoding generally produced higher quality dimensionality-reduced features than PCA. This is most easily explained by PCA working across time while the autoencoder transforms each sample independently. This likely preserved the time-based relations better than PCA. In future work, autoencoding should be used to reduce dimensionality.

# 4 Limitations and Future Work

## 4.1 Limitations

The relative novelty of deep learning applications for wildfires as well as the large spatiotemporal domain of this study resulted in a number of technical and conceptual limitations.

A significant source of limitations for the study centered around computing resources. Computational limits on the data pipeline first impacted the re-gridding and interpolation process. Re-gridding all 11 input features for each of the 3,652 days took several days on multiple computers. Interpolation required even greater computing resources, as spatiotemporal kriging took at least 3 hours for each day of MODIS AOD data, while even simple spatial interpolation took up to an hour per day of data. Spatiotemporal interpolation would have resulted in a dataset more reflective of real-world conditions and possibly better predictive models, but this was not a feasible option for this project given computational and time limits.

The constraints imposed by computing resources most acutely affected the model training process. The AI Platform provided by Google Cloud, along with the generous provision of free credits, enabled extensive model training and hyperparameter tuning on high CPU and memory machines. However, due to the large input and output dimensions along with the inability to train models simultaneously in one Google Cloud project, running all models on Google Cloud would have been infeasible within the timeframe of this study. Thus, the majority of model training and hyperparameter tuning was conducted on personal workstations, which have much lower memory, CPU, and GPU specifications compared to Google Cloud machines. This prevented training of the full set of models and hyperparameter combinations, among which more accurate model specifications may have existed.

Pre-training feature engineering and post-training analysis of feature importance were limited by time and computing resources. Possible future work involving these steps is detailed in the Future Work section below.

Another limitation related to both computational and theoretical aspects of the study is the use of ground-based AOD measurements to fill missing MODIS AOD data. AERONET AOD monitoring stations located worldwide provide ground AOD measurements, but they are spatially sparse and have temporal inconsistencies. However, a number of previous papers use AERONET data to fill in missing values in MODIS AOD, as ground-based AOD measurements are not hindered by cloud cover or reflective land surfaces [1][45][47]. During the data processing stage of this study, AERONET AOD data for Northern California was analyzed and regression models relating AERONET and MODIS AOD values were trained. However, regression models showed that AERONET accounts for less than 5% of the variability of MODIS AOD. Furthermore, due to the lack of AERONET monitoring stations in Northern California during the early years of this study, many days had less than five and often had no AERONET monitoring stations with values. The lack of significant correlation between AERONET and MODIS AOD and the small number of values indicated the minimal utility of ground-based AOD for this study. AERONET data may be useful as fill data for a region of study with more monitors and with the use of a satellite AOD product which has a greater correlation with the ground measurements.

An initial lack of sophisticated domain knowledge, both in relation to wildfires and building machine learning pipelines, resulted in steady but slow progress. A significant amount of time was used in gaining basic background knowledge about wildfires and the various climatic and geographic factors affecting the spread of wildfires and air quality. This process involved a broad study of the literature and consultation with domain experts. Once the important factors for air quality impact of wildfires were identified, additional literature review was required to determine suitable interpolation and transformation methods for meaningful and accurate results. The training of machine and deep learning models on complex datasets was a novel process for most authors. Thus, much time at the beginning of the project was spent studying texts and tutorials regarding data processing and model training [5][39]. The significant amount of time needed to build a foundation of intuition limited the complexity of the results and analysis.

## 4.2 Future Work

This study encompasses a broad spatial and temporal range as well as a diverse feature space. However, several aspects of the study present opportunities for further expansion and exploration.

The region of study, Northern California, includes remote arid basins, dense coastal forests, and the sprawling San Francisco Metropolitan Area. Each of these distinct regions will likely have varying levels and in-

teractions of fire, air quality, meteorological, and land cover variables. Thus, a comparative study that divides Northern California into smaller sub-grids, each having more homogeneous characteristics, may yield more tailored and accurate air quality predictions. This approach could be conducted using the same methods as this study, except with smaller overall grid sizes. There is also potential to introduce higher spatial resolution with the smaller areas of study, which would further increase the robustness of air quality predictions.

The analysis and engineering of features provide an opportunity for further work. One potential area of expansion is the inclusion of additional features which are used in related literature. Such features could include land use (agricultural, urban, undeveloped) and vehicle traffic levels, which would account for additional sources of air pollution. Including elevation and planetary boundary layer height may account for the geographical and meteorological factors influencing AOD [33][41][48].

These variables, while generally found to be of lesser importance in predicting air quality, may improve model accuracy by at least a modest amount. Beyond individual variables, a broader improvement to the feature set could involve data from atmospheric simulation models. One of the most prominent models is WRF-Chem, which provides complex global simulations of atmospheric gases and aerosols [44]. A number of related studies use WRF-Chem AOD simulations as an input feature [33][41]. The high resolution simulated AOD values would be especially useful in filling the 77% of grid cells without a satellite AOD measurement. While WRF-Chem was not used in this study due to the lack of extensive prerequisite knowledge, more complete AOD input features will likely improve the predictive performance of models.

Another area of future work is the analysis of feature importance. The authors plan to prioritize feature importance analysis in related future studies because it provides enhanced interpretability of results and is a common step in related studies [48]. Less important features could be removed to create more parsimonious models, and the most important features could be further analyzed for greater applicability. Feature importance metrics are established for tree-based ensemble methods such as random forests and gradient boosting [31][35]. However, determining feature importance for neural networks is much more difficult, because of the "black-box" nature of their learning process.

The methods used in this project are easily transferable to other spatiotemporal domains and aerosol pollution sources. All datasets used in this study except NARR provide global coverage, and several global meteorological datasets can be explored as alternatives to NARR. Thus, studies of other regions with prominent wildfire events such as the Amazon rainforest and British Columbia are feasible. The air quality impact of other high-polluting point source events besides wildfires could also be predicted with similar methods. Such sources may include fossil fuel power plants and volcanic eruptions, although the former would have a lower amount of emissions and the latter a higher amount relative to a standard wildfire event. Future studies exploring these sources can be conducted using similar datasets and methods to this study.

A final aspect for future work involves integrating prediction models with a publicly accessible interface. This integration could enable users to simply view model predictions for a certain time step in the future based on current conditions. A more interactive and comprehensive interface would enable users to add custom wildfires and view corresponding air quality predictions. Providing the ability for the general public to easily obtain real and hypothetical information about air quality as a result of wildfires will increase awareness about the impact of the climate on human health and enable avoidance of poor air quality conditions.

# 5 Societal Aspects

Wildfires in combination with anthropogenic climate change will continue to worsen the air quality experienced by human populations. These air pollutants can have especially detrimental impacts on the health of the dense urban and suburban populations who reside in the path of wildfire smoke dispersion. However, poor air quality does not have equal impacts on all demographics. The area studied in this paper, Northern California, has especially stark socioeconomic inequalities [36]. Air pollution, as well as climate change in general, has a disproportionate impact on those with less socioeconomic capital. The inability to relocate to less urban neighborhoods with lower air pollution, lack of personal transportation to avoid outdoors exposure to air pollutants, and lower awareness of the dangers of poor air quality are a few of the plethora of factors which increases the health risk of air pollution for the less fortunate. Thus, accurate predictions of air quality must be communicated in an easily accessible format. As mentioned in the Future Work section, the production of an Internet-accessible interface that incorporates the data and model pipeline of this study would be one option for providing useful air quality forecasts. Ensuring that the language and layout used in the interface is easily comprehensible by the general public will be especially important in creating an equitable product.

The methodology used in this study has the potential to provide a cost-effective way to spatiotemporally forecast air quality at a level of detail that direct measurement cannot attain. By predicting the movement of air particles from various sources of air pollution, such models can achieve high levels of accuracy with proper real time prediction of air quality in certain areas. Such models can also help to predict air quality in the long term by modeling the long term stagnation of air pollutant particles in certain urban and suburban

areas. Furthermore, this study trains models to predict the impact of wildfires in rural areas on urban populations at a relatively high spatial resolution. Thus our study could help densely populated areas that are in need of fine-grained air quality predictions.

# 6 Conclusion

This paper indicates the potential feasibility and difficulties of applying deep learning techniques to predict wildfire-induced air quality on a large spatiotemporal scale. Both the baseline persistence and naive VAR results demonstrate the need for machine learning methods. Out of all the models, the artificial neural network model with two hidden layers had the best performance among all models and hyperparameter combinations. Autoencoding the input features also showed promise in improving model performance compared to principal component analysis. While the results of this paper are certainly insufficient for real-world air quality prediction, they present not only some of the challenges of building effective models for this problem, but also provide some insight into the potential these models have. A generalized model may be difficult due to data constraints, but with a smaller more specific scope and less sparse data, the models and results demonstrated in this report can certainly be expanded upon. As anthropogenic climate change continues to occur, wildfires will only become a larger issue, especially in regions with intense dry seasons such as Northern California. While these disasters have a slew of harmful effects, their detrimental impact on air quality should become of increasing importance. Models that are able to accurately predict these changes will become invaluable, and the results of this paper offer guidance in that direction.

# 7 Acknowledgements

This work was submitted as a final report for the University of Toronto AI Club's inaugural ProjectX competition. The authors thank the organizers and sponsors of ProjectX for the opportunity to learn and apply machine learning techniques to help solve an issue of global importance. The authors would like to express their gratitude towards Dr. Vahe Vardhanyan, who provided guidance for each step of the study, from topic formulation to model training. The following domain experts at the University of Wisconsin-Madison took time to communicate with the authors and answer many questions in detail: Professors Ankur Desai, Tracey Holloway, Jonathan Martin and Tristan L'ecuyer, Dr. Tyler Caraza-Harter, Feng He and James Kossin, and Mr. Gaurav Doshi.

# References

[1] Adhikary, Brijesh Kulkarni, Sarika D'Allura, Alessio Tang, Youhua Chai, Tianfeng Leung, L. Qian, Y. Chung, C.E. Ramanathan, V. Carmichael, Gregory. (2008). A regional scale chemical transport modeling of Asian aerosols with data assimilation of AOD observations using optimal interpolation technique. Atmospheric Environment. 42. 8600-8615. 10.1016/j.atmosenv.2008.08.031.

[2] Ayanlade, Sina Atai, Godwin Jegede, Margaret. (2019). Variability in atmospheric aerosols and effects of humidity, wind and InterTropical Discontinuity over different ecological zones in Nigeria. Atmospheric Environment. 201. 10.1016/j.atmosenv.2018.12.039.

[3] Baudin, Michaël Lebrun, Régis Iooss, Bertrand Popelin, Anne-Laure. (2017). OpenTURNS: An Industrial Software for Uncertainty Quantification in Simulation. 10.1007/978-3-319-12385-1_64.

[4] Black, Carolyn Tesfaigzi, Yohannes Bassein, Jed Miller, Lisa. (2017). Wildfire Smoke Exposure and Human Health: Significant Gaps in Research for a Growing Public Health Issue. Environmental Toxicology and Pharmacology. 55. 10.1016/j.etap.2017.08.022.

[5] Brownlee, Jason. (2018). Deep Learning for Time Series Forecasting. https://machinelearningmastery.com/deep-learning-for-time-series-forecasting/.

[6] Beer, T. The interaction of wind and fire. Boundary-Layer Meteorol 54, 287–308 (1991). https://doi.org/10.1007/BF00183958.

[7] "Climate Change Indicators: Wildfires." EPA, Environmental Protection Agency, 23 Oct. 2020, www.epa.gov/climate-indicators/climate-change-indicators-wildfires.

[8] Cressie, Noel. (2015). Spatial Prediction and Kriging. 10.1002/9781119115151.ch3.

[9] Donkelaar, Aaron Martin, Randall Park, Rokjin. (2006). Estimating ground-level PM 2.5 using aerosol optical depth determined from satellite remote sensing. Journal of Geophysical Research. 111. 10.1029/2005JD006996.

[10] Duchi, John Hazan, Elad Singer, Yoram. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. Journal of Machine Learning Research. 12. 2121-2159.

[11] Flannigan, M.D., Logan, K.A., Amiro, B.D. et al. Future Area Burned in Canada. Climatic Change 72, 1–16 (2005). https://doi.org/10.1007/s10584-005-5935-y.

[12] Freeborn, Patrick Wooster, Martin Roy, David Cochrane, Mark. (2014). Quantification of MODIS fire radiative power (FRP) measurement uncertainty for use in satellite-based active fire characterization and biomass burning estimation. Geophysical Research Letters. 41. 10.1002/2013GL059086.

[13] Fuentes, Sigfredo Tongson, Eden Bei, Roberta Gonzalez Viejo, Claudia Ristic, R. Tyerman, Stephen Wilkinson, Kerry. (2019). Non-Invasive Tools to Detect Smoke Contamination in Grapevine Canopies, Berries and Wine: A Remote Sensing and Machine Learning Modeling Approach. Sensors. 19. 10.3390/s19153335.

[14] Giglio, Louis. MODIS Collection 6 Active Fire Product User's Guide Revision A. 18 Mar. 2015, lpdaac.usgs.gov/documents/88/MOD14_User_Guide_v6.pdf.

[15] Gillett, N. P., Weaver, A. J., Zwiers, F. W., and Flannigan, M. D. (2004), Detecting the effect of climate change on Canadian forest fires, Geophys. Res. Lett., 31, L18211, doi:10.1029/2004GL020876.

[16] Holdaway, MR. (1996). Spatial modeling and interpolation of monthly temperature using kriging. Climate Research - CLIMATE RES. 6. 215-225. 10.3354/cr006215.

[17] Holden, Zachary Swanson, Alan Luce, Charles Jolly, William Maneta, Marco Oyler, Jared Warren, Dyer Parsons, Russell Affleck, David. (2018). Decreasing fire season precipitation increased recent western US forest wildfire activity. Proceedings of the National Academy of Sciences. 115. 201802316. 10.1073/pnas.1802316115.

[18] Hu, Xuefei Waller, Lance Lyapustin, Alexei Wang, Yujie Al-Hamdan, Mohammad Crosson, William Estes Jr, Maurice Estes, Sue Quattrochi, Dale Puttaswamy, Sweta Liu, Yang. (2014). Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. Remote Sensing of Environment. 140. 220–232. 10.1016/j.rse.2013.08.032.

[19] lu, Xugang Tsao, Yu Matsuda, Shigeki Hori, C.. (2013). Speech enhancement based on deep denoising Auto-Encoder. Proc. Interspeech. 436-440.

[20] Jain, Piyush et al. "A Review of Machine Learning Applications in Wildfire Science and Management." Environmental Reviews (2020): n. pag. Crossref. Web.

[21] Kamel Didan - University of Arizona, Alfredo Huete - University of Technology Sydney and MODAPS SIPS - NASA. (2015). MOD13C1 MODIS/Terra Vegetation Indices 16-Day L3 Global 0.05Deg CMG. NASA LP DAAC. http://doi.org/10.5067/MODIS/MOD13C1.006.

[22] Levin, Noam Heimowitz, Aliza. (2012). Mapping spatial and temporal patterns of Mediterranean wildfires from MODIS. Remote Sensing of Environment. 126. 12-26. 10.1016/j.rse.2012.08.003.

[23] Levy, R., Hsu, C., et al., 2015. MODIS Atmosphere L2 Aerosol Product. NASA MODIS Adaptive Processing System, Goddard Space Flight Center, USA: http://dx.doi.org/10.5067/MODIS/MOD04$_L$2.061.

[24] Li, Jing Carlson, Barbara Lacis, Andrew. (2015). How well do satellite AOD observations represent the spatial and temporal variability of PM2.5 concentration for the United States?. Atmospheric Environment. 102. 260-273. 10.1016/j.atmosenv.2014.12.010.

[25] Li, Lei Prakash, B.. (2011). Time Series Clustering: Complex is Simpler!. Proceedings of the 28th International Conference on Machine Learning, ICML 2011. 185-192.

[26] Liu, Xiaoxiao Bertazzon, Stefania Villeneuve, Paul Johnson, Markey Stieb, Dave Coward, Stephanie Tanyingoh, Divine Windsor, Joseph Underwood, Fox Hill, Michael Rabi, Doreen Ghali, William Wilton, Stephen James, Matthew Graham, Michelle McMurtry, M.Sean Kaplan, Gilaad. (2020). Temporal and spatial effect of air pollution on hospital admissions for myocardial infarction: a case-crossover study. CMAJ Open. 8. E619-E626. 10.9778/cmajo.20190160.

[27] Liu, Yongqiang Stanturf, John Goodrick, Scott. (2010). Trends in global wildfire potential in a changing climate. Forest Ecology and Management. 259. 685-697. 10.1016/j.foreco.2009.09.002.

[28] Lozhkin, V Tarkhov, Dmitriy Timofeev, V Lozhkina, O Vasilyev, Alexander. (2016). Differential neural network approach in information process for prediction of roadside air pollution by peat fire. IOP Conference Series: Materials Science and Engineering. 158. 012063. 10.1088/1757-899X/158/1/012063.

[29] Medina, Susan Vicente, Rubén Nieto-Taladriz, María Aparicio, Nieves Chairi, Fadia Vergara Diaz, Omar Araus, Jose. (2019). The Plant-Transpiration Response to Vapor Pressure Deficit (VPD) in Durum Wheat Is Associated With Differential Yield Performance and Specific Expression of Genes Involved in Primary Metabolism and Water Transport. Frontiers in Plant Science. 9. 1994. 10.3389/fpls.2018.01994.

[30] Olorunfemi Adetona, Timothy E. Reinhardt, Joe Domitrovich, George Broyles, Anna M. Adetona, Michael T. Kleinman, Roger D. Ottmar Luke P. Naeher (2016) Review of the health effects of wildland fire smoke on wildland firefighters and the public, Inhalation Toxicology, 28:3, 95-139, DOI: 10.3109/08958378.2016.1145771.

[31] Pan, Feng Converse, Tim Ahn, David Salvetti, Franco Donato, Gianluca. (2009). Feature selection for ranking using boosted trees. 2025-2028. 10.1145/1645953.1646292.

[32] Piñol, Josep Terradas, Jaume Lloret, Francisco. (1998). Climate Warming, Wildfire Hazard, and Wildfire Occurrence in Coastal Eastern Spain. Climatic Change. 38. 345-357. 10.1023/A:1005316632105.

[33] Reid, Colleen Jerrett, Michael Petersen, Maya Pfister, Gabriele Morefield, Philip Tager, Ira Raffuse, Sean Balmes, John. (2015). Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning. Environmental Science Technology. 49. 10.1021/es505846r.

[34] Ryan, Patrick Lemasters, Grace. (2007). A Review of Land-use Regression Models for Characterizing Intraurban Air Pollution Exposure. Inhalation toxicology. 19 Suppl 1. 127-33. 10.1080/08958370701495998.

[35] Saeys, Yvan Abeel, Thomas Van de Peer, Yves. (2008). Robust Feature Selection Using Ensemble Feature Selection Techniques. Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Part II. 5212. 313-325. 10.1007/978-3-540-87481-2$_2$1.

[36] Sengupta, Somini, and Nadja Popovich. "Wildfire Smoke Is Poisoning California's Kids. Some Pay a Higher Price." The New York Times, The New York Times, 27 Nov. 2020, www.nytimes.com/interactive/2020/11/26/climate/california-smoke-children-health.html.

[37] Shi, H., Jiang, Z., Zhao, B., Li, Z., Chen, Y., Gu, Y., et al. (2019). Modeling study of the air quality impact of record-breaking Southern California wildfires in December 2017. Journal of Geophysical Research: Atmospheres, 124, 6554– 6570. https://doi.org/10.1029/2019JD030472.

[38] Spracklen, Dominick Mickley, L.J. Logan, Jennifer Hudman, R.C. Yevich, R. Flannigan, Mike Westerling, A.. (2009). Impacts of climate change from 2000 to 2050 on wildfire activity and carbonaceous aerosol concentrations in the western United States. Journal of Geophysical Research. 114. 10.1029/2008JD010966.

[39] TensorFlow. (2020). Time Series Forecasting: TensorFlow Core. www.tensorflow.org/tutorials/structured_data/time_series.

[40] Urbanski, S. Baker, Stephen. (2008). Chemical Composition of Wildland Fire Emissions. https://www.fs.fed.us/rm/pubs_other/rmrs_ 2009_urbanski_s001.pdf.

[41] Watson, Gregory Telesca, Donatello Reid, Colleen Pfister, Gabriele Jerrett, Michael. (2019). Machine learning models accurately model ozone exposure during wildfire events. Environmental Pollution. 254. 10.1016/j.envpol.2019.06.088.

[42] Williams, A. P., Abatzoglou, J. T., Gershunov, A., Guzman-Morales, J., Bishop, D. A., Balch, J. K., Lettenmaier, D. P. (2019). Observed impacts of anthropogenic climate change on wildfire in California. Earth's Future, 7, 892– 910. https://doi.org/10.1029/2019EF001210.

[43] "Worldview: Explore Your Dynamic Planet." NASA, NASA, worldview.earthdata.nasa.gov/.

[44] WRF-Chem. (2020). Atmospheric Chemistry Observations amp; Modeling, www2.acom.ucar.edu/wrf-chem.

[45] Xiao, Qingyang Wang, Yujie Chang, Howard Meng, Xia Geng, Guannan Lyapustin, Alexei Liu, Yang. (2017). Full-coverage high-resolution daily PM 2.5 estimation using MAIAC AOD in the Yangtze River Delta of China. Remote Sensing of Environment. 199. 10.1016/j.rse.2017.07.023.

[46] Xie, Yuanyu Wang, Yuxuan Zhang, Kai Dong, Wenhao Lv, Baolei Bai, Yuqi. (2015). Daily estimation of ground-level PM2.5 concentrations over Beijing using 3 km resolution MODIS AOD. Environmental science technology. 49. 10.1021/acs.est.5b01413.

[47] Yang, Jing Hu, Maogui. (2018). Filling the missing data gaps of daily MODIS AOD using spatiotemporal interpolation. The Science of the total environment. 633. 677-683. 10.1016/j.scitotenv.2018.03.202.

[48] Yao, Jiayun Angela Brauer, Michael Raffuse, Sean Henderson, Sarah. (2018). Machine Learning Approach To Estimate Hourly Exposure to Fine Particulate Matter for Urban, Rural, and Remote Populations during Wildfire Seasons. Environmental Science Technology. 52. 10.1021/acs.est.8b01921.

[49] Yao, Jiayun Angela Raffuse, Sean Brauer, Michael Williamson, Grant Bowman, David Johnston, Fay Henderson, Sarah. (2018). Predicting the minimum height of forest fire smoke within the atmosphere using machine learning and data from the CALIPSO satellite. Remote Sensing of Environment. 206. 98-106. 10.1016/j.rse.2017.12.027.

[50] Yeo, In-Kwon Johnson, Richard. (2000). A new family of power transformations to improve normality or symmetry. Biometrika. 87. 10.1093/biomet/87.4.954.

[51] Yu, Chao Chen, Liangfu su, Lin Fan, Meng Li, Shenshen. (2011). Kriging interpolation method and its application in retrieval of MODIS aerosol optical depth. Proceedings - 2011 19th International Conference on Geoinformatics, Geoinformatics 2011. 1-6. 10.1109/GeoInformatics.2011.5981052.

[52] Zou, Y.; O'Neill, S.M.; Larkin, N.K.; Alvarado, E.C.; Solomon, R.; Mass, C.; Liu, Y.; Odman, M.T.; Shen, H. Machine Learning-Based Integration of High-Resolution Wildfire Smoke Simulations and Observations for Regional Health Impact Assessment. Int. J. Environ. Res. Public Health 2019, 16, 2137.

[53] Yufang Goulden, Michael Faivre, Nicolas Veraverbeke, Sander Sun, Fengpeng Hall, Alex Hand, Michael Hook, Simon Randerson, James. (2015). Identification of two distinct fire regimes in Southern California: Implications for economic impact and future change. Environmental Research Letters. 10. 10.1088/1748-9326/10/9/094005.