Comparison of deep 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*

November 10, 2020

1 Introduction

Wildfires release smoke particulates that can greatly harm human health. Robust predictions of wildfire-induced air pollutant dispersion are necessary to plan for the worst-case scenarios and minimize detrimental impacts. Health officials and residents 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 with varying chemical properties. 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) [25]. 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 [2]. 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. For example, the increased spread of wildfires is expected to cause a 40% increase in organic carbonaceous aerosol concentrations in the western US by 2050 [22]. Air pollution from the fires (such as smoke particles and ozone) has been linked to cardiovascular and respiratory diseases, thus increasing health risks for affected populations [19]. Spikes in hospital admissions one to two weeks after nearby wildfire events have also been observed in the literature [16].

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 [8][12][27]. Higher temperatures reduce atmospheric moisture content, 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 [15]. The increasing trend of fire activity observed across the western United States since the 1980s is strongly correlated with temperature as well as vapor pressure deficit, which determines plant transpiration rates [18]. This phenomenon is particularly evident in areas with higher biomass density, such as our region of focus, 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 [27]. Thus global warming, one of the prominent features of climate change, increases the potential spatial growth of wildfires.

Our work focuses on providing accurate and timely predictions of daily fine particulate and air pollutant concentrations as a result of wildfire incidents in Northern California. Northern California was chosen as the spatial area of study both because it is prevalent in the wildfire literature and seen severe wildfire seasons with major human impacts in recent years [7].

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

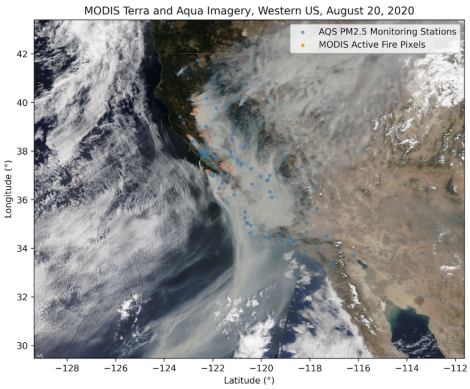

Figure 1: *[28] The above map depicts AQS monitor stations and MODIS fire pixels. The release of smoke from fire pixels and the distant dispersion of the air pollutants can clearly be seen*

wildfire effects) are largely deterministic (equation-based). These models, however, can be significantly inaccurate as they fail to account for the number and sheer complexity of the relevant parameters [21].

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 very little to do with a specific focus on air quality. A recent field review [14] found nearly 300 papers that all used machine learning algorithms to model different subdomains of the wildfire problem. Of the 35 papers classified under the "Fire Effects" section, only seven dealt with modeling "Smoke and Particulate" levels, which are measures of air quality. While a couple of these did share some similarities in input parameters and output measures such as PM2.5 [30], in general, all the models were relatively disconnected in their main focus, location studied, and the model used [9][17][20][26][29][30].

In our work, we aim to expand upon existing research by introducing other learning algorithms, different air quality measures (AOD), 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 [20][29][30][33]. We believe deep learning models such as recurrent neural networks and convolutional neural networks are particularly suited for modeling this subject due to the complexity of the temporal nature of the data as well as the prominence of satellite data. By implementing statistical, machine, and deep learning algorithms, we are able to learn more about the relative performance of each type of model. Our 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 spatial grid chosen spans from 33^o N to 43^o N and 116^o W to 126^o W with grid squares of side length 0.05^o, resulting in a final grid of 200 by 200 square cells. This consists of almost the entirety of California and most of Nevada, excluding the area south of Los Angeles. While this is larger than the intended region of study, Northern California, it ensures that the impact of air quality and fire events outside of the region are fully captured in the data. Temporally, data spanning from 2010-2019 is used to train and evaluate each of the models. Note that all preliminary results (reported below) are for models trained and tested on 2019 data, and do not represent finalized (or even meaningful) results.

The following data products were used to obtain the input features for our learning models.

The North American Regional Reanalysis (NARR) dataset contains information about the temperatures, winds, moistures, soil, and various other meteorological conditions at 29 pressure points using the ETA 32 km grid point. For our purposes, we imported the dataset in netCDF format and converted it into CSV format while extracting the essential features within the spatial region of study.

The following input features from NARR data were chosen because of the established impact of these variables on the spread of wildfires and the dispersion of air pollutants: Temperature (Kelvin), Humidity (kg/kg), Wind velocity (m/s), and Wind velocity vector components (m/s) [20][26][33].

One feature that was not included among the input features thus far that will be incorporated for future model iterations is precipitation.

The MODIS (Moderate Resolution Imaging Spectroradiometer) Land Cover Type Product includes several datasets containing annual land cover classifications at a 0.05o by 0.05o (latitude, longitude) spatial resolution with global coverage. Land cover classifications for the product are determined from MODIS satellite imagery and measurements, supervised classification, and probabilistic algorithms. This product provides a file for each year from 2001 to 2019 [23]. The dataset from the MODIS Land Cover Product used in this project is the classification system established by the International Geosphere-Biosphere Programme (IGBP), which includes 17 different land cover classifications.

Land cover data products which provide continuous data will be used in the final model iterations. The categorical MODIS Land Cover data, when one-hot encoded as detailed later in this section, comprise a disproportionately large portion of the input dataset. Continuous land cover data will enable lower computation times and condense land cover information from several low-importance variables to one or two higher-importance features in the final dataset.

The Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Fire MOD14 Ver-

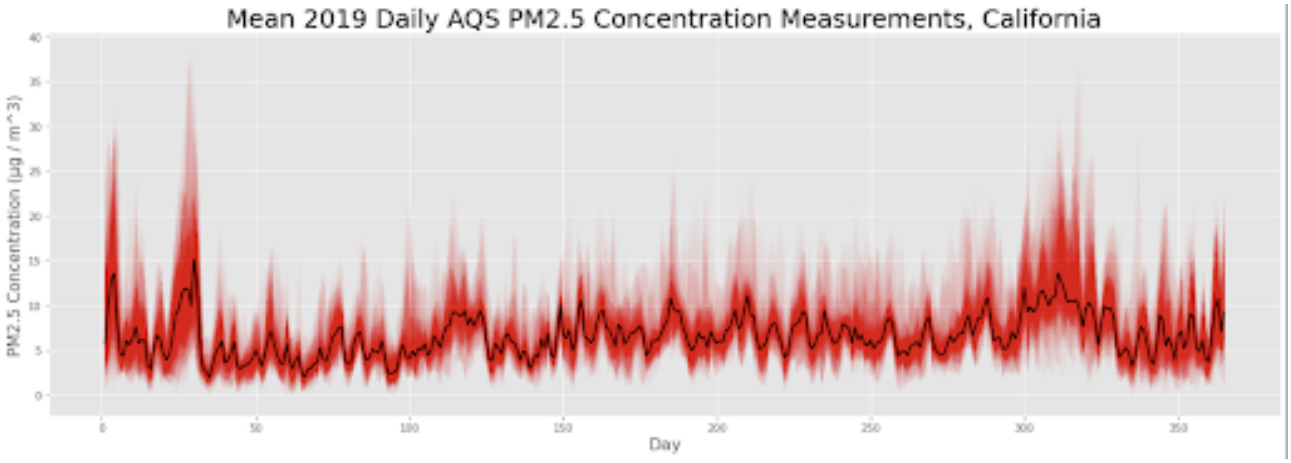

Figure 2: *Air Quality at* $(39.76^oN, 121.84^oW)$

sion 6 product are produced daily [11]. The MOD14 product is used to identify fires in the spatial area of study. Each pixel within the data covers a grid cell with dimensions of approximately 1km by 1km. It has the ability to detect small fires using dynamic thresholding. The data is maintained in HDF-EOS format but was extracted as CSV through the FIRMS server for an easier processing step.

The following input features from MODIS Active Fire data were used: Fire radiative power (megawatts) and fire confidence (%)

Fire confidence measures the percent confidence of the existence of a fire within a particular MODIS grid cell. This fire data also includes brightness temperature data, another variable which measures the intensity of fire. However, fire radiative power is a linear function of brightness temperature [10], so fire radiative power was chosen as the measure of fire intensity in order to avoid collinearity within the dataset.

The Air Quality Service (AQS) maintained by the Environmental Protection Agency (EPA) provides several data products based on the EPA's ground monitoring station measurements [1]. These data products include hourly measurements of PM2.5, PM10, ozone, sulfur dioxide, nitrogen dioxide, and carbon monoxide, as well as several meteorological and toxin measurements. The 2019 hourly PM2.5 data is the air quality measure used in this project thus far. The spatial domain of study includes 39 EPA air quality monitoring stations, which on average provide measurements for 97.8% of the possible measurement times. Thus, the AQS PM2.5 data has suitable temporal density. The EPA AQS PM2.5 data was used to generate the output feature for the models. The hourly PM2.5 data were averaged to a daily time scale, to coincide with the temporal resolution of the input features. The final shape of the output feature was 365 rows by 39 columns, representing the 365 days in 2019 and the 39 air quality monitoring stations in the region of study.

The MODIS Land Cover grid cells exactly match the custom grid, so no re-gridding was necessary. The MODIS Active Fire data points were assigned to the nearest grid cell in the custom grid. This fire data is very spatially sparse and inconsistent over time. Thus, re-gridding was applied to ensure a degree of spatial consistency in the data. The NARR data was not re-gridded, because this data has a lower spatial resolution than the intended grid. Thus, re-gridding would require interpolation, which would introduce extraneous noise to the dataset.

The Yeo-Johnson Power transform (equation number) was used for quantitative features (all but land cover) and labels to normalize each to a mean of 0 and standard deviation of 1.

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

ϕ is the transformation of the data, while λ and y are the optimization parameter and dataset respectively [31]. For the categorical land-cover data, one hot encoding was used to create a feature vector for each land classification. Each vector was then included in the final training set.

After the features were standardized, Principal Component Analysis (PCA) was applied to reduce the dimensionality of our dataset while preserving as much of the original variance as possible. The principal components are found by decomposing the matrix into its singular value decomposition. From this decomposition, the vectors that explain the most variance in the data can be found and used to form the most accurate low-rank approximation of the original data. This process was essential for building and training the models, since it significantly reduced the size of the dataset and thus reduced the required computational time and resources.

Two baseline time series models were implemented to evaluate the performance of the learning models:

Persistence and Autoregressive Integrated Moving Average (ARIMA) models. Persistence models are 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). Integrated Autoregressive Moving Average (ARIMA) models non-stationary time series via differencing. The autocorrelation and partial autocorrelation plots for several PM2.5 monitoring station time series data indicated that an autoregressive model of order 1 would be most appropriate. Thus, vector autoregression (VAR), a multivariate variant of ARIMA which uses the autoregressive terms but does not difference the data, was determined to be most appropriate for the PM2.5 time series data [32]. A VAR model of order 1 was fitted on the normalized PM2.5 dataset using R and the MTS package [24].

Linear Regression fits the line which minimizes the least-squared error between estimated and true output values. A linear regression model was fitted to the PCA-transformed dataset and used as a baseline comparison for other models.

SVR, the extension of SVM for quantitative data, was implemented using linear, polynomial, sigmoid, and radial basis (RBF) kernels. RBF kernels generally outperformed the others, and were for the results section.

Random Forests have been one of the most prominent models in this field and have generally performed well compared to other models [5]. They are created by sequentially training many "weak learners", which are typically small decision trees, in a process called boosting. The primary hyperparameter for this model is the number of trees to train, but there are various regularization parameters that can also be adjusted. These hyperparameters were tuned as described later in the section.

Multilayer neural network models were used to create a deep learning air quality prediction model. Neural networks learn a complex, non-deterministic function by training a set of weights to transform the input features [4]. An artificial neural network using sigmoid activation functions, stochastic gradient descent, and the mean-squared error cost function was trained on the PCA-transformed input data. The network outputted a matrix with the predicted air quality measurements for each monitoring station and each day of 2019. The adjustable hyperparameters for the neural network are the learning rate, number of layers, and nodes per layer. These hyperparameters will be systematically tuned using the validation method described later in this section.

Recurrent neural networks are a type of neural network that are especially well suited for sequential data. Unlike standard feed-forward networks, RNNs are trained sequentially on the data and pass a "hidden state" to the next iteration that represents the information learned in the previous steps. As such, RNN suits the time-series nature of the data well and was trained on the PCA-reduced data. There are many hyperparameters for RNNs, including but not limited to the number of RNN nodes per layer, the number of layers, learning rate and optimization function. These hyperparameters were tuned as described in the section below. One more important model design choice is the type of RNN node the model uses; standard RNNs have issues with "short-term memory", since the hidden state naturally weighs more recent information more heavily, so long-short-term memory (LSTM) nodes and gated recurrent units (GRU) were developed to solve this problem [6][13]. LSTMs have thus far performed the best so far, and were used for the results section.

Many of the models require several parameters to be set before the training of the model. Tuning these parameters can significantly affect the model's performance and is known as hyperparameter tuning. Since our data is time-series based, the typical k-fold cross-validation methods cannot be employed because the training data must be in one continuous time block. Instead, "walk-forward" validation will be used, which is a process by which the models are trained on varying time ranges, then some later time range is used as the validation set. For instance, the models could be trained on the first two years and predict on the third year, then they could train on the first four years and predict on the fifth, and so on. The RMSE values of these predictions are averaged and serve as the score for the specific hyperparameters of this model. The hyperparameters are tuned by choosing the combination of hyperparameters that leads to the lowest score (lowest error).

3 Results

Currently iterations of both the dataset and the learning algorithms do not have verified results. So, this section will focus on chosen model evaluation metrics and present some preliminary results for a proof-of-concept model used to establish our overall project pipeline. All preliminary results are trained on just data from 2019; a small part of the complete data set. Thus, they are meant to be interpreted as a proof-of-concept, as opposed to being meaningful results. For the preliminary results, each algorithm was trained on the first 60% of the 2019 data, and performance evaluated on the latter 40%. This train-test split ensures that significant parts of the fire season, May through November [3], are included in both the training and testing datasets. The training dataset includes data for January 1st through June 30th, 2019, while the test dataset contains data for June 30th through December 31st, 2019. The consistency between the training and test datasets will ensure that model evaluation metrics represent the strength of the algorithm and not the differing trends within the output data.

Two statistical measures, root mean square error (RMSE) and R^2 will be used to evaluate the performance of each of the algorithms. RMSE is a standard absolute measurement of the error for each predictive

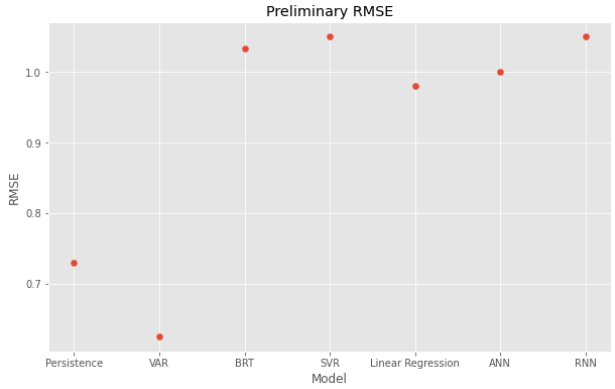

Figure 3: *Preliminary RMSE results*

model. Due to an incomplete dataset and limited time frame, only the preliminary RMSE scores of the completed models predicting one time step forward have been produced (see Fig. 3). In the future, models will be tested for both single-step predictions as well as multi-step predictions and evaluated using RMSE and R-squared. In the final deliverable, there will be plots of the scores of all the models for both types of prediction. Further, for the best performing models, there will be plots showing how their predictions compare to the actual air quality values.

References

[1] AQS (Air Quality System) User Guide. Environmental Protection Agency, 2018, www.epa.gov/sites/production/files/2018-07/documents/aqs_user_guide_2018_2.pdf.

[2] 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.

[3] Borunda, Alejandra. "Climate Change Is Contributing to California's Fires." California's Fires Are Partly Fueled by Climate Change, National Geographic Society, 30 Oct. 2019, www.nationalgeographic.com/science/2019/10/climate-change-california-power-outage/.

[4] Braspenning, Peter. (1995). Introduction: Neural Networks as Associative Devices.. 1-9. 10.1007/BFb0027020.

[5] Breiman, L. (2001). Random Forests. Machine Learning. 45. 5-32. 10.1023/A:1010950718922.

[6] Chung, Junyoung & Gulcehre, Caglar & Cho, KyungHyun & Bengio, Y.. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.

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

[8] 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

[9] 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.

[10] 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.

[11] Giglio, Louis. MODIS Collection 6 Active Fire Product User's Guide Revision A. 18 Mar. 2015, lpdaac.usgs.gov/documents/88/MOD14.

[12] 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.

[13] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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

[20] 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.

[21] 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

[22] 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.

[23] Sulla-Menashe, Damien, and Mark A Friedl. User Guide to Collection 6 MODIS Land Cover (MCD12Q1 and MCD12C1) Product. 14 May 2018, lpdaac.usgs.gov/documents/101/MCD12.

[24] Tsay, Ruey S, and David Wood. "All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models." Package 'MTS', CRAN, 8 Oct. 2018, cran.r-project.org/web/packages/MTS/MTS.pdf.

[25] Urbanski, S. & Baker, Stephen. (2008). Chemical Composition of Wildland Fire Emissions.

[26] 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.

[27] 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

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

[29] 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.

[30] 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.

[31] 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.

[32] Zivot, Eric, and Jiahui Wang. Modeling Financial Time Series with S-PLUS®. Springer New York, 2006.

[33] 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.