Comparison of Machine Learning Methods to Predict the Air Quality Impact of Wildfires

Gautam Agarwal, Jack Cai, Collin Frink, Brian Hu, Eliot Kim, Shreyans Saraogi
UW Madison ProjectX Deliverable 1, September 27, 2020

---

1 Introduction and Research Question

Rising temperatures caused by anthropogenic climate change have increased the area burned by wildfires [21]. This trend has been emphasized by the devastating wildfires in western North America in the summer of 2020. One significant effect of this is the unhealthy air quality experienced by people far beyond the immediate region of the fire. Air pollution from the fires (such as smoke particles and ozone) has been linked to cardiovascular and respiratory diseases, thus creating health risks for said populations [13].

Therefore, reliable predictions of air quality are critical for protecting vulnerable populations by communicating timely advisories and warnings. In our study, we aim to improve upon existing deterministic methods of forecasting wildfire-induced air quality impacts by applying and assessing machine learning methods to predict air quality. Our initial region of focus will be Northern California, as this area has particularly been faced with devastating wildfires in recent years. These facets lead us to our research question: *How do various machine learning methods compare when predicting the air quality impact of wildfires in Northern California?*

2 Relevance to Climate Change and Potential Impact

There is significant evidence in the literature that indicates the burned area of wildfires today is exacerbated by anthropogenic climate change. For example, higher global temperatures due to greenhouse gas emissions are strongly linked to increases in the area burned by wildfires [5][7]. This is because higher temperatures reduce atmospheric moisture content, causing increased transpiration rates from plants and reducing precipitation amounts. Thus, the vegetation which feeds wildfires becomes drier as temperatures rise, enabling fires to draw on more plentiful and flammable fuel [1]. 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 [21]. Thus global warming, one of the prominent features of climate change, increases the potential spatial growth of wildfires.

As human activity continues to emit increasing amounts of greenhouse gases, wildfires will likely spread further in certain areas, exposing more people to the fires. Beyond the direct impact of the fires, distant populations will face the health impacts of higher concentrations and further spread of fine particulates and air pollutants produced by wildfires. 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 [18]. Fine particulate and air pollution exposure have long-term adverse health impacts, such as chronic cardiovascular and respiratory illnesses [13]. As air quality worsens due to more far-reaching wildfires, then, these health conditions will impact more people and become more severe. In order to mitigate the potential damage of poor air quality caused by wildfires, accurate forecasts of pollutant dispersion are critical.

Our research will focus on providing accurate and timely predictions of fine particulate and air pollutant concentrations after wildfire incidents. As wildfires become more frequent and intense, health officials and residents alike will benefit from a better understanding of the risk in air pollution caused by wildfires. This is because firstly, turbulence in the atmosphere carries smoke plumes across far distances and into unexpected regions. Secondly, in most cases, increases in harmful aerosols are undetectable by human senses, so smoke forecasts are essential to put out safety measures and warn the public. Similar to weather forecasting, we can only plan for worst-case scenarios if we know what is coming in a timely manner and with high confidence.

3 Machine Learning Justification and Existing Research

Both air pollution and fire modeling are data-intensive due to a large number of parameters and the amount of available data. Machine learning is capable of analyzing this data and finding meaningful patterns that humans and deterministic models cannot. With so many potentially impactful variables in this research question, machine learning will be a productive tool in creating an accurate and consistent model.

Currently, the methods involved in air quality prediction (with and without wildfire effects) are largely deterministic (equation-based). These models, however, are often inaccurate as they fail to account for the number and sheer complexity of all the relevant parameters. For example, a paper from 2017 attempted to model air quality using a probabilistic model. In this case, the model failed to account for key characteristics such as burned forest area, wind speeds, and biofuel, resulting in an uncertainty of nearly 50% [17].

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 found nearly 300 papers that all used machine learning algorithms to model different subdomains of the wildfire problem [9]. 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 [23], in general, all the models were relatively disconnected in their main focus, location studied, and the model used [6][10][15][20][22][23].

In our work, we aim to expand upon said existing research by applying new learning algorithms, introducing different air quality measures, and including more recent data. Most prior research in this field has emphasized more traditional machine learning models, such as random forests, generalized boosted regression, and multivariate linear regression [15][22][23][24], but deep learning applications are uncommon. 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 both machine and deep learning algorithms, we will be able to learn more about the relative performance of each type of model. Finally, we also plan to use meteorological and land cover data up to 2019, with a focus on including data more recent than that of the existing literature. This will ensure that our air quality forecasts are pertinent to present conditions.

4 Methods and Datasets

We plan to derive four general types of inputs – terrain, meteorological, air quality, and fire-related – from the following datasets. For terrain features and characteristics, the USGS Landsat provides extremely precise spatial resolution data, but only once every 16 days [16]. For meteorological data, we plan to consider several sources: North American Regional Reanalysis (NARR) data, PRISM, and DAYMET [4][12][14]. NARR provides relatively coarse spatial resolution data but offers all potentially relevant variables and easy comparison to satellite data. On the other hand, PRISM provides much higher spatial resolution data and is especially effective in mountainous areas, such as Northern California, our region of study. However, it provides a very limited set of variables that we can use. DAYMET is similar to PRISM in terms of spatial resolution, but is generally easier to use and provides more variables. However, it does not function as well with elevation changes, which could be of importance to our project. For air quality measures, we plan to use either the aerosol optical depth (AOD) or the Environmental Protection Agency's (EPA) Air Quality Index (AQI) [2][3]. AOD is measured by various satellites and has a high spatial resolution, but it is a proxy for pollutant concentration within the entire column of the atmosphere. AQI is measured by ground-level stations and is thus much more pertinent to humans, but is limited by the low spatial density of stations. Finally, for fire-related data, we plan to consider the satellites GOES-R, MODIS, and VIIRS as data sources [8][11][19]. As a geostationary satellite, GOES-R provides much more frequent updates, but at the cost of spatial resolution. On the other hand, MODIS and VIIRS provide daily updates, but at higher spatial resolution. We will evaluate these sources as we determine the time frame we want to predict over.

Once we have collected all the necessary data, we will preprocess this data so it can be applied to the following models: random forest (RF), boosted regression tree (BRT), recurrent neural network (RNN), convolutional neural network (CNN), and a reinforcement agent-based learning model (RL). RF and BRT are models that are prevalent in the existing literature on wildfire modeling. These models have had varying degrees of success, and we aim to expand upon this by comparing them to deep learning models. The output measure of our models will be either the AQI or AOD in our area of interest, Northern California. For each learning model, we will use a cross-validation method to evaluate the best training hyperparameters. Finally, for testing, we will compare our predictions to real measurements of air quality to evaluate their accuracy.

5 Domain Expertise

We have reached out to the following UW-Madison experts from climate change backgrounds: Professors Tracey Holloway, Ankur Desai, Jonathan Martin, and Tristan L'ecuyer, Dr. Feng He and James Kossin, and Mr. Gaurav Doshi. Through our discussions with them, we gained valuable insights on potential directions for research.

We selected our research question in large part due to our conversation with Professor Desai, during which he mentioned the abundance of air quality-related data and the high potential for machine learning applications in this area of study. In addition, they also helped us critically evaluate the impacts of our potential research. Specifically, Professor Holloway emphasized the expansive room for innovation in the field of forecasting wildfire-induced air quality impacts. Additionally, many of the experts elaborated on the challenges we may face while conducting research in the climate change "arena" and the best places to find datasets. Their advice regarding research logistics and impact analysis helped us make a final choice of our research topic, while their feedback on the relative merits of different satellite data products helped us narrow down our datasets.

6 Current and Future Work

Currently, we have a strong sense of the overall aim of the project, as well as how our project will innovate and expand upon previous work. This is a result of exploring the existing literature surrounding the connections between climate change and wildfires, machine learning modeling of wildfires, and connections between wildfires and air quality. We also looked into the existing models (both ML and deterministic) that predict different measures of air quality as a function of wildfire parameters. In regards to datasets, we have explored data from the MODIS Terra AOD product and the CMAQ air quality model among others, as multiple measurement sources will be necessary. While we haven't specifically chosen said datasets, we have concluded that there are enough relevant sources of data to make the project feasible.

One question we are yet to resolve is the selection of a suitable time window for the air quality predictions. If we decide to focus on a shorter time window, we can estimate local air quality around a wildfire incident much faster, enabling timely and high spatial resolution air quality forecasts for the most-affected populations. Alternatively, if we decide to focus on long-term effects, we can explore the high-level dispersion trajectories a few weeks into the future. We will most likely pursue the first option, but the second option may be a potential future addition to the project.

In order to determine the appropriate time window (and other more nuanced aspects of our research question), we will need to conduct further exploration of datasets and models. We plan to conduct further investigations of the datasets described in Section III. We will also study the implementation of convolutional neural networks and recurrent neural networks using various programming frameworks. Further review of previous literature applying deep learning techniques (RNN, CNN) to climate and weather-related phenomena will be conducted to determine appropriate methods of treating meteorological and land cover input variables.

We are confident that the scope of our project is attainable and that the data we need exists and is accessible. However, many more nuanced details remain flexible. Our dataset choices require further research, specifically in what features will suit our project best, and how compatible each dataset is with other datasets. We also have not completely finalized the list of learning algorithms or the specific evaluation techniques of the models.

7 Possible Alternative Research Directions

Initially, we started out with four broad climate change-related topics of interest: hurricanes, floods, flash droughts/heatwaves, and wildfires. Further investigation of literature, discussion, and possible machine learning applications narrowed our efforts to the latter two. While we have clearly focused on wildfires, some combination of flash droughts and heat waves does remain of interest.

More practical flexibility exists within the wildfire domain, however, as we have explored pursuing a number of different projects not relating to air quality prediction. For example, we considered using machine learning to optimize firefighting resource deployment based on population density, socioeconomic assets, and susceptibility to fires. We concluded that this area of study would be highly interesting and applicable, but the broad scope and variety of implicated variables would make achieving results in the given time frame difficult. This topic, however, led to another project of interest within the wildfire domain: using machine learning to predict fire spread. While we consider this as another potentially beneficial avenue of research, this field has been extensively explored, as many learning-based fire-spread models already exist [9]. Thus, a combination of further literature review and including more sophisticated learning algorithms would be necessary to determine potential areas for innovation.

In all, if our proposed project on air quality prediction doesn't work out, we would be comfortable with switching to one of the topics listed above. As of now, however, our air quality prediction problem seems the most interesting and promising. While some of the more nuanced details are still flexible within the scope of our current proposed project (such as adding or removing datasets, learning algorithms, etc), we feel confident in the project's purpose, direction, and overall scope.

References

[1] Abatzoglou, John Williams, Alton. (2016). Impact of anthropogenic climate change on wildfire across western US forests. Proceedings of the National Academy of Sciences. 113. 11770-11775. 10.1073/pnas.1607171113.

[2] Aerosol Optical Depth. National Aeronautics and Space Administration. Accessed from https://aeronet.gsfc.nasa.gov/new_web/Documents/Aerosol_Optical_Depth.pdf

[3] Air Data: Air Quality Data Collected at Outdoor Monitors Across the US. Environmental Protection Agency. Accessed from https://www.epa.gov/outdoor-air-quality-data

[4] DAYMET. Oak Ridge National Laboratory. Accessed from https://daymet.ornl.gov/

[5] 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

[6] 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.

[7] 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.

[8] GOES-R Series Satellites. National Oceanic and Atmospheric Association. Accessed from https://www.ncdc.noaa.gov/data-access/satellite-data/goes-r-series-satellitesFDC

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

[10] 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.

[11] MODIS. National Aeronautics and Space Administration. Accessed from https://modis.gsfc.nasa.gov/data/

[12] North American Regional Reanalysis. National Oceanic and Atmospheric Association. Accessed from https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/north-american-regional-reanalysis-narr

[13] 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

[14] PRISM Climate Group. Northwest Alliance for Computational Science and Engineering. Accessed from https://prism.oregonstate.edu/normals/

[15] 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.

[16] Remote Land Sensing and Landsat. USGS. Accessed from https://www.usgs.gov/products/data-and-tools/real-time-data/remote-land-sensing-and-landsat

[17] 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

[18] 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.

[19] VIIRS NASA Active Fire Product. National Aeronautics and Space Administration. Accessed from https://viirsland.gsfc.nasa.gov/Products/NASA/FireESDR.html

[20] 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.

[21] 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

[22] 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.

[23] 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.

[24] 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.