Using machine learning to estimate wildfire $PM_{2.5}$ at California ZIP codes (2006-2020)

Rosana Aguilera[1], Nana Luo[1], Rupa Basu[2], Jun Wu[3], Alexander Gershunov[1], Tarik Benmarhnia[1,4]

[1] Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA

[2] Cal EPA/OEHHA, Oakland, CA, USA

[3] Department of Environmental and Occupational Health, Program in Public Health, University of California, Irvine, CA, USA

[4] Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, La Jolla, CA, USA

Abstract

Epidemiological studies on the detrimental health impacts of exposure to fine particulate matter ($PM_{2.5}$) from different sources of emission can inform regulatory policy and identify vulnerable communities. Though $PM_{2.5}$ has decreased in the U.S. in the two past decades, the increasing frequency and severity of wildfires contribute to episodically impair air quality in wildfire-prone regions and beyond. Monitoring air quality extensively is challenging. Since government-operated monitors are sparsely located across California and the U.S., several regions and populations remain unmonitored. Current approaches to estimate $PM_{2.5}$ concentrations in unmonitored areas often rely on gathering large amounts of data, such as satellite-derived aerosol properties and meteorological variables. and direct use of low-cost air sensor measurements that may be associated with substantial uncertainty Furthermore, modelling wildfire-specific $PM_{2.5}$ is often based on chemical transport model predictions, which results in highly computationally intensive efforts. Our study used an ensemble model that integrated multiple machine learning algorithms and a large set of predictor variables to estimate daily $PM_{2.5}$ at the ZIP code level, a relevant spatio-temporal resolution for epidemiological and public health studies. Our models achieved comparable results to previous machine learning studies for $PM_{2.5}$ prediction, but avoided processing larger, computationally

intensive datasets. In addition, we use machine learning to estimate the wildfire-specific $PM_{2.5}$ concentrations through a novel multiple imputation approach.

Introduction

Exposure to fine particulate matter with aerodynamic diameter smaller than 2.5 µm ($PM_{2.5}$) is associated with a range of acute and chronic adverse health effects (Xing et al., 2016; Pope et al., 2006), including increased risk of mortality and hospitalization. Wildfire smoke is a source of $PM_{2.5}$ air pollution, with potential differential impacts on respiratory health when compared to ambient pollution (Wegesser et al., 2009, Aguilera et al., 2021). $PM_{2.5}$ in the United States has decreased since the early 2000s due to environmental regulations (Schwarzman et al., 2021), with the exception of wildfire-prone areas like the western United States (McClure & Jaffe, 2018). In this region, wildfires have been increasing in severity and frequency (Westerling and Bryant, 2007; Goss et al., 2020) impacting $PM_{2.5}$ levels (McClure & Jaffe, 2018), and this trend is predicted to continue (Ford et al., 2018; Neumann et al., 2021).

Smoke $PM_{2.5}$ has been linked to respiratory health impacts and higher hospitalization rates (Gan et al., 2017; Liu et al., 2015; Reid et al., 2016; Gan et al., 2017; Liu et al., 2017). However, quantifying the extent and variety of health impacts due to wildfire smoke is challenging due to the episodic nature of these events, as well as data and methodological limitations that hinder the accurate estimation of exposure (Liu et al., 2015). Studies that isolate the $PM_{2.5}$ concentrations attributable to wildfire smoke to study the effects on increased respiratory hospitalizations are scarce (Liu et al., 2017; Aguilera et al., 2021).

Accurate estimation of $PM_{2.5}$ exposures at a high spatiotemporal resolution is important for evaluating its health effects, particularly at small temporal (days to weeks) and spatial (neighborhood) scales. Although many countries have a substantial network of regulatory $PM_{2.5}$ monitoring stations that are routinely operated by government agencies, their spatial coverage is still very limited in terms of accurately representing population exposures, especially in regions of the world that have complex spatiotemporal variability in emissions, topography, meteorology, land-use and population density, such as the state of California (CA) in the United States (US) (Lee, 2019; Liu et al., 2009). Therefore, studies based only on $PM_{2.5}$ measured from the regulatory monitors would inevitably exclude many communities, including those mostly exposed to wildfire smoke.

One approach to resolving this issue has been developing models to predict local $PM_{2.5}$ based on satellite, meteorological, and land use data. This process typically involves developing a prediction model that relies on large amounts of input data and it is highly computationally intensive to predict air pollution levelsin unmonitored areas. Various approaches have been proposed to model $PM_{2.5}$ in the recent decade, with satellite-derived aerosol optical depth, land-use variables, chemical transport model output, and several meteorological variables as major predictor variables. In addition, some researchers have combined spatiotemporal data sets to perform sophisticated modeling of $PM_{2.5}$ exposure from wildfire smoke using data-adaptive machine learning coupled with empirical and deterministic model output (Fadaru et al., 2020). Each approach has its strengths and weaknesses, which affect the interpretation of study findings and the translation of research to public health practice (Fadaru et al., 2020).

Most of these studies estimate $PM_{2.5}$ at a resolution of 1 km x 1 km grid cell resolution to provide fine spatial granularity (Di et al., 2019; Lee, 2019; Li et al., 2020). One of the main challenges of this approach is the differing spatial resolution of available datasets, making necessitating the implementation of downscaling methods and similar steps to prepare predictor datasets at a comparable spatial scale. In addition, working with datasets of 1 km^2 cells comprising large areas, such as California, translates into issues with big data handling, storage, and computing capabilities. Since these issues can run into system limits, research on methods of air pollutants like $PM_{2.5}$ must consider the technical limitations in the existing methodologies. Furthermore, most of previous studies focused on estimating overall $PM_{2.5}$ concentrations, without distinguishing among sources of emission such as wildfire smoke and non-smoke sources. In addition to implementing approaches based on physical processes (e.g., chemical transport models), statistical approaches can also be employed to isolate wildfire-specific $PM_{2.5}$. In this paper, we propose using multiple imputation to estimate wildfire-specific $PM_{2.5}$ based on a counterfactual approach.

Our study used an ensemble model that integrated multiple machine learning algorithms and predictor variables to first estimate daily $PM_{2.5}$ at a ZIP code level, a relevant spatial resolution for public health and epidemiological studies. We then apply the multiple imputation approach, which uses machine learning to impute non-smoke $PM_{2.5}$ concentrations for ZIP code days categorized as exposed to wildfire smoke. Our study design allows for environmental health researchers to construct and train machine learning models capable of predicting $PM_{2.5}$ at specific locations, such as ZIP code population-weighted centroids, thus avoiding highly computationally intensive efforts of predicting into unmonitored gridded space in large regions where no people live. Furthermore, we expanded this approach to isolate wildfire-specific $PM_{2.5}$ in California for the 2006-2020 period.

Materials and Methods

The data used in the estimation of daily, ZIP code level $PM_{2.5}$ using machine learning techniques covered the period 2006-2019 and are described below. A brief summary of continuous variables used is included in Table 1. Satellite-derived data were pre-processed using the Google Earth Engine (GEE) API (Google Earth Engine Team, 2015). GEE makes it possible to rapidly process vast amounts of satellite imagery at large scale with the power of cloud computing (Gorelick et al., 2017). In addition, we utilize H2O (Cook, 2016), an open-source big data platform, to achieve higher performance and reduce processing time in our analysis using R (version 4.0.3; R Core Team, 2020). Specifically, training and data processing is done in the high-performance H2O cluster rather than in R memory on a local computer.

$PM_{2.5}$ measurements

We used in situ daily $PM_{2.5}$ measurements (2006–2019) from the United States Environmental Protection Agency (EPA) Air Quality System (AQS) (https://www.epa.gov/aqs) that were collected by state, local, and tribal air pollution control agencies. The AQS $PM_{2.5}$ network includes both continuous monitoring and 24-hour sampling on a 1-in-6 day, 1-in-3 day and everyday schedule. Measurements were taken from sites using the federal reference method (FRM) (EPA parameter code 88101) and acceptable non-FRM methods (EPA parameter code 88502) in California monitoring sites (n = 219; location shown in Figure S1 in Sup. Info) and we used 24 hr averaged data.

Aerosol Optical Depth

Aerosol Optical Depth (AOD) is a satellite-derived parameter measuring the degree to which suspended particles affect the transmission of light by absorption or scattering. Therefore, it is an indirect measure of the particles present in the column of air on a given time. AOD can be used to fill spatial gaps but does not distinguish surface-level aerosols. The Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm has been recently developed to retrieve AOD measurements from raw Moderate Resolution Imaging Spectroradiometer (MODIS) data at 1 km×1 km resolution (Lyapustin et al., 2018). As an advanced algorithm, MAIAC leverages a spatial and temporal algorithms to simultaneously retrieve atmospheric aerosols and bidirectional reflectance from MODIS data. MAIAC further detects clouds and corrects atmospheric effects over both dark vegetated surfaces and bright desert targets to obtain better daily AOD values at a higher spatial resolution (1 km × 1 km) (Lyapustin et al., 2018). The algorithm is also tuned to reduce masking of wildfire smoke as clouds (Lyapustin et al., 2012). Since absorption optical depth of aerosol species varies with wavelength (Bergstrom et al., 2007), AOD measurements at different wavelengths can account for representing different chemical compositions of $PM_{2.5}$ and thus be potentially helpful to achieve accurate modeling. We therefore included AOD measurements at 470 nm and 550 nm from both the Aqua and Terra satellites.

Plume Height

The recently developed MAIAC algorithm (Lyapustin et al., 2018) offers a unique tool for smoke detection and characterization. Plume height is reported near detected fire hot spots when the smoke plume is optically thick and exhibits a brightness temperature contrast with an unobscured neighboring land surface. Plume height observations may provide constraints on the vertical distribution of smoke and its impact on surface concentrations (Cheeseman et al., 2020). MAIAC retrieval only provides a single plume height altitude, regardless of the plume vertical depth (Lyapustin et al., 2019).

Meteorological Variables

Meteorological conditions such as precipitation, minimum and maximum temperature, surface shortwave radiation, specific humidity, and wind speed and wind direction were extracted from the high-resolution Gridded Surface Meteorological dataset (gridMET; Abatzoglou, 2013). The gridMET dataset blends the high resolution spatial data from PRISM with the high temporal resolution data from the National Land Data Assimilation System (NLDAS) to produce spatially and temporally continuous, complete, high-resolution (1/24th degree ~4-km) gridded dataset of surface meteorological variables across the contiguous United States.

Land-use variables

Land-use variables are proxies for local emissions and air pollution levels. Land-use variables approximate emission of air pollutants, often at kilometer or sub-kilometer scale. We prepared (1) land-use coverage types, (2) distance to nearest highway, (3) distance to coastline, (4) elevation, and (5) NDVI (normalized difference vegetation index), to capture the impact of emissions from neighboring areas.

Land cover variables, including forest cover and impervious surfaces, were retrieved from the National Land Cover Database (NLCD, https://catalog.data.gov/dataset/usgs-2011-nationallandcover). The spatial resolution of the NLCD coverage is 30 × 30 m^2 and data are available roughly every 3-5 years (2001, 2004, 2006, 2008, 2011, 2016). Since land-surface characteristics can be assumed to change gradually, we use simple linear interpolation to fill in missing values in gap years.

Distance to the nearest highway was computed using Caltrans - State Highway Network using geographic information system (GIS). Similarly, we estimated the distance from the California coastline with respect to the location of monitoring points and population-weighted ZIP code centroids. Elevation was derived from the 3-arc-second (90-meter) Shuttle Radar Topography Mission (SRTM) dataset distributed by USGS Earth Resources Observation and Science (EROS) Data Center (https://www.usgs.gov/centers/eros).

The NOAA Climate Data Record (CDR) of AVHRR Normalized Difference Vegetation Index (NDVI; Vermote et al., 2014) contains gridded daily NDVI derived from the NOAA AVHRR

Surface Reflectance product. It provides a measurement of surface vegetation coverage activity, gridded at a resolution of 0.05° and computed globally over land surfaces.

Smoke plumes

Smoke plumes were obtained from the NOAA Hazard Mapping System (HMS), available from September 2005 onward. The HMS product uses visible satellite imagery and trained satellite analyst skills to estimate the spatial extent of smoke, though it cannot discern whether a given plume is at ground level or higher in the atmosphere (Rolph et al., 2009). In addition, the HMS smoke-plume extent data has not been validated and could thus have systematic biases because discrimination of smoke can vary by region, season, and weather conditions (Brey et al., 2018). However, HMS smoke plumes remain a common binary metric used to determine if smoke is present in the atmospheric column on a given day (Lipner et al., 2019). The HMS smoke products are stored as polygon shapefiles representing the spatial extent of daily smoke plumes (ftp://satepsanone.nesdis.noaa.gov/volcano/FIRE/HMS_ARCHIVE/). A smoke binary variable was created by intersecting zip code polygons with smoke polygons, which was then used as an indication of daily exposure to wildfire $PM_{2.5}$.

Missing values

Missing values occur among predictor variables. To predict $PM_{2.5}$ concentration for the entire study area and during the entire study period, it is essential to fill in the missing values. We identified variables with no missing values, namely land-use types and meteorological variables, and used these as predictors in a random forest model to impute missing values for other variables such as AOD. We used the R Package missRanger (Mayer, 2019) to do fast missing value imputation by trained random forest. Using this method, each variable is imputed by predictions from a random forest using all other variables as covariates. The algorithm iterates multiple times over all variables until the average out-of-bag prediction error of the models stops to improve (Mayer, 2019).

Machine Learning for $PM_{2.5}$ estimation

We assembled daily values for response (measured PM$_{2.5}$) and explanatory variables for each of the air quality monitoring points (n = 219) available in California. Of the 1,766,053 resulting observations available, 60% (n = 1,059,965) was used for training our machine learning models and 20% (n = 353,043) for validation and prediction testing, each.

Table 1: Summary statistics of daily values for the response (PM$_{2.5}$) and explanatory variables at air quality monitoring locations

	Units	*Minimum*	*Maximum*	*Mean*	*Median*
PM$_{2.5}$	μg m^{-3}	0.1	557	10.7	8.6
Wind Speed	m s^{-1}	0.3	21.1	3.27	2.92
Maximum Temperature	K	261	325	297	297
Minimum Temperature	K	247	313	283	283
Precipitation	mm	0	271	1.07	0
Specific Humidity	kg kg^{-1}	0.00015	0.0247	0.00644	0.00630
Shortwave Radiation	W m^{-2}	1.90	391	229	235
Wind Direction	Degrees clockwise from North	0	360	167	240
NDVI	-	-1,000	9,603	2,292	2,249
AOD 470nm	-	0	4,000	174	153
AOD 550nm	-	6	2,960	129	113
Plume Height	m	0.0282	4,110	809	658

Base Learners

We used four base learners available within the H2O framework for machine learning: generalized linear models, deep learning, distributed random forest, and gradient boosting (Cook, 2016). Generalized Linear Models (GLM) estimate regression models for outcomes following exponential distributions; in our case, Gaussian (i.e. normal) distribution. H2O's Deep Learning is based on a multi-layer feedforward artificial neural network that can contain a large number of hidden layers of neurons. Distributed Random Forest (DRF) generates a forest of regression trees, rather than a single one. The regression algorithm takes the average prediction over all of their trees (more trees will reduce the variance) to make a final prediction. Gradient Boosting Machine (GBM) is a forward learning ensemble method. Within the H2O framework, GBM sequentially builds regression trees on all the features of the dataset in a fully distributed way - each tree is built in parallel. We trained these base learner models individually on all response ($PM_{2.5}$) and explanatory variables, with parameters of each machine learning algorithm selected manually. To avoid overfitting, we validated our models with 10-fold cross-validation.

Ensemble model

Stacking involves training a learning algorithm to combine the predictions of several other learning algorithms. First, all of the base learners are trained using the available data, then a combiner algorithm, the metalearner, is trained to make a final prediction using all the predictions of the other algorithms as additional inputs. Stacking typically yields a better performance than any single one of the trained models in the ensemble (Yang, 2017). We used stacking to ensemble the base learners described above to generate $PM_{2.5}$ predictions. We used the ensemble model to estimate daily $PM_{2.5}$ at the ZIP code level within 2006-2020 in California.

Once $PM_{2.5}$ estimates were obtained, we compared our estimates with those obtained by Di et al. (2019). We extracted the estimated concentrations at California ZIP code locations from the 1km x 1km dataset available online for years 2000-2016 (Di et al., 2019). The resulting effort is presented in Supplementary Information.

Wildfire $PM_{2.5}$ estimation

Multiple imputation approach

We used a multiple imputation approach to estimate $PM_{2.5}$ concentrations attributable to non-smoke sources in ZIP code/days identified as exposed to wildfire smoke by comparing observed $PM_{2.5}$ values to estimated counterfactual values in the absence of wildfire smoke. More specifically, we followed these steps: 1) We define the exposure to wildfire for a given ZIP code day if the smoke plume polygon intersects with the ZIP code polygon. In addition, we used the plume height presence as an indicator of exposure for a given ZIP code-day. 2) Based on the above exposure definition, we temporarily remove the ZIP code days exposed to wildfire smoke from our original $PM_{2.5}$ dataset. 3) Using the multiple imputation approach via fast random forest, we impute the values of non-smoke $PM_{2.5}$ on all ZIP code days categorized as exposed to smoke. This step provided estimates of ambient $PM_{2.5}$ unrelated to wildfire smoke. 4) We then subtract all non-smoke $PM_{2.5}$ values from the original daily $PM_{2.5}$ concentrations to obtain the levels of $PM_{2.5}$ attributable to wildfire smoke in ZIP code days previously categorized as exposed.

Results

Base and Ensemble Models Fit

Of the four base learner algorithms, model performance was highest for random forest (DRF) and gradient boosting (GBM). Results and model fit metrics are presented in Tables 2 and 3. In terms of variables and their degree of importance in explaining $PM_{2.5}$ variation, wind velocity appeared to be the most important in both DRF and GBM models (Figures 1 and 2). Stacking both base learners produced better results, particularly for the training dataset (R^2 = 0.97, Table 4), with a prediction R^2 of 0.86. The ensemble model appears to underpredict $PM_{2.5}$ concentrations in some instances, as seen in the comparison between observed and predicted $PM_{2.5}$ in monitoring sites across California (Figure 3).

Table 2: Model Metrics for Distributed Random Forest

Distributed Random Forest (DRF)

Regression Model Metrics	Training*	Validation	Cross-validation**
MSE	12.22308	13.38268	12.74663
RMSE	3.496152	3.658234	3.570242
Mean Absolute Error (MAE)	1.758955	1.849317	1.794321
Mean Residual Deviance	12.22308	13.38268	12.74663
R-squared	**0.845501**	**0.8557143**	**0.8528437**

** Metrics reported on Out-Of-Bag training samples*

*** 10-fold cross-validation on training data*

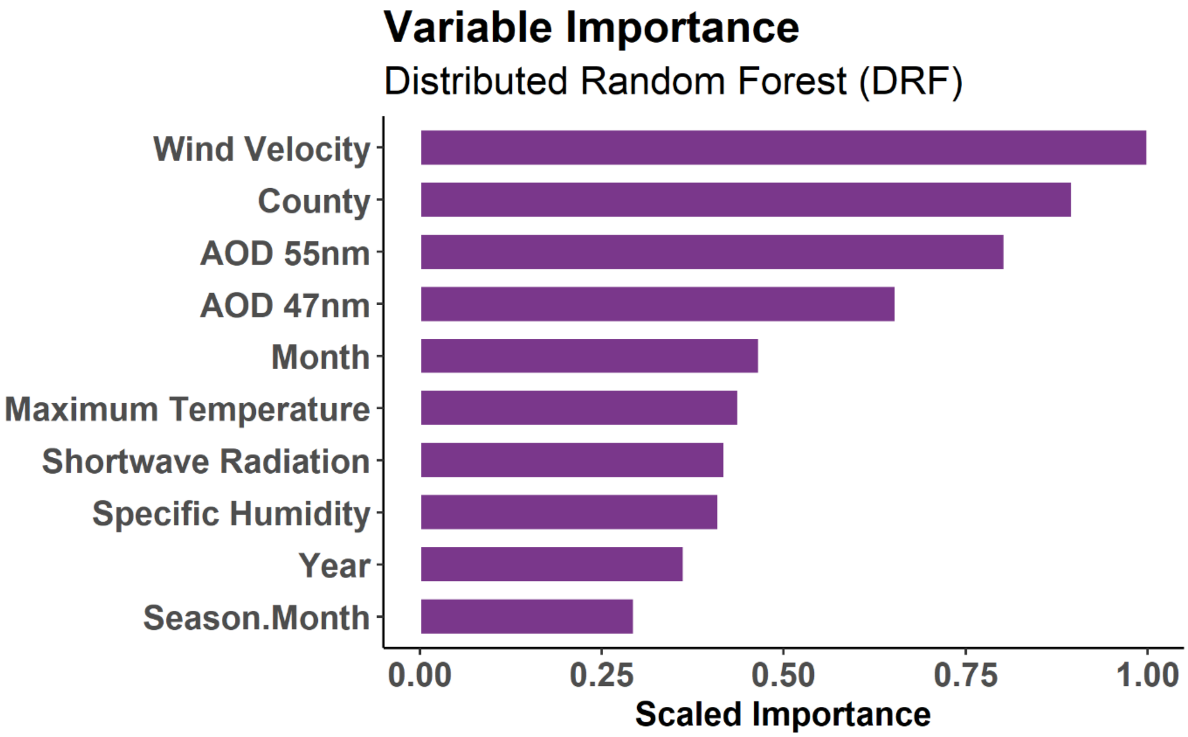


Figure 1: Variable Importance for the top 10 explanatory variables in the Distributed Random Forest (DRF) model

Table 3: Model Metrics for Gradient Boosting

Gradient Boosting (GBM)

Regression Model Metrics	Training	Validation	Cross-validation**
MSE	3.273814	13.56439	14.12381
RMSE	1.809368	3.682987	3.758166
MAE	0.8516281	1.905206	1.937965
Mean Residual Deviance	3.273814	13.56439	14.12381
R-squared	**0.9622047**	**0.8398809**	**0.8369445**

*** 10-fold cross-validation on training data*

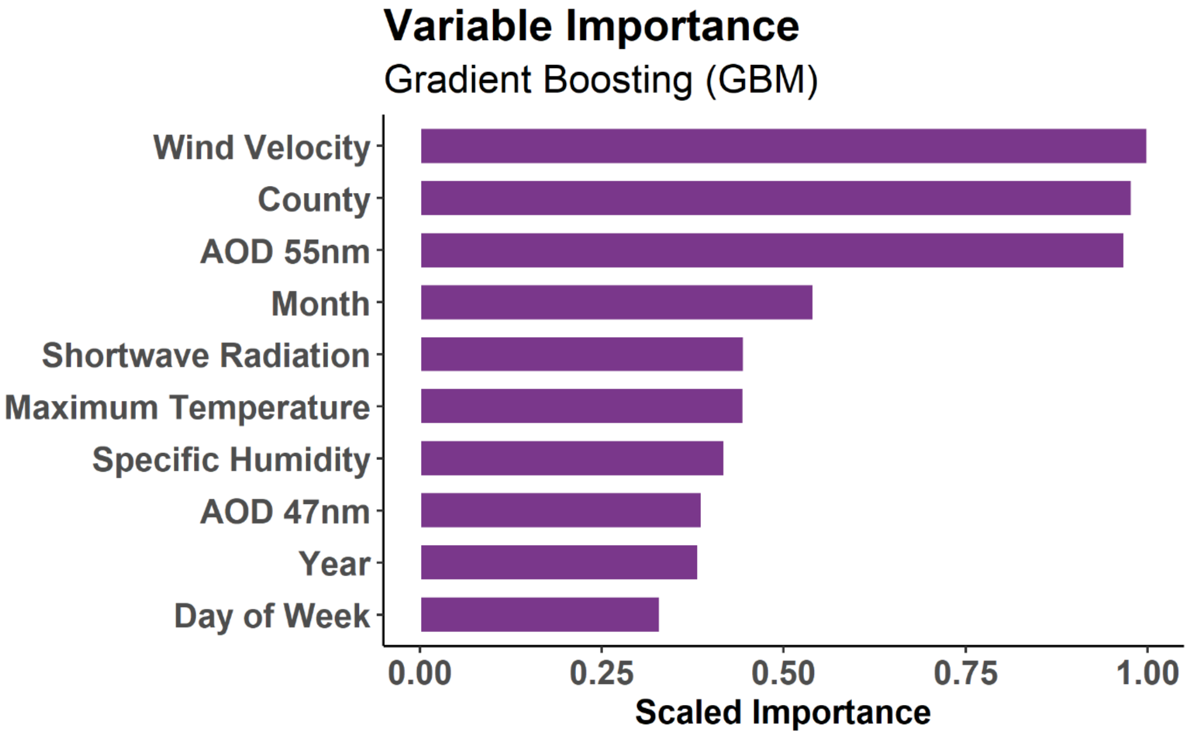


Figure 2: Variable Importance for the top 10 explanatory variables in the Gradient Boosting (GBM) model

Table 4: Model Metrics for Ensemble Model

Ensemble Model

Regression Model Metrics	Training	Validation	Prediction
MSE	2.651724	11.818	11.8096
RMSE	1.628411	3.437732	3.43651
Mean Absolute Error (MAE)	0.7703117	1.762147	1.752857
Mean Residual Deviance	2.651724	11.818	11.8096
R-squared	**0.9705874**	**0.8604959**	**0.8645956**

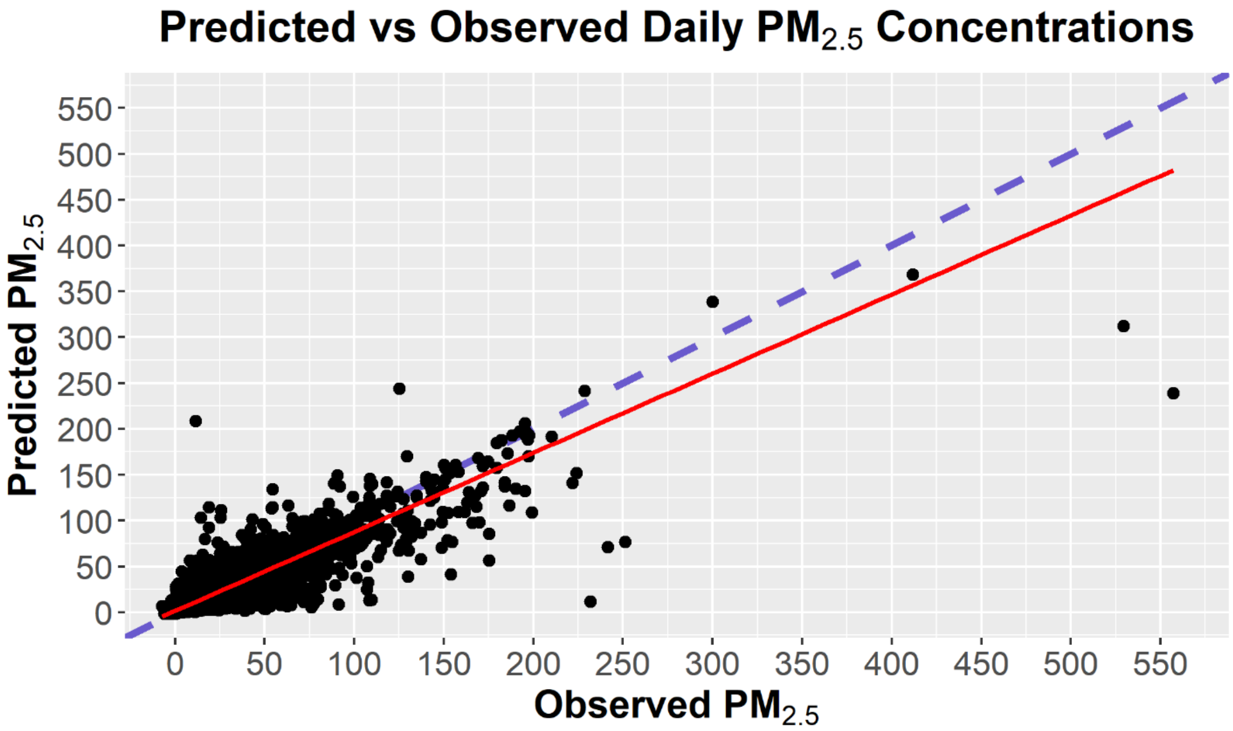


Figure 3: Observed versus Predicted $PM_{2.5}$ Concentrations at Monitoring Sites (R^2 = 0.86). Dashed blue line corresponds to the reference (1-to-1) line; red line is the linear model fit.

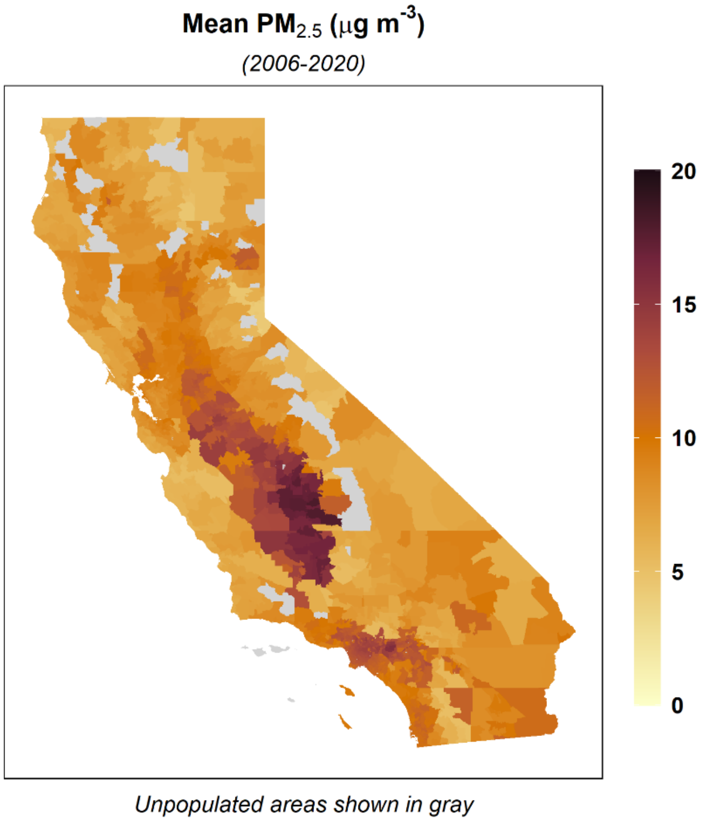
Predictions at the ZIP code level in California

Mean $PM_{2.5}$ concentrations predicted at the ZIP code level are shown in Figure 4. These averages over the 15-year study period (2006-2020) tend to be highest around the Central Valley region, as well as in highly populated areas in Southern California coastal ZIP codes. In Figure S2 (Sup. Info.), the highest $PM_{2.5}$ mean concentrations in the Central Valley occurred during Fall and Winter months. Table 5 shows seasonal differences in mean $PM_{2.5}$, with lowest mean (and maximum) concentration observed in Spring.

Table 5: Summary Statistics for $PM_{2.5}$ predictions (μg m^{-3}) at ZIP code centroids by season within 2006-2020

Season	*Mean*	*Minimum*	*Maximum*	*Median*	*IQR*
Fall	10.8	<1	328	9.21	5.72
Winter	11.0	<1	263	9.16	7.39

Spring	7.52	<1	86.0	6.86	4.41
Summer	10.1	<1	167	9.15	5.20


Mean $PM_{2.5}$ ($\mu g\ m^{-3}$)

(2006-2020)

Unpopulated areas shown in gray

Figure 4: Mean $PM_{2.5}$ concentrations at ZIP codes within the 2006-2020 study period, predicted by the ensemble model with predictors at the ZIP code level.

Comparing our $PM_{2.5}$ predictions to those obtained by Di et al. (2019) showed a few spatial differences. Figure S4 in Sup. Info. shows an R^2 of 0.6 for a linear model between these two datasets for a comparison of 2007 $PM_{2.5}$ concentrations at California ZIP codes. Spatially, higher R^2 values are observed in the Central Valley region and in more densely populated areas in coastal California ZIP codes (Figure S5). Overall mean concentrations appeared to be slightly higher in our models when compared to Di et al. (2019) (Figure S6). Perhaps the most interesting difference was that of maximum concentrations predicted in 2007, where our estimates showed the highest PM2.5 in ZIP codes affected by wildfires, whereas highest values for the Di et al. (2019) dataset were observed in Central California (Figure S7). Differences in methodology, predictors used and spatial scales between the two modeling efforts can account for the differences observed.

Wildfire-specific $PM_{2.5}$ at ZIP code level

As mentioned above, the ensemble model results tended, for the most part, to underpredict $PM_{2.5}$ concentrations (Figure 3). Figure 5 below shows the 15-year mean concentrations of wildfire-specific $PM_{2.5}$ estimated by the multiple imputation method. The highest mean concentrations are observed in Northern California. Concentrations for other wildfire-prone areas like Southern California (SoCal), where major firestorms and wildfire events occurred in 2007, 2008 and 2017, are lower than expected. Nonetheless, a closer look at the wildfire events in August and September in 2020, when the entire state was practically covered by smoke at some point in time, showed that wildfire-specific $PM_{2.5}$ were well represented spatially (Figure S3 in Sup. Info.). Table 6 shows a seasonal summary of wildfire-specific $PM_{2.5}$ over the 15-year period, with highest values observed during Fall months (September, October, November).

Table 6: Summary statistics for wildfire-specific $PM_{2.5}$ (µg m^{-3}) at ZIP code centroids by season within 2006-2020

Season	*Mean*	*Minimum*	*Maximum*	*Median*	*IQR*
Fall	8.84	<1	310	3.07	5.61
Winter	5.20	<1	254	2.55	3.88
Spring	1.82	<1	39.1	1.37	1.77
Summer	5.93	<1	158	2.40	4.12

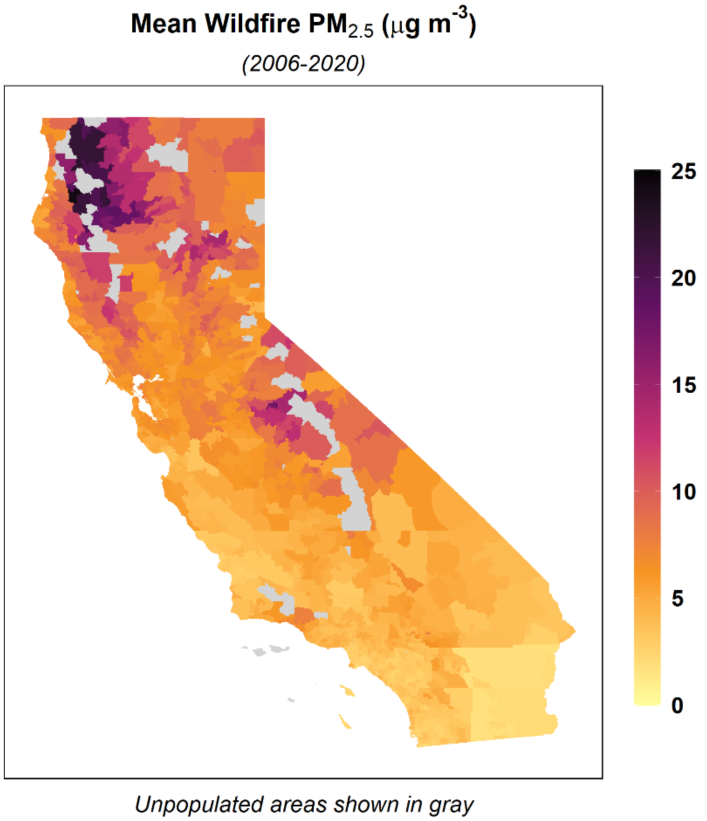


Figure 5: Mean wildfire-specific $PM_{2.5}$ concentrations at ZIP codes within the 2006-2020 study period, estimated via cubic spline imputation.

Discussion

Our final ensemble model incorporates $PM_{2.5}$ predictions from two machine learning algorithms, random forest and gradient boosting, achieving excellent predictive performance (R^2 of 0.86 and RMSE of 3.44 µg m^{-3}). The two machine learning algorithms used approximately 50 predictor variables, ranging from satellite-derived aerosol properties, land-use and meteorological data, with cross-validation controlling for overfitting. With the trained model, we predicted daily $PM_{2.5}$ within a 15-year period (2006-2020) at ZIP code population-weighted centroids in California. Daily, ZIP code level predictions indicated that our model was successful in capturing the spatial distribution and temporal peaks in wildfire-related $PM_{2.5}$.

Our ensemble model metrics above compare with previous efforts of $PM_{2.5}$ estimation in California (e.g., Li et al., 2020) and the US (Di et al., 2019) using a 1km x 1km grid for prediction.

For instance, Li et al., 2020 reported a prediction R^2 of 0.87 (RMSE = 2.29 μg m^{-3}) for weekly $PM_{2.5}$ concentrations in California within 2008-2017. Stowell et al., (2020), who focused on Southern California, demonstrated the usefulness of remote sensing products such as MAIAC AOD to achieve better exposure data in unmonitored regions. In fact, in our models, AOD was among the most important variables in explaining $PM_{2.5}$ variability.

Except for a few recent studies (Liu et al., 2017; Lipner et al., 2019; Aguilera et al., 2021; Sorensen et al., 2021; Heft-Neal et al., 2021), isolating wildfire-specific $PM_{2.5}$ is still an uncommon practice when estimating $PM_{2.5}$ exposure datasets. For instance, Li et al., 2020 looked at wildfire-related weekly concentrations of $PM_{2.5}$ in California and assessed their spatiotemporal patterns within their 10-year span study. These weekly concentrations included other sources of $PM_{2.5}$, in addition to wildfire smoke. However, since different sources of $PM_{2.5}$ might have differential impacts on human health (Wegesser et al., 2009; Ostro et al., 2016; Aguilera et al., 2021), it is important to isolate wildfire-specific concentrations from e.g., other sources of $PM_{2.5}$.

For the estimation of wildfire-specific concentrations, authors like Liu et al., (2017) relied on chemical transport models (CTM), which can be data and computationally intensive. Most studies mentioned above (i.e, Lipner et al., 2019; Aguilera et al., 2021; Sorensen et al., 2021; Heft-Neal et al., 2021) have relied on using HMS smoke plumes and seasonal background $PM_{2.5}$ to estimate wildfire-specific concentrations, among other similar methods. In our current study, which also uses HMS smoke plumes as an initial binary classification for exposure, we implemented a fast random forest algorithm for the imputation of background $PM_{2.5}$ on ZIP code days classified as exposed to wildfire smoke. In addition to $PM_{2.5}$ from all sources, our current efforts provide daily wildfire-specific $PM_{2.5}$ estimates for the entire region of California within a 15-year span, directly estimated at the location of population-weighted centroids of individual ZIP codes.

We acknowledge that our approach has limitations. For instance, the number and extent of smoke plumes used to categorize exposed ZIP code days represent a conservative estimate due to the limitations of visible satellite data. In addition to all the above, our definition of smoke exposure may have also misclassified some of the smoke $PM_{2.5}$ as non-smoke $PM_{2.5}$ and vice versa. Regarding our implementation of machine learning algorithms, we note that a limited number of these is currently implemented within the H2O framework. Thus, the reliance on H2O is also a limitation. Other non-supported algorithms such as extreme gradient boosting (XGBoost) would also be worth considering as it has demonstrating high predicting capabilities in other studies estimating $PM_{2.5}$ concentrations (e.g., Just et al., 2020). We also note that we did not differentiate other specific sources of $PM_{2.5}$ (e.g., traffic emissions, agricultural burns, prescribed fires, etc.) besides wildfire-specific concentrations. Moreover, though relevant in the study of impacts on public and environmental health, we do not consider the chemical speciation of $PM_{2.5}$, as data is scarce and it is also beyond the scope of this work.

Our statistical method can be generalized to other large heterogenous regions with high variability in emission sources, land-use, topography, meteorology and population growth. Using multisource data integrated into an ensemble machine learning framework allowed us to capture temporal and spatial trends over our study region, including days where wildfires were present, and isolating the wildfire-specific contribution as a source of $PM_{2.5}$ pollution in California ZIP codes.

Acknowledgements

This work was funded by the Office of Environmental Health Hazard Assessment #19-E0022. We thank Sara McElroy for helping with an introduction to Google Earth Engine.

References

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. International Journal of Climatology, 33(1), 121-131.

Aguilera, R., Corringham, T., Gershunov, A., & Benmarhnia, T. (2021). Wildfire smoke impacts respiratory health more than fine particles from other sources: observational evidence from Southern California. Nature communications, 12(1), 1-8.

Brey, S. J., Ruminski, M., Atwood, S. A. & Fischer, E. V. Connecting smoke plumes to sources using Hazard Mapping System (HMS) smoke and fire location data over North America, Atmos. Chem. Phys. 18, 1745 – 1761 (2018).

Cheeseman, M., Ford, B., Volckens, J., Lyapustin, A., & Pierce, J. R. (2020). The Relationship Between MAIAC Smoke Plume Heights and Surface PM. Geophysical Research Letters, 47(17), e2020GL088949.

Cook, D. (2016). Practical machine learning with H2O: powerful, scalable techniques for deep learning and AI. " O'Reilly Media, Inc.".

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., ... & Schwartz, J. (2019). An ensemble-based model of PM2. 5 concentration across the contiguous United States with high spatiotemporal resolution. Environment international, 130, 104909.

Fadadu, R. P., Balmes, J. R., & Holm, S. M. (2020). Differences in the Estimation of Wildfire-Associated Air Pollution by Satellite Mapping of Smoke Plumes and Ground-Level Monitoring. International Journal of Environmental Research and Public Health, 17(21), 8164.

Ford, B., Val Martin, M., Zelasky, S. E., Fischer, E. V., Anenberg, S. C., Heald, C. L., & Pierce, J. R. (2018). Future fire impacts on smoke concentrations, visibility, and health in the contiguous United States. GeoHealth, 2(8), 229-247.

Gan, R. W., Ford, B., Lassman, W., Pfister, G., Vaidyanathan, A., Fischer, E., ... & Magzamen, S. (2017). Comparison of wildfire smoke estimation methods and associations with cardiopulmonary-related hospital admissions. GeoHealth, 1(3), 122-136.

Google Earth Engine Team (2015). Google Earth Engine: A Planetary-scale Geospatial Analysis Platform. https://earthengine.google.com

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote sensing of Environment, 202, 18-27.

Goss, M., Swain, D. L., Abatzoglou, J. T., Sarhadi, A., Kolden, C. A., Williams, A. P., & Diffenbaugh, N. S. (2020). Climate change is increasing the likelihood of extreme autumn wildfire conditions across California. Environmental Research Letters, 15(9), 094016.
Lee, H.J., 2019. Benefits of high resolution PM2.5 prediction using satellite MAIAC AOD and land use regression for exposure assessment: California examples. Environ. Sci. Technol. 53, 12774–12783.

Just, A. C., Arfer, K. B., Rush, J., Dorman, M., Shtein, A., Lyapustin, A., & Kloog, I. (2020). Advancing methodologies for applying machine learning and evaluating spatiotemporal models of fine particulate matter (PM2.5) using satellite data over large regions. Atmospheric Environment, 239, 117649.

Li, L., Girguis, M., Lurmann, F., Pavlovic, N., McClure, C., Franklin, M., ... & Habre, R. (2020). Ensemble-based deep learning for estimating PM2. 5 over California with multisource big data including wildfire smoke. Environment International, 145, 106143.

Liu, Y., Paciorek, C.J., Koutrakis, P., 2009. Estimating regional spatial and temporal variability of PM2.5 concentrations using satellite data, meteorology, and land use information. Environ. Health Perspect. 117, 886–892. Livingston, J.M., Redemann, J., Shinozuka, Y., Johnson, R.,

Lipner, E. M., O'Dell, K., Brey, S. J., Ford, B., Pierce, J. R., Fischer, E. V., & Crooks, J. L. (2019). The associations between clinical respiratory outcomes and ambient wildfire smoke exposure among pediatric asthma patients at National Jewish Health, 2012–2015. GeoHealth, 3(6), 146-159.

Liu, J. C., Pereira, G., Uhl, S. A., Bravo, M. A. & Bell, M. L. (2015). A systematic review of the physical health impacts from non-occupational exposure to wildfire smoke. Environ. Res. 136, 120 ‑ 132.

Liu, J. C., Wilson, A., Mickley, L. J., Dominici, F., Ebisu, K., Wang, Y., ... & Bell, M. L. (2017). Wildfire-specific fine particulate matter and risk of hospital admissions in urban and rural counties. Epidemiology (Cambridge, Mass.), 28(1), 77.

Lyapustin, A., Korkin, S., Wang, Y., Quayle, B., and Laszlo, I. (2012). Discrimination of biomass burning smoke and clouds in MAIAC algorithm, Atmos. Chem. Phys., 12, 9679–9686.

Lyapustin, A., Wang, Y. (2018). MCD19A2 MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2G Global 1km SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 2021-02-22 from https://doi.org/10.5067/MODIS/MCD19A2.006

Lyapustin, A., Wang, Y., Korkin, S., Kahn, R., & Winker, D. (2019). MAIAC thermal technique for smoke injection height from MODIS. IEEE Geoscience and Remote Sensing Letters, 17(5), 730– 734.

Mayer, M. (2019), MissRanger: Fast Imputation of Missing Values, R package version 2.1.0

McClure, C. D. & Jaffe, D. A. (2018). US particulate matter air quality improves except in wildfire-prone areas. Proc. Natl Acad. Sci. USA 115, 7901 ‑ 7906.

Neumann, J. E., Amend, M., Anenberg, S., Kinney, P. L., Sarofim, M., Martinich, J., ... & Roman, H. (2021). Estimating PM2. 5-related premature mortality and morbidity associated with future wildfire emissions in the western US. Environmental Research Letters, 16(3), 035019.

Pope, C. A. III & Dockery, D. W. Health effects of fine particulate air pollution: lines that connect. J. Air Waste Manag. Assoc. 56, 709–742 (2006).

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reid, C. E., Brauer, M., Johnston, F. H., Jerrett, M., Balmes, J. R., & Elliott, C. T. (2016). Critical review of health impacts of wildfire smoke exposure. Environmental health perspectives, 124(9), 1334-1343.

Rolph, G. D., Draxler, R. R., Stein, A. F., Taylor, A., Ruminski, M. G., Kondragunta, S., ... & Davidson, P. M. (2009). Description and verification of the NOAA smoke forecasting system: the 2007 fire season. Weather and Forecasting, 24(2), 361-378.

Schwarzman, M., Schildroth, S., Bhetraratana, M., Alvarado, Á., & Balmes, J. (2021). Raising standards to lower diesel emissions. Science, 371(6536), 1314-1316.

Vermote, E.; Justice, C.; Csiszar, I.; Eidenshink, J.; Myneni, R. B.; Baret, F.; Masuoka, E.; Wolfe, R.E.; Claverie, M.; NOAA CDR Program. (2014): NOAA Climate Data Record (CDR) of Normalized Difference Vegetation Index (NDVI), Version 4. NOAA National Centers for Environmental Information. https://doi.org/10.7289/V5PZ56R6.

Wegesser, T. C., Pinkerton, K. E. & Last, J. A. (2009) California wildfires of 2008: coarse and fine particulate matter toxicity. Environ. Health Perspect. 117, 893‑897.

Westerling A L and Bryant B P 2007 Climate change and wildfire in California Clim. Change 87 231–49

Yang, Y. (2017). Ensemble learning. In temporal data mining via unsupervised ensemble learning (pp. 35-56). Elsevier.

Xing, Y. F., Xu, Y. H., Shi, M. H. & Lian, Y. X. The impact of PM2.5 on the human respiratory system. J. Thorac. Dis. 8, E69 (2016).