Title

A novel ensemble-based statistical approach to estimate daily wildfire-specific PM2.5 in California (2006-2020).

Authors

Clemesha, Rachel
Gershunov, Alexander
Benmarhnia, Tarik
et al.

Publication Date

2023

DOI

10.1016/j.envint.2022.107719

Peer reviewed

Published in final edited form as:
Environ Int. 2023 January ; 171: 107719. doi:10.1016/j.envint.2022.107719.

A novel ensemble-based statistical approach to estimate daily wildfire-specific $PM_{2.5}$ in California (2006–2020)

Rosana Aguilera[a,*], **Nana Luo**[a], **Rupa Basu**[b], **Jun Wu**[c], **Rachel Clemesha**[a], **Alexander Gershunov**[a], **Tarik Benmarhnia**[a]

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

[b]Office of Environmental Health Hazard Assessment, California Environmental Protection Agency, Oakland, CA, USA

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

Abstract

Though fine particulate matter ($PM_{2.5}$) has decreased in the United States (U.S.) in the past two decades, the increasing frequency, duration, and severity of wildfires significantly (though episodically) impairs air quality in wildfire-prone regions and beyond. Increasing $PM_{2.5}$ concentrations derived from wildfire smoke and associated impacts on public health require dedicated epidemiological studies. Main sources of $PM_{2.5}$ data are provided by government-operated monitors sparsely located across U.S., leaving several regions and potentially vulnerable populations unmonitored. Current approaches to estimate $PM_{2.5}$ concentrations in unmonitored areas often rely on big data, such as satellite-derived aerosol properties and meteorological variables, apply computationally-intensive deterministic modeling, and do not distinguish wildfire-specific $PM_{2.5}$ from other sources of emissions such as traffic and industrial sources. Furthermore, modelling wildfire-specific $PM_{2.5}$ presents a challenge since measurements of the smoke contribution to $PM_{2.5}$ pollution are not available. Here, we aim to use statistical methods to isolate wildfire-specific $PM_{2.5}$ from other sources of emissions. Our study presents an ensemble model that optimally combines multiple machine learning algorithms (including gradient boosting machine, random forest and deep learning), and a large set of explanatory variables to, first, estimate daily $PM_{2.5}$ concentrations at the ZIP code level, a relevant spatiotemporal resolution for epidemiological studies. Subsequently, we propose a novel implementation of an imputation

[*]Corresponding author at: Scripps Institution of Oceanography, University of California San Diego, 9500 Gilman Drive #0230, La Jolla, CA 92093, USA. r1aguilerabecker@ucsd.edu (R. Aguilera).

CRediT authorship contribution statement
Rosana Aguilera: Data curation, Methodology, Formal analysis, Writing – original draft. **Nana Luo:** Data curation, Writing – review & editing. **Rupa Basu:** Conceptualization, Writing – review & editing. **Jun Wu:** Methodology, Writing – review & editing. **Rachel Clemesha:** Data curation, Writing – review & editing. **Alexander Gershunov:** Conceptualization, Writing – review & editing. **Tarik Benmarhnia:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2022.107719.

approach to estimate the wildfire-specific $PM_{2.5}$ concentrations that could be applied geographical regions in the US or worldwide. Our ensemble model achieved comparable results to previous machine learning studies for $PM_{2.5}$ prediction while avoiding processing larger, computationally intensive datasets. Our study is the first to apply a suite of statistical models using readily available datasets to provide daily wildfire-specific $PM_{2.5}$ at a fine spatial scale for a 15-year period, thus providing a relevant spatiotemporal resolution and timely contribution for epidemiological studies.

Keywords

Air pollution; Wildfire; $PM_{2.5}$; Human health; Machine learning

1. Introduction

Exposure to fine particulate matter with aerodynamic diameter smaller than 2.5 μm ($PM_{2.5}$) is associated with a wide range of acute and chronic adverse health effects (Xing et al., 2016; Pope and Dockery, 2006), including increased risk of mortality and hospitalization. Although $PM_{2.5}$ has decreased in the United States (U.S.) in the two past decades due to stricter air quality policy, wildfires and associated smoke pollution in the western U.S. have contributed to poor air quality in wildfire-prone regions and beyond (Schwarzman et al., 2021; McClure & Jaffe, 2018). Wildfires are becoming more severe and frequent (Westerling and Bryant, 2007; Williams et al., 2019; Goss et al., 2020), impacting $PM_{2.5}$ levels (McClure & Jaffe, 2018) and this trend is predicted to continue in the context of climate change (Ford et al., 2018; Williams et al., 2019; Neumann et al., 2021).

Wildfire smoke and the resulting $PM_{2.5}$ air pollution detrimentally impact respiratory health, and evidence has shown that it might be more harmful than non-smoke $PM_{2.5}$ pollution (Wegesser et al., 2009, Aguilera et al., 2021). Wildfire $PM_{2.5}$ has been associated with respiratory disease impacts and high 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). Moreover, studies that isolate $PM_{2.5}$ concentrations attributable to wildfire smoke to study the effects on increased respiratory and other disease hospitalizations are scarce (Liu et al., 2017; Aguilera et al., 2021; Marlier et al., 2022) and are often limited to small regions and short time scales (e.g., Stowell et al., 2019; Cleland et al., 2021).

Accurate estimation of $PM_{2.5}$ exposures from different sources and at a high spatiotemporal resolution is critical for evaluating its health effects, particularly at small temporal (days to weeks) and spatial (neighborhood) scales. Although many regions in the U.S. and in the world 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 with complex spatiotemporal variability in emissions, topography, geography, meteorology, land-use and population density, such as the state of California (U.S.) (Lee, 2019; Liu et al., 2009).

Therefore, studies based only on $PM_{2.5}$ measured from regulatory monitors would inevitably exclude many communities, potentially those most exposed to wildfire smoke.

Various approaches have been proposed to model $PM_{2.5}$ (from any source) in the recent decade, mainly using Chemical Transport Models (CTMs), statistical models or a combination of the two approaches. Datasets used to predict $PM_{2.5}$ vary from satellite-derived aerosol optical depth, land-use variables, chemical transport models output, and relevant meteorological variables such as temperature and wind velocity (Di et al., 2019; Lee, 2019; Li et al., 2020; Reid et al., 2021; Yu et al. 2022). Some studies have combined spatiotemporal datasets to perform sophisticated modeling of $PM_{2.5}$ exposure to wildfire smoke using data-adaptive machine learning, coupled with output from empirical and deterministic models (Fadadu et al., 2020).

Recent studies typically estimate $PM_{2.5}$ at a 1 km × 1 km grid cell resolution to provide fine spatial granularity (Di et al., 2019; Lee, 2019; Li et al., 2020; Yu et al. 2022). However, several challenges arise when using gridded space. First, the differing spatial resolution of available datasets oftentimes makes necessary the implementation of downscaling methods and similar steps to prepare predictor datasets at a comparable spatial scale. Secondly, 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 that might not be available to most researchers and air quality policy makers. A recent study by Reid et al. (2021) provided estimates at the ZIP code level (among other spatial scales).ZIP code level estimation is appropriate for public health applications given that data records are usually associated with patients' address of residence (with ZIP code information as the smallest geographical scale). Nonetheless, research on methods of air pollutants like $PM_{2.5}$ involving large datasets and computationally-intensive deterministic models must consider the technical limitations in the existing methodologies.

In addition, most 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 $PM_{2.5}$. Accurately measuring the location and severity of wildfire $PM_{2.5}$ exposure is key for evaluating impacts on public health, but currently remains empirically challenging. For this reason, we propose a novel approach using spatiotemporal multiple imputation to estimate wildfire-specific $PM_{2.5}$ based on a counterfactual approach. Specifically, we used a suite of machine learning algorithms and readily available data to estimate daily wildfire-specific $PM_{2.5}$ at ZIP code level, a relevant spatial resolution for public health and epidemiological studies. We apply and test our methodology in California ZIP codes for the 2006–2020 period.

2. Materials and methods

Our study region is the state of California, located on the West Coast in the U.S. (Fig. 1). California coastal areas are, for the most part, densely populated, contrasting with inland regions. The data used in the estimation of daily, ZIP code level $PM_{2.5}$ using a set of machine learning techniques covered the period 2006–2020 and are described in detail in the subsequent sections. A summary of statistics for continuous variables used is included in Table S1 in Supplemental Material.

Satellite-derived data were pre-processed using the Google Earth Engine (GEE; 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). All values for response and explanatory variables (see below) were extracted at the location of $PM_{2.5}$ monitoring sites, and at population-weighted centroids for the estimation of ZIP code level $PM_{2.5}$.

Figure S1 (Supplemental Material) summarizes the main steps involved in estimating daily $PM_{2.5}$ concentrations at ZIP codes in California. Once we obtain the $PM_{2.5}$ concentrations from all sources, we isolate the wildfire-specific concentrations on ZIP codes and days in California exposed to smoke. Briefly, we apply a multiple imputation approach, which iteratively fits random forest models to impute non-smoke $PM_{2.5}$ concentrations for a given ZIP code and day categorized as exposed to wildfire smoke, and then subtract them from counterfactual $PM_{2.5}$ concentrations (i.e., what would be measured in the absence of wildfire smoke). All methods used in our approach are further described in the sections below.

2.1. Response variable: $PM_{2.5}$ measurements

We used in situ daily $PM_{2.5}$ measurements (2006–2020) 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 daily monitoring and 24-hour sampling on a 1-in-6 day, 1-in-3 day and everyday schedule. Measurements (n = 575,582) were taken from California monitoring sites (n = 219; locations shown in Fig. 1).

2.2. Explanatory variables

We included several potential explanatory variables for the estimation of $PM_{2.5}$ in our study region. Time-varying variables such as satellite-derived aerosol properties and fixed properties such as elevation are described in detail in the following sections. In addition, we included information for explanatory variables such as county, air basin within California (Fig. 1; https://ww2.arb.ca.gov/applications/emissions-air-basin), day of the week, month, season, and site latitude and longitude.

2.2.1. Aerosol optical depth—Aerosol Optical Depth (AOD), a satellite-derived parameter measuring the degree to which suspended particles affect the transmission of light, is an indirect measure of the particles present in a column of air at a given time. 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 and Wang, 2018). MAIAC leverages 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 high spatial resolution (1 km × 1 km) (Lyapustin and Wang, 2018). The algorithm is also tuned to reduce masking of wildfire smoke as clouds (Lyapustin et al., 2012) and we use the binary MODIS Quality Assurance (QA) flags to select the cloud-free data with optimal quality. Since absorption optical depth of aerosol species varies with wavelength (Bergstrom et al., 2007), AOD measurements at different wavelengths can account for 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. We used the average of these AOD measurements obtained by Terra (passing time roughly at 10:30 am, local time) and Aqua (1:30 pm, local time) satellites.

2.2.2. Meteorological variables—Meteorological conditions such as wind velocity and temperature can affect $PM_{2.5}$ concentrations (Tai et al., 2010; Chen et al. 2020). Precipitation, minimum and maximum temperatures, surface shortwave radiation, specific humidity, wind velocity and wind direction data were extracted from the high-resolution Gridded Surface Meteorological dataset (gridMET; Abatzoglou, 2013). The gridMET dataset blends the high-resolution spatial data from Parameter-elevation Regressions on Independent Slopes Model (PRISM) with the high temporal resolution data from the National Land Data Assimilation System (NLDAS) to produce a spatially and temporally continuous, complete, high-resolution (1/24th degree ~ 4-km) gridded dataset of daily surface meteorological variables across the contiguous US.

2.2.3. Land-use variables—Land-use variables are proxies for local emissions and background 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. Since land-surface characteristics can be assumed to change gradually, missing values in gap years are replaced by linear interpolation between neighboring values (Verbesselt et al., 2010) using data available in the last 20 years (2001, 2004, 2006, 2008, 2011 and 2016). The linear interpolation of land cover data has been previously used in other modeling efforts (e.g., Di et al., 2019; Li et al., 2020).

Distance to the nearest highway was computed using Caltrans - State Highway Network using a geographic information system (GIS; ArcGIS Pro version 2.6.3.; ESRI, 2020). 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.

2.2.4. Thermal inversions—Radiosondes, routinely launched using helium balloons twice daily, provide a vertical profile of temperature measurements and are well suited for estimation of low-level temperature inversions. We used radiosonde measurements taken at both 0 UTC (4 pm local time) and 12 UTC (4 am), available at NOAA/ESRL Radiosonde Database (esrl.noaa.gov/raobs) three sites in California (locations shown in Fig. 1). By considering inversions that were present at both soundings on a given day, we excluded short-lived nighttime surfaced-based inversions produced by intense radiational cooling overnight (Gillies et al. 2010, Beard et.al 2012). Measurements at Oakland (Northern California) and Miramar (Southern California) are representative of coastal California and the Central Valley (as shown by Iacobellis et al., 2010), and the Edwards Airforce Base location was included to represent inland/desert regions. Inversions are stable features acting as lids on upward motion trapping pollution close to the surface. Daytime temperature inversions are typically produced by large-scale subsidence as the major ingredient and tend to be homogeneous over large areas (Iacobellis et al., 2010; Clemesha et al., 2017), so the sparse observational radiosonde network is largely adequate to resolve California inversions.

An inversion is defined to be present when a temperature at a given altitude in the sounding is warmer than the temperature at an altitude below it. The temperature profile is examined between the surface and 700 hPa level. Inversion strength (DT) is calculated as the temperature difference between the top and bottom of the inversion region (Iacobellis et al. 2010); Clemesha et al. 2017). In cases where more than one inversion is observed, the inversion having the largest value of DT is used (Clemesha et al., 2017). Inversion strength is an important factor determining pollution concentration levels in the Central Valley and in southern coastal California (Iacobellis et al., 2010). We also included the inversion base (ZBASE), which is the lowest elevation before warming begins (i.e., at the bottom of the inversion; Iacobellis et al. 2010; Clemesha et al. 2017).

2.2.5. Smoke plumes and wildfire data—Smoke plumes were obtained from the NOAA Hazard Mapping System (HMS; Ruminski et al., 2006), available in the region of North America from September 2005 onward (https://satepsanone.nesdis.noaa.gov/pub/FIRE/web/HMS/). 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. 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 $PM_{2.5}$.

We included additional wildfire related variables such as distance to and maximum area of the nearest wildfire perimeter within a 100 km radius. For this, we used wildfire perimeter polygons provided by the Fire and Resource Assessment Program (FRAP) of the California Department of Forestry and Fire Protection (CalFire; https://frap.fire.ca.gov/frap-projects/fire-perimeters/) and the fire points (i.e., thermal anomalies identified by trained analysts) from the above HMS dataset.

2.3. Missing values

Missing values occurred among both response and explanatory variables. To estimate $PM_{2.5}$ concentration at all ZIP codes in California and during the entire study period, it is essential to fill in the missing values. In the case of satellite-derived data, missing measurements mainly occur due to cloud cover on a given day. We identified explanatory variables with no missing values, namely land-use types and meteorological variables, and used these as predictors in a iterative and fast implementation of random forest models to impute missing values for other explanatory variables such as AOD and NDVI. We used the R Package missRanger (Mayer, 2019) to do fast missing value imputation by chained 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). We report the out-of-bag error for imputed variables in Supplementary Information (Table S2), as a measure of imputation (prediction) accuracy.

2.4. Machine learning for $PM_{2.5}$ estimation

We assembled daily values for response (observed $PM_{2.5}$) and explanatory variables for each of the air quality monitoring points (n = 219) available in California. We first set aside observations from 5 monitoring sites (selected to represent different areas of our study region; see Fig. 1) for independent testing of our model as a hold-out, test dataset to evaluate the model performance. Of the remaining observations, 80% was used for training our machine learning models and 20% for validation (i.e, optimization of the machine learning model parameters obtained during the training process, using 10-fold cross-validation). We ran three base learner algorithms (detailed below) and then stacked these into an ensemble model to generate estimates of $PM_{2.5}$ for prediction testing and to eventually estimate daily $PM_{2.5}$ concentrations at the ZIP code level.

2.4.1. Base learners

2.4.1. Base learners—To run the machine learning algorithms, we used H2O (Cook, 2016), an open-source big data platform, to achieve higher performance and reduce processing time in our analysis using R software (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.

We used three base learners available within the H2O framework for machine learning: deep learning, distributed random forest, and gradient boosting (Cook, 2016). Unlike linear regression, these algorithms are non-parametric and have no requirements concerning the form of the probability density function of the response variable·H2O's Deep Learning (DL) is based on a multi-layer, feedforward artificial neural network that is trained with stochastic gradient descent (iterative method for optimizing loss function) by updating weight using back-propagation (fine-tuning the weights of a neural net based on the error rate (i.e. loss) obtained in the previous epoch or iteration). The network may contain many hidden layers consisting of nodes, and there may be intermediate layers between the input and output layers. Variable importance for DL is estimated using the Gedeon method, which considers the weights connecting the input features to the first two hidden layers (Candel et al. 2016).

Distributed Random Forest (DRF; Breiman, 2001) generates a forest of de-correlated regression trees and then averages them for reducing the variance of an estimated prediction function. A bootstrap sample is chosen at random with replacement from the data. Some observations end up in the bootstrap sample more than once, while others are not included ("out-of-bag", OOB). The excluded OOB data are predicted from the bootstrap samples and by combining the OOB predictions from all trees. DRF allows estimation of the variable importance by calculating percentage increase in mean square error by shuffling the values of the OOB samples.

Gradient Boosting Machine (GBM; Friedman, 2001) is a forward learning ensemble method that uses a tree-based ensemble of weak models (decision trees). Whereas random forest builds an ensemble of deep independent trees, GBMs build an ensemble of shallow trees (weak learner or regression tree) in sequence with each tree learning and improving on the previous one. A gradient descent procedure is used to minimize the loss when adding trees. The general idea of gradient descent is to adjust parameter(s) iteratively to minimize a loss (or cost) function, the error between predicted values, and the actual values.

We trained these models individually on all response ($PM_{2.5}$) and explanatory variables, with optimal parameters of each machine learning algorithm selected by conducting a grid-search. We chose a cartesian grid search where we specified a set of values for each hyperparameter of interest and trained and validated each base learner for every combination of the hyperparameter values. Once the grid search was completed, we used MSE as performance metric to choose the optimal model hyperparameters (see Table S3 in Supplemental Information).

We validated our base learner models with k-fold cross-validation (k = 10), a standard method for estimating the performance of a machine learning algorithm on a dataset. In k-fold cross-validation, the original dataset is divided into groups of observations (or folds) of approximately equal size, whereas the first fold is treated as a validation set (i.e., the set that is held out), while the remainder constitutes the training set (James et al., 2014). The prediction error is estimated on the held-out folds.

2.4.2. Ensemble model—H2O's Stacked Ensemble method is a supervised ensemble machine learning algorithm that finds the optimal combination of a collection of prediction algorithms using a process called stacking. Unlike bagging and boosting, the goal in stacking is to ensemble strong, diverse sets of learners together. Specifically, stacking involves training a learning algorithm to combine the predictions of multiple learning algorithms. First, all base learner algorithms 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. It has been shown that stacking typically yields a better performance than any single one of the trained models in the ensemble (Yang, 2017). We used stacking to combine the base learners described above to generate $PM_{2.5}$ predictions. We then 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 from the ensemble model with $PM_{2.5}$ concentrations obtained by Di et al. (2019) and Reid et al., (2021). For this purpose, we extracted the estimated concentrations at California ZIP code locations from the 1 km × 1 km dataset available online for years 2000–2016 (Di et al., 2019). Reid et al. (2021) estimated $PM_{2.5}$ concentrations at ZIP code level, using the location of population-weighted centroids as we did in our study.

2.5. Wildfire $PM_{2.5}$ estimation

After estimating daily $PM_{2.5}$ from any source at the ZIP code level within 2006–2020, we used a multiple imputation approach to estimate counterfactual $PM_{2.5}$ concentrations that would have been observed in the absence of wildfire smoke, and then compared our observed values to estimated counterfactual $PM_{2.5}$ concentrations to estimate wildfire smoke specific $PM_{2.5}$.

More specifically, we followed these steps (summarized in Fig. 2): 1) We define the exposure to wildfire for a given ZIP code day if the smoke plume polygon intersects with the ZIP code polygon (i.e., if the ZIP code was covered by smoke on that 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 imputed the values of non-smoke $PM_{2.5}$ on all ZIP code days categorized as exposed to smoke. Briefly, the algorithm is based on an iterative imputation approach that uses a fast implementation of random forest to fill missing values, in this case, the non-smoke PM2.5 concentrations on ZIP code days categorized as exposed to wildfire smoke. Thus, this step provided estimates of background $PM_{2.5}$ unrelated to wildfire smoke contribution. We impute the non-smoke concentrations by a) using only the PM2.5 concentrations available for unexposed ZIP code days, and b) using $PM_{2.5}$. As before with covariates such as day of the week, month of the year and year. The out-of-bag prediction error resulting from the imputation approach acts as an indication of imputation accuracy. 4) Finally, we then 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.

Since there is no gold-standard dataset for the validation of wildfire-specific $PM_{2.5}$ concentrations, we relied on the HMS smoke products described in Section 2.2.5. We used the smoke density classification reported for HMS smoke plumes starting in 2010 and plotted the distribution of our resulting wildfire-specific $PM_{2.5}$ concentrations that corresponded to the ZIP code days impacted by the three classes of smoke density: light, medium and heavy.

3. Results

3.1. Base learners and ensemble model performance

Model performance was relatively similar for all three base learners: deep learning (DL), random forest (DRF) and gradient boosting (GBM). Resulting optimal hyperparameters using the grid search are detailed for each base learner in Table S3. Model fit metrics for the base learners are presented in Table S4 in Supplemental Information.

In terms of explanatory variables and their degree of importance in explaining $PM_{2.5}$ variation, wind velocity, inversion strength and aerosol optical depth (AOD) appeared to be among the most influential in both DRF and GBM models (Fig. 3). For the deep learning algorithm, wind direction appears as the most influential, followed by land cover variables. Both wind direction and velocity are expected to explain the variation in $PM_{2.5}$ transport from upwind and/or surrounding sources, e.g., wildfire or traffic emissions. Air temperature and thermal inversions can affect the formation of particles by promoting photochemical reactions between precursors, or by allowing the accumulation of unhealthy levels of $PM_{2.5}$ close to ground level. In addition, AOD indirectly measures aerosols in the atmospheric column and thus typically explains an important portion of the variation in $PM_{2.5}$ concentrations.

Stacking all three base learners in our ensemble model improved model prediction capabilities, with a prediction R^2 of 0.83 for all sites and 0.78 for the hold-out test dataset (Table 1). The ensemble model appears to underpredict high values (roughly, > 300 $\mu g\ m^{-3}$) of $PM_{2.5}$ concentrations, as seen in the comparison between observed and predicted $PM_{2.5}$ in monitoring sites across California (Fig. 4). However, like in most statistical approaches, it is expected to underpredict at very high values (Reid et al., 2021), particularly since these $PM_{2.5}$ concentrations might be associated with episodic events such as wildfires.

3.2. $PM_{2.5}$ estimated at the ZIP code level in California

Mean $PM_{2.5}$ concentrations from all sources estimated at the ZIP code level are shown in Fig. 5. 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. The highest non-smoke $PM_{2.5}$ median concentrations in the Central Valley, where agricultural activities are concentrated, occurred during Fall and Winter months (Figure S2 in Sup. Information).

We compared our $PM_{2.5}$ estimates derived from the ensemble model to those previously obtained by Di et al., (2019) and Reid et al., (2021). Figure S3 (displaying the correlation matrix between datasets; found in Sup. Information) shows that our estimates were most highly correlated with the values obtained by Di et al., 2019 (correlation of 0.78) and one of the efforts involving a random forest model by Reid et al., 2021 (correlation of 0.73). Differences in methodology, predictors used and spatial scales between the two modeling efforts can account for the differences observed.

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

Fig. 6 shows the 15-year mean concentrations of wildfire-specific $PM_{2.5}$ estimated by the multiple imputation method. The out-of-bag error in the univariate imputation step for the non-smoke $PM_{2.5}$, used in the estimation of wildfire-specific $PM_{2.5}$, was 0.007849 (i.e., error rate 0.78 %) when relying on only $PM_{2.5}$ concentrations for unexposed ZIP codes in the imputation. Comparable results were obtained when adding covariates such as day of the week, month of the year and year in the imputation approach. Results from this sensitivity analysis and imputation accuracy metrics (i.e., oob) is shown in Table S5.

The highest mean concentrations for wildfire-specific $PM_{2.5}$ are observed in Northern California, which were widely affected by extreme wildfire events during 2008 and the last four years in our study period. Concentrations for other wildfire-prone areas like Southern California (SoCal), where major wildfire events occurred in the fall of 2007 (also shown in Figure S4) and 2008 were lower than in the Northern counterpart. The high value in the Southern coast shown in Fig. 6 corresponds to an extreme and prolonged (month-long) wildfire event (the Thomas Fire) that occurred in December 2017 (wildfire $PM_{2.5}$ for this month shown in Figure S4). In addition, a closer look at the wildfire events in September 2020, when practically the entire state of California was covered by smoke for several days at a time, and November 2018 (Northern California) showed that wildfire-specific $PM_{2.5}$ were well represented spatially (Figure S4 in Sup. Information). When compared and related to the smoke density reported in the HMS datasets for smoke plumes on a given day, we observed that higher concentrations of wildfire PM2.5 were associated with heavy smoke density plumes, whereas the opposite occurred with smoke plumes with light smoke density (Figure S5). Lastly, Table S6 shows a summary of wildfire-specific $PM_{2.5}$ over the 15-year period, as well as the estimated daily $PM_{2.5}$ from non-smoke and all sources in California ZIP codes.

4. Discussion

Our final ensemble model incorporated $PM_{2.5}$ predictions from three machine learning algorithms, random forest, deep learning and gradient boosting, achieving excellent predictive performance (R^2 of 0.78 and RMSE of 3.51 $\mu g\ m^{-3}$). These machine learning algorithms used approximately 50 predictor variables, ranging from satellite-derived aerosol properties, land-use and meteorological data. With the trained model, we predicted daily all-sources $PM_{2.5}$ within a 15-year period (2006–2020) at ZIP code population-weighted centroids in California (n > 9 million observations). 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 U.S. (Di et al., 2019) using a 1 km × 1 km grid for prediction. For instance, Li et al., (2020) reported a prediction R^2 of 0.87 (RMSE = 2.29 $\mu g\ m^{-3}$) for weekly $PM_{2.5}$ concentrations in California within 2008–2017. Reid et al. (2021) reported prediction R^2 values ranging between 0.52 and 0.73 for $PM_{2.5}$ estimation in the Western U.S within 2008–2018. In addition, 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 studied 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 particularly important to isolate wildfire-specific concentrations from other sources of $PM_{2.5}$, such as traffic emissions, for prospective epidemiological studies.

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 and based on several assumptions. In addition, CTM model results tend to be limited to estimates at the county level (e.g., Liu et al., 2017), shorter time periods (e.g., months) or at non-daily temporal resolutions. Most studies mentioned above (i.e, Lipner et al., 2019; Aguilera et al., 2021; Sorensen et al., 2021; Heft-Neal et al., 2021) have used 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 classifier for exposure, we implemented a fast random forest algorithm for the imputation of background (non-smoke) $PM_{2.5}$ on given ZIP codes and 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 (e.g., cloud cover and consideration of the entire atmospheric column). In addition to all the above, our definition of smoke exposure may have caused misclassification, to a small extent, of smoke $PM_{2.5}$ as non-smoke $PM_{2.5}$ and vice versa. Related issues that can arise are the occurrence of negative values in the estimation of wildfire-specific $PM_{2.5}$, which can occur when most ZIP codes on a given day (or consecutive days) are covered by smoke. Though these cases are rare (in our case, only 1.5% of exposed ZIP code days were covered by an extent of 90% of more within the study region), we can suggest alternative approaches such as relying on seasonal patterns of background $PM_{2.5}$ or applying metrics such as robust differences as proposed by Bekbulat et al. (2021) for days with high percent coverage of smoke plumes in the area of interest.

Regarding our implementation of machine learning algorithms, we note that a limited number of these is currently implemented within the H2O framework (Cook, 2016). Thus, the reliance on H2O is also a limitation. Other non-supported algorithms in H2O such as extreme gradient boosting (XGBoost) would also be worth considering as they have demonstrated high predicting capabilities in other studies estimating $PM_{2.5}$ concentrations (e.g., Just et al., 2020). In relation to the suite of statistical methods used in our study, we acknowledge that the uncertainties from previous steps can cascade in the subsequent steps in our approach. However, it is worth noting that each model output in our approach is validated with the best available dataset and that model fit and performance metrics for the machine learning algorithms were comparable to similar studies in the literature.

Lastly, we also note that we did not differentiate other specific sources of $PM_{2.5}$ (e.g., traffic emissions, agricultural burns, prescribed forest fires, etc.) besides wildfire-specific concentrations, which could be addressed in future work. 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 such data is scarce, though this will be addressed in future work.

5. Conclusion

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. For this reason, it is imperative to isolate the contribution of wildfire smoke from other sources, such as traffic and industrial emissions. Our study design allows 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, mainly inhabited regions.

Our statistical approach can be generalized to other large heterogenous regions with high variability in emission sources, land-use, topography, meteorology, and population growth. Using multisource and readily available data integrated into an ensemble machine learning framework allowed us to capture temporal and spatial variability 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was funded by the Office of Environmental Health Hazard Assessment #19-E0022 and the National Institute of Aging #1RF1AG071024. We thank Sara McElroy for helping with an introduction to Google Earth Engine.

Data availability

Wildfire PM2.5 data are available at https://github.com/benmarhnia-lab/Wildfire_PM25_California_ZIP.

References

Abatzoglou JT, 2013. Development of gridded surface meteorological data for ecological applications and modelling. Int. J. Climatol 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. Nat. Commun 12 (1), 1–8. [PubMed: 33397941]

Bekbulat B, Apte JS, Millet DB, Robinson AL, Wells KC, Presto AA, Marshall JD, 2021. Changes in criteria air pollution levels in the US before, during, and after COVID-19 stay-at-home orders: Evidence from regulatory monitors. Sci Total Environ. 769, 144693 10.1016/j.scitotenv.2020.144693. [PubMed: 33736238]

Bergstrom RW, Pilewskie P, Russell PB, Redemann J, Bond TC, Quinn PK, Sierau B, 2007. Spectral absorption properties of atmospheric aerosols. Atmospheric Chemistry and Physics 7 (23), 5937–5943.

Breiman L, 2001. Random forests. Mach. Learn 45, 5–32.

Brey SJ, Ruminski M, Atwood SA, Fischer EV, 2018. Connecting smoke plumes to sources using Hazard Mapping System (HMS) smoke and fire location data over North America. Atmos. Chem. Phys 18, 1745–1761.

Candel A, Parmar V, LeDell E, Arora A, 2016. Deep learning with H2O. H2O. ai Inc, pp. 1–21.

Clemesha RES, Gershunov A, Iacobellis SF, Cayan DR, 2017. Daily Variability of California Coastal Low Cloudiness: A Balancing Act between Stability and Subsidence. Geophys. Res. Lett 44, 3330–3338. 10.1002/2017GL073075.

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

Chen Z, Chen D, Zhao C, Kwan MP, Cai J, Zhuang Y, Xu B, 2020. Influence of meteorological conditions on PM2. 5 concentrations across China: A review of methodology and mechanism. Environ. Int 139, 105558. [PubMed: 32278201]

Cleland SE, Serre ML, Rappold AG, West JJ, 2021. Estimating the acute health impacts of fire-originated PM2. 5 exposure during the 2017 California Wildfires: Sensitivity to choices of inputs. GeoHealth 5 (7) e2021GH000414.

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. Environ. Int 130, 104909. [PubMed: 31272018]

Esri, 2020. ArcGIS Pro: Release 2.6.3 Environmental Systems Research Institute, Redlands, CA.

Fadadu RP, Balmes JR, Holm SM, 2020. Differences in the Estimation of Wildfire-Associated Air Pollution by Satellite Mapping of Smoke Plumes and Ground-Level Monitoring. Int. J. Environ. Res. Public Health 17 (21), 8164. [PubMed: 33167314]

Ford B, Val Martin M, Zelasky SE, Fischer EV, Anenberg SC, Heald CL, Pierce JR, 2018. Future fire impacts on smoke concentrations, visibility, and health in the contiguous United States. GeoHealth 2 (8), 229–247. [PubMed: 32159016]

Friedman JH, 2001. Greedy function approximation: A gradient boosting machine. Ann. Stat 29, 1189–1232.

Gan RW, Ford B, Lassman W, Pfister G, Vaidyanathan A, Fischer E, Volckens J, Pierce JR, Magzamen S, 2017. Comparison of wildfire smoke estimation methods and associations with cardiopulmonary-related hospital admissions. GeoHealth 1 (3), 122–136. [PubMed: 28868515]

Gillies RR, Wang SY, Booth MR, 2010. Atmospheric scale interaction on wintertime intermountain west low-level inversions. Weather and Forecasting 25 (4), 1196–1210.

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 Sens. Environ 202, 18–27.

Goss M, Swain DL, Abatzoglou JT, Sarhadi A, Kolden CA, Williams AP, Diffenbaugh NS, 2020. Climate change is increasing the likelihood of extreme autumn wildfire conditions across California. Environ. Res. Lett 15 (9), 094016.

Heft-Neal S, Driscoll A, Yang W, Shaw G, Burke M, 2022. Associations between wildfire smoke exposure during pregnancy and risk of preterm birth in California. Environmental Research 203, 111872. [PubMed: 34403668]

Iacobellis SF, Cayan DR, Norris JR, Kanamitsu M, 2010. Impact of climate change on the frequency and intensity of low-level temperature inversions in California. Final Report to the California Air Resources Board Project 06–319. http://www.arb.ca.gov/research/apr/past/06-319.pdf.

James G, Witten D, Hastie T, Tibshirani R, 2014. An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, New York.

Just AC, Arfer KB, 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. Atmos. Environ 239, 117649.

Lee HJ, 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 (21), 12774–12783. [PubMed: 31566957]

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. Environ. Int 145, 106143. [PubMed: 32980736]

Liu Y, Paciorek CJ, 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 (6), 886–892. [PubMed: 19590678]

Lipner EM, O'Dell K, Brey SJ, Ford B, Pierce JR, Fischer EV, Crooks JL, 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. [PubMed: 32159037]

Liu JC, Pereira G, Uhl SA, Bravo MA, Bell ML, 2015. A systematic review of the physical health impacts from non-oecupational exposure to wildfire smoke. Environ. Res 136, 120–132. [PubMed: 25460628]

Liu JC, Wilson A, Mickley LJ, Dominici F, Ebisu K, Wang Y, … & Bell ML (2017). Wildfire-specific fine particulate matter and risk of hospital admissions in urban and rural counties. Epidemiology (Cambridge, Mass.), 28(1), 77. [PubMed: 27648592]

Lyapustin A, Korkin S, Wang Y, Quayle B, 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 . NASA EOSDIS Land Processes DAAC. Accessed 2021-02-22 from 10.5067/MODIS/MCD19A2.006.

Marlier ME, Crnosija N, & Benmarhnia T (2022). Wildfire smoke exposures and adult health outcomes. Preprint available at https://www.essoar.org/doi/abs/10.1002/essoar.l0510602.1. Accessed: May 1, 2022.

Mayer M, 2019. MissRanger: Fast Imputation of Missing Values. R package version 2(1).

McClure CD, Jaffe DA, 2018. US particulate matter air quality improves except in wildfire-prone areas. Proc. Natl Acad. Sci. USA 115 (31), 7901–7906. [PubMed: 30012611]

Neumann JE, Amend M, Anenberg S, Kinney PL, Sarofim M, Martinich J, Roman H, 2021. Estimating PM2. 5-related premature mortality and morbidity associated with future wildfire emissions in the western US. Environ. Res. Lett 16 (3), 035019.

Ostro B, Malig B, Hasheminassab S, Berger K, Chang E, Sioutas C, 2016. Associations of source-specific fine particulate matter with emergency department visits in California. American journal of epidemiology 184 (6), 450–459. [PubMed: 27605585]

Pope CA, Dockery DW, 2006. Health effects of fine particulate air pollution: lines that connect. J. Air Waste Manag. Assoc 56 (6), 709–742. [PubMed: 16805397]

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

Reid CE, Brauer M, Johnston FH, Jerrett M, Balmes JR, Elliott CT, 2016. Critical review of health impacts of wildfire smoke exposure. Environ. Health Perspect 124 (9), 1334–1343. [PubMed: 27082891]

Reid CE, Considine EM, Maestas MM, et al. , 2021. Daily PM2.5 concentration estimates by county, ZIP code, and census tract in 11 western states 2008–2018. Sci. Data 8, 112. 10.1038/s41597-021-00891-l. [PubMed: 33875665]

Rolph GD, Draxler RR, Stein AF, Taylor A, Ruminski MG, Kondragunta S, Zeng J, Huang H-C, Manikin G, McQueen JT, Davidson PM, 2009. Description and verification of the NOAA smoke forecasting system: the 2007 fire season. Weather Forecast. 24 (2), 361–378.

Ruminski M, Kondragunta S, Draxler R, & Zeng J (2006, May). Recent changes to the hazard mapping system. In Proceedings of the 15th International Emission Inventory Conference (Vol. 15, p. 18).

Schwarzman M, Schildroth S, Bhetraratana M, Alvarado Á, Balmes J, 2021. Raising standards to lower diesel emissions. Science 371 (6536), 1314–1316. [PubMed: 33766877]

Sorensen C, House JA, O'Dell K, Brey SJ, Ford B, Pierce JR, et al. , 2021. Associations between wildfirerelated PM2.5 and Intensive Care Unit admissions in the United States, 2006–2015. GeoHealth 5, e2021GH000385. 10.1029/2021GH000385.

Stowell JD, Geng G, Saikawa E, Chang HH, Fu J, Yang CE, Strickland MJ, 2019. Associations of wildfire smoke PM2. 5 exposure with cardiorespiratory events in Colorado 2011–2014. Environ. Int 133, 105151. [PubMed: 31520956]

Stowell JD, Bi J, Al-Hamdan MZ, Lee HJ, Lee SM, Freedman F, Liu Y, 2020. Estimating PM2. 5 in Southern California using satellite data: factors that affect model performance. Environ. Res. Lett 15 (9), 094004.

Tai AP, Mickley LJ, Jacob DJ, 2010. Correlations between fine particulate matter (PM2. 5) and meteorological variables in the United States: Implications for the sensitivity of PM2. 5 to climate change. Atmos. Environ 44 (32), 3976–3984.

Vermote E; Justice C; Csiszar I; Eidenshink J; Myneni RB; Baret F; Masuoka E; Wolfe RE; 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. 10.7289/V5PZ56R6.

Verbesselt J, Hyndman R, Newnham G, Culvenor D, 2010. Detecting trend and seasonal changes in satellite image time series. Remote Sens. Environ 114 (1), 106–115.

Wegesser TC, Pinkerton KE, Last JA, 2009. California wildfires of 2008: coarse and fine particulate matter toxicity. Environ. Health Perspect 117 (6), 893–897. [PubMed: 19590679]

Westerling AL, Bryant BP, 2007. Climate change and wildfire in California. Clim Change 87 (S1), 231–249.

Williams AP, Abatzoglou JT, Gershunov A, Guzman-Morales J, Bishop DA, Balch JK, Lettenmaier DP, 2019. Observedimpacts of anthropogenic climatechange on wildfire in California. Earth'sFuture 7 (8), 892–910.

Yang Y, 2017. Ensemble learning. In: In temporal data mining via unsupervised ensemble learning. Elsevier, pp. 35–56.

Yu W, Li S, Ye T, et al. , 2022. Deep Ensemble Machine Learning Framework for the Estimation of $PM_{2.5}$ Concentrations. Environ. Health Perspect 130 (3), 37004. [PubMed: 35254864]

Xing YF, Xu YH, Shi MH, Lian YX, 2016. The impact of PM2.5 on the human respiratory system. J. Thorac. Dis 8, E69. [PubMed: 26904255]

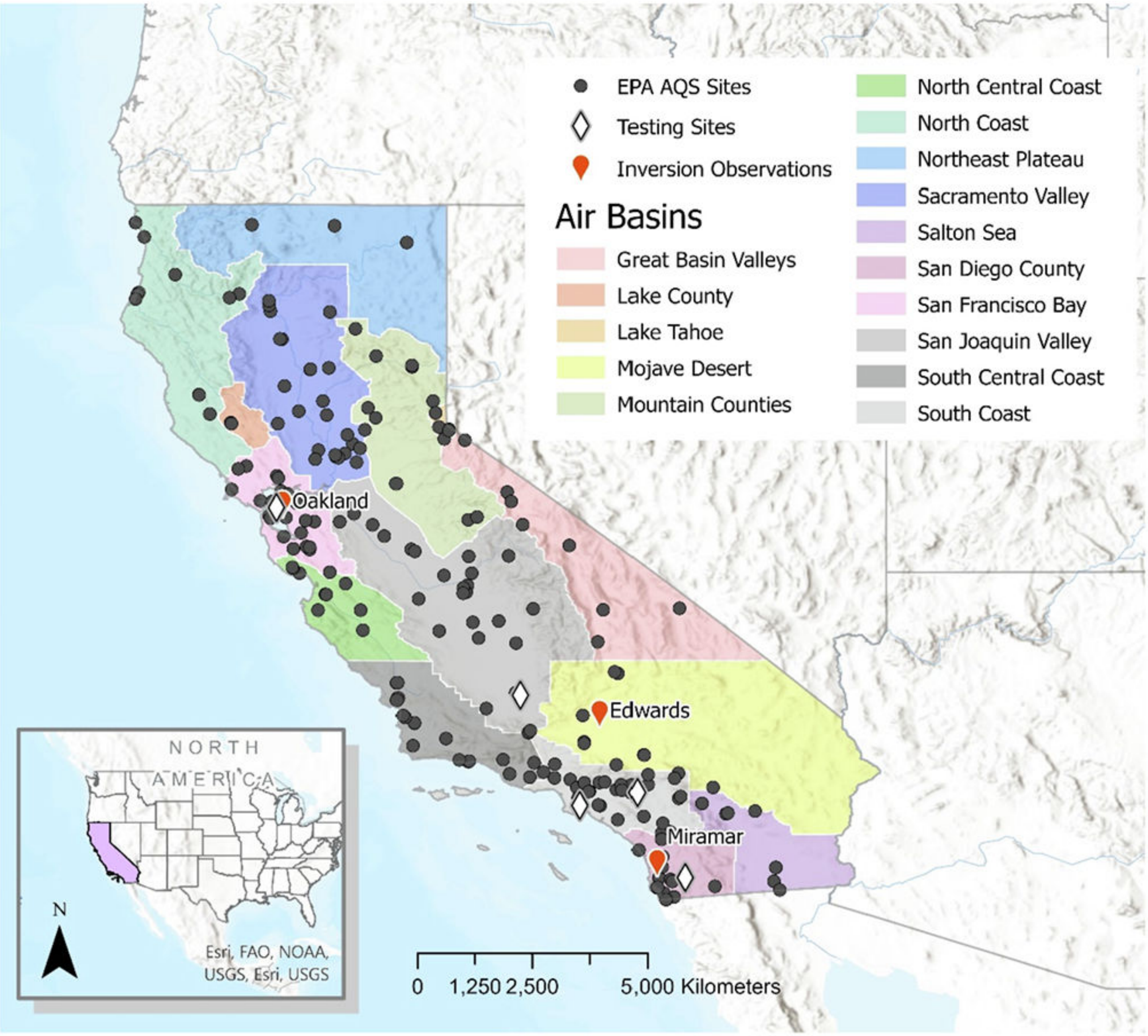
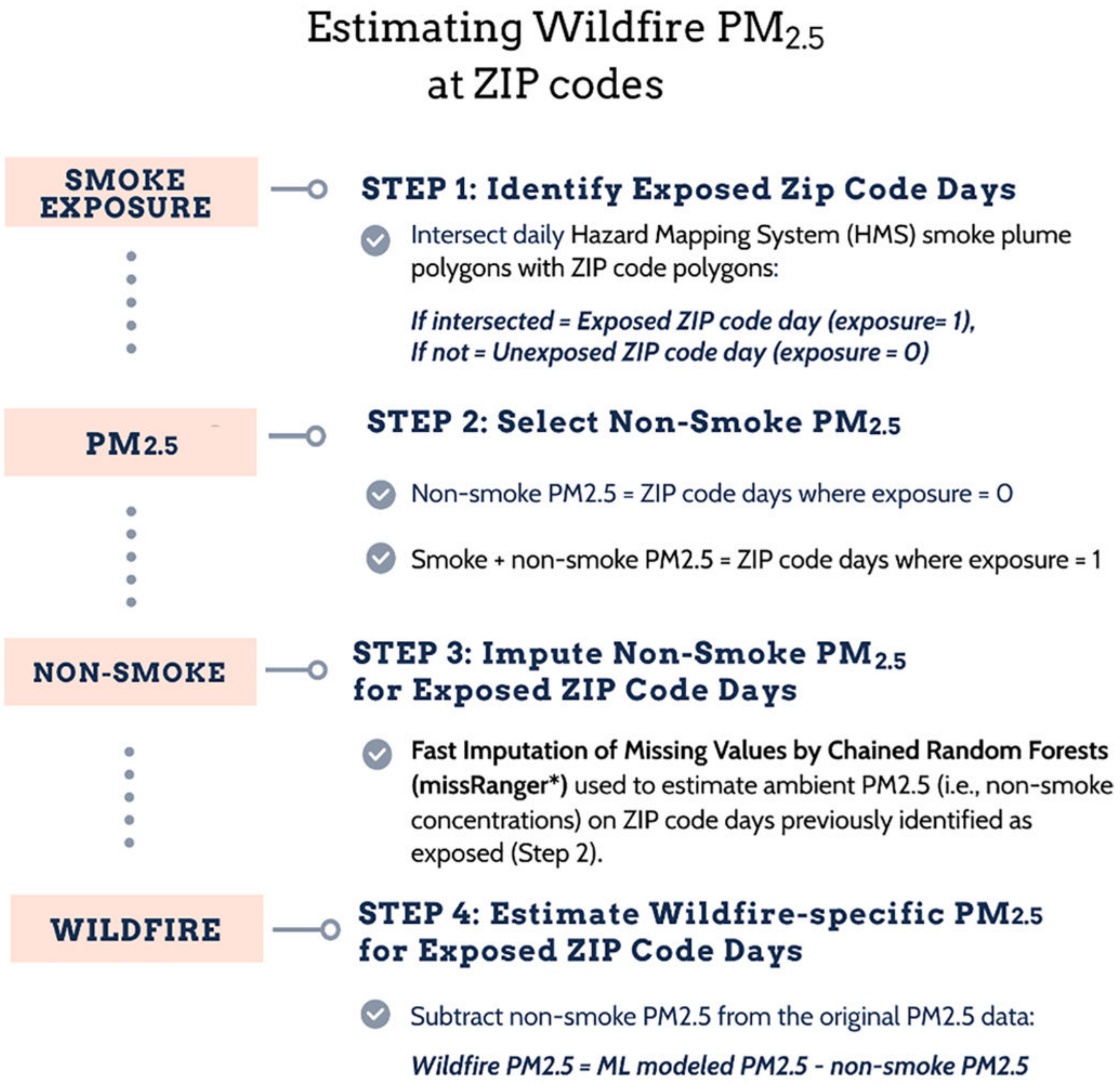


Fig. 1.

Study area (California, USA; see inlet figure) divided into 15 Air Basins by the California Air Resources Board. The Environmental Protection Agency's Air Quality System monitoring sites (EPA AQS Sites, grey dots) are also shown. Of these sites, 5 testing sites (shown here as white diamonds) were selected to evaluate the performance of our machine learning ensemble model. Locations (n = 3) of radiosonde measurements for inversion properties are also shown.



Estimating Wildfire $PM_{2.5}$ at ZIP codes

SMOKE EXPOSURE

STEP 1: Identify Exposed Zip Code Days

✓ Intersect daily Hazard Mapping System (HMS) smoke plume polygons with ZIP code polygons:

If intersected = Exposed ZIP code day (exposure= 1),
If not = Unexposed ZIP code day (exposure = 0)

$PM_{2.5}$

STEP 2: Select Non-Smoke $PM_{2.5}$

✓ Non-smoke PM2.5 = ZIP code days where exposure = 0

✓ Smoke + non-smoke PM2.5 = ZIP code days where exposure = 1

NON-SMOKE

STEP 3: Impute Non-Smoke $PM_{2.5}$ for Exposed ZIP Code Days

✓ **Fast Imputation of Missing Values by Chained Random Forests (missRanger*)** used to estimate ambient PM2.5 (i.e., non-smoke concentrations) on ZIP code days previously identified as exposed (Step 2).

WILDFIRE

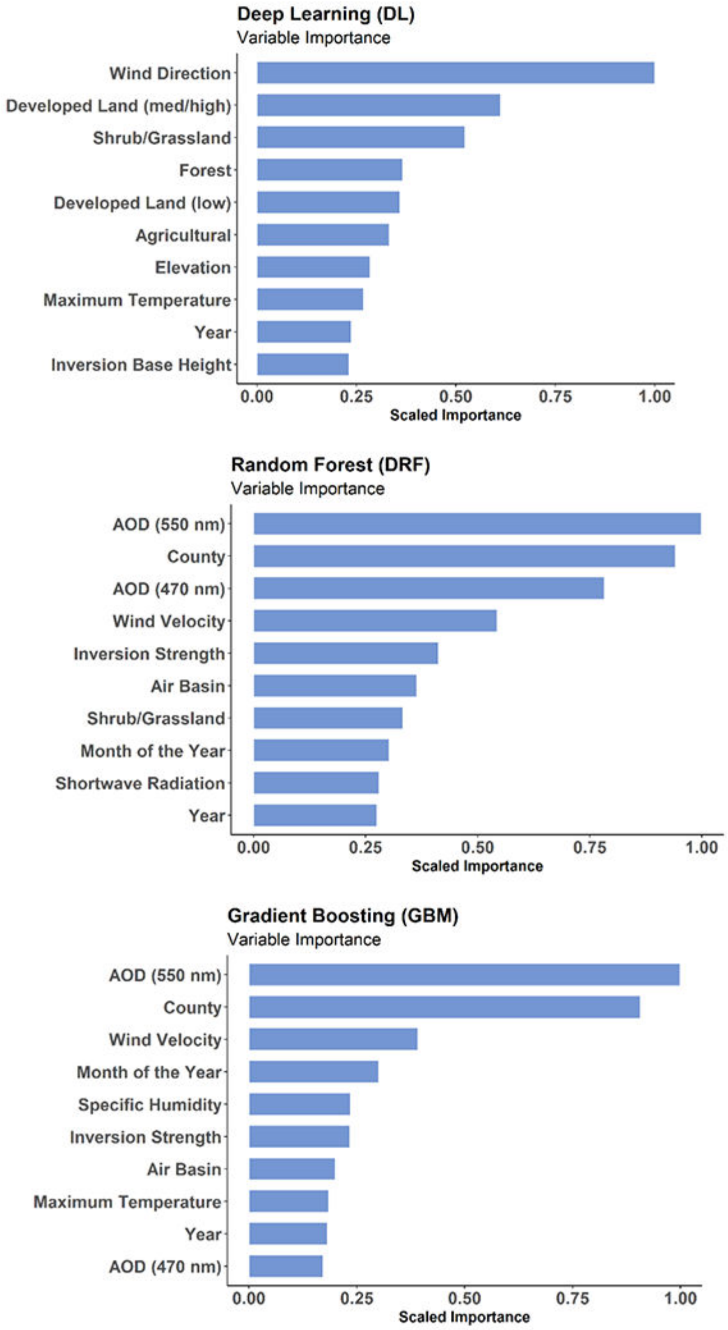
STEP 4: Estimate Wildfire-specific $PM_{2.5}$ for Exposed ZIP Code Days

✓ Subtract non-smoke PM2.5 from the original PM2.5 data:

Wildfire PM2.5 = ML modeled PM2.5 - non-smoke PM2.5

Fig. 2.

Flowchart of steps followed to estimate daily wildfire-specific $PM_{2.5}$ at ZIP code population weighted-centroids in California within 2006–2020.



Deep Learning (DL)
Variable Importance

Random Forest (DRF)
Variable Importance

Gradient Boosting (GBM)
Variable Importance

Fig. 3.

Variable Importance for the top 10 explanatory variables in the Base Learner Models.

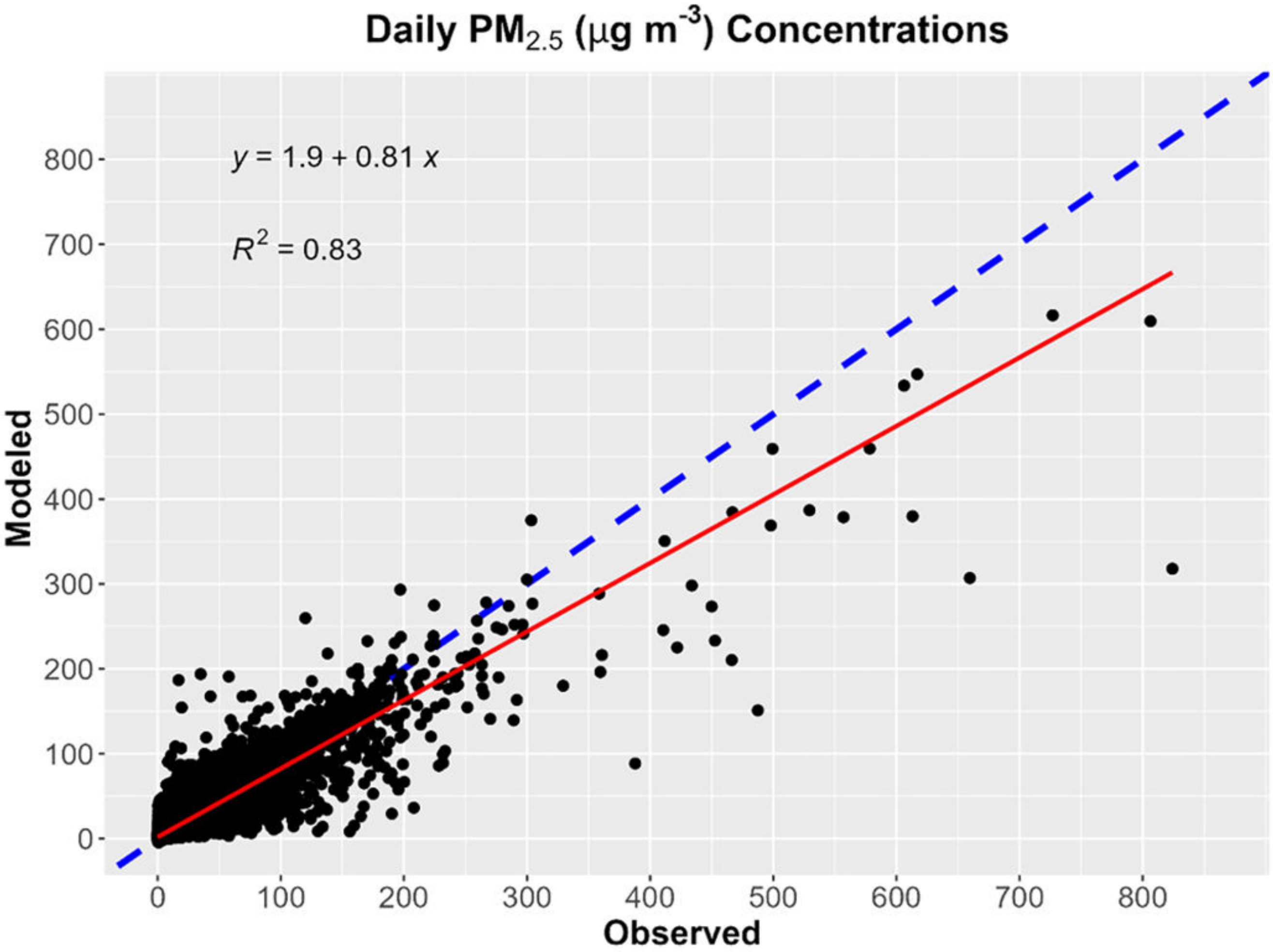


Fig. 4.

Observed versus Modeled $PM_{2.5}$ Concentrations at all EPA AQS Monitoring Sites ($R^2 = 0.83$). Dashed blue line corresponds to the reference (1-to-1) line; red line is the linear model fit. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

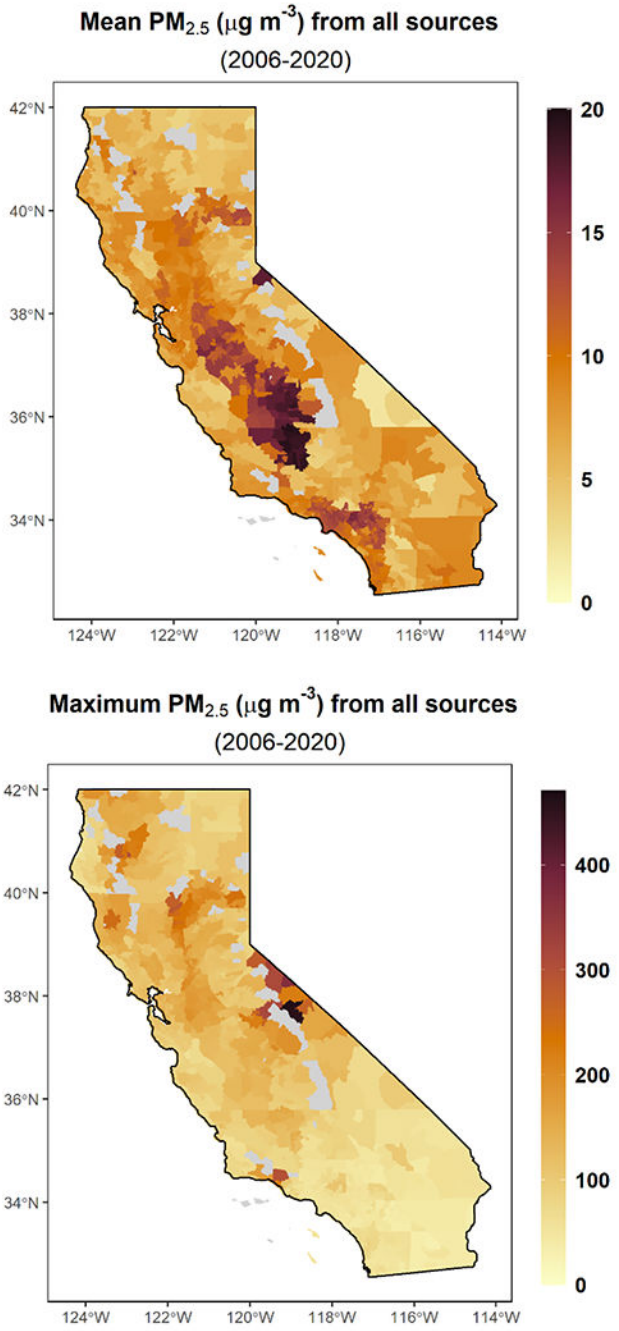


Fig. 5.

Mean (top) and maximum (bottom) $PM_{2.5}$ concentrations from all sources at ZIP codes within the 2006–2020 study period, estimated by the ensemble model with predictors at ZIP code population-weighted centroids. Uninhabited ZIP codes are shown in gray.

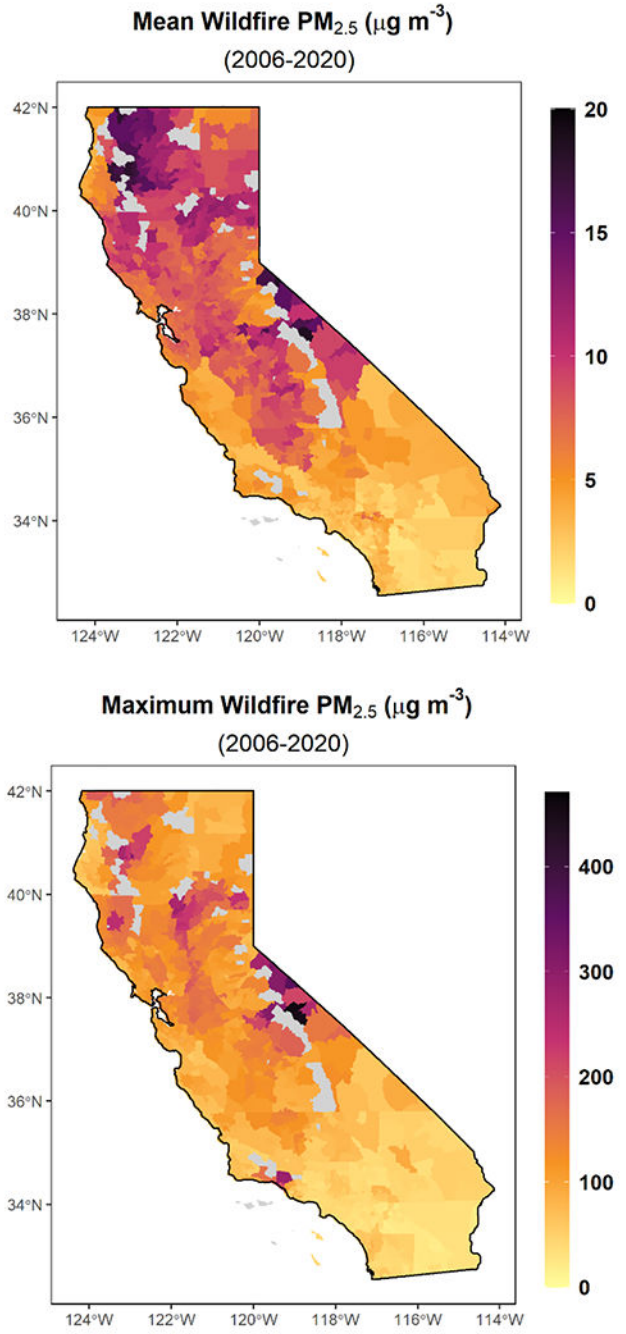


Fig. 6.

Mean (top) and maximum (bottom) wildfire-specific $PM_{2.5}$ concentrations at ZIP codes within the 2006–2020 study period.

Table 1

Model Performance Metrics for Ensemble Model using an optimal combination of the three base learners (Deep Learning, Random Forest and Gradient Boosting). Units for RMSE and MAE are µg m$^{-3.}$

Ensemble Model				
Model Performance Metrics	**Training**	**Validation**	**Hold-out Test (5 Sites)**	**Prediction at all Sites**
RMSE	3.40	5.61	3.51	4.37
Mean Absolute Error (MAE)	2.19	2.90	2.39	2.39
R-squared	**0.87**	**0.67**	**0.78**	**0.83**