Environmental Science & Technology

Article

pubs.acs.org/est

Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning

Colleen E Reid,*[,†,◆] Michael Jerrett,[†,¶] Maya L Petersen,[‡,§] Gabriele G Pfister,[∥] Philip E. Morefield,[⊥] Ira B Tager,[‡] Sean M Raffuse,[#] and John R Balmes[†,▽]

[†]Environmental Health Sciences Division, School of Public Health, University of California, Berkeley, California 94720, United States

[‡]Epidemiology Division, School of Public Health, University of California, Berkeley, California 94720, United States

[§]Biostatistics Division, School of Public Health, University of California, Berkeley, California 94720, United States

[∥]Atmospheric Chemistry Division, National Center for Atmospheric Research, Boulder, Colorado 80301, United States

[⊥]National Center for Environmental Assessment, U.S. Environmental Protection Agency, Washington, D.C. 20460, United States

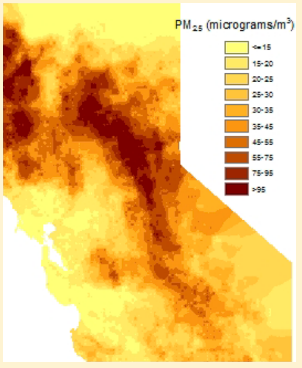
[#]Sonoma Technology, Inc., Petaluma, California 94954, United States

[▽]Department of Medicine, University of California, San Francisco, California 94143, United States

[¶]Environmental Health Sciences Department, Fielding School of Public Health, University of California, Los Angeles, California 90095, United States

Supporting Information

ABSTRACT: Estimating population exposure to particulate matter during wildfires can be difficult because of insufficient monitoring data to capture the spatiotemporal variability of smoke plumes. Chemical transport models (CTMs) and satellite retrievals provide spatiotemporal data that may be useful in predicting $PM_{2.5}$ during wildfires. We estimated $PM_{2.5}$ concentrations during the 2008 northern California wildfires using 10-fold cross-validation (CV) to select an optimal prediction model from a set of 11 statistical algorithms and 29 predictor variables. The variables included CTM output, three measures of satellite aerosol optical depth, distance to the nearest fires, meteorological data, and land use, traffic, spatial location, and temporal characteristics. The generalized boosting model (GBM) with 29 predictor variables had the lowest CV root mean squared error and a CV-R^2 of 0.803. The most important predictor variable was the Geostationary Operational Environmental Satellite Aerosol/Smoke Product (GASP) Aerosol Optical Depth (AOD), followed by the CTM output and distance to the nearest fire cluster. Parsimonious models with various combinations of fewer variables also predicted $PM_{2.5}$ well. Using machine learning algorithms to combine spatiotemporal data from satellites and CTMs can reliably predict $PM_{2.5}$ concentrations during a major wildfire event.



INTRODUCTION

The frequency and severity of wildfires are projected to increase in many parts of the world due to alterations of temperature and precipitation patterns related to climate change.[1] Although numerous studies have investigated the acute health effects of exposure to urban particulate matter (PM), few have investigated the health impacts of exposure to wildfire PM on the general population.[2] Increasing evidence suggests that wildfire PM causes adverse respiratory health effects,[3−6] with some evidence of increased mortality.[7,8] The research shows conflicting evidence for cardiovascular health effects,[9,10] despite coherent evidence of such effects from exposure to other sources of PM.[11]

The lack of consistency in findings could be due to difficulties in population exposure assessment to wildfire smoke. Many $PM_{2.5}$ (PM with aerodynamic diameter ≤2.5 μm) monitors measure only every 3 or 6 days, which requires either averaging over time or imputing values on missing days. Most health effect studies also assign all individuals to the same exposure, either from one monitor,[12−15] or from an average of all monitors in the proximate area.[7,16] Even in locations with dense monitoring networks, smoke plumes vary on spatial scales smaller than monitors can capture. Thus, assigning one value to all exposed individuals likely leads to oversmoothing of exposure estimates, which can bias results, often toward the null, can increase variance, or both, depending on the type of error,[17] thereby making it harder to discern a true causal health effect. Improved modeling of air pollution exposures is also important for risk assessment that relies on exposure-response estimates from epidemiological studies.[18]

Recent studies of the health effects of wildfires have begun to include information on air pollution from satellites, dispersion models, and chemical transport models (CTMs) to estimate

Received: December 1, 2014
Revised: January 29, 2015
Accepted: February 3, 2015

DOI: 10.1021/es505846r
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

population exposure. Some studies have used visible satellite imagery to classify regions as exposed or unexposed,[10] classify days as wildfire-affected,[19] and assign monitors to areas without monitoring data by similarity in smokiness.[20] Others have used quantitative satellite data on atmospheric aerosol loading[9,21] or fire radiative power estimates[22] to classify regions as smoke-exposed. These dichotomizations simplify exposure and could miss gradation in effects associated with concentrations of smoke exposure in a population during a wildfire event.

A few wildfire health studies have used air pollution dispersion models[10,23−25] or CTMs[26,27] to estimate air pollution levels in space and time. Interestingly, both studies that used CTM output combined it with satellite aerosol optical depth (AOD) data,[26,27] but neither included other variables in their analyses despite evidence that meteorological parameters can help to scale vertically full-column AOD measures to ground-level $PM_{2.5}$ estimates.[28−30]

Satellite AOD and CTMs provide quantitative, spatially continuous information about air pollution; however, currently they have spatial resolutions too coarse for estimating human exposures that may vary on small spatial scales during wildfires. Additionally, the relationships between AOD and $PM_{2.5}$ are spatially and temporally heterogeneous.[31−33] Horizontal scaling to smaller spatial resolution can be achieved by using air pollution measurements at monitoring stations as the response variable and AOD, CTM output, and other data as predictors. Coefficients from the fitted statistical model can be applied to predict exposures at unknown locations.[34] Often called land-use regression, this method is used traditionally to create spatial models to estimate long-term average air pollution exposures.[35] Recently, researchers have shown that satellite-based AOD observations can improve the predictive power of $PM_{2.5}$ land-use regression models while also contributing temporal information that is lacking when only considering temporally invariant land-use variables.[36−40]

One limitation of these regression-based exposure models is that they assume a priori a specific type of statistical model for their data, such as a linear model,[41−43] a generalized additive model (GAM),[38] or mixed models.[36,37] Choosing one statistical model may limit the ability to find the best predictive model for the data. In data-adaptive methods, the data inform the choice of model rather than imposing a specific model a priori. V-fold cross-validation provides one data-adaptive method for choosing between candidate estimators while avoiding over-fitting to the data.[44]

Within the air quality literature, researchers have begun to use nonlinear models to predict PM concentrations,[45,46] although few studies have employed robust machine learning techniques such as cross-validation to select among optimal models based on performance metrics.[47−49] We aim to improve exposure assessment to PM during wildfires by using a data-adaptive method that selects among a wider group of statistical algorithms than previous studies to combine an optimal set of variables to best approximate concentrations of total $PM_{2.5}$ during the 2008 northern California wildfires. The optimal model will then be used to estimate spatiotemporal exposures to these wildfires for use in subsequent epidemiological analyses.

MATERIALS AND METHODS

Setting. During the weekend of June 20−21, 2008, over 6000 lightning strikes ignited thousands of fires in 26 counties in northern California.[50] Meteorological conditions and difficulty with fire suppression contributed to very high air pollution levels throughout the state.[51] Our study period is June 20 to July 31, 2008 (N = 42 days), the period when air pollution levels were elevated. These fires contributed to numerous monitor-days that exceeded the U.S. Environmental Protection Agency (USEPA) 24 h average $PM_{2.5}$ standard (35 $\mu g/m^3$).

Data Sources. We collected ground-based monitoring data for $PM_{2.5}$ from the US EPA, the California Air Resources Board (CARB), and the AirNow (http://www.airnow.gov/) and AirFire (http://www.airfire.org/) databases. We used 37 Federal Reference Monitors (FRM), 12 other gravimetric monitors, and 63 beta-attenuation monitors (31 used for regular monitoring by CARB, 9 from the US Forest Service, and 20 that were deployed during these fires to regions without continuously operating monitors). We used FRM monitors, which are deployed for compliance with the USEPA National Ambient Air Quality Standards, and Federal Equivalent Method (FEM) monitors which provide measurements on days when the FRMs are not recording (most FRM monitors collect samples only every six or three days) and at 42 locations without FRMs. Data from colocated FRM and FEM monitors were highly correlated (r = 0.94 to 1) with a mean difference in values of −1.964 $\mu g/m^3$ (range: −19.830, 35.880). We performed sensitivity analyses to compare results using just the FRM monitoring data and all but the FRM monitoring data to our main results.

After first cleaning the data of monitoring data values that had quality control flags demonstrating machine errors, two values were removed from the analysis because they were outliers; one was a value of zero that was surrounded by values close to 20 $\mu g/m^3$ and the other was over 400 $\mu g/m^3$, which was determined to be too high to be accurately measured by a beta-attenuation monitor (BAM).[52]

The National Center for Atmospheric Research (NCAR) provided $PM_{2.5}$ concentration estimates from the Weather Research and Forecasting with Chemistry (WRF-Chem) 3.2 model. WRF-Chem 3.2 is a regional CTM based on the chemical, spatial, and temporal boundary conditions from the Model for OZone And Related chemical Tracers (MOZART)-4, a global CTM (see Pfister et al.[53]). Inputs included meteorology, physical and chemical atmospheric processes, emissions from a California-specific emissions inventory for 2008, online biogenic emissions, and fire emissions estimated with the Fire Inventory from NCAR (FINN) V1, (see Wiedinmyer et al.[54]). We used 24 h averages of the hourly output of $PM_{2.5}$ at the lowest vertical level of the model, where population exposure occurs.

We obtained AOD measurements from the Geostationary Operational Environmental Satellite (GOES) West Aerosol Smoke Product (GASP) from the National Oceanic and Atmospheric Administration (NOAA) using their January 7, 2009 revised algorithm. The GASP product has a spatial resolution of 4 km pixels at nadir and has daily retrievals every 30 min during daylight; approximately 24 retrievals per day. We assessed all GASP retrievals for data quality and removed any null values and scenes with too few pixel values. NOAA's quality control process removed some pixels from the center of dense smoke plumes either because these were assumed to be clouds or the signal was too low or negative.[55] We estimated these missing values by fitting an optimal radial basis function (RBF) because: (i) an optimal RBF would be selected by minimizing the root mean squared error (RMSE) of the

DOI: 10.1021/es505846r
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Table 1. Variables Used To Predict $PM_{2.5}$ during the 2008 Northern California Wildfires

variables	data source	temporal resolution	spatial resolution
Dependent Variable			
$PM_{2.5}$ from monitoring stations (N = 112)	USEPA, California Air	daily or hourly	
37 Federal Reference monitors	Resources Board, Air Districts, and U.S. Forest Service		
12 other gravimetric monitors			
43 BAM monitors			
20 eBAMs (just for fire)			
Spatiotemporal Variables			
GASP AOD	National Oceanic and Atmospheric Administration	half-hourly, daylight	4 km
MODIS AOD	NASA	twice daily	10 km
Local AOD	Sonoma Technology, Inc. (derived from raw MODIS retrievals)	daily	0.5 km
WRF-Chem $PM_{2.5}$ ($\mu g/m^3$)	National Center for Atmospheric Research	hourly	12 km
distance to nearest cluster of active fires (m)	Derived from USDA Forest Service Remote Sensing Applications Center	daily	
counts of fires in nearest cluster/distance			
relative humidity (%)	rapid update cycle	daily	13 km
sea level pressure (Pa)			
surface pressure (Pa)			
Planetary boundary layer height (m)			
U-component of wind speed (m/s)			
V-component of wind speed (m/s)			
dew point temperature (K)			
Temperature at 2 m (K)			
Spatial Variables			
x-coordinate (m)	U.S. Environmental Protection		
y-coordinate (m)	Agency Air Quality System		
counts of traffic within 1 km	Dynamap 2000, TeleAtlas	annual	1 km
% of urban land use within 1km	2006 National Land Cover Database		1 km
% of agricultural land use within 1km			
% of vegetation land use within 1km			
any high intensity land use within 1 km			
elevation (m)	National Elevation Data set 2010		
binary indicator variables for air basin (San Francisco Bay Area, Sacramento Valley, San Joaquin Valley, and Mountain Counties)	California Air Resources Board		air basin
population density	U.S. Census 2000		block group
Temporal Variables			
Julian date and weekend		daily	

interpolated surface, (ii) RBF allows interpolation of values greater than the input values which is important given that the missing values had higher reflectance from smoke than surrounding values that were not removed by NOAA, and (iii) RBF is an exact interpolator and thus all observed data points are retained. Cloud cover in the summer in California is not a major impediment to retrieval except along the Pacific coast, where we did not interpolate missing values. We calculated daily average surface values for all days during which there were at least 12 successful GASP retrievals.

The MODIS (MOderate Resolution Imaging Spectroradiometer) AOD product has a spatial resolution of 10 km and temporal resolution of at most two retrievals per day. All MODIS data were processed with the same data cleaning, RBF interpolation, and daily averaging as the GASP data. The average RMSE from the RBF functions were 0.086 for GASP and 0.054 for MODIS (AOD values are unitless but tend to range from 0 to 1 over the U.S.).

Sonoma Technology Inc. and the University of Southern California created a high-resolution (500 m) kernel-smoothed AOD for northern California during these wildfires using raw MODIS data. They used a local estimate of surface brightness, a local AOD algorithm for fresh smoke plumes, and a less restrictive cloud filter that does not screen out pixels that are part of smoke plumes[56] to create AOD estimates more refined to local conditions than the standard MODIS AOD product.

We downloaded temperature, relative humidity, sea level pressure, surface pressure, planetary boundary layer height, dew point temperature, and the U and V components of wind speed from the National Climatic Data Center's Rapid Update Cycle (RUC) Model (http://ruc.noaa.gov/) and calculated 24 h averages from hourly data.

Researchers at NCAR provided cumulative daily sums of fire points from MODIS Fire Detection points from the Remote Sensing Applications Center of the US Forest Service (http://activefiremaps.fs.fed.us/gisdata.php). From these data, we calculated two daily metrics: the distance from each monitoring site to the nearest cluster of fire points (those within 5 km of each other) and the number of fire points within each cluster divided by the distance.

To account for other sources of $PM_{2.5}$ during the wildfires that would contribute to monitored $PM_{2.5}$ values during the fires, we included traffic and land use information. We calculated the sum of all traffic counts within 1 km of a

DOI: 10.1021/es505846r
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Table 2. CV-RMSE and CV-R^2 Values for the Best Model Across the 11 Algorithms

	model with smallest CV-RMSE for subsets of variables			model with fewer variables whose CV-RMSE was within 1.5% of the smallest CV-RMSE		
	CV-RMSE ($\mu g/m^3$)	CV-R^2	no. of variables selected	CV-RMSE ($\mu g/m^3$)	CV-R^2	no. of variables selected
random forest	1.513	0.796	20	1.521	0.790	14
bagged trees	1.687	0.672	27	1.696	0.665	15
generalized boosting model	1.489	0.803	29	1.495	0.799	13
elastic net regression	1.848	0.538	28	1.852	0.535	27
multivariate adaptive regression splines	1.642	0.701	28	1.648	0.696	26
lasso regression	1.821	0.558	28	1.834	0.548	23
support vector machines	1.556	0.761	16	1.561	0.758	15
gaussian processes	1.580	0.746	16	1.591	0.739	14
generalized linear model	1.821	0.558	29	1.834	0.549	23
K-nearest neighbors	2.030	0.387	2	2.044	0.374	1
generalized additive model	1.607	0.725	26	1.609	0.724	25

$PM_{2.5}$ monitor from Dynamap 2000.[57] We used the National Land Cover Database for 2006[58] to calculate within 1 km of each monitor the percentage of urban development (codes 22, 23, and 24), agriculture (codes 81 and 82), other vegetated area (codes 21, 41, 42, 43, 52, and 71), and to create a binary indicator of whether any Developed High Intensity land use (code 24) occurred.

We used the National Elevation Data set for California from 2010 and population density estimates by block group from the 2000 U.S. census. We extracted the x- and y-coordinates for each monitor in the California Teale Albers projection, and created indicator variables for each of the following air basins: San Francisco Bay, Sacramento Valley, San Joaquin Valley, and Mountain Counties. We also created a continuous variable of Julian date and a binary variable denoting if the day was a weekend.

Statistical Analysis. We used 10-fold cross-validation to determine which of the following 11 algorithms, chosen to reflect a diversity of statistical algorithm types, resulted in the best predictor of $PM_{2.5}$ in these data: generalized linear models (GLM),[59] random forest (RF),[60] bagged trees,[61] generalized boosting models (GBM),[62] generalized additive models (GAM),[63] multivariate adaptive regression splines,[64] elastic nets,[65] support vector machines with a radial basis kernel,[66] Gaussian processes with a radial basis kernel,[66] k nearest neighbors regression,[61] and lasso regression.[67] Nested within this 10-fold cross-validation was another level of 10-fold cross-validation for each of 29 subsets of predictor variables (e.g., all 29, the 28 best, the 27 best,...) from the list of 29 independent variables in Table 1, thus running 10-fold cross-validation 319 (29×11) times. The log of $PM_{2.5}$ for all monitor-days (N = 1540) was the dependent variable. Within this nested 10-fold cross-validation, parameters for the models that required them (i.e., interaction depth and shrinkage for GBM) were estimated using an additional layer of 10-fold cross-validation (see Kuhn[68] for details).

In 10-fold cross-validation, each model is trained on 90% of the data and then evaluated on the 10% the data that is left out (the validation set), in our case a random sample of our data. This process is repeated 10 times and the resulting performance metric (i.e., the cross-validation root mean square error (CV-RMSE)) is averaged across the 10 exhaustive and mutually exclusive validation sets. As a result, performance is always evaluated based on data not used to train the model, with each observation contributing exactly once to validation. For each algorithm, we selected two "best" models: (1) the model with the smallest CV-RMSE and (2) a more parsimonious model whose CV-RMSE was within 1.5% of the smallest CV-RMSE. We then compared the smallest CV-RMSE from each algorithm to choose which algorithm best fit our data.

To further analyze fit of the models with the lowest CV-RMSE, we inspected residual plots for lack of heteroskedasticity and assessed agreement between monitoring data and predicted values at the monitoring sites with Bland-Altman plots. We assessed bias by the slope of a linear regression with zero intercept on the predicted compared to the observed data. Further, we examined spatial autocorrelation in the residuals using Moran's I, compared the range and distribution of predicted and observed values, and visualized predicted values across the study area to determine if the model predictions captured the spatial characteristics of the smoke plume as seen in visible imagery from the MODIS satellite.

When variables are correlated, the subset of variables chosen is dependent on how the folds are created, which is determined by a random seed. The optimal model should still have similar performance even with a different set of predictor variables. If certain variables were better predictors regardless of the composition of the folds, these variables would be repeatedly selected under different internal data splits. We therefore ran our data-adaptive method five times with different seeds for sorting the observations and assessed the average relative importance of each variable in the model with the lowest CV-RMSE across these five runs. We used the GBM's calculation of relative importance, which is essentially the empirical improvement of the model for splitting on that variable summed over all nodes within a tree and averaged over all trees within the boosted model. Additionally, we investigated if fewer variables could predict $PM_{2.5}$ concentrations well during the wildfires by only allowing the algorithm to select among smaller subsets of variables.

We used R v.2.15.3[59] for all statistical analyses, GeoDa v.1.2.0[69] for Moran's I and ArcGIS 10.1[70] for spatial data processing and map creation.

RESULTS

The variable most correlated with the outcome was the GASP AOD, followed by the distance to the nearest active fire cluster (negatively), and then equally by the WRF-Chem model's $PM_{2.5}$ estimate and the local AOD product (Supporting Information (SI) Table S1). Many of the predictor variables were correlated with each other (SI Table S2). Because we were

DOI: 10.1021/es505846r
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

interested in prediction not inference, collinearity was not a primary concern.

Table 2 shows the CV-RMSE, CV-R^2, and number of variables chosen for the prediction model with the lowest CV-RMSE and for the more parsimonious model. Across algorithms, GBM fit the data the best, but the RF model was a close second; for both methods, the model with the optimal set of variables had a CV-R^2 that rounds to 0.80. The parameters chosen for GBM by another layer of nested 10-fold cross-validation were interaction depth = 9, number of trees = 500 and shrinkage = 0.1.

We compared the out-of-sample predicted values with the observed values of the RF and GBM models. The GBM's predicted-observed plot shows the values more evenly distributed across the line of unity ($y = x$) at the low and high values where the RF model overpredicts and underpredicts, respectively. The Bland-Altman plots, however, demonstrate slightly tighter agreement for the RF model than the GBM with fewer large negative residuals (Figure 1). We found little evidence of bias in either model with a slope of 1.005 (SE = 0.003) for the RF model and 0.999 (SE = 0.003) for the GBM. Moran's I based on a queen's contiguity matrix of the first-order nearest neighbors revealed no evidence of spatial autocorrelation in the residuals for either algorithm (SI Table S3).

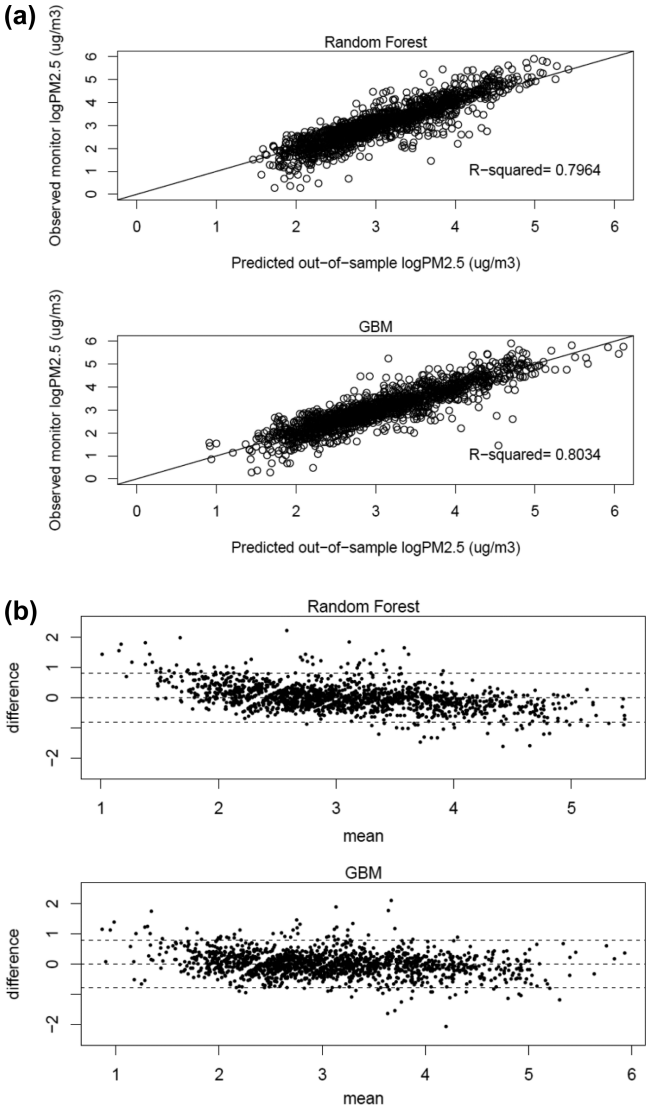


Figure 1. Model diagnostic plots for the optimal model based on 10-fold cross-validation using RF and GBM, respectively.

Figure 2 shows the satellite images and the predicted grids (5 km) for RF and GBM models on June 29, a day with minimal smoke, and July 11, a day with smoke covering most areas. This comparison is limited in two main ways: (1) the visible imagery are from one time point, whereas the model predictions represent 24 h days, and (2) the satellite images shows total atmospheric column smoke and our model predicts at ground level. With these limitations in mind, each model appears to capture some of the spatial variability in the smoke plume evident in the visible imagery.

A comparison of the predicted values for the 5 km grid over the study area demonstrated that the RF model predicted values across a smaller range (min = 3.4 $\mu g/m^3$, max = 188.4 $\mu g/m^3$) than the GBM model (min = 2.0 $\mu g/m^3$, max = 337.4 $\mu g/m^3$). The latter was closer to the full range of the observed monitoring data (min = 1.5 $\mu g/m^3$, max = 364.8 $\mu g/m^3$) (SI Figure S1).

SI Figure S2 shows the CV-RMSE for every subset of variables run for GBM. The first few variables had the most impact on the CV-RMSE, and although the model with all 29 variables had the smallest CV-RMSE, the model with only 13 variables has a CV-RMSE less than 1.5% greater. These 13 variables were, in order of importance: GASP AOD, distance to the nearest fire cluster, WRF-Chem, Julian date, surface pressure, local AOD, sea level pressure, relative humidity, v-component of wind speed, u-component of wind speed, x-coordinate, MODIS AOD, and temperature. This more parsimonious model fit the observed data well (SI Figure S3) and was comparable to the model with all 29 variables with the greatest difference occurring for the extremely high values (SI Figure S4).

When we ran the GBM five times allowing different random seeds, the CV-R^2 values for the best models rounded to 0.80 or 0.81. In each run, the model with the smallest CV-RMSE selected between 20 and 29 variables with 19 chosen by all five; the parsimonious models selected between 14 and 28 variables. The average relative importance of each variable across the five runs (SI Table S4) demonstrated that GASP AOD was the most influential variable in creating our optimal exposure model. The rank ordering of the variables was fairly consistent; GASP AOD, Julian date, and WRF-Chem were chosen as the first, second, and third variables, respectively, for each of the five runs.

Although the run that allowed selection among all 29 variables had the lowest CV-RMSE and highest CV-R^2, many of the other subsets with important variables removed approximated the fit of the optimal model (Table 3), possibly due to high collinearity among the spatiotemporal variables. Pearson correlations between MODIS AOD, local AOD, and WRF-Chem with GASP AOD were 0.712, 0.705, and 0.483, respectively. The CV-R^2 for the model with only universally available variables (i.e., those not specific to our study domain such as x- and y-coordinates, dummies for air basin, and Julian date) was 0.77, only slightly lower than that of the model with all of the variables.

Our results from analyses with just FRM monitors and all but the FRM monitors showed that the model using only FRM data had the smallest CV-RMSE (SI Table S5). The FRM monitoring data did not have as large a variance, likely due to

DOI: 10.1021/es505846r
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

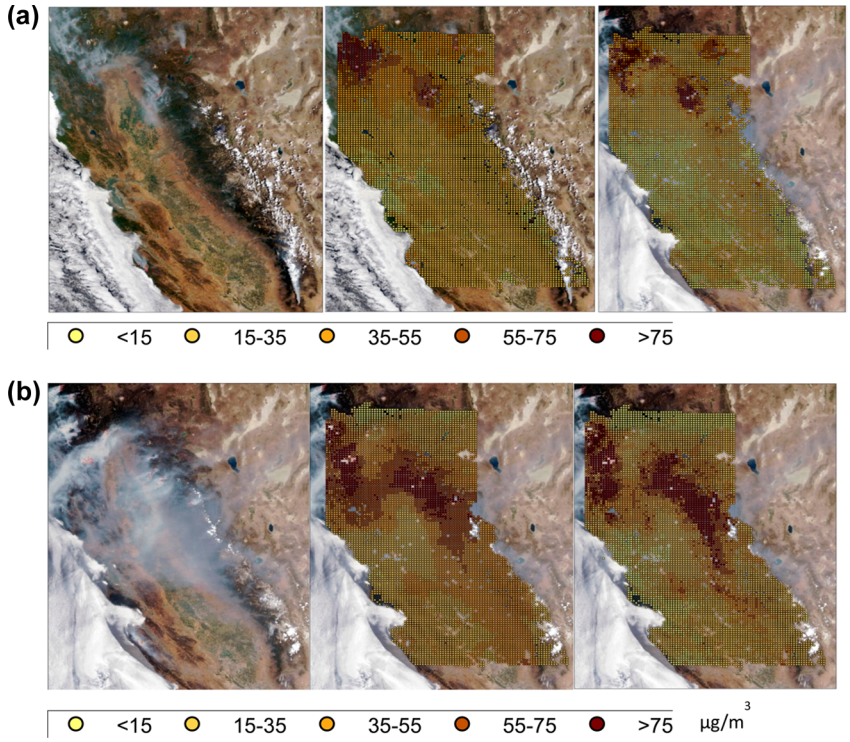


Figure 2. (a) Satellite image, predicted grid from RF and from GBM on June 29, 2008. (b) Satellite image and predicted grids from RF and GBM on July 11, 2008.

Table 3. CV-R^2 and CV-RMSE for GBM Models with Different Subsets of Variables

	all variables	GASP AOD plus[a]	WRF-Chem plus[a]	emissions[b] plus[a]	just plus[a]	MODIS AOD plus[a]	local AOD plus[a]	universal variables[c]
CV-RMSE	1.489	1.495	1.531	1.556	1.542	1.548	1.520	1.542
CV-R^2	0.803	0.800	0.774	0.757	0.768	0.764	0.784	0.770
no. of variables chosen	29 out of 29	25 out of 26	25 out of 26	18 out of 26	19 out of 25	22 out of 26	16 out of 26	19 out of 20

[a]Plus means the following variables: temperature, relative humidity, sea level pressure, surface pressure, planetary boundary layer height, dew point temperature, and the U and V components of wind speed, distance to the nearest fire cluster, counts of fires in nearest cluster/distance, *x*-coordinate, *y*-coordinate, counts of traffic within 1 km, % of urban land use within 1 km, % of agricultural land use within 1 km, % of vegetation land use within 1 km, any high intensity land use within 1 km, elevation, indicator variables for air basin, population density, Julian date, and an indicator variable for weekend. [b]The emissions plus model allowed selection from the plus variables and the estimated total emissions per day from the FINN model. [c]The universal variables include: GASP AOD, WRF-Chem, MODIS AOD, temperature, relative humidity, sea level pressure, surface pressure, planetary boundary layer height, dew point temperature, and the U and V components of wind speed, distance to the nearest fire cluster, counts of fires in nearest cluster/distance, counts of traffic within 1 km, % of urban land use within 1 km, % of agricultural land use within 1km, % of vegetation land use within 1 km, any high intensity land use within 1 km, elevation, and population density.

the fewer monitor-days (N = 277), compared to all of the data (N = 1540), and the limited locations of the FRM data farther from the fires. Inclusion of other monitors including eBAMs that were deployed to areas closer to the fires and more daily monitors supplied better data support for concentration prediction by increasing spatial and temporal coverage, and in the case of eBAMs, also information on concentrations closer to the fires. By including all monitoring data, estimated concentrations more likely matched the true exposures.

DISCUSSION

Our analyses demonstrate the utility of using data-adaptive approaches (i.e., machine learning algorithms) to combine spatial, temporal, and spatiotemporal data to improve concentration estimates for $PM_{2.5}$ from satellite data and CTMs. Our best model had a CV-R^2 of 0.803 with little heteroskedasticity or autocorrelation in the residuals, good agreement with the observed data, and predicted values that captured the variability evident in visible satellite imagery of the smoke plume on high and low smoke days.

Had we assumed one statistical algorithm a priori, it likely would have yielded inferior results. The best CV-R^2 value from a GLM model was 0.558 compared to 0.803 from GBM. Even GAM, which performed better (CV-R^2 = 0.725) than a linear model, still did not perform as well as GBM or RF.

GBM is a generalization of tree boosting that provides an accurate and effective model for data mining.[71] Boosting combines many weak tree-based models into a powerful committee of models. The method requires each iterative model to better predict previously poorly predicted observations by up-weighting those observations and down-weighting well-predicted observations. By combining all of the weak models together, the boosted model predicts well over the range of observations.[71]

DOI: 10.1021/es505846r
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

GASP AOD was the most predictive variable of surface-level $PM_{2.5}$ concentration. Its variable importance factor in the GBM was three times that of the next most important variable (distance to the nearest fire cluster). GASP AOD has corresponded well to a ground-based measure of AOD (AERONET)[72] and predicted in situ $PM_{2.5}$ concentrations well in the eastern U.S., but correlations were weaker in the West.[33] GASP AOD had the finest temporal resolution, every half hour during daylight hours compared to twice daily for the other two AOD sources, but intermediate spatial resolution (4 km compared to 10 km for MODIS and 0.5 km for local AOD), suggesting that temporal rather than spatial resolution of AOD is important for predicting $PM_{2.5}$ during wildfires. Although research has shown that MODIS AOD corresponds better to ground-based AOD measurements than GASP AOD,[72,73] statistical models that incorporate meteorological and land-use data with GASP have yielded good results.[38]

Previous research demonstrates that $PM_{2.5}$ and AOD are more correlated when more particles are in the fine mode,[73] and when PM concentrations are higher, particularly during wildfires.[74] Our results corroborate these findings, but also demonstrate improved performance when other spatial, temporal, and spatiotemporal data are combined with AOD to predict $PM_{2.5}$.

WRF-Chem, a CTM, also predicted ground-level $PM_{2.5}$ concentrations well during the fires. Our model run with just WRF-Chem and other variables had a CV-R^2 value of 0.774 and WRF-Chem had the third highest variable importance across GBM runs. CTMs have been used to assess the impacts of wildfires on air quality,[75−77] and have been combined with satellite data in health risk assessments of fires,[26,27] but have not yet been used for exposure assessment in epidemiological analyses, although dispersion models have.[10,25] Dispersion models, however, may lack chemical reactions and thus underestimate total particulate matter during wildfires.

Although some satellite data products have recently been released with finer spatial resolution,[78] most CTMs and satellite data are too spatially coarse for exposure estimation for epidemiological analyses. Our method of incorporating local land use and traffic information provides one framework for how to spatially downscale coarse spatiotemporal data sets to increase relevance for epidemiological analyses. Our results also add to the growing literature that combines satellite retrievals with other data to estimate air pollution exposures;[26,27,79,80] such analyses combine the observational strengths of the satellite data with the ground-level estimates of CTMs to better predict population air pollution exposures.

To approximate the true data-generating process that created $PM_{2.5}$ concentrations during these wildfires, we would want to select from the largest library of algorithms possible. The 11 algorithms we used represent a large range of statistical models and our list of predictor variables is more extensive than those previously used to estimate wildfire $PM_{2.5}$ exposure. Although land use and traffic variables are important predictors of $PM_{2.5}$ during normal conditions, these variables were not strong predictors during these wildfires compared to AOD measures and the WRF-Chem output. Interestingly, when we excluded all AOD and CTM data, the resulting model of just meteorological, spatial, and temporal variables predicted our observed data well (Table 3, "Just plus" model).

An important finding from our work is that models with variables that are not specific to these fires but could be obtained for any location, those included in our "universal model", performed almost as well as the best performing model. In follow-up research, we are now investigating whether the models generated for these fires predict monitored $PM_{2.5}$ levels well when applied to other fires with different characteristics. Results from these ongoing analyses could yield a prediction model that could be used to estimate $PM_{2.5}$ concentrations during wildfires in places with little to no monitoring data.

Two recent studies have demonstrated the ability of remotely sensed fire information, prior day air quality, and meteorological data to predict air quality the next day.[82,83] These models had lower performance than our model, which could be due to the diversity of modeling algorithms or input variables used, the fact that we were predicting same day rather than next day $PM_{2.5}$, or due to differences in location or fire characteristics. Further research into the use of our method to forecast $PM_{2.5}$ could inform public health efforts during wildfire events.

Our model performed very well compared to out-of-sample $PM_{2.5}$ measurements. It provides estimates of daily $PM_{2.5}$ concentrations during a significant wildfire smoke episode. By combining data with broad coverage, such as that from satellites and CTMs, with local small-area spatial and temporal information, this method could be applied in other regions that experience regular wildfires but have fewer monitoring stations.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information contains the following tables: Pearson correlations between variables, results from the Moran's *I* test for autocorrelation, variable importance in the final prediction model, and the results from using FRM monitors only. The Supporting Information contains the following figures: a boxplot comparing predicted values for the RF and the GBM, diagnostic plots of the GBM with only 13 predictor variables, and a boxplot of predicted values from the 29 variable and the 13 variable GBM. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone:617-495-8108; fax:617-495-5418; e-mail: coreid@hsph.harvard.edu.

Present Address

◆C.E.R.: Robert Wood Johnson Health and Society Scholar, 9 Bow St., Cambridge, Massachusetts 02138, United States.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Ricardo Cisneros of the University of California, Merced for providing monitoring data from the eBAMs, AirFire, and AirNow networks. A version of this work was previously presented as a poster whose abstract was published in Environmental Health Perspectives.[84] This research was supported under a cooperative agreement from the Centers for Disease Control and Prevention through the Association of Schools of Public Health Grant Number CD300430, the Joint Fire Science Program, Bureau of Land Management

DOI: 10.1021/es505846r
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

(L14AC00173), and an EPA STAR Fellowship Assistance Agreement no. FP-91720001-0 awarded by the USEPA. The WRF-Chem modeling work was supported by NASA grant NNX08AD22G. The local AOD estimates were developed under NIEHS grant 1R21ES016986. Maya Petersen is a recipient of a Doris Duke Clinical Scientist Development Award. NCAR is operated by the University Corporation of Atmospheric Research under sponsorship of the National Science Foundation. The views expressed in this paper are solely those of the authors and not those of the funding agencies.

■ ABBREVIATIONS

AOD	aerosol optical depth
CARB	California Air Resources Board
CTM	chemical transport model
CV	cross-validated
FEM	federal equivalent method
FRM	federal reference method
GAM	generalized additive model
GASP	GOES aerosol smoke product
GBM	generalized boosting model
GLM	generalized linear model
GOES	geostationary operational environmental satellite
FINN	fire inventory from NCAR
MODIS	moderate resolution imaging spectroradiometer
NCAR	National Center for Atmospheric Research
NOAA	National Oceanic and Atmospheric Administration
PM	particulate matter
$PM_{2.5}$	particulate matter less than or equal to 2.5 μm in aerodynamic diameter
RBF	radial basis function
RMSE	root mean squared error
RUC	rapid update cycle
USEPA	United States Environmental Protection Agency
WRF-Chem	Weather Research and Forecasting with Chemistry model

■ REFERENCES

(1) Confalonieri, U.; Menne, B.; Akhtar, R.; Ebi, K. L.; Hauengue, M.; Kovats, R. S.; Revich, B.; Woodward, A. Human Health. In *Climate Change 2007: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to teh Fourth Assessment Report of the Intergovernmental Panel on Climate Change*; Parry, M. L., Canziani, O. R., Palutikof, J. P., Linden, P. J. v. d., Hanson, C. E., Eds.; Cambridge University Press: Cambridge, UK, 2007; pp 391−431.

(2) Naeher, L. P.; Brauer, M.; Lipsett, M.; Zelikoff, J. T.; Simpson, C. D.; Koenig, J. Q.; Smith, K. R. Woodsmoke health effects: A review. *Inhal Toxicol* **2007**, *19* (1), 67−106.

(3) Delfino, R. J.; Brummel, S.; Wu, J.; Stern, H.; Ostro, B.; Lipsett, M.; Winer, A.; Street, D. H.; Zhang, L.; Tjoa, T.; Gillen, D. L. The relationship of respiratory and cardiovascular hospital admissions to the southern California wildfires of 2003. *Occup Environ. Med.* **2009**, *66* (3), 189−97.

(4) Henderson, S. B.; Johnston, F. H. Measures of forest fire smoke exposure and their associations with respiratory health outcomes. *Curr. Opin Allergy Clin. Immunol.* **2012**, *12* (3), 221−7.

(5) Kunzli, N.; Avol, E.; Wu, J.; Gauderman, W. J.; Rappaport, E.; Millstein, J.; Bennion, J.; McConnell, R.; Gilliland, F. D.; Berhane, K.; Lurmann, F.; Winer, A.; Peters, J. M. Health effects of the 2003 Southern California wildfires on children. *Am. J. Respir. Crit. Care Med.* **2006**, *174* (11), 1221−8.

(6) Morgan, G.; Sheppeard, V.; Khalaj, B.; Ayyar, A.; Lincoln, D.; Jalaludin, B.; Beard, J.; Corbett, S.; Lumley, T. Effects of bushfire smoke on daily mortality and hospital admissions in Sydney, Australia. *Epidemiology* **2010**, *21* (1), 47−55.

(7) Johnston, F.; Hanigan, I.; Henderson, S.; Morgan, G.; Bowman, D. Extreme air pollution events from bushfires and dust storms and their association with mortality in Sydney, Australia 1994−2007. *Environ. Res.* **2011**, *111* (6), 811−6.

(8) Sastry, N. Forest fires, air pollution, and mortality in southeast Asia. *Demography* **2002**, *39* (1), 1−23.

(9) Rappold, A. G.; Stone, S. L.; Cascio, W. E.; Neas, L. M.; Kilaru, V. J.; Carraway, M. S.; Szykman, J. J.; Ising, A.; Cleve, W. E.; Meredith, J. T.; Vaughan-Batten, H.; Deyneka, L.; Devlin, R. B. Peat bog wildfire smoke exposure in rural north carolina is associated with cardiopulmonary emergency department visits assessed through syndromic surveillance. *Environ. Health Perspect.* **2011**, *119* (10), 1415−20.

(10) Henderson, S. B.; Brauer, M.; Macnab, Y. C.; Kennedy, S. M. Three measures of forest fire smoke exposure and their associations with respiratory and cardiovascular health outcomes in a population-based cohort. *Environ. Health Perspect.* **2011**, *119* (9), 1266−71.

(11) Brook, R. D.; Rajagopalan, S.; Pope, C. A., 3rd; Brook, J. R.; Bhatnagar, A.; Diez-Roux, A. V.; Holguin, F.; Hong, Y.; Luepker, R. V.; Mittleman, M. A.; Peters, A.; Siscovick, D.; Smith, S. C., Jr.; Whitsel, L.; Kaufman, J. D. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. *Circulation* **2010**, *121* (21), 2331−78.

(12) Chen, L.; Verrall, K.; Tong, S. Air particulate pollution due to bushfires and respiratory hospital admissions in Brisbane, Australia. *Int. J. Environ. Health Res.* **2006**, *16* (3), 181−91.

(13) Tham, R.; Erbas, B.; Akram, M.; Dennekamp, M.; Abramson, M. J. The impact of smoke on respiratory hospital outcomes during the 2002−2003 bushfire season, Victoria, Australia. *Respirology* **2009**, *14* (1), 69−75.

(14) Lee, T. S.; Falter, K.; Meyer, P.; Mott, J.; Gwynn, C. Risk factors associated with clinic visits during the 1999 forest fires near the Hoopa Valley Indian Reservation, California, USA. *Int. J. Environ. Health Res.* **2009**, *19* (5), 315−27.

(15) Kolbe, A.; Gilchrist, K. L. An extreme bushfire smoke pollution event: Health impacts and public health challenges. *N S W Public Health Bull.* **2009**, *20* (1−2), 19−23.

(16) Analitis, A.; Georgiadis, I.; Katsouyanni, K. Forest fires are associated with elevated mortality in a dense urban setting. *Occup Environ. Med.* **2012**, *69* (3), 158−62.

(17) Zeger, S. L.; Thomas, D.; Dominici, F.; Samet, J. M.; Schwartz, J.; Dockery, D.; Cohen, A. Exposure measurement error in time-series studies of air pollution: Concepts and consequences. *Environ. Health Perspect.* **2000**, *108* (5), 419−426.

(18) Fann, N.; Bell, M. L.; Walker, K.; Hubbell, B. Improving the linkages between air pollution epidemiology and quantitative risk assessment. *Environ. Health Perspect* **2011**, *119* (12), 1671−5.

(19) Johnston, F. H.; Hanigan, I. C.; Henderson, S. B.; Morgan, G. G.; Portner, T.; Williamson, G. J.; Bowman, D. M. Creating an integrated historical record of extreme particulate air pollution events in Australian cities from 1994 to 2007. *J. Air Waste Manag Assoc* **2011**, *61* (4), 390−8.

(20) Wu, J.; Winer, A.; Delfino, R. Exposure assessment of particulate matter air pollution before, during, and after the 2003 Southern California wildfires. *Atmos. Environ.* **2006**, *40* (18), 3333−3348.

(21) Frankenberg, E.; McKee, D.; Thomas, D. Health consequences of forest fires in Indonesia. *Demography* **2005**, *42* (1), 109−29.

(22) Elliott, C. T.; Henderson, S. B.; Wan, V. Time series analysis of fine particulate matter and asthma reliever dispensations in populations affected by forest fires. *Environ. Health* **2013**, *12*, 11.

(23) Rappold, A. G.; Cascio, W. E.; Kilaru, V. J.; Stone, S. L.; Neas, L. M.; Devlin, R. B.; Diaz-Sanchez, D. Cardio-respiratory outcomes associated with exposure to wildfire smoke are modified by measures of community health. *Environ. Health* **2012**, *11*, 71.

(24) Yao, J.; Brauer, M.; Henderson, S. B. Evaluation of a wildfire smoke forecasting system as a tool for public health protection. *Environ. Health Perspect* **2013**, *121* (10), 1142−7.

DOI: 10.1021/es505846r
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

(25) Thelen, B.; French, N. H.; Koziol, B. W.; Billmire, M.; Owen, R. C.; Johnson, J.; Ginsberg, M.; Loboda, T.; Wu, S. Modeling acute respiratory illness during the 2007 San Diego wildland fires using a coupled emissions-transport system and generalized additive modeling. *Environ. Health* **2013**, *12* (1), 94.
(26) van Donkelaar, A.; Martin, R. V.; Levy, R. C.; da Silva, A. M.; Krzyzanowski, M.; Chubarova, N. E.; Semutnikova, E.; Cohen, A. J. Satellite-based estimates of ground-level fine particulate matter during extreme events: A case study of the Moscow fires in 2010. *Atmos. Environ.* **2011**, *45* (34), 6225−6232.
(27) Johnston, F. H.; Henderson, S. B.; Chen, Y.; Randerson, J. T.; Marlier, M.; Defries, R. S.; Kinney, P.; Bowman, D. M.; Brauer, M. Estimated global mortality attributable to smoke from landscape fires. *Environ. Health Perspect.* **2012**, *120* (5), 695−701.
(28) Gupta, P.; Christopher, S. A., Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. *J. Geophys. Res.: Atmos.* **2009**, *114* (D14).
(29) Gupta, P.; Christopher, S. A.; Wang, J.; Gehrig, R.; Lee, Y.; Kumar, N. Satellite remote sensing of particulate matter and air quality assessment over global cities. *Atmos. Environ.* **2006**, *40* (30), 5880−5892.
(30) Koelemeijer, R. B. A.; Homan, C. D.; Matthijsen, J. Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. *Atmos. Environ.* **2006**, *40* (27), 5304−5315.
(31) Weber, S. A.; Engel-Cox, J. A.; Hoff, R. M.; Prados, A. I.; Zhang, H. An improved method for estimating surface fine particle concentrations using seasonally adjusted satellite aerosol optical depth. *J. Air Waste Manage. Assoc.* **2010**, *60* (5), 574−85.
(32) Zhang, H.; Hoff, R. M.; Engel-Cox, J. A. The relation between Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol optical depth and $PM_{2.5}$ over the United States: A geographical comparison by U.S. Environmental Protection Agency regions. *J. Air Waste Manage. Assoc.* **2009**, *59* (11), 1358−69.
(33) Paciorek, C. J.; Liu, Y.; Moreno-Macias, H.; Kondragunta, S. Spatiotemporal associations between GOES aerosol optical depth retrievals and ground-level $PM_{2.5}$. *Environ. Sci. Technol.* **2008**, *42* (15), 5800−6.
(34) Briggs, D. J.; de Hoogh, C.; Gulliver, J.; Wills, J.; Elliott, P.; Kingham, S.; Smallbone, K. A regression-based method for mapping traffic-related air pollution: Application and testing in four contrasting urban environments. *Sci. Total Environ.* **2000**, *253* (1−3), 151−67.
(35) Jerrett, M.; Arain, A.; Kanaroglou, P.; Beckerman, B.; Potoglou, D.; Sahsuvaroglu, T.; Morrison, J.; Giovis, C. A review and evaluation of intraurban air pollution exposure models. *J. Exposure Anal. Environ. Epidemiol.* **2005**, *15* (2), 185−204.
(36) Kloog, I.; Koutrakis, P.; Coull, B. A.; Lee, H. J.; Schwartz, J. Assessing temporally and spatially resolved $PM_{2.5}$ exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmos. Environ.* **2011**, *45* (35), 6267−6275.
(37) Kloog, I.; Nordio, F.; Coull, B. A.; Schwartz, J. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal $PM_{2.5}$ exposures in the Mid-Atlantic states. *Environ. Sci. Technol.* **2012**, *46* (21), 11913−21.
(38) Liu, Y.; Paciorek, C. J.; Koutrakis, P. Estimating regional spatial and temporal variability of PM(2.5) concentrations using satellite data, meteorology, and land use information. *Environ. Health Perspect* **2009**, *117* (6), 886−92.
(39) Hu, X.; Waller, L. A.; Al-Hamdan, M. Z.; Crosson, W. L.; Estes, M. G.; Estes, S. M.; Quattrochi, D. A.; Sarnat, J. A.; Liu, Y. Estimating ground-level $PM_{2.5}$ concentrations in the southeastern US using geographically weighted regression. *Environ. Res.* **2013**, *121* (Complete), 1−10.
(40) Chang, H. H.; Hu, X.; Liu, Y. Calibrating MODIS aerosol optical depth for predicting daily $PM_{2.5}$ concentrations via statistical downscaling. *J. Exposure Anal. Environ. Epidemiol.* **2014**, *24* (4), 398−404.
(41) Henderson, S. B.; Beckerman, B.; Jerrett, M.; Brauer, M. Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environ. Sci. Technol.* **2007**, *41* (7), 2422−8.
(42) Moore, D. K.; Jerrett, M.; Mack, W. J.; Kunzli, N. A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA. *J. Environ. Monit* **2007**, *9* (3), 246−52.
(43) Ross, Z.; Jerrett, M.; Ito, K.; Tempalski, B.; Thurston, G. A land use regression for predicting fine particulate matter concentrations in the New York City region. *Atmos. Environ.* **2007**, *41* (11), 2255−2269.
(44) Zhang, P. Model selection via multifold cross-validation. *Ann. Stat* **1993**, *21* (1), 299−313.
(45) Hou, W. Z.; Li, Z. Q.; Zhang, Y. H.; Xu, H.; Zhang, Y.; Li, K. T.; Li, D. H.; Wei, P.; Ma, Y., Using support vector regression to predict PM_{10} and $PM_{2.5}$. In *35th International Symposium on Remote Sensing of Environment (Isrse35)*, 2014; Vol. *17*.
(46) Lu, W. Z.; Wang, W. J. Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. *Chemosphere* **2005**, *59* (5), 693−701.
(47) Beckerman, B. S.; Jerrett, M.; Martin, R. V.; van Donkelaar, A.; Ross, Z.; Burnett, R. T. Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California. *Atmos. Environ.* **2013**, *77*, 172−177.
(48) Pandey, G.; Zhang, B.; Jian, L. Predicting submicron air pollution indicators: A machine learning approach. *Environ. Sci.: Processes Impacts* **2013**, *15* (5), 996−1005.
(49) Sayegh, A. S.; Munir, S.; Habeebullah, T. M. Comparing the performance of statistical models for predicting PM_{10} concentrations. *Aerosol Air Qual. Res.* **2014**, *14* (3), 653−665.
(50) CARB. *$PM_{2.5}$ and PM_{10} Natual Event Document Summer 2008 Northern California Wildfires June/July/August 2008*; California Air Resources Board, 2009.
(51) Reid, S. B.; Huang, S.; Pollard, E. K.; Craig, K. J.; Sullivan, D. C.; Zahn, P. H.; MacDonald, C. P.; Raffuse, S. M. *An Almanac for Understanding Smoke Persistence During the 2008 Fire Season*; Sonoma Technology, Inc. Prepared for U.S. Department of Agriculture—Forest Service Pacific Southwest Region, 2009.
(52) McDougall, M., Personal Communication. 2011.
(53) Pfister, G.; Avise, J.; Wiedinmyer, C.; Edwards, D.; Emmons, L.; Diskin, G.; Podolske, J.; Wisthaler, A. CO source contribution analysis for California during ARCTAS-CARB. *Atmos. Chem. Phys.* **2011**, *11* (15), 7515−7532.
(54) Wiedinmyer, C.; Akagi, S. K.; Yokelson, R. J.; Emmons, L. K.; Al-Saadi, J. A.; Orlando, J. J.; Soja, A. J. The Fire INventory from NCAR (FINN): A high resolution global model to estimate the emissions from open burning. *Geosci. Model Dev.* **2011**, *4* (3), 625−641.
(55) Kondragunta, S.; Seybold, M. *Revisions to GOES Aerosol and Smoke Product (GASP) Algorithm*, 2009.
(56) Raffuse, S. M.; McCarthy, M. C.; Craig, K. J.; DeWinter, J. L.; Jumbam, L. K.; Fruin, S.; James Gauderman, W.; Lurmann, F. W. High-resolution MODIS aerosol retrieval during wildfire events in California for use in exposure assessment. *J. Geophys. Res.: Atmos.* **2013**, *118* (19), 11,242−11,255.
(57) *Dynamap/Traffic Counts*; Spatial Insights, Inc., 2000.
(58) Fry, J. A.; Xian, G.; Jin, S.; Dewitz, J. A.; Homer, C. G.; LIMIN, Y.; Barnes, C. A.; Herold, N. D.; Wickham, J. D. Completion of the 2006 national land cover database for the conterminous United States. *Photogramm. Eng. Remote Sens.* **2011**, *77* (9), 858−864.
(59) R Core Team *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
(60) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2* (3), 18−22.
(61) Kuhn, M.Contributions from Jed Wing, S. W., Andre Williams, Chris Keefer and Allan Engelhardt. In *CARET: Classification and Regression Training*, R package version 5, 2012; pp 15−023.

DOI: 10.1021/es505846r
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

(62) Ridgeway, G. *gbm: Generalized Boosted Regression Models.*, R package version 1, 2007; pp 6−3.
(63) Hastie, T. *gam: Generalized Additive Models*; R package version 1.06.2, 2011.
(64) Milborrow, S. *Earth: Multivariate Adaptive Regression Spline Models. Derived from mda:mars by Trevor Hastie and Rob Tibshirani*; R package version 3.2−3, 2012.
(65) Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Software* **2010**, *33* (1), 1−22.
(66) Karatzolou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab—An S4 package for kernel methods in R. *J. Stat. Software* **2004**, *11* (9), 1−20.
(67) Hastie, T.; Efron, B. *lars: Least Angle Regression, Lasso and Forward Stagewise*; R package version 1.1., 2012.
(68) Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **2008**, *28* (5), 1−26.
(69) Anselin, L.; Syabri, I.; Kho, Y. GeoDa: An Introduction to Spatial Data Analysis. *Geogr. Anal.* **2006**, *38* (1), 5−22.
(70) ESRI. *ArcGIS 10.1*; Redlands, CA, 2012.
(71) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed.; Springer-Verlag, 2009; p 763.
(72) Prados, A.; Kondragunta, S.; Ciren, P.; Knapp, K., GOES Aerosol/Smoke Product(GASP) over North America: Comparisons to AERONET and MODIS observations. *J. Geophys. Res.: Atmos.* **2007**, *112*.
(73) Green, M.; Kondragunta, S.; Ciren, P.; Xu, C. Comparison of GOES and MODIS aerosol optical depth (AOD) to aerosol robotic network (AERONET) AOD and IMPROVE $PM_{2.5}$ mass at Bondville, Illinois. *J. Air Waste Manag Assoc* **2009**, *59* (9), 1082−91.
(74) Gupta, P.; Christopher, S. A.; Box, M. A.; Box, G. P. Multi year satellite remote sensing of particulate matter air quality over Sydney, Australia. *Int. J. Remote Sens.* **2007**, *28* (20), 4483−4498.
(75) Pfister, G. G.; Wiedinmyer, C.; Emmons, L. K., Impacts of the fall 2007 California wildfires on surface ozone: Integrating local observations with global model simulations. *Geophys. Res. Lett.* **2008**, *35* (19).
(76) Hu, Y.; Odman, M. T.; Chang, M. E.; Jackson, W.; Lee, S.; Edgerton, E. S.; Baumann, K.; Russell, A. G. Simulation of air quality impacts from prescribed fires on an urban area. *Environ. Sci. Technol.* **2008**, *42* (10), 3676−82.
(77) Choi, Y. J.; Fernando, H. J. Simulation of smoke plumes from agricultural burns: Application to the San Luis/Rio Colorado airshed along the U.S./Mexico border. *Sci. Total Environ.* **2007**, *388* (1−3), 270−89.
(78) Chudnovsky, A. A.; Lee, H. J.; Kostinski, A.; Kotlov, T.; Koutrakis, P. Prediction of daily fine particulate matter concentrations using aerosol optical depth retrievals from the Geostationary Operational Environmental Satellite (GOES). *J. Air Waste Manage. Assoc.* **2012**, *62* (9), 1022−31.
(79) van Donkelaar, A.; Martin, R. V.; Pasch, A. N.; Szykman, J. J.; Zhang, L.; Wang, Y. X.; Chen, D. Improving the accuracy of daily satellite-derived ground-level fine aerosol concentration estimates for North America. *Environ. Sci. Technol.* **2012**, *46* (21), 11971−8.
(80) van Donkelaar, A.; Martin, R. V.; Brauer, M.; Kahn, R.; Levy, R.; Verduzco, C.; Villeneuve, P. J. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. *Environ. Health Perspect.* **2010**, *118* (6), 847−55.
(81) Engel-Cox, J. A.; Hoff, R. M.; Rogers, R.; Dimmick, F.; Rush, A. C.; Szykman, J. J.; Al-Saadi, J.; Chu, D. A.; Zell, E. R. Integrating lidar and satellite optical depth with ambient monitoring for 3-dimensional particulate characterization. *Atmos. Environ.* **2006**, *40* (40), 8056−8067.
(82) Yao, J.; Henderson, S. B. An empirical model to estimate daily forest fire smoke exposure over a large geographic area using air quality, meteorological, and remote sensing data. *J. Exposure Sci. Environ. Epidemiol.* **2014**, *24* (3), 328−35.
(83) Price, O. F.; Williamson, G. J.; Henderson, S. B.; Johnston, F.; Bowman, D. M. The Relationship between particulate pollution levels in Australian Cities, meteorology, and landscape fire activity detected from MODIS hotspots. *PLoS One* **2012**, *7* (10), e47327.
(84) Reid, C. E.; Jerrett, M.; Tager, I.; Petersen, M.; Morefield, P.; Balmes, J. R. Spatiotemporal modeling of wildfire smoke exposure in Northern California using satellite data and chemical transport models. In *Abstracts of the 2013 Conference of the International Society of Environmental Epidemiology (ISEE)*; The International Society of Exposure Science (ISES), and the International Society of Indoor Air Quality and Climate (ISIAQ), August 19−23, 2013; Basel, Switzerland.

DOI: 10.1021/es505846r
Environ. Sci. Technol. XXXX, XXX, XXX−XXX