Published in final edited form as:

Environ Int. 2019 September ; 130: 104909. doi:10.1016/j.envint.2019.104909.

An Ensemble-based Model of $PM_{2.5}$ Concentration across the Contiguous United States with High Spatiotemporal Resolution

Qian Di[a,b], **Heresh Amini**[a], **Liuhua Shi**[a], **Itai Kloog**[c], **Rachel Silvern**[d], **James Kelly**[e], **M. Benjamin Sabath**[f], **Christine Choirat**[f], **Petros Koutrakis**[a], **Alexei Lyapustin**[g], **Yujie Wang**[h], **Loretta J. Mickley**[i], **Joel Schwartz**[a]

[a.]Department of Environmental Health, Harvard T.H. Chan School of Public Heath, Boston, Massachusetts, United States

[b.]Research Center for Public Health, Tsinghua University, Beijing, China

[c.]Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer Sheva, Israel.

[d.]Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts, United States

[e.]U.S. Environmental Protection Agency, Office of Air Quality Planning & Standards, Research Triangle Park, North Carolina, United States

[f.]Department of Biostatistics, Harvard T.H. Chan School of Public Heath, Boston, Massachusetts, United States

[g.]NASA Goddard Space Flight Center, Greenbelt, Maryland, United States

[h.]University of Maryland, Baltimore County, Baltimore, Maryland, United States

[i.]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge Massachusetts, United States

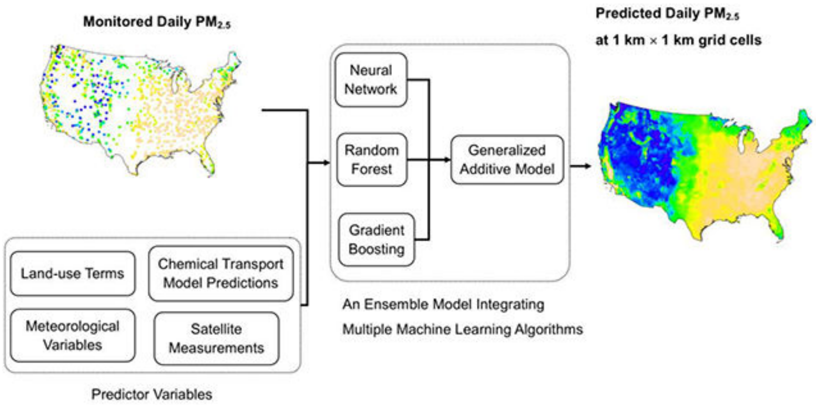
Abstract

Various approaches have been proposed to model $PM_{2.5}$ in the recent decade, with satellite-derived aerosol optical depth, land-use variables, chemical transport model predictions, and several meteorological variables as major predictor variables. Our study used an ensemble model that integrated multiple machine learning algorithms and predictor variables to estimate daily $PM_{2.5}$ at a resolution of 1 km×1 km across the contiguous United States. We used a generalized additive model that accounted for geographic difference to combine $PM_{2.5}$ estimates from neural network, random forest, and gradient boosting. The three machine learning algorithms were based on multiple predictor variables, including satellite data, meteorological variables, land-use variables, elevation, chemical transport model predictions, several reanalysis datasets, and others. The model training results from 2000 to 2015 indicated good model performance with a 10-fold cross-validated R^2 of 0.86 for daily $PM_{2.5}$ predictions. For annual $PM_{2.5}$ estimates, the cross-validated

Competing Financial Interests Declaration:
There are no competing financial interests for the authors of this paper.

R^2 was 0.89. Our model demonstrated good performance up to 60 μg/m^3. Using trained $PM_{2.5}$ model and predictor variables, we predicted daily $PM_{2.5}$ from 2000 to 2015 at every 1 km×1 km grid cell in the contiguous United States. We also used localized land-use variables within 1 km×1 km grids to downscale $PM_{2.5}$ predictions to 100 m × 100 m grid cells. To characterize uncertainty, we used meteorological variables, land-use variables, and elevation to model the monthly standard deviation of the difference between daily monitored and predicted $PM_{2.5}$ for every 1 km×1 km grid cell. This $PM_{2.5}$ prediction dataset, including the downscaled and uncertainty predictions, allows epidemiologists to accurately estimate the adverse health effect of $PM_{2.5}$. Compared with model performance of individual base learners, an ensemble model would achieve a better overall estimation. It is worth exploring other ensemble model formats to synthesize estimations from different models or from different groups to improve overall performance.

Graphical Abstract


Monitored Daily $PM_{2.5}$
Predicted Daily $PM_{2.5}$
at 1 km × 1 km grid cells
Neural Network
Random Forest
Gradient Boosting
Generalized Additive Model
An Ensemble Model Integrating
Multiple Machine Learning Algorithms
Land-use Terms
Chemical Transport Model Predictions
Meteorological Variables
Satellite Measurements
Predictor Variables

Keywords

Fine Particulate Matter ($PM_{2.5}$); Ensemble Model; Neural Network; Gradient Boosting; Random Forest

1. Introduction

$PM_{2.5}$, or fine particulate matter, is a major public health concern (Seaton, MacNee et al. 1995, Cohen, Brauer et al. 2017), and is associated with multiple adverse health outcomes, including mortality (Di, Wang et al. 2017), morbidity (Lippmann, Ito et al. 2000), cardiovascular disease (Pope 2003), respiratory symptoms (Dominici, Peng et al. 2006), diabetes (Yang, Guo et al. 2018), and others. Both long-term and short-term $PM_{2.5}$ exposures are associated with adverse health effects (Kloog, Ridgway et al. 2013, Shi, Zanobetti et al. 2016). Recent studies suggest that $PM_{2.5}$ may penetrate blood brain barrier and enter the brain, causing various neurodegenerative diseases (Maher, Ahmed et al. 2016, de Prado Bert, Mercader et al. 2018).

Accurate estimation of $PM_{2.5}$ is a prerequisite of related epidemiological analyses. As $PM_{2.5}$ concentrations decline in the United States, there is a growing need for accurate estimation of $PM_{2.5}$ at lower levels. A variety of methods have been used to model $PM_{2.5}$, evolving from typical linear regressions to machine learning approaches. In the following, we briefly summarized major methods used for $PM_{2.5}$ modeling to pave the way for our own method.

Linear Regression and Correlational Study:

Ever since Wang et al. (2003) proposed that satellite-derived aerosol optical depth (AOD) could be used to model air quality (Wang and Christopher 2003), AOD has remained an important predictor for $PM_{2.5}$ modeling. Early studies used simple regression models and correlational studies to model $PM_{2.5}$ with AOD (Engel-Cox, Holloman et al. 2004, Koelemeijer, Homan et al. 2006). This approach soon lost popularity and almost disappeared from recent literature, since the relationship between $PM_{2.5}$ and AOD is not straightforward. Explanatory power from a simple linear model was too low to be useful.

Mixed-Effect Model:

Kloog et al. blended AOD with local land-use data and used a mixed-effect model to account for the temporally changing slope between $PM_{2.5}$ and AOD due to meteorological conditions (Kloog, Koutrakis et al. 2011). Use of a mixed-effect model partially resolves the complex relationship between AOD and $PM_{2.5}$. Mixed-effect models received wide application ever since, for its simplicity and flexibility. Chudnovsky et al. used similar methodology on AOD to predict $PM_{2.5}$ concentration within the Northeastern United States (Chudnovsky, Lee et al. 2012). Similar mixed-effect models have been re-applied with slightly different predictor variables, to different study areas, including Mid-Atlantic States in the United States (Kloog, Nordio et al. 2014), Southeastern United States (Lee, Kloog et al. 2015), Mexico City (Just, Wright et al. 2015), Israel (Kloog, Sorek-Hamer et al. 2015), Beijing, China (Xie, Wang et al. 2015), Beijing and neighboring areas of China (Zheng, Zhang et al. 2016), entire China (Ma, Hu et al. 2014), and London, United Kingdom (Beloconi, Kamarianakis et al. 2016),

Geographically Weighted Regression:

Mixed-effect models grant flexibility in the temporal dimension, and their counterparts, geographically weighted regression (GWR) models grant flexibility in the spatial dimension. Hu (2009) found a spatially inconsistent relationship between AOD and $PM_{2.5}$ across the contiguous United States, and used geographically weighted regression to account for the spatially heterogeneous relationship (Hu 2009), and later updated the model for the Southeastern United States (Hu, Waller et al. 2013). van Donkelaar et al. (2015) used AOD and simulation data to obtain high-resolution $PM_{2.5}$ estimation across the North America using geographically weighted regression (van Donkelaar, Martin et al. 2015). Similar methods have been repeated in Pearl River Delta Region (Song, Jia et al. 2014), Beijing and its neighboring areas (Zou, Pu et al. 2016), Central China (Bai, Wu et al. 2016), and entire China (You, Zang et al. 2016).

Generalized Additive Model:

Both mixed-effect models and geographically weighted regression assume a linear relationship between predictor variables and the dependent variable, although the coefficients may vary. Generalized additive models use smoothing functions to account for nonlinear relationships. Paciorek et al. (2008) used generalized additive model and found nonlinear relationships between relative humidity, planetary boundary layer height, $PM_{2.5}$ monitoring data and AOD (Paciorek, Liu et al. 2008). Using a similar method, Liu et al. (2009) applied smoothing functions on AOD and several meteorological variables, and predicted $PM_{2.5}$ for the Northeastern United States (Liu, Paciorek et al. 2009). Similar models have been applied to $PM_{2.5}$ modeling in California (Strawa, Chatfield et al. 2013), and Northern China (Liu, He et al. 2012).

Machine Learning Algorithms:

Neural networks are able to model any kind of nonlinear and interactive relationship given enough data, suitable for modeling $PM_{2.5}$, where the underlying atmospheric dynamics are elusive, and variables have complex interactions (Bishop 1995, Haykin and Network 2004). Gupta et al. (2009) used a neural network and included AOD, relative humidity, planetary boundary layer height, temperature, location, and month. The neural network was trained with monitored 1-hour averaged and 24-hour averaged $PM_{2.5}$ (Gupta and Christopher 2009). Our previous $PM_{2.5}$ model in the contiguous United States was also based on neural networks, but with a larger number of predictor variables. (Di, Kloog et al. 2016). Besides neural networks, other machine learning algorithms have been used in $PM_{2.5}$ modeling for different study areas around the world, such as random forest for a Europe-wide model (Chen, de Hoogh et al. 2018), a US-wide model (Hu, Belle et al. 2017), a regional model for China (Wu, Guo et al. 2011), and Northern China (Huang, Xiao et al. 2018), and some localized models (Brokamp, Jandarov et al. 2017), boosted regression tree (Suleiman, Tight et al. 2016); support vector machine at city level (Sotomayor-Olmedo, Aceves-Fernández et al. 2013, Weizhen, Zhengqiang et al. 2014) and at national level (de Hoogh, Héritier et al. 2018); and gradient boosting (Zhan, Luo et al. 2017).

In summary, we found the following patterns from existing studies, which motivate us to propose our own ensemble-based $PM_{2.5}$ model. **First**, most models used a few predictor variables, but additional variables did contribute to better modeling. Adding extra variables, even when they are intuitively less relevant to $PM_{2.5}$ modeling, improves model performance. For example, NO_2 column measurement (Zheng, Zhang et al. 2016) and other OMI (Ozone Monitoring Instrument) measurements (Strawa, Chatfield et al. 2013) were found to improve models. Penalty terms can avoid overfitting by the use of too many covariates. **Second**, the model performance of different algorithms seems to vary by location and concentration. For example, Chen et al. (2018) found that all machine learning algorithms performed similarly, and no method demonstrated superior external validity (Chen, de Hoogh et al. 2018). A neural network only performed slightly better than a boosting regression tree in London (Suleiman, Tight et al. 2016). Further, geographic weighted regressions indicated spatial variability in predictive performance. Thus, it is theoretically infeasible to have a single model optimally fit all regions. Consequently, there is a recent tendency to use hybrid models instead of a single model, expecting that multiple methods would "complement" each other. A hybrid model usually combines variables or fitting algorithms. Some examples include a hybrid model combining mixed-effect model with land-use regression (Kloog, Chudnovsky et al. 2014), autoregressive-moving-average model with support vector machine (Wang, Zhang et al. 2017), geographically weighted regression with gradient boosting (Zhan, Luo et al. 2017).

Therefore, in this study, we fit a nationwide $PM_{2.5}$ model using a large number of data set as predictor variables and multiple learning methods. We incorporated three machine learning algorithms: neural network, random forest, and gradient boosting. We modeled $PM_{2.5}$ separately with each algorithm and used a geographically weighted generalized additive model as an ensemble model to obtain an overall $PM_{2.5}$ prediction. Predictor variables included satellite-derived AOD, other satellite-based measurements, chemical transport model predictions, land-use variables, meteorological variables, and many others. The ensemble model was validated with 10-fold cross-validation. After model validation, we made predictions of daily $PM_{2.5}$ at 1 km × 1 km grid cells in the contiguous United States from 2000 to 2015. We also predicted the monthly standard deviation of the difference between monitored and predicted $PM_{2.5}$ to quantify the uncertainty of $PM_{2.5}$ modeling in the same 1 km × 1 km grid cells. Finally, we downscaled $PM_{2.5}$ predictions from 1 km × 1 km grid cells to 100 m × 100 m grid cells, using selected downscaling predictor variables. The high resolution $PM_{2.5}$ estimates allow epidemiologists to assess the health effects of $PM_{2.5}$ with higher accuracy, and the $PM_{2.5}$ estimates further help correct residual exposure measurement errors. The ensemble model framework is also useful to combining air pollution models from different sources for future studies.

2. Data

2.1. Study Area

The study area was constrained to the contiguous United States due to data availability, from January 1st, 2000 to December 31st, 2015, in total 5,844 days.

2.2. $PM_{2.5}$ Monitoring Data

Monitoring data for $PM_{2.5}$ were obtained from the Air Quality System (AQS) operated by the Environmental Protection Agency (EPA), The Interagency Monitoring of Protected Visual Environments (IMPROVE), Clean Air Status and Trends Network (CASTNET), and other regional or local monitoring data sets, with 2,156 monitoring sites in our study area from 2000 to 2015. Not all monitoring sites were operating throughout the study period, and many of them operated every 3 or 6 days. We obtained or calculated 24-hour averaged $PM_{2.5}$ and converted the unit into microgram per cubic meter ($\mu g/m^3$) if needed. Monitoring sites were not equally distributed across the study area, with more sites in the Eastern United States, Western coast, and urban areas. Besides, the monitoring network does not adequately sample the full range of concentration scales, due to the limited number of monitoring sites and various monitoring siting criteria. For example, mountainous regions and rural areas had fewer sites.

The $PM_{2.5}$ distribution exhibits some degree of spatial autocorrelation. Monitoring data from nearby monitoring sites are more correlated than data from faraway sites. To leverage spatial autocorrelation and improve model performance, we incorporated spatially lagged monitored $PM_{2.5}$ into the model as predictor variables. Spatially lagged variables were weighted averages of monitored $PM_{2.5}$ from nearby monitoring sites, and weights were inversely proportional to distance and distance squared. We also incorporated one-day lagged and three-day lagged values of spatially lagged terms.

2.3. AOD Measurements and Related Satellite Data

The Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard the Earth Observing System (EOS) satellite has been widely used to measure AOD (Salomonson, Barnes et al. 1989, King, Kaufman et al. 1992). The Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm has been developed to retrieve AOD measurements from raw MODIS data at 1 km × 1 km resolution, especially for conditions of dark vegetated surfaces or bright backgrounds, where conventional AOD retrieval algorithms may have issues (Sayer, Hsu et al. 2013). The absorption optical depth of aerosol species varies with wavelength (Bergstrom, Pilewskie et al. 2007); thus, AOD measurements at different wavelengths are informative to account for different chemical compositions of $PM_{2.5}$ and potentially helpful to achieve accurate modeling. We therefore included AOD measurements at 470 nm and 550 nm from both the Aqua and Terra satellites. AOD measurements retrieved from the deep-blue algorithm (with 10 km resolution) were also included to provide complementary information. Similar to previous studies, AOD measurements (1) with high uncertainty, (2) over water bodies, (3) over snow coverage, and (4) above 1.5 were excluded from modeling, based on the quality assurance flags (Kloog, Koutrakis et al. 2011, Kloog, Chudnovsky et al. 2014).

Previous studies have documented the association between errors in AOD retrievals and surface characteristics (i.e., surface brightness) (Drury, Jacob et al. 2008). MAIAC algorithm, although designed to retrieve AOD from complex background conditions, is not immune to distortion from surface brightness (Lyapustin, Wang et al. 2011). Thus, we incorporated surface reflectance (MOD09A1) from MODIS measurements to control for the impact of surface brightness on AOD data quality (Vermote 2015).

$PM_{2.5}$ composition affects AOD measurements. For example, absorbing aerosols include aerosols from biomass burning, desert dust, and volcanic ash (Herman, Bhartia et al. 1997). The presence of absorbing aerosol components may lead to different AOD values compared with values for the same mass concentration of predominantly scattering components. Similarly, $PM_{2.5}$ with larger portion of black carbon leads to higher AOD values in MODIS retrievals. To address this behavior, we used the absorbing aerosol index (AAI), which measures the portion of absorbing aerosol. We used two AAI measurements at UV and visible ranges (OMAERUVd, OMAEROe) from OMI (Herman, Bhartia et al. 1997, Torres, Bhartia et al. 1998). We also incorporated other measurements from OMI, such as column NO_2 and SO_2 measurements.

In addition to satellite-derived AOD, some reanalysis datasets provide aerosol estimation, such as MERRA2 (Modern-Era Retrospective analysis for Research and Applications, Version 2). Aerosol products from MERRA2 are not measured but simulated, and prone to simulation errors, but they have almost no missing values and serve as good complement to satellite-derived AOD. We used MERRA2 variables related to aerosol concentrations, including sulfate aerosol, hydrophilic black carbon, hydrophobic black carbon, hydrophilic organic carbon, and hydrophobic organic carbon (Buchard, Randles et al. 2017).

2.4. Meteorological Conditions

Meteorological conditions were retrieved from the North American Regional Reanalysis (NOAA) data set, with more details about this data set found elsewhere (Kalnay, Kanamitsu et al. 1996). The full list of 16 meteorological variables can be found in the supplementary material (Section 1).

2.5. Chemical Transport Model

A chemical transport model (CTM) is a numerical simulation model that incorporates emission inventories and simulates the chemical formation, loss, transportation, and deposition of trace gases and particles for a 3-D Eulerian gridded domain. CTMs simulate the concentration of air pollutants at the surface level, and their vertical distribution using process-based mechanistic parameterizations. Predicted vertical profiles allow us to estimate the ground-level contribution to column aerosol concentrations. Since AOD is a column measurement of aerosols, and not the surface-level aerosol concentration, CTM predictions of $PM_{2.5}$ vertical profiles can help calibrate AOD to ground-based monitors (Liu, Park et al. 2004).

CTMs are also able to capture secondary aerosol formation due to complex gas, particle, and multiphase reactions. For example, isoprene is an important precursor of $PM_{2.5}$, but relevant reactions were not well modeled in most CTMs. Recently, GEOS-Chem, a global CTM (http://www.geos-chem.org) (Bey, Jacob et al. 2001), developed a new aqueous-phase mechanism to simulate secondary $PM_{2.5}$ coupled to a detailed gas-phase isoprene oxidation scheme (Marais, Jacob et al. 2016). We obtained daily predictions of total $PM_{2.5}$ mass and mass concentration of several $PM_{2.5}$ components from this new version of GEOS-Chem at 0.5° × 0.625° resolution, as well as from the Community Multiscale Air Quality (CMAQ, www.epa.gov/cmaq) model, a regional CTM that is commonly run with 12-km horizontal resolution over the U.S. (Appel, Napelenok et al. 2017, Kelly, Jang et al. 2018).

2.6. Land-use Variables

Land-use variables are proxies for local emissions and air pollution levels. Chemical transport models are generally unable to simulate air pollution at fine spatial scales, due to the high computational cost and the lack of availability of fine scale emission inventories. Land-use variables approximate emission of air pollutants, often at kilometer or sub-kilometer scale. Previous models often incorporated land-use variables directly, as an approximation of those localized air pollution dynamics. Land-use variables have been used for long-term (e.g., annual or seasonal) exposure assessment in local to continental (Eeftens, Beelen et al. 2012), and global scales (Larkin, Geddes et al. 2017). Recent developments included short-term predictions (e.g. hourly or daily) as well (Son, Osornio-Vargas et al. 2018).

We prepared (1) land-use coverage types, (2) road density, (3) restaurant density, (4) elevation, and (5) NDVI (Normalized difference vegetation index) at 1 km × 1 km grid cells, and also aggregated them to 10 km × 10 km grid cells, to capture the impact of emissions from neighboring areas. Details of data preparation can be found in the supplementary material (Section 2). We used both 1-km and 10-km land-use gridded variables in the model as separate predictor variables. We also aggregated land-use coverage type and elevation from its original resolution to 100 m × 100 m resolution to use as downscaling variables in the localized model.

3. Methods

3.1. Overview

We trained a neural network, random forest, and gradient boosting on all input variables, with parameters of each machine learning algorithm selected by cross-validated grid search processes. We obtained predicted $PM_{2.5}$ from each learner; and then used a geographically weighted generalized additive model as an ensemble model to combine $PM_{2.5}$ estimation. Furthermore, $PM_{2.5}$ concentration exhibits some degree of spatial and temporal autocorrelation, which can be used to improve model performance. We calculated spatially and temporally lagged $PM_{2.5}$ predictions from nearby monitoring sites and neighboring days, treated them as additional input variables, and fit the above models again alongside with existing predictor variables (Figure 1).

To avoid overfitting, we validated our model with 10-fold cross-validation. After splitting all monitoring sites into 10 splits, we trained with 90% of the data and predicted $PM_{2.5}$ at the remaining 10% of monitoring sites. The same process also held for other splits. After assembling $PM_{2.5}$ predictions from all ten splits, we calculated R^2, spatial R^2, and temporal R^2 between predicted and monitored $PM_{2.5}$ at each monitoring site. We also reported R^2 by year.

3.2. Base Learners and Ensemble Model

The details of neural network, random forest and gradient boosting algorithms can be found elsewhere (Bishop 2006). A simple explanation is that all three machine learning algorithms attempt to model the complex relationship between input variables (X’s, predictor variables of $PM_{2.5}$ model) and the dependent variable (Y, or monitored $PM_{2.5}$ for this study) with different approaches. In previous studies, the three algorithms had different model performance under different conditions, probably because of different study areas. By combining the three complementary algorithms, we expect to obtain a better estimate of $PM_{2.5}$.

The performance of each learner depends on hyper parameters, such as how many trees in a random forest, depth of tree; number of layers in a neural network, number of neurons in each layer, lasso penalty, etc. We chose values of hyper parameters for each base learner in a grid search process (Table S1).

A common approach for ensemble averaging is to regress the monitor measurements against the estimates from base learners, with the regression coefficients representing the weight given to each learner. This approach assumes those weights are constant across the country, and do not depend on the pollution concentration. We relaxed these assumptions by regressing the monitored values against thin plate splines of latitude, longitude, and the predicted concentration for each learner. This allows, for example, one learner to have more weight at higher concentrations, but not at lower concentrations; or to have more weight in particular regions of the country.

3.3. Missing Values

Missing values occur among predictor variables. Sometimes missing values can be quite common. Missing values arose from different sources: (1) some predictor variables were unavailable for early years, such as OMI measurements and AOD measurements from the Aqua satellite, because the satellites were launched in 2005 and 2002, respectively. (2) Measurements were not possible over some locations and time, such as AOD over clouds or snow, or soil moisture near a waterbody. (3) Data processing removed some measurements due to high uncertainty, such as AOD measurements greater than 1.5, which were excluded. To predict $PM_{2.5}$ concentration for the entire study area and during the entire study period, some method is required to fill in the missing values.

A good method should predict what the values would have been had they not been missing, so we used variables without missing values to predict each variable with missing values. We identified all variables with no missing values and used them as predictor variables of a random forest. For example, AOD measurement at 550 nm (AOD 550) has more than 50% missing values. We used CMAQ and GEOS-Chem predictions, land-use types, and meteorological variables as predictors of a random forest, since these variables have no missing values, and trained the random forest when AOD 550 were available. Then we predicted AOD 550 when AOD 550 were missing to fill in the missing values. Again, a grid search was used to obtain the best parameter combinations.

Some land-use measurements were intermittent and unavailable over a certain period. For example, NDVI and surface reflectance measurements are available every 16 and 8 days; all land-use terms from the NLCD were available every 5 years. Since land-surface characteristics can be assumed to change gradually, we use simple linear interpolation to fill in missing values.

3.4. Model Prediction

After filling in missing values and interpolating, all input variables were available across the study area. We trained the three base learners and the ensemble model with input variables and monitored $PM_{2.5}$ as the dependent variable, and then used trained models and predictor variables at each 1 km × 1 km grid cell to predict $PM_{2.5}$.

We also downscaled the 1-km-level predictions to 100 m × 100 m grid cells. We took the difference between monitored and predicted $PM_{2.5}$ at monitoring sites, and used downscaling predictors within 100 m of the monitor as predictors in a random forest to predict the difference. The downscaling predictors include NLCD land-use cover types, road density, and elevation at 100-m level, as well as meteorological variables: air temperature, humidity, wind speed, and planetary boundary layer height. We trained the random forest at each monitoring site and predicted the within-cell variations at every 100 m × 100 m grid cell.

Finally, we estimated the spatial and temporal pattern of model uncertainty. Model performance is determined by various factors and varies by location and time. By referring to previous studies and using prior knowledge, we identified several variables that edict model performance and used them to model monthly standard deviation of the difference between daily monitored and predicted $PM_{2.5}$ at each monitoring site:

$$sd_{PM_{ij}} = f_1(Location_i) + f_2\left(Location_i, \widehat{PM}_{ij}\right) + f_3(elevation) + f_4(\ surface\ reflectance\ ) + f_5(humidity) \\ + f_6(\ tree\ canopy\ ) + f_8(NDVI) + f_9(urban) + Year + e_{ij}$$

where $sd_{PM_{ij}}$ is the monthly standard deviation of the difference between daily monitored and predicted $PM_{2.5}$ at location i and in the month j; f_1 denotes a penalized spline for location i; f_2 denotes a thin-plate spline for an interaction between location i and predicted $PM_{2.5}$ at location i and in the month j; $f_3 \sim f_9$ denote splines for predictors at location i; $\widehat{PM}_{ij}$ is the mean predicted $PM_{2.5}$ at a location i in the month j; and e_{ij} is the error term.

4. Results

Table 1 presents the cross-validated R^2 by year. R^2 values ranged from 0.75 to 0.90, with an average of 0.86, indicating good model performance. The spatial R^2 ranged from 0.73 to 0.91, with an average of 0.89, demonstrating that our model can well capture the spatial variation of long-term $PM_{2.5}$. The average RMSE (root mean square error) was 1.26 μg/m^3 spatially, and 2.53 μg/m^3 temporally. There is a noticeable improvement compared with our previous model (Di, Kloog et al. 2016). Of the three machine learning algorithms, model performance of neural network and random forest was quite close, and better than gradient boosting. Overall, neural network outperformed random forest (R^2: 0.855 vs. 0.854, Table 1); but random forest outperformed for some years (Table 1), some regions (Table 2), all seasons except summer (Table 3), and spatially (Table S2). The overall model performance from the ensemble model outperformed that from any single algorithm.

Figure 2 displays the cross-validated R^2 at each monitoring site, with high R^2 in most areas of the Eastern United States and parts of the West Coast. For mountainous regions, especially the Appalachian and Rocky Mountains, we obtained lower R^2, indicating that mountainous terrain has an important influence on model performance (Table 2). The spatial pattern of model performance for this study was similar to our previous model and other previous studies (Engel-Cox, Holloman et al. 2004). The predicted uncertainty demonstrated a similar spatial pattern. If examined by season, model performed well in the autumn and less unsatisfyingly in winter.

While the incremental R^2 from ensemble averaging compared to the best single learner was not large overall, it affected the linearity of the association between measured and predicted $PM_{2.5}$. For the ensemble, a spline fit between daily predicted and monitored $PM_{2.5}$ is almost a straight 1:1 line up to 60 μg/m^3, a concentration seldom seen in the United States, demonstrating satisfying performance even at high concentration, when monitoring data were scarce (Figure 3). The performance of the individual base learners was worse than the ensemble average. The random forest overestimated at high concentrations. Overestimation was even more serious for the gradient boosting. Overall, the ensemble model improved model performance although quite close to the neural network. This pattern exemplifies the usefulness of ensemble averaging, and the use of splines on individual learners to do the averaging. The spline comparing monitored to predicted annual $PM_{2.5}$ was almost a straight 1:1 line, indicating good fit at the annual level (Figure 4).

We reported the variable importance of predictor variables from three machine learning algorithms in Table 4. Spatially lagged $PM_{2.5}$ from nearby monitoring sites was clearly an important predictor. For the random forest and gradient boosting, spatially lagged $PM_{2.5}$ contributed significantly to model performance, followed by CMAQ predictions. The relative contribution of predictor variables varies by algorithms, and the contribution was spread out across more predictor variables for the neural network than random forest and gradient boosting. AOD related variables contributed the most to the neural network, along with latitude, longitude and other land-use variables.

For spatial distribution, $PM_{2.5}$ concentrations were higher in populous places, such as Los Angeles, and the entire Eastern United States, excluding the Appalachian Mountains and some remote areas in Northern Maine and Florida. The Central Valley of California also had high concentrations. $PM_{2.5}$ concentration dropped significantly after 2008 (Figure 5). The hotspots on the 2015 map were almost indistinguishable under the same color scale (Figure 6).

The local regression predicting address-specific differences from the 1-km average was examined in the Boston metropolitan area. Figur e 7 shows the estimated concentrations on 100 m × 100 m grid.

5. Discussion

Our ensemble model incorporates $PM_{2.5}$ predictions from three machine learning algorithms, including neural network, random forest, and gradient boosting, and achieved excellent performance, with a spatial R^2 of 0.89 and spatial RMSE of 1.26 μg/m^3. Temporal R^2 was also good (0.85). The three machine learning algorithms used more than 100 predictor variables, ranging from satellite data, land-use data, meteorological data, and CTM predictions, with cross-validation controlling for overfitting. A generalized additive model combined $PM_{2.5}$ estimates from machine learning algorithms and allowed the contribution of each algorithm to vary by location. With the trained model, we predicted daily $PM_{2.5}$ for the entire contiguous United States from 2000 to 2015 at every 1 km × 1 km grid cell. This high-resolution accurate estimation allows the estimation of both short-term and long-term exposures to $PM_{2.5}$. The modeled uncertainty of $PM_{2.5}$ enables to further correct exposure assessment errors in epidemiology.

Our $PM_{2.5}$ model outperforms previous models and our own previous model in the following ways. First, our model achieved high R^2, with better agreement between monitored $PM_{2.5}$ and predicted $PM_{2.5}$. The cross-validated R^2 (0.86) was higher compared with our previous model that was solely based on neural network (R^2 0.84), a geographically weighted regression model with AOD as input (R^2 0.67 in the east and 0.22 in the west) (Hu

2009), another a geographically weighted regression model with GEOS-Chem and AOD as inputs (R^2 0.82) (van Donkelaar, Martin et al. 2015), and other studies estimating $PM_{2.5}$ for the entire contiguous United States (Liu, Park et al. 2004). Moreover, our model achieved large spatiotemporal coverage and high spatiotemporal resolution at the same time, and demonstrated a potential to downscale to 100-meter level. Previous studies either achieved large coverage or high resolution, and few of them achieve both at the same time. Finally, our model estimated the monthly uncertainty of $PM_{2.5}$ prediction for every 1 km grid cell, being the first study of this kind.

Our model outperformed previous models for the following three reasons. **First**, we incorporated a larger number of predictor variables, including two CTMs. Compared with our previous model, we added more detailed classification of land-use types, and new variables such as restaurant density. For different land-use types, the emission profiles and rates vary, and such differences are informative for modeling $PM_{2.5}$ at local scales. The variable importance of land-use types also proved the importance of such detailed classification (Table 4). **Second**, we developed an approach to fill in missing values. Missing values in satellite data, due to cloud or snow coverage, is a concern for $PM_{2.5}$ modeling, particularly since they are not missing at random. Previous studies either ignored the missing values (potentially biasing long-term averages), or used smoothing with inverse probability of missingness weights (Kloog, Koutrakis et al. 2011), or multiple imputation (Huang, Xiao et al. 2018). Our previous model used a simple interpolation to fill in missing values for some predictor variables. For this study, we developed a separate prediction model for each variable and filled in the missing values before model training and model prediction. This strategy is computationally intensive, but improves model performance, with most of the gain in spatial R^2, indicating better model performance at annual level. **Third**, the three machine learning algorithms complement each other. While the neural network and random forest had similar overall performance, they did not perform equally well in every location (Table 2). The spline plots also demonstrate that the three machine learning algorithms do not perform equally well at all concentration levels (Figure 3). The neural network was better at capturing temporal variation of $PM_{2.5}$, while the random forest modeled spatial variation better (Table S2). By non-linearly combining different base learners and allowing their contributions to vary by location and concentration, the base learner that performs better at a specific location or concentration level contributes more to the ensemble model in that instance. Therefore, the ensemble model can outperform the individual machine learning algorithms.

Our $PM_{2.5}$ model revealed spatial and temporal trends in $PM_{2.5}$ levels in the contiguous United States. Overall, there was an east-west gradient in $PM_{2.5}$ concentration, with the Eastern United States, except the Appalachian Mountains and some remote areas of Maine, having relatively higher $PM_{2.5}$ concentration than the Western United States, where most areas are either mountainous or covered by desert (Figure 6). There are some hotspots in the Western U.S., such as the Central Valley of California. At a small spatial scale, $PM_{2.5}$ concentration is primarily driven by land-use (Figure 7). As revealed by our localized modeling, $PM_{2.5}$ concentrations near highways are elevated, consistent with recent near-road $PM_{2.5}$ monitoring (DeWinter, Brown et al. 2018). In terms of temporal pattern, $PM_{2.5}$

concentrations decreased noticeably after 2008 (Figure 5). This may be due to a combination of economic recession and emission controls, particularly the cross-state air pollution rule, which reduced emissions from coal fired power plants. The time series of $PM_{2.5}$ concentration also indicate strong seasonal patterns, with peaks in the summer (Figure 5). High summer concentrations were observed particularly in the Southeastern United States, likely due to increased secondary organic aerosol associated with isoprene and monoterpene emissions from trees under conducive summer conditions (Figure 6) (Sharkey, Singsaas et al. 1996, Sharkey, Wiberley et al. 2008, Zhang, Yee et al. 2018). Relatively low regional $PM_{2.5}$ was observed in winter due to reduced photochemistry, associated with shorter daytime, lower sunlight intensity, colder temperatures, and reduced biogenic and wildfire emissions. However, elevated wintertime $PM_{2.5}$ did occur in localized areas due to increased emission from home heating in combination with reduced meteorological mixing. For example, Salt Lake City and Central Valley experienced wintertime $PM_{2.5}$ episodes due to temperature inversions trapping emission from residential wood combustion and other resources in the valley (Franchin, Fibiger et al. 2018, Kelly, Parworth et al. 2018). Elevated wintertime $PM_{2.5}$ was also observed in the Ohio River Valley, where power plants are concentrated. Model performance degraded slightly in recent years (Table 1). Although the reason for this trend is not entirely clear, it may be associated with CMAQ, an important predictor variable. Annual R^2 values for CMAQ sulfate predictions in the Ohio River Valley were negatively correlated with year during 2007–2015 ($r = -0.74$). Although the mean bias in CMAQ predictions improved over these years, the degradation in R^2 suggests that predicting the variability of some pollutants with CTMs may be more challenging under lower air pollution concentrations.

The relative importance of predictor variables varied by machine learning algorithm. Gradient boosting heavily depended on the spatially lagged $PM_{2.5}$, followed by CMAQ predictions, standard deviation of elevation, and land-use terms. The random forest demonstrated a similar pattern with less contribution from spatially lagged $PM_{2.5}$. $PM_{2.5}$ distribution demonstrates a high degree of spatial and temporal autocorrelation, and that is why spatially lagged $PM_{2.5}$ could be an important predictor variable for both gradient boosting and random forest. But for the neural network, AOD variables were the primary predictor, followed by spatially lagged $PM_{2.5}$, road density, latitude, and longitude. The contribution of spatially lagged $PM_{2.5}$ was negligible for the neural network (2.68%), suggesting that the neural network may find some complex and nonlinear associations between AOD and other predictors to predict $PM_{2.5}$. For example, the AOD-$PM_{2.5}$ relationship not only relies on temperature, humidity and other meteorological conditions, but also demonstrates regional difference. The AOD-$PM_{2.5}$ relationship in the Southeastern United States, where primary source of $PM_{2.5}$ is from tree emission, is different from the Northeastern United States, where primary source is from power plants and vehicles. Also, the elevation variation in the Mountainous region imposes challenge to modeling, and simulating such complex relationship is the strength of neural network method. In comparison, without the ability of uncovering such complex relationship, gradient boosting and random forest use the superficial and obvious auto-correlation to estimate $PM_{2.5}$.

Our study suggests that the best learning algorithm varies based on context (e.g., year, season, location, concentration, etc.), and a single fitting method will not be optimal for air pollution modeling in all situations. Approaches that integrate and synthesize individual base learners could be developed to achieve a better overall estimation. Performance improvement for ensemble model in this study seems to be negligible, because the three base learners all performed relatively well and used the same predictor predictors, and there is little extra information by combining them. In practice, if combining models from different research groups or with different predictor variables, performance improvement would be more obvious. Our ensemble model demonstrates features of both geographically weighted regression and generalized additive models, showing flexibility and good performance. As more researchers explored different approaches of air pollution modeling, it is worth exploring other ensemble model formats to synthesize different models or estimation from different research groups in order to obtain an optimized overall estimation.

Our $PM_{2.5}$ model can benefit subsequent epidemiological studies in multiple ways. First, our $PM_{2.5}$ model exhibits high overall model performance. As environmental epidemiological studies are important to inform air quality standard setting and receive more scrutiny, accurate exposure assessment is both essential and critical, especially as pollution concentrations decrease. Moreover, our $PM_{2.5}$ model performs particularly well at predicting annual averages, the standard metric used to assess long-term health effects of $PM_{2.5}$. Also, the performance is good at both low and high concentrations, including daily levels up to 60 $\mu g/m^3$. Finally, we have quantified model uncertainty in the $PM_{2.5}$ prediction, which will allow subsequent studies to take into account exposure measurement error (Spiegelman 2016).

Conclusion

We used an ensemble model to integrate neural network, random forest and gradient boosting to estimate daily $PM_{2.5}$ from 2000 to 2015 for the entire contiguous United States. Predictor variables included satellite measurements, chemical transport model predictions, land-use terms, meteorological variables, etc. After cross-validation, the mean R^2 between daily predicted and monitored $PM_{2.5}$ was 0.86, with RMSE 2.79 $\mu g/m^3$. R^2 was 0.89 at annual level, indicating good model performance. After model training, the model produced daily $PM_{2.5}$ predictions at 1 km × 1 km grid cells. We further downscaled the 1-km-level predictions to 100 m × 100 m levels, with additional downscaling predictors. We also predicted monthly standard deviation of the difference between daily monitored and predicted $PM_{2.5}$ at 1 km × 1 km grid cells. By comparing model performance of individual machine learning algorithms, we found that a single machine learning algorithm may underperform at a particular year, season, location, pollution concentration etc., and an ensemble model incorporating estimation from these multiple machine learning algorithms can achieve a superior model performance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements, Including Grant Information

This publication was made possible by U.S. EPA grant numbers RD-834798, RD-835872, and 83587201; HEI grant 4953-RFA14-3/16-4. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the U.S. EPA. The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency. Further, the U.S. EPA does not endorse the purchase of any commercial products or services mentioned in the publication. Research described in this article was also conducted under contract to the Health Effects Institute (HEI), an organization jointly funded by the United States Environmental Protection Agency (EPA) (Assistance Award No.CR-83467701) and certain motor vehicle and engine manufacturers. The contents of this article do not necessarily reflect the views of HEI, or its sponsors, nor do they necessarily reflect the views and policies of the EPA or motor vehicle and engine manufacturers. The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

References

Appel KW, Napelenok SL, Foley KM, Pye HOT, Hogrefe C, Luecken DJ, Bash JO, Roselle SJ, Pleim JE, Foroutan H, Hutzell WT, Pouliot GA, Sarwar G, Fahey KM, Gantt B, Gilliam RC, Heath NK, Kang D, Mathur R, Schwede DB, Spero TL, Wong DC and Young JO (2017). “Description and evaluation of the Community Multiscale Air Quality (CMAQ) modeling system version 5.1.” Geosci Model Dev 10(4): 1703–1732. [PubMed: 30147852]

Bai Y, Wu L, Qin K, Zhang Y, Shen Y and Zhou Y (2016). “A geographically and temporally weighted regression model for ground-level PM2. 5 estimation from satellite-derived 500 m resolution AOD.” Remote Sensing 8(3): 262.

Beloconi A, Kamarianakis Y and Chrysoulakis N (2016). “Estimating urban PM10 and PM2. 5 concentrations, based on synergistic MERIS/AATSR aerosol observations, land cover and morphology data.” Remote Sensing of Environment 172: 148–164.

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

Bey I, Jacob DJ, Yantosca RM, Logan JA, Field BD, Fiore AM, Li Q, Liu HY, Mickley LJ and Schultz MG (2001). “Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation.” Journal of Geophysical Research 106: 23073.

Bishop CM (1995). Neural networks for pattern recognition, Oxford university press.

Bishop CM (2006). Pattern recognition and machine learning. New York, Springer.

Brokamp C, Jandarov R, Rao M, LeMasters G and Ryan P (2017). “Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches.” Atmospheric Environment 151: 1–11. [PubMed: 28959135]

Buchard V, Randles CA, da Silva AM, Darmenov A, Colarco PR, Govindaraju R, Ferrare R, Hair J, Beyersdorf AJ, Ziemba LD and Yu H (2017). “The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part II: Evaluation and Case Studies.” Journal of Climate 30(17): 6851–6872.

Chen J, de Hoogh K, Strak M, Kerckhoffs J, Vermeulen R, Brunekreef B and Hoek G (2018). OP III–4 Exposure assessment models for no2 and pm2. 5 in the elapse study: a comparison of supervised linear regression and machine learning approaches, BMJ Publishing Group Ltd.

Chudnovsky AA, Lee HJ, Kostinski A, Kotlov T and Koutrakis P (2012). “Prediction of daily fine particulate matter concentrations using aerosol optical depth retrievals from the Geostationary Operational Environmental Satellite (GOES).” Journal of the Air & Waste Management Association 62: 1022–1031. [PubMed: 23019816]

Cohen AJ, Brauer M, Burnett R, Anderson HR, Frostad J, Estep K, Balakrishnan K, Brunekreef B, Dandona L and Dandona R (2017). “Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015.” The Lancet 389(10082): 1907–1918.

de Hoogh K, Héritier H, Stafoggia M, Künzli N and Kloog I (2018). “Modelling daily PM2.5 concentrations at high spatio-temporal resolution across Switzerland.” Environmental Pollution 233: 1147–1154. [PubMed: 29037492] de Prado Bert P, Mercader EMH, Pujol J, Sunyer J and Mortamais M (2018). “The Effects of Air Pollution on the Brain: a Review of Studies Interfacing Environmental Epidemiology and Neuroimaging.” Current environmental health reports 5(3): 351–364. [PubMed: 30008171]

DeWinter JL, Brown SG, Seagram AF, Landsberg K and Eisinger DS (2018). “A national-scale review of air pollutant concentrations measured in the US near-road monitoring network during 2014 and 2015.” Atmospheric Environment 183: 94–105.

Di Q, Kloog I, Koutrakis P, Lyapustin A, Wang Y and Schwartz J (2016). “Assessing PM2. 5 Exposures with High Spatiotemporal Resolution across the Continental United States.” Environmental science & technology 50(9): 4712–4721. [PubMed: 27023334]

Di Q, Wang Y, Zanobetti A, Wang Y, Koutrakis P, Choirat C, Dominici F and Schwartz JD (2017). “Air pollution and mortality in the Medicare population.” New England Journal of Medicine 376(26): 2513–2522. [PubMed: 28657878]

Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL and Samet JM (2006). “Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases.” Jama 295(10): 1127–1134. [PubMed: 16522832]

Drury E, Jacob DJ, Wang J, Spurr RJ and Chance K (2008). “Improved algorithm for MODIS satellite retrievals of aerosol optical depths over western North America.” Journal of Geophysical Research: Atmospheres (1984–2012) 113(D16).

Eeftens M, Beelen R, de Hoogh K, Bellander T, Cesaroni G, Cirach M, Declercq C, Dedele A, Dons E and de Nazelle A (2012). “Development of land use regression models for PM2. 5, PM2. 5 absorbance, PM10 and PMcoarse in 20 European study areas; results of the ESCAPE project.” Environmental science & technology 46(20): 11195–11205. [PubMed: 22963366]

Engel-Cox JA, Holloman CH, Coutant BW and Hoff RM (2004). “Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality.” Atmospheric Environment 38(16): 2495–2509.

Franchin A, Fibiger DL, Goldberger L, McDuffie EE, Moravek A, Womack CC, Crosman ET, Docherty KS, Dube WP and Hoch SW (2018). “Airborne and ground-based observations of ammonium-nitrate-dominated aerosols in a shallow boundary layer during intense winter pollution episodes in northern Utah.” Atmospheric Chemistry and Physics 18(23): 17259–17276.

Gedeon TD (1997). “Data mining of inputs: analysing magnitude and functional measures.” International Journal of Neural Systems 8(02): 209–218. [PubMed: 9327276]

Gupta P and Christopher SA (2009). “Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach.” Journal of Geophysical Research: Atmospheres 114(D20).

Haykin S and Network N (2004). “A comprehensive foundation.” Neural Networks 2(2004).

Herman J, Bhartia P, Torres O, Hsu C, Seftor C and Celarier E (1997). “Global distribution of UV-absorbing aerosols from Nimbus 7/TOMS data.” J. Geophys. Res 102(16): 911–916.

Hu X, Belle JH, Meng X, Wildani A, Waller LA, Strickland MJ and Liu Y (2017). “Estimating PM2. 5 concentrations in the conterminous United States using the random forest approach.” Environmental Science & Technology 51(12): 6936–6944. [PubMed: 28534414]

Hu X, Waller LA, Al-Hamdan MZ, Crosson WL, Estes MG Jr, Estes SM, Quattrochi DA, Sarnat JA and Liu Y (2013). “Estimating ground-level PM2. 5 concentrations in the southeastern US using geographically weighted regression.” Environmental Research 121: 1–10. [PubMed: 23219612]

Hu Z (2009). “Spatial analysis of MODIS aerosol optical depth, PM 2.5, and chronic coronary heart disease.” International journal of health geographics 8(1): 27. [PubMed: 19435514]

Huang K, Xiao Q, Meng X, Geng G, Wang Y, Lyapustin A, Gu D and Liu Y (2018). “Predicting monthly high-resolution PM2. 5 concentrations with random forest model in the North China Plain.” Environmental pollution 242: 675–683. [PubMed: 30025341]

Just AC, Wright RO, Schwartz J, Coull BA, Baccarelli AA, Tellez-Rojo MM, Moody E, Wang Y, Lyapustin A and Kloog I (2015). “Using high-resolution satellite aerosol optical depth to estimate daily PM2. 5 geographical distribution in Mexico City.” Environmental science & technology 49(14): 8576–8584. [PubMed: 26061488]

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo

KC, Ropelewski C, Wang J, Jenne R and Joseph D (1996). "The NCEP/NCAR 40-Year Reanalysis Project." Bulletin of the American Meteorological Society 77: 437–471.

Kelly JT, Jang CJ, Timin B, Gantt B, Reff A, Zhu Y, Long S and Hanna A (2018). "A system for developing and projecting PM2.5 spatial fields to correspond to just meeting National Ambient Air Quality Standards." Atmospheric Environment.

Kelly JT, Parworth CL, Zhang Q, Miller DJ, Sun K, Zondlo MA, Baker KR, Wisthaler A, Nowak JB and Pusede SE (2018). "Modeling NH4NO3 over the San Joaquin Valley during the 2013 DISCOVER‐AQ campaign." Journal of Geophysical Research: Atmospheres 123(9): 4727–4745.

King MD, Kaufman YJ, Menzel WP and Tanre D (1992). "Remote sensing of cloud, aerosol, and water vapor properties from the Moderate Resolution Imaging Spectrometer (MODIS)." Geoscience and Remote Sensing, IEEE Transactions on 30(1): 2–27.

Kloog I, Chudnovsky AA, Just AC, Nordio F, Koutrakis P, Coull BA, Lyapustin A, Wang Y and Schwartz J (2014). "A new hybrid spatio-temporal model for estimating daily multi-year PM 2.5 concentrations across northeastern USA using high resolution aerosol optical depth data." Atmospheric Environment 95: 581–590. [PubMed: 28966552]

Kloog I, Koutrakis P, Coull BA, Lee HJ and Schwartz J (2011). "Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements." Atmospheric Environment 45: 6267–6275.

Kloog I, Nordio F, Zanobetti A, Coull BA, Koutrakis P and Schwartz JD (2014). "Short term effects of particle exposure on hospital admissions in the Mid-Atlantic states: a population estimate." PloS one 9(2): e88578. [PubMed: 24516670]

Kloog I, Ridgway B, Koutrakis P, Coull BA and Schwartz JD (2013). "Long-and short-term exposure to PM2. 5 and mortality: using novel exposure models." Epidemiology (Cambridge, Mass.) 24(4): 555.

Kloog I, Sorek-Hamer M, Lyapustin A, Coull B, Wang Y, Just AC, Schwartz J and Broday DM (2015). "Estimating daily PM2. 5 and PM10 across the complex geo-climate region of Israel using MAIAC satellite-based AOD data." Atmospheric Environment 122: 409–416. [PubMed: 28966551]

Koelemeijer R, Homan C and Matthijsen J (2006). "Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe." Atmospheric Environment 40(27): 5304–5315.

Larkin A, Geddes JA, Martin RV, Xiao Q, Liu Y, Marshall JD, Brauer M and Hystad P (2017). "Global land use regression model for nitrogen dioxide air pollution." Environmental science & technology 51(12): 6957–6964. [PubMed: 28520422]

Lee M, Kloog I, Chudnovsky A, Lyapustin A, Wang Y, Melly S, Coull B, Koutrakis P and Schwartz J (2015). "Spatiotemporal prediction of fine particulate matter using high-resolution satellite images in the Southeastern US 2003–2011." Journal of Exposure Science and Environmental Epidemiology.

Lippmann M, Ito K, Nadas A and Burnett R (2000). "Association of particulate matter components with daily mortality and morbidity in urban populations." Research report (Health Effects Institute)(95): 5–72, discussion 73–82.

Liu Y, He K, Li S, Wang Z, Christiani DC and Koutrakis P (2012). "A statistical model to evaluate the effectiveness of PM2. 5 emissions control during the Beijing 2008 Olympic Games." Environment international 44: 100–105. [PubMed: 22406019]

Liu Y, Paciorek CJ and Koutrakis P (2009). "Estimating Regional Spatial and Temporal Variability of PM2.5 Concentrations Using Satellite Data, Meteorology, and Land Use Information." Environmental Health Perspectives 117: 886–892. [PubMed: 19590678]

Liu Y, Park RJ, Jacob DJ, Li Q, Kilaru V and Sarnat JA (2004). "Mapping annual mean ground‐level PM2. 5 concentrations using Multiangle Imaging Spectroradiometer aerosol optical thickness over the contiguous United States." Journal of Geophysical Research: Atmospheres 109(D22).

Lyapustin A, Wang Y, Laszlo I, Kahn R, Korkin S, Remer L, Levy R and Reid J (2011). "Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm." Journal of Geophysical Research: Atmospheres (1984–2012) 116(D3).

Ma Z, Hu X, Huang L, Bi J and Liu Y (2014). “Estimating ground-level PM2. 5 in China using satellite remote sensing.” Environmental science & technology 48(13): 7436–7444. [PubMed: 24901806]

Maher BA, Ahmed IA, Karloukovski V, MacLaren DA, Foulds PG, Allsop D, Mann DM, Torres-Jardón R and Calderon-Garciduenas L (2016). “Magnetite pollution nanoparticles in the human brain.” Proceedings of the National Academy of Sciences 113(39): 10797–10801.

Marais EA, Jacob DJ, Jimenez JL, Campuzano-Jost P, Day DA, Hu W, Krechmer J, Zhu L, Kim PS and Miller CC (2016). “Aqueous-phase mechanism for secondary organic aerosol formation from isoprene: application to the southeast United States and co-benefit of SO 2 emission controls.” Atmospheric Chemistry and Physics 16(3): 1603–1618.

Paciorek CJ, Liu Y, Moreno-Macias H and Kondragunta S (2008). “Spatiotemporal associations between GOES aerosol optical depth retrievals and ground-level PM2. 5.” Environmental science & technology 42(15): 5800–5806. [PubMed: 18754512]

Pope CA (2003). “Cardiovascular Mortality and Long-Term Exposure to Particulate Air Pollution: Epidemiological Evidence of General Pathophysiological Pathways of Disease.” Circulation 109: 71–77. [PubMed: 14676145]

Rifkin R and Klautau A (2004). “In defense of one-vs-all classification.” Journal of machine learning research 5(Jan): 101–141.

Salomonson VV, Barnes W, Maymon PW, Montgomery HE and Ostrow H (1989). “MODIS: Advanced facility instrument for studies of the Earth as a system.” Geoscience and Remote Sensing, IEEE Transactions on 27(2): 145–153.

Sayer A, Hsu N, Bettenhausen C and Jeong MJ (2013). “Validation and uncertainty estimates for MODIS Collection 6 “Deep Blue” aerosol data.” Journal of Geophysical Research: Atmospheres 118(14): 7864–7872.

Seaton A, MacNee W, Donaldson K and Godden D (1995). “Particulate air pollution and acute health effects.” Lancet 345: 176–178. [PubMed: 7741860]

Sharkey TD, Singsaas EL, Vanderveer PJ and Geron C (1996). “Field measurements of isoprene emission from trees in response to temperature and light.” Tree Physiology 16(7): 649–654. [PubMed: 14871703]

Sharkey TD, Wiberley AE and Donohue AR (2008). “Isoprene emission from plants: why and how.” Annals of Botany 101(1): 5–18. [PubMed: 17921528]

Shi L, Zanobetti A, Kloog I, Coull BA, Koutrakis P, Melly SJ and Schwartz JD (2016). “Low-concentration PM2. 5 and mortality: Estimating acute and chronic effects in a population-based study.” Environmental health perspectives 124(1): 46. [PubMed: 26038801]

Son Y, Osornio-Vargas ÁR, O’Neill MS, Hystad P, Texcalac-Sangrador JL, Ohman-Strickland P, Meng Q and Schwander S (2018). “Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters.” Science of The Total Environment 639: 40–48. [PubMed: 29778680]

Song W, Jia H, Huang J and Zhang Y (2014). “A satellite-based geographically weighted regression model for regional PM2. 5 estimation over the Pearl River Delta region in China.” Remote Sensing of Environment 154: 1–7.

Sotomayor-Olmedo A, Aceves-Fernández MA, Gorrostieta-Hurtado E, Pedraza-Ortega C, Ramos-Arreguín JM and Vargas-Soto JE (2013). “Forecast urban air pollution in Mexico City by using support vector machines: A kernel performance approach.” International Journal of Intelligence Science 3(03): 126.

Spiegelman D (2016). “Evaluating Public Health Interventions: 4. The Nurses’ Health Study and Methods for Eliminating Bias Attributable to Measurement Error and Misclassification.” American Journal of Public Health 106(9): 1563–1566. [PubMed: 27509282]

Strawa A, Chatfield R, Legg M, Scarnato B and Esswein R (2013). “Improving retrievals of regional fine particulate matter concentrations from Moderate Resolution Imaging Spectroradiometer (MODIS) and Ozone Monitoring Instrument (OMI) multisatellite observations.” Journal of the Air & Waste Management Association 63(12): 1434–1446. [PubMed: 24558706]

Suleiman A, Tight M and Quinn A (2016). “Hybrid neural networks and boosted regression tree models for predicting roadside particulate matter.” Environmental Modeling & Assessment 21(6): 731–750.

Torres O, Bhartia P, Herman J, Ahmad Z and Gleason J (1998). “Derivation of aerosol properties from satellite measurements of backscattered ultraviolet radiation: Theoretical basis.” Journal of Geophysical Research: Atmospheres (1984–2012) 103(D14): 17099–17110.

van Donkelaar A, Martin RV, Spurr RJ and Burnett RT (2015). “High-Resolution Satellite-Derived PM2.5 from Optimal Estimation and Geographically Weighted Regression over North America.” Environ Sci Technol 49(17): 10482–10491. [PubMed: 26261937]

Vermote E (2015). MOD09A1 MODIS/Terra Surface Reflectance 8-Day L3 Global 500m SIN Grid V006. N. E. L. P. DAAC

Wang J and Christopher SA (2003). “Intercomparison between satellite‐derived aerosol optical thickness and PM2. 5 mass: implications for air quality studies.” Geophysical research letters 30(21).

Wang P, Zhang H, Qin Z and Zhang G (2017). “A novel hybrid-Garch model based on ARIMA and SVM for PM2. 5 concentrations forecasting.” Atmospheric Pollution Research 8(5): 850–860.

Weizhen H, Zhengqiang L, Yuhuan Z, Hua X, Ying Z, Kaitao L, Donghui L, Peng W and Yan M (2014). Using support vector regression to predict PM10 and PM2. 5 IOP Conference Series: Earth and Environmental Science, IOP Publishing.

Wu Y, Guo J, Zhang X and Li X (2011). Correlation between PM concentrations and aerosol optical depth in eastern China based on BP neural networks Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International, Ieee.

Xie Y, Wang Y, Zhang K, Dong W, Lv B and Bai Y (2015). “Daily estimation of ground-level PM2. 5 concentrations over Beijing using 3 km resolution MODIS AOD.” Environmental science & technology 49(20): 12280–12288. [PubMed: 26310776]

Yang Y, Guo Y, Qian ZM, Ruan Z, Zheng Y, Woodward A, Ai S, Howard SW, Vaughn MG and Ma W (2018). “Ambient fine particulate pollution associated with diabetes mellitus among the elderly aged 50 years and older in China.” Environmental Pollution.

You W, Zang Z, Zhang L, Li Y, Pan X and Wang W (2016). “National-scale estimates of ground-level PM2. 5 concentration in China using geographically weighted regression based on 3 km resolution MODIS AOD.” Remote Sensing 8(3): 184.

Zhan Y, Luo Y, Deng X, Chen H, Grieneisen ML, Shen X, Zhu L and Zhang M (2017). “Spatiotemporal prediction of continuous daily PM2. 5 concentrations across China using a spatially explicit machine learning algorithm.” Atmospheric environment 155: 129–139.

Zhang H, Yee LD, Lee BH, Curtis MP, Worton DR, Isaacman-VanWertz G, Offenberg JH, Lewandowski M, Kleindienst TE and Beaver MR (2018). “Monoterpenes are the largest source of summertime organic aerosol in the southeastern United States.” Proceedings of the National Academy of Sciences 115(9): 2038–2043.

Zheng Y, Zhang Q, Liu Y, Geng G and He K (2016). “Estimating ground-level PM2. 5 concentrations over three megalopolises in China using satellite-derived aerosol optical depth measurements.” Atmospheric Environment 124: 232–242.

Zou B, Pu Q, Bilal M, Weng Q, Zhai L and Nichol JE (2016). “High-resolution satellite mapping of fine particulates based on geographically weighted regression.” IEEE Geoscience and Remote Sensing Letters 13(4): 495–499.

Highlights

- An ensemble model integrates three machine learning algorithms and estimates $PM_{2.5}$;
- Satellite measurements, land-use terms, and many variables were predictors;
- Model predicts daily $PM_{2.5}$ at 1 km × 1 km grid cells in the entire United States;
- Model predictions were downscaled to 100 m × 100 m level;
- Monthly uncertainty level of prediction was also estimated.

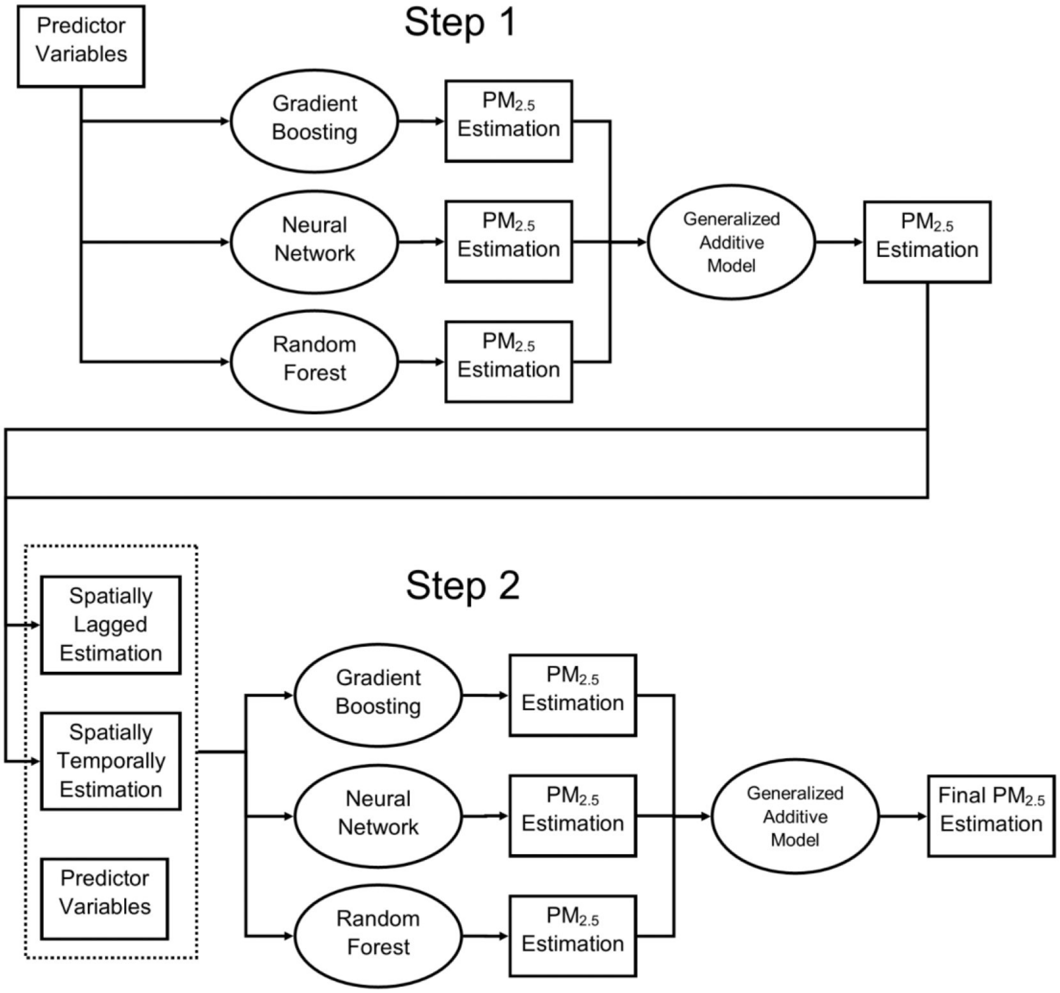


Figure 1.

Flowchart of Model Training Process

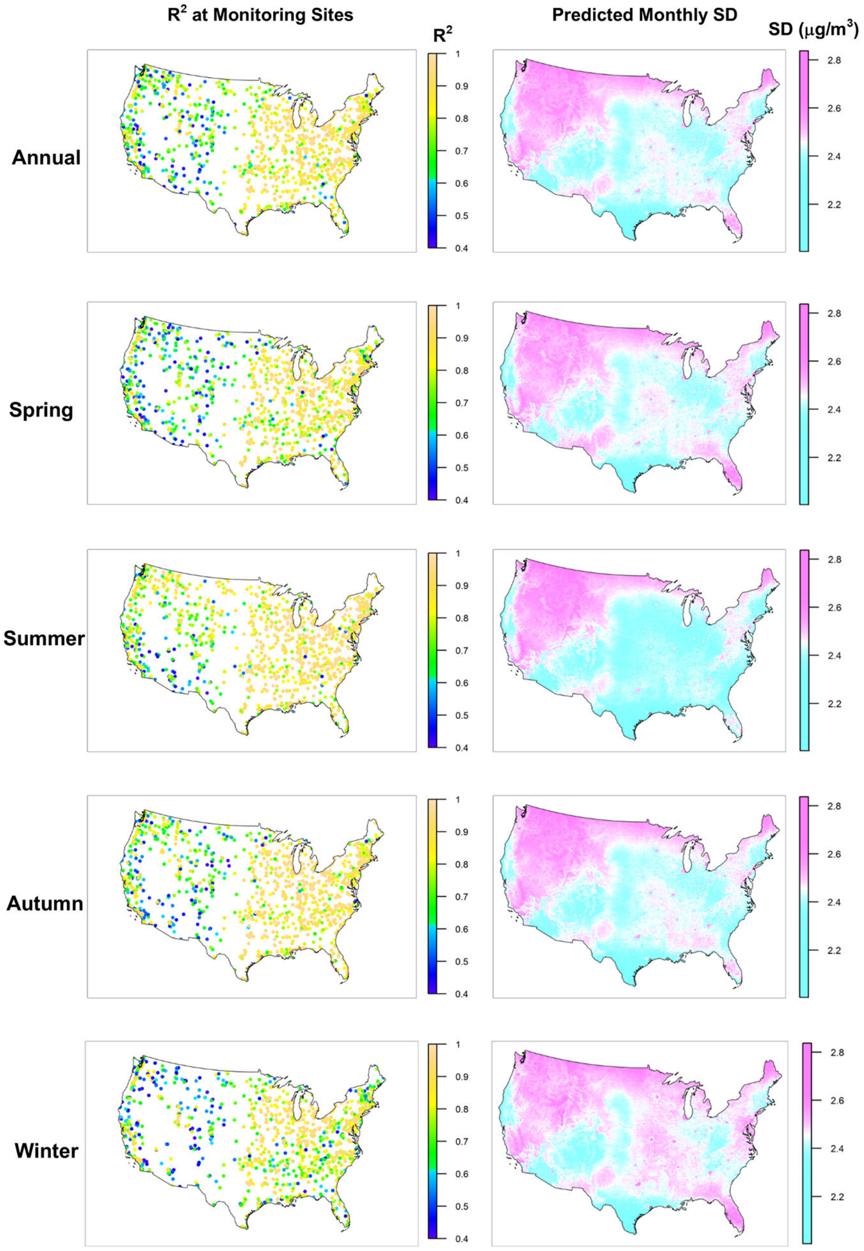


Figure 2. Cross-Validated R^2 at Monitoring Sites and Predicted Monthly Standard Deviation

The left figures present cross-validated R^2 at each monitoring site; the right figures present predicted monthly standard deviation at 1 km × 1 km grid cells, averaged over the entire years and four seasons. All maps were plotted at the same color scale.

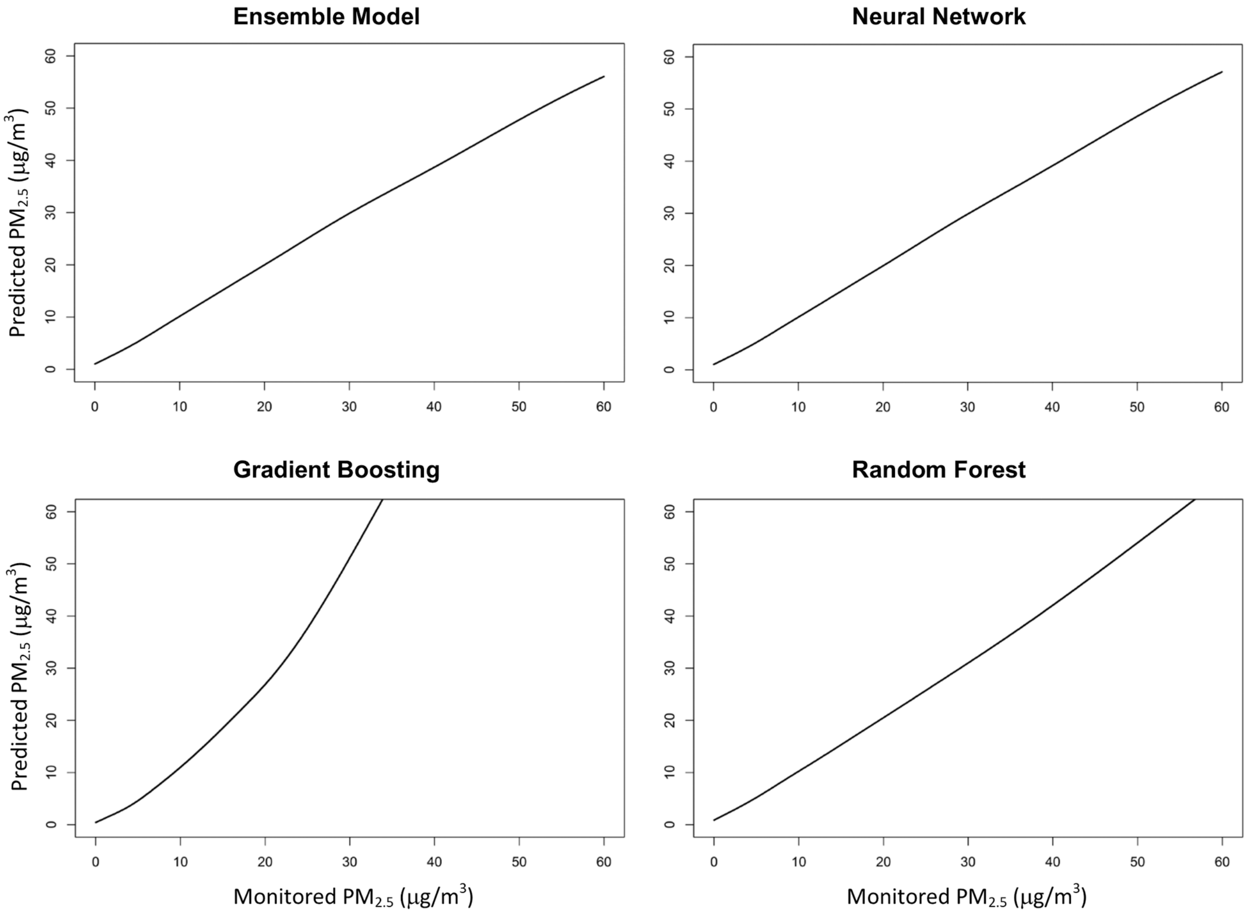


Figure 3. Relationship between Monitored and Predicted $PM_{2.5}$ from the Ensemble Model and Three Machine Learning Algorithms

We regressed daily predicted $PM_{2.5}$ from the ensemble model, neural network, gradient boosting, and random forest against monitored $PM_{2.5}$ in a generalized additive model, with spline on the monitored $PM_{2.5}$. Dashed lines represent 95% confidence interval. All plots were truncated to 60 μg/m^3, since 99.99% of daily $PM_{2.5}$ monitoring values from 2000 to 2015 were below 60 μg/m^3.

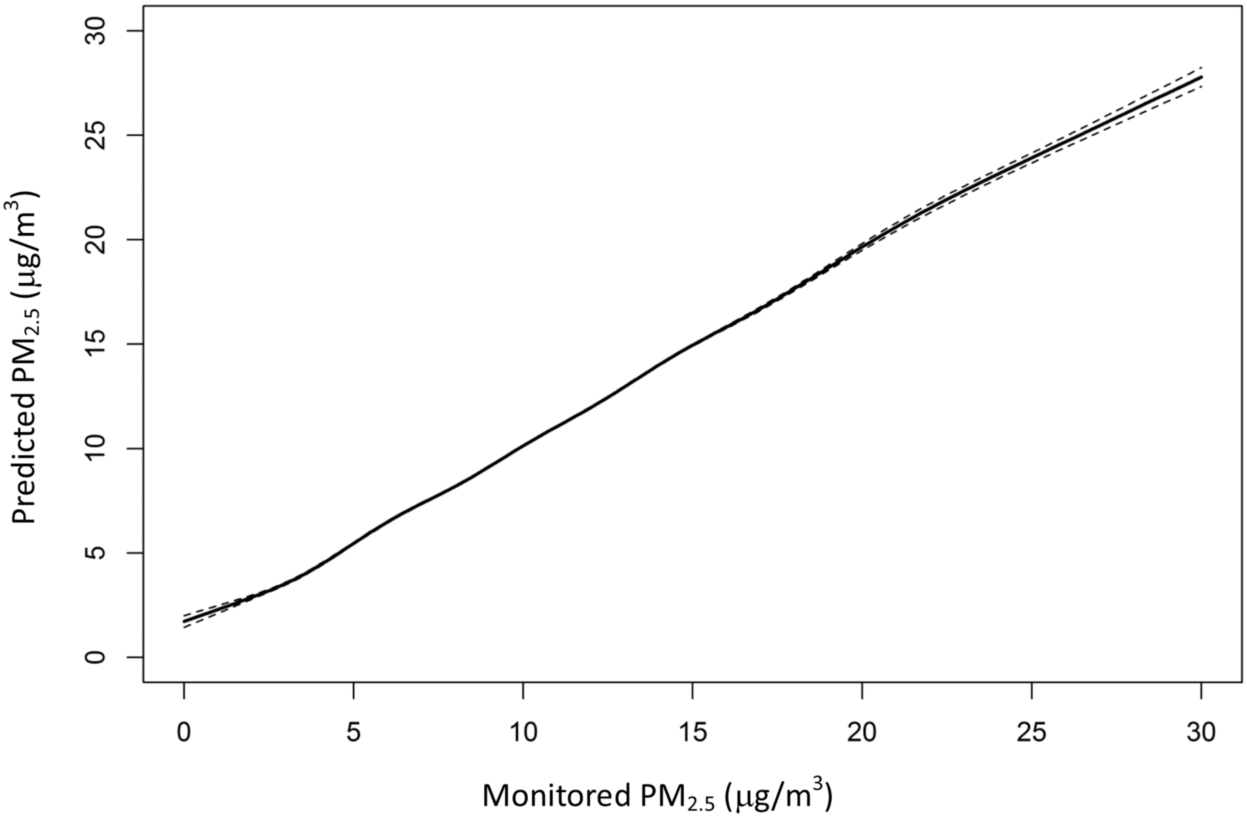


Figure 4. Relationship between Monitored and Predicted $PM_{2.5}$ at Annual Level

We regressed annual averaged $PM_{2.5}$ predictions from the ensemble model against annual averaged monitored $PM_{2.5}$ in a generalized additive model, with spline on the monitored $PM_{2.5.}$ Dashed lines represent 95% confidence interval.

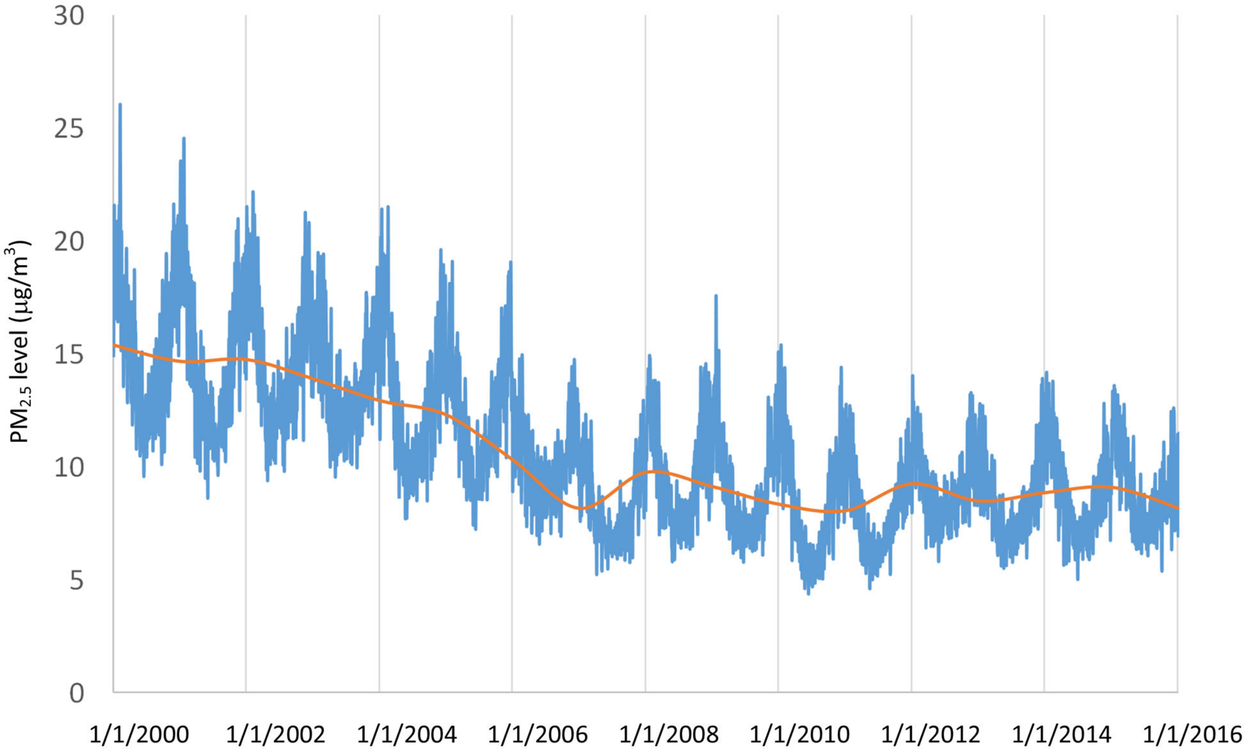


Figure 5. Temporal Trend of $PM_{2.5}$

We calculated the daily nationwide averages (blue line), by averaging daily predictions at all 1 km × 1 km grid cells; then we calculated nationwide annual averages (orange line).

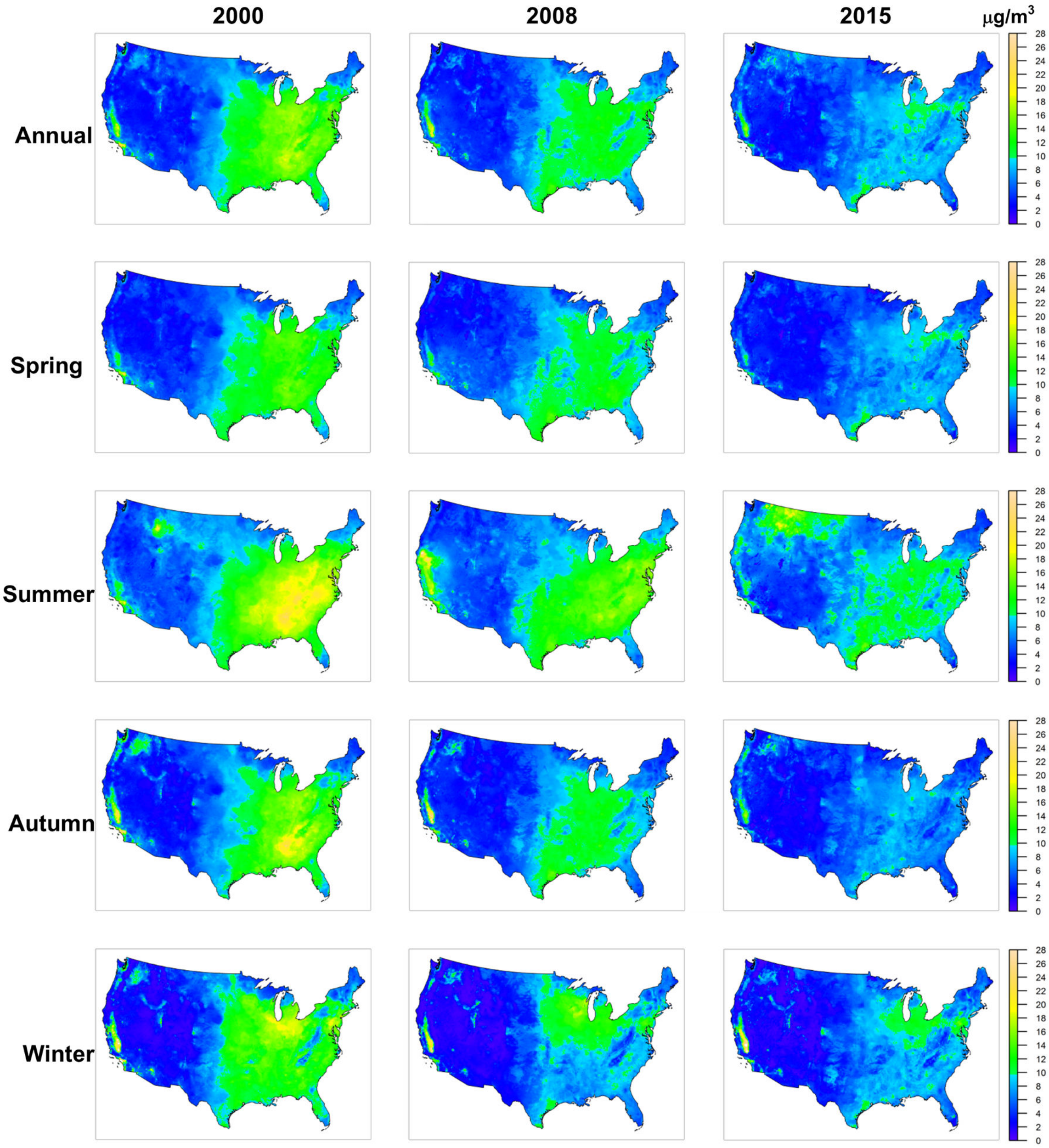


Figure 6. Spatial Distribution of Predicted $PM_{2.5}$

We predicted daily $PM_{2.5}$ at each 1 km × 1 km grid cell in the contiguous United States and calculated annual and seasonal averages for each grid cell. All maps were plotted at the same color scale.

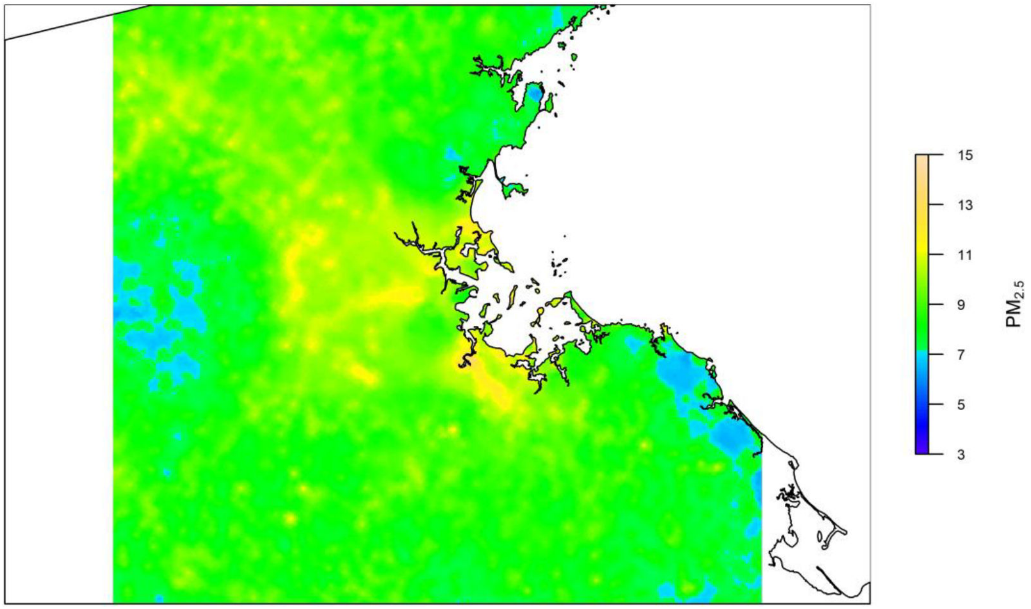


Figure 7. Downscaled $PM_{2.5}$ levels in the Great Boston Area

First, we made daily predictions of $PM_{2.5}$ at 1 km × 1 km grid cells in the Great Boston Area; and then we predicted residuals at 100 m × 100 m grid cells using our localized downscaling model; finally, we added residual with 1-km-level prediction to obtain 100-m-level predictions for the year of 2010. We visualized the annual averages at 100-m level for the year 2010.

Table 1

Cross-Validated R^2 for the Entire Study Area

	Ensemble Model						Neural Network	Random Forest	Gradient Boosting
Year	**R^2**	**RMSE**	**Bias**	**Slope**	**Spatial R^2**	**Temporal R^2**	**R^2**	**R^2**	**R^2**
2000	0.868	3.189	0.805	0.953	0.904	0.855	0.865[a]	0.863	0.836
2001	0.854	3.385	0.626	0.964	0.897	0.835	0.849[a]	0.849	0.822
2002	0.892	2.808	0.590	0.960	0.894	0.888	0.884	0.891[a]	0.860
2003	0.885	2.706	0.547	0.965	0.883	0.877	0.877	0.881[a]	0.853
2004	0.883	2.660	0.629	0.955	0.885	0.873	0.879	0.882[a]	0.854
2005	0.902	2.670	0.494	0.971	0.905	0.894	0.901	0.901[a]	0.880
2006	0.884	2.496	0.506	0.969	0.890	0.877	0.881[a]	0.876	0.855
2007	0.884	2.696	0.483	0.976	0.905	0.877	0.879	0.880[a]	0.859
2008	0.876	2.417	0.440	0.972	0.890	0.867	0.872[a]	0.865	0.834
2009	0.861	2.404	0.341	0.981	0.883	0.851	0.855[a]	0.847	0.817
2010	0.849	2.538	0.538	0.965	0.872	0.844	0.842[a]	0.835	0.809
2011	0.832	2.670	0.742	0.940	0.871	0.832	0.829[a]	0.822	0.792
2012	0.818	2.656	0.921	0.914	0.884	0.809	0.814[a]	0.805	0.744
2013	0.781	3.020	0.929	0.908	0.734	0.789	0.777[a]	0.775	0.718
2014	0.751	2.940	0.792	0.936	0.772	0.752	0.746[a]	0.734	0.701
2015	0.783	2.851	0.804	0.922	0.824	0.786	0.767	0.774[a]	0.711
Overall	**0.860**	**2.786**	**0.625**	**0.956**	**0.894**	**0.847**	**0.855**[a]	**0.854**	**0.818**

[a] the learner outperformed other learners in that year.

Note: All presented R^2 values were based on 10-fold cross-validation from the ensemble model. We used the trained ensemble model to make prediction at 1-km level. The calculation of daily R^2, spatial R^2, and temporal R^2 have been described elsewhere (Kloog, Koutrakis et al. 2011). We regressed predicted $PM_{2.5}$ against monitored $PM_{2.5}$ in a linear regression model, and obtained the slope and intercept (bias in the Table).

Table 2

Cross-Validated R^2 for Different Regions

	Ensemble Model						**Neural Network**	**Random Forest**	**Gradient Boosting**
Region	**R^2**	**RMSE**	**Bias**	**Slope**	**Spatial R^2**	**Temporal R^2**	**R^2**	**R^2**	**R^2**
East North Central	0.924	2.106	0.344	0.981	0.915	0.925	0.917	0.921[a]	0.899
East South Central	0.894	2.214	0.533	0.964	0.897	0.891	0.888	0.891[a]	0.872
Middle Atlantic	0.893	2.512	0.420	0.978	0.870	0.900	0.887[a]	0.883	0.863
Mountain	0.769	3.337	0.891	0.911	0.769	0.764	0.765	0.769[a]	0.694
New England	0.889	2.122	0.356	0.979	0.878	0.894	0.883[a]	0.875	0.852
Pacific	0.802	4.045	0.929	0.936	0.850	0.776	0.797[a]	0.797	0.750
South Atlantic	0.895	2.117	0.441	0.974	0.907	0.890	0.888	0.893[a]	0.872
West North Central	0.862	2.326	0.408	0.966	0.875	0.854	0.856[a]	0.851	0.824
West South Central	0.850	2.246	0.717	0.942	0.788	0.857	0.841[a]	0.841	0.824

[a] the learner outperformed other learners in that location.

Note: Region division was based on U.S. Census Bureau. New England: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont; Middle Atlantic: New Jersey, New York, Pennsylvania; East North Central: Indiana, Illinois, Michigan, Ohio, Wisconsin; West North Central: Iowa, Nebraska, Kansas, North Dakota, Minnesota, South Dakota, Missouri; South Atlantic: Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia; East South Central: Alabama, Kentucky, Mississippi, Tennessee; West South Central: Arkansas, Louisiana, Oklahoma, Texas; Mountain: Arizona, Colorado, Idaho, New Mexico, Montana, Utah, Nevada, Wyoming; Pacific: Alaska, California, Hawaii, Oregon, Washington. Although the Pacific Region includes Alaska and Hawaii, both states were not included in our modeling.

Table 3

Cross-Validated R^2 for Four Seasons

	Ensemble Model						Neural Network	Random Forest	Gradient Boosting
Season	**R^2**	**RMSE**	**Bias**	**Slope**	**Spatial R^2**	**Temporal R^2**	**R^2**	**R^2**	**R^2**
Spring	0.853	2.801	0.604	0.955	0.890	0.837	0.848	0.848[a]	0.812
Summer	0.858	2.271	0.421	0.972	0.895	0.841	0.852[a]	0.847	0.826
Autumn	0.901	2.445	0.544	0.962	0.939	0.881	0.896	0.896[a]	0.859
Winter	0.825	3.484	0.888	0.940	0.829	0.813	0.819	0.820[a]	0.777

[a] the learner outperformed other learners in that season.

Table 4

Relative Contribution of Predictor Variables for Three Machine Learning Algorithms

Gradient Boosting	%	Neural Network	%	Random Forest	%
Spatially Lagged Monitored $PM_{2.5}$	46.52%	AOD related variables[c]	9.25%	Spatially Lagged Monitored $PM_{2.5}$	28.96%
CMAQ $PM_{2.5}$	11.58%	Spatially Lagged Monitored $PM_{2.5}$	2.68%	CMAQ $PM_{2.5}$	16.51%
CMAQ $PM_{2.5}$ Sulfate	4.89%	Road Density (All Roads)	2.10%	CMAQ $PM_{2.5}$ Elemental Carbon	14.80%
Standard Deviation of Elevation	3.58%	Longitude	2.02%	CMAQ $PM_{2.5}$ Organic Carbon	6.28%
NLCD Developed Area[a]	2.79%	Latitude	1.99%	CMAQ $PM_{2.5}$ Sulfate	5.88%
CMAQ $PM_{2.5}$ Elemental Carbon	2.74%	Standard Deviation of Elevation	1.93%	Spatially Lagged Monitored $PM_{2.5}$[b]	3.29%
CMAQ $PM_{2.5}$ Organic Carbon	2.65%	NLCD Planted Land Coverage[a]	1.93%	AOD related variables[c]	2.65%
Spatially Lagged Monitored $PM_{2.5}$[b]	2.22%	Soil moisture	1.92%	CMAQ NO_2	1.86%
Longitude	1.93%	Road Density (Pri-Secondary Road)	1.80%	Latitude	1.69%
NLCD Impervious Land Coverage[a]	1.82%	NLCD Developed Area	1.72%	NLCD Impervious Land[a]	1.40%
Standard Deviation of Elevation[a]	1.51%	NLCD Waterbody Coverage[a]	1.63%	NLCD Developed Area[a]	1.37%
Road Density (All Roads)	1.51%	NLCD Tree Canopy[a]	1.57%	Road Density (All Roads)	1.32%
AOD related variables[c]	1.50%	Standard Deviation of Elevation[a]	1.55%	Longitude	0.99%
Latitude	1.36%	Road Density (Pri-Secondary Road)[a]	1.55%	Standard Deviation of Elevation	0.91%
NLCD Planted Land Coverage[a]	1.03%	NLCD Herbaceous Land[a]	1.53%	Upward Longwave Radiation	0.59%
NLCD Tree Canopy Coverage	0.81%	NLCD Wetland Coverage[a]	1.51%	CMAQ $PM_{2.5}$ Nitrate	0.59%
Road Density (Pri-Secondary Road) [a]	0.72%	NLCD Tree Canopy Coverage	1.51%	Daily Maximal Air Temperature	0.54%
NLCD Barren Land Coverage[a]	0.72%	Road Density (Primary Road)[a]	1.46%	Road Density (Pri-Secondary Roads)[a]	0.54%
MERRA2 Sulfate Aerosol	0.59%	Road Density (Primary Road)	1.45%	MODIS Daytime Surface Temperature	0.54%
OMI NO_2 Column Concentration	0.57%	Spatially Lagged Monitored NO_2	1.41%	NLCD shrubland[a]	0.44%

[a] These land-use variables were averaged over 10000 m × 10000 m;

[b] These were 1-day lagged values;

[c] AOD related variables include: AOD at 470 nm from the Terra and Aqua satellites (both retrieved by the MAIAC algorithms), AOD at 550 nm from the Terra and Aqua satellites (both retrieved by the MAIAC algorithms), AOD at 550 nm (retrieved by the deep blue algorithm), OMI AAI (at the visible and UV spectrum), surface reflectance, and aerosol products from MERRA2 (sulfate aerosol, hydrophilic black carbon, hydrophobic black carbon, hydrophilic organic carbon, and hydrophobic organic carbon)

Note: For each learning algorithm, we ranked the importance of predictor variables and listed the first 20 of them. For random forest and gradient boosting, we calculated how much the squared error over all trees decreased after a variable was selected to split on in the tree building process. The decreased squared error was determined as the relative influence of that variable (Rifkin and Klautau 2004). For neural network, the sensitivity of output to input was used to assess variable importance (Gedeon 1997).