Environmental Science & Technology

Article

Cite This: Environ. Sci. Technol. 2018, 52, 13260−13269

pubs.acs.org/est

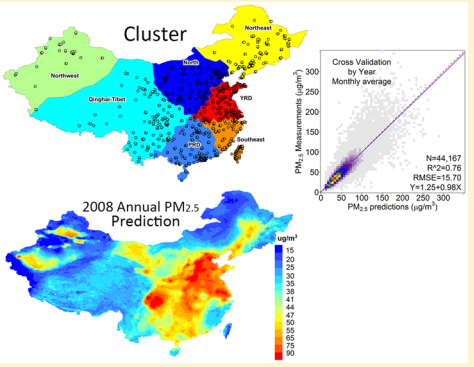
An Ensemble Machine-Learning Model To Predict Historical $PM_{2.5}$ Concentrations in China from Satellite Data

Qingyang Xiao,[†] Howard H. Chang,[‡] Guannan Geng,[†] and Yang Liu*[,†]

[†]Department of Environmental Health and [‡]Department of Biostatistics and Bioinformatics, Emory University, Rollins School of Public Health, Atlanta, Georgia 30322, United States

Supporting Information

ABSTRACT: The long satellite aerosol data record enables assessments of historical $PM_{2.5}$ level in regions where routine $PM_{2.5}$ monitoring began only recently. However, most previous models reported decreased prediction accuracy when predicting $PM_{2.5}$ levels outside the model-training period. In this study, we proposed an ensemble machine learning approach that provided reliable $PM_{2.5}$ hindcast capabilities. The missing satellite data were first filled by multiple imputation. Then the modeling domain, China, was divided into seven regions using a spatial clustering method to control for unobserved spatial heterogeneity. A set of machine learning models including random forest, generalized additive model, and extreme gradient boosting were trained in each region separately. Finally, a generalized additive ensemble model was developed to combine predictions from different algorithms. The ensemble prediction characterized the spatiotemporal distribution of daily $PM_{2.5}$ well with the cross-validation (CV) R^2 (RMSE) of 0.79 (21 $\mu g/m^3$). The cluster-based subregion models outperformed national models and improved the CV R^2 by ∼0.05. Compared with previous studies, our model provided more accurate out-of-range predictions at the daily level (R^2 = 0.58, RMSE = 29 $\mu g/m^3$) and monthly level (R^2 = 0.76, RMSE = 16 $\mu g/m^3$). Our hindcast modeling system allows for the construction of unbiased historical $PM_{2.5}$ levels.



INTRODUCTION

$PM_{2.5}$ (fine particulate matter with an aerodynamic diameter of 2.5 μm or less) is a critical air pollutant that is associated with adverse health outcomes.[1,2] However, the ground monitoring of $PM_{2.5}$ is limited in most developing regions. For instance, in China, the national air quality monitoring network was established in 2013 such that $PM_{2.5}$ measurements before 2013 were unavailable, making it difficult to assess long-term $PM_{2.5}$ trends. To extend ground air quality monitoring networks, satellite-retrieved aerosol optical depth (AOD) has been increasingly used for air pollution monitoring in the past decade. Satellite data with broad spatial coverage, a long data record, and high spatial resolutions could support the assessment of historical air pollution levels in developing regions.

Previous studies have presented various statistical models to describe the nonlinear relationship between satellite AOD and ground $PM_{2.5}$ concentration, addressing the effects of meteorological parameters, emission sources, and land use information.[3−7] Benefitting from a long $PM_{2.5}$ ground monitoring record, most US-based models have aimed to extend the spatial coverage of $PM_{2.5}$ monitoring networks rather than to generate historical $PM_{2.5}$ levels. These models often included daily random effects or day-stratification to improve performance. Although day-specific intercepts and slopes can capture the unobserved fine-scale temporal trends in the associations between $PM_{2.5}$ concentration and explanatory variables, applying the daily effects outside the model fitting period imposes a strong and often unrealistic assumption that the estimated daily effects during the model fitting period will remain constant during the hindcast period. Thus, the model performance degraded significantly outside the model fitting period. For example, Ma et al.[19] reported that when using a model fitted with 2013 data to predict daily $PM_{2.5}$ concentrations in 2014, the R^2 was 0.41 compared to the 10-fold cross-validation (CV) R^2 of 0.79. Weng et al.[8] also reported that using a model fitted with 2015 data to predict daily $PM_{2.5}$ concentrations in 2014 had R^2 of 0.47, wheare the model CV R^2 was 0.80.

Another $PM_{2.5}$ modeling approach, driven by atmospheric chemical transport model (CTM) simulations, has also been reported.[9,10] This approach estimated the scaling factor between AOD and $PM_{2.5}$ from model simulations and applied the scaling factor to satellite retrieved AOD to get $PM_{2.5}$ estimations. This approach can estimate historical $PM_{2.5}$ levels at the global scale; however, the relatively low accuracy of CTM simulations limited the performance of this approach and the prediction accuracy was not comparable to statistical models.[11,12]

Received: June 4, 2018
Revised: September 3, 2018
Accepted: October 24, 2018
Published: October 24, 2018

DOI: 10.1021/acs.est.8b02917
Environ. Sci. Technol. 2018, 52, 13260−13269

Most recently, machine learning algorithms have been applied to $PM_{2.5}$ prediction. Machine learning algorithms can deal with complex nonlinear relationships with interactions, making them promising in air pollution prediction. Neural network,[7] random forest,[13] and deep belief network[14] have been presented in the U.S. and in China, but these models relied on spatial and/or temporal correlations of $PM_{2.5}$ estimated from ground observations to improve model performance. As a result, these models cannot estimate historical $PM_{2.5}$ levels when ground measurements were unavailable. Gradient boosting[15] and generalized regression neural network[16] have also been employed to predict daily $PM_{2.5}$ concentrations in China, with the 10-fold CV R^2 of 0.76 and 0.67, respectively. Although these models did not rely on $PM_{2.5}$ measurements to construct input variables, neither of these two studies reported their models' hindcast ability.

Previous studies revealed significant spatial heterogeneity in relationships between $PM_{2.5}$, satellite AOD, and meteorological parameters.[7,17] Thus, dividing a large modeling domain and training regional models could help control for unobserved spatial features and improve model performance.[13] Previous studies in the U.S. divided study domains according to climate regions defined by the National Oceanic and Atmospheric Administration (NOAA),[13,18] but it is not clear how to divide China into reasonable subregions. Ma et al.[19] fitted their multistage model for each province in China, but provincial areas vary dramatically, ranging from 0.07 million km^2 (Ningxia) to 1.7 million km^2 (Xinjiang). In addition, provincial boundaries do not necessarily reflect any geographic or emission patterns, and observations from one province are generally insufficient to support a complex model. Thus, researchers had to select different buffer radii manually to ensure sufficient model fitting data in each province-based region.

In this study, we proposed a machine-learning approach that provided reliable historical $PM_{2.5}$ concentration estimates. We developed a clustering method to divide China into seven temporally stable regions. Then we trained a set of machine learning models in each region that did not rely on daily effects during 2013−2016. We finally combined predictions from various models by an additive ensemble model.

METHODS

Data. The study domain covers mainland China, Hong Kong special administrative region, and Taiwan (Figure 1). The China map with province outlines is downloaded from http://www.resdc.cn/. We constructed a 0.1° modeling grid covering this study domain for data integration. We used data during 2013−2016 for model training and data during the first seven months of 2017 for hindcast evaluation. We also obtained $PM_{2.5}$ concentrations in 2008 in Beijing as a case to evaluate the model hindcast performance.

$PM_{2.5}$ Measurements. Hourly $PM_{2.5}$ concentrations in 2013−2017 were measured at ~1,593 air quality monitoring stations across mainland China (Figure 1). Since the air quality monitoring network was under development during the study period, the number of monitoring stations increased over the years. Measurements are published by the China National Environmental Monitoring Center (CNEMC, http://www.cnemc.cn/) and were downloaded from PM25.in (http://pm25.in/). Additionally, we collected $PM_{2.5}$ measurements in Hong Kong and Taiwan from the Hong Kong environmental protection department (http://epic.epd.gov.hk/) and the Taiwan environmental protection agency (http://taqm.epa.gov.tw/), respectively. We removed repeated identical measurements for at least three continuous hours, assuming that such repetition was due to instrument malfunction. Daily average concentrations with less than 18 hourly measurements were excluded. Thus, during the study period, 2.3% of daily measurements were removed due to the exclusion of consecutive identical hourly measurements. Daily average $PM_{2.5}$ measurements from stations located within the same grid cell were averaged, and we got as many as 1214 grid cells with $PM_{2.5}$ measurements. Finally, we analyzed $PM_{2.5}$ concentrations during 2008 in Beijing measured at Tsinghua University, Daxing District, and Miyun District during June to October, 2008,[20] as well as measured at Peking University and at the U.S. embassy (https://china.usembassy-china.org.cn/embassy-consulates/beijing/air-quality-monitor/) as a test of model hindcast capabilities. The location of these sampling sites do not coincide with any later established regulatory monitors (Figure 1).

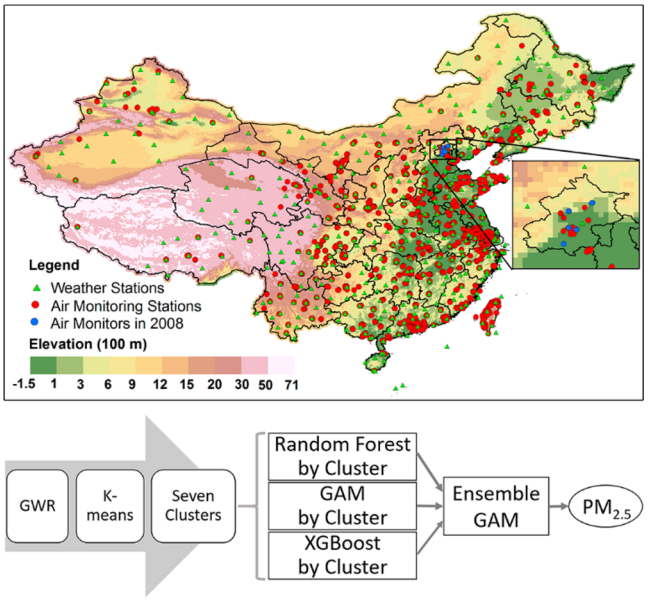


Figure 1. Map of the study domain (above) and model structure (below). Air quality monitors are shown as red dots, and the weather stations included in the National Centers for Environmental Information (NCEI) data set are shown as green triangles. The China map with province outlines was downloaded from http://www.resdc.cn/, and the elevation data were obtained from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM) version 2.

Satellite Data. The Moderate Resolution Imaging Spectroradiometer (MODIS) Collection 6 level 2 aerosol products at 10 km resolution from Aqua and Terra satellites were downloaded from the Atmospheric Archive and Distribution System (http://ladsweb.nascom.nasa.gov/). Since MODIS retrievals were affected by the bow-tie effect (pixels were stretched at the border of each scan), to correctly assign AOD retrievals to the 0.1° grid cell, we created Thiessen polygons from centroid of AOD pixels. Two retrieval algorithms, Deep Blue (DB) and Dart Target (DT), have been developed to retrieve MODIS AOD at 10 km resolution.[21,22] These two algorithms use different methods to characterize surface reflectance and are suitable for different land surfaces. DT provides high quality retrievals over

DOI: 10.1021/acs.est.8b02917
Environ. Sci. Technol. 2018, 52, 13260−13269

vegetation covered land, while DB is able to retrieve AOD over bright land, e.g. urban regions. Additionally, the parameter "combined AOD" combines high quality retrievals from DB and DT, accounting for surface situations. Since this combination only includes high quality retrievals, its coverage is very limited. In this study, we included all three AOD parameters as separate inputs in our machine-learning models.

Due to cloud cover or high surface reflectance, about 40−70% of satellite retrievals are missing on average in East Asia.[23] To improve the coverage of satellite retrievals without decreasing retrieval quality, we filled data gaps in DB AOD, DT AOD, and combined AOD separately using multiple imputation. The details of this method are provided elsewhere, and here is a brief summary.[5] We first fitted daily linear regressions between AOD retrievals from the Aqua satellite (overpass time at 1:30 pm local time) and Terra satellite (overpass time at 10:30 am local time) and used the regression coefficients to estimate the missing Aqua/Terra AOD when only one of them is present. Then the observed and predicted AOD values were averaged to reflect daily aerosol loadings.[24] We then filled the missing daily average AOD by multiple imputation with an additive model driven by chemical transport model AOD simulations, temperature, and humidity in the boundary layer, elevation, and MODIS cloud fraction.[25] Each missing daily AOD was imputed five times to account for the additional uncertainty due to imputation, and the average of the five imputed AODs served as a predictor in machine learning models.

The MODIS active fire data were obtained from the Fire Information for Resource Management System (FIRMS, https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms). We developed buffers with various radii, including 20, 30, 50, and 75 km, to assign fire information to the corresponding 0.1° modeling grid cell. Specifically, we searched and summed the number of fire spots within each buffer centered on the centroid of each grid cell. We extracted *cloud_fraction_day* from Aqua and Terra Collection 6 level 2 cloud products (MYD06_L2 and MOD06_L2), at 5 km spatial resolution. The daily cloud fraction was calculated as the average of Aqua and Terra cloud fraction that were interpolated to 0.1° grid cell by the nearest neighbor approach. Normalized Difference Vegetation Index (NDVI) data were obtained from Terra MODIS monthly global NDVI data set at 1 km resolution (MOD13A3). The NDVI value of each grid cell was assigned as the average of NDVI pixels falling within the corresponding grid cell. Missing data in NDVI were interpolated by inverse distance weighting.

The tropospheric vertical column NO_2 density and absorbing aerosol index (AAI) data in visible light and UV light from Ozone Monitoring Instrument (OMI) was downloaded from the Goddard Earth Sciences Data and Information Services Center (https://mirador.gsfc.nasa.gov/). We extracted and processed the parameters *ColumnAmountNO2Trop* from the OMI NO2 level 2 data (OMNO2), *AerosolIndexUV* and *AerosolIndexVIS* from the OMI Aerosol Extinction Optical Depth and Aerosol Types level 2 data (OMAERO), and *UVAerosolIndex* from the OMI Near-UV Aerosol Absorption and Extinction Optical Depth and Single Scattering Albedo level 2 data (OMAERUV). Due to a row anomaly starting from 2007, retrievals with the cross track anomaly flag as nonzero were removed and oversampling was conducted to smooth the systematic noise. Regarding the NO_2 column density, the value of each 0.1° grid cell was assigned as the average of samples from a 20 km-radius buffer centered on this grid cell during each season. Regarding the AAI parameters, retrievals with lower than 0.5% percentile were removed and the values of each 0.1° grid cell were assigned as the average of samples from a 30 km-radius buffer centered on this grid cell during each season.

Meteorological and Land Use Data. Meteorological parameters in 2013−2017 were extracted from the Goddard Earth Observing System Data Assimilation System GEOS-5 Forward Processing (GEOS 5-FP) at 0.25° latitude × 0.3125° longitude resolution. Meteorological parameters in 2008 were extracted from the Goddard Earth Observing System Model, Version 5 (GEOS 5) at 0.5° × 0.5° resolution. The meteorological data were downscaled to a 0.1° grid cell by inverse distance weighting. The elevation data were obtained from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM) version 2 at 30 m resolution. Population density data were obtained from the LandScan Global Population Database at 1 km resolution.[26]

Since we extracted various wind parameters at different heights of the atmosphere (wind direction, u and v component of wind speed at 10 m, averaged in the boundary layer, and at 500 mb), to reduce feature space and avoid the curse of dimensionality,[27] we applied dimension reduction by linear discriminant analysis (LDA) on these wind parameters.[28] We extracted the first and second components from LDA that cumulatively explained over 95% of variabilities in all wind parameters.

MERRA-2 $PM_{2.5}$ Reanalysis Data. We obtained daily $PM_{2.5}$ simulations from the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2).[29] The MERRA-2 $PM_{2.5}$ simulations have relatively high accuracy[30] at 0.5° latitude × 0.625° longitude resolution. MERRA-2 data provided additional information on $PM_{2.5}$ distribution at a broad scale. The total concentration of $PM_{2.5}$ was calculated using the following equation:[31,32]

$$PM_{2.5} = 1.375 \times SO_4 + 2.1 \times OC + BC + Dust_{2.5} + Seasalt_{2.5}$$

where SO_4, OC, BC represent the MERRA-2 concentration of sulfate ion, organic carbon, and black carbon, respectively. $Dust_{2.5}$ and $Seasalt_{2.5}$ are the concentration of dust and sea salt with a radius less than 2.5 μm. Specifically, we summed dust concentrations of bin 1 (radius 0.1−1.0 μm), 2 (radius 1−1.5 μm), and 3 (radius 1.5−3.0 μm) and sea salt concentrations of bin 1 (radius 0.03−0.1 μm), 2 (radius 0.1−0.5 μm), and 3 (radius 0.5−1.5 μm). We multiplied SO_4 by 1.375 to get the concentration of sulfate aerosol, assuming that sulfate is primarily presented as ammonium sulfate. The ratio between organic carbon and organic matter, 2.1, was estimated from $PM_{2.5}$ observations and MERRA-2 organic carbon simulations in China during 2013−2016. The MERRA-2 aerosol simulation product does not provide particle nitrate concentration. The MERRA-2 $PM_{2.5}$ simulations at 50 km resolution was interpolated by inverse distance weighting to the 0.1° modeling grid.

Visibility Data. The visibility data were extracted from the Integrated Surface Data set (IDS) from the U.S. National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information (NCEI). Visibility was measured at 407 stations in China (Figure 1). The daily

DOI: 10.1021/acs.est.8b02917
Environ. Sci. Technol. 2018, 52, 13260−13269

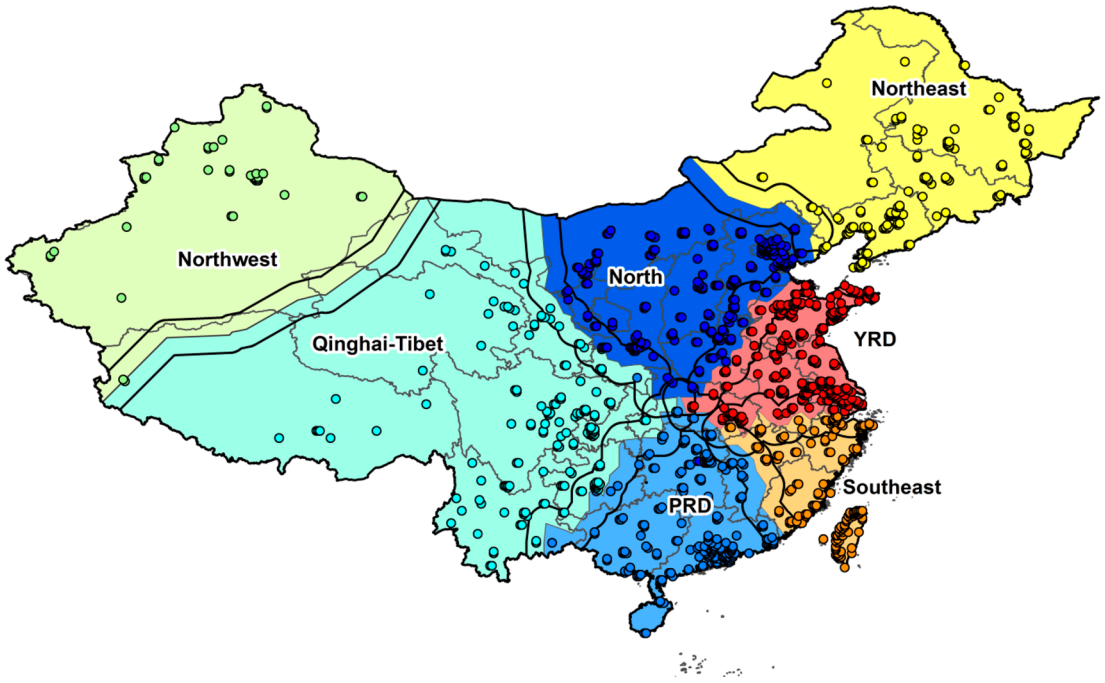


Figure 2. Seven clusters covering the study domain. The China map with province outlines was downloaded from http://www.resdc.cn/.

average visibility was interpolated by inverse distance weighting and assigned to the 0.1° modeling grid.

MODELS

A diagram of our modeling method is shown in Figure 1. First, we divided our study domain according to the coefficient surface estimated from geographically weighted regression (GWR) by the K-Means algorithm. Then we trained three machine learning models, including random forest, extreme gradient boosting (XGBoost), and generalized additive model (GAM) in each region, separately. The decision tree based algorithms, random forest and XGboost, provided the estimated importance of predictors that guided parameter selection.[13,33] The GAM model has been widely used to characterize the spatial distribution of $PM_{2.5}$.[6,34] Finally, to improve the hindcast accuracy and robustness, we combined predictions from the three individual machine learning models by a GAM ensemble model. We trained and evaluated prediction models at the daily level.

The R package "mlr" was used to optimize hyperparameters of each algorithm through 5-fold cross validation (CV) and fixed holdout (Text S1). We evaluated the model performance by 10-fold CV at daily level that we randomly selected 90% of data to train individual models and the ensemble model, and then, we used the remaining 10% of data to examine the model performance. This process was repeated 10 times so that each data record was left for testing once. Because, in such a standard CV, the randomly selected training data set usually contains enough observations to estimate local spatial and temporal trends that may not hold constant outside the model fitting domain and period, we also conducted 10-fold CV spatially and temporally to detect potential spatial and temporal overfitting. For the spatial CV, we randomly selected 10% of monitors to test the model. Similarly, for the temporal CV, we randomly selected 10% of days to test the model. Additionally, we conducted a by-year CV using data outside the training period (i.e., 2017) and data during 2013−2016 to evaluate our model's hindcast performance. We held one year's worth of data at a time for testing and used remaining years for model training. Additionally, we used measurements outside the existing monitoring network (temporary research stations in 2008 in Beijing) to further characterize the hindcast prediction error.

Cluster Analysis. First, we fitted a GWR model with the annual average $PM_{2.5}$ concentrations together with annual mean DB AOD, meteorological variables, population density, and elevation (eq 1). DB AOD was included because it had the highest coverage before gap-filling. GWR has been widely used to analyze spatially varying relationships.[35] It generates a continuous surface of regression coefficients through a spatial weighting mechanism. Since we aimed to control the spatial trend by clustering, we used annual average values for GWR fitting and ignored the temporal variations to avoid short-term fluctuations in cluster patterns.

$$\begin{aligned} PM2.5_{t,i} \sim {} & Elev_i + DB\_AOD_{t,i} + Pop_{t,i} + Tem_{t,i} \\ & + Humidity_{t,i} + Prec_{t,i} + PBLH_{t,i} + AAI\_UV_{t,i} \\ & + Column\_NO2_{t,i} + e_{t,i} \end{aligned} \quad (1)$$

where $PM2.5_{t,i}$ represents the annual average $PM_{2.5}$ concentrations of year t at grid cell i; $Elev_i$ represents the elevation of grid cell i; $DB\_AOD_{t,i}$, $Pop_{t,i}$, $Tem_{t,i}$, $Humidity_{t,i}$, $Prec_{t,i}$, $PBLH_{t,i}$, $AAI\_UV_{t,i}$, and $Column\_NO2_{t,i}$ represent the annual average DB AOD, population, temperature, humidity, precipitation, planetary boundary layer height, AAI in UV light, and tropospheric vertical column NO_2 density of year t at grid cell i, respectively.

After fitting the GWR, we clustered $PM_{2.5}$ monitors according to the vector of estimated coefficients by the K-Means algorithm and assigned $PM_{2.5}$ monitoring stations to different clusters. The number of clusters (k) was decided after comparing the by-year CV results using various values of k, ranging between one (national model) and 15 (Figure S1).

DOI: 10.1021/acs.est.8b02917
Environ. Sci. Technol. 2018, 52, 13260−13269

Table 1. Model Fitting and 10-Fold CV Results at the Daily Level for Individual Cluster-Based Models, Individual National Models, and the Ensemble Model

	individual model				
R^2 (RMSE ($\mu g/m^3$))	XGBoost	random forest	GAM	LEM+GAM	ensemble model
model fitting (cluster)	0.84 (18)	0.77 (22)	0.65 (28)	0.64 (27)	0.85 (18)
standard CV (cluster)	0.78 (21)	0.77 (22)	0.65 (28)	0.63 (27)	0.79 (21)
standard CV (national)	0.72 (24)	0.71 (25)	0.60 (29)	0.57 (30)	
temporal CV (cluster)	0.71 (25)	0.72 (25)	0.65 (28)	0.48 (33)	0.73 (24)
spatial CV (cluster)	0.74 (22)	0.75 (23)	0.58 (30)	0.63 (28)	0.76 (22)

The estimated coefficients from GWR were normalized before clustering, and we gave longitude and latitude twice the weight to favor spatially continuous clusters. To examine the effects of randomization on the clustering results, we examined 20 different random seeds when selecting initial centroids and used the most common clustering pattern for the following analysis. Thiessen polygons were generated from monitors, and we assigned grid cells within each Thiessen polygon to the same cluster of the corresponding monitor in the center (Figure 2). We added a 1° buffer to each region and averaged $PM_{2.5}$ predictions in the buffer to ensure that the daily $PM_{2.5}$ predictions are spatially continuous. To examine the long-term stability of the clusters, we also estimated the clustering pattern by year as a sensitivity analysis.

Generalized Additive Model. GAM is a nonparametric model where the dependent variable depends linearly on smooth functions of predictors. We log transformed $PM_{2.5}$ concentrations to improve the prediction accuracy of high $PM_{2.5}$ values. The GAM model is shown as

$$\begin{aligned} \mathrm{lg\_PM2.5}_{i,j} &= s((\mathrm{Lon}, \mathrm{Lat})_i) + s(\mathrm{DB\_AOD}_{i,j}) \\ &+ s(\mathrm{DT\_AOD}_{i,j}) + s(\mathrm{AAI\_UV}_{i,j}) + s(\mathrm{Prec}_{i,j}) \\ &+ s(\mathrm{Prec\_lag1}_{i,j}) + s(\mathrm{Column\_NO2}_{i,j}) + s(\mathrm{Humidity}_{i,j}) \\ &+ s(\mathrm{Tem}_{i,j}) + s(\mathrm{Visibility}_{i,j}) + s(\mathrm{MERRA2\_PM2.5}_{i,j}) \\ &+ s(\mathrm{Pop}_{t,i}) + \mathrm{PBLH}_{i,j} + e_{i,j} \end{aligned} \quad (2)$$

where $\mathrm{lg\_PM2.5}_{i,j}$ represents the log of $PM_{2.5}$ concentrations on day j at grid cell i; $s((\mathrm{Lon}, \mathrm{Lat})_i)$ represents a thin plate surface of longitude and latitude of grid cell i; $s()$ represents a smooth function of the corresponding parameter.

Random Forest Model. Initially proposed by Breiman,[36] the random forest algorithm is a bagged classifier based on decision tree. The random forest algorithm offers several advantages over other machine learning algorithms: it allows both continuous and categorical input variables; it is robust to outliers; and it provides variable importance as well as out of bag error for model evaluation. One limitation of the random forest algorithm is that with the increase of number of trees and complexity of each tree, the model training and prediction time can increase significantly. Since the contribution of each predictor varied across regions, we selected predictors separately in each region.

Extreme Gradient Boosting Model. The XGBoost algorithm is developed from gradient boosting.[37] Gradient boosting model has been shown to outperform various statistical and machine learning models in predicting $PM_{2.5}$ levels during a wildfire event.[33] XGBoost requires less training and predicting time than random forest and has been widely used in data mining competitions.[38,39] The R package, xgboost, was used to train the XGBoost model.[40] The hyperparameters of XGBoost model were selected by grid search (Text S1). To avoid overfitting, only parameters with the evaluation statistic Gain, which describes the improvement in accuracy after splitting on the corresponding feature, larger than 0.01 were included in the model.

Ensemble Model. To ensure a spatially continuous prediction surface, we fitted a national GAM model including predictions from the three individual models during the 4-year modeling period, 2013−2016. Predictions from the GAM model were transformed to normal scale before training the ensemble model. The ensemble model is shown as

$$\begin{aligned} \mathrm{PM2.5}_{i,j} &= s(\mathrm{Pred\_RandomForest}_{i,j}) + s(\mathrm{Pred\_XGBoost}_{i,j}) \\ &+ s(\mathrm{Pred\_GAM}_{i,j}) + e_{i,j} \end{aligned} \quad (3)$$

where $\mathrm{Pred\_RandomForest}_{i,j}$, $\mathrm{Pred\_XGBoost}_{i,j}$, and $\mathrm{Pred\_GAM}_{i,j}$ are the predictions of $PM_{2.5}$ concentrations on day j at grid cell i from random forest, XGBoost, and GAM, respectively.

Two-stage Statistical Model. One of the previously used $PM_{2.5}$ prediction methods is the two stage statistical framework with a first stage linear mixed effects model (LME) driven by AOD and meteorological variables, and a second stage generalized additive model (GAM) driven by land use information. To compare the performance of our machine learning algorithm with the statistical model, we fitted a LME +GAM model with the seven clusters constructed in this study and a similar set of predictors. The model structure is similar to that previously reported in the work of Ma et al.,[19] and details of the model are provided in Text S2.[5,19]

RESULTS AND DISCUSSION

Cluster Analysis. The estimated cluster map is shown in Figure 2. As expected, the separation of clusters did not follow provincial boundaries. Three northeastern provinces, i.e., Heilongjiang, Jilin, and Liaoning, as well as the northern Inner Mongolia constituted the Northeast cluster, characterized by its long winter/heating season and large presence of heavy industry. The North China Plain constituted the North cluster, characterized by its coal consumption[41,42] and stagnant atmospheric conditions in winter, contributing to frequent regional haze events.[43,44] The Yangtze River Delta was separated into two clusters: the relatively cold north (YRD) with central heating in winter and the relatively warm south without central heating (Southeast). The Pearl River Delta (PRD) was another cluster, located on the coast with warm weather. The PRD and Southeast clusters also produce more hydroelectricity than other regions.[45] The Qinghai-Tibetan Plateau, Sichuan, Yunnan, and Gansu province constituted the largest cluster (West) with a high altitude and low population density. Xinjiang Uygur Autonomous Region dominated the

DOI: 10.1021/acs.est.8b02917
Environ. Sci. Technol. 2018, 52, 13260−13269

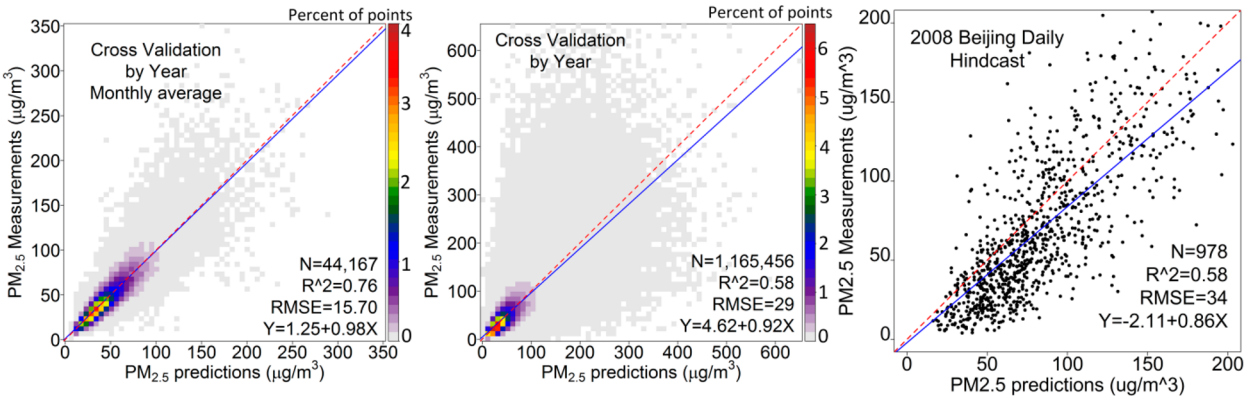


Figure 3. By-year CV results and the hindcast performance of the ensemble model.

Northwest cluster, characterized by substantial dust emissions from the Taklamakan Desert. The number of clusters affected the model hindcast performance (Figure S1). With the increase in number of clusters, the by-year CV accuracy increased at first and then decreased or remained relatively constant. The optimal number of clusters in this study is estimated as seven. Changing the initial randomly selected centroid only led to slightly different cluster patterns (Figure S2). This cluster pattern was also stable across years (Figure S3), slightly affected by changes in the number and the spatial distribution of monitors during the modeling period.

Individual Machine-Learning Model Performance. Table 1 shows the model fitting and CV performance of individual cluster-based models and reference national models at the daily level. The density plots of model fitting performance and CV performance are shown in Figures S4 and S5, respectively. The machine learning algorithms outperformed the statistical model. Although when fitted the LME+GAM model in 2013 only, the model fitting R^2 (0.80) is comparable with the results in a previous national study.[19] When fitting the LME+GAM model with four years of data (2013−2016), model performance was significantly worse (R^2 = 0.64). One reason, as we discussed in the introduction, is that the random effect on day of year was not sufficient to describe the temporal variations in the AOD-$PM_{2.5}$ relationship. Thus, using the LME+GAM model to predict $PM_{2.5}$ levels outside the modeling period may lead to larger prediction errors.

The cluster-based approach outperformed the national approach with all three machine-learning algorithms as well as the two-stage statistical model (Table 1). The CV R^2 values of cluster-based XGBoost, random forest, GAM, and LME +GAM models were 0.06, 0.06, 0.05, and 0.06 higher than their national counterparts. Thus, spatial clustering could improve model performance for both machine learning algorithms and statistical models. This is expected because the relations between $PM_{2.5}$ and its predictors would vary across our large spatial domain. By controlling unobserved spatial heterogeneity, the cluster-based models are able to capture the spatiotemporal variation in $PM_{2.5}$ more accurately than the national model. To our knowledge, this is the first data-driven method that divided China into temporally stable regions for $PM_{2.5}$ modeling purpose. This clustering approach could aid modeling efforts in the future by other researchers.

The ensemble prediction outperformed all individual models in cross-validation, with a CV R^2 of 0.79, RMSE of 21 $\mu g/m^3$, slope of 1.00, and intercept of 0.0 at the daily level (Table 1, Figure S5). The XGBoost model had the lowest CV RMSE (21 $\mu g/m^3$) and the highest CV R^2 (0.78) among individual models, followed by random forest (CV R^2 0.77, RMSE 22 $\mu g/m^3$). Since we excluded measurement-based predictors, the CV R^2 of our model was lower than some previous machine learning models that included spatial or/and temporal smooth surfaces of $PM_{2.5}$ estimated from ground measurements.[7,14] Our models are suitable for $PM_{2.5}$ hindcast prediction.

As expected, prediction error increased in temporal and spatial CV relative to the standard CV, indicating that unobserved spatial and temporal trends contributed to the prediction of $PM_{2.5}$ (Table 1). The random forest algorithm, the XGBoost algorithm, and the two-stage statistical model relied more on the temporal trend: the R^2 in spatial CV was higher than the R^2 in temporal CV. On the contrary, GAM relied more on the spatial trend and showed no temporal overfitting with the spatial CV R^2 (0.58) lower than the temporal CV R^2 (0.65). Additionally, we noticed that while generating accurate $PM_{2.5}$ estimates, decision tree based machine learning algorithms, e.g. random forest and XGBoost, have difficulties handling spatial predictors and including time-fixed spatial parameters led to unsmooth prediction maps. Thus, we combined predictions from different models that characterized different aspects of the complex relationships between $PM_{2.5}$ and predictors to improve model performance.

We observed spatial heterogeneity in parameter importance. Visibility and MERRA-2 $PM_{2.5}$ simulations are two of the most important parameters in all clusters regarding the importance index provided by random forest and XGBoost. Different from satellite AOD that describes vertical column aerosol loading, visibility is an indicator of horizontal aerosol loading and are associated with ground $PM_{2.5}$ concentrations.[46] MERRA-2 $PM_{2.5}$ simulations integrate data from various sources and have been shown to accurately describe large-scale $PM_{2.5}$ distributions in the U.S. and Europe.[31,32] However, both parameters are at relatively low spatial resolutions: the visibility data was measured at ∼400 stations in China and MERRA-2 simulations are at 0.625° × 0.5° resolution. The tropospheric vertical column NO_2 density and AAI from OMI also contributed significantly in $PM_{2.5}$ predictions, but resampling of the OMI data is necessary due to a row anomaly, leading to reduced temporal resolution. Although satellite AOD retrievals were not the most important variables in random forest and XGBoost models, they provided valuable information describing the fine-resolution spatial distribution of $PM_{2.5}$ at the daily level.

DOI: 10.1021/acs.est.8b02917
Environ. Sci. Technol. 2018, 52, 13260−13269

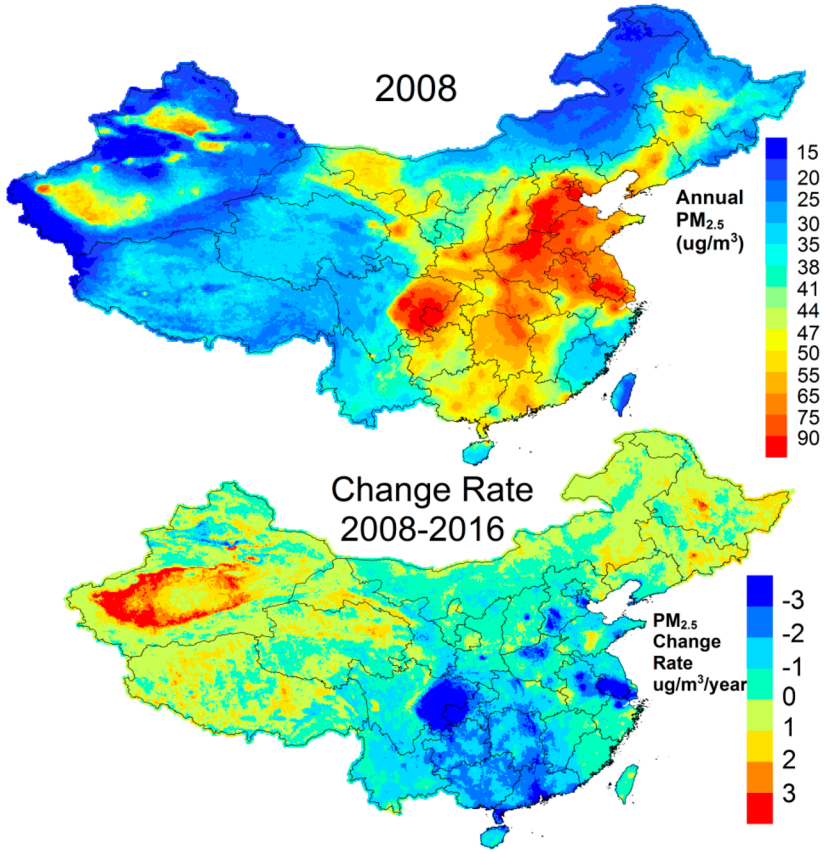


Figure 4. Annual $PM_{2.5}$ distribution in 2008 (above) and the estimated $PM_{2.5}$ change rate during 2008−2016 (below). The China map with province outlines was downloaded from http://www.resdc.cn/.

Model Hindcast Performance. Figure 3 shows the hindcast performance of the ensemble model. In the by-year CV, at monthly level, the hindcast predictions matched well with measurements, with a R^2 of 0.76 and a RMSE of 15.7 $\mu g/m^3$. The linear regression between the ensemble hindcast predictions and ground measurements produced a slope closest to unity (0.98) and a intercept closest to zero (1.25 $\mu g/m^3$), indicating a very minor prediction bias. We noticed that the model temporal 10-fold CV error still underestimated the daily hindcast prediction error. For example, the ensemble model had the temporal 10-fold CV R^2 (RMSE) of 0.73 (24 $\mu g/m^3$), while the daily hindcast R^2 was 0.58 (29 $\mu g/m^3$). This result suggested that intra-annual changes in $PM_{2.5}$ emission sources due to economic development and policy changes might affect the relationships between $PM_{2.5}$ and its predictors, but such changes in emission profiles were not well characterized in our current model. Comparing the by-year CV results of our machine learning algorithms with those of the LME+GAM model, the LME+GAM model had R^2 of 0.49 (RMSE of 32 $\mu g/m^3$) at a daily level and 0.71 (RMSE of 17.4 $\mu g/m^3$) at a monthly level (Figure S6). Thus, our ensemble model outperformed previous statistical models in hindcast predictions at both daily and monthly levels.[8,19]

We observed large variations in model hindcast performance across clusters. In general, clusters in the north and west had lower prediction accuracy than the other clusters (Table S1), partly due to increased missing AOD retrievals during the long (up to five months) winter in the north and west of China as well as the high surface reflectance over the desert in the west. Missing satellite data due to snow cover and bright surfaces can hardly be accurately imputed by the current imputation model that is designed to fill missing data due to cloud cover. Model performance using different algorithms remained relatively stable across the clusters, i.e., the XGBoost model outperformed the GAM model in all clusters. However, model performance statistics varied by cluster. For example, all the algorithms captured $PM_{2.5}$ levels in the Southeast well, but the GAM model generated significantly worse hindcasts in the Northwest.

To better evaluate our model's hindcast performance, we predicted $PM_{2.5}$ levels in 2008 (Figure 3). The daily hindcast predictions in 2008 were comparable with the by-year CV results, with an R^2 value of 0.58 and RMSE of 34 $\mu g/m^3$. Thus, the model performance did not appear to deteriorate in time. To increase the robustness of our model, we included 4 y worth of data for model training, whereas previous studies only used 1 or 2 y data or trained a separate model for each year.[7,14,19] Similar to the spatial clustering, training a model during a short time period, or temporal clustering, can better characterize short-term relationships. However, these annual models may estimate temporally unstable relationships that cannot be applied outside the modeling year. For example, when using only 2013 data to fit the models, the model fitting R^2 increased to 0.87, 0.88, and 0.75 for XGBoost, random forest, and GAM model, respectively. However, the hindcast R^2 decreased to 0.45, 0.49, 0.32, and 0.43 for XGBoost, random forest, GAM, and ensemble model, respectively (Figure S7). When fitting models with 1 y worth of data, the hindcast performance improved as the model-training year gets closer to 2017, the hindcast year. For example, the ensemble prediction

DOI: 10.1021/acs.est.8b02917
Environ. Sci. Technol. 2018, 52, 13260−13269

R^2 was 0.43, 0.45, 0.49, and 0.55 using models fitted with data of years 2013, 2014, 2015, and 2016, respectively (Figure S7). This result suggested a long-term trend in $PM_{2.5}$-predictor relationships, and the hindcast ability of the annual model deteriorated when predicting $PM_{2.5}$ levels long before the model-training year. On the contrary, our ensemble hindcast prediction agreed well with ground measurements in 2008, 5 y before the model-training period. Similarly, to ensure a robust modeling system in space and time, we preferred low-complexity models, e.g. trees with smaller height and smaller number of leaves. We noticed that although increasing model complexity to a certain degree improved model standard CV performance, it also increased the risk of spatial and temporal overfitting (i.e., lower spatial and temporal CV R^2 values).

The annual $PM_{2.5}$ distribution map in 2008 (Figure 4) indicated some hot spots of $PM_{2.5}$ in Beijing, Tianjin, Hebei province, and Henan province. As a demonstration of our ensemble hindcast model, we estimated the annual $PM_{2.5}$ change rate during 2008−2016 with linear regression and noticed that the air quality at these hot spots was significantly improved during this 8 y period. During this 8 y period, $PM_{2.5}$ levels decreased or remained constant in most parts of China.[47] The largest improvement in annual average $PM_{2.5}$ concentration occurred in the Sichuan basin, followed by Henan province, Hebei province, Tianjin City, Taiyuan City, the Yangtze River Delta, and Pearl River Delta, at more than 3 $\mu g/m^3$ per year. However, $PM_{2.5}$ levels in Northeast and Western China increased. For example, $PM_{2.5}$ levels in Heilongjiang, Jilin, Gansu, Qinghai, and Shandong province have increased at approximately 1−2 $\mu g/m^3$ per year. The Taklimakan Desert also experienced an increasing trend of $PM_{2.5}$ levels that was possibly due to the increased frequency of blowing dust events.[12] A previous study also reported an increase in PM_{10} levels measured in this region after 2008.[48]

One limitation of our ensemble prediction model is the underestimation of some high $PM_{2.5}$ values (Figure 3), which could be attributed to the retrieval error in AOD and the relatively coarse resolution of our national model. Previous studies indicated that MODIS collection 6 AOD retrievals tend to overestimate AOD values.[49,50] Calibrating satellite AOD against ground measurements from NASA's Aerosol Robotic Network (AERONET) may further improve the accuracy of AOD retrievals. However, there were only 10 operational AERONET stations in China during 2013−2016; this cannot support a reliable nationwide calibration. Assuming that the quality of satellite retrievals remains constant in time, extending the study period to include more AERONET stations could support a reliable calibration of satellite AOD and, therefore, improve the performance of our $PM_{2.5}$ prediction models. Regarding model resolution, we constructed a national 0.1° × 0.1° grid for data integration and model fitting since the highest resolution predictors, MODIS level 2 AOD retrievals, are at a 10 km nominal resolution. Additionally, for a national model, the 0.1° grid cells revealed enough spatial variations. However, some abnormally high $PM_{2.5}$ concentrations due to local emission sources can hardly be captured at this spatial scale. As shown in Figure S9, although the residual distribution did not show any spatial patterns, suggesting that the model had no systematic bias, we observed considerable spatial variations in $PM_{2.5}$ residual within the 0.1° grid cell. As a result, the misalignment between grid level $PM_{2.5}$ predictions and point measurements may lead to underestimate of very high $PM_{2.5}$ measurements. Employing AOD products with a higher spatial resolution, e.g. MAIAC aerosol products,[51,52] and constructing a finer modeling grid could result in better model performance at high $PM_{2.5}$ levels.

Another limitation of the machine learning model is the lack of explanatory capability. In this study we selected decision tree based algorithms (random forest and XGBoost) that estimated an importance index of each predictor to guide predictor selection.[39] Previous studies reported bias in the estimated predictor importance and presented various ways to adjust the estimated importance index.[27,38] Since we aimed to predict historical $PM_{2.5}$ concentrations, these methods and further analyses of predictor contribution are beyond the scope of this study.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.8b02917.

Text, figures, and tables that provide additional information on hyperparameter optimization, performance of individual models, and sensitivity analyses of the spatial clustering method (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Mailing address: Emory University, Rollins School of Public Health, 1518 Clifton Road NE, Atlanta, GA 30322, USA. Phone: (404) 727-2131. E-mail: yang.liu@emory.edu.

ORCID

Qingyang Xiao: 0000-0001-5910-884X
Guannan Geng: 0000-0002-1605-8448
Yang Liu: 0000-0001-5477-2186

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the NASA Applied Sciences Program (Grant no. NNX16AQ28G, PI: Liu) and the National Institutes of Health (Grant no. R01ES027892, PI: Chang). We thank Fengchao Liang of Fuwai Hospital for providing part of the $PM_{2.5}$ measurements in Beijing during 2008.

■ REFERENCES

(1) Sorek-Hamer, M.; Just, A. C.; Kloog, I. The Use of Satellite Remote Sensing in Epidemiological Studies. *Curr. Opin. Pediatr.* **2016**, *28* (2), 228.

(2) Brauer, M.; Amann, M.; Burnett, R. T.; Cohen, A.; Dentener, F.; Ezzati, M.; Henderson, S. B.; Krzyzanowski, M.; Martin, R. V.; Van Dingenen, R. Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. *Environ. Sci. Technol.* **2012**, *46* (2), 652−660.

(3) Hu, X.; Waller, L.; Lyapustin, A.; Wang, Y.; Liu, Y. 10-year spatial and temporal trends of PM 2.5 concentrations in the southeastern US estimated using high-resolution satellite data. *Atmos. Chem. Phys.* **2014**, *14* (12), 6301−6314.

(4) Ma, Z.; Hu, X.; Huang, L.; Bi, J.; Liu, Y. Estimating ground-level PM2. 5 in China using satellite remote sensing. *Environ. Sci. Technol.* **2014**, *48* (13), 7436−7444.

(5) Xiao, Q.; Wang, Y.; Chang, H. H.; Meng, X.; Geng, G.; Lyapustin, A.; Liu, Y. Full-coverage high-resolution daily PM2. 5 estimation using MAIAC AOD in the Yangtze River Delta of China. *Remote Sensing of Environment* **2017**, *199*, 437−446.

DOI: 10.1021/acs.est.8b02917
Environ. Sci. Technol. 2018, 52, 13260−13269

(6) Kloog, I.; Chudnovsky, A. A.; Just, A. C.; Nordio, F.; Koutrakis, P.; Coull, B. A.; Lyapustin, A.; Wang, Y.; Schwartz, J. 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. *Atmos. Environ.* **2014**, *95*, 581−590.
(7) Di, Q.; Kloog, I.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Schwartz, J. Assessing PM2. 5 exposures with high spatiotemporal resolution across the continental United States. *Environ. Sci. Technol.* **2016**, *50* (9), 4712−4721.
(8) Weng, Q.; Xu, B.; Hu, X.; Liu, H. Use of earth observation data for applications in public health. *Geocarto International* **2014**, *29* (1), 3−16.
(9) 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.
(10) Van Donkelaar, A.; Martin, R. V.; Park, R. J. Estimating ground-level PM2. 5 using aerosol optical depth determined from satellite remote sensing. *J. Geophys. Res.* **2006**, *111*, D21.
(11) Geng, G.; Zhang, Q.; Martin, R. V.; van Donkelaar, A.; Huo, H.; Che, H.; Lin, J.; He, K. Estimating long-term PM 2.5 concentrations in China using satellite-based aerosol optical depth and a chemical transport model. *Remote Sensing of Environment* **2015**, *166*, 262−270.
(12) van Donkelaar, A.; Martin, R. V.; Brauer, M.; Hsu, N. C.; Kahn, R. A.; Levy, R. C.; Lyapustin, A.; Sayer, A. M.; Winker, D. M. Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. *Environ. Sci. Technol.* **2016**, *50* (7), 3762.
(13) Hu, X.; Belle, J. H.; Meng, X.; Wildani, A.; Waller, L.; Strickland, M.; Liu, Y. Estimating PM2. 5 Concentrations in the Conterminous United States Using the Random Forest Approach. *Environ. Sci. Technol.* **2017**, *51* (12), 6936.
(14) Li, T.; Shen, H.; Yuan, Q.; Zhang, X.; Zhang, L. Estimating ground-level PM2. 5 by fusing satellite and station observations: A geo-intelligent deep learning approach. *Geophys. Res. Lett.* **2017**, *44* (23), 11985.
(15) Zhan, Y.; Luo, Y.; Deng, X.; Chen, H.; Grieneisen, M. L.; Shen, X.; Zhu, L.; Zhang, M. Spatiotemporal prediction of continuous daily PM 2.5 concentrations across China using a spatially explicit machine learning algorithm. *Atmos. Environ.* **2017**, *155*, 129−139.
(16) Li, T.; Shen, H.; Zeng, C.; Yuan, Q.; Zhang, L. Point-surface fusion of station measurements and satellite observations for mapping PM 2.5 distribution in China: Methods and assessment. *Atmos. Environ.* **2017**, *152*, 477−489.
(17) 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 PM2. 5 exposures in the Mid-Atlantic states. *Environ. Sci. Technol.* **2012**, *46* (21), 11913−11921.
(18) Kloog, I.; Sorek-Hamer, M.; Lyapustin, A.; Coull, B.; Wang, Y.; Just, A. C.; Schwartz, J.; Broday, D. M. Estimating daily PM 2.5 and PM 10 across the complex geo-climate region of Israel using MAIAC satellite-based AOD data. *Atmos. Environ.* **2015**, *122*, 409−416.
(19) Ma, Z.; Hu, X.; Sayer, A. M.; Levy, R.; Zhang, Q.; Xue, Y.; Tong, S.; Bi, J.; Huang, L.; Liu, Y. Satellite-based spatiotemporal trends in PM2. 5 concentrations: China, 2004−2013. *Environ. Health Perspect.* **2016**, *124* (2), 184.
(20) Liu, Y.; He, K.; Li, S.; Wang, Z.; Christiani, D. C.; Koutrakis, P. A statistical model to evaluate the effectiveness of PM 2.5 emissions control during the Beijing 2008 Olympic Games. *Environ. Int.* **2012**, *44*, 100−105.
(21) Levy, R.; Mattoo, S.; Munchak, L.; Remer, L.; Sayer, A.; Patadia, F.; Hsu, N. The Collection 6 MODIS aerosol products over land and ocean. *Atmos. Meas. Tech.* **2013**, *6*, 2989−3034.
(22) Hsu, N.; Jeong, M. J.; Bettenhausen, C.; Sayer, A.; Hansell, R.; Seftor, C.; Huang, J.; Tsay, S. C. Enhanced Deep Blue aerosol retrieval algorithm: The second generation. *Journal of Geophysical Research: Atmospheres* **2013**, *118* (16), 9296−9315.
(23) Xiao, Q.; Zhang, H.; Choi, M.; Li, S.; Kondragunta, S.; Kim, J.; Holben, B.; Levy, R.; Liu, Y. Evaluation of VIIRS, GOCI, and MODIS Collection 6 AOD retrievals against ground sunphotometer observations over East Asia. *Atmos. Chem. Phys.* **2016**, *16* (3), 1255−1269.
(24) Jinnagara Puttaswamy, S.; Nguyen, H. M.; Braverman, A.; Hu, X.; Liu, Y. Statistical data fusion of multi-sensor AOD over the Continental United States. *Geocarto International* **2014**, *29* (1), 48−64.
(25) Platnick, S.; King, M. D.; Ackerman, S. A.; Menzel, W. P.; Baum, B. A.; Riédi, J. C.; Frey, R. A. The MODIS cloud products: Algorithms and examples from Terra. *IEEE Transactions on Geoscience and Remote Sensing* **2003**, *41* (2), 459−473.
(26) Dobson, J. E.; Bright, E. A.; Coleman, P. R.; Durfee, R. C.; Worley, B. A. LandScan: a global population database for estimating populations at risk. *Photogram. Eng. Remote Sens.* **2000**, *66* (7), 849−857.
(27) Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional variable importance for random forests. *BMC Bioinf.* **2008**, *9*, 307−307.
(28) Friedman, J.; Hastie, T.; Tibshirani, R. *The elements of statistical learning*; Springer series in statistics: New York, 2001; Vol. *1*.
(29) Randles, C.; da Silva, A. M.; Buchard, V.; Colarco, P.; Darmenov, A.; Govindaraju, R.; Smirnov, A.; Holben, B.; Ferrare, R.; Hair, J.; et al. The MERRA-2 aerosol reanalysis, 1980 onward. Part I: System description and data assimilation evaluation. *J. Clim.* **2017**, *30* (17), 6823−6850.
(30) Buchard, V.; Randles, C.; da Silva, A.; Darmenov, A.; Colarco, P.; Govindaraju, R.; Ferrare, R.; Hair, J.; Beyersdorf, A.; Ziemba, L.; et al. The MERRA-2 aerosol reanalysis, 1980 onward. Part II: Evaluation and case studies. *J. Clim.* **2017**, *30* (17), 6851−6872.
(31) Buchard, V.; da Silva, A.; Randles, C.; Colarco, P.; Ferrare, R.; Hair, J.; Hostetler, C.; Tackett, J.; Winker, D. Evaluation of the surface PM 2.5 in Version 1 of the NASA MERRA Aerosol Reanalysis over the United States. *Atmos. Environ.* **2016**, *125*, 100−111.
(32) Provençal, S.; Buchard, V.; da Silva, A. M.; Leduc, R.; Barrette, N. Evaluation of PM surface concentrations simulated by Version 1 of NASA's MERRA Aerosol Reanalysis over Europe. *Atmos. Pollut. Res.* **2017**, *8* (2), 374−382.
(33) Reid, C. E.; Jerrett, M.; Petersen, M. L.; Pfister, G. G.; Morefield, P. E.; Tager, I. B.; Raffuse, S. M.; Balmes, J. R. Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. *Environ. Sci. Technol.* **2015**, *49* (6), 3887−3896.
(34) Yanosky, J. D.; Paciorek, C. J.; Laden, F.; Hart, J. E.; Puett, R. C.; Liao, D.; Suh, H. H. Spatio-temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors. *Environ. Health* **2014**, *13* (1), 63.
(35) Brunsdon, C.; Fotheringham, A. S.; Charlton, M. E. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis* **1996**, *28* (4), 281−298.
(36) Breiman, L. Random forests. *Machine learning* **2001**, *45* (1), 5−32.
(37) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*; ACM: 2016; pp 785−794.
(38) Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinf.* **2007**, *8*, 25−25.
(39) Archer, K. J.; Kimes, R. V. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis* **2008**, *52* (4), 2249−2260.
(40) Chen, T.; He, T. *Xgboost: extreme gradient boosting*; R package version 0.4-2, 2015.
(41) Li, H.; Zhang, Q.; Zhang, Q.; Chen, C.; Wang, L.; Wei, Z.; Zhou, S.; Parworth, C.; Zheng, B.; Canonaco, F.; et al. Wintertime aerosol chemistry and haze evolution in an extremely polluted city of the North China Plain: significant contribution from coal and biomass combustion. *Atmos. Chem. Phys.* **2017**, *17* (7), 4751−4768.

DOI: 10.1021/acs.est.8b02917
Environ. Sci. Technol. 2018, 52, 13260−13269

(42) Huang, L.; Hu, J.; Chen, M.; Zhang, H. Impacts of power generation on air quality in China—part I: an overview. *Resources, Conservation and Recycling* **2017**, *121*, 103−114.
(43) Xu, W.; Zhao, C.; Ran, L.; Deng, Z.; Liu, P.; Ma, N.; Lin, W.; Xu, X.; Yan, P.; He, X.; et al. Characteristics of pollutants and their correlation to meteorological conditions at a suburban site in the North China Plain. *Atmos. Chem. Phys.* **2011**, *11* (9), 4353−4369.
(44) Zhao, X.; Zhao, P.; Xu, J.; Meng, W.; Pu, W.; Dong, F.; He, D.; Shi, Q. Analysis of a winter regional haze event and its formation mechanism in the North China Plain. *Atmos. Chem. Phys.* **2013**, *13* (11), 5685−5696.
(45) Liu, J.; Zhao, D.; Gerbens-Leenes, P.; Guan, D. China's rising hydropower demand challenges water sector. *Sci. Rep.* **2015**, *5*, 11446.
(46) Liu, M.; Bi, J.; Ma, Z. Visibility-Based PM2. 5 Concentrations in China: 1957−1964 and 1973−2014. *Environ. Sci. Technol.* **2017**, *51* (22), 13161−13169.
(47) Di, Q.; Wang, Y.; Zanobetti, A.; Wang, Y.; Koutrakis, P.; Choirat, C.; Dominici, F.; Schwartz, J. D. Air Pollution and Mortality in the Medicare Population. *N. Engl. J. Med.* **2017**, *376* (26), 2513−2522.
(48) Zhang, X.-X.; Sharratt, B.; Chen, X.; Wang, Z.-F.; Liu, L.-Y.; Guo, Y.-H.; Li, J.; Chen, H.-S.; Yang, W.-Y. Dust deposition and ambient PM 10 concentration in northwest China: spatial and temporal variability. *Atmos. Chem. Phys.* **2017**, *17* (3), 1699−1711.
(49) Fan, A.; Chen, W.; Liang, L.; Sun, W.; Lin, Y.; Che, H.; Zhao, X. Evaluation and Comparison of Long-Term MODIS C5. 1 and C6 Products against AERONET Observations over China. *Remote Sensing* **2017**, *9* (12), 1269.
(50) de Leeuw, G.; Sogacheva, L.; Rodriguez, E.; Kourtidis, K.; Georgoulias, A. K.; Alexandri, G.; Amiridis, V.; Proestakis, E.; Marinou, E.; Xue, Y.; van der A, R. Two decades of satellite observations of AOD over mainland China using ATSR-2, AATSR and MODIS/Terra: data set evaluation and large-scale patterns. *Atmos. Chem. Phys.* **2018**, *18* (3), 1573−1592.
(51) Lyapustin, A.; Martonchik, J.; Wang, Y.; Laszlo, I.; Korkin, S. Multiangle implementation of atmospheric correction (MAIAC): 1. Radiative transfer basis and look-up tables. *J. Geophys. Res.* **2011**, DOI: 10.1029/2010JD014985.
(52) Lyapustin, A.; Wang, Y.; Laszlo, I.; Kahn, R.; Korkin, S.; Remer, L.; Levy, R.; Reid, J. Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. *J. Geophys. Res.* **2011**, DOI: 10.1029/2010JD014986.

DOI: 10.1021/acs.est.8b02917
Environ. Sci. Technol. 2018, 52, 13260−13269