



Space-time trends of PM_{2.5} constituents in the conterminous United States estimated by a machine learning approach, 2005–2015



Xia Meng^a, Jenny L. Hand^b, Bret A. Schichtel^c, Yang Liu^{a,*}

^a Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA

^b Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA

^c National Park Service, Air Resources Division, Lakewood, CO, USA

ARTICLE INFO

Handling Editor: X Querol

Keywords:

PM_{2.5}
Sulfate
Nitrate
Elemental carbon
Organic carbon
Random forest

ABSTRACT

Particulate matter with aerodynamic diameter less than 2.5 μm (PM_{2.5}) is a complex mixture of chemical constituents emitted from various emission sources or through secondary reactions/processes; however, PM_{2.5} is regulated mostly based on its total mass concentration. Studies to identify the impacts on climate change, visibility degradation and public health of different PM_{2.5} constituents are hindered by limited ground measurements of PM_{2.5} constituents. In this study, national models were developed based on random forest algorithm, one of machine learning methods that is of high predictive capacity and able to provide interpretable results, to predict concentrations of PM_{2.5} sulfate, nitrate, organic carbon (OC) and elemental carbon (EC) across the conterminous United States from 2005 to 2015 at the daily level. The random forest models achieved high out-of-bag (OOB) R² values at the daily level, and the mean OOB R² values were 0.86, 0.82, 0.71 and 0.75 for sulfate, nitrate, OC and EC, respectively, over 2005–2015. The long-term temporal trends of PM_{2.5} sulfate, nitrate, OC and EC predictions agreed well with their corresponding ground measurements. The annual mean of predicted PM_{2.5} sulfate and EC levels across the conterminous United States decreased substantially from 2005 to 2015; while concentrations of predicted PM_{2.5} nitrate and OC decreased and fluctuated during the study period. The annual prediction maps captured the characterized spatial patterns of the PM_{2.5} constituents. The distributions of annual mean concentrations of sulfate and nitrate were generally regional in the extent that sulfate decreased from east to west smoothly with enhancement in California and nitrate had higher concentration in Midwest, Metro New York area, and California. OC and EC had regional high concentrations in the Southeast and Northwest as well as localized high levels around urban centers. The spatial patterns of PM_{2.5} constituents were consistent with the distributions of their emission sources and secondary processes and transportation. Hence, the national models developed in this study could provide supplementary evaluations of spatio-temporal distributions of PM_{2.5} constituents with full time-space coverages in the conterminous United States, which could be beneficial to assess the impacts of PM_{2.5} constituents on radiation budgets and visibility degradation, and support exposure assessment for regional to national health studies at county or city levels to understand the acute and chronic toxicity and health impacts of PM_{2.5} constituents, and consequently provide scientific evidence for making targeted and effective regulations of PM_{2.5} pollution.

1. Introduction

Particulate matter (PM) is a complex mixture of chemical constituents generated by direct emissions from various sources as well as through photochemical reactions in the atmosphere (Hand et al., 2014). The impacts of ambient PM on climate change, visibility degradation and public health have drawn concerns worldwide (Beelen et al., 2014; Chen et al., 2017; Cohen et al., 2017; Di et al., 2017; IPCC, 2013; Kaufman et al., 2012; Lepeule et al., 2012; Malm et al., 1994). PM could

affect climate and impair visibility by scattering and absorbing sunlight, and modifying the Earth's radiative transfer (IPCC, 2013; Khan et al., 2010; Malm et al., 1994). PM constituents play different roles in climate change and visibility degradation. Non-absorbing components, such as particle sulfate and nitrate, scatter the solar radiation and have the effect of cooling the atmosphere; while absorbing components, such as elemental carbon and brown organic carbon, absorb the solar radiation and produce the warming effect. The light extinction efficiencies also vary significantly among PM constituents (Hand et al., 2011), which

* Corresponding author at: Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA.

E-mail address: yang.liu@emory.edu (Y. Liu).

<https://doi.org/10.1016/j.envint.2018.10.029>

Received 3 August 2018; Received in revised form 14 October 2018; Accepted 15 October 2018

Available online 07 November 2018

0160-4120/ © 2018 Elsevier Ltd. All rights reserved.

affects their capacity of visibility reduction. PM with aerodynamic diameter less than $2.5\mu\text{m}$ ($\text{PM}_{2.5}$) has become the fifth-ranking mortality risk factor globally in 2015 (Cohen et al., 2017); however, $\text{PM}_{2.5}$ is currently regulated and its health effects are mostly studied based on its total mass concentration and the knowledge regarding differential toxicities of $\text{PM}_{2.5}$ constituents is limited and the results are inconsistent based on studies with different research designs and usually in small study domains (Achilleos et al., 2017; Bergen et al., 2013; Chung et al., 2015; EPA, 2009; Rohr and Wyzga, 2012). The concentrations of PM constituents are extremely variable in space and time. Therefore, understanding the spatial and temporal distributions of the PM constituents is crucial to identify the PM components and emission sources responsible for local and regional atmospheric warming and cooling effects as well as visibility impairment of PM, and to support toxicological and epidemiological studies in large spatial scales (e.g., regional or national scales) to better understand the most harmful $\text{PM}_{2.5}$ constituents to public health. Finally, all these scientific evidence could be able to support policy making to control $\text{PM}_{2.5}$ pollution effectively (EPA, 2008; Fuzzi et al., 2015; IPCC, 2013; Malm et al., 1994).

However, understanding the spatial and temporal distributions of the PM constituents in large spatial scales are often hindered by the sparse or non-existent coverage of ground measurements of $\text{PM}_{2.5}$ constituents (Pope and Dockery, 2006). Even in the United States, which has the most extensive ground monitoring networks of $\text{PM}_{2.5}$ constituents in the world with 300+ monitoring sites, the monitoring sites of $\text{PM}_{2.5}$ constituents are still too sparse comparing with those of monitoring $\text{PM}_{2.5}$ mass concentrations to fully illustrate the variabilities of $\text{PM}_{2.5}$ constituents in space and time, especially in western United States and rural areas. Therefore, national-scale models that predict concentrations of $\text{PM}_{2.5}$ constituents are highly needed. Several modeling methods have been proposed in previous studies to estimate concentrations or population exposure levels of $\text{PM}_{2.5}$ constituents in large spatial scales. The nearest neighbor method was used to study the association between long-term exposure to $\text{PM}_{2.5}$ constituents and mortality in the Medicare population in the eastern United States. Exposure was assigned to observations from a monitoring site within six miles of the enrollee's ZIP code centroid (Chung et al., 2015). This simple approach ignores the spatial heterogeneity of different $\text{PM}_{2.5}$ constituents and is not suitable in areas with sparse monitoring sites, e.g. in the western United States. Land use regression (LUR) and Universal Kriging (UK) were also developed to estimate concentrations of $\text{PM}_{2.5}$ constituents in Europe and the United States (Bergen et al., 2013; Wang et al., 2014). Despite their good performance, LUR and UK models are normally developed to estimate concentrations of $\text{PM}_{2.5}$ constituents at the annual level for studies on chronic health effects other than those on the acute or sub-chronical health impacts of $\text{PM}_{2.5}$ constituents, since the models are mainly developed on geographic variables of low temporal resolutions. Statistical models driven by remote sensing data are able to estimate concentrations of $\text{PM}_{2.5}$ constituents at the daily level (Liu et al., 2007a, 2007b; Meng et al., 2018). For example, Liu et al. (2007b) developed regression models with fractional aerosol optical depth (AOD) values derived from the Multi-angle Imaging Spectroradiometer (MISR) aerosol retrievals to estimate concentrations of $\text{PM}_{2.5}$ sulfate, nitrate and OC in the United States (Liu et al., 2007b). However, the long-term average predictions based on MISR data may be biased because MISR's sampling frequency is low and missing AOD caused by cloud or high surface reflectance is not random. Chemical transport models (CTMs) are able to predict $\text{PM}_{2.5}$ constituents with full temporal and spatial coverages. The accuracy of CTMs simulation is relatively low because of outdated emissions inventory, e.g. the National Emissions Inventory (NEI) are updated every three years, and imperfect understanding of the formation mechanisms of aerosol components (Baek et al., 2011). A comparison between simulated concentrations from a global 3-D CTM (GEOS-Chem) and ground measurements showed that the R^2 values between the two datasets were 0.22, 0.25, 0.28 and 0.30 for $\text{PM}_{2.5}$ EC, OC, nitrate and

sulfate, respectively, in the northeastern United States between 2005 and 2010 (Di et al., 2016b); therefore, the low accuracy limits the application of GEOS-Chem simulations to directly evaluate concentrations of $\text{PM}_{2.5}$ constituents quantitatively, and further corrections of the simulations are usually needed. Scaling the CTM simulations by satellite AOD data improved the correlations between simulated $\text{PM}_{2.5}$ constituents and the corresponding ground measurements, but the correlations between ground measurements and the scaled nitrate, organic mass (OM) and EC were still low (Philip et al., 2014). Further research is needed to develop models with high predicting accuracy, temporal flexibility and spatial resolution to estimate concentrations of $\text{PM}_{2.5}$ constituents in large spatial scales, subsequently provide short- and long-term exposure assessments for future epidemiological studies, especially in areas with none or few monitoring sites of $\text{PM}_{2.5}$ constituents.

In order to generate accurate predictions and capture the spatial-temporal characteristics of ground PM levels, more predictor variables are involved in modeling strategies, including land cover parameters, meteorological variables, remote sensing data, simulated results from CTMs, and spatial and temporal indicators (e.g. spatial coordinates and index for day of the year) (Brauer et al., 2012; Hu et al., 2014; Hu et al., 2017; Li et al., 2017; Liu et al., 2009a; Liu et al., 2018). The relationships between PM concentrations and these predictor variables are complex and might not be fully explained by statistical modeling methods. For example, linear mixed effects (LME) method assumes the linear relationship between $\text{PM}_{2.5}$ and predictor variables, and requires low collinearity among predictor variables. Therefore, machine learning algorithms are increasingly being explored as they require fewer and less restrictive assumptions with regard to the independence of observations and distributions of predictor variables (Di et al., 2016a; Hu et al., 2017; Liu et al., 2018). For examples, Hu et al. predicted $\text{PM}_{2.5}$ mass concentrations in the United States with random forest method. Di et al. employed neural network method to calibrate GEOS-Chem simulated $\text{PM}_{2.5}$ constituents in the northeastern United States. Unlike other machine learning algorithms, e.g. neural network approach, random forest algorithm is highlighted by being able to produce importance ranking of all predictor variables (Gromping, 2009; Liaw and Wiener, 2002) to measure their predicting strengths. The importance ranking of predictors could make the modeling results more interpretable by helping understand the spatial and temporal characteristics of PM constituents, and identify their potential emission sources.

To date, efforts on predicting concentrations of $\text{PM}_{2.5}$ constituents by combining multiple predictors and based on machine learning algorithms are still very limited, especially in the national scale for characterizing the nationwide spatial and temporal variabilities of $\text{PM}_{2.5}$ constituents. In this study, we developed national models by incorporating land use parameters, meteorological variables and GEOS-Chem simulated $\text{PM}_{2.5}$ constituents to predict concentrations of $\text{PM}_{2.5}$ sulfate, nitrate, OC and EC in the conterminous United States from 2005 to 2015. Firstly, random forest models were developed for $\text{PM}_{2.5}$ sulfate, nitrate, OC and EC in each year. Temporal and spatial cross validations were followed to evaluate the capabilities of the models for capturing the temporal and spatial variabilities of $\text{PM}_{2.5}$ constituents. Then the spatial patterns as well as long-term temporal trends of $\text{PM}_{2.5}$ sulfate, nitrate, OC and EC were illustrated based on the predictions in the Results section and were further discussed in Discussion section. Finally, we proposed perspectives for future applications of our models and predictions.

2. Data and methods

Our study domain was defined as the 48 conterminous states and Washington DC, and the study period was from year 2005 to 2015. We used several similar predictors as in Hu et al. (2017), including meteorological data, land use and population data, and spatial and

temporal indicators (Hu et al., 2017). The first main difference between our dataset and the dataset in Hu et al. (2017) is that the previous study focused on PM_{2.5} mass concentrations while we tried to predict concentrations of PM_{2.5} constituents, including PM_{2.5} sulfate, nitrate, OC and EC. Therefore, they used measured concentrations of total PM_{2.5} mass of 1200+ sites from EPA while we used measurements of only 300+ sites from both the U.S. Environmental Protection Agency's Chemical Speciation Network (EPA CSN) and Interagency Monitoring of Protected Visual Environments (IMPROVE) network. Secondly, to spatially align various input datasets and make predictions, we created a modeling grid covering our study domain with a spatial resolution of $0.25^\circ \times 0.3125^\circ$, which was coarser than the 10 km resolution in Hu et al. (2017). There are far fewer PM_{2.5} speciation sites than those of PM_{2.5} mass concentrations; hence, the $0.25^\circ \times 0.3125^\circ$ spatial resolution is reasonable to capture details of regional and local spatial patterns of PM_{2.5} constituents. Additionally, in this conceptual experimental study, we mainly focused on model development and evaluation; therefore, a relatively coarse modeling grid was used as a trade-off for computational efficiency as we included 11 years and four PM_{2.5} species in this study. Thirdly, satellite remote sensed AOD data were not included in this study since few of them were designed to distinguish PM_{2.5} constituents. Although MISR fractional AODs could provide information regarding aerosols microphysical properties, its low sampling frequency would limit our sample size and our models' capability of producing daily predictions. Description of input datasets and modeling strategy were provided below.

2.1. Ground measurements of PM_{2.5} constituents

Daily-averaged concentrations of PM_{2.5} sulfate (SO₄²⁻), nitrate (NO₃⁻), organic carbon (OC) and elemental carbon (EC) in 2005–2015 were obtained from the CSN (<http://aqsd1.epa.gov>) and IMPROVE (<http://views.cira.colostate.edu/fed>) network. Together, there are more than 300 speciation monitoring sites in the study domain (Fig. 1). IMPROVE sites are primarily located in rural areas including many National Parks and Wilderness Areas, while most of the CSN sites are located in urban and suburban areas. The sampling frequency is 1-in-3 or 1-in-6 days for CSN and 1-in-3 days for IMPROVE. Different samplers and analytical methods cause systemic differences between

measurements of PM_{2.5} carbonaceous components in CSN and IMPROVE (Solomon et al., 2014). As a result, PM_{2.5} EC and OC data from the CSN need to be adjusted for compatibility with the IMPROVE (Hand et al., 2011; Malm et al., 2011) before model development. The details of the adjustment were described in Text 1 of supplementary material.

2.2. GEOS-Chem simulated PM_{2.5} constituents

GEOS-Chem is a global 3-D chemical transport model for atmospheric composition driven by meteorological input from the Goddard Earth Observing System (GEOS) of the National Aeronautics and Space Administration (NASA) Global Modeling Assimilation Office (GMAO). Details of GEOS-Chem can be found at its official website <http://acmg.seas.harvard.edu/geos/>. The details of the simulations used in this study were summarized in Text 2 of supplementary material. The simulated concentrations of PM_{2.5} constituents were assigned to the ground monitoring site if the site fell inside the GEOS-Chem simulation grid when preparing modeling dataset, and were assigned to the grid if the centroid of the grid fell inside the GEOS-Chem simulation grid when preparing predicting dataset.

2.3. Meteorological data

Meteorological variables were obtained from the North American Regional Reanalysis (NARR) data products at 32 km spatial resolution (<https://www.esrl.noaa.gov/psd/data/gridded/data.narr.html>). Air temperature (AT), dew point temperature (DPT), relative humidity (RH), surface pressure (SP), visibility (VIS), planetary boundary layer height (PBLH), east-west component of the wind vector (u-wind), north-south component of the wind vector (v-wind), downward long-wave radiation flux (DLWRF), and downward shortwave radiation flux (DSWRF) were included in the models. Meteorological data were interpolated to ground monitoring sites by inverse-distance weighting method when preparing modeling dataset, and were assigned to predicting grids by nearest neighbor method.

2.4. Land use and population data

Elevation data were extracted from Advanced Spaceborne Thermal

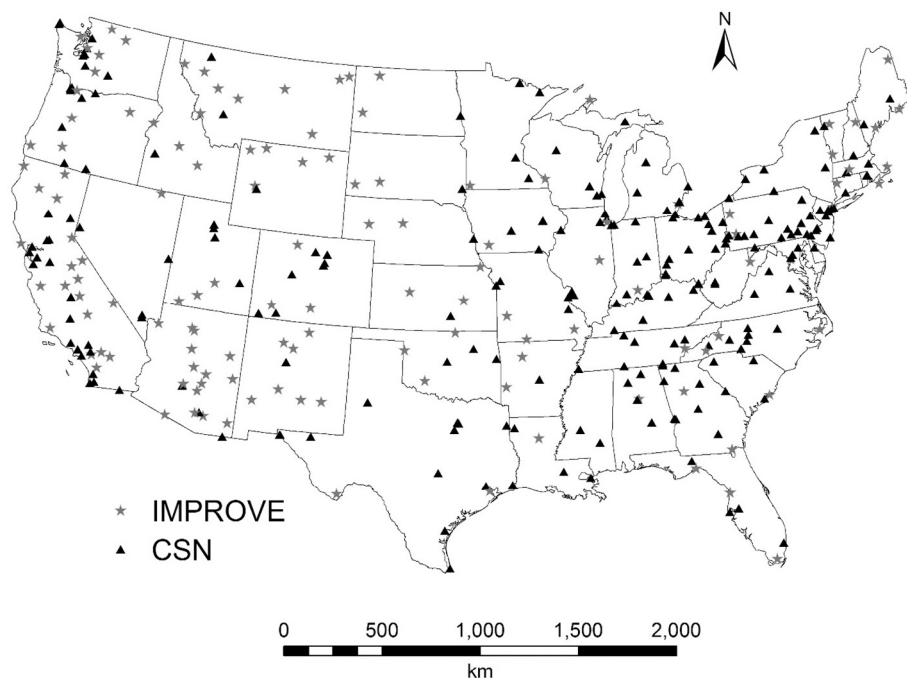


Fig. 1. Locations of CSN and IMPROVE monitoring sites for PM_{2.5} constituents in 2010.

Emission and Reflection Radiometer (ASTER) Global Digital Elevation Map (GDEM) (<https://asterweb.jpl.nasa.gov/gdem.asp>) version 2 to indicate topography. Major roads in 2011 were obtained from ESRI StreetMap USA (Environmental Systems Research Institute, Inc., Redlands, CA) to indicate local traffic emissions. Impervious surface, cultivated land cover and forest cover information at 30 m resolution were obtained from the National Land Cover Database 2011 (NLCD 2011, <https://www.mrlc.gov>). The population data from 2010 census were obtained from the United States Census Bureau (<https://www.census.gov>). 1 km × 1 km buffers were created around ground monitoring sites to aggregate population number, major road length, elevation and land cover data. The same procedure was applied for preparing predicting dataset.

2.5. Spatial and temporal indicators

The geographic coordinates of ground monitoring sites (x, y) were included as spatial indicators to reflect the potential spatial distributions of PM_{2.5} constituents. Month of the year (Month) was included as covariate to account for systematic seasonal trend within the year.

2.6. Modeling methods

Random forest models were employed for PM_{2.5} sulfate, nitrate, OC and EC in each year over 2005–2015, by incorporating GEOS-Chem simulated concentrations of PM_{2.5} constituents, meteorological variables, land use data, population data, spatial and temporal indicators, as well as spatial lagging terms of measured PM_{2.5} constituents and land use parameters in this study. Previous studies have found that including spatial autocorrelation terms (named as spatial lagging terms in this study) of measurements of PM_{2.5} constituents, and spatial trends of land use data as input predictor variables can improve model performance by considering potential influence from neighboring areas (Di et al., 2016b; Hu et al., 2017). Spatial lagging terms of measurements of PM_{2.5} constituents and land use data were calculated by the distance-inversed weighted average function for each ground monitoring site in modeling dataset and each grid in predicting dataset (details were summarized in supplementary material Text 3).

Details of random forest models have been given elsewhere (Breiman, 2001; Hu et al., 2017; Liaw and Wiener, 2002) and were summarized here. Generally, random forest is an ensemble method based on decision trees. For growing each tree in the forest, bootstrap sampling is used to draw the training set from the original dataset with replacement. Therefore, prediction can be made on the data not in the training set (called “out-of-bag” data, aka OOB data); then the OOB error estimate (expressed as mean squared error, MSE, defined as the average of the squares of the differences between the predictions and measurements) of the random forest model can be calculated by aggregating the OOB predictions from all trees and comparing the predictions with their corresponding measurements. The OOB prediction errors can be used to select model parameters, validate model's performance and produce importance ranking of predictors. Two parameters need to be set for random forest modeling based on package “randomForest” in R: m_{try} , which means the number of variables selected randomly for modeling at each node of the tree, and n_{tree} , which means the number of trees in the forest. We set m_{try} as 9 which is approximately one third of total predictor variables number (Liaw, 2015). The OOB MSE values of our random forest models decreased very slightly and almost kept stable when n_{tree} was greater than 300; therefore we set n_{tree} as 500 to get the balance of prediction accuracy and computing time. We calculated R^2 values (mentioned as OOB R^2 in the following parts) and normalized mean errors (NME, defined as the absolute value of difference between prediction and measurement divided by the measurement) between OOB predictions and their corresponding measurements to evaluate the model performance since the OOB predicting is also out-of-sample predictions. On the average, each

data point would be out-of-bag around 36% of the times (Liaw and Wiener, 2002). Random forest is able to measure the variables' predicting strength (expressed as importance ranking) by comparing the MSE between predictions before and after that variable is permuted while all others are left unchanged, with the assumption that permuting the values of one variable won't produce more prediction errors if the variable dose not perform significantly in the model.

For spatial cross validation (CV), all sites were divided into 10 subsets randomly and equally, data of the 90% sites were used as training dataset each time to develop model and data of the remaining 10% sites were used to produce predictions. This CV procedure should be repeated for 10 times, then R^2 was calculated between predictions and measurements of all data. Similarly, temporal CV was conducted based on days in each year.

3. Results

3.1. Descriptive statistics of the model fitting dataset

Annual mean ground-level measurements of PM_{2.5} sulfate, nitrate, OC and EC across in our study domain in 2005–2015 were summarized in Table 1. The annual mean concentrations of all four PM_{2.5} constituents decreased gradually during the study period based on ground measurements of all sites. From 2005 to 2015, the annual mean concentrations decreased from 2.69 $\mu\text{g}/\text{m}^3$ to 0.96 $\mu\text{g}/\text{m}^3$ for sulfate, from 1.08 $\mu\text{g}/\text{m}^3$ to 0.66 $\mu\text{g}/\text{m}^3$ for nitrate, from 1.69 $\mu\text{g}/\text{m}^3$ to 1.47 $\mu\text{g}/\text{m}^3$ for OC and from 0.57 $\mu\text{g}/\text{m}^3$ to 0.32 $\mu\text{g}/\text{m}^3$ for EC across the conterminous United States.

Table 1

Annual mean measurements of PM_{2.5} sulfate, nitrate, OC and EC from CSN and IMPROVE networks in conterminous United States in 2005–2015 (unit: $\mu\text{g}/\text{m}^3$).

Year	Sulfate	Nitrate	OC	EC
2005	2.69	1.08	1.69	0.57
2006	2.10	0.91	1.60	0.53
2007	2.23	1.01	1.65	0.52
2008	1.89	0.88	1.58	0.46
2009	1.55	0.84	1.39	0.38
2010	1.47	0.83	1.44	0.40
2011	1.44	0.77	1.48	0.40
2012	1.22	0.70	1.47	0.38
2013	1.17	0.77	1.37	0.35
2014	1.15	0.79	1.32	0.33
2015	0.96	0.66	1.47	0.32

3.2. Evaluation of model performance

Performance of our random forest models was evaluated by OOB R^2 at the daily level for models in each year (Table 2). Overall, the OOB R^2 values were greater than 0.7 for all four constituents, which were higher for sulfate and nitrate, and lower for OC and EC. The mean OOB R^2 values based on daily predictions during the study period were 0.86, 0.82, 0.71 and 0.75 for PM_{2.5} sulfate, nitrate, OC and EC, respectively. The mean OOB R^2 values based on annual predictions were 0.98, 0.97, 0.96 and 0.97 for sulfate, nitrate, OC and EC, respectively (shown in Table S2 in supplementary material).

The results of spatial and temporal CVs were shown in Table 3. The spatial CV R^2 values were the highest and most comparable with OOB R^2 for sulfate, slightly lower for nitrate, and were the lowest for OC and EC (~0.15 lower than OOB R^2). Fig. 4 showed that the spatial distributions were much smoother for sulfate than for OC and EC; therefore, it should be easier for models to capture the overall patterns of sulfate and predict more accurately at places without measurements, especially when the spatial lagging term of measurements and CTM-simulated concentrations were included as predictors. The temporal CV R^2 values for the four PM_{2.5} constituents were mostly comparable with

Table 2The out-of-bag (OOB) R^2 values for PM_{2.5} sulfate, nitrate, OC and EC at the daily levels over 2005–2015.

Year	Sulfate			Nitrate			OC			EC		
	R^2	Slope	Intercept	R^2	Slope	Intercept	R^2	Slope	Intercept	R^2	Slope	Intercept
2005	0.90	1.02	−0.06	0.83	1.04	−0.06	0.73	1.05	−0.11	0.75	1.05	−0.04
2006	0.89	1.02	−0.04	0.82	1.05	−0.06	0.72	1.05	−0.10	0.75	1.05	−0.04
2007	0.89	1.02	−0.05	0.81	1.05	−0.07	0.64	1.04	−0.09	0.69	1.04	−0.03
2008	0.88	1.02	−0.05	0.80	1.04	−0.05	0.67	1.05	−0.10	0.74	1.04	−0.03
2009	0.85	1.02	−0.04	0.81	1.05	−0.06	0.69	1.04	−0.08	0.75	1.05	−0.03
2010	0.85	1.02	−0.04	0.82	1.04	−0.05	0.74	1.05	−0.09	0.77	1.05	−0.03
2011	0.87	1.02	−0.04	0.82	1.04	−0.05	0.70	1.05	−0.10	0.76	1.05	−0.03
2012	0.85	1.02	−0.03	0.83	1.04	−0.04	0.68	1.06	−0.11	0.74	1.05	−0.03
2013	0.84	1.02	−0.03	0.83	1.05	−0.05	0.76	1.05	−0.09	0.78	1.05	−0.02
2014	0.83	1.02	−0.03	0.84	1.04	−0.04	0.71	1.05	−0.09	0.74	1.05	−0.02
2015	0.82	1.02	−0.03	0.83	1.05	−0.04	0.73	1.07	−0.13	0.75	1.05	−0.02
Mean	0.86	1.02	−0.04	0.82	1.04	−0.05	0.71	1.05	−0.10	0.75	1.05	−0.03

Table 3 R^2 values of spatial and temporal cross validation for PM_{2.5} sulfate, nitrate, OC and EC during 2005–2015.

Year	Spatial CV R^2				Temporal CV R^2			
	Sulfate	Nitrate	OC	EC	Sulfate	Nitrate	OC	EC
2005	0.90	0.77	0.58	0.57	0.89	0.81	0.72	0.74
2006	0.89	0.75	0.58	0.58	0.88	0.80	0.71	0.73
2007	0.88	0.73	0.55	0.56	0.88	0.79	0.63	0.68
2008	0.87	0.73	0.57	0.61	0.86	0.78	0.65	0.73
2009	0.83	0.72	0.57	0.61	0.83	0.79	0.68	0.74
2010	0.83	0.76	0.62	0.62	0.83	0.80	0.72	0.75
2011	0.85	0.74	0.58	0.62	0.85	0.80	0.69	0.75
2012	0.83	0.76	0.57	0.59	0.84	0.81	0.65	0.73
2013	0.82	0.76	0.64	0.65	0.82	0.82	0.75	0.77
2014	0.81	0.78	0.60	0.61	0.82	0.82	0.69	0.73
2015	0.80	0.78	0.63	0.61	0.81	0.82	0.71	0.74
Mean	0.85	0.75	0.59	0.60	0.85	0.80	0.69	0.74

their OOB R^2 values during the study period, indicating that our models were stable and efficient in capturing the temporal trends of PM_{2.5} constituents.

Fig. S1 shows the linear regressions between ground observations (y-axis) and averaged OOB predictions of PM_{2.5} constituents (x-axis) in 2005, 2010 and 2015 as examples. Predicted concentrations were highly consistent with observations as most of points were distributed along the 1:1 line with regression intercepts closed to 0 and slopes closed to 1 (Table 2). However, for a few points with high measured concentrations, our models tended to underestimate the concentrations of PM_{2.5} constituents.

The mean NME values based on daily predictions were 21% for sulfate, and 33%, 29%, and 32% for nitrate, OC and EC, respectively (Table S3). The NME based on month-averaged and annual-averaged predictions were significantly smaller for all four PM_{2.5} constituents. The mean NME values were 10%, 16%, 15% and 16% for sulfate, nitrate, OC and EC at the monthly level, respectively; and 5%, 9%, 8%, and 8% for sulfate, nitrate, OC and EC at the annual level, respectively (Table S3). The high precision and low uncertainty of sulfate measurement might partly contribute to the low predicting error of sulfate (Hyslop and White, 2008).

3.3. Importance ranking of predictor variables

The importance rankings of predictor variables included in the random forest models varied slightly for the same PM_{2.5} constituent in different years, but differed substantially among different PM_{2.5} constituents. As an example, importance rankings of all predictor variables used in model development for PM_{2.5} sulfate, nitrate, OC and EC in 2010 are shown in Fig. 2. Variables in top ranks for all four PM_{2.5}

constituents in common were spatial lagging terms of ground observations and GEOS-Chem simulated concentrations. For sulfate, variables in top ranks also included predictors reflecting effects of regional and surrounding environment on plume dispersion, such as spatial indicators (x and y coordinate of monitoring sites) and meteorological variables. For nitrate, besides spatial indicators and meteorological variables, top ranking variables included population and impervious land surface, which are indicators of urban areas and anthropogenic activities. At the same time, spatial lagging term of cultivated land, which indicates the agricultural sources of nitrogen, also ranked high. The importance rankings of variables were similar for OC and EC that predictors of local sources had higher ranks, such as impervious land surface, population density and length of major roads, which reflect influences of anthropogenic activities and traffic emissions on OC and EC levels. Additionally, forest land ranked higher in OC and EC lists than that in sulfate and nitrate lists, indicating potential effects of wild fires and biogenic secondary organic aerosols on OC and EC concentrations.

3.4. Long-term temporal trends

Long-term temporal trends of PM_{2.5} sulfate, nitrate, OC and EC levels based on predictions nationwide in 2005–2015 were shown in Fig. 3. PM_{2.5} sulfate and EC levels decreased substantially during the study period. Annual mean concentrations of PM_{2.5} sulfate and EC in 2015 were only 40% and 60% approximately as high as those in 2005. Concentrations of PM_{2.5} nitrate and OC decreased but fluctuated during 2005–2015. For example, OC concentrations have decreased since 2005 but rebounded in 2011, 2012 and 2015. Measurements and predictions displayed similar long-term trends for the four PM_{2.5} constituents (Fig. S2 in supplementary material).

3.5. Spatial distributions

PM_{2.5} sulfate, nitrate, OC and EC showed their featured characteristics of spatial variability in Fig. 4. Sulfate concentrations distributed more smoothly than other three constituents. The concentrations of PM_{2.5} constituents were significantly higher and decreased more sharply in the eastern United States than those in the West.

High sulfate concentrations were also seen along southwest coastline, especially in Metro Los Angeles area. States located in the Midwest area (Michigan, Ohio, Indiana, Illinois, Wisconsin, and Iowa), Metro New York area, as well as Central Valley and Metro Los Angeles in California experienced apparently high nitrate concentrations than other areas. OC concentrations were higher in the Southeast, Northwest and California. EC concentrations were generally higher in the East than in the West. Both OC and EC had localized higher concentrations around urban centers.

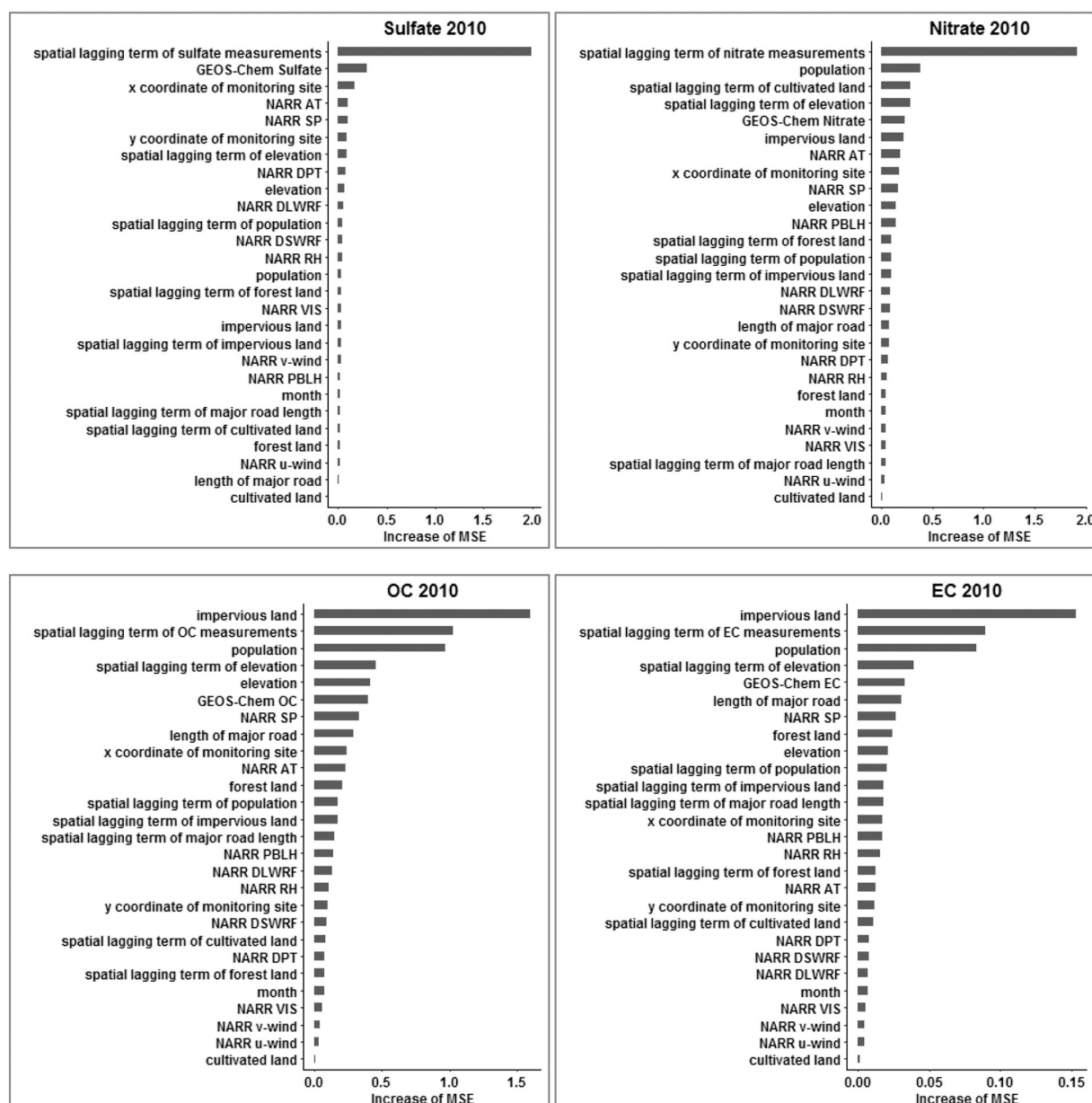


Fig. 2. Importance rankings of all variables based on mean square errors (MSE) for PM_{2.5} sulfate, nitrate, organic carbon (OC) and elemental carbon (EC) in 2010.

3.6. Sensitive analysis

It can be seen that the spatial lagging terms of ground measurements were significant predictors for all four PM_{2.5} constituents from Fig. 2. Therefore, we did a sensitivity analysis to test the stability of our random forest models by removing the spatial lagging terms of measurements, which were only available 1-in-3 days and limited the models' capacity to produce predictions every day. Table S4 shows that OOB R² values of models without spatial lagging terms of measurements only slightly decreased by 0.02–0.09 for the four constituents, indicating that our simplified models could produce every-day predictions with relatively high accuracy.

4. Discussion

In this study, our random forest models achieved relatively high predicting accuracy and the mean OOB R² values were 0.86, 0.82, 0.71 and 0.75 for predicting concentrations of PM_{2.5} sulfate, nitrate, OC and EC at the daily level over 2005–2015. The prediction errors were much

lower at the monthly and annual levels comparing with those at the daily level (Table S3). The long-term temporal trends of predictions matched well with ground measurements, and the spatial distributions of predictions captured the spatial variabilities and hotspots in areas with significant high concentrations. The prediction maps reflected the effects of potential emission sources on concentrations of PM_{2.5} sulfate, nitrate, OC and EC with the help of importance rankings of random forest.

4.1. Strengths of random forest models

We chose random forest method for modeling because it has several strengths. Firstly, comparing with parametric statistical regression models, the nonparametric algorithm is powerful in capturing complex and non-linear relationships between PM_{2.5} and predictors, such as related geographic variables (Brokamp et al., 2017; Liu et al., 2018). Secondly, comparing with other machine learning methods such as neural network, random forest is able to provide importance estimates of predictor variables to help understand the “black-box” results of

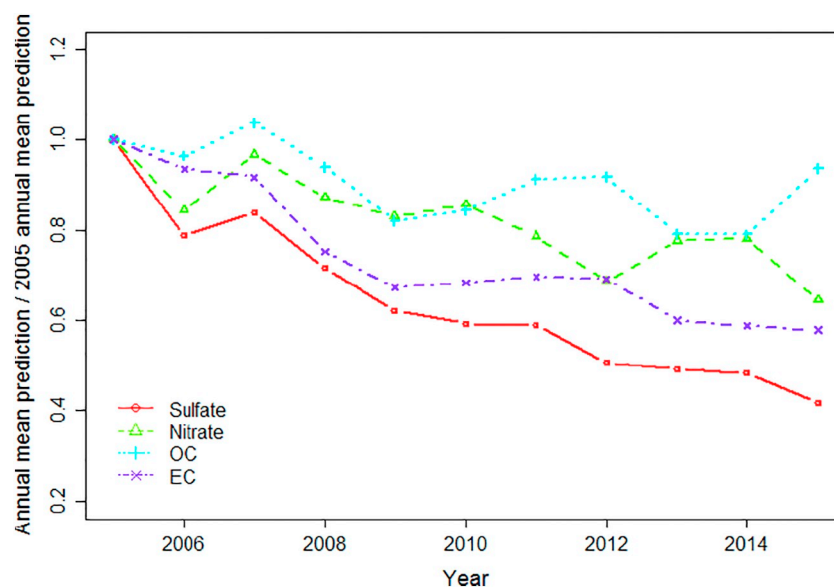


Fig. 3. Temporal trends of national wide annual mean of predicted concentrations for PM_{2.5} sulfate, nitrate, organic carbon (OC) and elemental carbon (EC) from 2005 to 2015. X-axis is the year index, and y-axis is the ratio of annual mean predictions of the year divided by annual mean predictions of 2005.

random forest (Breiman, 2001). In our study, the importance rankings are different for four PM_{2.5} constituents and reflect their featured characteristics and emission sources. For example, spatial lagging terms of measurements ranked higher for sulfate and nitrate than for OC and EC, since sulfate and nitrate are secondary aerosols that are smoother in spatial and temporal distributions. Cultivated land cover ranked higher for nitrate, indicating that agriculture is one of the emission sources of nitrate precursor and could be part of a management plan to decrease particulate nitrate and PM_{2.5} levels; meanwhile, predictor variables of local sources related to anthropogenic activities and traffic emissions, including impervious land, population and length of major roads, had higher ranks for OC and EC, suggesting that mobile sources and combustion from anthropogenic activities are important contributors to OC and EC levels in urban areas. Additionally, comparing with standard decision trees that each node is split using the best split among all variables, each node in a random forest is split using the best among m_{try} of predictors randomly chosen at that node, which saves time and is proven to perform well and be robust against overfitting (Breiman, 2001). Finally, random forest produces out-of-sample predictions while developing models, which can be used to evaluate model performance directly and the results are similar with standard 10-fold CV, which could save plenty of computing time.

4.2. Comparison of model performance with similar studies

Our national random forest models performed comparably with a study using backward propagation neural network in the northeastern United States (Di et al., 2016b), which R^2 values were 0.85 for sulfate, 0.83 for nitrate, 0.69 for OC and 0.71 for EC at the daily level. Our OOB R^2 values were higher than previous national models at either the daily or the annual level, especially for predicting OC and EC concentrations (Bergen et al., 2013; Liu et al., 2007a; Liu et al., 2007b; Liu et al., 2009b; Philip et al., 2014). For example, although MISR fractional AODs could provide rich information about aerosol shape, size and extinction, the model adjusted R^2 values based on daily predictions were 0.62, 0.13 and 0.19 for sulfate, nitrate and OC in the eastern United States, respectively, and 0.40, 0.54 and 0.28 for sulfate, nitrate and OC in the western United States, respectively (Liu et al., 2007b). The Pearson correlation coefficients (r) of AOD-scaled simulated concentrations from GEOS-Chem and measured concentrations of PM_{2.5} constituents at the daily level across North America among 2005–2008

were 0.95 and 0.68 for sulfate and nitrate, respectively, but only 0.56 and 0.45 for black carbon and organic mass, respectively (Philip et al., 2014). The R^2 values of national models with universal kriging method in the Multi-Ethnic Study of Atherosclerosis (MESA) study based on annual pollutant levels were 0.69, 0.82 and 0.96 for OC, EC, and sulfur, respectively (Bergen et al., 2013); while the mean OOB CV R^2 values based on annual concentrations in our study were 0.96, 0.97 and 0.98 for OC, EC and sulfate, respectively.

4.3. Temporal and spatial trends of PM_{2.5} sulfate, nitrate, OC and EC

Our random forest models captured not only long-term temporal trends of the four PM_{2.5} constituents, but also their spatial patterns in both national and local scales. Predictions of the four PM_{2.5} constituents were highly consistent with their ground measurements in long-term temporal trends in that PM_{2.5} sulfate, nitrate, OC and EC concentrations all decreased during the past decade. Meanwhile, these four PM_{2.5} constituents showed east-west gradient pattern with enhancement in California. The pattern is quite similar with distribution of coal power plants (<https://www.eia.gov/state/maps.php>), which is one of main emission sources of SO₂ – the precursor of sulfate and NO_x – the precursor of nitrate. Additionally, the eastern United States also has higher city density and population density, which are highly associated with emissions from anthropogenic sources.

Sulfate, as one main secondary component of PM_{2.5}, its concentrations decreased from east to west smoothly but with an enhancement in California, especially around Los Angeles area. The high sulfate concentrations may reflect the regional influence from power plants (Hand et al., 2012b) and from ship emission especially around several major ports located along the coastline of Southern California (Vutukuru and Dabdub, 2008). In particular, the Ports of Los Angeles and Long Beach comprise the largest port complex and handle one-fourth of all container cargo traffic in the United States (<http://www.dot.ca.gov/>); on the other hand, marine fuels contain high levels of sulfur (Vutukuru and Dabdub, 2008). Therefore, the high concentrations of sulfate around Los Angeles area might indicate the contribution of ship emissions on sulfate pollution in that area. The sulfate concentrations were higher and decreased more rapidly in East where SO₂ emissions from power plants were also high (Hand et al., 2012a), showing the efficiency of improving air quality by controlling the power plant emissions. At the same time, the significant decrease of sulfate in southern California may

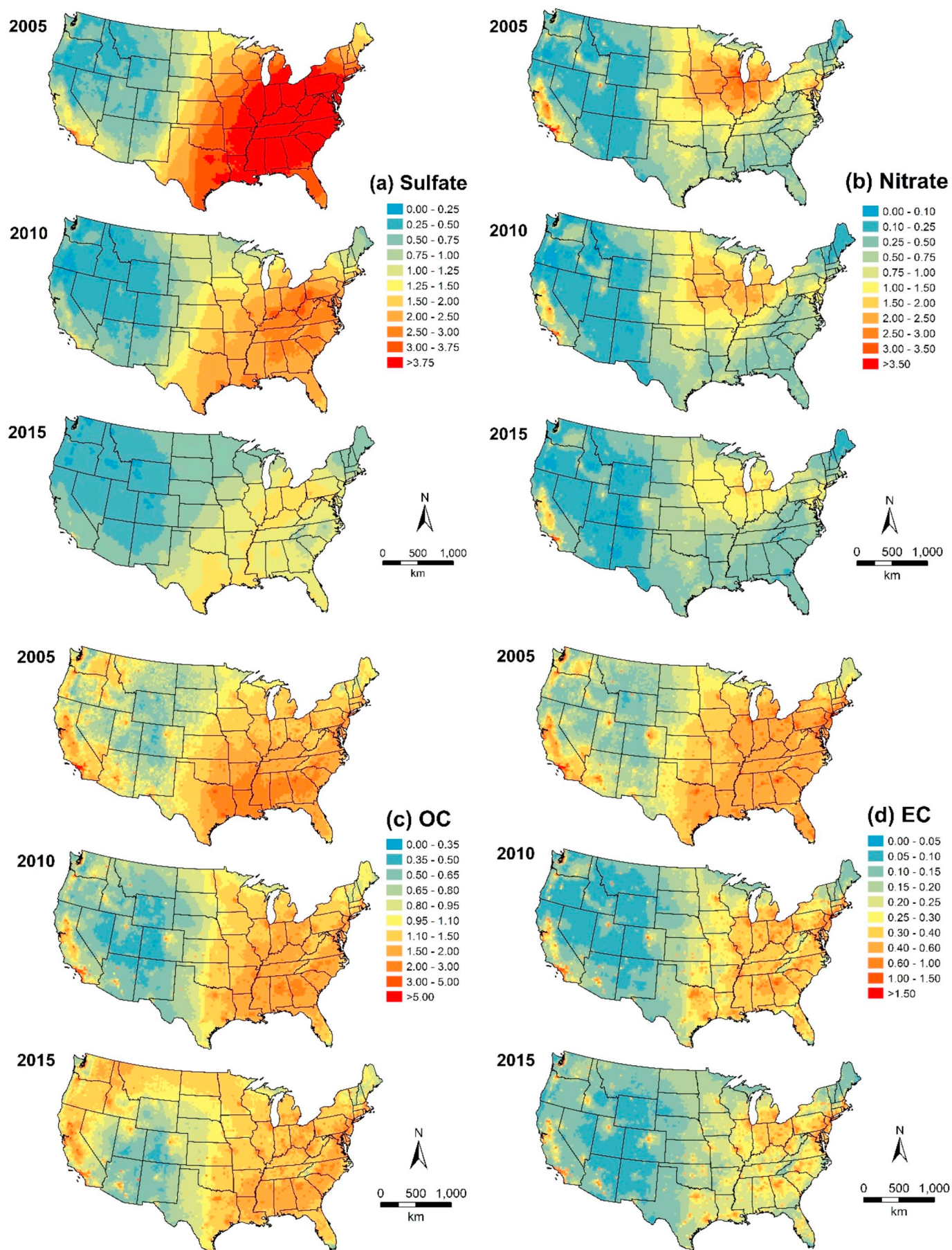


Fig. 4. Annual mean predictions of $PM_{2.5}$ (a) sulfate, (b) nitrate, (c) OC and (d) EC across conterminous United States in 2005, 2010, 2015 (unit: $\mu g/m^3$).

also attribute to the sulfur fuel regulations off the coasts since 2008 (<https://www.arb.ca.gov/ports/marinevess/ogv.htm>).

Nitrate also decreased gradually during 2005–2015. Higher concentrations were also shown in upper and middle Midwest regions as well as central valley area of California, where the area is mostly covered by cultivated land. The agriculture is one of the main emission sources of ammonia that can neutralize HNO_3 to form particulate nitrate. Meanwhile, some urban areas, such as New York-Philadelphia-Washington D.C. areas, Great Lakes areas, and Los Angeles areas, also had higher nitrate levels. The spatial pattern of nitrate was consistent with the importance ranking that population density, cultivated land and impervious land had higher ranks.

National averaged predicted concentrations of $\text{PM}_{2.5}$ carbonaceous components decreased during 2005–2015 whereas national averaged OC concentrations rebounded in year 2011, 2012 and 2015. The national averaged EC level decreased substantially but OC decreased slightly during 2005–2015. We calculated the primary OC (POC) and secondary OC (SOC) by the EC tracer method based on our predictions and GEOS-Chem simulated POC and EC concentrations (details of this method was provided in Supplementary Text 4) (Yu et al., 2004), and found that POC tended to decrease while SOC tended to increase (Fig. S3) during the study period. Zhao et al. (2017) found that SOA formation from gasoline vehicles increased with higher ratio of non-methane organic gas to oxides of nitrogen (Zhao et al., 2017). Therefore, the stricter emission standards of gasoline vehicles might lead to higher SOC while decreased NO_x level. On the other hand, according to data released by the National Interagency Fire Center (NIFC), the top five worst years for wildfires (2015, 2006, 2007, 2012, and 2011) have all occurred in the most recent decade regarding both total acres burned and acres burned per fire (Fig. S4, data was downloaded from <https://www.ncdc.noaa.gov/sotc/fire/201513>), which is consistent with the temporal trend of OC concentrations during our study period. Especially in 2015, wildfires burned over 10 million acres of land mostly in western United States, which potentially explained the high OC levels in Washington, California, Oregon and Idaho. Predictions of $\text{PM}_{2.5}$ OC showed both regional and localized high concentrations while EC tended to have significantly higher concentrations localized around urban centers, since EC and OC are indicators of primary emissions from combustion in addition that OC is also from secondary organic aerosols formed in the atmosphere. The regional high concentrations of OC in the West and Southeast were likely associated with biomass burning emissions and biogenic emissions (Hand et al., 2014). Hotspots of $\text{PM}_{2.5}$ carbonaceous components in urban areas, especially in urban centers, suggested that $\text{PM}_{2.5}$ carbonaceous levels are highly related with anthropogenic activities.

4.4. Perspectives for future applications

The high spatial CV R^2 values indicate that our national models could predict the concentrations of $\text{PM}_{2.5}$ constituents accurately in locations where ground measurements are not available in the national scale. The predictions with full spatial coverage could provide additional features of the spatial patterns and variability of $\text{PM}_{2.5}$ constituents across the nation, especially in the South, Southwest and West North Central regions where the monitoring network is quite sparse. Our prediction maps displayed similar spatial patterns to those interpolated from measurements (Hand et al., 2014). Furthermore, our models could fill the gap in areas without measurements and show more details. For example, by comparing Fig. 4 in this study with Fig. 1 in Hand et al. (2014), it can be seen that our prediction maps were able to capture hotspots where no measurements exist, and to illustrate localized high levels at finer scales than with smooth and attenuated splines. In addition, our models are able to provide daily predictions with relatively low prediction errors in a long period, and the prediction accuracy is much higher at the monthly and annual levels, which could better estimate the long-term trends.

Overall, our models might help assess concentrations of $\text{PM}_{2.5}$ constituents where and when monitoring data are not available from current monitoring networks. The full coverage predictions with relatively high accuracy could help evaluate quality of model simulations of surface concentrations of $\text{PM}_{2.5}$ constituents. The predictions may also help calculate regional net radiation budget and light extinction to identify the most responsible $\text{PM}_{2.5}$ components for atmospheric warming effects and visibility degradation (Khan et al., 2010; Malm et al., 1994). Targeted and effective regulations of such $\text{PM}_{2.5}$ components and their emission sources could be mutually beneficial to improve air quality and slow down global warming. Gridded predictions of $\text{PM}_{2.5}$ mass concentrations have been used in epidemiological studies worldwide to provide exposure assessment in areas without ground measurements (Cohen et al., 2017; Di et al., 2017; Liu et al., 2016; Xiao et al., 2018). However, the application of gridded predictions of $\text{PM}_{2.5}$ constituents is still in its early stage, and it could be a promising method of exposure assessment for epidemiological studies on the toxicity and health impacts of $\text{PM}_{2.5}$ constituents. Our models could predict $\text{PM}_{2.5}$ constituents with relatively low prediction errors compared to CTM models, and are of full spatial coverage at the national scale compared to ground measurements. Given the relatively coarse spatial resolution, the current predictions are not appropriate to assess individual exposure levels at residential addresses due to potential exposure misclassification. However, the current predictions are still useful for health studies conducted in large spatial scales and focusing on health data integrated at the county, city or state levels, such as studies similar to the national study with time series method to analyze the associations between PM and mortality risk at the city level (Dai et al., 2014). Our long-term predictions showed the changes in $\text{PM}_{2.5}$ constituents, which could help evaluate the health benefits and changes of PM toxicity caused by environmental policies of controlling $\text{PM}_{2.5}$. For example, the recent study conducted in New York State found that although hospitalization rates and total $\text{PM}_{2.5}$ mass concentrations both decreased after the implementation of environmental policies; however, comparing with the period of 2005–2013, the hospitalization rates of ischemic heart disease and myocardial infarction associated with increased $\text{PM}_{2.5}$ was higher in period of 2014–2016, when policies of improving air quality and an economic recession occurred (Zhang et al., 2018). Their further analysis suggested that the changes in PM constituents might change the toxicity of $\text{PM}_{2.5}$, and the SOA might be a particularly important trigger of acute cardiovascular events (Zhang et al., 2018). This study also provided a clue that improving the accuracy of estimating the concentrations of POA and SOA based on our current results might be an important direction in the future.

There are two limitations of our models. Firstly, the spatial resolution of current prediction maps is $0.25^\circ \times 0.3125^\circ$, which captured the spatial variability of $\text{PM}_{2.5}$ constituents in inter-city/county scales and in areas without measurements but was coarse to predict fine spatial gradients at intra-city/county level. The relatively coarse resolution limits the capacity of our current predictions to provide accurate exposure assessments for individuals in health studies based on residential addresses. Simulations with higher spatial resolutions may help illustrate finer spatial variabilities of $\text{PM}_{2.5}$ constituents and produce more accurate estimations of the health effects related with $\text{PM}_{2.5}$ constituents. For example, Pungert and West (2013) found that mortality risk tended to be underestimated based on exposure assessments of coarser grid resolutions than that based on exposure assessment of finer grid resolution, and the underestimations were less for secondary components (sulfate, nitrate, and OC) and more for elemental carbon (Pungert and West, 2013). However, most predictor variables included in our modeling dataset are able to support predictions with higher spatial resolutions, e.g., the original land use parameters are at 30 m resolution; and meteorological parameters and spatial lagging terms can be interpolated to point level, which enables our modeling strategy to produce high-resolution predictions when needed. Secondly, correlations between GEOS-Chem simulated concentrations of $\text{PM}_{2.5}$

constituents and ground measurements were relatively low. Simulations with higher accuracy are expected to improve the performance of our models. Whether and how the simulations of PM_{2.5} constituents from other CTMs affect our model performance could be evaluated in future when more data are available.

5. Conclusions

In summary, our models based on random forest algorithm predicted concentrations of PM_{2.5} sulfate, nitrate, OC and EC with relatively high accuracy and stability in the national scale and in a long temporal period. The predictions matched well with measurements in term of long-term trends of PM_{2.5} sulfate, nitrate, OC and EC in 2005–2015 in conterminous United States and captured both regional patterns and localized hotspots of the PM_{2.5} constituents. The importance rankings of predictor variables from random forest models made the results interpretable and indicated the effects of potential emission sources on concentrations of PM_{2.5} constituents. Hence, the national models developed in this study could provide supplementary evaluations of spatio-temporal distributions of PM_{2.5} constituents with full time-space coverages, which could be beneficial to assess the impacts of PM_{2.5} constituents on radiation budgets and visibility degradation, and support regional to national health studies at county or city levels to understand the acute and chronic toxicity and health impacts of PM_{2.5} constituents, and consequently provide scientific evidence for making targeted and effective regulations of PM_{2.5} pollution.

Acknowledgment

The work of X. Meng and Y. Liu is partially supported by the MISRC science team at the JPL, California Institute of Technology, led by D. Diner (subcontract 1588347), and by the NASA Applied Sciences Program, managed by J. Haynes (grant no. NNX16AQ28G, PI: Y. Liu).

IMPROVE is a collaborative association of state, tribal, and federal agencies, and international partners. US Environmental Protection Agency is the primary funding source, with contracting and research support from the National Park Service. The Air Quality Group at the University of California, Davis is the central analytical laboratory, with ion analysis provided by Research Triangle Institute, and carbon analysis provided by Desert Research Institute.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.envint.2018.10.029>.

References

- Achilleos, S., Kioumourtoglou, M.A., Wu, C.D., Schwartz, J.D., Koutrakis, P., Papatheodorou, S.I., 2017. Acute effects of fine particulate matter constituents on mortality: a systematic review and meta-regression analysis. *Environ. Int.* 109, 89–100.
- Baek, J., Hu, Y.T., Odman, M.T., Russell, A.G., 2011. Modeling secondary organic aerosol in cmaq using multigenerational oxidation of semi-volatile organic compounds. *J. Geophys. Res.-Atmos.* 116.
- Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z.J., Weinmayr, G., Hoffmann, B., et al., 2014. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre escape project. *Lancet* 383, 785–795.
- Bergen, S., Sheppard, L., Sampson, P.D., Kim, S.Y., Richards, M., Vedal, S., et al., 2013. A national prediction model for pm2.5 component exposures and measurement error-corrected health effect inference. *Environ. Health Perspect.* 121, 1017–1025.
- Brauer, M., Amann, M., Burnett, R.T., Cohen, A., Dentener, F., Ezzati, M., et al., 2012. Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. *Environ. Sci. Technol.* 46, 652–660.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32.
- Brokamp, C., Jandarov, R., Rao, M.B., LeMasters, G., Ryan, P., 2017. Exposure assessment models for elemental components of particulate matter in an urban environment: a comparison of regression and random forest approaches. *Atmos. Environ.* 151, 1–11.
- Chen, R.J., Yin, P., Meng, X., Liu, C., Wang, L.J., Xu, X.H., et al., 2017. Fine particulate air pollution and daily mortality a nationwide analysis in 272 Chinese cities. *Am. J. Respir. Crit. Care Med.* 196, 73–81.
- Chung, Y., Dominici, F., Wang, Y., Coull, B.A., Bell, M.L., 2015. Associations between long-term exposure to chemical constituents of fine particulate matter (pm2.5) and mortality in medicare enrollees in the eastern United States. *Environ. Health Perspect.* 123, 467–474.
- Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., et al., 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. *Lancet* 389, 1907–1918.
- Dai, L.Z., Zanobetti, A., Koutrakis, P., Schwartz, J.D., 2014. Associations of fine particulate matter species with mortality in the United States: a multicity time-series analysis. *Environ. Health Perspect.* 122, 837–842.
- Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y.J., Schwartz, J., 2016a. Assessing pm2.5 exposures with high spatiotemporal resolution across the continental United States. *Environ. Sci. Technol.* 50, 4712–4721.
- Di, Q., Koutrakis, P., Schwartz, J., 2016b. A hybrid prediction model for pm2.5 mass and components using a chemical transport model and land use regression. *Atmos. Environ.* 131, 390–399.
- Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., et al., 2017. Air pollution and mortality in the medicare population. *N. Engl. J. Med.* 376, 2513–2522.
- EPA US, 2008. Integrated Review Plan for the National Ambient Air Quality Standards for Particulate Matter. Research Triangle Park, NC.
- EPA US, 2009. Integrated Science Assessment for Particulate Matter. U.S. Environmental Protection Agency, Washington, DC.
- Fuzzi, S., Baltensperger, U., Carslaw, K., Decesari, S., van der Gon, H.D., Facchini, M.C., et al., 2015. Particulate matter, air quality and climate: lessons learned and future needs. *Atmos. Chem. Phys.* 15, 8217–8299.
- Gromping, U., 2009. Variable importance assessment in regression: linear regression versus random forest. *Am. Stat.* 63, 308–319.
- Hand, J.L., Copeland, S.A., Day, D.E., Dillner, A.N., Indresand, H., Malm, W.C., McDade, C.E., Moore, C.T., Pitchford, M.L., Schichtel, B.A., Watson, J.G., 2011. Spatial and Seasonal Patterns and Temporal Variability of Haze and Its Constituents in the United States: Report v June 2011. IMPROVE Reports.
- Hand, J.L., Schichtel, B.A., Malm, W.C., Pitchford, M.L., 2012a. Particulate sulfate ion concentration and so2 emission trends in the United States from the early 1990s through 2010. *Atmos. Chem. Phys.* 12, 10353–10365.
- Hand, J.L., Schichtel, B.A., Pitchford, M., Malm, W.C., Frank, N.H., 2012b. Seasonal composition of remote and urban fine particulate matter in the United States. *J. Geophys. Res.-Atmos.* 117.
- Hand, J.L., Schichtel, B.A., Malm, W.C., Pitchford, M., Frank, N.H., 2014. Spatial and seasonal patterns in urban influence on regional concentrations of speciated aerosols across the United States. *J. Geophys. Res.-Atmos.* 119, 12832–12849.
- Hu, X.F., Waller, L.A., Lyapustin, A., Wang, Y.J., Al-Hamdan, M.Z., Crosson, W.L., et al., 2014. Estimating ground-level pm2.5 concentrations in the southeastern United States using maia aod retrievals and a two-stage model. *Remote Sens. Environ.* 140, 220–232.
- Hu, X.F., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., et al., 2017. Estimating pm2.5 concentrations in the conterminous United States using the random forest approach. *Environ. Sci. Technol.* 51, 6936–6944.
- Hyslop, N.P., White, W.H., 2008. An evaluation of interagency monitoring of protected visual environments (improve) collocated precision and uncertainty estimates. *Atmos. Environ.* 42, 2691–2705.
- IPCC, 2013. IPCC: Fifth Assessment Report: Climate Change 2013: The Physical Science Basis, Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.
- Kaufman, J.D., Adar, S.D., Allen, R.W., Barr, R.G., Budoff, M.J., Burke, G.L., et al., 2012. Prospective study of particulate air pollution exposures, subclinical atherosclerosis, and clinical cardiovascular disease the multi-ethnic study of atherosclerosis and air pollution (mesa air). *Am. J. Epidemiol.* 176, 825–837.
- Khan, A.J., Li, J.J., Dutkiewicz, V.A., Husain, L., 2010. Elemental carbon and sulfate aerosols over a rural mountain site in the northeastern United States: regional emissions and implications for climate change. *Atmos. Environ.* 44, 2364–2371.
- Lepeule, J., Laden, F., Dockery, D., Schwartz, J., 2012. Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard six cities study from 1974 to 2009. *Environ. Health Perspect.* 120, 965–970.
- Li, C., Martin, R.V., van Donkelaar, A., Boys, B.L., Hammer, M.S., Xu, J.W., et al., 2017. Trends in chemical composition of global and regional population-weighted fine particulate matter estimated for 25 years. *Environ. Sci. Technol.* 51, 11185–11195.
- Liaw, A., 2015. Package 'randomforest'.
- Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. In: *R News*. 2/3, pp. 18–22.
- Liu, Y., Koutrakis, P., Kahn, R., 2007a. Estimating fine particulate matter component concentrations and size distributions using satellite-retrieved fractional aerosol optical depth: part 1 - method development. *J. Air Waste Manage. Assoc.* 57, 1351–1359.
- Liu, Y., Koutrakis, P., Kahn, R., Turquet, S., Yantosca, R.M., 2007b. Estimating fine particulate matter component concentrations and size distributions using satellite-retrieved fractional aerosol optical depth: Part 2 - a case study. *J. Air Waste Manage. Assoc.* 57, 1360–1369.
- Liu, Y., Paciorek, C.J., Koutrakis, P., 2009a. Estimating regional spatial and temporal variability of pm2.5 concentrations using satellite data, meteorology, and land use information. *Environ. Health Perspect.* 117, 886–892.
- Liu, Y., Schichtel, B.A., Koutrakis, P., 2009b. Estimating particle sulfate concentrations using misr retrieved aerosol properties. *IEEE J-Stars* 2, 176–184.
- Liu, C., Yang, C.Y., Zhao, Y.H., Ma, Z.W., Bi, J., Liu, Y., et al., 2016. Associations between

- long-term exposure to ambient particulate air pollution and type 2 diabetes prevalence, blood glucose and glycosylated hemoglobin levels in China. *Environ. Int.* 92–93, 416–421.
- Liu, Y., Cao, G., Zhao, N., Mulligan, K., Ye, X., 2018. Improve ground-level pm 2.5 concentration mapping using a random forests-based geostatistical approach. *Environ. Pollut.* 235, 272–282.
- Malm, W.C., Sisler, J.F., Huffman, D., Eldred, R.A., Cahill, T.A., 1994. Spatial and seasonal trends in particle concentration and optical extinction in the United-States. *J. Geophys. Res.-Atmos.* 99, 1347–1370.
- Malm, W.C., Schichtel, B.A., Pitchford, M.L., 2011. Uncertainties in pm2.5 gravimetric and speciation measurements and what we can learn from them. *J. Air Waste Manage. Assoc.* 61, 1131–1149.
- Meng, X., Garay, M.J., Diner, D.J., Kalashnikova, O.V., Xu, J., Liu, Y., 2018. Estimating pm2.5 speciation concentrations using prototype 4.4 km-resolution misr aerosol properties over southern California. *Atmos. Environ.* 181, 70–81.
- Philip, S., Martin, R.V., van Donkelaar, A., Lo, J.W.H., Wang, Y.X., Chen, D., et al., 2014. Global chemical composition of ambient fine particulate matter for exposure assessment. *Environ. Sci. Technol.* 48, 13060–13068.
- Pope, C.A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: lines that connect. *J. Air Waste Manage. Assoc.* 56, 709–742.
- Punger, E.M., West, J.J., 2013. The effect of grid resolution on estimates of the burden of ozone and fine particulate matter on premature mortality in the USA. *Air Qual. Atmos. Health* 6, 563–573.
- Rohr, A.C., Wyzga, R.E., 2012. Attributing health effects to individual particulate matter constituents. *Atmos. Environ.* 62, 130–152.
- Solomon, P.A., Crumpler, D., Flanagan, J.B., Jayanty, R.K.M., Rickman, E.E., McDade, C.E., 2014. Us national pm2.5 chemical speciation monitoring networks-csn and improve: description of networks. *J. Air Waste Manage. Assoc.* 64, 1410–1438.
- Vutukuru, S., Dabdub, D., 2008. Modeling the effects of ship emissions on coastal air quality: a case study of southern California. *Atmos. Environ.* 42, 3751–3764.
- Wang, M., Beelen, R., Stafoggia, M., Raaschou-Nielsen, O., Andersen, Z.J., Hoffmann, B., et al., 2014. Long-term exposure to elemental constituents of particulate matter and cardiovascular mortality in 19 European cohorts: results from the escape and transphorm projects. *Environ. Int.* 66, 97–106.
- Xiao, Q.Y., Chen, H.Y., Strickland, M.J., Kan, H.D., Chang, H.H., Klein, M., et al., 2018. Associations between birth outcomes and maternal pm2.5 exposure in shanghai: a comparison of three exposure assessment approaches. *Environ. Int.* 117, 226–236.
- Yu, S.C., Dennis, R.L., Bhavsar, P.V., Eder, B.K., 2004. Primary and secondary organic aerosols over the United States: estimates on the basis of observed organic carbon (oc) and elemental carbon (ec), and air quality modeled primary oc/ec ratios. *Atmos. Environ.* 38, 5257–5268.
- Zhang, W., Lin, S., Hopke, P.K., Thurston, S.W., van Wijngaarden, E., Croft, D., et al., 2018. Triggering of cardiovascular hospital admissions by fine particle concentrations in New York state: before, during, and after implementation of multiple environmental policies and a recession. *Environ. Pollut.* 242, 1404–1416.
- Zhao, Y.L., Saleh, R., Saliba, G., Presto, A.A., Gordon, T.D., Drozd, G.T., et al., 2017. Reducing secondary organic aerosol formation from gasoline vehicle exhaust. *Proc. Natl. Acad. Sci. U. S. A.* 114, 6984–6989.