Random forest models for PM$_{2.5}$ speciation concentrations using MISR fractional AODs

Guannan Geng$^{1,2}$, Xia Meng$^{1,3}$, Kebin He$^1$ and Yang Liu$^1$

1 Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322, United States of America
2 State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, People’s Republic of China
3 Department of Environmental Health, School of Public Health, Fudan University, Shanghai 200032, People’s Republic of China

E-mail: yang.liu@emory.edu

Keywords: MISR, random forest, fine particulate matter speciation, exposure assessment, satellite remote sensing

Supplementary material for this article is available online

Abstract

It is increasingly recognized that various chemical components of PM$_{2.5}$ might have differential toxicities to human health, although such studies are hindered by the sparse or non-existent coverage of ground PM$_{2.5}$ speciation monitors. The Multi-angle Imaging SpectroRadiometer (MISR) onboard the Terra satellite has an innovative design to provide information about aerosol shape, size and extinction that are more related to PM$_{2.5}$ speciation concentrations. In this study, we developed random forest models that incorporated ground measurements of PM$_{2.5}$ species, MISR fractional AODs, simulated PM$_{2.5}$ speciation concentrations from a chemical transport model (CTM), land use variables and meteorological fields, to predict ground-level daily PM$_{2.5}$ sulfate, nitrate, organic carbon (OC) and elemental carbon (EC) concentrations in California between 2005 and 2014. Our models had out-of-bag $R^2$ of 0.72, 0.70, 0.68 and 0.70 for sulfate, nitrate, OC and EC, respectively. We also conducted sensitivity tests to explore the influence of variable selection on model performance.

Results show that if there are sufficient ground measurements and predictor data to support the most sophisticated model structure, fractional AODs and total AOD have similar predicting power in estimating PM$_{2.5}$ species. Otherwise, models using fractional AODs outperform those with total AOD. PM$_{2.5}$ speciation concentrations are more sensitive to land use variables than other supporting data (e.g., CTM simulations and meteorological information).

1. Introduction

Numerous studies have reported the associations between adverse health effects and exposure to ambient fine particulate matter (particles with a diameter less than 2.5 $\mu$m, PM$_{2.5}$) [1–3]. It is also increasingly recognized that various chemical components of PM$_{2.5}$ might have differential toxicities to human health [4–6]. For example, combustion-related species such as elemental carbon (EC) and Potassium (K) were found to have a stronger association with mortality than other species [4]. However, information regarding the health effects of PM$_{2.5}$ components is limited, mainly hindered by the sparse or non-existent coverage of ground PM$_{2.5}$ speciation monitors. Meanwhile, the changes in PM$_{2.5}$ mass concentrations are driven by the variations of its chemical composition that are related directly to precursor emissions [7, 8], thus, understanding the spatiotemporal changes of PM$_{2.5}$ speciation also helps designing more effective air pollution mitigation measures.

In recent years, many sophisticated models have been reported in the literature to incorporate information from ground monitors, satellite remote sensing and chemical transport models (CTMs) in order to obtain improved estimates of PM$_{2.5}$ mass concentrations [9–11]. However, attempts to estimate high-resolution PM$_{2.5}$ speciation concentrations have been few. Some works [7, 12, 13] used satellite-based total aerosol optical depth (AOD) and composition information from the CTMs to estimate PM$_{2.5}$ speciation concentrations. For example, global PM$_{2.5}$ chemical...
composition were inferred by combining AOD with conversion factors between AOD and PM$_{2.5}$ species provided by the GEOS-Chem model [12]. Such a method does not rely on ground observation data; however, without calibration against observation data, the estimated results had relatively low accuracy (e.g., $r^2$ between 0.65 and 0.75 for different species in Geng et al [7]). Some studies [14, 15] utilized ground observations, meteorological fields and land use variables to calibrate and downscale the CTM-simulated results and obtained high-resolution PM$_{2.5}$ speciation data. For instance, Di et al [14] used a backward propagation neural network to calibrate GEOS-Chem simulated PM$_{2.5}$ speciation concentrations and improved their spatial resolution from 0.50° × 0.66° to 1 km by incorporating land use variables. However, temporal variabilities in PM$_{2.5}$ species concentrations cannot be readily captured by land use terms, indicated by significantly lower temporal $R^2$ (by 0.02–0.26) compared to total $R^2$ in Di et al [14].

The Multi-angle Imaging SpectroRadiometer (MISR) onboard the Terra satellite has an innovative design of viewing the Earth simultaneously at nine widely spaced angles [16]. The change in reflection at different view angles affords the means to distinguish different types of atmospheric particles that are more related to PM$_{2.5}$ composition [17, 18]. MISR provides fractional AODs defined by a size distribution, shape, complex index of refraction and scale height. Such data has been used in statistical models to predict PM$_{2.5}$ speciation and was proved to have better predicting power than the total AOD [19, 20]. However, those studies either used MISR AOD at a coarse resolution of 17.6 km [19] or used experimental data products that have limited spatial coverage [20–22]. Meanwhile, the model structures developed for PM$_{2.5}$ speciation concentrations before were simple and only involved a few predictors [19, 20]. For example, Meng et al [20] developed generalized additive models (GAMs) using MISR 4.4 km fractional AODs and some meteorological and geographical indicators to predict the concentrations of four PM$_{2.5}$ species. The leave-one-out-cross-validation (LOOCV) $R^2$ for different species varied from 0.49 to 0.62. The recently released operational MISR products at 4.4 km resolution make it possible to develop a high-resolution model with inputs from multiple sources and conduct comprehensive analyses to reveal the importance of fractional AODs in predicting PM$_{2.5}$ speciation concentrations.

Moreover, the availability of input predictors for the PM$_{2.5}$ speciation models varies among regions. In data-rich regions like the United States, there are established networks of PM$_{2.5}$ speciation monitors such as the Chemical Speciation Network (CSN) and the Interagency Monitoring of Protected Visual Environments network (IMPROVE) and high-quality supporting information for the model development. While in most other regions in the world, lack of access to and poor quality of input data might not support sophisticated model structures. Therefore, it is crucial to understand the impact of variable selection on developing PM$_{2.5}$ speciation models. The influence of different variable choices on model performance has been studied for PM$_{2.5}$ mass concentrations before [23], and results showed that models without land use variables lacked fine spatial details in predictions, but could still capture the temporal variability of PM$_{2.5}$. However, the results might be different for PM$_{2.5}$ species due to their different spatial and temporal characteristics.

In this work, we selected the state of California as a case to develop PM$_{2.5}$ speciation models using the newly released MISR Level 2 AOD product, CTM simulations and other supporting information. Then we conducted a series of sensitivity scenarios to examine the impacts of different variable choices on model’s performance to provide guidance for model development in data-poor regions. Finally, suggestions are made based on our results for other regions to build PM$_{2.5}$ speciation concentrations models.

2. Materials and methods

Our study domain covers the state of California with an 80 km buffer along its inland border, as shown in figure 1. We constructed a 1 km × 1 km resolution modeling grid for data integration. The study period is from 2005 to 2014.

2.1. Datasets

2.1.1. PM$_{2.5}$ speciation monitoring data

The 24 h mean concentrations of PM$_{2.5}$ sulfate, nitrate, organic carbon (OC) and EC within the study domain during 2005–2014 were downloaded from CSN (http://aqdcr1.epa.gov) and IMPROVE (http://views.cira.colostate.edu/ed). There are 55 PM$_{2.5}$ speciation monitors located in our domain and their spatial distribution is shown in figure 1. The sampling frequency is 1-in-3 or 1-in-6 d for CSN and 1-in-3 d for IMPROVE, and the total numbers of observations in each locations are shown in figure S1 (available online at stacks.iop.org/ERL/15/034056/mmedia). CSN and IMPROVE have different samplers and analytical methods, which caused systematic differences in the measurements of carbonaceous compounds in PM$_{2.5}$. Following our previous works [15, 20], we adjusted OC and EC data from CSN for compatibility with IMPROVE data. More details of the adjustment could be found in Meng et al [15]. The monitors were assigned to 1 km grid cells where they fell.

2.1.2. MISR 4.4 km aerosol product

MISR views the Earth at nine separate angles simultaneously, which provides information to distinguish different types of aerosol particles. MISR has a global coverage of every nine days, with repeat coverage between 2 and 9 d based on the latitude. A new version
(V23) of the MISR operational aerosol product has been released recently, which offers substantial improvements compared to the previous version (V22), including a higher spatial resolution of 4.4 km, improved performance evaluated against ground measurements [24], improved cloud screening and additional diagnostic fields. We downloaded the V23 Level 2 aerosol data for 2005–2014 from NASA Earthdata portal (https://search.earthdata.nasa.gov/), which contain the same 74 aerosol mixtures as in previous versions [25]. The eight fractional AOD components used to construct a total of 74 mixtures were utilized in this study and details of their descriptions could be found elsewhere [15, 18, 26, 27]. These eight fractional AOD components (i.e., component 1, 2, 3, 6, 8, 14, 19 and 21) represent different particle shape, scattering property and effective radius for a number-weighted log-normal distribution (table S1). The fractional AOD components could be obtained from any MISR aerosol observation using the following equation [18]:

$$\text{AOD}_i = \frac{\sum_{j=1}^{74} \alpha \text{AOD}_{\text{mixture } j} \times \text{Fraction}_{\text{component } i \text{ in mixture } j}}{\text{Number of successful mixtures}}$$

where $\text{AOD}_{\text{mixture } j}$ is the AOD mixture $j$, $\text{Fraction}_{\text{component } i \text{ in mixture } j}$ is the contribution of component $i$ to the AOD for mixture $j$; if mixture $j$ is retrieved successfully, then $\alpha = 1$, otherwise $\alpha = 0$.

All the AOD data were interpolated to the centroid of the 1 km grid cells using the inverse distance weighting (IDW) approach. Since AOD has missing values, we only kept data in those 1 km grid cells that fell inside the 4.4 km MISR pixels where AOD data was available.

2.1.3. CTM simulations

The CTM simulations used in this study were conducted by Community Multiscale Air Quality (CMAQ) model version 5.0.2 [28] driven by the Weather Research and Forecasting (WRF) model version v3.4. Details of the model configurations for WRF and CMAQ could be found in Zhang et al [29]. Briefly, the model was conducted over the continental United States from 2005 to 2014 at 12 km spatial resolution with 35 vertical layers from the surface to the top of the free troposphere. National Land Cover Database (NLCD) were used as inputs for the WRF model, and the simulated meteorological parameters were processed using the Meteorology-Chemistry Interface Processor version 4.1.3 [30] to create inputs for the CMAQ model. The chemical boundary conditions for the CMAQ model were taken from year-specific simulations of the global GEOS-Chem model [31]. The anthropogenic emission inputs were based on 2005, 2008 and 2011 National Emission Inventory data. The biogenic emissions were provided by the Biogenic Emissions Inventory System version 3.14 [32].

Daily mean concentrations of sulfate, nitrate, OC and EC at the surface level were extracted to represent
the ambient PM$_{2.5}$ speciation concentrations. All the concentration data were interpolated to the 1 km grid using IDW approach.

2.1.4. Meteorological fields
Temperature, wind speed and specific humidity data at ~13 km spatial resolution were obtained from the hourly North American Land Data Assimilation Systems phase 2 (NLDAS-2, http://ldas.gsfc.nasa.gov/nldas/) datasets. Planetary boundary layer height at 32 km spatial resolution was taken from the 3 hourly North American Regional Reanalysis (NARR, http://emc.ncep.noaa.gov/mmb/rreanl/) products. All meteorological data were averaged between 9 am and 12 pm over satellite overpass time and were also interpolated to the 1 km grid using IDW approach.

2.1.5. Land use variables
Elevation data at 30 m spatial resolution were extracted from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Map (GDEM) (https://asterweb.jpl.nasa.gov/gdem.asp) version 2. Road networks were obtained from ESRI StreetMap USA (Environmental Systems Research Institute, Inc., Redlands, CA). Impervious surface, forest cover, shrub cover and cultivated land cover information at 30 m spatial resolution were taken from NLCD (https://mrlc.gov) 2006 and 2011. Population data at 1 km spatial resolution were extracted from the LandScan Global Population Database (https://landsat.eio.usgs.gov/). Elevation, impervious surface, forest cover, shrub cover and cultivated land cover were averaged while population and road lengths were summed within the 1 km grid.

2.2. Modeling
Random forest models were built separately for sulfate, nitrate, OC and EC over 2005–2014 using MISR fractional AODs, CMAQ simulations, meteorological fields and land use variables. Initially proposed by Breiman [33], random forest algorithm is an ensemble learning method based on decision trees, which has the advantage of allowing both continuous and categorical input variables and are quite robust to outliers. It also provides variable importance rankings as well as out-of-bag errors for variable selection and model evaluation. It has been used in previous studies to estimate air pollutant concentrations and shown excellent performances [10, 15, 34].

In this study, four out-of-sample cross-validation (CV) experiments were considered to evaluate the prediction performance of our models. First, $R^2$ values and root mean square errors (RMSE) between out-of-bag predictions and observations were calculated. Second, a temporal CV was conducted to evaluate the model’s ability to perform temporal prediction when certain days were missing. After all the observation days were randomly divided into 10 subsets, we used data from nine subsets of the observation days to train the model and predicted PM$_{2.5}$ speciation data in the remaining subset. This process was repeated for 10 times after all the observation days were tested. We also did a spatial CV to evaluate the model’s ability to make predictions at locations without monitors. Similarly, all the ground monitors were randomly divided into 10 subsets. The model was trained using data from nine subsets of the monitors, and then used to make predictions on the remaining subset. This process was also repeated for 10 times. Finally, a spatio-clustered CV was performed following Murray et al [35] to test the model’s robustness when a large spatial group of data were missing instead of a single location. A hierarchical clustering method by proximity of monitoring locations was used to define 10 clusters of all the monitors. Then data from nine clusters was used to predict PM$_{2.5}$ speciation data in the remaining cluster and this process was also repeated for 10 times.

2.3. Sensitivity tests of variable selection
In data-rich regions, there are sufficient input dataset to support the most sophisticated model structure (e.g., the full model in our study). While for data-poor regions, either the PM$_{2.5}$ speciation monitors or other supporting data are limited, which could only allow for a reduced form of the full model developed in the data-rich regions. As indicated in previous studies, different predictors affect the model’s performance in predicting PM$_{2.5}$ mass concentrations [23]. In this study, we considered our models’ robustness to different choices of input parameters when estimating PM$_{2.5}$ speciation concentrations. Previous works [20] compared model performance when using fractional AODs and total AOD, and found that models with fractional AODs outperformed models with total AOD when other predictors were the same. Meanwhile, CMAQ simulations, meteorological fields and land use variables are also important predictors because they could provide insights into local-scale spatial/temporal variability of PM$_{2.5}$ speciation concentrations. In this study, we designed different model structures to compare the predicting power of fractional AODs, total AOD and no AOD involved, and also test different combinations of other predictors, as shown in table S2. The REF case is the full model in our study with fractional AODs and all the supporting data. The comparison between REF, M1 and M2 shows the influence of different AOD data. M3–M5, M6–M8 and M9–M11 are models with only CMAQ, only land use variables and only meteorological fields as the supporting data, respectively. M12 and M13 are two AOD only models.
3. Results and discussion

3.1. Descriptive statistics of the training dataset
On average, MISR had 323 d of observations during 2005–2014 in our study domain (figure S2), due to the low sampling frequency of MISR instrument. There were more missing AOD in cloudy and rainy winter and the mountainous regions. After integration of all the dataset, we obtained 3,543, 3,505, 3,474 and 3,474 site-days of data over 2005–2014 in the model training dataset for sulfate, nitrate, OC and EC, respectively. The amount of training dataset was mainly limited by the low sampling frequency of MISR instrument. There were more missing AOD in cloudy and rainy winter and the mountainous regions. After integration of all the dataset, we obtained 3,543, 3,505, 3,474 and 3,474 site-days of data over 2005–2014 in the model training dataset for sulfate, nitrate, OC and EC, respectively. The amount of the training dataset was mainly limited by the low sampling frequency of MISR. The 10 yr mean concentrations of sulfate, nitrate, OC and EC for California were 1.11 μg m⁻³, 0.68 μg m⁻³, 0.68 μg m⁻³, and 0.68 μg m⁻³, respectively.

3.2. Model evaluation
Table 1 and figure S3 show the CV results of our random forest models at the daily level for the entire study period of the four PM₂.₅ species, including the values of \( R^2 \) and RMSE between the CV estimates and observations. The out-of-bag predictions from the random forest models showed relatively good agreements with the observations, with out-of-bag \( R^2 \) of 0.72, 0.70, 0.68 and 0.70 for sulfate, nitrate, OC and EC, respectively. The model for sulfate had the best performance among all species, with RMSE of 0.54 μg m⁻³, while the model for OC had a lower \( R^2 \) and a RMSE of 1.30 μg m⁻³. Results also showed that our models had comparable performance in different seasons (Table S3).

The temporal CV \( R^2 \) values of the PM₂.₅ speciation models were generally comparable to their out-of-bag \( R^2 \), with slight decreases by 0.01–0.02 for different species, indicating that our models were stable and efficient in capturing the temporal variability of PM₂.₅ speciation concentrations. The spatial CV \( R^2 \) values, however, decreased by 0.08–0.14 for the four species when compared to the out-of-bag results, with values for sulfate and nitrate higher than those for OC and EC. This has also been found in previous studies, and the reason is that the spatial distribution of sulfate is much smoother than OC and EC, which is easier for the models to predict at locations where ground monitors are absent. The models’ prediction performance further decreased from spatial CV to spatial-clustered CV, because information from nearby monitors to aid spatial interpolation are missing.

### Table 1. \( R^2 \) and RMSE (μg m⁻³) values of out-of-bag predictions, temporal CV, spatial CV and spatial-clustered CV for sulfate, nitrate, OC and EC at daily level for the entire study period.

| Species | Out-of-bag | Temporal CV | Spatial CV | Spatial-clustered CV |
|---------|------------|-------------|------------|----------------------|
| Sulfate | \( R^2 \) | 0.72 | 0.70 | 0.62 | 0.52 |
| Nitrate | \( R^2 \) | 0.70 | 0.54 | 0.62 | 0.70 |
| OC      | \( R^2 \) | 0.68 | 0.68 | 0.62 | 0.62 |
| EC      | \( R^2 \) | 0.70 | 1.23 | 0.62 | 1.36 |
|         | RMSE      | 0.54 | 1.32 | 1.36 | 0.55 |
|         | RMSE      | 1.30 | 0.37 | 1.55 | 0.43 |
|         | RMSE      | 0.69 | 0.37 | 0.46 | 0.53 |

3.3. Importance ranking of model predictors
Figure 2 shows the importance rankings of all predictors used in model development for sulfate, nitrate, OC and EC. CMAQ simulations played important roles in all four models with high importance rankings. Meteorological fields, such as air temperature, was among the highest ranking predictors for the sulfate model, because it could affect the formation of sulfate via chemical reactions. The fractional AOD components, AOD2 and AOD6, were also important predictors for the estimation of sulfate concentrations. These two components both represent spherical and non-absorbing particles that are related to sulfate concentrations. For the nitrate model, top ranking variables included population and impervious surface, which are both related to anthropogenic fossil fuel combustion that could release NOₓ, precursor gases of nitrate. Shrub cover, which reflects the emission sources of soil NOₓ [36, 37], also ranked high. Variables of high importance rankings were similar for OC and EC, which were mostly land use variables that could represent the spatial patterns of local-scale emissions sources. For example, impervious surface and population density are significant indicators of anthropogenic emissions. Length of major road could reflect the emissions from on-road transportation on OC and EC concentrations. Notably, shrub cover and forest cover also ranked high for OC and EC, which might represent the contribution of wildfires to OC and EC levels in California [38, 39].

3.4. Spatial and temporal trends of PM₂.₅ speciation predictions
Figure 3 shows the spatial distributions of the predicted annual mean concentrations at 1 km spatial resolution for sulfate, nitrate, OC and EC in 2005, 2009 and 2014. Generally, the four PM₂.₅ species have similar spatial patterns, with high values concentrated in the Central Valley, the Bay area and Metro Los Angeles, which are all heavily populated regions affected by anthropogenic emissions. Sulfate concentrations had smoother patterns than other three species, and the most polluted regions were along the coastal area in Southern California, where several major ports lied and SO₂ was emitted from ship engines. Nitrate concentrations are enhanced in the Central Valley, where mobile and urban emissions are...
major NOx sources. The spatial distributions of the carbonaceous species were more heterogeneous than sulfate, with hotspots located in urban centers. OC and EC mainly come from primary emission sources and the spatial patterns were highly related to human activities. Meanwhile, there were also some high values in the northern part of California, which might be contributed by wildfires. During 2005–2014, the concentrations of the four PM2.5 species showed a decreasing trend, especially after the year 2009. Since that MISR AOD had a lot of missingness in its spatial and temporal coverage, which might affect the trends presented in our study, the temporal trends are not further discussed.

3.5. Sensitivity to variable selection

Table 2 shows the out-of-bag $R^2$ values under different model structures for the four PM$_{2.5}$ species. When involving all the supporting data including CMAQ simulations, land use variables and meteorological fields, fractional AODs and total AOD had similar predicting performance for all PM$_{2.5}$ species judged by $R^2$ (REF and M1), which were slightly better than M2 without AOD data. This indicates that CMAQ simulations, land use variables and meteorological fields together could explain the majority of the variability in PM$_{2.5}$ speciation concentrations. When developing PM$_{2.5}$ speciation models in data-rich regions, using total AOD or even no AOD could obtain similar model performance as fractional AODs. Moreover, total AOD or no AOD have the advantage of greater spatial/temporal coverage, since total AOD could be taken from MODIS and full coverage were achieved with no AOD involved.

When the model structure was simplified and only incorporated some of the supporting data, fractional AODs usually had better performance than total AOD, which is consistent with the findings in previous

Figure 2. Importance rankings of variables based on mean square errors (MSE) for (a) sulfate, (b) nitrate, (c) OC and (d) EC.
For example, M6 with fractional AODs and land use variables within the model had an out-of-bag $R^2$ of 0.58 for nitrate, much higher than the $R^2$ of 0.39 from M7. When developing PM$_{2.5}$ speciation models at data-poor regions where only some of the supporting data are available, using fractional AODs would be a better choice to more accurately estimate PM$_{2.5}$ speciation concentrations, although MISR AOD have relatively low sampling frequency. Table 2 also compares the model’s performance when using different combinations of supporting information, which could provide guidance for the data-poor regions about variable selections when developing their own models. Unlike the United States, many regions around the world do not have enough PM$_{2.5}$ speciation monitors that could support complex model structures. A simplified model is necessary and it is crucial to select the most useful
parameters to be included in the model. In addition, some developing regions also lack access to high quality air quality model simulations since they do not have local emission inventories or sufficient computing resources. Therefore, understanding the influences from different kinds of supporting data on model performance are important. Results of our sensitivity scenarios showed that, in general, models with land use variables had higher $R^2$ than other models, which is consistent with the importance ranking results mentioned above. For example, M6 had an out-of-bag $R^2$ of 0.62 for sulfate, higher than 0.55 from M3, 0.52 from M9 and 0.33 from M12. Models with only CMAQ simulations (e.g., M3) had good predicting power for nitrate, with $R^2$ higher than M6 with land use variables and M9 with meteorological fields. When choosing supporting data for a simplified model, the prior choice should be land use data, followed by CMAQ simulations.

We further compare the relative differences between the predicted concentrations from M3, M6, M9, M12 and the REF case at different impervious surface levels, which were models using fractional AODs and different combinations of supporting predictors (figure 4). For sulfate, M3 and M6 had very similar predicted results compared to the REF case, while the relative differences between M9, M12 and the REF case were generally within 25%. This confirms the findings in previous studies that sulfate is usually the best model among the four species. For nitrate, OC and EC, predictions from the models with land use variables (i.e., M6) are closest to those from the REF case, followed by the models with CMAQ simulations (i.e., M3). With these two kinds of predictors, the models could capture the majority of the spatial patterns of the PM$_{2.5}$ species, especially in urban areas with high percentages of impervious surface. When land use data and CMAQ simulations are absent, predictions from M9 and M12 had overestimated nitrate by ~150%, OC by ~100% and EC by ~200%.

### 3.6. Comparison with previous studies

Our random forest models outperformed the GAMs developed in Meng et al [20] using an experimental version of MISR 4.4 km fractional AODs. Their LOOCV $R^2$ values were 0.62, 0.56, 0.49 and 0.53 for sulfate, nitrate, OC and EC, respectively, compared to the out-of-bag $R^2$ between 0.68 and 0.72 of our models. Our models included more predictors (~20 in total) than the GAMs, including CMAQ simulations and several land use variables. As indicated in section 3.5, CMAQ and land use data had great predicting powers in the estimation of PM$_{2.5}$ speciation concentrations. Besides, random forest models can better resolve the complex relationship between ground observations and multiple predictors [34] than the statistical models.

Compared to Di et al [14] and Meng et al [15], which were models without any inputs from satellite AOD, our models had relatively lower out-of-bag $R^2$. Di et al [14] used a backward propagation neural network to calibrate the GEOS-Chem simulations and their mean total $R^2$ for sulfate, nitrate, OC and EC were 0.81, 0.83, 0.69 and 0.71, respectively. Meng et al [15] utilized random forest models and their out-of-bag $R^2$ varied between 0.71 and 0.86. Based on our sensitivity test results, when excluding AOD data from our full models, $R^2$ values decreased slightly. However, the number of paired data records in the training dataset would increase significantly if no AOD data were involved, which would improve model training results. Also, their studies had larger domain with greater contrasts in PM$_{2.5}$ species that could lead to higher $R^2$. In addition, these two works used spatially and temporally lagged terms in their models to take advantage of spatial and temporal auto-correlation. By incorporating information from nearby monitors and adjacent days, the models’ performance were highly improved. However, this might not work for data-poor regions where the monitors are too sparse or the observations are made far apart in time. Therefore, the

### Table 2. $R^2$ values of the sensitivity test results for sulfate, nitrate, OC and EC.

| Case  | AOD data          | Other data                      | Sulfate | Nitrate | OC   | EC   |
|-------|-------------------|---------------------------------|---------|---------|------|------|
| REF   | Fractional AODs   | CMAQ + Land use + Meteorological fields | 0.72    | 0.70    | 0.68 | 0.70 |
| M1    | Total AOD         | CMAQ + Land use + Meteorological fields | 0.72    | 0.69    | 0.68 | 0.70 |
| M2    | —                 | CMAQ + Land use + Meteorological fields | 0.69    | 0.67    | 0.67 | 0.69 |
| M3    | Fractional AODs   | CMAQ                            | 0.55    | 0.59    | 0.45 | 0.50 |
| M4    | Total AOD         | CMAQ                            | 0.48    | 0.48    | 0.36 | 0.40 |
| M5    | —                 | CMAQ                            | 0.29    | 0.34    | 0.24 | 0.30 |
| M6    | Fractional AODs   | Land use                        | 0.62    | 0.58    | 0.55 | 0.58 |
| M7    | Total AOD         | Land use                        | 0.60    | 0.39    | 0.43 | 0.47 |
| M8    | —                 | Land use                        | 0.37    | 0.30    | 0.40 | 0.44 |
| M9    | Fractional AODs   | Meteorological fields           | 0.52    | 0.45    | 0.34 | 0.33 |
| M10   | Total AOD         | Meteorological fields           | 0.48    | 0.35    | 0.24 | 0.25 |
| M11   | —                 | Meteorological fields           | 0.36    | 0.22    | 0.16 | 0.19 |
| M12   | Fractional AODs   | —                               | 0.33    | 0.32    | 0.20 | 0.20 |
| M13   | Total AOD         | —                               | 0.12    | 0.03    | 0.01 | 0.01 |
spatial-temporal lagged terms were not considered as input variables in our study.

3.7. Limitation
Our work was subject to some limitations. MISR has a low sampling frequency, resulting in a lot of missingness in the data that might affect the spatial and temporal trends presented in our study. However, our work presented a feasible method to use fractional AODs, CMAQ simulations, meteorological fields and land use variables to estimate PM2.5 speciation concentrations. Future improvements in the satellite instrument could bring more data and also improve the performance of our model. For example, the Multi-Angle Imager for Aerosols (MAIA) instrument [40], currently in development, has a legacy of multi-angle imaging from MISR but includes 14 bands in the ultraviolet to shortwave infrared including three polarimetric bands to obtain higher sensitivity to aerosol composition [40, 41]. MAIA will also make its optical and PM type retrievals at a higher spatial resolution and sampling frequency. The MAIA AOD retrievals are expected to have better ability in estimating PM2.5 speciation concentrations than MISR. In the sensitivity analyses, we compared the model performance between fractional AODs and total AOD based on an assumption that they have the same spatiotemporal coverage. However, total AOD from other instrument such as MODIS might have much higher data coverage than MISR, which could improve the model’s performance as the number of input data are larger. The findings in our study are applicable in regions similar to California, however, for regions with different emission sources or patterns, results remain unknown. Therefore, more efforts are needed to further understand the influence of data from different sources in different types of places.

Acknowledgments
The work of G Geng and Y Liu is partially supported by the MAIA science team and MISR science team, both at the JPL, California Institute of Technology, led by D Diner (Subcontract #1588347 and #1363692), and by the NASA Applied Sciences Program, managed by J Haynes (Grant # NNX16AQ28G, PI: Y Liu). We thank James Mulholland and Nirupama Senthilkumar from Georgia Tech for providing the CMAQ simulations.
Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] Raaschou-Nielsen O, Andersen Z J, Beelen R, Samoli E, Stafoggia M, Weinmayr G, Hofmann B, Fischer P, Nieuwlaat R and Brunekreef B 2013 Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE) Lancet Oncol. 14 813–22

[2] Turner M C, Krewski D, Pope C A III, Chen Y, Gapstur S M and Thun M J 2011 Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers Am. J. Respiratory Crit. Care Med. 184 1574–81

[3] Brook R D, Rajagopalan S, Pope C A III, Brook J R, Bhatnagar A, Diez-Roux A V, Holguin F, Hong Y, Luepker R V and Mittleman M A 2010 Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association Circulation 121 2331–78

[4] Achilleos S, Kioumourtzoglou M-A, Wu C-D, Schwartz J D, Koutrakis P and 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

[5] Chung Y, Dominici F, Wang Y, Coull B A and 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. 123 467–74

[6] Rohr A C and Wynga R E 2012 Attributing health effects to individual particulate matter constituents Atmos. Environ. 62 130–52

[7] Geng G, Zhang Q, Tong D, Li M, Zheng Y, Wang S and He K 2017 Chemical composition of ambient PM2.5 over China and relationship to precursor emissions during 2005–2012 Atmos. Chem. Phys. 17 9187–203

[8] Li C, Martin R V, van Donkelaar A, Boys B L, Hammer M S, Xu J W, Marais E A, Reff A, Strum M and Ridley D A 2017 Trends in chemical composition of global and regional population-weighted fine particulate matter estimated for 25 years Environ. Sci. Technol. 51 11185–95

[9] Geng G, Murray N L, Tong D, Fu J S, Hu X, Lee P, Meng X, Chang H H and Liu Y 2018 Satellite-based daily PM2.5 estimates during fire seasons in Colorado J. Geophys. Res.: Atmos. 123 8159–71

[10] Xiao Q, Chang H H, Geng G and Liu Y 2018 An ensemble machine-learning model to predict historical PM2.5 concentrations in China from satellite data Environ. Sci. Technol. 52 13260–9

[11] Xue T, Zheng Y, Tong D, Zheng B, Li X, Zhu T and Zhang Q 2019 Spatiotemporal continuous estimates of PM2.5 concentrations in China, 2000–2016: a machine learning method with inputs from satellites, chemical transport model, and ground observations Environ. Int. 123 345–57

[12] Philip S, Martin R V, van Donkelaar A, Lo J W-H, Wang Y, Chen D, Zhang L, Kashibata P S, Wang S and Zhang Q 2014 Global chemical composition of ambient fine particulate matter for exposure assessment Environ. Sci. Technol. 48 13060–8

[13] van Donkelaar A, Martin R V, Li C and Burnett R T 2019 Regional estimates of chemical composition of fine particulate matter using a combined geocscience-statistical method with information from satellites, models, and monitors Environ. Sci. Technol. 53 2395–411

[14] Di Q, Koutrakis P and Schwartz J 2016 A hybrid prediction model for PM2.5 mass and components using a chemical transport model and land use regression Atmos. Environ. 131 390–9

[15] Meng X, Hand J L, Schichtel B A and Liu Y 2018 Space-time trends of PM2.5 constituents in the conterminous United States estimated by a machine learning approach, 2003–2015 Environ. Int. 121 1137–47

[16] Diner D J, Beckert J C, Reilly T H, Bruegge C J, Conel J E, Kahn R A, Martonchik J V, Ackerman T P, Davies R and Gerssl A 1998 Multi-angle Imaging Spectro Radiometer (MISR) instrument description and experiment overview IEEE Trans. Geosci. Remote Sens. 36 1072–87

[17] Martonchik J V, Kahn R A and Diner D J 2009 Retrieval of aerosol properties over land using MISR observations Satellite Aerosol Remote Sensing Over Land (Berlin: Springer) pp 267–93

[18] Liu Y, Koutrakis P and Kahn R 2007 Estimating fine particulate matter component concentrations and size distributions using satellite-retrieved fractional aerosol optical depth: I. Method development J. Air Waste Manage. Assoc. 57 1351–9

[19] Liu Y, Koutrakis P, Kahn R, Turquety S and Yantosca R M 2007 Estimating fine particulate matter component concentrations and size distributions using satellite-retrieved fractional aerosol optical depth: II. A case study J. Air Waste Manage. Assoc. 57 1360–9

[20] Meng X, Garry M J, Diner D J, Kalashnikova O V, Xu J and 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

[21] Franklin M, Kalashnikova O V and Garry M J 2017 Size-resolved particulate matter concentrations derived from 4.4 km resolution size-fractionated Multi-angle Imaging Spectro Radiometer (MISR) aerosol optical depth over Southern California Remote Sens. Environ. 196 312–23

[22] Friberg M D, Kahn R A, Limbacher J A, Appel K W and Mulloolland J A 2018 Constraining chemical transport PM2.5 modeling outputs using surface monitor measurements and satellite retrievals: application over the San Joaquin Valley Atmos. Chem. Phys. 18 12891–913

[23] Geng G, Murray N L, Chang H H and Liu Y 2018 The sensitivity of satellite-based PM2.5 estimates to its inputs: implications to model development in data-poor regions Environ. Int. 121 550–69

[24] Geng G, Murray N L, Chang H H and Liu Y 2018 The sensitivity of satellite-based PM2.5 estimates to its inputs: implications to model development in data-poor regions Environ. Int. 121 550–69

[25] Garry M J, Kalashnikova O V and Bull M A 2017 Development of a higher-spatial-resolution (4.4 km) MISR aerosol optical depth product using AERONET–DRAGON data Atmos. Chem. Phys. 17 5095–106

[26] Martonchik J V, Diner D J, Kahn R A, Ackerman T P, Verstraete M M, Finty B and Gordon H R 1998 Techniques for the retrieval of aerosol properties over land and ocean using multiple imaging Imaging IEEE Trans. Geosci. Remote Sens. 36 1212–27

[27] Liu Y, Kahn R and Koutrakis P 2007 Estimating PM2.5 component concentrations and size distributions using satellite-retrieved fractional aerosol optical depth: II. A case study J. Air Waste Manage. Assoc. 57 1351–9

[28] Liu Y, Kahn R, Turquety S, Yantosca R M and Koutrakis P 2007 Estimating PM2.5 component concentrations and size distributions using satellite-retrieved fractional aerosol optical depth: II. A case study J. Air Waste Manage. Assoc. 57 1351–9

[29] Appel K W, Pouliot G A, Simon H, Sarwar G, Pye H O T, Napeleon S L, Akhtar F and Roselle S J 2013 Evaluation of dust and trace metal estimates from the community multiscale air quality (CMAQ) model version 5.0 Geosci. Model Dev. 6 883–914

[30] Zhang Y, Foley K M, Schwede B D, Bash J O, Pinto J P and Dennis R L 2019 A measurement-model fusion approach for improved wet deposition maps and trends J. Geophys. Res.: Atmos. 124 4237–51

[31] Otte T L and Pleim J E 2010 The meteorology–chemistry interface processor (MCIP) for the CMAQ modeling system: updates through MCIv3.4 J. Geosci. Model Dev. 3 243–56

[32] Bey I, Jacob D J, Yantosca R M, Logan J A, Field B D, Fiore A M, Li Q, Liu H Y, Mickley J L and Schultz M G 2001 Global modeling of tropospheric chemistry with assimilated
meteorology: model description and evaluation J. Geophys. Res. 106 23073–95

[32] Carlton A G and Baker K R 2011 Photochemical modeling of the ozark isoprene volcano: MEGAN, BEIS, and their impacts on air quality predictions Environ. Sci. Technol. 45 4438–45

[33] Breiman L 2001 Random forests Mach. Learn. 45 5–32

[34] Hu X F, Belle J H, Meng X, Wildani A, Waller L A, Strickland M J and Liu Y 2017 Estimating PM2.5 concentrations in the conterminous united states using the random forest approach Environ. Sci. Technol. 51 6936–44

[35] Murray N L, Holmes H A, Liu Y and Chang H H 2019 A Bayesian ensemble approach to combine PM2.5 estimates from statistical models using satellite imagery and numerical model simulation Environ. Res. 178 108601

[36] Almaraz M, Bai E, Wang C, Trousdell J, Conley S, Faloona I and Houlton B Z 2018 Agriculture is a major source of NOx pollution in California Sci. Adv. 4 eaao3477

[37] Williams E J, Guenther A and Fehsenfeld F C 1992 An inventory of nitric oxide emissions from soils in the United States J. Geophys. Res.: Atmos. 97 7511–9

[38] Clinton N E, Gong P and Scott K 2006 Quantification of pollutants emitted from very large wildland fires in Southern California, USA Atmos. Environ. 40 3686–95

[39] Mühle J, Luiker T J, Su Y, Miller B R, Prather K A and Weiss R F 2007 Trace gas and particulate emissions from the 2003 southern California wildfires J. Geophys. Res.: Atmos. 112 D03307

[40] Diner D J et al 2018 Advances in multiangle satellite remote sensing of speciated airborne particulate matter and association with adverse health effects: from MISR to MAIA J. Appl. Remote Sens. 12 042603

[41] Liu Y and Diner D J 2017 Multi-angle imager for aerosols: a satellite investigation to benefit public health Public Health Rep. 132 14–7