Published in final edited form as:
Environ Pollut. 2018 November ; 242(Pt A): 675–683. doi:10.1016/j.envpol.2018.07.016.

Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China Plain

Keyong Huang[a,b], **Qingyang Xiao**[b], **Xia Meng**[b], **Guannan Geng**[b], **Yujie Wang**[c,d], **Alexei Lyapustin**[c], **Dongfeng Gu**[a,*], **Yang Liu**[b,*]

[a]Department of Epidemiology, State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100037, China

[b]Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA

[c]NASA Goddard Space Flight Center, Greenbelt, MD, USA

[d]Goddard Earth Sciences and Technology Center, University of Maryland Baltimore County, Baltimore, MD, US

Abstract

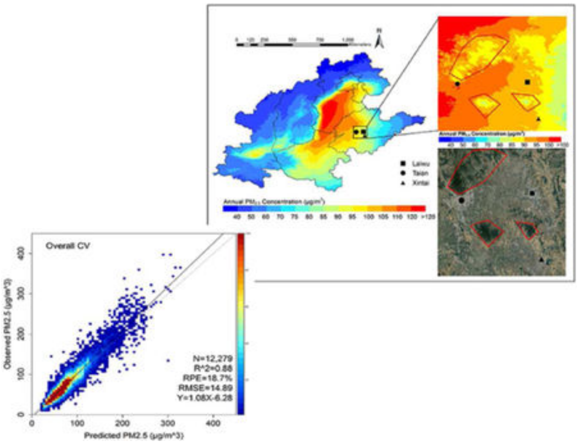
Exposure to fine particulate matter ($PM_{2.5}$) remains a worldwide public health issue. However, epidemiological studies on the chronic health impacts of $PM_{2.5}$ in the developing countries are hindered by the lack of monitoring data. Despite the recent development of using satellite remote sensing to predict ground-level $PM_{2.5}$ concentrations in China, methods for generating reliable historical $PM_{2.5}$ exposure, especially prior to the construction of $PM_{2.5}$ monitoring network in 2013, are still very rare. In this study, a high- performance machine-learning model was developed directly at monthly level to estimate $PM_{2.5}$ levels in North China Plain. We developed a random forest model using the latest Multi-angle implementation of atmospheric correction (MAIAC) aerosol optical depth (AOD), meteorological parameters, land cover and ground $PM_{2.5}$ measurements from 2013 to 2015. A multiple imputation method was applied to fill the missing values of AOD. We used 10-fold cross-validation (CV) to evaluate model performance and a separate time period, January 2016 to December 2016, was used to validate our model's capability of predicting historical $PM_{2.5}$ concentrations. The overall model CV R^2 and relative prediction error (RPE) were 0.88 and 18.7%, respectively. Validation results beyond the modeling period (2013 to 2015) shown that this model can accurately predict historical $PM_{2.5}$ concentrations at the monthly (R^2 = 0.74, RPE = 27.6%), seasonal (R^2 = 0.78, RPE = 21.2%) and annual (R^2 = 0.76, RPE = 16.9%) level. The annual mean predicted $PM_{2.5}$ concentrations from 2013 to 2016 in our

[*]**Address for correspondence**: Yang Liu PhD, Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA, 1518 Clifton Road NE, Atlanta, GA 30322, USA, yang.liu@emory.edu. Dongfeng Gu PhD, Department of Epidemiology, State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100037, China 167 Beilishi Rd, Beijing, 100037, PRC, gudongfeng@vip.sina.com.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

study domain was 67.7 μg/m3 and Southern Hebei, Western Shandong and Northern Henan were the most polluted areas. Using this computationally efficient, monthly and high-resolution model, we can provide reliable historical $PM_{2.5}$ concentrations for epidemiological studies on $PM_{2.5}$ health effects in China.

Graphical abstract



Capsule:

Random forest model developed at monthly level using satellite data can be applied to estimate long-term $PM_{2.5}$ concentrations in North China Plain.

Keywords

$PM_{2.5}$; MAIAC AOD; Machine Learning; North China Plain

1 Introduction

Numerous epidemiological studies have found that long-term exposure to fine particulate matter ($PM_{2.5}$) was associated with higher risk of cardiovascular diseases and respiratory diseases, which were mainly conducted in western countries(Di et al., 2017; Miller et al., 2007). Compared with cross-sectional or time-series studies, epidemiological cohort studies can clearly identify the temporal sequence between exposure and outcome, and provide more stable results after long-term follow-up surveys, thus allow more accurate health effect estimates(Di et al., 2017; Pope et al., 2002). With the rapid industrialization and economic development, severe $PM_{2.5}$ pollution episodes frequently occurred in China and the annual $PM_{2.5}$ concentrations were much higher than the WHO standards (10 μg/m3). However, epidemiological studies of $PM_{2.5}$ health impacts were very rare in China, because they required long-term, accurate $PM_{2.5}$ exposure data, which was not available until the establishment of ground monitoring network in 2013.

With a high spatiotemporal coverage, satellite-derived aerosol optical depth (AOD) has been increasingly used to predict $PM_{2.5}$ concentrations and can supplement ground $PM_{2.5}$ monitors for health studies. Several large-scale epidemiological cohort studies of $PM_{2.5}$

health effects have used the satellite AOD-$PM_{2.5}$ prediction models to assess the exposure levels, such as the Global Burden of Diseases study (GBD), US Medicare study and Canadian National-level Cohort study(Cohen et al., 2017; Crouse et al., 2012; Di et al., 2017).

There were two major approaches which have been used to estimate ground $PM_{2.5}$ concentrations using satellite AOD: the scaling approach and the statistical models(Liu et al., 2009; van Donkelaar et al., 2010). The scaling approach obtained the ratios of $PM_{2.5}$/AOD from the chemical transport model (CTM), then applied this ratio to predict ground $PM_{2.5}$(van Donkelaar et al., 2010). This method didn't require ground $PM_{2.5}$ measurements and thus can provide $PM_{2.5}$ estimations for areas or time periods without ground monitoring data. However, the accuracy of the scaling approach is limited by the emission inventory, and the uncertainties in the model parameterization(Xiao et al., 2017). For the statistical model approach, numerous regression models have been developed to define the association between $PM_{2.5}$ and AOD. These models have become increasingly complicated, from simple linear regression to much complex models, including linear mixed effects models (LME), generalized additive models (GAM), geographically weighted regression (GWR), hierarchical models and Bayesian models(Ma et al., 2014; Ma et al., 2016; Wang, 2003; Xie et al., 2015; Yu et al., 2017). However, the statistical model approach required large amount of ground $PM_{2.5}$ monitoring data to develop and validate the models, therefore studies of developing AOD-$PM_{2.5}$ models were limited in China before 2013.

Benefiting from recently established ground monitoring networks, several advanced statistical models between AOD and $PM_{2.5}$ have been established in China at the regional or at national levels(Ma et al., 2016; Xiao et al., 2017; Xie et al., 2015). For example, Ma et al., (2016) built a two-stage statistical model including LME and GAM models to build the relations between $PM_{2.5}$ and AOD in China. He and Huang, (2018) developed a geographically and temporally weighted regression model (GTWR) to estimate $PM_{2.5}$ exposure in China. However, the coarse spatial resolution (10 km or 3 km) in those studies cannot be used to support urban-scale or smaller area exposure assessment for epidemiological studies(Hu et al., 2014a). The resolution of AOD data was the main reason for the coarse spatial resolution. For instance, AOD products from Multiangle Imaging SpectroRadiometer (MISR), Moderate Resolution Imaging Spectroradiometer (MODIS) Collection 5 (C5), and MODIS C6 have a spatial resolution of 17.6 km, 10 km, and 3 km, respectively(Hu et al.,2014a). Recently, a new Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm was developed for MODIS AOD at 1 km resolution. MAIAC AOD has been shown to be highly correlated with $PM_{2.5}$ levels and been increasingly used to predict $PM_{2.5}$ concentrations in North America(Di et al., 2016; Hu et al., 2014a; Kloog et al., 2014). However, similar studies were rare in China because the MAIAC data was still not publicly available.

Another limitation of previous studies was the non-random missingness of AOD, which was mainly caused by cloud cover, bright surfaces and extremely high aerosol loadings which were incorrectly regarded as cloud(Xiao et al., 2017). Without considering the non- random missingness of AOD, previous satellite AOD-$PM_{2.5}$ prediction models cannot be directly used to provide exposure assessment for epidemiological studies. Otherwise, it may lead to exposure bias when estimating long-term concentrations of $PM_{2.5}$ in epidemiological studies. For instance, Zheng et al., (2016) reported that missing AOD values introduce negative biases in predicting annual $PM_{2.5}$ levels in Beijing-Tianjin-Hebei region (BTH), while lead to positive biases in Pearl River Delta region (PRD), probably due to distinct mechanisms of missing AOD in these two region. These findings were supported by those of Xie et al., (2015), who also reported an underestimate of long-term $PM_{2.5}$ concentrations due to missing AOD in Beijing.

To investigate the health effects caused by $PM_{2.5}$ in China, it was crucial to develop $PM_{2.5}$ prediction models with capabilities of predicting historical $PM_{2.5}$ concentrations prior to establishment of ground monitoring network in 2013. However, many previous studies had a strong model assumption that the daily $PM_{2.5}$-AOD relationship remained constant for the same day of year across different years, thus leading to low prediction accuracy when used to predict $PM_{2.5}$ concentrations beyond their modeling period(Liang et al., 2018; Ma et al., 2016; Xiao et al., 2017).

Furthermore, most studies developed their statistical models at the daily level, then averaged to corresponding longer time scales to study the chronic $PM_{2.5}$ health effects(e.g. previous 3 months, previous year)(Cohen et al., 2017; Crouse et al., 2012; Di et al., 2017), which would require a high computational cost and take a very long time to calculate when developing daily statistical models at high spatial resolution (1 km) over a large area for a long time period. Except for the traditional statistical models, recent studies have also attempted to use the machine learning algorithm models to make $PM_{2.5}$ estimations(Di et al., 2016; Hu et al., 2017; Zhan et al., 2017). With the capabilities of handling nonlinear relations and interaction effects between variables, the machine learning methods generally shown comparable or superior performance to traditional statistical models. For example, Hu et al., (2017) used the random forest algorithm to predict daily ground PM $_{2.5}$ levels at 10 km resolution in the United States, with a cross validation (CV) R^2 of 0.80. Di et al., (2016) developed a neural network-based model to estimate daily $PM_{2.5}$ in the United States at 1 km resolution, achieving a R^2 of 0.84. In China, a geographically weighted gradient boosting machine (GW-GBM) was developed to estimate $PM_{2.5}$ levels at 50 km resolution(Zhan et al., 2017). To our knowledge, no other machine learning methods used for estimating $PM_{2.5}$ concentrations have been reported in China.

In the current study, we aimed to develop a random forest model to provide historical, high-resolution and unbiased monthly $PM_{2.5}$ concentrations for epidemiological studies. We first adopted the multiple imputation method to fill the missing MAIAC AOD values considering the cloud cover, meteorological data and spatial-temporal autocorrelation of AOD. Then we developed a random forest model directly at the monthly level incorporating 1-km MAIAC AOD, land use information, meteorological variables, and demographics as predictors, to estimate the monthly $PM_{2.5}$ concentrations from 2013 to 2015. We used 10-fold overall CV and spatial CV to evaluate the model's prediction accuracy. Additionally, a separate time period, January 2016 to December 2016, was used to assess our model's capability of predicting historical $PM_{2.5}$ concentrations.

2 Materials and methods

2.1 Study area

Figure 1 shows the North China study area, including seven provinces or municipalities (Beijing, Tianjin, Hebei Province, Shanxi Province, Shaanxi Province, Shandong Province, and Henan Province). The study region has an area of 1.1 million km^2 with a total population over 367 million in 2010. It is characterized by the heavy industries in this area, including the coal-fired power plant, cement factories and iron and steel factories, which are the main sources of $PM_{2.5}$ emissions. In addition, heating by fossil fuels burning in winter and a rapidly increasing vehicle fleet further aggravate the air pollution in this area. We build a 50-km buffer to obtain the $PM_{2.5}$ predictions near the boundary with similar accuracy to those from the remained locations in our study area.

2.2 Data

2.2.1 Ground measurements—Ground $PM_{2.5}$ measurements were collected from China National Environmental Monitoring Center (http://www.cnemc.cn/) and monitoring stations controlled by local governments in our study region. In total, daily average $PM_{2.5}$ monitoring data from 704 air quality monitors in this area from January 2013 to December 2016 were collected for this study (Figure 1). After excluding those months with less than 10 days of $PM_{2.5}$ measurements, we calculated the monthly average $PM_{2.5}$ concentrations.

With a high accuracy, Aerosol Robotic Network (AERONET) AOD has been used for satellite AOD validation and calibration in several studies. In this study, we downloaded AERONET Level 1.5 data from 11 sites in the study region ranging from 2013 to 2016 (https://aeronet.gsfc.nasa.gov/new_web/aerosols.html) and used it to calibrate the MAIAC AOD data. AERONET AOD at 550 nm was interpolated from the AOD at 440 and 675 nm using the Angstrom Exponent.

Land based visibility data from 2013 to 2016 was downloaded from the National Centers for Environmental Information (NCEI, ftp://ftp.ncdc.noaa.gov/pub/data/noaa/). There were 113 stations measuring visibility in our study domain.

2.2.2 Satellite data—In the current study, we used the MAIAC AOD for $PM_{2.5}$ modeling. MAIAC is an algorithm used to retrieve AOD from Aqua (cross at 1:30 pm local time) and Terra (cross at 10:30 local time) at 1 km resolution(Lyapustin et al., 2011). We got the AOD data from the MAIAC team covering the year of 2013 to 2016. We used the MAIAC 1-km grid for data integration. Within each grid, we conducted a simple linear regression between MAIAC AOD and AERONET AOD within the hour of satellite crossover time for each season. Then we used the established relations to calibrate the MAIAC AOD. To increase the MAIAC coverage, we performed simple linear regression between Aqua and Terra AOD for each day and used the relations to predict missing AOD values when there is only one of them present. Then we used the average of Aqua and Terra AOD for $PM_{2.5}$ modeling.

MODIS Aqua and Terra daily cloud fraction data (MYD06_L2 and MOD06_L2, 5-km resolution) and monthly Normalized Difference Vegetation Index (NDVI) data (MOD13A3, 1-km resolution) was downloaded from the NASA website (https://ladsweb.modaps.eosdis.nasa.gov/). In addition, we downloaded the MODIS fire data from NASA fire information for resource management system (https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms). We obtained the urban cover, forest cover and crop land cover at 300 m resolution from the European Space Agency (ESA) GlobalCover Portal (http://due.esrin.esa.int/page_globcover.php).

2.2.3 Assimilated dataset—Meteorological data were obtained from the European Center for Medium-Range Weather Forecast (ECMWF) Re-analysis Interim (ERA-Interim) (Dee et al., 2011). The spatial resolution of this dataset is 0.125 degree and the temporal resolution is per 3 or 6 hours. All meteorological measurements for the period from 8:00 am to 2:00 pm local time were extracted and averaged to represent the weather conditions at the Aqua and Terra overpass time. The Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) is the latest atmospheric reanalysis dataset produced by NASA, at a spatial resolution of 0.5°×0.625°(Gelaro et al., 20 17). Three-hourly AOD data were downloaded from MERRA-2 website (https://disc.gsfc.nasa.gov/daac-bin/FTPSubset2.pl) and averages of AOD values from 8:00 am to 2:00 pm local time were calculated to represent MERRA-2 AOD at the satellite overpass time. In addition, MERRA-2 simulates five $PM_{2.5}$ species (black carbon, organic carbon, sea salt, dust, sulfate) based on the Goddard Chemistry, Aerosol, Radiation, and Transport (GOCART) aerosol model. We calculated the daily MERRA-2 $PM_{2.5}$ concentrations based on above $PM_{2.5}$ species, according to the method used in previous studies(Provençal et al., 2017),

2.2.4 Other datasets—The Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) at 1 km resolution was used to obtain the elevation. The population density data in 2010 (1 km resolution) was downloaded from LandScan (http://wms.cartographic.com/LandScan2010/). The road network information in 2014 was obtained from Beijing NavInfo Technology Company.

2.3 Data integration

The MAIAC 1-km grid was used to integrate all the aforementioned datasets. Ground $PM_{2.5}$ data were averaged within each grid, if there were multiple monitors in this grid. The monthly means of ECMWF meteorological parameters and visibility were interpolated to the MAIAC grid using the inverse distance weighting (IDW) method. We used the nearest neighbor approach to match daily cloud fraction to each MAIAC grid. The major road length (road length multiple by lane number) and the nearest distance to major roads, population density, elevation and NDVI were calculated for each MAIAC grid. MERRA-2 AOD and MERRA-2 $PM_{2.5}$ were matched to each MAIAC grid if this MAIAC grid fell into a given MERRA-2 grid. Monthly counts of fire spots for each MAIAC grid within 20- km, 30-km, 50-km and 75-km radius buffer were calculated, respectively. The exploratory analysis suggested that counts of fire spots within 75-km buffer contributed more to the model prediction accuracy than other buffer lengths, thus we selected fire counts within 75-km buffer in the final model(Hu et al., 2014b).

2.4 Methods

2.4.1 MAIAC AOD gap filling by multiple imputation—We used the AOD gap filling method proposed recently by Xiao, et al(Xiao et al., 2017). The detailed procedure of our AOD gap filling was articulated in the supplemental material. In brief, the missing MAIAC AOD was filled by an additive imputation model. This model included cloud fraction, elevation, MERRA-2 AOD, meteorological parameters and spatial- temporal trends of AOD as predictors (Xiao et al., 2017). After the inclusion of meteorological variables and cloud fraction data, our imputation method considered the aerosol-cloud interaction effect on AOD. Within each rolling 5-day period, we imputed the missing AOD values on the 3rd day by including AOD measurements on two days prior to and two days after that day.

2.4.2 Random forest model development and validation—Random forest is a machine learning method for classification and regression which uses an ensemble of decision trees(Hu et al., 2017). Each tree is grown by a bootstrap sample, and a random subset of predictors is selected at each split. Predictions are obtained by averaging results of different trees. Instead of being a black box like other machine learning methods, the random forest model can provide variable importance measures which make our model results more interpretable. Our random forest model incorporated MAIAC AOD, meteorological parameters, elevation, land use information, population and MERRA-2 $PM_{2.5}$ to estimate monthly ground $PM_{2.5}$ concentrations. The detailed predictors were shown in online supplemental Table S1.

We conducted an overall 10-fold CV to evaluate the model performance. The entire model-fitting dataset was randomly split into 10 groups, with each group containing about 10% of the data. In each time of cross validation, nine groups of the data were selected to fit the model, which was then used to make predictions on the remaining group. This process was repeated 10 times until every group was predicted. In addition, to validate the prediction accuracy of $PM_{2.5}$ in unmonitored locations, we performed a spatial CV, in which the training 1-km grid cells were randomly split into 10 groups first, with each group containing 10% of the grid cells. In each time of cross validation, we select nine groups of the grid cells, then all the data from these grid cells were used to fit the model and make predictions on the remaining group. This process was repeated 10 times until every group was predicted. Furthermore, a separate time period, January 2016 to December 2016, was used to evaluate the model's accuracy of predicting historical $PM_{2.5}$ levels. Coefficient of determination (R^2), root mean squared prediction error (RMSE) and relative prediction error (RPE) between model predictions and observations were calculated to evaluate the model performance.

3 Results

3.1 Descriptive analysis

Finally, there were 12,279 observations included in the 2013–2015 model fitting dataset. The descriptive statistics for all parameters were shown in Table S2. Overall, the average of ground $PM_{2.5}$ was 81 μg/m3, and the mean gap filled AOD value wa s 0.85. These estimates were more than 6 times higher than those reported in the continental United States(Hu et al., 2017; Liu et al., 2009).

3.2 Multiple imputation

The percentage of missing MAIAC AOD was 45.5%, 49.8%, 52.6% for year 2013–2015, respectively. After multiple imputation, the coverage of AOD increased to 100%. The average fitting R^2 of our daily multiple imputation model was 0.79, ranging from 0.52 to 0.94. The annual mean AOD distribution without versus with imputation for each year was shown in Figure S1. The spatial contrast of AOD after imputation was consistent with that before imputation: the highest AOD values occurred at Southern Hebei, Western Shandong and Northern Henan Province. Nevertheless, the annual mean AOD values increased by 0.23 after imputation compared with that of observed AOD. Temporally, the annual AOD values decreased gradually from 0.79 in 2013 to 0.74 in 2015 (Table S3).

3.3 Results of model validation

The overall CV and spatial CV analysis results were shown in Figure 2. Our random forest model achieved a high prediction accuracy with an overall 10-fold CV R^2 of 0.88. In addition, the RPE and RMSE for monthly $PM_{2.5}$ predictions were 18.7% and 14.89 $\mu g/m^3$, respectively, implying a relatively good agreement between model predictions and ground measurements in North China area. The imputation procedure usually causes extra variability because of the random error(Xiao et al., 2017). However, the overall CV R^2 from models using original MAIAC AOD was 0.88 as well, indicating that our multiple imputation method did not decrease the model's accuracy. The 10-fold spatial CV validation obtained a similar R^2 of 0.88, while RMSE slightly increased to 15.06 $\mu g/m^3$.

3.4 Evaluation of historical $PM_{2.5}$ predictions

A separate time period was applied to assess if our random forest model can predict historical $PM_{2.5}$ accurately. The random forest model fitted by data of 2013 to 2015 was used to predict $PM_{2.5}$ concentrations in 2016 at monthly, seasonal and annual level (Figure 3). Results shown that our model can predict historical $PM_{2.5}$ concentrations with high accuracy at the monthly level (R^2=0.74, RMSE=17.80 $\mu g/m\ ^3$ and RPE=27.6%), seasonal level (R^2=0.78, RMSE=13.75 $\mu g/m^3$ and RPE=21.2%) and annual level (R^2=0.76, RMSE=11.35 $\mu g/m^3$ and RPE=16.9%). The current model underestimated $PM_{2.5}$ concentrations at higher concentrations (>180 μg/m3). This could be because more than 96% of the monthly $PM_{2.5}$ measurements were below the 180μg/m ^{3} in the modeling dataset. After we removed the monthly $PM_{2.5}$ observations greater than 180 $\mu g/m^3$, the model performance improved with slope much closer to one than before (Figure 3).

3.5 Variable importance

Figure 4 illustrated the variable importance metrics for predictors in our random forest model. It demonstrated that the MERRA-2 $PM_{2.5}$ measurements, 10-meter wind speed, visibility, surface albedo and MAIAC AOD are the five most important predictors for monthly $PM_{2.5}$ concentrations.

3.6 Spatial and temporal distributions of $PM_{2.5}$

The spatial distributions of annual mean $PM_{2.5}$ concentrations from 2013 to 2016 in the study area were shown in Figure S2. The annul mean predicted $PM_{2.5}$ concentrations from

2013 to 2016 was 67.7 μg/m3 in North China area. Th e most polluted areas were in Southern Hebei, Northern Henan and Western Shandong areas. In terms of temporal trend, the $PM_{2.5}$ concentrations decreased by 10.4 μg/m3 from 2013 t o 2016. In addition, our high resolution prediction model successfully displayed the local $PM_{2.5}$ gradients in our study area. For instance, Figure 5 showed that urbanized regions, such as Tai' an, Laiwu and Xintai city, had higher $PM_{2.5}$ concentrations, while regions covered by forest (red polygon) around these three cities had lower $PM_{2.5}$ concentrations.

4 Discussion

The current study developed a random forest model to estimate the monthly $PM_{2.5}$ concentrations at 1 km resolution and achieved a high prediction accuracy (cross validation R^2 is 0.88), which was better than existing models(Liang et al., 2018; Xiao et al., 2017; Yanosky et al., 2014; Yanosky et al., 2009). To the best of our knowledge, this is the first study in China to develop a prediction model of $PM_{2.5}$ concentrations directly at the monthly level covering a large area (more than 1.1 million km^2). As discussed below, our high resolution, monthly and computationally efficient model will help epidemiologists assess the long term exposure of $PM_{2.5}$ concentrations with better accuracy and thus be beneficial to epidemiological studies of health effects caused by $PM_{2.5}$.

Our models have several strengths. First, we developed our random forest model directly at the monthly level, which not only took much less time to compute but also achieved a high prediction accuracy. With a high spatial and temporal coverage of satellite data, several $PM_{2.5}$-AOD statistical models have been developed and ultimately applied to the epidemiological studies(Cohen et al., 2017; Di et al., 2017). But most previous studies developed their $PM_{2.5}$ prediction models at the daily level, then averaged to corresponding longer time scales to study chronic health effects(e.g., previous 3 months, previous year) (Cohen et al., 2017; Crouse et al., 2012; Di et al., 2017). It would take a long time to run a complex model at daily level, if conducted over a large area at a high spatial resolution covering a long time period. The only study we found which also built $PM_{2.5}$ prediction models at the monthly level was from the Nurses' Health Study, a large-scale prospective cohort study in the United States(Yanosky et al., 2014; Yanosky et al., 2009). They developed two separate generalized additive mixed models to predict monthly $PM_{2.5}$ concentrations for 1988–1998 and 1999–2007, respectively and both obtained a high predictive accuracy (CV R^2=0.77), and ultimately used this monthly model in cohort studies(Zhang et al., 2016). It indicated the feasibility of developing $PM_{2.5}$ prediction models at monthly level for long-term epidemiological studies. However, for the time period before 1998 when there were few $PM_{2.5}$ monitoring stations in the United States, they used the ratio of $PM_{2.5}$ to PM_{10} predicted from other time period to predict $PM_{2.5}$ levels(Yanosky et al., 2014). Thus the model accuracy before 1998 was unknown and they didn't report the validation accuracy beyond their modeling period. In the current study, we selected a separate time period to assess the models' capability of predicting historical $PM_{2.5}$ concentrations, and achieved a relatively high accuracy (Figure 3).

Second, our model can provide reliable historical $PM_{2.5}$ exposure data when there was no ground monitoring data. Several satellite-driven statistical models based on MODIS AOD have been developed at daily level in China, and they showed the ability of generating accurate $PM_{2.5}$ estimations(Guo et al., 2017; Liang et al., 2018; Ma et al., 2016; Xie et al., 2015). For example, Guo et al., (2017) developed a satellite-based GTWR model based on 3-km MODIS AOD to estimate daily ground $PM_{2.5}$ in Beijing and achieved a CV R^2 of 0.58. Similarly, Xie et al., (2015) developed a mixed effects model at 3 km resolution in Beijing and obtained a CV R^2 of 0.79. However, most previous studies focused on developing association between $PM_{2.5}$ and AOD within their modeling periods and ignored the model's capability of predicting historical PM $_{2.5}$ concentrations. It is crucial to develop $PM_{2.5}$-AOD prediction models which can provide accurate historical $PM_{2.5}$ estimations for epidemiological studies in China, owing to the lack of $PM_{2.5}$ monitoring data before 2013. Only a few studies have adopted a separate time period to validate their model's performance of estimating historical $PM_{2.5}$ concentrations. Liang et al., (2018) developed a three-stage statistical model (LME + Generalized additive mixed model + Kriging model) to predict daily $PM_{2.5}$ in Beijing using the 1-km MAIAC AOD, and obtained a CV R^2 of 0.79 to 0.82. However, the model's accuracy decreased (R^2 of 0.42 to 0.55), when it was used to predict historical $PM_{2.5}$ concentrations at monthly level(Liang et al., 2018). In the current study, we built our model directly at the monthly level, and still obtained a relatively good performance when used to predict historical $PM_{2.5}$ data at the monthly level (R^2 = 0.74).

Third, we improved the satellite AOD coverage to 100% by multiple imputation methods. The non-random AOD missing values might introduce exposure misclassification in epidemiological studies(Lv et al., 2016; Xiao et al., 2017). For instance, during the first quarter of 2015 (January to March), without AOD gap-filling, our model would underestimate the $PM_{2.5}$ concentrations in our study domain (Figure S3). The degree of the differences in predicted $PM_{2.5}$ concentrations were generally consistent with the spatial distributions of missing rate of AOD. Lower AOD coverage rates were associated with higher differences of $PM_{2.5}$ concentrations. In Southern Shaanxi and Southern Henan province with less than 30% AOD coverage, the $PM_{2.5}$ concentrations were underestimated by 5% to 10%. In addition, although the AOD coverage was about 60% in southern Hebei Province, it still underestimated $PM_{2.5}$ concentrations by 4% to 5%. Severe particulate matter pollution episodes frequently occurred in winter in Southern Hebei (Wang et al., 2014). Extremely high aerosol loadings might be incorrectly classified as clouds, which in turn leads to considerable AOD missingness(Zheng et al., 2016). Therefore, AOD missingness in areas with severe particulate matter air pollution tend to underestimate the $PM_{2.5}$ concentrations as well. Our results were consistent with previous studies conducted in North China areas(Lv et al., 2016; Xie et al., 2015; Zheng et al., 2016). Moreover, to increase the coverage of predicted $PM_{2.5}$ estimations, several gap-filling approaches have been proposed(Kloog et al., 2014; Lv et al., 2016). For instance, Lv et al., (2016) developed a two-step method to predict the missing AOD (first used season-specific AOD-$PM_{2.5}$ relation to fill missing AOD, then used ordinary Kriging to interpolate rest missing AOD). Kloog et al., (2014) considered the non-random missingness of AOD in the model fitting process using inverse probability weighting method, and filled the missing $PM_{2.5}$ predictions with spatial smoothing using values from surrounding grids. However, the major limitation of previous methods was that they depend on ground $PM_{2.5}$ data. Thus, they were not applicable to China in predicting historical $PM_{2.5}$ concentrations, because there were no ground monitoring data before 2013. Without depending on the ground measurements, our gap-filling method considered the cloud fraction, meteorological information and land use type, thus can be used to improve the prediction coverage when there were no or sparse $PM_{2.5}$ measurements.

Finally, the random forest model provided the variable importance measures, which made our results more interpretable and provided information for future studies to improve the prediction accuracy of $PM_{2.5}$. In the current study, the five most important variables were MERRA-2 $PM_{2.5}$, wind speed at 10 meters, visibility, surface albedo and MAIAC AOD. Using the GOCART model, MERRA-2 simulates the concentrations of the five $PM_{2.5}$ species. Several studies have evaluated the components of MERRA simulations in different regions of the world and they found that the concentrations of $PM_{2.5}$, PM_{10} were generally well simulated in both the U.S and Europe(Provençal et al., 2017). However, few studies have applied the MERRA-2 components in developing $PM_{2.5}$ prediction models in China. Our study found that MERRA-2 PM $_{2.5}$ was among the most important predictors for monthly $PM_{2.5}$ concentrations in North China. More research is needed to evaluate the contributions of MERRA-2 components to the performance of $PM_{2.5}$ prediction models. It is noteworthy that the local land cover and road length did not contributed too much to the model performance, which was different from a previous study conducted in the continental United States(Hu et al., 2017). Hu et al., (2017) reported that convolutional layer of $PM_{2.5}$ measurements, MODIS AOD, population density, local land cover and roads were important predictors for $PM_{2.5}$ estimations in the continental United States. It may result from the differences in the source profile of $PM_{2.5}$ between North China and the United States. Air quality in North China is highly influenced by strong point sources from heavy industries, including coal-fired power plant, cement factories and iron and steel factories(Lv et al., 2016). Several $PM_{2.5}$ source apportionment studies reported that residential, industrial and agricultural emissions were the most important contributors to primary and secondary $PM_{2.5}$ in North China(Li et al., 2017; Zhang et al., 2015). For instance, Li, et al. simultaneously used source apportionment and source sensitivity methods to identify the sources of $PM_{2.5}$ exposure in North China in 2013, and both methods shown that local emissions including industrial (42.7%), residential (36.9%), and agricultural (9.7%) emissions were the most important sources to $PM_{2.5}$, while transportation contributed less than 10%(Li et al., 2017). Additionally, measures of road length may not fully reflect the $PM_{2.5}$ emissions from traffic, and more accurate indicators, such as the traffic volume, should be used to develop $PM_{2.5}$ prediction models in future studies. In contrast, for the United States with much lower levels of $PM_{2.5}$ concentrations, local population activity and traffic were more important factors to nearby $PM_{2.5}$ concentrations than emissions from factories located a long distance away.

Except for the above strengths, our study found a decreasing trend of $PM_{2.5}$ levels from 2013 to 2016 in North China. The decreasing trend was consistent with observations from ground $PM_{2.5}$ monitors. In addition, previous studies also found a downward trend in $PM_{2.5}$ levels since 2013 in North China(He and Huang, 2018; Ma et al., 2016). This decline may be due to the stricter policies for energy conservation and emissions reductions from the Chinese government. For example, the State Council issued the ‘China National Action Plan on Air Pollution Prevention and Control 2013–2017’ in 2013(Jin et al., 2016). This action plan for the first time set quantitative air quality improvement goals with a clear time table and proposed ten key strategies to control the air pollution in China(Jin et al., 2016). Moreover, in 2016, as the first year of 13th Five-Year Plan, the new revised Air Pollution Control Law was implemented in China. Further, the regional coordination and integrated regional environmental management were well implemented in 2016 in Beijing-Tianjin-Hebei region and its surrounding areas.

5 Conclusions

In the current study, a random forest model was developed to estimate monthly $PM_{2.5}$ concentrations in North China including gap filled AOD, MERRA-2 simulations, meteorological parameters and land cover as predictors. Using this computationally efficient model, we can provide high resolution (1 km), historical monthly $PM_{2.5}$ concentrations with high accuracy covering the whole North China area (1.1 million km^2). The AOD gap-filling method used in this study can substantially increase the coverage of $PM_{2.5}$ predictions and reduce exposure assessment bias for epidemiological studies. Our prediction model will provide data support for epidemiological studies on $PM_{2.5}$ health effects. In the future, we will try to extrapolate our monthly models to other areas for $PM_{2.5}$ estimations and apply our predicted $PM_{2.5}$ estimations in epidemiological studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The work of Q. Xiao, G. Geng and Y. Liu was partially supported by Assistance Agreement No. 83586901 awarded by the U.S. Environmental Protection Agency to Y. Liu. It has not been formally reviewed by EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication. The work of X. Meng and Y. Liu was partially supported by the National Institutes of Health (Grant # 1R01ES027892). The work of D. Gu was supported by National Key Research and Development Program of China (2017YFC0211700) from Ministry of Science and Technology of China. The work of K. Huang was supported by the China Scholarship Council (201706210381).

Reference

1. Cohen AJ, Brauer M, Burnett R, Anderson HR, Frostad J, Estep K, Balakrishnan K, Brunekreef B, Dandona L, Dandona R, Feigin V, Freedman G, Hubbell B, Jobling A, Kan H, Knibbs L, Liu Y, Martin R, Morawska L, Pope CA, Shin H, Straif K, Shaddick G, Thomas M, van Dingenen R, van Donkelaar A, Vos T, Murray CJL, Forouzanfar MH, 2017 Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet 389, 1907–1918.

2. Crouse DL, Peters PA, van Donkelaar A, Goldberg MS, Villeneuve PJ, Brion O, Khan S, Atari DO, Jerrett M, Pope CA, Brauer M, Brook JR, Martin RV, Stieb D, Burnett RT, 2012 Risk of nonaccidental and cardiovascular mortality in relation to long-term exposure to low concentrations of fine particulate matter: a Canadian national- level cohort study. Environ Health Perspect 120, 708–714. [PubMed: 22313724]

3. Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Hólm EV, Isaksen L, Kållberg P, Köhler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette JJ, Park BK, Peubey C, de Rosnay P, Tavolato C, Thépaut JN, Vitart F, 2011 TheERA-Interim reanalysis: configuration

and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society 137, 553–597.

4. Di Q, Kloog I, Koutrakis P, Lyapustin A, Wang Y, Schwartz J, 2016 Assessing PM2.5 Exposures with High Spatiotemporal Resolution across the Continental United States. Environ Sci Technol 50, 4712–4721. [PubMed: 27023334]
5. Di Q, Wang Y, Zanobetti A, Wang Y, Koutrakis P, Choirat C, Dominici F, Schwartz JD, 2017 Air Pollution and Mortality in the Medicare Population. N Engl J Med 376, 2513–2522. [PubMed: 28657878]
6. Gelaro R, McCarty W, Suárez MJ, Todling R, Molod A, Takacs L, Randles CA, Darmenov A, Bosilovich MG, Reichle R, Wargan K, Coy L, Cullather R, Draper C, Akella S, Buchard V, Conaty A, da Silva AM, Gu W, Kim G-K, Koster R, Lucchesi R, Merkova D, Nielsen JE, Partyka G, Pawson S, Putman W, Rienecker M, Schubert SD, Sienkiewicz M, Zhao B, 2017 The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). Journal of Climate 30, 5419–5454.
7. Guo Y, Tang Q, Gong D-Y, Zhang Z, 2017 Estimating ground-level PM 2.5 concentrations in Beijing using a satellite-based geographically and temporally weighted regression model. Remote Sensing of Environment 198, 140–149.
8. He Q, Huang B, 2018 Satellite-based high-resolution PM2.5 estimation over the Beijing-Tianjin-Hebei region of China using an improved geographically and temporally weighted regression model. Environ Pollut 236, 1027–1037. [PubMed: 29455919]
9. Hu X, Belle JH, Meng X, Wildani A, Waller LA, Strickland MJ, Liu Y, 2017 Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. Environ Sci Technol 51, 6936–6944. [PubMed: 28534414]
10. Hu X, Waller LA, Lyapustin A, Wang Y, Al-Hamdan MZ, Crosson WL, Estes MG, Estes SM, Quattrochi DA, Puttaswamy SJ, Liu Y, 2014a Estimating ground- level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. Remote Sensing of Environment 140, 220–232.
11. Hu X, Waller LA, Lyapustin A, Wang Y, Liu Y, 2014b Improving satellite-driven PM2.5 models with Moderate Resolution Imaging Spectroradiometer fire counts in the southeastern U.S. J Geophys Res Atmos 119, 11375–11386. [PubMed: 28967648]
12. Jin Y, Andersson H, Zhang S, 2016 Air Pollution Control Policies in China: A Retrospective and Prospects. Int J Environ Res Public Health 13.
13. Kloog I, Chudnovsky AA, Just AC, Nordio F, Koutrakis P, Coull BA, Lyapustin A, Wang Y, Schwartz J, 2014 A New Hybrid Spatio-Temporal Model For Estimating Daily Multi-Year PM2.5 Concentrations Across Northeastern USA Using High Resolution Aerosol Optical Depth Data. Atmos Environ (1994) 95, 581–590.
14. Li X, Zhang Q, Zhang Y, Zhang L, Wang Y, Zhang Q, Li M, Zheng Y, Geng G, Wallington TJ, Han W, Shen W, He K, 2017 Attribution of PM 2.5 exposure in Beijing–Tianjin–Hebei region to emissions: implicat ion to control strategies. Science Bulletin 62, 957–964.
15. Liang F, Xiao Q, Wang Y, Lyapustin A, Li G, Gu D, Pan X, Liu Y, 2018 MAIAC-based long-term spatiotemporal trends of PM2.5 in Beijing, China. Sci Total Environ 616-617, 1589–1598. [PubMed: 29055576]
16. Liu Y, Paciorek CJ, Koutrakis P, 2009 Estimating regional spatial and temporal variability of PM(2.5) concentrations using satellite data, meteorology, and land use information. Environ Health Perspect 117, 886–892. [PubMed: 19590678]
17. Lv B, Hu Y, Chang HH, Russell AG, Bai Y, 2016 Improving the Accuracy of Daily PM2.5 Distributions Derived from the Fusion of Ground-Level Measurements with Aerosol Optical Depth Observations, a Case Study in North China. Environ Sci Technol 50, 4752–4759. [PubMed: 27043852]
18. Lyapustin A, Wang Y, Laszlo I, Kahn R, Korkin S, Remer L, Levy R, Reid JS, 2011 Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. Journal of Geophysical Research 116.
19. Ma Z, Hu X, Huang L, Bi J, Liu Y, 2014 Estimating ground-level PM2.5 in China using satellite remote sensing. Environ Sci Technol 48, 7436–7444. [PubMed: 24901806]

20. Ma Z, Hu X, Sayer AM, Levy R, Zhang Q, Xue Y, Tong S, Bi J, Huang L, Liu Y, 2016 Satellite-Based Spatiotemporal Trends in PM2.5 Concentrations: China, 2004–2013. Environ Health Perspect 124, 184–192. [PubMed: 26220256]

21. Miller KA, Siscovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, Kaufman JD, 2007 Long-term exposure to air pollution and incidence of cardiovascular events in women. N Engl J Med 356, 447–458. [PubMed: 17267905]

22. Pope CA 3rd, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD, 2002 Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA 287, 1132–1141. [PubMed: 11879110]

23. Provençal S, Buchard V, Silva A.M.d., Ledu c R, Barrette N, Elhacham E, Wang S-H, 2017 Evaluation of PM2.5 Surface Concentrations Simulated by Version 1 of NASA's MERRA Aerosol Reanalysis over Israel and Taiwan. Aerosol and Air Quality Research 17, 253–261. [PubMed: 29670645]

24. van Donkelaar A, Martin RV, Brauer M, Kahn R, Levy R, Verduzco C, Villeneuve PJ, 2010 Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environ Health Perspect 118, 847–855. [PubMed: 20519161]

25. Wang J, 2003 Intercomparison between satellite-derived aerosol optical thickness and PM2.5mass: Implications for air quality studies. Geophysical Research Letters 30.

26. Wang LT, Wei Z, Yang J, Zhang Y, Zhang FF, Su J, Meng CC, Zhang Q, 2014 The 2013 severe haze over southern Hebei, China: model evaluation, source apportionment, and policy implications. Atmospheric Chemistry and Physics 14, 3151–3173.

27. Xiao Q, Wang Y, Chang HH, Meng X, Geng G, Lyapustin A, Liu Y, 2017 Full-coverage high-resolution daily PM 2.5 estimation using MAIAC AOD in the Yangtze River Delta of China. Remote Sensing of Environment 199, 437–446.

28. Xie Y, Wang Y, Zhang K, Dong W, Lv B, Bai Y, 2015 Daily Estimation of Ground-Level PM2.5 Concentrations over Beijing Using 3 km Resolution MODIS AOD. Environ Sci Technol 49, 12280–12288. [PubMed: 26310776]

29. Yanosky JD, Paciorek CJ, Laden F, Hart JE, Puett RC, Liao D, Suh HH, 2014 Spatio-temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors. Environ Health 13, 63. [PubMed: 25097007]

30. Yanosky JD, Paciorek CJ, Suh HH, 2009 Predicting chronic fine and coarse particulate exposures using spatiotemporal models for the Northeastern and Midwestern United States. Environ Health Perspect 117, 522–529. [PubMed: 19440489]

31. Yu W, Liu Y, Ma Z, Bi J, 2017 Improving satellite-based PM2.5 estimates in China using Gaussian processes modeling in a Bayesian hierarchical setting. Sci Rep 7, 7048. [PubMed: 28765549]

32. Zhan Y, Luo Y, Deng X, Chen H, Grieneisen ML, Shen X, Zhu L, Zhang M, 2017 Spatiotemporal prediction of continuous daily PM 2.5 concentrations across China using a spatially explicit machine learning algorithm. Atmospheric Environment 155, 129–139.

33. Zhang L, Liu L, Zhao Y, Gong S, Zhang X, Henze DK, Capps SL, Fu T-M, Zhang Q, Wang Y, 2015 Source attribution of particulate matter pollution over North China with the adjoint method. Environmental Research Letters 10, 084011.

34. Zhang Z, Laden F, Forman JP, Hart JE, 2016 Long-Term Exposure to Particulate Matter and Self-Reported Hypertension: A Prospective Analysis in the Nurses' Health Study. Environ Health Perspect 124, 1414–1420. [PubMed: 27177127]

35. Zheng Y, Zhang Q, Liu Y, Geng G, He K, 2016 Estimating ground-level PM 2.5 concentrations over three megalopolises in China using satellite-derived aerosol optical depth measurements. Atmospheric Environment 124, 232–242.

Highlights:

1. MAIAC AOD at 1-km resolution was used to predict $PM_{2.5}$ levels in North China Plain
2. A high performance machine learning model was developed directly at monthly level
3. This model can predict historical $PM_{2.5}$ with high accuracy at monthly, seasonal and annual level
4. The multiple imputation method substantially increased $PM_{2.5}$ coverage to 100%

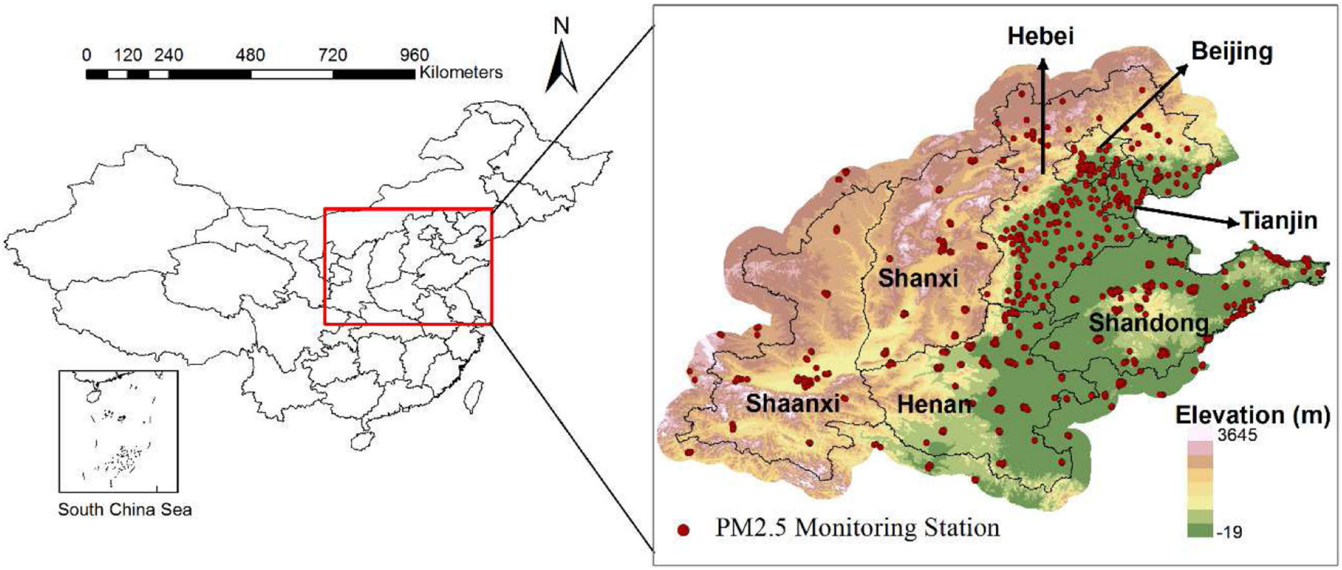


Figure 1.

Study area with a 50-km buffer, showing locations of ground $PM_{2.5}$ monitoring stations.

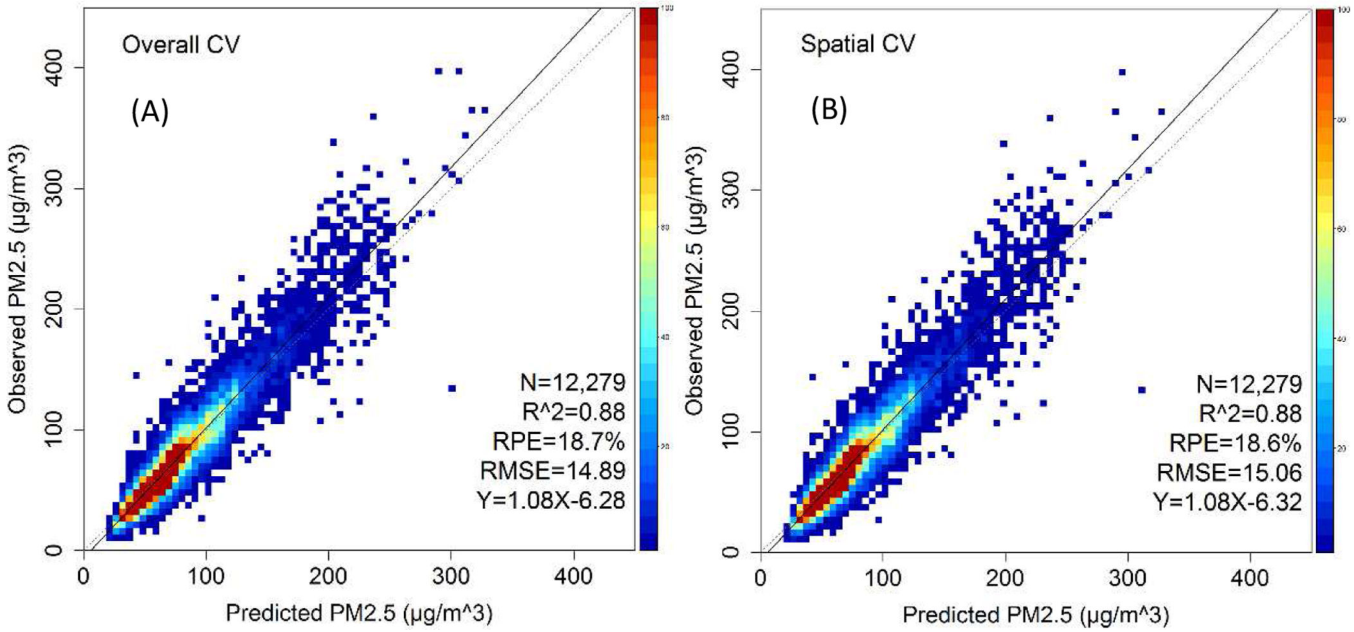


Figure 2.

10-fold cross validation. (A) Overall cross validation; (B) Spatial cross validation.

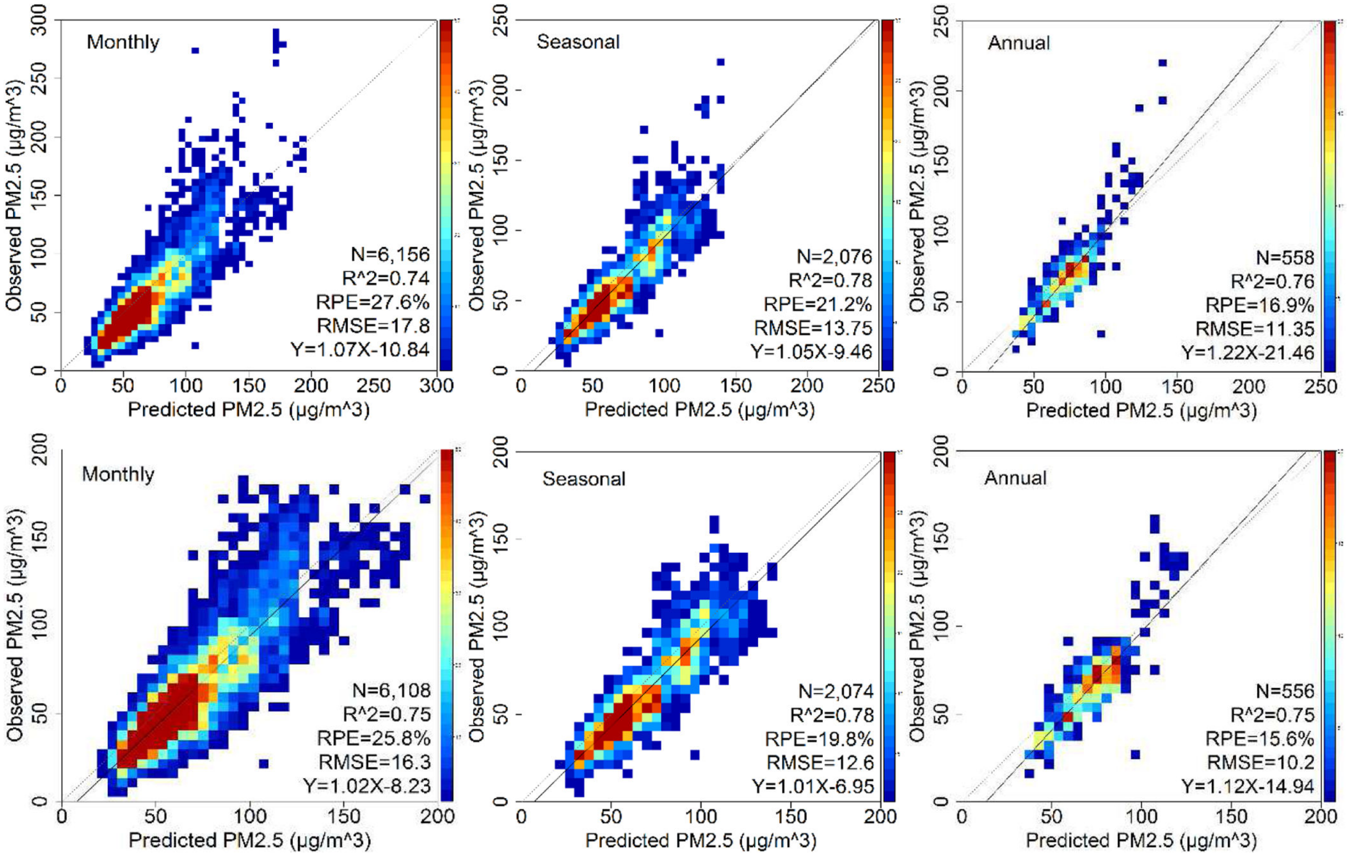


Figure 3.

Results of predicting $PM_{2.5}$ concentrations in 2016 at monthly, seasonal and annual level with models fitted from data of year 2013 to 2015. Upper panel: for the entire dataset; Lower panel: for dataset removing monthly $PM_{2.5}$ concentrations greater than 180 μg/m3.

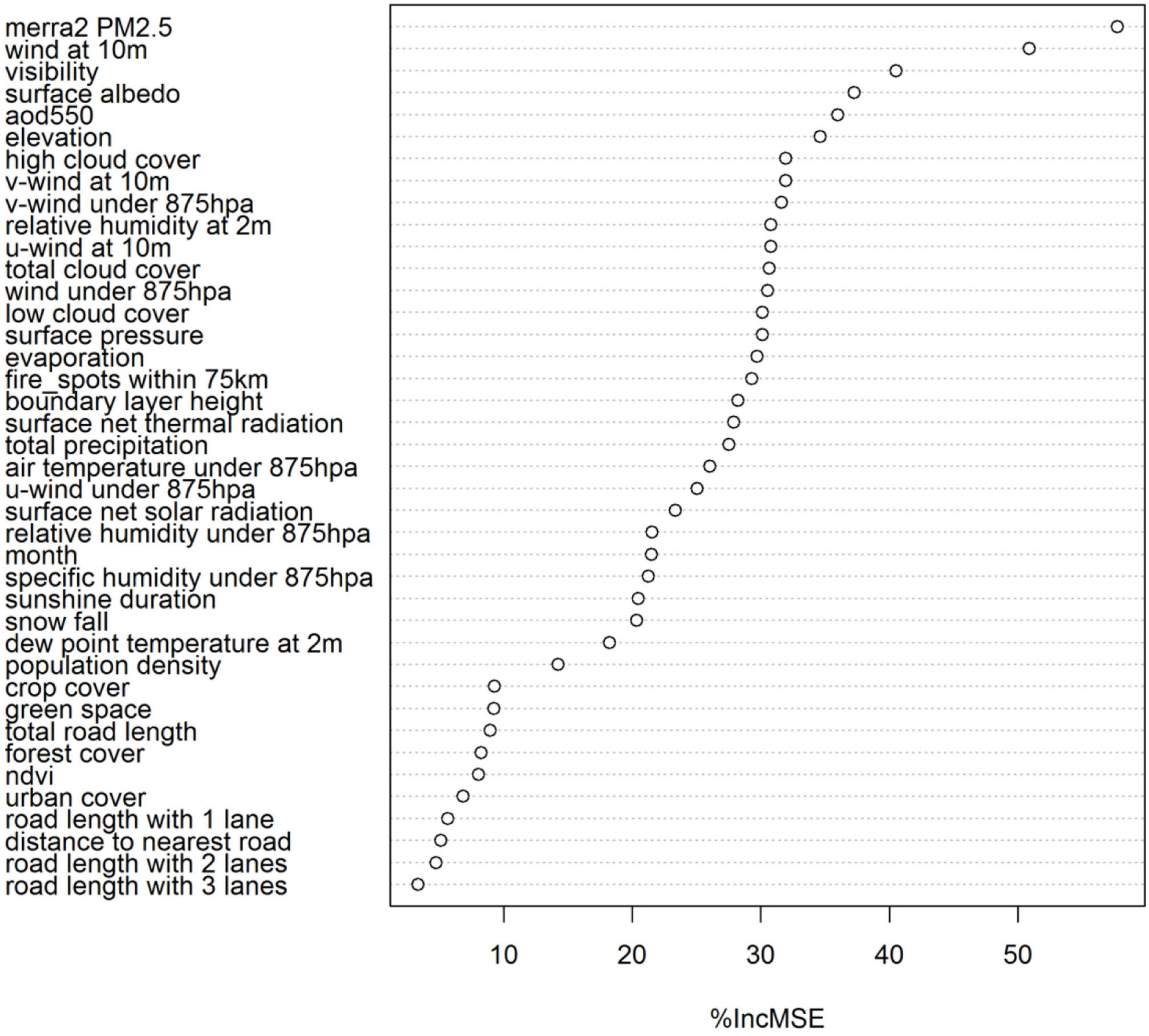


Figure 4.

Variable importance plot for the random forest model predicting the monthly $PM_{2.5}$ concentrations in North China.

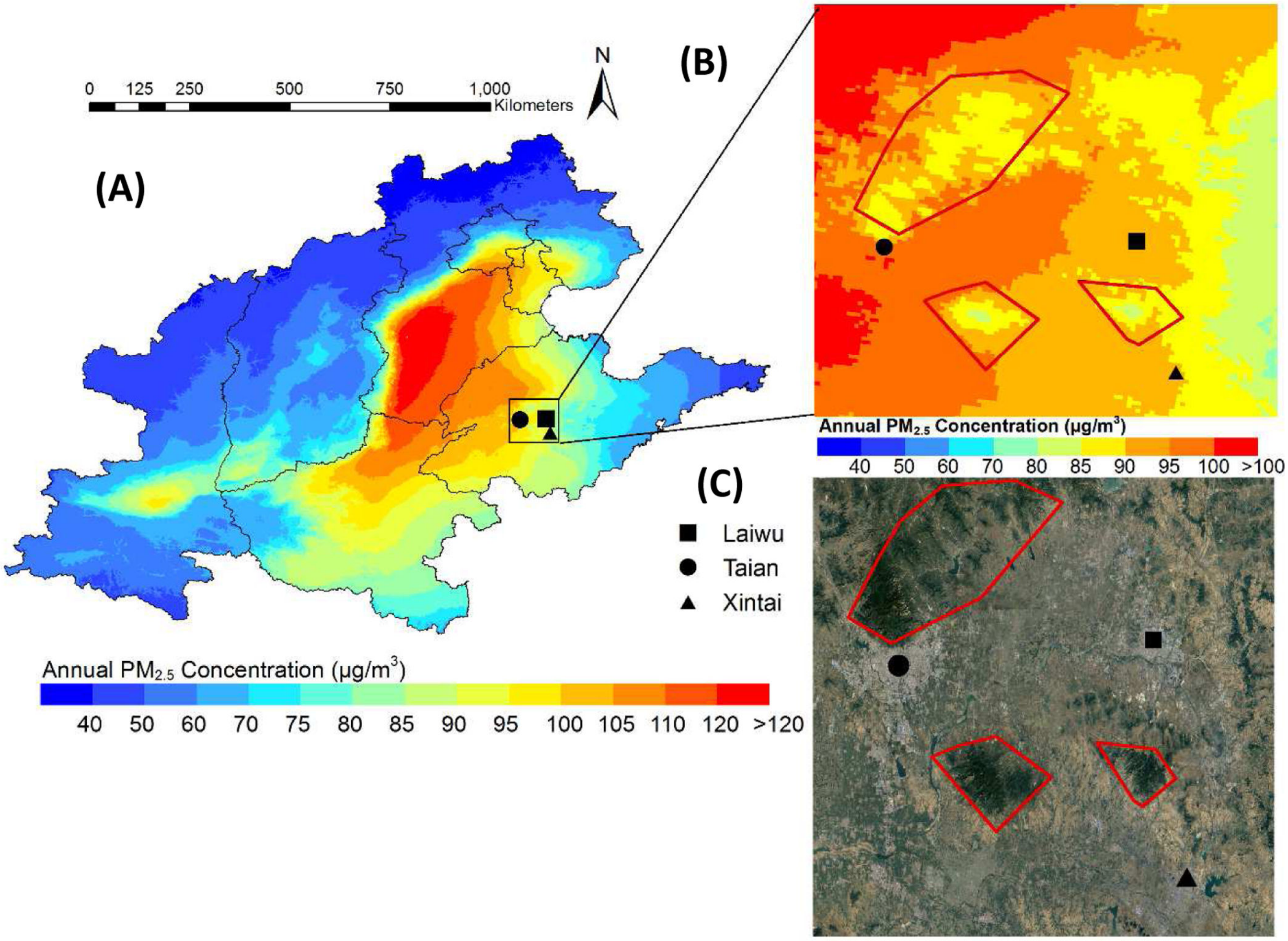


Figure 5.

PM$_{2.5}$ gradients under high resolution. A: annual PM$_{2.5}$ predictions in 2013; B: zoom in map of annual PM$_{2.5}$ predictions in Tai’an, Laiwu and Xintai City; C: satellite photo of Tai’an, Laiwu and Xintai City. Map data: Google Earth. Red polygon represented the forest cover.