Meteorological parameters and gaseous pollutant concentrations as predictors of ground-level PM2.5 concentrations in the Beijing–Tianjin–Hebei Region, China

Xinpeng Wang*, Wenbin Sun, Zhen Wang, Yahui Wang, Hongkang Ren

College of Geoscience and Surveying Engineering, China University of Mining and Technology-Beijing, D11 Xueyuan Road, Beijing 100083, People Republic of China

ABSTRACT

Ground-level PM$_{2.5}$ concentrations are severely underestimated by mixed-effects model that ignores the effects of primary pollutant emissions and secondary pollutant conversion. The model, in particular, underestimates the ground-level PM$_{2.5}$ concentrations associated with periods of heavy pollution. In this work, meteorological parameters and NO$_2$, SO$_2$, CO, and O$_3$ concentrations are introduced as predictors into a mixed-effects model to improve the estimation of PM$_{2.5}$ concentration based on Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol optical depth (AOD). The Beijing–Tianjin–Hebei (JingJinJi) Region is taken as the study area. The model provides an overall cross-validation (CV) $R^2$ of 0.84 and root-mean-square prediction error (RMSE) of 33.91 µg m$^{-3}$. The CV $R^2$ and RMSE of the proposed model are higher by 0.11 and lower by 9.16 µg m$^{-3}$, respectively, than those of the model that lacks gaseous pollutants as predictors. The $R^2$ and RMSE of the model increases and decreases by 0.14 and 13.37 µg m$^{-3}$, respectively, when PM$_{2.5}$ concentrations exceed the secondary standards set by the Ministry of Environment Protection of China (PM$_{2.5}$ > 75 µg m$^{-3}$). High PM$_{2.5}$ concentrations are associated with drastic improvements in the underestimation of PM$_{2.5}$ concentrations. The spatial distribution of PM$_{2.5}$ during periods of heavy pollution predicted by the proposed model is highly consistent with that inferred from monitoring data. Thus, the proposed model can be used to generate highly accurate maps of PM$_{2.5}$ distribution for long-term and short-term PM$_{2.5}$ exposure studies and can help reduce the misclassification of PM$_{2.5}$ exposure in heavily polluted areas.

Keywords: PM$_{2.5}$; Aerosol optical depth; Gaseous pollutant; Heavy pollution; Mixed-effects model

* Corresponding author:
Tel: +86-10-62339355
E-mail address: wangxinpeng0512@163.com
Epidemiological studies have illustrated that PM$_{2.5}$, defined as particulate matter with an aerodynamic diameter of $<2.5$ $\mu$m, has adverse effects on human health (Song et al., 2014). Long-term exposure studies have shown that the risks of ischemic heart disease (Crouse et al., 2012), lung cancer (Pope III et al., 2002), and cardiovascular mortality (Pope III et al., 2004) increase by 16%, 8%, and 10%, respectively, with every 10 µg m$^{-3}$ increment in PM$_{2.5}$ concentration. Short-term PM$_{2.5}$ exposure studies have emphasized that every 10 µg m$^{-3}$ elevation in 2 day mean PM$_{2.5}$ concentration increases the incidences of respiratory disease (Kloog et al., 2012) and myocardial infarction (Zanobetti et al., 2009) by 0.7% and 2.25%, respectively. Each 10 µg m$^{-3}$ increment in daily PM$_{2.5}$ concentration is associated with elevated hospital admission rates and increases the incidences of heart failure (Dominici et al., 2006), acute coronary syndrome (Belleudi et al., 2010), and pediatric asthma (Strickland et al., 2015) by 1.28%, 2.3%, and 1.3%, respectively. The accurate estimation of PM$_{2.5}$ concentration is a prerequisite of studies on the effects of long-term and short-term PM$_{2.5}$ exposure on human health.

The PM$_{2.5}$ exposure level of a certain population has traditionally been estimated on the basis of PM$_{2.5}$ concentration data acquired through ground monitoring over a certain distance (Laden et al., 2006). However, the sparsity and uneven spatial distribution of PM$_{2.5}$ monitoring sites will introduce uncertainty to the estimation of PM$_{2.5}$ exposure and result in the underestimation of health risks (Hu et al., 2014). Remotely sensed data for aerosol optical depth (AOD) have been
widely used to estimate PM$_{2.5}$ concentrations because of their spatial and temporal continuity (Donkelaar et al., 2011). Different statistical models, such as linear regression models (Gupta et al., 2006), generalized linear regression models (Liu et al., 2007; You et al., 2015), linear mixed models (Lee et al., 2011; Xie et al., 2015), geographically weighted regression models (Hu et al., 2014; You et al., 2016), generalized additive models (Paciorek et al., 2008; Strawa et al., 2013), and Bayesian statistical models (Chang et al., 2014; Lv et al., 2016), have been developed for the estimation of PM$_{2.5}$ concentration from satellite-derived AOD data. Nevertheless, AOD-based methods severely underestimate high PM$_{2.5}$ concentrations. To illustrate, PM$_{2.5}$ concentrations exceeding 40 µg m$^{-3}$ in the United States and 60 µg m$^{-3}$ in China are severely underestimated by these methods (Gupta and Christopher 2009a; Li et al., 2017; Liu et al., 2007; Ma et al., 2014).

The underestimation of PM$_{2.5}$ concentration will introduce uncertainty to long-term and short-term PM$_{2.5}$ exposure studies. The Beijing–Tianjin–Hebei Region (JingJinJi) has a large urban scale and human population and is characterized by high energy consumption (Wang et al., 2016). Thus, residents of this region are exposed to high concentrations of pollutant emissions. In recent years, several incidences of extremely high PM$_{2.5}$ concentrations that exceed 500 µg m$^{-3}$ and that persist for several days have been reported in the North China Plain, which is represented by the JingJinJi Region (Andersson et al., 2015). The underestimation of high PM$_{2.5}$ concentrations in heavily polluted areas will increase the exposure risk associated with PM$_{2.5}$ and result in the severe miscalculation of the effects of long-term and short-term PM$_{2.5}$ exposure on public health. Therefore, improving the accuracy of PM$_{2.5}$ estimation is crucial for reducing the misclassification of PM$_{2.5}$ exposure levels and promoting epidemiological research in heavily
Early studies used AOD as the sole predictor of surface PM$_{2.5}$ concentration. AOD is the extinction coefficient of light that originates from particle scattering over the entirety of a vertical column. PM$_{2.5}$ concentration is defined as the mass concentration of dry particles measured near the surface of the column. Thus, AOD values and PM$_{2.5}$ concentration are not strictly correlated (Chudnovsky et al., 2013; Li et al., 2015). Therefore, the AOD-based estimation of PM$_{2.5}$ concentration may be inaccurate (Saunders et al., 2014). Meteorological parameters (MET), such as wind direction, wind speed, temperature, humidity, and boundary layer height, are used as predictors to improve the accuracy of surface PM$_{2.5}$ concentration estimation (Liu et al., 2009; Paciorek et al., 2008). The characteristics of surface weather, however, are not the only determinants of air pollution development in the JingJinJi Region (Miao et al., 2015). In addition to unfavorable weather conditions, the excessive emission of primary pollutants and the secondary conversion of pollutants to particles are the main causes of heavy pollution incidences in the JingJinJi Region (Sun et al., 2014; Wang et al., 2014). Thus, additional predictors should be introduced to resolve the underestimation of high PM$_{2.5}$ concentration.

PM$_{2.5}$ is a complex mixture of particles that is primarily composed of SO$_4^{2-}$, NO$_3^-$, NH$_4^+$, elemental C, and organic C. The source emission of atmospheric pollutants and the conversion of gas pollutants are important sources of PM$_{2.5}$ (Kloog et al., 2012; Meng et al., 2014). Numerous scholars have improved the accuracy of PM$_{2.5}$ concentration estimation by introducing point emission and area-source emission data for North America into estimation models (Kloog et al., 2012; Lee et al., 2016; Strawa et al., 2013). Nevertheless, PM$_{2.5}$ concentration in North America...
is lower than that in the JingJinJi Region. Thus, the effects of source data on the estimation of PM$_{2.5}$ concentration have to be determined. The gaseous precursors of water-soluble inorganic salts include NO$_2$ and SO$_2$. These pollutants are homologous to primary pollutant emissions and can be oxidized into SO$_4^{2-}$ and NO$_3^{-}$ in the atmosphere. Atmospheric SO$_4^{2-}$ and NO$_3^{-}$, in turn, are the main sources of fine particulates in the JingJinJi Region (Zhang et al., 2007). Song et al. (2015) developed a statistical model for PM$_{2.5}$ concentration in Xi’an, China. Their model indicated that PM$_{2.5}$ concentration is strongly correlated with the concentrations of the gaseous pollutants (GASs) NO$_2$, SO$_2$, CO, and O$_3$. Thus, these pollutants can be used as auxiliary variables for PM$_{2.5}$ prediction. Zheng et al. (2016) corrected the model for the estimation of the annual average PM$_{2.5}$ concentration in the JingJinJi, Yangtze River, and Pearl River Delta Regions by introducing the values of annual average NO$_2$. Nevertheless, their approach is suitable only for the estimation of the long-term effects of PM$_{2.5}$ concentration.

The mixed-effects model proposed by Lee et al. (2011) has greatly improved the accuracy of PM$_{2.5}$ estimation by accounting for the temporal heterogeneity of PM$_{2.5}$ concentrations and AOD values. In this work, we used AOD, MET, and GASs as predictive variables in the construction of a linear mixed model to resolve the underestimation of high PM$_{2.5}$ concentration. We take the JingJinJi Region as the study area. Our approach increases the accuracy of PM$_{2.5}$ concentration estimation and reduces PM$_{2.5}$ exposure misclassification by accounting for the spatial and temporal distribution of PM$_{2.5}$ concentration. The rest of this paper is organized as follows: The study area, input datasets, and model structure are described in the Materials and Methods section. The results for model fitting, cross-validation (CV), PM$_{2.5}$ spatiotemporal distribution, and
MATERIALS AND METHODS

Study Area

The JingJinJi Region is located north of the North China Plain, east of Taihang Mountain, and south of Yanshan. Its terrain is elevated in the northwest and depressed in the southeast (Fig. 1[b]). The JingJinJi Region has an area of 218,000 square km and a population of 110,000,000. It is a core economic area in northern China. Energy consumption has gradually intensified in the region with the development of the national economy. The region has experienced several incidences of persistent heavy-pollution weather during which PM$_{2.5}$ concentrations exceeded 500 µg m$^{-3}$.

Input Data

**MODIS AOD Data**

MODIS satellite data were retrieved from Terra and Aqua, the earth observation system satellite of the National Aeronautics and Space Administration. Terra and Aqua cross the equator at approximately 10:30 and 13:30 local time, respectively. Collection 6 (C6) is the latest version of MODIS Aerosol Data in 2014. The Deep Blue algorithm (DB) of C6 has better AOD coverage and higher data quality than Collection 5.1 (Lee et al., 2016). Therefore, the AOD data collected by Aqua C6 DB over the period of January 1$^{th}$ 2014 to December 31$^{th}$ 2014 were used in this work. The data had a resolution of 10 km (MODIS parameter name:
Ground Monitoring Data

The ground monitoring data used in this work included data for PM$_{2.5}$, NO$_2$, SO$_2$, CO, and O$_3$ concentrations. The hourly ground monitoring data collected from January 1$^{\text{st}}$ 2014 to December 31$^{\text{th}}$ 2014 were retrieved from the official website of the China Environmental Monitoring Center (CEMC) (http://113.108.142.147:20035/emcpublish/) and the Beijing Environmental Monitoring Center (BJEMC) (http://zx.bjemc.com.cn/). CEMC had 80 monitoring sites during the period of January 1$^{\text{st}}$ to April 31$^{\text{th}}$ 2014. Since May 1$^{\text{st}}$, 23 monitoring sites have been added in Beijing. Thus, CMEC and BJEMC have 103 monitoring sites in 2014. The distribution map of the PM$_{2.5}$ monitoring sites is shown in Fig. 1(c).

PM$_{2.5}$, NO$_2$, SO$_2$, O$_3$, and CO data collected at 13:00 and 14:00 were extracted to match the satellite transit time and averaged as the data for PM$_{2.5}$ prediction.

Meteorological Data

Hourly MET data collected over the period of January 1$^{\text{st}}$ 2014, to December 31$^{\text{th}}$ 2014 by 171 meteorological monitoring stations (Fig. 1[b]) were retrieved from the Public Weather Service Center of China Meteorological Administration. The data included wind direction, wind speed, pressure, temperature, and humidity. Daily 13:00 and 14:00 meteorological data were extracted to match the satellite transit time and averaged as the data for PM$_{2.5}$ prediction.

Data Processing and Integration

Grid cells covering the whole JingJinJi Region were first constructed with a resolution of 10
km for data integration. The 103 PM$_{2.5}$ monitoring sites were matched to the corresponding grid cells on the basis of latitude and longitude, and the grid wherein PM$_{2.5}$ monitoring sites were located was designated as the grid monitoring site. If multiple PM$_{2.5}$ monitoring sites were located in the same grid, all monitoring site data located in that grid were averaged as the grid monitoring result. A total of 65 PM$_{2.5}$ grid monitoring sites were finally included. The value of AOD that exists over a window size of 3 × 3 grids centered on the grid monitoring site was averaged as the matching result when PM$_{2.5}$ and AOD data were matched. Meteorological monitoring data were assigned to each grid through the nearest neighbor method to represent the meteorological conditions of the grid monitoring site. NO$_2$, SO$_2$, CO, and O$_3$ were interpolated onto the 10 km grid data through the Kriging method. DEM data with a resolution of 30 m were resampled into data with a resolution of 10 km and matched with the PM$_{2.5}$ data of the grid monitoring site. In addition, the days with less than three matched data records were discarded during the data match process.

**Method**

Given the temporal heterogeneity of PM$_{2.5}$ concentrations, AOD values, and GASs concentrations, a linear mixed model was developed to account for the daily variability in the relationship of PM$_{2.5}$ with AOD and GASs. Furthermore, given the limited geographical area of the JingJinJi Region, PM$_{2.5}$–(AOD, GASs) relationships were assumed to exhibit minimal spatial variations on a given day, and spatial nonstationary variations on a regional scale were ignored (Lee et al., 2011; Zheng et al., 2016). The linear mixed model that used AOD, GASs, and MET
as predictors was designated as the AGM model and is described below:

$$\text{PM}_{i,j} = (\beta_0 + \beta_{0,i,j}) + (\beta_1 + \beta_{1,i,j})AOD_{i,j} + (\beta_2 + \beta_{2,i,j})NO2_{i,j} + (\beta_3 + \beta_{3,i,j})SO2_{i,j} + (\beta_4 +$$

$$\beta_{4,i,j})CO_{i,j} + (\beta_5 + \beta_{5,i,j})O3_{i,j} + \beta_6PRS_{i,j} + \beta_7WD_{i,j} + \beta_8WS_{i,j} + \beta_9TMP_{i,j} + \beta_{10}RH_{i,j} +$$

$$\beta_{11}ALT_{i,j} + \varepsilon_{i,j} (\beta_{0,j}\beta_{1,j}\beta_{2,j}\beta_{3,j}\beta_{4,j}\beta_{5,j}) \sim N([000000], \Psi) \tag{1}$$

where $PM_{i,j}$ is the PM$_{2.5}$ concentration at grid monitoring site $i$ on day $j$; $AOD_{i,j}$, $NO2_{i,j}$, $SO2_{i,j}$, $CO_{i,j}$, $O3_{i,j}$, $PRS_{i,j}$, $WD_{i,j}$, $WS_{i,j}$, $TMP_{i,j}$, $RH_{i,j}$, and $ALT_{i,j}$ are the AOD, NO$_2$, SO$_2$, CO, O$_3$, pressure, wind direction, wind speed, temperature, humidity, and elevation value at grid $i$ on day $j$, respectively; and $\beta_0$, $\beta_{0,i}$ are the fixed and random intercepts, respectively. $\beta_k$ ($k = 1, 2, \ldots, 5$) and the $\beta_{k,i,j}$ ($k = 1, 2, \ldots, 5$) are the fixed and day-specific random slopes for the predictors, respectively; $\varepsilon_{i,j}$ is the error term at grid $i$ on day $j$; and $\Psi$ is the variance-covariance matrix for day-specific random effects. Fixed effects represent the average effects of the predictor on PM$_{2.5}$ over the entirety of the study period, whereas random effects account for the daily variability in relationships between independent and dependent variables (Hu et al., 2014; Zheng et al., 2016).

The model was fitted with AOD and MET (AM model) parameters to assess the improvements resulting from the introduction of GASs as predictors.

**Model Validation**

A 10-fold CV based on grid monitoring sites was performed to evaluate the performance of the model with the final goal of predicting the grid PM$_{2.5}$ concentration. In verification, data from 90% of the grid monitoring sites were selected as the training data, and data from the remaining 10% of the grid monitoring sites were used as the validation data. This process was repeated 10 times, and all grid monitoring sites were tested. In addition, the spatial interpolation of GASs has
been conducted based on training data. That is, the GASs in the validation data are interpolated from training data. RMSE, coefficient of determination \( R^2 \), slope, and intercept were used to estimate the CV performance of the model.

**RESULTS**

**Descriptive Statistics**

Table 1 shows the descriptive statistics of the independent and dependent variables used in model fitting and validation. In 2014, the maximum PM\(_{2.5}\) concentration in the JingJinJi Region was 858 \( \mu g m^{-3} \) with an annual mean value of 77.3 \( \mu g m^{-3} \) and a standard deviation of 83.54. The average value and standard deviation of AOD were 0.82 and 0.87, respectively. The average PM\(_{2.5}\) concentration at each monitoring site and the annual mean AOD are shown in Fig.S2. The average PM\(_{2.5}\) concentration in the JingJinJi Region increased from the northwest to the southeast under the influence of topography and land-use type. The lowest and highest PM\(_{2.5}\) values of 29 and 139 \( \mu g m^{-3} \), respectively, were recorded at Zhangjiakou in the northwest and at Xingtai in the south, respectively. The spatial distribution of AOD in the JingJinJi Region was consistent with that of PM\(_{2.5}\). Specifically, PM\(_{2.5}\) concentrations were high in areas with high AOD values and vice versa. However, the spatial distribution of PM\(_{2.5}\) and AOD during different seasons contradicted the above situation (Fig.S2). The variable seasonal spatial distribution of PM\(_{2.5}\) and AOD indicates that the relationship between PM\(_{2.5}\) and AOD changed with time.

The results for Pearson’s correlation analysis between PM\(_{2.5}\) and independent variables are shown in Table 2. Correlation analysis revealed that NO\(_2\) and CO had the strongest correlations
with PM$_{2.5}$ concentrations with the correlation coefficients of 0.748 and 0.697, respectively. AOD and SO$_2$ had the next strongest correlations with PM$_{2.5}$ concentrations with correlation coefficients of 0.6 and 0.565, respectively. The results of correlation analysis illustrate the feasibility of introducing GASs as predictors into PM$_{2.5}$ prediction models. The multiple collinearity problems that exist among GASs were ignored given that NO$_2$, SO$_2$, CO, and O$_3$ represent different pollutant constituents (Zhang et al., 2007) and PM$_{2.5}$ concentration prediction is the focus of this study (Reid et al., 2015).

**Model Fitting and Validation**

Fig. 2 show the scatter plots for the model fitting and CV of the AGM and AM models. The CV R$^2$ of the AM and the AGM models were 0.73 and 0.84, respectively. The CV RMSE of the AM and AGM models were 43.07 and 33.91 µg m$^{-3}$, respectively. The CV R$^2$ and RMSE obtained by the models were lower and higher those obtained through model fitting, respectively. These results indicate that the proposed model was slightly overfitting. However, when GASs are introduced, the R2 of model fitting and CV increased by 0.012 and 0.11, respectively, and the RMSE of model fitting and CV decreased by 12.54 µg m$^{-3}$ and 9.16 µg m$^{-3}$, respectively. These results indicate that the performances of both models have greatly improved. In addition, the scatter plot shows that when the PM$_{2.5}$ concentrations exceeded 300 µg m$^{-3}$, the AM model severely underestimated PM$_{2.5}$ concentrations, whereas the AGM model greatly improved PM$_{2.5}$ concentration estimation. The improvement in high-value underestimation was reflected by the increase in slope from 0.74 to 0.86 and the reduction in intercept value from 20.11 to 12.15.

The performances of the AM and the AGM models in the estimation of high PM$_{2.5}$...
concentrations are compared in Table 3. When the observed PM$_{2.5}$ concentrations exceeded 0, 35, 75, and 100 µg m$^{-3}$, the R$^2$ of the AGM model was higher than that of the AM model by 0.11, 0.12, 0.14, and 0.15, respectively, whereas the RMSE of the AGM model was lower than that of the AM model by 9.16, 11.17, 13.37, 15.8 µg m$^{-3}$, respectively. High PM$_{2.5}$ concentrations were associated with high increments in R$^2$ and considerable decrements in RMSE. These associations indicate that high PM$_{2.5}$ concentrations were associated with the drastic improvement in the performance of the AGM model.

Seasonal performance statistics are shown in Table 4. The performance of the AGM model for all seasons was better than that of the AM model. The performance improvement was the greatest in heavy-pollution winter, with RMSE reduced by 15.28 µg m$^{-3}$ and R$^2$ increased by 0.14. The performance improvement was the least in summer, with RMSE reduced by 4.3 µg m$^{-3}$ and R$^2$ increased by 0.09.

**Signs and magnitudes of the model coefficients**

Table S2 and Fig. S1 show the fixed effects and daily variations of slopes and intercepts of the AGM model. The fixed coefficients of AOD and GASs are positive. However, as can be seen from Fig. S1, the slope of these variables varies daily. The slope of AOD varies the most from -41.35 to 118.41, followed by CO, which varies from -15.23 to 77.26. It should be noted that the slopes of variables are positive in most days, indicating that PM$_{2.5}$ is positively correlated with these variables most of the time. The daily variation of slope and intercept shows that the relationship between PM$_{2.5}$ and predictor vary with time, indicating the importance of temporal heterogeneity in PM$_{2.5}$ concentration prediction.
Variation in the Spatial Distribution of PM$_{2.5}$ during a Typical Heavy Pollution Period

Fig. 3 shows the variation in the spatial distribution of PM$_{2.5}$ during a typical heavy pollution period from October 6$^{th}$ 2014 to October 12$^{th}$ 2014. The spatial distribution variation of PM$_{2.5}$ during the heavy pollution process is described in Text S1. This typical heavy pollution process illustrates that the spatial distribution of PM$_{2.5}$ predicted by the AGM model was consistent with that inferred from monitoring data. Thus, the AGM model can well reflect the occurrence, diffusion, and disappearance of PM$_{2.5}$ and provides strong data support to studies on the diffusion of PM$_{2.5}$ during periods of heavy pollution.

Prediction of Seasonal and Annual Average PM$_{2.5}$ Concentrations

The spatial distribution of seasonal and annual average PM$_{2.5}$ concentration is shown in Fig. 4. The JingJinJi Region experienced the heaviest pollution during winter. The average concentration of PM$_{2.5}$ in the JingJinJi Region exceeded 120 µg m$^{-3}$ during winter. Average PM$_{2.5}$ concentrations in the southeast of Handan and east of Shijiazhuang, Xingtai, and Handan were higher than those in other areas. In addition to industrial emissions and motor vehicle exhaust emissions, residential coal-fired heating contributes to the high PM$_{2.5}$ concentrations in these areas. During autumn, heavily polluted areas were mainly located in the south of Beijing; southeast of Baoding; central Shijiazhuang, Xingtai, and Handan; and other areas. Average PM$_{2.5}$ concentrations were 90–120 µg m$^{-3}$ and may exceed 120 µg m$^{-3}$ in some areas. The lowest average PM$_{2.5}$ concentration in the JingJinJi Region was recorded during summer and was less than 75 µg m$^{-3}$ in most areas. This value meets the secondary PM$_{2.5}$ concentration standard set by the Environmental Protection Department. The average PM$_{2.5}$ concentration in most areas of
Zhangjiakou was less than 30 µg m\(^{-3}\), which meets the primary PM\(_{2.5}\) concentration standard set by the Environmental Protection Department.

The average annual PM\(_{2.5}\) concentration in the entire JingJinJi Region exceeded the primary PM\(_{2.5}\) concentration standard. Ten areas failed to meet the secondary PM\(_{2.5}\) concentration standard. PM\(_{2.5}\) concentrations in Baoding, Shijiazhuang, Xingtai, Handan, and other areas may even exceed 90 µg m\(^{-3}\). Therefore, the overall environmental situation in the JingJinJi Region is not optimistic.

**DISCUSSION**

In this study, gaseous pollutant data are introduced into the mixed effects model to improve the underestimation of PM\(_{2.5}\) in JingJinJi based on satellite remote sensing and meteorological data. There could be many causes for the underestimation of PM\(_{2.5}\). One of the reasons may be that there are fewer matching data at high concentrations, as discussed by Liu et al. (Liu *et al.*, 2005). While the study of Gupta and Christopher has shown that a small range of high PM\(_{2.5}\) mass concentration corresponds to a large range of AOT values may resulting in an underestimation of high PM\(_{2.5}\) concentrations (Gupta and Christopher 2009b). For further analysis, the correlation statistics between PM\(_{2.5}\) and predictors under different PM\(_{2.5}\) concentrations are shown in Table S1. The correlation between PM\(_{2.5}\) and AOD decreases as the concentration of PM\(_{2.5}\) increases. Therefore, AOD cannot adequately represent the correlation between high PM\(_{2.5}\) concentrations and independent variables (Gupta and Christopher 2009b; Liu *et al.*, 2005), resulting in an underestimation of high PM\(_{2.5}\) concentration.
The development of heavy pollution in the JingJinJi Region is closely related to atmospheric pollutant concentrations, atmospheric oxidation, and MET (Wang et al., 2015). Atmospheric pollutants include primary pollutant emissions and converted secondary pollutants. Primary pollutants mainly originate from coal fires and traffic emissions, and GASs such as NO₂ and SO₂ have homology with the primary pollutants, which can represent primary pollutant emissions to a certain extent (Wang et al., 2014). Secondary pollutants are mainly derived from secondary inorganic salts (SO₄²⁻ and NO₃⁻) in the particulate state (Sun et al., 2014). As an important component of PM₂.₅, SO₄²⁻ and NO₃⁻ are mainly oxidized by SO₂ and NO₂ in the atmosphere. The oxidation efficiency of SO₂ and NO₂ is related to the atmospheric oxidant Ox (NO₂+O₃). As an indicator of atmospheric oxidation capacity, a high Ox concentration can promote the secondary conversion of SO₂ and NO₂ to SO₄²⁻ and NO₃⁻, thereby increasing PM₂.₅ concentrations. We can draw the conclusion that NO₂, SO₂, CO, and O₃ not only represent the emission of primary pollutants in the process of heavy pollution formation, but they are also the precursor and oxidant of secondary pollutant transformation. In addition, GASs have a good correlation with high PM₂.₅ concentration as shown in Table S1, which shows that they can more adequately represent the correlation between high PM₂.₅ concentrations and independent variables. And the results in Table 3 have shown that high PM₂.₅ concentrations were associated with high increments in R² and considerable decrements in RMSE. Therefore, they can be used as auxiliary variables to resolve the underestimation of high PM₂.₅ concentration and improve the accuracy of AOD-based PM₂.₅ estimation.

Comparing the RMSE differences of the AGM and AM models for each season revealed that
the performance of the AGM model exhibited the greatest improvement in winter with an RMSE difference of 15.28 µg m$^{-3}$. Nevertheless, the AGM model showed negligible improvements in its performance in the prediction of summer PM$_{2.5}$ concentrations and presented a RMSE difference of 4.3 µg m$^{-3}$. During winter, the increased prevalence of coal-fire heating increases the emission of PM$_{2.5}$ and its precursors, such as NO$_2$ and SO$_2$. The formation of fine particulates through the secondary conversion of NO$_2$ and SO$_2$ to SO$_4^{2-}$ and NO$_3^-$ is the main cause of heavy pollution in the JingJinJi Region during winter (Wang et al., 2016). The introduction of NO$_2$ and other GASs as independent variables enabled the AGM model to fully reflect the correlation between high PM$_{2.5}$ concentration and independent variables during winter and improved the performance of the AGM model in the estimation of winter PM$_{2.5}$ concentration.

By introducing GASs, the AGM model gives higher CV $R^2$ (0.84) than previously published studies in JingJinJi, e.g., the mixed effects model for Beijing ($R^2=0.79$) (Xie et al., 2015) and JingJinJi ($R^2=0.77$) (Zheng et al., 2016), the Bayesian model for northern China ($R^2=0.68$) (Lv et al., 2016). However, the RMSE in this paper is higher than that in these studies, which may be due to the reason that the concentration of PM$_{2.5}$ was lower in previous studies. The small range of PM$_{2.5}$ concentration indicates that the concentration of PM$_{2.5}$ is relatively concentrated, which results in the smaller RMSE value. To compare with the most commonly used machine learning models, we also experimented with these models and the results are shown in Table 5. The performance of all models has been greatly improved by introducing GASs, which shows the feasibility of introducing GASs as predictors. After introducing GASs, gradient boosting decision tree (GBDT) performs the best among the three machine learning models, with CV $R^2$ and RMSE
being 0.79 and 38.08µg m$^{-3}$, respectively. However, the $R^2$ and RMSE are lower by 0.05 and higher by 6.49µg m$^{-3}$, respectively, than AGM model, which may be due to GBDT ignoring the temporal heterogeneity between PM$_{2.5}$ and predictors.

Although our model demonstrated drastically improved PM$_{2.5}$ prediction performance, it continued to exhibit a certain degree of underestimation because of the following reasons: First, we used Kriging interpolation to obtain the spatial distribution of GASs in the JingJinJi Region; interpolation, however, is associated with a certain degree of uncertainty (Oliver and Webster 1990). Second, a complex nonlinear relationship exists between high PM$_{2.5}$ concentrations and AOD (Liu et al., 2005; Ma et al., 2014; Zheng et al., 2016). The linear mixed model presupposes that a linear relationship exists between PM$_{2.5}$ and predictive variables (Lee et al., 2011). This assumption, however, will oversimplify the complex relationship between PM$_{2.5}$ and predictors (Reid et al., 2015). Furthermore, in addition to meteorological conditions, primary pollutant discharge, secondary pollutant conversion, dust, and building dust are sources of PM$_{2.5}$ in the JingJinJi Region. However, we ignored the effects of these factors on the development of heavy pollution. Finally, we only predicted PM$_{2.5}$ concentrations for available gridded AOD values. Nevertheless, AOD data are often missing because of the effects of clouds, high surface reflection, and high PM$_{2.5}$ concentrations (Levy et al., 2010; Tao et al., 2012). If the missing AOD values are related to high PM$_{2.5}$ concentrations, PM$_{2.5}$ concentration may be underestimated if only the retrieved AOD fields are used (Lv et al., 2016).

**CONCLUSION**
AOD and MET cannot fully represent the correlation between high PM$_{2.5}$ concentrations and predictive variables. Thus, estimation models that use AOD and MET as predictors may seriously underestimate PM$_{2.5}$ concentration. We improved the accuracy of AOD-based PM$_{2.5}$ prediction by introducing MET parameters and NO$_2$, SO$_2$, CO, and O$_3$ concentrations as predictors into a mixed-effects model. Our model provided 10-fold CV $R^2$ and RMSE values of 0.84 and 33.91 µg m$^{-3}$, respectively. The maps of PM$_{2.5}$ concentration during heavy-pollution weather and seasons predicted by our model are consistent with those inferred from monitoring data. Our model can well reflect the appearance, diffusion, and disappearance of PM$_{2.5}$ during periods of heavy pollution. Therefore, the introduction of GASs as independent variables into the prediction model can improve the accuracy of predicting the spatial distribution of PM$_{2.5}$. Our proposed model can be used to generate highly accurate maps of PM$_{2.5}$ distribution for long-term and short-term PM$_{2.5}$ exposures studies and can help reduce PM$_{2.5}$ exposure misclassification in heavily polluted areas.

ACKNOWLEDGEMENTS

This study was supported by the National Natural Science Foundation of China (41671383).

REFERENCE


Belleudi, V., Faustini, A., Stafoggia, M., Cattani, G., Marconi, A., Perucci, C. A. and Forastiere,


integrated surface, satellite, and meteorological products: Multiple regression approach.

*Journal of Geophysical Research Atmospheres* 114:-.


Using aerosol optical thickness to predict ground-level PM2.5 concentrations in the St. Louis area: A comparison between MISR and MODIS. *Remote Sens. Environ.* 107:33-44.


associations between GOES aerosol optical depth retrievals and ground-level PM2.5. *Environ.

(2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air

Pope III, C. A., Burnett, R. T., Thurston, G. D., Thun, M. J., Calle, E. E., Krewski, D. and
Godleski, J. J. (2004). Cardiovascular mortality and long-term exposure to particulate air
pollution: epidemiological evidence of general pathophysiological pathways of disease.
*Circulation* 109:71-77.

M. and Balmes, J. R. (2015). Spatiotemporal prediction of fine particulate matter during the
49:3887-3896.


regression model for regional PM 2.5 estimation over the Pearl River Delta region in China.


Table 1. Descriptive statistics of dependent variable and independent variables (N=12955).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Units</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Median</th>
<th>Std.Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM$_{2.5}$</td>
<td>µg m$^{-3}$</td>
<td>1</td>
<td>858</td>
<td>77.3</td>
<td>49</td>
<td>83.54</td>
</tr>
<tr>
<td>AOD</td>
<td>unitless</td>
<td>0.014</td>
<td>3.5</td>
<td>0.82</td>
<td>0.51</td>
<td>0.87</td>
</tr>
<tr>
<td>NO$_2$</td>
<td>µg m$^{-3}$</td>
<td>1</td>
<td>253</td>
<td>35.06</td>
<td>25</td>
<td>30.82</td>
</tr>
<tr>
<td>SO$_2$</td>
<td>µg m$^{-3}$</td>
<td>1</td>
<td>596</td>
<td>42.2</td>
<td>24</td>
<td>51.49</td>
</tr>
<tr>
<td>O$_3$</td>
<td>µg m$^{-3}$</td>
<td>2</td>
<td>362</td>
<td>97.05</td>
<td>83</td>
<td>58.59</td>
</tr>
<tr>
<td>CO</td>
<td>mg m$^{-3}$</td>
<td>0.0015</td>
<td>18.42</td>
<td>1.15</td>
<td>0.86</td>
<td>1.08</td>
</tr>
<tr>
<td>PRS</td>
<td>hPa</td>
<td>875.1</td>
<td>1038.5</td>
<td>1002.85</td>
<td>1008.4</td>
<td>26.92</td>
</tr>
<tr>
<td>WD</td>
<td>º</td>
<td>0.6</td>
<td>358</td>
<td>164.51</td>
<td>158.8</td>
<td>69.64</td>
</tr>
<tr>
<td>WS</td>
<td>m/s</td>
<td>0.1</td>
<td>9.4</td>
<td>1.61</td>
<td>1.3</td>
<td>1</td>
</tr>
<tr>
<td>TMP</td>
<td>ºC</td>
<td>-17.5</td>
<td>32.8</td>
<td>12.09</td>
<td>13.5</td>
<td>11.41</td>
</tr>
<tr>
<td>RH</td>
<td>%</td>
<td>9</td>
<td>99.9</td>
<td>55.73</td>
<td>56.9</td>
<td>19.84</td>
</tr>
<tr>
<td>ALT</td>
<td>m</td>
<td>7</td>
<td>908</td>
<td>124.7</td>
<td>39</td>
<td>207.06</td>
</tr>
</tbody>
</table>
Table 2. Variable correlation statistics (N=12955).

<table>
<thead>
<tr>
<th></th>
<th>PM2.5</th>
<th>NO2</th>
<th>SO2</th>
<th>O3</th>
<th>CO</th>
<th>AOD</th>
<th>PRS</th>
<th>WD</th>
<th>WS</th>
<th>TMP</th>
<th>RH</th>
<th>ALT</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM2.5</td>
<td>1</td>
<td>0.748</td>
<td>0.565</td>
<td>-0.12</td>
<td>0.697</td>
<td>0.6</td>
<td>0.177</td>
<td>-0.08</td>
<td>-0.07</td>
<td>-0.07</td>
<td>0.358</td>
<td>-0.15</td>
</tr>
<tr>
<td>NO2</td>
<td>1</td>
<td>0.683</td>
<td>-0.4</td>
<td>0.748</td>
<td>0.313</td>
<td>0.224</td>
<td>-0.08</td>
<td>-0.11</td>
<td>-0.29</td>
<td>-0.222</td>
<td>-0.14</td>
<td></td>
</tr>
<tr>
<td>SO2</td>
<td>1</td>
<td>-0.32</td>
<td>0.705</td>
<td>0.249</td>
<td>0.088</td>
<td>-0.01</td>
<td>-0.03</td>
<td>-0.34</td>
<td>0.126</td>
<td>-0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O3</td>
<td>1</td>
<td>-0.25</td>
<td>0.124</td>
<td>-0.24</td>
<td>-0.09</td>
<td>0.03</td>
<td>0.763</td>
<td>0.104</td>
<td>0.007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO</td>
<td>1</td>
<td>0.358</td>
<td>0.16</td>
<td>-0.03</td>
<td>-0.04</td>
<td>-0.24</td>
<td>0.242</td>
<td>-0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AOD</td>
<td>1</td>
<td>0.121</td>
<td>-0.133</td>
<td>-0.072</td>
<td>0.193</td>
<td>0.524</td>
<td>-0.186</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRS</td>
<td>1</td>
<td>0.063</td>
<td>-0.18</td>
<td>-0.01</td>
<td>0.116</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WD</td>
<td>1</td>
<td>0.33</td>
<td>-0.163</td>
<td>-0.282</td>
<td>0.116</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WS</td>
<td>1</td>
<td>-0.031</td>
<td>-0.389</td>
<td>-0.035</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TMP</td>
<td>1</td>
<td>0.247</td>
<td>-0.107</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RH</td>
<td>1</td>
<td>-0.111</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALT</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Comparison of the performances of the two models under different PM$_{2.5}$ concentrations.

<table>
<thead>
<tr>
<th>Observed PM$_{2.5}$ (µg m$^{-3}$)</th>
<th>AM_RMSE</th>
<th>AGM_RMSE</th>
<th>AM_R$^2$</th>
<th>AGM_R$^2$</th>
<th>RMSE_Residual</th>
<th>R$^2$_Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;0</td>
<td>43.07</td>
<td>33.91</td>
<td>0.73</td>
<td>0.84</td>
<td>-9.16</td>
<td>0.11</td>
</tr>
<tr>
<td>&gt;35</td>
<td>52.95</td>
<td>41.78</td>
<td>0.65</td>
<td>0.77</td>
<td>-11.17</td>
<td>0.12</td>
</tr>
<tr>
<td>&gt;75</td>
<td>62.63</td>
<td>49.26</td>
<td>0.55</td>
<td>0.69</td>
<td>-13.37</td>
<td>0.14</td>
</tr>
<tr>
<td>&gt;100</td>
<td>70.07</td>
<td>54.27</td>
<td>0.49</td>
<td>0.64</td>
<td>-15.8</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Note: AM_RMSE and AGM_RMSE are the RMSE of the AM and AGM models, respectively; AM_R$^2$ and AGM_R$^2$ are the R$^2$ of the AM and AGM models, respectively.
Table 4. Performance statistics of the two models in different seasons.

<table>
<thead>
<tr>
<th>Season</th>
<th>AGM_RMSE</th>
<th>AM_RMSE</th>
<th>AGM_R²</th>
<th>AM_R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring</td>
<td>26.44</td>
<td>31.33</td>
<td>0.81</td>
<td>0.74</td>
</tr>
<tr>
<td>Summer</td>
<td>24.03</td>
<td>28.33</td>
<td>0.76</td>
<td>0.67</td>
</tr>
<tr>
<td>Autumn</td>
<td>33.62</td>
<td>41.92</td>
<td>0.88</td>
<td>0.81</td>
</tr>
<tr>
<td>Winter</td>
<td>46.21</td>
<td>61.49</td>
<td>0.81</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Note: AM_RMSE and AGM_RMSE are the RMSE of AM and AGM models, respectively; AM_R² and AGM_R² are the R² of AM and AGM models, respectively.
Table 5. Performance comparison of the AGM with machine learning models.

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2_{gas}$</th>
<th>$R^2_{nogas}$</th>
<th>RMSE$_{gas}$</th>
<th>RMSE$_{nogas}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0.77</td>
<td>0.63</td>
<td>39.54</td>
<td>50.7</td>
</tr>
<tr>
<td>GBDT</td>
<td>0.79</td>
<td>0.65</td>
<td>38.08</td>
<td>49.56</td>
</tr>
<tr>
<td>DNN</td>
<td>0.79</td>
<td>0.65</td>
<td>38.32</td>
<td>49.7</td>
</tr>
<tr>
<td>This Study</td>
<td>0.84</td>
<td>0.73</td>
<td>33.91</td>
<td>43.07</td>
</tr>
</tbody>
</table>

*RF: Random Forest Model; GBDT: Gradient Boosting Decision Tree Model; DNN: Deep Neural Network Model.*

$R^2_{gas}$ and $R^2_{nogas}$ represent the $R^2$ of models with or without GAS, respectively. RMSE$_{gas}$ and RMSE$_{nogas}$ are the RMSE of models with or without GAS, respectively.
Figure captions

**Fig. 1.** Study area and the locations of PM$_{2.5}$ monitoring sites and meteorological stations. (a) Location of JingJinJi in China. (b) Locations of meteorological stations and elevation of JingJinJi. (c) Locations of PM$_{2.5}$ monitoring sites.

**Fig. 2.** Results of model fitting and CV. The dash line is the 1:1 line as a reference. (a)(c) are fitting results of the AM and AGM models, respectively. (b)(d) are CV results of the AM and AGM models, respectively.

**Fig. 3.** Grid site monitoring value and prediction results of the AGM model for a period of heavy pollution from October 6$^{th}$ to 12$^{th}$ 2014. (a1)–(g1) Prediction results of the AGM model. (a2)–(g2) Grid site monitoring results. Given the lack of AOD data for October 12$^{th}$ and the similarity between site monitoring data for October 12$^{th}$ and October 13$^{th}$, the data from October 13$^{th}$ was used instead of the October 12$^{th}$ data for prediction.

**Fig. 4.** Comparisons of seasonal and annual average predicted and grid monitoring PM$_{2.5}$ concentrations. (a)–(e) Prediction results of the AGM model. (f)–(j) Grid site monitoring results.
Fig. 1.

![Map of China with highlighted regions](image1.png)
Fig. 2.

(a) Predicted = 0.76 x Observed + 18.89
R² = 0.77
RMSE = 40.4
N = 12955

(b) Predicted = 0.74 x Observed + 20.11
R² = 0.73
RMSE = 43.07
N = 12955

(c) Predicted = 0.88 x Observed + 9.32
R² = 0.89
RMSE = 27.86
N = 12955

(d) Predicted = 0.86 x Observed + 12.15
R² = 0.84
RMSE = 33.91
N = 12955
Fig. 3.
Fig. 4.

(a) Spring  (b) Summer  (c) Autumn  (d) Winter

(e) Annual PM$_{2.5}$ ($\mu$g/m$^3$)
- <30
- 30 - 45
- 45 - 60
- 60 - 75
- 75 - 90
- 90 - 105
- 105 - 120
- 120 - 135
- 135 - 150
- 150 - 165
- 165 - 180
- >180

(f) Spring  (g) Summer  (h) Autumn  (i) Winter

(j) Annual PM$_{2.5}$ ($\mu$g/m$^3$)
- <30
- 30 - 45
- 45 - 60
- 60 - 75
- 75 - 90
- 90 - 105
- 105 - 120
- 120 - 135
- 135 - 150
- 150 - 165
- 165 - 180
- >180