Environmental Pollution 249 (2019) 735–749

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

Predicting ground-level $PM_{2.5}$ concentrations in the Beijing-Tianjin-Hebei region: A hybrid remote sensing and machine learning approach☆

Xintong Li, Xiaodong Zhang*

School of Environmental Science and Engineering, Shandong University, Qingdao, Shandong, 266237, China

ARTICLE INFO

Article history:
Received 27 November 2018
Received in revised form
13 February 2019
Accepted 17 March 2019
Available online 22 March 2019

Keywords:
Remote sensing
Aerosol optical depth
Machine learning
$PM_{2.5}$
Random forest

ABSTRACT

An accurate estimation of $PM_{2.5}$ (fine particulate matters with diameters ≤ 2.5 μm) concentration is critical for health risk assessment and generating air pollution control strategies. In this study, a hybrid remote sensing and machine learning approach, named RSRF model is proposed to estimate daily ground-level $PM_{2.5}$ concentrations, which integrates Random Forest (RF), one of machine learning (ML) models, and aerosol optical depth (AOD), one of remote sensing (RS) products. The proposed RSRF model provides an opportunity for an adequate characterization of real-time spatiotemporal $PM_{2.5}$ distributions at uninhabited places and complex surfaces. It also offers advantages in handling complicated non-linear relationships among a large number of meteorological, environmental and air pollutant factors, as well as ever-increasing environmental data sets. The applicability of the proposed RSRF model is tested in the Beijing-Tianjin-Hebei region (BTH region) during 2015–2017. Deep Blue (DB) AOD from Aqua-retrieved Collection 6.1 (C_61) aerosol products of Moderate Resolution Imaging Spectroradiometer (MODIS) is validated with Aerosol Robotic Network. The validation results indicate C_61 DB AOD has a high correlation with ground based AOD in the BTH region. The proposed RSRF model performed well in characterizing spatiotemporal variations of annual and seasonal $PM_{2.5}$ concentrations. It not only is useful to quantify the relationships between $PM_{2.5}$ and relevant factors such as DB AOD, meteorological and air pollutant variables, but also can provide decision support for air pollution control at a regional environment during haze periods.

1. Introduction

Fine particulate matters, also called $PM_{2.5}$ whose aerodynamic equivalent diameter is less than 2.5 μm, has been proved to be associated with many cardiovascular and respiratory diseases and even does damage to nervous systems (Kioumourtzoglou et al., 2016; Madrigano et al., 2013; Xing et al., 2016). Particularly in China, with rapid economic development and urbanization, the $PM_{2.5}$ pollution is the main concern of air pollution at present. A major cause for deteriorating air quality in urban agglomeration is due to the extensive anthropogenic and industrial emissions. An accurate prediction of $PM_{2.5}$ concentrations is essential for environmental and public health risk analysis. Although more than 1,400 ground-based air monitoring stations have been established to provide hourly averaged $PM_{2.5}$ concentration data in China, these measurements based on points are not continuous. They are unequally distributed on a spatial scale to capture variations in the study of air pollution analysis, especially for the stations located at uninhabited places and complex surfaces (Chu et al., 2016).

With the development of aviation and aerospace technology, the ability to monitor environmental quality has been enhanced, such as satellites which have a comprehensive spatiotemporal coverage. Combining with aerosol optical depth (AOD), a satellite product which can quantify the extinction effect of particulate matter suspended in the atmosphere, the capability to estimate ground-level $PM_{2.5}$ concentration from ground monitoring networks has been expanded. AOD has been increasingly used as an important predictor for $PM_{2.5}$ concentration estimation locally or globally (Beloconi et al., 2016; Bilal et al., 2017; Chen et al., 2017a; Li et al., 2015; Ma et al., 2014; Péré et al., 2009; Xie et al., 2015; Zhang and Li, 2015). The AOD products from Moderate Resolution Imaging

☆ This paper has been recommended for acceptance by Dr. Haidong Kan.

* Corresponding author.
E-mail address: xdzhang@sdu.edu.cn (X. Zhang).

https://doi.org/10.1016/j.envpol.2019.03.068
0269-7491/

Spectroradiometer (MODIS) launched by NASA in 1999 and 2002 are mostly used. MODIS AOD products can provide us with credible long-term monitoring records for atmospheric studies. Their retrieval algorithms are continuously updated. As a result, MODIS AOD is more widely considered in predicting concentrations of particulate matter. Atmosphere is a complex system and the extinction effect is affected by multiple substances such as air molecules, solid and liquid particles. It is desired to consider many other auxiliary environmental data such as emitted air pollutant concentrations and weather conditions as supplementary data to help predict the $PM_{2.5}$ concentrations.

Previously, various mathematical models have been developed for predicting the spatiotemporal distributions of $PM_{2.5}$, including chemical transport models and statistical models (Chu et al., 2016). The chemical transport models based on the physical-chemical mechanisms of reaction, transport, and deposition processes in the atmosphere can simulate the concentration and species composition of fine particulate particles. Their basic idea is to estimate $PM_{2.5}$ concentrations through multiplying AOD by local scaling factors derived from a global atmospheric chemistry model (Liu et al., 2004a; van Donkelaar et al., 2006). This method could predict $PM_{2.5}$ concentrations in the regions without historical observations. However, a large uncertainty may be brought to the prediction results by uncertain initial boundary conditions and uncertain model structures (Bergin et al., 1999; Mallet and Sportisse, 2006). For example, the complex environmental physical-chemical processes could not be completely expressed in chemical transport models. A lack of reliable air pollution emission inventory data would also bring difficulties in simulation. More and more historical data have been accumulated owing to the gradually expanding air monitoring networks. That provides convenience for estimating $PM_{2.5}$ concentrations by developing statistical methods with high computational efficiency and simple modeling principles. Linear regression models such as multiple linear regression models have been developed to build the relationship between $PM_{2.5}$ and AOD through incorporation of multiple covariates such as weather and social variables (Wang and Christopher, 2003; Liu et al., 2005). However, the relationship between $PM_{2.5}$ and AOD is not just linear. More recently, advanced statistical models have been proposed to study the relationship between $PM_{2.5}$ and satellite AOD such as generalized linear model (Liu et al., 2007), land use regression model (Hystad et al., 2011), geographically weighted regression (Ma et al., 2014), mixed-effected model (Kloog et al., 2011). The increasing availability of big spatial data has the potential to improve the accuracies of $PM_{2.5}$ concentration assessment models. That also leads to complex relationships among predication variables in big data sets such as multicollinearity. Since statistical models based on probability theory and multivariate statistical analyses need a series of assumptions in advance, their extension capability is limited. Due to increasing complexity, these statistical models could encounter difficulties in handling data sets with a large number of predictors, especially when these predictors were dependent.

Machine Learning (ML) is interdisciplinary involving statistics, data science, and computing which has been widely used in many applications (Alpaydin, 2009). The ML methods which build models from a data point of view have strengths in handling complicated non-linear relationships among a large number of environmental data sets and do not need to address many of the assumptions required for statistical models such as sample normality, homoscedasticity, multicollinearity, independence and other strict parametric assumptions (Grange et al., 2018). Many ML methods have been proposed to predict concentrations of air pollutants, such as neural networks (NN) (Arhami et al., 2013; Gupta and Christopher, 2009; Pérez et al., 2000; Voukantsis et al., 2011; Zou et al., 2015), support vector machines (SVM) (Liu et al., 2017; Lu and Wang, 2005; Suárez-Sánchez et al., 2011; Wang et al., 2017; Wang et al., 2008), decision tree (DT) (Reid et al., 2015; Zhan et al., 2017), and ensemble methods. Some hybrid ML methods combined with principal component analysis (PCA) (Voukantsis et al., 2011), genetic algorithm (GA) (Antanasijević et al., 2013) were also developed. In addition, some studies also applied a stepwise-cluster analysis technology whose essence is a classification tree in the sense of probability, to predict environmental variables to overcome the sophisticated structure of NN (Fan et al., 2015; Li et al., 2015; Li et al., 2016). Among ML methods, Random Forest (RF) is an ensemble decision tree approach which shows less overfitting. There are two advantages of RF: randomness and ensemble. RF gives a chance to every predictor to appear with different randomly selected covariates in different context, and can thus better reflect the potentially complex effect of predictors on the projection (Strobl et al., 2009). Besides this, the randomness of input samples has an effect of denoising. The ensemble strategy where the RF is consisted by irrelevant and randomly constructed trees can reduce overfitting due to the ability to assimilate predictions from those tree classifiers. A robust projection result can be provided by RF. The number of decision trees can be adjusted to reduce the training time based on required accuracy and computing resources (Hu et al., 2017), which show an effective approach to tune training time without too much performance reduction. Compared with SVR with a specific kernel function and sophisticated structure of NN, the tree structure of a RF model is easy to understand since it is similar to the bisection method. Although the interpretability of random forest which is consist by many decision trees is weaker than that of only a decision tree, it can avoid overfitting. In addition, an RF model can provide users with statistical metrics such as variable importance to investigate the strength of relationships between projections and various predictors during the training process (Hastie et al., 2009). Different RF models were developed in the United States to estimate concentrations of $PM_{2.5}$ (Hu et al., 2017; Brokamp et al., 2018; Liu et al., 2018). A meteorological normalization technique utilizing RF models was proposed for Swiss PM_{10} trend-analysis (Grange et al., 2018). Zhan et al. (2018a, 2018b) used RF methods to estimate concentrations of O_3 and NO_2 in China. However, there are few studies on development of an RF model to estimate daily $PM_{2.5}$ concentrations in China.

Therefore, the objective of this study is to propose a hybrid remote sensing and random forest method, named RSRF to predict the spatiotemporal distributions of daily $PM_{2.5}$ concentrations across the BTH region. The proposed RSRF method entails: a) combining the dynamic and wide monitoring ability of AOD and the advantage of ML method in handling complicated non-linear relationships among large environmental data sets to estimate daily $PM_{2.5}$ concentrations; b) using the variable importance to help analyze the variable contributions in the formation of $PM_{2.5}$; c) comparing the performance of RSRF to other ML methods. The proposed RSRF model is based on AOD retrievals of MODIS Aqua satellite with relevant covariates such as weather variables and air pollutants. Variable importance is employed to evaluate the effects of each predictor on prediction of the $PM_{2.5}$ concentrations. The proposed RSRF method is compared with other ML methods including Multiple Linear Regression (MLR), Multivariate Adaptive Regression Splines (MARS) and Support Vector Regression (SVR). Using the developed RSRF model, variations of the spatiotemporal pollution patterns of $PM_{2.5}$ erupting during heavy haze pollution periods are analyzed. The results are helpful to provide decision support for air pollution control at a regional environment during the haze periods.

2. Materials and methods

2.1. Overview of the study area

A severe haze pollution which main component is fine particulate particles would lead to a series of economic, environmental and health problems such as adverse effects on human health, and economic losses from air and ground traffic accidents due to reduced visibility. The BTH region (Fig. 1) is China's political and cultural center, which consists of Beijing, Tianjin, and 11 prefecture-level cities. It is an important core area of North China's economy, suffering from the most serious haze pollution due to the high aerosol loadings from both local and regional sources. Developing an accurate model for prediction of $PM_{2.5}$ concentrations in the BTH region is essential for air pollution control, regional health risk assessment and economic development. In this study, the study area is divided into 72 × 73 grids, with approximately 0.1° × 0.1° of every grid.

2.2. Data collection

2.2.1. $PM_{2.5}$ and other air pollutant measurements

The hourly ground-level $PM_{2.5}$ observations at 78 monitoring sites throughout the BTH region during 2015–2017 were collected from China National Environmental Monitoring Center (http://106.37.208.233:20035/). These monitoring sites are relatively evenly distributed in the entire BTH region covering urbans, suburbs and mountains but are sparse in the northern mountains and southern plain (Fig. 1). According to the Chinese National Ambient Air Quality Standard and Environmental Protection Standard (MEP, 2012; MEP, 2013a; MEP, 2013b), the measurement methods of $PM_{2.5}$ and other air pollutants are showed in Table S1. Air pollutants are important precursor pollutants for the formation of fine particulate particles, and thus selected as predictor variables. A reason of selecting ground monitoring air pollution data instead of satellite monitoring data as predictors is that the pollution concentration estimated by satellite is an eventual distribution of entire atmospheric layer and cannot represent the distribution near ground. Besides this, the ground monitoring data have a high precision. The measured ground value is an average of a previous hour. For example, the value at 13 o'clock is an averaged value of measurements during 12:00–13:00. As a correspondence to the satellite passing local time (13:30), the hourly monitoring data were averaged from 13 o'clock to 15 o'clock. If any monitoring value was lacked due to an instrument failure, the rest hourly monitoring values would be averaged as the corresponding monitoring value. If all three monitoring values were lacked, the sample would be neglected.

2.2.2. MODIS AOD

There are two satellite products of MODIS AOD, including one from Terra and the other from Aqua. Terra has more severe sensor aging than Aqua, such as the sensor aging in the blue band. Multiple MODIS products such as AOD based on dark target algorithm over land and Ångström exponent over the ocean show a global underestimation (Lyapustin et al., 2014). There have been some efforts trying to reduce those negative effects from collection 5 to collection 6.1. To minimize the uncertainties brought by this sensor aging as much as possible, only Aqua AOD was used in this study.

According to the differences of resolution and retrieval algorithms, MODIS AOD products are divided into four products, including 10 km Dark Target (DT) retrievals, 10 km Deep Blue (DB) retrievals, 10 km DB/DT combined retrievals and 3 km DT retrievals. According to the study of Chen (Chen et al., 2017b), 10 km DB product in collection 6 (C06) was more suitable to considerable retrievals under all air quality conditions and performed better than 10 km DT retrievals in Beijing. The study of Tao et al. (2015) indicated AOD retrieved by dark target method was overestimated and usually missed regional haze pollution in China and DB retrievals obviously had a higher accuracy than DT retrievals in northern China. Under a condition of low AOD load, DB AOD has a wider space coverage and more accurate than DT AOD (Sayer et al., 2014). The DT at 3 km has a lower reliability than MODIS C06 AOD products at 10 km due to the problem of pixel screening method (Nichol and Bilal, 2016). Moreover, the enhanced DB method adopted by C06 is suitable for not only bright areas but also dark targets, covering a wider monitoring range than previous DB

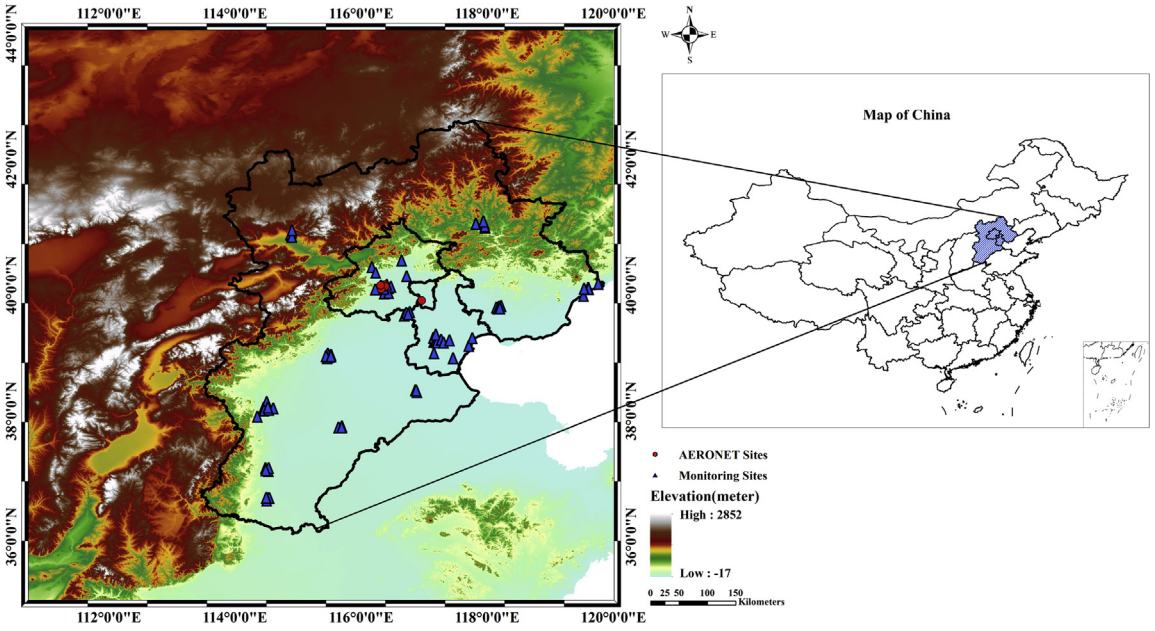

Fig. 1. Topographic map of the BTH region (thick black lines), highlighting the locations of 5 AERONET sites (red circles) and 78 monitoring sites (blue triangles). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

versions. Collection 6.1 (C061) released in late 2017 is an improvement of C06. Although C061 reduces the effects of heterogeneous terrain and the uncertainty in elevated terrain, it results in a slight decrease in spatial coverage compared to C06 (Hsu, 2017). We made a statistic of days with valid AOD for eastern China. The result (Fig. S1) indicates the number of days with valid AOD values in the BTH region for a year slightly increases in C6.1, especially for Beijing and the mountainous region of northeast BTH region. The C061 AOD can provide enough AOD data series more than 120 days a year for most BTH region. Thus, level 2 MODIS DB aerosol products (hereinafter referred to as MODIS AOD) at 10 km with best estimate (Quality Assurance, QA = 2, 3) of Aqua (MYD_04_L2_C061) from January 1st, 2015 to December 31st, 2017 were downloaded from The Level-1 and Atmosphere Archive & Distribution System (LAADS). AOD is a variable depending on wavelength and the AOD of 550 nm channel is used in this study.

The spatial resolution of MODIS AOD is 10×10 km^2 at nadir. In the following processing, for convenience all the datasets were resampled into $0.1^\circ \times 0.1^\circ$. To reduce the abnormal values, the AOD is averaged by window sizes of 1×1 (win1), 3×3 (win3), and 5×5 (win5) centered at the monitoring sites. A missing of the AOD values due to clouds or poor quality of retrievals leads to a large uncertainty for pixels around it. Based on the consideration of the noise level introduced by spatial averaging, the average of a ground monitoring site was done only when the corresponding position of the center site had valid AOD values in satellite products.

2.2.3. Meteorological fields

The meteorological data from January 1st, 2015 to December 31st, 2017 were obtained from National Centers for Environmental Prediction/National Center for Atmospheric Research Reanalysis Project (NCEP/NCAR Reanalysis 1) (Kalnay et al., 1996) provided by the NOAA (https://www.esrl.noaa.gov/psd/). The input data sets were selected based on their presumed influences on $PM_{2.5}$. Meteorological conditions are critical in estimating $PM_{2.5}$ concentrations, because it can affect the formation and transport process of air pollutants (Laakso et al., 2003). Meteorological variables used in this study included air temperature (AT), relative humidity (RH), wind speed (WS), wind direction (WD) and pressure (P). Detailed information of meteorological variables is shown in Table S2. The spatial resolution was also resampled into $0.1^\circ \times 0.1^\circ$ in the next process considering its coarse resolution to keep consistent with other datasets. The daily reanalysis data is available four times a day at 0Z, 6Z, 12Z, and 18Z, based on Coordinated Universal Time (UTC) (Kalnay et al., 1996). These variables are instantaneous values at the reference time. The reanalysis data at 6Z were extracted using the positions of 78 air monitoring sites because 6Z was the closest time to the satellite passing time.

2.2.4. Aerosol Robotic Network

Aerosol Robotic Network (AERONET) is a ground-based aerosol network which can provide a long-term and readily accessible public domain database of aerosol optical (Giles et al., 2018). Due to the low uncertainty of approximately 0.01–0.02 in AOD (wavelength dependent) (Holben, 2001), AERONET is a widely used product to validate satellite AOD retrievals. Higher quality ground AOD data with a stricter quality control are available in Version 3 databases, where more accurate temperature characterization and automatic cloud screening are adopted (Giles et al., 2018). There are three data quality levels: Level 1.0 (unscreened), Level 1.5 (cloud-screened and quality controlled), and Level 2.0 (quality-assured). Level 2.0 datasets are recommended to be used for scientific research. Quality-assured Level 2.0 utilizing Level 1.5 cloud screened and quality-controlled data set from deployment can be obtained after pre- and pro-field calibration. A lot of studies applied level 1.5 dataset as true ground monitoring value to evaluate satellite AOD. Xiao et al. (2016) used both level 2.0 and level 1.5 AOD datasets as the ground truth value to evaluation the quality of MODIS C6 AOD of east Asia including BTH region, their preliminary results indicate that the level 1.5 daily average AOD values agreed well with the level 2.0 data, with a slope of 1.0 and zero intercept. Huang et al. (2016) also used level 1.5 data to as a supplement to make validation of satellite AOD. In this study, to maintain the quantities of data pairs, both Level 1.5 and Level 2.0 datasets of 5 AERONET sites over the BTH region were employed. Our primary results indicate that the quality of level 2.0 is indeed stricter than that of level 1.5 but their difference is small. Their locations and surface types are listed in Table S3.

Since AERONET AOD measurements do not include the 550 nm channel, AOD at 550 nm needs to be interpolated using Ångström Exponent $\alpha_{440-675\ nm}$ and AOD at 440 nm and 675 nm channels provided in the AERONET data sets. Ångström Exponent (Ångström, 1964) is used to describe the dependency of the aerosol optical thickness, or aerosol extinction coefficient on wavelength, and can be expressed as follows (Liu et al., 2004b):

$$\alpha_{\lambda_1-\lambda_3} = -\frac{\ln\left(\frac{\tau_{\lambda_1}}{\tau_{\lambda_3}}\right)}{\ln\left(\frac{\lambda_1}{\lambda_3}\right)} \quad (1)$$

where $\alpha_{\lambda_1\sim\lambda_3}$ is the Ångström Exponent of wavelength λ_1 and λ_3, τ_{λ_i} is the aerosol optical depth at wavelength λ_i (in this study, $\lambda_1 = 440$ nm, $\lambda_2 = 550$ nm, and $\lambda_3 = 675$ nm). Any AOD of wavelength between λ_1 and λ_3 can be interpolated straightforwardly (Liu et al., 2004b). Deduced from Eq. (1), the AOD at 550 nm can be interpolated as follows:

$$\tau_{\lambda_2} = \tau_{\lambda_3} * \frac{\lambda_2^{-\alpha_{\lambda_1-\lambda_3}}}{\lambda_3} \quad (2)$$

As AOD changes very small (<0.01) in this wavelength range, the uncertainty introduced by this interpolation can be neglected (Sayer et al., 2014). The errors in the interpolation vary from 0% to 10% corresponding to different aerosol types (Yan et al., 2015).

AERONET AOD data are acquired at 15-min intervals on the average. On time scale, statistics of the AERONET measurements were calculated for a 1-h time window centered on satellite overpass time (approximately 13:30 local time) and at least two AERONET measurements to be available during this period.

2.2.5. Data integration and normalization

The input variables were divided into two sets, including air pollutants and weather variables. AOD was included in the set of air pollutants. Before training, the data were resampled into $0.1^\circ \times 0.1^\circ$ grids, and normalized to help improve both the convergence speed and the accuracy of the model. The frequency distributions of variables are shown in Fig. S2, most of which follow a skewed distribution or normal distribution. Specially, the distributions of some meteorological variables such as AT, P and WD, follow a bimodal distribution which is a mixed distribution consisting of multiple normal distributions due to their seasonal variations. In this study, the Z-score standardization was used. Standardization of the input data has insignificant effects on the RSRF model since there is no analogue of a regression coefficient that is affected by variable measurement scales. Actually, any algorithm based on recursive partitioning such as decision trees and regression trees does not require inputs (features) to be normalized, since it is invariant to monotonic transformations of the features. However, for the sake of convenience for comparisons among different

models, all the input data for every model were standardized.

2.3. Model development

2.3.1. Hybrid RSRF model for predicting $PM_{2.5}$ concentrations

Random Forest (RF) is an ensemble machine learning method consisting of many individual decision trees growing from bagged data and its prediction is a vote result of those trees (i.e. an average of those trees for a regression problem). There are many various decision tree algorithms such as Iterative Dichotomiser 3 (ID3), C4.5, C5.0 and Classification and Regression Trees (CART). A CART tree has a structure like a tree to divide the samples into two groups in which the samples have similar features (see Fig. 2). The training data at node can be represented by Θ. The internal node θ is also called splitting node to partition the input samples into two sets $\Theta_L(\theta)$ and $\Theta_R(\theta)$ subsets which are sent to the appropriate child nodes. Terminal nodes called leaf nodes are predictions which have no child nodes. The criterion for determining the best split θ^* at node n is based on max information gain. In CART trees, Gini impurity $G(\Theta,\theta)$ is used to represent the idea of information gain (De'ath and Fabricius, 2000):

$$\theta^* = \mathrm{arg}min_\theta G(\Theta,\theta) \tag{3}$$

In a regression tree whose target is a continuous value, Gini impurity is calculated by impurity function $H(\Theta)$ as follows (Pedregosa et al., 2011):

$$G(\Theta,\theta) = \frac{n_L}{N_n} H(\Theta_L(\theta)) + \frac{n_R}{N_n} H(\Theta_R(\theta)) \tag{4}$$

where N_n is the number of samples at node n, n_L and n_R are the numbers of samples at left child node and right child node, respectively. Mean squared error which is equal to variance reduction is used as criteria for determining locations for future splits (Pedregosa et al., 2011).

$$H(\Theta_n) = \frac{1}{N_n} \sum_{i \in N_n} (y_i - c_n)^2 \tag{5}$$

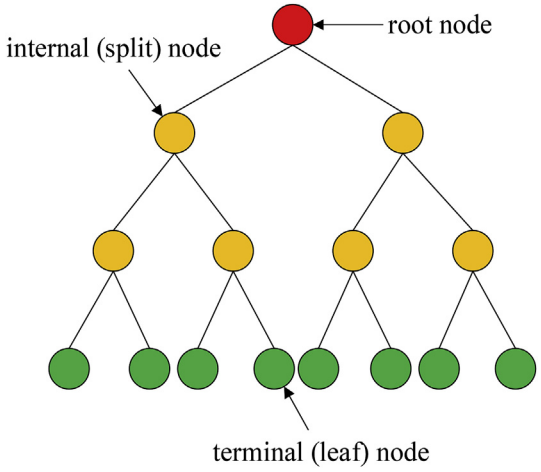

Fig. 2. A common structure of a binary decision tree. The red circle is root node, and there is only one root node in a decision tree. Yellow circles are internal nodes used to make splitting. Green circles are terminal nodes, also called leaf nodes which are the prediction results. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

$$c_n = \frac{1}{N_n} \sum_{i \in N_n} y_i \tag{6}$$

where c_n is mean value of terminal nodes. When the node purity is achieved or the tree depth is the maximum allowed tree depth D or there are not enough data used to split, this splitting is stopped (Grange et al., 2018; Pedregosa et al., 2011).

A random forest is an integration of many trees. Randomness of RF refers to randomly sampling observations with replacement (bootstrapping) from the training set along with sampling of predictors during splitting (Breiman, 1996). Datasets resampled with a same size of original datasets used to build trees are called bootstrap datasets. The combination process of taking bootstrap sample sets and aggregating the models trained on the basis of each bootstrap sample set is called bagging (abbreviation of bootstrap aggregating) (Strobl et al., 2009). Sampling training data set randomly is one of the most popular ways to add randomness into trees during the training phase. Another way is to make a non-repeated random selection of features to split each node. That means the number of randomly preselected splitting predictors can be unequal to the total number of predictors. This can reveal more interactional relationships among variables which would be neglected when a powerful predictive variable exists (Strobl et al., 2009).

In general, there are two main steps in growing a RF with k trees. The first step is to generate k new training sets S_k with replacement from the original training set. Every new training set has the same size with original training set. Then kth tree is grown on the training set S_k using non-repeated random feature selection to produce a predictor $h(\boldsymbol{x}, S_k)$ where $\boldsymbol{x}$ is an input dataset. Eventually, k predictors are averaged to obtain the final prediction. In the repeating processing, k trees grown from k data sets contain different observations and predictors. This reduce the correlation among trees and improve generalization ability of model. The combination of bagging process and ensemble predictions make RF to produce a robust estimation result.

In the developed RSRF model, the dynamic and wide monitoring ability of AOD and the advantages of RF method in handling complicated non-linear relationships and supplements of meteorological and air contaminant variables are integrated into a general framework to estimate daily $PM_{2.5}$ concentrations. The Python package named Scikit-learn (Pedregosa et al., 2011) was used to train the RSRF model. The metrics used to evaluate regression quality are goodness of fit (R^2) and root-mean square error (RMSE). The 10-fold cross-validation was used to adjust model parameters to avoid overfitting. The developed RSRF model was compared with other three ML models including multiple linear regression (MLR), multivariate adaptive regression splines (MARS) and support vector regression (SVR).

2.3.2. Variable importance

A variable's importance is a metric of strength of the relationships between predictors and projections. The relative importance of each input parameter from the trained RSRF model was calculated. It helps to analyze the importance of the effects of input variables on $PM_{2.5}$ pollution. For the SVR model with a nonlinear kernel function, it is difficult to evaluate the importance of variables. That is since we don't know the concrete form of nonlinear mapping function, and the weight vector ω cannot be computed directly. Random Forest method can easily rank variable importance in the training process utilizing the out of bag samples to estimate errors to construct a variable importance measure which has been applied to help choose important features in other studies (Genuer et al., 2010).

3. Results and discussion

3.1. AOD validation and window size selection

According to the study of Ichoku et al. (2002), applying single MODIS pixel values directly to ground point measurements is incongruous. To ensure the accuracy of AOD and evaluate the effects of windows sizes on parameter statistics, three window sizes including win1 (1 pixel × 1 pixel), win3 (3 pixel × 3 pixel) and win5 (5 pixel × 5 pixel) were tested. The square window (not circle or other shapes) is attribute to the pixel feature of satellite data. Histograms (Fig. S3) showed that both MODIS and AERONET AOD measurements exhibited similar right-skewed, monomodal distributions which agreed with the previous studies. The distribution trend of win5 is similar to that of AERONET.

The preliminary statistics information is shown in Table S4. The mean value of AERONET is 0.367 with a range of 0.020–2.479. The mean value and value range of all three window sizes of MODIS are larger than those of AERONET. Those mean MODIS AOD is a bit overestimated compared with AERONET.

The uncertainties of AERONET and MODIS AOD were evaluated using Expected Error (EE). As shown in Table 1 and Fig. 3, more than 70% of samples fall within the interval of EE, indicating that the MODIS AOD is well correlated with AERONET AOD and less than 30% of samples with large uncertainties. The fact that about 20% of samples fall above EE interval indicates the existence of a little overestimation in MODIS. Less samples falling below the EE interval suggests less underestimation. Compared with other similar studies (Chen et al., 2017b; Tao et al., 2015), in the BTH region, the underestimation trend in DB retrievals of C6.1 may be smaller than the previous versions but the overestimation still remains. However, this uncertainty brought by AOD can inevitably propagate into the final results.

Linear regression analysis was also applied to compare the differences among different window sizes (Fig. S4). Generally, there were slight differences among the three size windows in this study region. The larger the window size, the more obvious smoothing effect for noise, especially for AOD in the range of 0.8–1.2 in Fig. S3. This effect could also be reflected by gradually smaller standard deviations in Table S4. With the increase of a window size, the mean and minimum of MODIS AOD increased, while maximum of MODIS AOD decreased. For high AOD values, a larger window size may suppress the noises, but introduce unexpected errors due to spatial or aerosol type heterogeneity. That meant selection of a window size for the smaller regions should be more cautious. The commonly used window size for the validation of 10 km MODIS AOD product is 5 pixels (i.e. 50 km × 50 km). This was deduced from visually estimated speed of aerosol fronts from the animated daily sequences of Total Ozone Mapping Spectrometer (TOMS) aerosol index images (Ichoku et al., 2002). This estimated speed is at a global scale. For a regional study, it is necessary to decide the averaging windows size according to the validation. In this study, based on a consideration of the effects of window sizes and convenience to compare our results with others, the averaging window of all MODIS AOD in the rest of this paper was based on the 5 × 5 subset grid boxes.

3.2. Model performance

3.2.1. Results of model validation

The model inputs consist of data sets for AOD, SO_2, NO_2, CO, O_3, AT, RH, WS, P and WD. The total number of the matching samples of all variables at 78 monitor sites are 25807. Annual and seasonal regression analyses were conducted and their results were showed in Tables 2 and 3. Results of model fitting vs. monitoring data are showed in Fig. 4.

For the RSRF model, the features are always random permuted at each split. As a result, the best-found split may vary even with the same training data and same number of features used to find the best split, but it had very tiny effects on the predictions (Pedregosa et al., 2011). The changes in this study occurred after two decimal points. The parameter selection of RSRF is set to n_eatimator = 500, min_samples_split = 20, max_features = 0.8. We used an exhaustive searching method, named grid-search to determine parameters' values within the ranges recommended by other studies based on two performance metrices of R^2 and RMSE. To realize randomly selecting of features to reveal more relationships among predictors and projection, the max_features is set to 0.8 (eighty percent of the features are used in a model). The results showed that the overall R^2 had a relatively high value of 0.933, and RMSE was 16.315 $\mu g/m^3$ in the BTH region during 2015–2017. Using 10-fold cross-validation (CV), R^2 was reduced to 0.843, and RMSE was increased to 25.320 $\mu g/m^3$. That indicated a good agreement among training estimates and CV estimates. When examined by seasons (Table 3), the model performed well across time with R^2 of 0.885–0.944 and RMSE of 13.924–19.197 $\mu g/m^3$. The prediction accuracy of the RSRF model varied by seasons. For example, the prediction ability of the RSRF model in summer was relatively weaker than in other three seasons. This could be attributed to less samples in summer which were almost half of the other seasons.

The RSRF model was compared with MLR, MARS and SVR models. The modeling results are shown in Tables 2 and 3 and Fig. 4. The RSRF, MARS and SVR models have higher prediction accuracies than the MLR model. The overall prediction accuracy of the RSRF model is superior to those of other three models. For example,

Table 1
Validation summary of different window size satellite AOD based on AERONET.

Year	Win_size	Pearson's r	mean	RMSE	<-EE	-EE~+EE	>+EE	N
2015	1*1	0.955	0.420	0.169	5.67%	74.50%	19.83%	353
	3*3	0.960	0.431	0.163	1.70%	74.79%	23.51%	353
	5*5	0.960	0.440	0.165	1.98%	71.67%	26.35%	353
2016	1*1	0.941	0.429	0.174	4.25%	76.22%	19.53%	471
	3*3	0.950	0.433	0.159	3.82%	77.28%	18.90%	471
	5*5	0.953	0.440	0.153	4.03%	73.89%	22.08%	471
2017	1*1	0.941	0.353	0.153	10.12%	80.00%	9.88%	405
	3*3	0.941	0.370	0.151	7.41%	80.25%	12.35%	405
	5*5	0.934	0.418	0.158	6.17%	76.79%	17.04%	405
all year	1*1	0.944	0.401	0.166	6.59%	76.97%	16.44%	1229
	3*3	0.950	0.412	0.158	4.39%	77.54%	18.06%	1229
	5*5	0.960	0.421	0.158	4.15%	74.21%	21.64%	1229

Notes: The statistical metrics for evaluation include the number of MODIS/AERONET matchups (N), Pearson's correlation coefficient, the root mean square error (RMSE), and the fraction of MODIS/AERONET in agreement within the DB algorithm's expected uncertainty EE = ± (0.05 + 0.20τ_A).

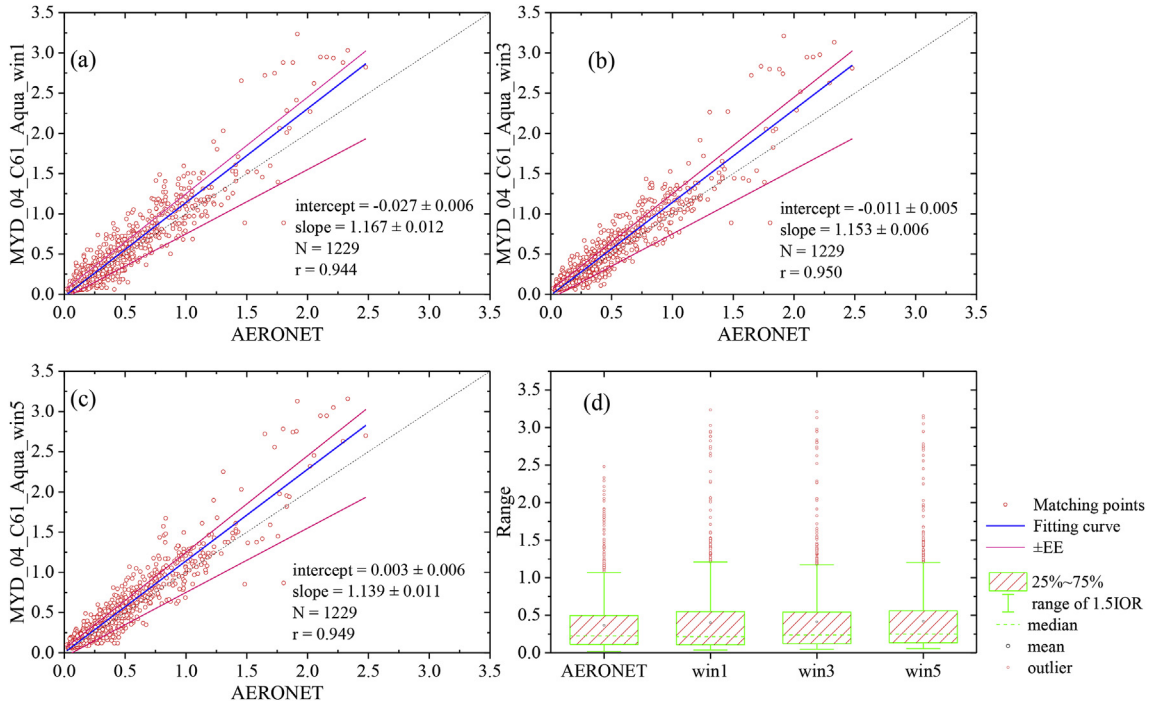

Fig. 3. Scatterplots of AERONET versus MODIS AOD.

Table 2
Annual prediction accuracies of RSRF, MLR, MARS, and SVR models.

Year		2015	2016	2017	All year
MLR	R^2	0.738/0.736	0.77/0.77	0.677/0.685	0.728/0.733
	RMSE (μg/m^3)	33.870/34.761	31.312/32.927	30.639/30.429	32.802/33.016
MARS	R^2	0.793/0.782	0.840/0.747	0.818/0.773	0.792/0.776
	RMSE (μg/m^3)	30.079/31.545	26.941/33.777	23.028/25.284	28.682/30.180
SVR	R^2	0.923/0.833	0.942/0.859	0.942/0.877	0.914/0.850
	RMSE (μg/m^3)	18.336/27.667	16.183/25.689	12.990/18.987	18.481/24.745
RSRF	R^2	0.932/0.835	0.938/0.859	0.941/0.872	0.933/0.843
	RMSE (μg/m^3)	17.331/27.516	16.744/25.657	12.682/19.378	16.315/25.320
N		8219	8399	9189	25807

Table 3
Seasonal prediction accuracies of RSRF, MLR, MARS, and SVR models.

		Spring	Summer	Autumn	Winter
MLR	R^2	0.654/0.646	0.608/0.591	0.750/0.751	0.803/0.799
	RMSE (μg/m^3)	31.185/31.744	19.372/20.106	27.072/27.207	36.048/36.837
MARS	R^2	0.795/0.754	0.719/0.654	0.836/0.805	0.845/0.805
	RMSE (μg/m^3)	24.021/26.445	16.411/18.447	21.954/24.070	31.941/35.789
SVR	R^2	0.924/0.849	0.912/0.697	0.947/0.856	0.948/0.883
	RMSE (μg/m^3)	14.582/20.737	9.159/17.289	12.414/20.671	18.576/28.024
RSRF	R^2	0.932/0.835	0.885/0.707	0.938/0.843	0.944/0.867
	RMSE (μg/m^3)	13.924/21.683	10.500/17.003	13.500/21.596	19.197/29.914
N		7874	3151	6241	8541

Notes: The training and cross-validation results are separated by a slash. For the instability of some models whose R^2 is negative during the process of 10-fold cross-validation, this value is marked by an underline. N is the sample size.

compared to the MLR model, the R^2 value of the RSRF model increased from 0.728 to 0.933, and its RMSE decreased from 32.802 μg/m^3 to 16.315 μg/m^3. MARS is not a robust method in this study because of negative R^2 appeared in the process of cross-validation for annual estimation in 2017 and seasonal estimation of spring. The differences between the results of the RSRF and the SVR models is tiny. The SVR model performed overfitting in summer and the RSRF model performed good in seasonal and annual simulations without too much overfitting. The results of the RSRF model agreed well with ground results as shown in Fig. 4. In addition, comparisons of residual error distributions were showed in Fig. S5. Both of range and standard deviation of residual errors of the RSRF model are smaller than those of other three models (Table S5). That indicated the estimation of $PM_{2.5}$ concentrations from the RSRF model is more precise.

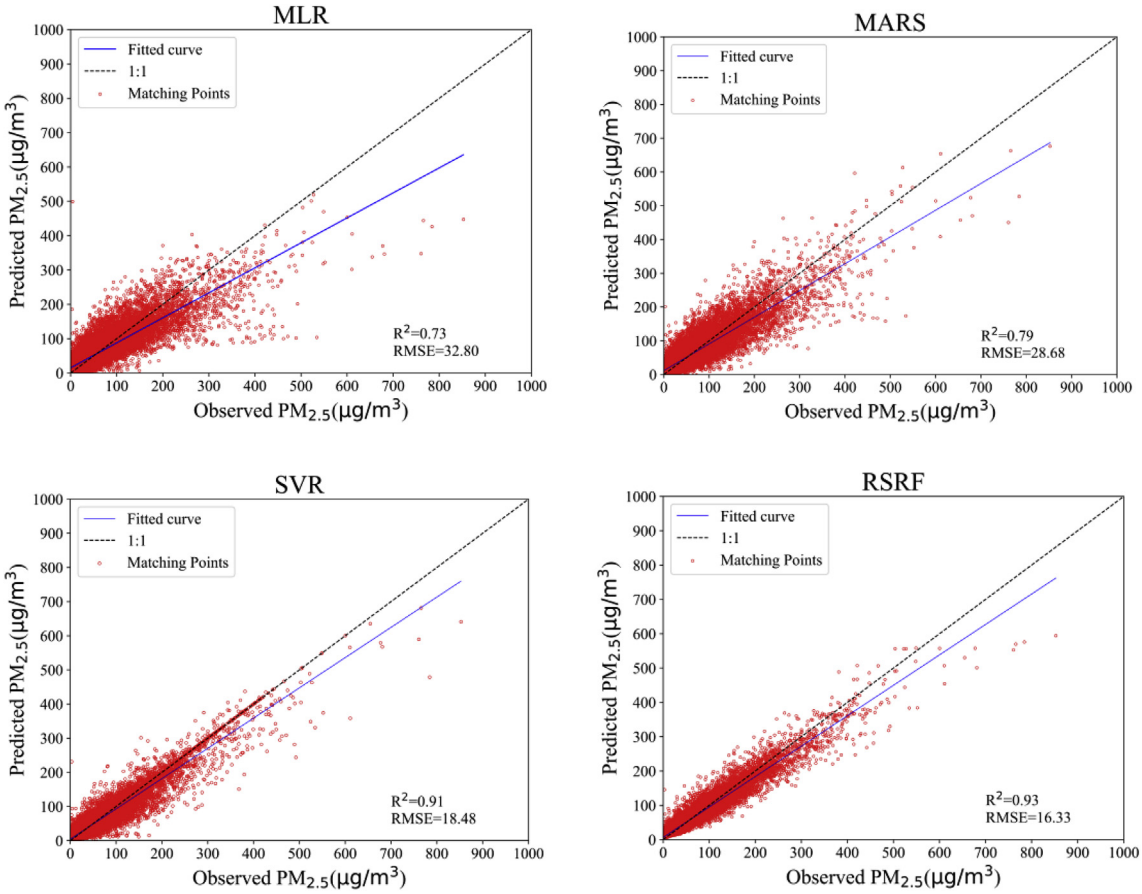

Fig. 4. The model performance comparation of MLR, MARS, SVR, RSRF.

3.2.2. Residuals and sensitivity analysis

The distributions of three-year $PM_{2.5}$ residuals obtained from the RSRF model with cross validation are shown in Fig. S6 with annual and different seasonal periods. Every period was divided into training and testing sets (i.e. spring_train, spring_test, summer_train, and summer_test). The residuals during the testing periods were always higher than those during the training periods. A positive and negative residual represents an overestimation and underestimation of $PM_{2.5}$ concentrations, respectively. The prediction capability of the RSRF model for peak concentrations of $PM_{2.5}$ is relatively weak (Figs. S5 and S6). This may be due to the effect of regression to the mean, resulting from the standard approach of minimizing the residual sum of squares in regression analysis (Zhan et al., 2018b). Another possible reason is the uncertainty of MODIS AOD detected under extremely polluted weather conditions. In summer, the residuals during both training and testing periods are smaller than other seasons. The fine particulate pollution in summer is not very serious and the number of samples in summer is less than other three seasons. That may be due to the weather conditions such as cloudy and precipitation. On the contrary, larger residuals appearing in winter were affected by the severe pollution of $PM_{2.5}$ and limited by the weak prediction capability under high pollution concentration of regression analysis. The median residual value is nearly zero for all cases and the distributions of residuals are approximately normal, indicating the prediction results are reasonable.

In order to quantitatively describe the effects of each parameter on the simulation results and make comparation with variable importance analysis, Sobol sensitivity analysis (Sobol, 1993) which is a widely used variance-based measure was conducted for all 10 variables using the SALib module in Python (Herman and Usher, 2017). A total of 22000 input model samples were generated using Saltelli's extension of the sobol sequence (Saltelli, 2002). First-order index, S1, is used to show the effects of a variable x_i on model outputs alone and second-order index, S2, shows the interactional effects of two variables x_{ij} on the model outputs. ST is the total-order index which is the sum of all indices including variable x_i. The results are shown in Table S6. AOD, NO_2 and CO exhibit high first-order sensitivities. That showed the uncertainty of the model may come mainly from the uncertainty of those input data. SO_2 and O_3 appear to have no first-order effects. For the weather variables, ST of air temperature and relative humidity are larger than S1 correspondingly. That showed those variables influence the model performance indirectly, mainly through higher-order interactions with other variables. In this study, three key input variables, including AOD, NO_2 and CO, can capture most characteristics of the relationship of input and output.

3.2.3. Variable importance

The relative importance of each input parameter for each model time step (annual and seasonal) was shown in Fig. 5. In general, the important input variables are AOD, NO_2, followed by CO, RH and AT. This is consistent with the results of sensitivity analysis. The variable importance of different time interval varies. Many factors can cause it, such as climate change, emission control, and weather condition. In spring, AOD is the most important input factor. In summer, the most important inputs are AOD and O_3. In autumn, AOD, NO_2 and CO are important variables in simulation of $PM_{2.5}$

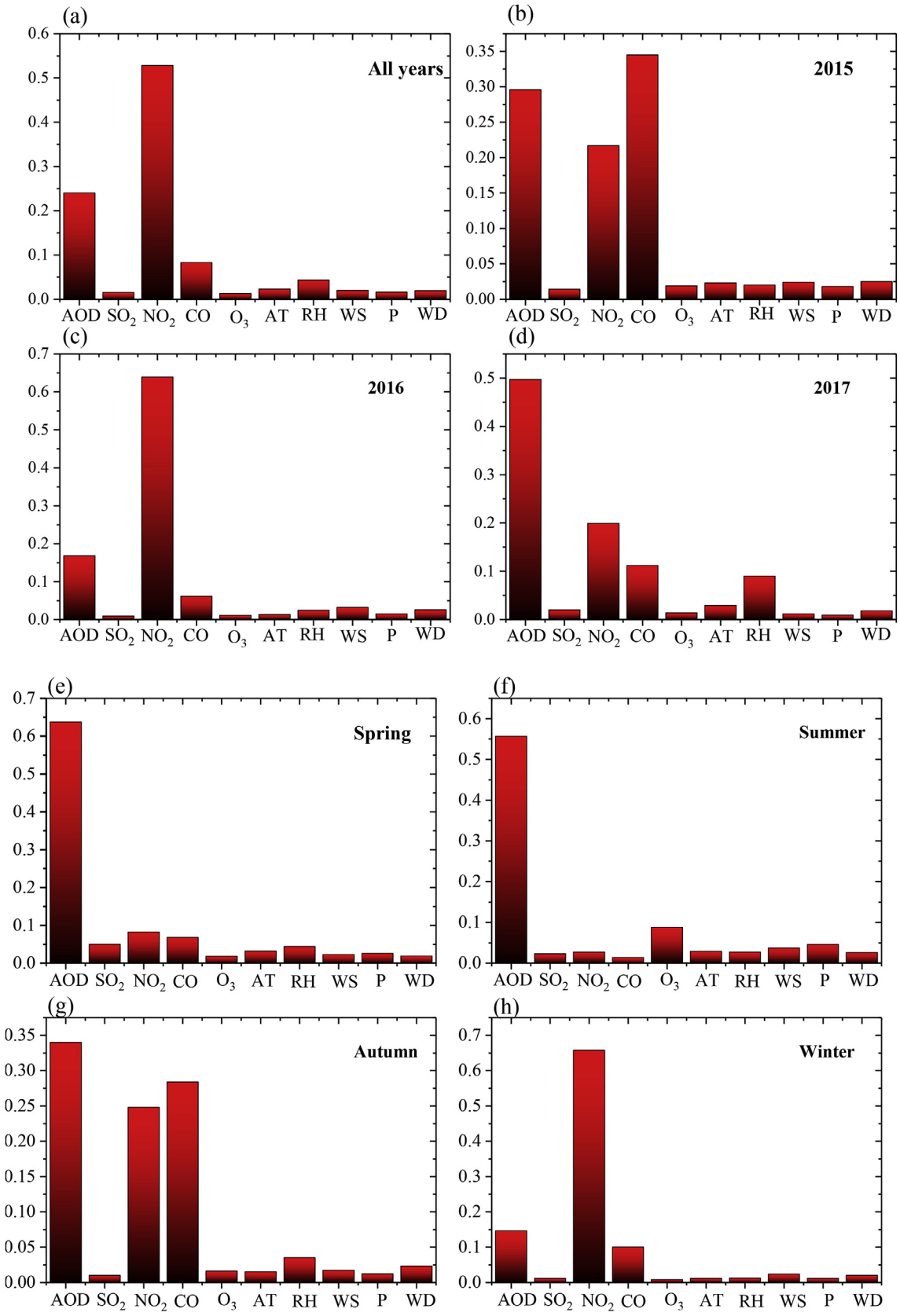

Fig. 5. Annual and seasonal variable importance analysis of the RSRF model.

concentrations. In winter, NO_2 is the most influential predictor. This indicates that the emission of air pollutants, especially NO_2, is responsible for the fine particle pollution in most cases. Although sulfate is one of the main compositions of fine particulate, SO_2 is not a very important variable in the RSRF model. Even after removing two important variables of NO_2 and CO, SO_2 was still not very important. This result is also consistent with sensitive analysis. This may be since the chemical property of SO_2 is less active than that of NO_2. On the other hand, the effects of the weather conditions on estimation of $PM_{2.5}$ concentrations are relatively weak in this study. When weather variables were not included in the model's input variables, the prediction capability of the RSRF model declined. $PM_{2.5}$ are not directly affected by the weather variables.

3.3. Predictor contribution

$PM_{2.5}$ is a main component of haze and the leading cause of haze weather. Haze weather can further aggravate the $PM_{2.5}$ accumulation. A better understanding of predictors contributions in the generation process of $PM_{2.5}$ through analyzing the relationship of $PM_{2.5}$ and other variables can help us make control strategy to reduce severe haze weather.

A correlation analysis of all variables is conducted and shown in Table S7. Air pollutants are important precursor substances of $PM_{2.5}$ formation. NO_2, CO, AOD and SO_2 all have close relationships with $PM_{2.5}$, and their seasonal distributions are shown in Fig. S7. These air pollutants are in the same atmospheric environment with similar distributions. The seasonal variation trends of NO_2, CO, and SO_2 are consistent with that of $PM_{2.5}$, except for AOD and O_3. This is partially consistent with the conclusion of variable importance.

The strength of relationship between AOD and $PM_{2.5}$ varies with seasons. In spring and summer, AOD is the most important indicator variable in the RSRF model. In autumn and winter, the prediction capability of AOD declines obviously. Although AOD have a high correlation with $PM_{2.5}$, it is not a unique indication for the fine particle pollution due to the increasingly complicated pollution situations. In spring, a lot of dust brought by sandstorms to the BTH region provide the sources of pollution of $PM_{2.5}$, which may lead to a strong prediction capability of AOD for the $PM_{2.5}$ pollution. However, the seasonal variation tendency of $PM_{2.5}$ in summer differs from that of AOD and is more similar to variation tendencies of other air pollutants except for O_3. The average monthly concentration of AOD in summer is relatively high but the concentration of $PM_{2.5}$ is not very high. In summer, a higher mixing layer and a low horizontal wind speed make the aerosol particles not easy for diffusion (Zhang and Cao, 2015). The air pollutants in North China can accumulate due to the activities of Eastern Asian summer monsoon which brings aerosol and moisture from the south and the north movement of summer subtropical high-pressure belt which prevents the diffusion of air pollutants (Qiu et al., 2017). Another possible reason of a high AOD load in summer is the photochemical reactions deduced from the corresponding high concentration of O_3 as shown in Fig. S7. Nitrogen oxides can react with volatile organic chemicals (VOCs) sharply under intense ultraviolet radiation to generate O_3 and many secondary particles which are the main components of photochemical smog. Sufficient solar radiation in summer and precursor pollutants such as NO_x emitted by anthropogenic sources, such as motor vehicle exhaust, would promote the photochemical reaction of O_3. In spring and summer, the concentration of O_3 is high in the BTH region, while the concentration of NO_2 is low due to consumption as shown in Fig. S7. O_3 has a strong positive correlation with AT and a negative correlation with NO_2 (shown in Table S7). Those can be evidences of the gradually intense photochemical reactions from spring to summer. The process of photochemical reactions may increase the load of AOD, but the pollution of fine particle is not very serious. Although AOD is an important predictor in prediction of $PM_{2.5}$ concentrations, a high load of AOD is not necessarily caused by a heavy pollution of $PM_{2.5}$. It is a precondition of $PM_{2.5}$ pollution. More other possible directly related predictors should be considered due to increasingly complex composition of atmospheric environment in China.

NO_2 and CO have high correlations with $PM_{2.5}$ in the BTH region and are important variables in the RSRF model especially in autumn and winter. Although $PM_{2.5}$ has a similar variation tendency to and a strong correlation with SO_2 (shown in Table S7), SO_2 is not an important predictor in the variable importance analysis. Effective pollution control measures for sulfide, changes in the structure of burning fossil fuels for winter heating, and heavy traffic pollution emissions may be the possible reasons (Dao et al., 2015). Besides these, the partial reason is that the amount and chemical mechanisms of different precursors in formation of $PM_{2.5}$ depend largely on atmospheric conditions, such as presence of atmospheric oxidants and water vapor (Hodan and Barnard, 2004). For example, CO is an important precursor of secondary pollution. It can affect the oxidation of the atmosphere by consuming hydroxyl radicals, and indirectly affect the transformation of other substances such as SO_2. Chemical properties of NO_2 and CO are more active than that of SO_2. NO_2 and CO may make more contributions than SO_2 to the formation of $PM_{2.5}$ in this study. This indicates the $PM_{2.5}$ pollution may be becoming dominant by nitric acid compound, which is consistent with the results by Dao et al. (2015). Enhanced anthropogenic emissions aggravates the fine particulate pollution. In winter, the concentration of CO whose main source is burning of fossil fuels and biomass fuels (Tian et al., 2017) is almost twice as high as that in other seasons (Fig. S7). Gaseous pollutants such as sulfur dioxide, nitrogen oxides and ammonia are important precursor substances of fine particles (Ye and Chen, 2013). They can be converted to sulfate and nitrate through heterogeneous phase chemical reaction under unfavorable meteorological conditions. The pollution of fine particulate particles would be worsened due to the new particle formation and secondary production of aerosol particles.

Compared with the set of air pollutants, the effect of the set of weather variables on $PM_{2.5}$ is not obvious in the RSRF model, with low correlation coefficients with $PM_{2.5}$. Atmospheric environment is a very complex system, and weather variables usually have an indirect effect on $PM_{2.5}$ that cannot be ignored. For the set of weather variables, air temperature has a negative correlation with $PM_{2.5}$. High temperature usually causes the ground pressure to be down (Fig. S7). This increases vertical convection of air mass which can promote the dilution of pollutants. RH is considered as an important factor which can promote the hygroscopic growth of fine particles and contribute to the building of environment of heterogeneous phase chemical reaction (Ye and Chen, 2013). In summer, the RH is high but the concentration of $PM_{2.5}$ is low. This may be since more precipitation in summer promotes the dilution of pollutants. The effect of wind is not distinct, unless the wind speed is very high.

3.4. Estimation of $PM_{2.5}$ concentrations during a haze period

The trained RSRF model is applied to predict the $PM_{2.5}$ concentrations from December 25 to 29, 2014. The time and space series of $PM_{2.5}$ concentration changes were shown in Fig. 6. In addition to AOD, data of air pollutants are generated using kriging interpolation of the site monitoring values. Kriging method considers not only the relationships between observation points and estimation points, but also the relationships among the relative position of each observation point. Its interpolation effect is better than the inverse distance weighted (IDW) method when the

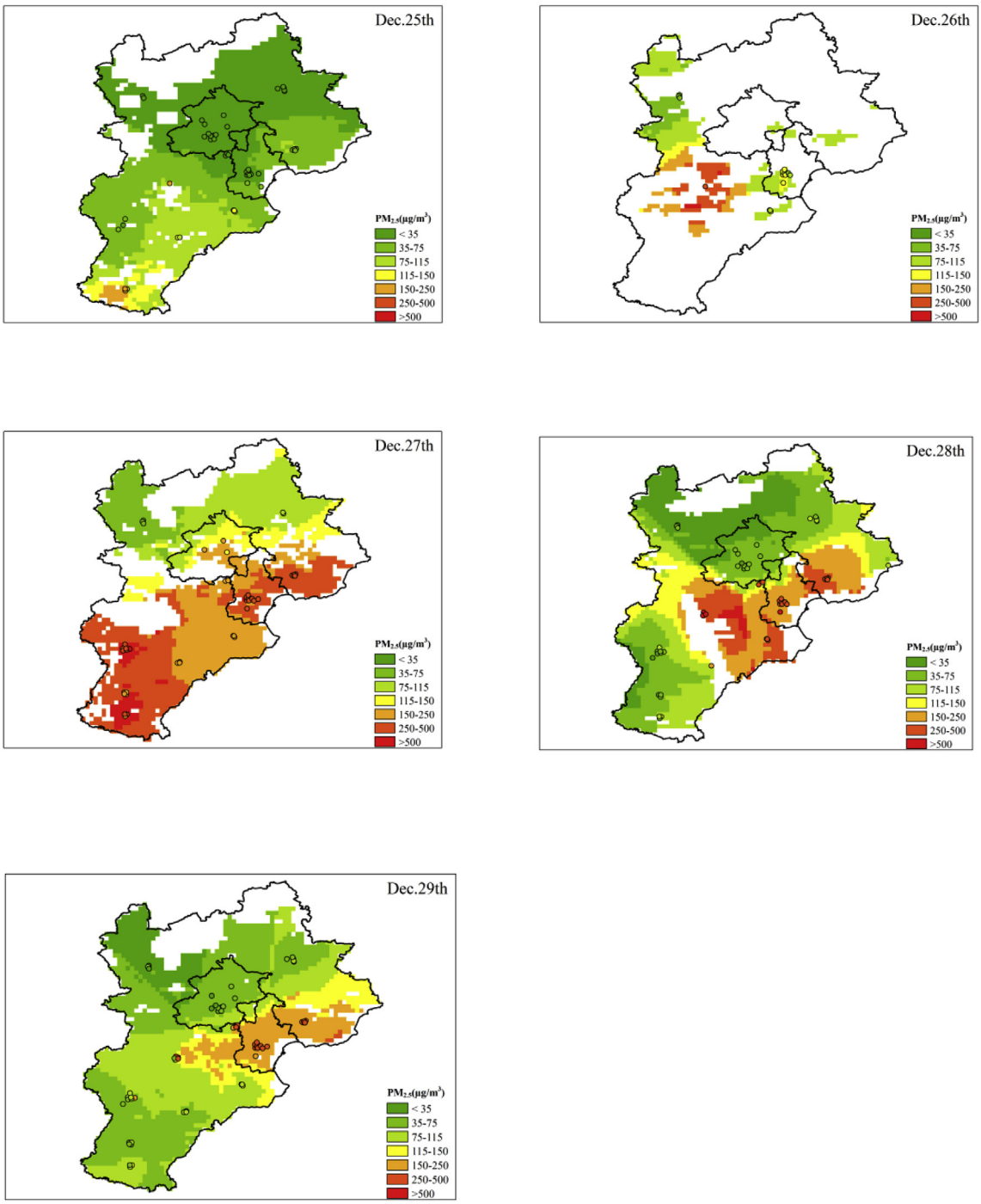

Fig. 6. Variations of $PM_{2.5}$ concentrations and comparisons with ground-measured concentrations during a period of serious air pollution in year 2014.

ground monitoring stations are rare (Wong et al., 2004). Therefore, using kriging method to interpolate spatial data often achieves ideal results. The distribution of air pollutants could also be replaced by precise ground layer products of regional climate models or satellites in the future.

The prediction results from the RSRF model are consistent with the true $PM_{2.5}$ concentrations on the ground (Fig. 7). The complete heavy haze period includes both clean and heavy pollution weather conditions. According to the model results we can better understand the air pollution process and the sources of potential pollution sources. The pollution zone extended from the southwest to the northeast of the BTH region. On December 25th, 2014, when is the first day of this period, most areas had a clean weather except for the southwestern BTH region. Combined with HYSPLIY (Stein et al., 2015) (Hybrid Single Particle Lagrangian Integrated Trajectory Model) backward trajectory (Fig. 7) of December 25th, 2014, which was carried out forward 48 h, we can know the major air mass transported to the BTH region came from Siberian plateau. This northwest air mass passing over mountainous terrain is usually relatively clean and the air quality on December 25th, 2014 in

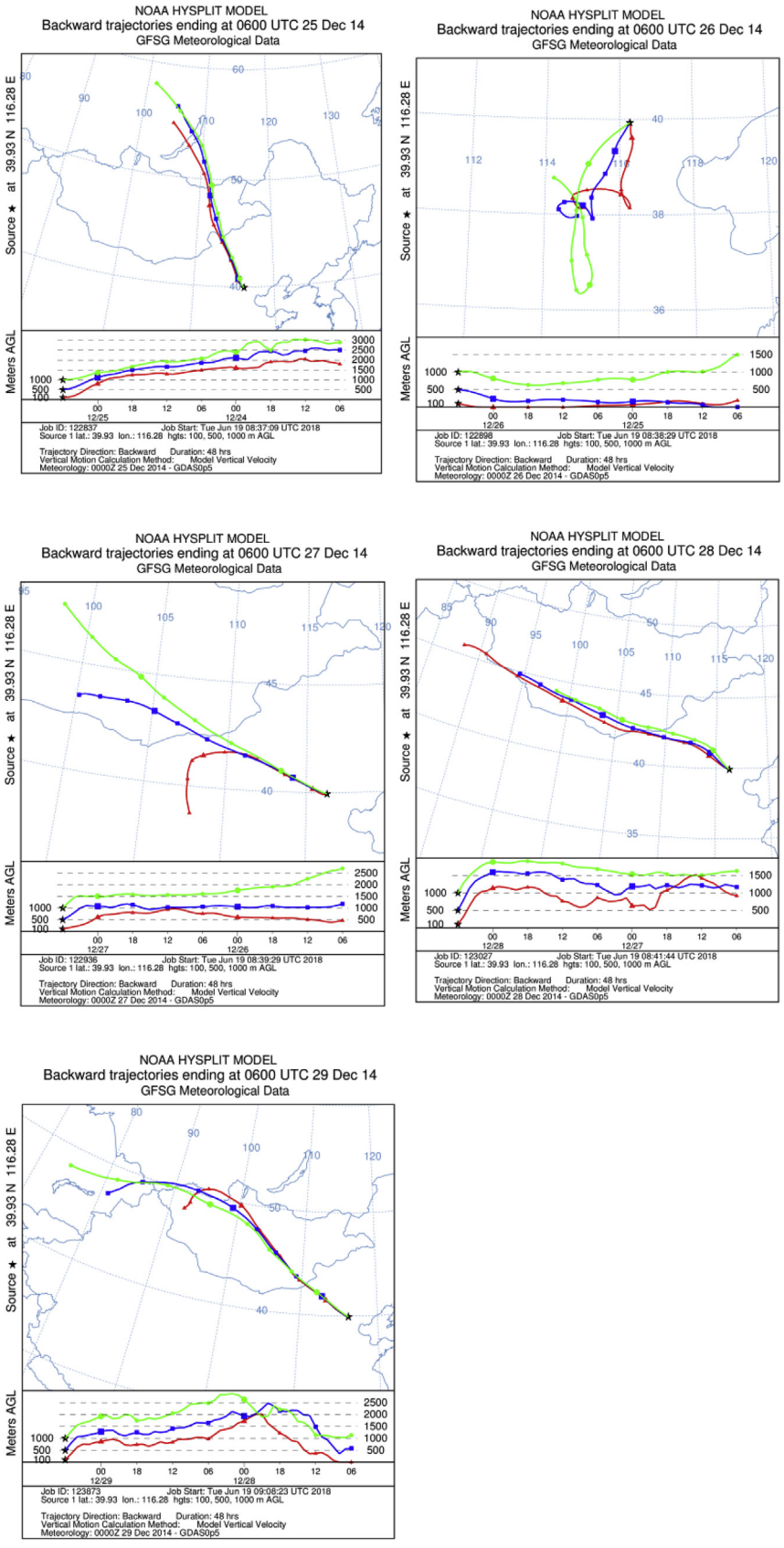

Fig. 7. HYSPLIY backward trajectory generated by Hybrid Single Particle Lagrangian Integrated Trajectory Model.

most of the BTH region was good except for the corner of the southern region. On December 26th, 2014, the $PM_{2.5}$ pollution broke out. The main air mass during 48 h before December 26th, 2014 came from the southern BTH region with high densities of population and industry, including Shijiazhuang, Baoding and Hengshui. The air mass trajectories were short, with a low altitude. That indicates the bad weather conditions, including poor diffusion conditions and a low boundary layer height. The air mass stayed for a long time at local region and carried a lot of air contaminants to the BTH region. From the backward trajectory of December 27th, 2014, the BTH region was still influenced by the air mass coming from southwestward. The pollution in this period are mainly from the emissions of the southern BTH region and partly influenced by the emissions of Shanxi and surrounding cities. The $PM_{2.5}$ pollution was still severe and spread throughout all the BTH region. The pollution started to dissipate on the December 28th, 2014 from southwest to northeast. Before December 28th, 2014 due to the bad weather conditions which were not conducive to the diffusion of pollutants, the pollution expanded to all the BTH region. As the weather condition improved, the pollution reduced from December 28th, 2014. The air condition of most BTH region is good, except for the eastern areas with some pollution on December 29th, 2014. This process indicates more attention on pollution emission control measures in the BTH region should be paid to the southwestern part.

4. Conclusion

A hybrid remote sensing and machine learning model, named RSRF, has been proposed which integrates high quality AOD, weather variables and air pollution variables into a general modeling framework to predict daily $PM_{2.5}$ concentrations in the BTH region. Variable importance of the proposed RSRF model was used to analyze the predictors contributions in the formation of $PM_{2.5}$. The proposed RSRF model was compared with the MLR, MARS and SVR models. The results indicated that the RSRF model had a relatively high prediction accuracy, outperforming other three models. The main conclusions derived for this study is: a) The prediction abilities of different predictors on $PM_{2.5}$ concentrations vary seasonally. AOD is not a unique indicator for the fine particle pollution. As a result, more directly relevant predictors such as important precursor pollutants should be considered. b) Although weather variables have less direct effects on $PM_{2.5}$ pollution, they are essential in the RSRF model. c) The $PM_{2.5}$ pollution in the BTH region is becoming dominant by nitrite acid compounds. More attention should be paid to the emission control of oxynitride. d) Ozone may be becoming the main air pollutant in summer in the BTH region.

The proposed RSRF model was applied to estimate a severe haze pollution in winter 2014 in the BTH region to demonstrate its applicability. The pollution was from the surrounding areas, especially the southern BTH region emitting a large amount of air pollution precursors of $PM_{2.5}$. Under an unfavorable weather condition, the emitted air pollutants could be transformed into fine particles, causing a severe haze pollution. When making environmental management policies, more attentions should be paid to the key variables, such as oxynitride, and the emission control for the southwestern BTH region.

The RSRF model can successfully address the daily spatiotemporal variations in the $PM_{2.5}$-AOD relationships and provide insight to relevant studies in the BTH region. The developed RSRF model could also be applied in many other air quality exploratory data analyses, such as description of the formation processes of regional $PM_{2.5}$ pollution episodes, evaluation of daily human exposure, and development of air pollution control measures.

Acknowledgment

We would like to acknowledge NASA Goddard Space Flight Center for offering MODIS sensor data. We thank the Principal Investigators and their staff for establishing and maintaining the 5 AERONET sites used in this investigation and the International AERONET Federation for their contributions to AERONET data. We also would like to thank *NOAA/OAR/ESRL PSD, Boulder, Colorado, USA* for providing NCEP Reanalysis data and China National Environmental Monitoring Centre for providing $PM_{2.5}$ and other air pollutant measurements. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2019.03.068.

References

Alpaydin, E., 2009. Introduction to Machine Learning. MIT press.
Ångström, A., 1964. The parameters of atmospheric turbidity. Tellus 16, 64–75. https://doi.org/10.3402/tellusa.v16i1.8885.
Antanasijević, D.Z., Pocajt, V.V., Povrenović, D.S., Ristić, M.Đ., Perić-Grujić, A.A., 2013. PM_{10} emission forecasting using artificial neural networks and genetic algorithm input variable optimization. Sci. Total Environ. 443, 511–519. https://doi.org/10.1016/j.scitotenv.2012.10.110.
Arhami, M., Kamali, N., Rajabi, M.M., 2013. Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. Environ. Sci. Pollut. Res. 20, 4777–4789. https://doi.org/10.1007/s11356-012-1451-6.
Beloconi, A., Kamarianakis, Y., Chrysoulakis, N., 2016. Estimating urban PM_{10} and $PM_{2.5}$ concentrations, based on synergistic MERIS/AATSR aerosol observations, land cover and morphology data. Remote Sens. Environ. 172, 148–164. https://doi.org/10.1016/j.rse.2015.10.017.
Bergin, M.S., Noblet, G.S., Petrini, K., Dhieux, J.R., Milford, J.B., Harley, R.A., 1999. Formal uncertainty analysis of a Lagrangian photochemical air pollution model. Environ. Sci. Technol. 33, 1116–1126. https://doi.org/10.1021/es980749y.
Bilal, M., Nichol, J.E., Spak, S.N., 2017. A new approach for estimation of fine particulate concentrations using satellite aerosol optical depth and binning of meteorological variables. Aerosol Air. Qual. Res. 17, 356–367. https://doi.org/10.4209/aaqr.2016.03.0097.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140. https://doi.org/10.1007/BF00058655.
Brokamp, C., Jandarov, R., Hossain, M., Ryan, P., 2018. Predicting daily urban fine particulate matter concentrations using a random forest model. Environ. Sci. Technol. 52, 4173–4179. https://doi.org/10.1021/acs.est.7b05381.
Chen, Q.X., Yuan, Y., Huang, X., Jiang, Y.Q., Tan, H.P., 2017a. Estimation of surface-level $PM_{2.5}$ concentration using aerosol optical thickness through aerosol type analysis method. Atmos. Environ. 159, 26–33. https://doi.org/10.1016/j.atmosenv.2017.03.050.
Chen, W., Fan, A., Yan, L., 2017b. Performance of MODIS C6 aerosol product during frequent haze-fog events: a case study of Beijing. Remote. Sens-Basel. 9, 496–515. https://doi.org/10.3390/rs9050496.
Chu, Y., Liu, Y., Li, X., Liu, Z., Lu, H., Lu, Y., Mao, Z., Chen, X., Li, N., Ren, M., Liu, F., Tian, L., Zhu, Z., Xiang, H., 2016. A review on predicting ground $PM_{2.5}$ concentration using satellite aerosol optical depth. Atmosphere-Basel 7, 129–154. https://doi.org/10.3390/atmos7100129.
Dao, X., Zhang, L.L., Wang, C., Chen, Y., Lv, Y.B., Teng, E.J., 2015. Characteristics of mass and ionic compounds of atmospheric particles in winter and summer of Beijing-Tian-Hebei area, China. Environ. Chem. 34, 60–69. https://doi.org/10.7524/j.issn.0254-6108.2015.01.2014032603.
De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192. CARTAP] 2.0.CO;2. https://doi.org/10.1890/0012-9658(2000)081[3178.
Fan, Y.R., et al., 2015. A stepwise-cluster forecasting approach for monthly streamflows based on climate teleconnections. Stoch. Env. Res. Risk. A. 29, 1557–1569. https://doi.org/10.1007/s00477-015-1048-y.
Genuer, R., Poggi, J.M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recogn. Lett. 31, 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014.
Giles, D., Holben, B., Eck, T., Smirnov, A., Sinyuk, A., Schafer, J., Sorokin, M., Slutsker, I., 2018. Aerosol Robotic Network (AERONET) version 3 aerosol optical depth and inversion products. In: AGU Fall Meeting Abstracts accessed 1 October 2018. https://aeronet.gsfc.nasa.gov/new_web/Documents/AeroCenter_Poster_Bash-AERONET.pdf.
Grange, S.K., Carslaw, D.C., Lewis, A.C., Boleti, E., Hueglin, C., 2018. Random forest

meteorological normalisation models for Swiss PM_{10} trend analysis. Atmos. Chem. Phys. 1–28. https://doi.org/10.5194/acp-18-6223-2018.
Gupta, P., Christopher, S.A., 2009. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. J. Geophys. Res. Atmos. 114 (14), D20205. https://doi.org/10.1029/2008JD011497.
Hastie, T., Tibshirani, R., Friedman, J., 2009. Random Forest. The Elements of Statistical Learning. Springer, New York, NY, pp. 587–604.
Herman, J., Usher, W., 2017. SALib: an open-source Python library for Sensitivity Analysis. In: Journal of Open Source Software, vol. 2, p. 97. https://doi.org/10.21105/joss.00097.
Hodan, W.B., Barnard, W.R., 2004. Evaluating the Contribution of $PM_{2.5}$ Precursor Gases and Re-entrained Road Emissions to Mobile Source $PM_{2.5}$ Particulate Matter Emissions. MACTEC Federal Programs, Research Triangle Park, NC. http://www.epa.gov/ttnchie1/conference/ei13/mobile/hodan.pdf.
Holben, B.N., 2001. An emerging ground-based aerosol climatology: aerosol optical depth from AERONET. J. Geophys. Res. Atmos. 106, 12067–12097. https://doi.org/10.1029/2001JD900014.
Hsu, N.C., 2017. Changes to MODIS Deep Blue Aerosol Products between Collection 6 and Collection 6.1. August 9 2017 accessed 1 October 2018. https://modis-atmos.gsfc.nasa.gov/sites/default/files/ModAtmo/modis_deep_blue_c61_changes2.pdf.
Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., Liu, Y., 2017. Estimating $PM_{2.5}$ concentrations in the Conterminous United States using the random forest approach. Environ. Sci. Technol. 51, 6936–6944. https://doi.org/10.1021/acs.est.7b01210.
Huang, J.F., Kondragunta, S., Laszlo, I., Liu, H.Q., Remer, L.A., Zhang, H., Superczynski, S., Ciren, P., Holben, B.N., Petrenko, M., 2016. Validation and expected error estimation of Suomi-NPP VIIRS aerosol optical thickness and Ångström exponent with AERONET. J. Geophys. Res. Atmos. 121, 7139–7160. https://doi.org/10.1002/2016JD024834.
Hystad, P., Setton, E., Cervantes, A., Poplawski, K., Deschenes, S., Brauer, M., van Donkelaar, A., Lamsal, L., Martin, R., Jerrett, M., Demers, P., 2011. Creating national air pollution models for population exposure assessment in Canada. Environ. Health Perspect. 119, 1123–1129. https://dx.doi.org/10.1289/ehp.1002976.
Ichoku, C., Chu, D.A., Mattoo, S., Kaufman, Y.J., Remer, L.A., Tanré, D., Slutsker, I., Holben, B.N., 2002. A spatio-temporal approach for global validation and analysis of MODIS aerosol products. Geophys. Res. Lett. 29. MOD1-1-MOD1-4. https://doi.org/10.1029/2001GL013206.
Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., et al., 1996. The NCEP/NCAR 40-Year reanalysis project. Bull. Am. Meteorol. Soc. 77, 437–472. https://doi.org/10.1175/1520-0477(1996)077%3e0437:TNYRP%3c2.0.CO;2.
Kioumourtzoglou, M.A., Schwartz, J.D., Weisskopf, M.G., Melly, S.J., Wang, Y., Dominici, F., Zanobetti, A., 2016. Long-term $PM_{2.5}$ exposure and neurological hospital admissions in the northeastern United States. Environ. Health Perspect. 124, 23–29. https://doi.org/10.1289/ehp.1408973.
Kloog, I., Koutrakis, P., Coull, B.A., Lee, H.J., Schwartz, J., 2011. Assessing temporally and spatially resolved $PM_{2.5}$ exposures for epidemiological studies using satellite aerosol optical depth measurements. Atmos. Environ. 45, 6267–6275. https://doi.org/10.1016/j.atmosenv.2011.08.066.
Laakso, L., Hussein, T., Aarnio, P., Komppula, M., Hiltunen, V., Viisanen, Y., Kulmala, M., 2003. Diurnal and annual characteristics of particle mass and number concentrations in urban, rural and Arctic environments in Finland. Atmos. Environ. 37, 2629–2641. https://doi.org/10.1016/S1352-2310(03)00206-1.
Li, J., Carlson, B.E., Lacis, A.A., 2015. How well do satellite AOD observations represent the spatial and temporal variability of $PM_{2.5}$ concentration for the United States? Atmos. Environ. Times 102, 260–273. https://doi.org/10.1016/j.atmosenv.2014.12.010.
Li, Z., Huang, G.H., Wang, X.Q., Han, J.C., 2015. Development of a stepwise-clustered hydrological inference model. J. Hydrol. Eng. 20 (10), 04015008. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001165.
Li, Z., Huang, G.H., Wang, X.Q., Han, J.C., Fan, Y.R., 2016. Impacts of future climate change on river discharge based on hydrological inference: a case study of the Grand River Watershed in Ontario, Canada. Sci. Total Environ. 548, 198–210. https://doi.org/10.1016/j.scitotenv.2016.01.002.
Liu, Y., Park, R.J., Jacob, D.J., Li, Q., Kilaru, V., Sarnat, J.A., 2004a. Mapping annual mean ground-level $PM_{2.5}$ concentrations using Multiangle Imaging Spectroradiometer aerosol optical thickness over the contiguous United States. J. Geophys. Res. Atmos. 109, D22206. https://doi.org/10.1029/2004JD005025.
Liu, Y., Sarnat, A.J., Coull, A.B., Koutrakis, P., Jacob, J.D., 2004b. Validation of Multiangle Imaging Spectroradiometer (MISR) aerosol optical thickness measurements using Aerosol Robotic Network (AERONET) observations over the contiguous United States. J. Geophys. Res. Atmos. 109, D06205. https://doi.org/10.1029/2003JD003981.
Liu, Y., Sarnat, J.A., Kilaru, V., Jacob, D.J., Koutrakis, P., 2005. Estimating ground-level $PM_{2.5}$ in the eastern United States using satellite remote sensing. Environ. Sci. Technol. 39, 3269–3278. https://doi.org/10.1021/es049352m.
Liu, Y., Franklin, M., Kahn, R., Koutrakis, P., 2007. Using aerosol optical thickness to predict ground-level $PM_{2.5}$ concentrations in the St. Louis area: a comparison between MISR and MODIS. Remote Sens. Environ. 107, 33–44. https://doi.org/10.1016/j.rse.2006.05.022.
Liu, B.C., Binaykia, A., Chang, P.C., Tiwari, M.K., Tsao, C.C., 2017. Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): a case study of Beijing-Tianjin-Shijiazhuang. PLoS One 12, e0179763. https://doi.org/10.1371/journal.pone.0179763.
Liu, Y., Cao, G., Zhao, N., Mulligan, K., Ye, X., 2018. Improve ground-level $PM_{2.5}$ concentration mapping using a random forests-based geostatistical approach. Environ. Pollut. 235, 272–282. https://doi.org/10.1016/j.envpol.2017.12.070.
Lu, W.Z., Wang, W.J., 2005. Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. Chemosphere 59, 693–701. https://doi.org/10.1016/j.chemosphere.2004.10.032.
Lyapustin, A., Wang, Y., Xiong, X., Meister, G., Platnick, S., Levy, R., Franz, B., Korkin, S., Hilker, T., Tucker, J., Hall, F., Sellers, P., Wu, A., Angal, A., 2014. Scientific impact of MODIS C5 calibration degradation and C6+ improvements. Atmos. Meas. Tech. 7, 4353–4365. https://doi.org/10.5194/amt-7-4353-2014.
Ma, Z., Hu, X., Huang, L., Bi, J., Liu, Y., 2014. Estimating ground-level $PM_{2.5}$ in China using satellite remote sensing. Environ. Sci. Technol. 48, 7436–7444. https://doi.org/10.1021/es5009399.
Madrigano, J., Kloog, I., Goldberg, R., Coull, B.A., Mittleman, M.A., Schwartz, J., 2013. Long-term exposure to $PM_{2.5}$ and incidence of acute myocardial infarction. Environ. Health Perspect. 121, 192–196. https://doi.org/10.1289/ehp.1205284.
Mallet, V., Sportisse, B., 2006. Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations: an ensemble approach applied to ozone modeling. J. Geophys. Res. Atmos. 111, D01302. https://dx.doi.org/10.1029/2005JD006149.
MEP, 2012. Ambient Air Quality Standards (GB3095-2012). China Environmental Science Press (MEP, Ministry of Ecology and Environment of the People's Republic of China).
MEP, 2013a. Technical Specifications for Installation and Acceptance of Ambient Air Quality Continuous Automated Monitoring System for SO_2, NO_2, O_3 and CO (HJ 193-2013). China Environmental Science Press (MEP, Ministry of Ecology and Environment of the People's Republic of China).
MEP, 2013b. Technical Specifications for Installation and Acceptance of Ambient Air Quality Continuous Automated Monitoring System for PM_{10} and $PM_{2.5}$ (HJ 655-2013. China Environmental Science Press (MEP, Ministry of Ecology and Environment of the People's Republic of China).
Nichol, J.E., Bilal, M., 2016. Validation of MODIS 3 km resolution aerosol optical depth retrievals over Asia. Remote. Sens-Basel. 8, 328–337. https://doi.org/10.3390/rs8040328.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Péré, J.C., Pont, V., Mallet, M., Bessagnet, B., 2009. Mapping of PM_{10} surface concentrations derived from satellite observations of aerosol optical thickness over South-Eastern France. Atmos. Res. 91, 1–8. https://doi.org/10.1016/j.atmosres.2008.05.001.
Pérez, P., Trier, A., Reyes, J., 2000. Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile. Atmos. Environ. 34, 1189–1196. https://doi.org/10.1016/S1352-2310(99)00316-7.
Qiu, J.Y., Sheng, L.F., Zhou, Y., Wang, W.C., Liu, Q., Li, X.D., Chen, Q., 2017. Temporal and spatial distribution of summer haze-fog and its increase in Eastern China from 1980 to 2012. Adv. Geophys. 7, 739–750. https://doi.org/10.12677/ag.2017.76075.
Reid, C.E., Jerrett, M., Petersen, M.L., Pfister, G.G., Morefield, P.E., Tager, I.B., Raffuse, S.M., Balmes, J.R., 2015. Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. Environ. Sci. Technol. 49, 3887–3896. https://doi.org/10.1021/es505846r.
Saltelli, A., 2002. Making best use of model evaluations to compute sensitivity indices. Comput. Phys. Commun. 145, 280–297. https://doi.org/10.1016/S0010-4655(02)00280-1.
Sayer, A.M., Munchak, L.A., Hsu, N.C., Levy, R.C., Bettenhausen, C., Jeong, M.J., 2014. MODIS Collection 6 aerosol products: comparison between Aqua's e-Deep Blue, Dark Target, and "merged" data sets, and usage recommendations. J. Geophys. Res. Atmos. 119 (13), 965–989. https://doi.org/10.1002/2014JD022453.
Sobol, I.M., 1993. Sensitivity estimates for nonlinear mathematical models. Math. Model Civ. Eng. 1, 407–414.
Stein, A.F., Draxler, R.R., Rolph, G.D., Stunder, B.J.B., Cohen, M.D., Ngan, F., 2015. NOAA's HYSPLIT atmospheric transport and dispersion modeling System. Bull. Am. Meteorol. Soc. 96, 2059–2077. https://doi.org/10.1175/BAMS-D-14-00110.1.
Strobl, C., Malley, J., Tut, G., 2009. An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. Psychol. Methods 14, 323–348. https://doi.org/10.1037/a0016973.
Suárez-Sánchez, A., García Nieto, P.J., Riesgo Fernández, P., del Coz Díaz, J.J., Iglesias-Rodríguez, F.J., 2011. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). Math. Comput. Model. 54, 1453–1466. https://doi.org/10.1016/j.mcm.2011.04.017.
Tao, M., Chen, L., Wang, Z., Tao, J., Che, H., Wang, X., Wang, Y., 2015. Comparison and evaluation of the MODIS Collection 6 aerosol data in China. J. Geophys. Res. Atmos. 120, 6992–7005. https://doi.org/10.1002/2015JD023360.
Tian, B., Ding, M.H., Sun, W.J., Tang, J., Wang, Y.T., Zhang, T., Xiao, C.D., Zhang, D.Q., 2017. Research progress of atmospheric carbon monoxide. Adv. Earth Sci. 32, 34–43.
van Donkelaar, A., Martin, R.V., Park, R.J., 2006. Estimating ground-level $PM_{2.5}$ using aerosol optical depth determined from satellite remote sensing. J. Geophys. Res. Atmos. 111, D21201. https://doi.org/10.1029/2005JD006996.
Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A.,

Kolehmainen, M., 2011. Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki. Environ. Sci. Technol. 409, 1266–1276. https://doi.org/10.1016/j.scitotenv.2010.12.039.

Wang, J., Christopher, S.A., 2003. Intercomparison between satellite-derived aerosol optical thickness and PM2.5 mass: implications for air quality studies. Geophys. Res. Lett. 30, 2095–2099. https://doi.org/10.1029/2003GL018174.

Wang, W., Men, C., Lu, W., 2008. Online prediction model based on support vector machine. Neurocomputing 71, 550–558. https://doi.org/10.1016/j.neucom.2007.07.020.

Wang, P., Zhang, H., Qin, Z., Zhang, G., 2017. A novel hybrid-Garch model based on ARIMA and SVM for $PM_{2.5}$ concentrations forecasting. Atmos. Pollut. Res. 8, 850–860. https://doi.org/10.1016/j.apr.2017.01.003.

Wong, D.W., Yuan, L., Perlin, S.A., 2004. Comparison of spatial interpolation methods for the estimation of air quality data. J. Expo. Anal. Env. Epid. 14, 404–415. https://doi.org/10.1038/sj.jea.7500338.

Xiao, Q., et al., 2016. Evaluation of VIIRS, GOCI, and MODIS Collection 6 AOD retrievals against ground sunphotometer observations over East Asia. Atmos. Chem. Phys. 16, 1255–1269. https://doi.org/10.5194/acp-16-1255-2016.

Xie, Y., Wang, Y., Zhang, K., Dong, W., Lv, B., Bai, Y., 2015. Daily estimation of ground-level $PM_{2.5}$ concentrations over Beijing using 3 km resolution MODIS AOD. Environ. Sci. Technol. 49, 12280–12288. https://doi.org/10.1021/acs.est.5b01413.

Xing, Y.F., Xu, Y.H., Shi, M.H., Lian, Y.X., 2016. The impact of $PM_{2.5}$ on the human respiratory system. J. Thorac. Dis. 8, E69–E74. https://doi.org/10.3978/j.issn.2072-1439.2016.01.19.

Yan, N., Wu, G., Zhang, X., Zhang, C., Xu, T., Lazhu, 2015. Variation of aerosol optical properties from AERONET observation at Mt. Muztagh Ata, Eastern Pamirs. Atmos. Res. 153, 480–488. https://doi.org/10.1016/j.atmosres.2014.10.013.

Ye, X.N., Chen, J.M., 2013. Haze and hygroscopic growth. Nat. Mag. 35, 337–341.

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M.L., 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. Atmos. Environ. 155, 129–139. https://doi.org/10.1016/j.atmosenv.2017.02.023.

Zhan, Y., Luo, Y., Deng, X., Grieneisen, M.L., Zhang, M., Di, B., 2018a. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ. Pollut. 233, 464–473. https://doi.org/10.1016/j.envpol.2017.10.029.

Zhan, Y., Luo, Y., Deng, X., Zhang, K., Zhang, M., Grieneisen, M.L., Di, B., 2018b. Satellite-based estimates of daily NO_2 xxposure in China using hybrid random forest and spatiotemporal kriging model. Environ. Sci. Technol. 52, 4180–4189. https://doi.org/10.1021/acs.est.7b05669.

Zhang, Y.L., Cao, F., 2015. Fine particulate matter ($PM_{2.5}$) in China at a city level. Sci. Rep-UK 5, 14884–14896. https://doi.org/10.1038/srep14884.

Zhang, Y., Li, Z., 2015. Remote sensing of atmospheric fine particulate matter ($PM_{2.5}$) mass concentration near the ground from satellite observation. Remote Sens. Environ. 160, 252–262. https://doi.org/10.1016/j.rse.2015.02.005.

Zou, B., Wang, M., Wan, N., Wilson, J.G., Fang, X., Tang, Y., 2015. Spatial modeling of $PM_{2.5}$ concentrations with a multifactoral radial basis function neural network. Environ. Sci. Pollut. Res. 22, 10395–10404. https://doi.org/10.1007/s11356-015-4380-3.