



Estimation of PM_{2.5} concentration in China using linear hybrid machine learning model

Zhihao Song¹, Bin Chen¹, Yue Huang¹, Li Dong¹, and Tingting Yang²

¹College of Atmospheric Sciences, Lanzhou University, Lanzhou 730000, China

²Gansu Seed General Station, Lanzhou 730030, China

Correspondence: Bin Chen (chenbin@lzu.edu.cn)

Received: 5 March 2021 – Discussion started: 30 March 2021

Revised: 11 June 2021 – Accepted: 5 July 2021 – Published: 4 August 2021

Abstract. Satellite remote sensing aerosol optical depth (AOD) and meteorological elements were employed to invert PM_{2.5} (the fine particulate matter with a diameter below 2.5 µm) in order to control air pollution more effectively. This paper proposes a restricted gradient-descent linear hybrid machine learning model (RGD-LHMLM) by integrating a random forest (RF), a gradient boosting regression tree (GBRT), and a deep neural network (DNN) to estimate the concentration of PM_{2.5} in China in 2019. The research data included Himawari-8 AOD with high spatiotemporal resolution, ERA5 meteorological data, and geographic information. The results showed that, in the hybrid model developed by linear fitting, the DNN accounted for the largest proportion, and the weight coefficient was 0.62. The R^2 values of RF, GBRT, and DNN were reported as 0.79, 0.81, and 0.8, respectively. Preferably, the generalization ability of the mixed model was better than that of each sub-model, and R^2 (determination coefficient) reached 0.84, and RMSE (root mean square error) and MAE (mean absolute error) were reported as 12.92 and 8.01 µg m⁻³, respectively. For the RGD-LHMLM, R^2 was above 0.7 in more than 70 % of the sites and RMSE and MAE were below 20 and 15 µg m⁻³, respectively, in more than 70 % of the sites due to the correlation coefficient having a seasonal difference between the meteorological factor and PM_{2.5}. Furthermore, the hybrid model performed best in winter (mean R^2 was 0.84) and worst in summer (mean R^2 was 0.71). The spatiotemporal distribution characteristics of PM_{2.5} in China were then estimated and analyzed. According to the results, there was severe pollution in winter with an average concentration of PM_{2.5} being reported as 62.10 µg m⁻³. However, there was only slight pollution in summer with an average concentration of PM_{2.5}

being reported as 47.39 µg m⁻³. The period from 10:00 to 15:00 LT (Beijing time, UTC+8 every day is the best time for model inversion; at this time the pollution is also high. The findings also indicate that North China and East China are more polluted than other areas, and their average annual concentration of PM_{2.5} was reported as 82.68 µg m⁻³. Moreover, there was relatively low pollution in Inner Mongolia, Qinghai, and Tibet, for their average PM_{2.5} concentrations were reported below 40 µg m⁻³.

1 Background

In recent years, pollutants have been discharged increasingly in China where air pollution is becoming worse than ever before due to rapid urbanization and industrialization (Wang et al., 2019a). The fine particulate matter (PM_{2.5}) with a diameter below 2.5 µm is the main component of air pollutants, having considerable impacts on human health, atmospheric visibility, and climate change (Gao et al., 2015; Pan et al., 2018; Pun et al., 2017; Qin et al., 2017). The global concern about PM_{2.5} has increased significantly since it was listed as a top carcinogen (Apte et al., 2015; Lim et al., 2020). Currently, ground monitoring is the most efficient method of measuring PM_{2.5} (Yang et al., 2018). However, monitoring stations are not evenly distributed due to terrain and construction costs; therefore, it is difficult to obtain a wide range of accurate PM_{2.5} concentration data (Han et al., 2015). To solve the problem, the method of estimating PM_{2.5} with satellite remote sensing was developed. Satellite remote sensing is characterized by a wide coverage and high resolution (Hoff and Christopher, 2009; Xu et al., 2021). There is

also a high correlation between aerosol optical depth (AOD), obtained from satellite remote sensing inversion, and PM_{2.5}; therefore, AOD is a very effective method of monitoring the spatiotemporal concentration characteristics of PM_{2.5}.

After Engel-Cox et al. (2004) proposed using satellite AOD to estimate PM_{2.5} concentration, several studies have been reported in the literature to address this theory. Based on the regression model, Liu et al. (2005) introduced AOD, boundary layer height, relative humidity, and geographical parameters as the main controlling factors to estimate PM_{2.5} in the eastern part of the United States, and the verification coefficient R^2 obtained was 0.46. Tian and Chen (2010) used AOD, PM_{2.5}, and meteorological parameters in Southern Ontario, Canada, to establish a semi-empirical model to predict PM_{2.5} concentration per hour, and the verification coefficient R^2 obtained in rural and urban areas was 0.7 and 0.64, respectively. Hu et al. (2013) proposed a geographically weighted regression model to estimate the surface PM_{2.5} concentration in southeastern America by combining AOD, meteorological parameters, and land use information. Their model average R^2 was 0.6. Lee et al. (2012) believed that the satellite remote sensing AOD data would face interference from clouds and snow and ice, and the reliability of the data was questionable. They proposed a mixed model based on AOD calibration to predict the ground PM_{2.5} concentration in New England, USA, and achieved good results ($R^2 = 0.83$). Li et al. (2016) used a PM_{2.5} remote sensing method with remote sensing of ground PM_{2.5}. Combined with MODIS (Moderate Resolution Imaging Spectroradiometer) AOD and ground observation data, Lv et al. (2017) estimated the daily surface PM_{2.5} concentration in the Beijing–Tianjin–Hebei region and improved the data resolution to 4 km. Using an interpretable self-adaptive deep neural network, Chen et al. (2021) estimated daily spatially continuous PM_{2.5} concentrations across China and analyzed the contribution of various characteristics to the PM_{2.5} model. The data used in these early studies are AOD products obtained from polar-orbit satellite sensors. The daily observation frequency is limited. Due to the influence of cloud and ground reflection, the dynamic change information of PM_{2.5} cannot be obtained. As a result, geostationary satellite observations can be used to overcome the problem of low temporal resolution for estimating surface PM_{2.5} (Emili et al., 2010).

The Himawari-8 satellite commonly used in the Asia-Pacific region is a geostationary satellite launched by the Japan Meteorological Agency in 2014. The observation frequency is 10 min, and the observation results can characterize the aerosols and provide AOD data with a resolution of 5 km (Bessho et al., 2016; Yumimoto et al., 2016). Due to its excellent performance, Wei et al. (2021a) use Himawari-8 data to estimate ground PM_{2.5}; results show that the CV R^2 (cross-validation coefficient of determination) is 0.85, with a root mean square error (RMSE) and mean absolute error (MAE) of 13.62 and 8.49 $\mu\text{g m}^{-3}$, respectively. Wang et al. (2017) proposed an improved linear

model and introduced AOD, meteorological parameters, and geographic information to estimate PM_{2.5} in the Beijing–Tianjin–Hebei region, and the verification coefficient R^2 was 0.86. T. X. Zhang et al. (2019) used the Himawari-8 hourly AOD product to estimate ground PM_{2.5} in China's four major urban agglomerations. The results showed significant diurnal, seasonal, and spatial changes and improved the temporal resolution of estimating PM_{2.5} concentration to the hourly level. Yin et al. (2021) used Himawari-8 hourly TOAR (top-of-the-atmosphere reflectance) data to estimate ground PM_{2.5} in China, improving the data coverage area.

As research into ground-based PM_{2.5} estimation deepens, traditional linear or nonlinear models cannot meet the requirements of large-scale estimation and are gradually being replaced by machine learning algorithms with strong nonlinear fitting abilities (Guo et al., 2021; Mao et al., 2021). Liu et al. (2018) combined Kriging interpolation and a random forest algorithm to obtain the concentration of high-resolution ground PM_{2.5} in the United States. To demonstrate the accuracy and superiority of the proposed method, the results were compared with the PM_{2.5} concentration in ground measurement stations. Chen et al. (2019) stacked and predicted PM_{2.5} concentration based on a variety of machine learning algorithms, discussed the influence of meteorological factors on PM_{2.5}, and achieved an R^2 of 0.85. Li et al. (2017a) established a GRNN (generalized regression neural network) model for the whole of China to estimate PM_{2.5} concentration, and the results demonstrated that the performance of the deep learning model was better than that of the traditional linear model. In addition, there are some novel algorithms such as the space-time extra-tree (STET) (Wei et al., 2021b) and space-time random forest (STRF) (Wei et al., 2019a) algorithms that are also used for PM_{2.5} inversion research.

A large number of existing studies in the broader literature have examined the estimation of ground PM_{2.5} concentrations using satellite remote sensing AOD. However, the performance of PM_{2.5} estimation models established in the existing studies varies greatly and is not stable in different seasons and regions. To overcome this limitation, in this paper, a restricted gradient-descent linear hybrid machine learning model (RGD-LHMLM) based on a random forest (RF), gradient lifting regression tree (GBRT), and deep neural network (DNN) is proposed to estimate ground PM_{2.5} concentration. The model performance is evaluated from time and space to analyze its causes. Finally, spatiotemporal distribution of PM_{2.5} concentration in China in 2019 is obtained.

2 Data

2.1 Ground PM_{2.5} monitoring data

PM_{2.5} concentration data for 2019 used in this study are available from the China National Environmental Monitor-

ing Center (CNEMC) real-time air quality publication system. The PM_{2.5} datasets are calibrated and quality-controlled according to national standard GB 3095-2012 (China's National Ambient Air Quality Standards) (China, 2012). The system extracts hourly mean PM_{2.5} data. By the end of 2019, China had 1641 monitoring stations built and in operation. Figure 1 shows the spatial distribution of monitoring stations in China.

2.2 Satellite AOD data

The Advanced Himawari Imager (AHI) on the Himawari-8 satellite launched by the Japan Meteorological Agency is a highly improved multi-wavelength imager. It adopts the whole disk observation method and has 16 visible and infrared channels. It has the characteristics of fast imaging speed and flexible observation area and time. Himawari-8 AOD is obtained by an aerosol retrieval algorithm based on a Lambertian surface assumption developed by Yoshida et al. (2018). The Level 3 hourly AOD product, released by the Japan Aerospace Exploration Agency (JAXA), provides 500 nm AOD data with a spatial resolution of 5 km during the day. In previous studies (Zang et al., 2018), Himawari-8 AOD was compared with the AOD data of AERONET (Aerosol Robotic Network) in China and achieved good performance (Z. Zhang et al., 2019), so the results show that they are consistent ($R^2 = 0.75$); RMSE and MAE were 0.39 and 0.21, respectively (Wei et al., 2019b). The AOD data used in this study are the Himawari-8 Level 3 hourly AOD data in 2019 obtained from the Himawari Monitor website of the Japan Meteorological Agency (see “Code and data availability” section). In the study, we selected AOD with strict cloud screening, that is, AOD data with low uncertainty.

2.3 Meteorological data

ERA5 reanalysis data comprise an hourly collection of atmospheric and land-surface meteorological elements that has taken place since 1979 that the European Centre for Medium-Range Weather Forecasts (ECMWF) has used with its prediction model and data assimilation system to reanalyze archived observations (Jiang et al., 2021). Data used in this paper include surface relative humidity (RH, expressed as a percentage), air temperature at a height of 2 m (T_{2m} , in K), wind speed (U_{10} , V_{10} ; in m s^{-1}), surface pressure (SP, in Pa), boundary layer height (BLH, in m), and cumulative precipitation (RAIN, in m) at 10 m above the ground. A series of studies have indicated that these parameters can affect the concentration of PM_{2.5} (Fang et al., 2016; Guo et al., 2017; Li et al., 2017b; Wang et al., 2019b; Zheng et al., 2017; Gui et al., 2019). Uncertainty estimation of ERA5 data has been described in detail on the following website: <https://confluence.ecmwf.int/display/CKB/ERA5%3A+uncertainty+estimation> (last access: 1 July 2021).

2.4 Auxiliary data

The auxiliary data used in this study include high and low vegetation indices (LH, LL), ground elevation data (DEM), and population density data (PD). The high and low vegetation indices are derived from ERA5 reanalysis data and represent half of the total green leaf area per unit level ground area of high and low vegetation type, respectively. The ground elevation data are derived from SRTM3 (Shuttle Radar Topography Mission, 90 m resolution) measurements jointly conducted by NASA and the US Department of Defense National Imagery and Mapping Agency (NIMA), with a spatial resolution of 90 m. The population data come from the 2015 United Nations GPWv4-Adjusted Population Density data provided by NASA's Socioeconomic Data and Applications Center (SEDAC), which is based on national censuses and adjusted for relative spatial distribution.

3 Method

3.1 Random forest

Random forests (RF) are built based on the combination of a bagging algorithm and decision trees (Breiman, 2001), which are an extended variant of the parallel ensemble learning method (Stafoggia et al., 2019). To construct a large number of decision trees, the random forest model takes multiple samples of the sample data. In the decision trees, the nodes are divided into sub-nodes by using the randomly selected optimal features until all the training samples of the node belong to the same class. Finally, all the decision trees are merged to form the random forest. This method has proved to be effective in regression and classification problems and is one of the best-known machine learning algorithms used in many different fields (Yesilkanat, 2020).

3.2 Gradient-boosted regression trees

Differently from the random forest, gradient boosting regression trees (GBRTs) are based on a boosting algorithm and decision trees (Friedman, 2001). The basic principle of GBRTs is to construct N different basic learners through multiple iterations and constantly add the weight of the learners with a small error probability to eventually generate a strong learner (Johnson et al., 2018). The core of this method is that after each iteration, a learner will be built in the direction of residual reduction (gradient direction) to make the residual decrease in the gradient direction (Schonlau, 2005). The basic learner of the GBRT is the regression tree in the decision tree. During the prediction, a predicted value is calculated according to the model obtained. The minimum square root error is used to select the optimal feature to split the dataset, and the average value of the child node is then taken as the predicted value.

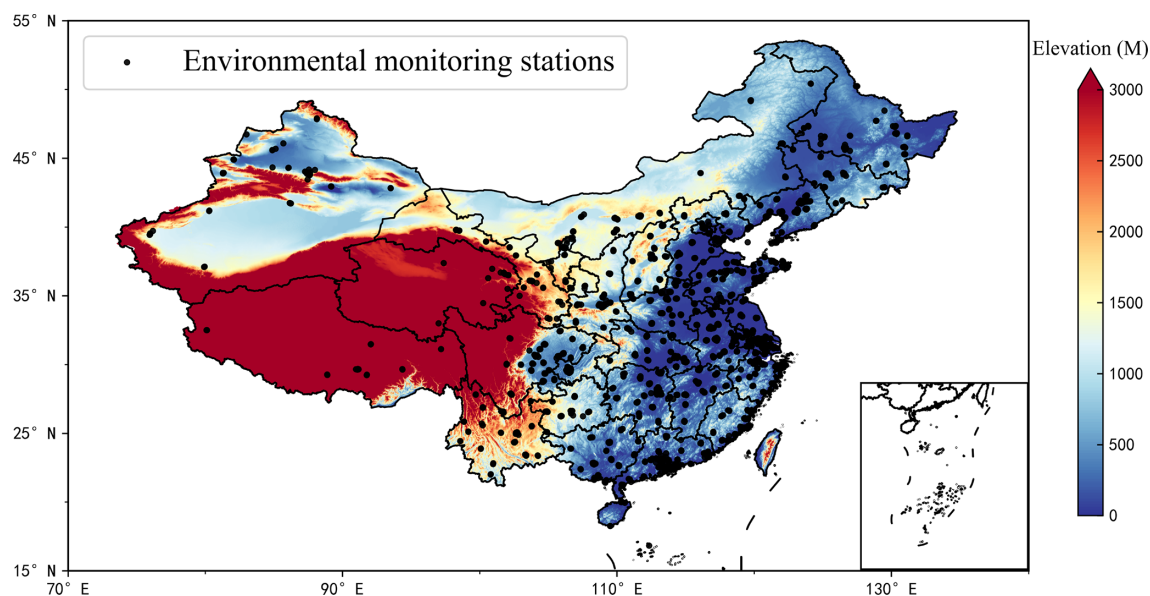


Figure 1. Distribution diagram of environmental monitoring stations in China.

3.3 Deep neural networks

Deep neural networks (DNNs) comprise a supervised learning technique that uses a backpropagation algorithm to minimize the loss function. They adjust the parameters through an optimizer and have high computational power, making them ideal for solving classification and regression problems (Wang and Sun, 2019). The structure of a DNN includes an input layer, an output layer, and several hidden layers. Each layer takes the output of all nodes of the previous layer as the input, and this process requires activation functions. Compared with other activation functions, the linear rectifying function (ReLU) has the advantages of simple derivation, faster convergence, and higher efficiency. At the same time, among the adaptive learning rate optimizers, the AdaMax optimizer performs the best. It not only has the advantages of Adam in determining the learning rate range and having stable parameters in each iteration but also simplifies the method of defining the upper limit range of the learning rate and improves the iteration efficiency (Diederik and Jimmy, 2014). Therefore, in this paper, we selected the AdaMax optimizer and ReLU activation function to train the DNN.

3.4 Model establishment and verification

After data processing, an RF, GBRT, and DNN are used for modeling.

$$\begin{aligned}
 \text{PM}_{2.5i,j} = & \text{AOD}_{i,j} + \text{BLH}_{i,j} + \text{RH}_{i,j} + \text{T}_{2\text{mi},j} + \text{LL}_{i,j} \\
 & + \text{LH}_{i,j} + \text{SP}_{i,j} + \text{RAIN}_{i,j} + \text{U}_{10i,j} + \text{V}_{10i,j} \\
 & + \text{PD}_{i,j} + \text{HEIGHT}_{i,j} + \text{LONG}_{i,j} + \text{LAT}_{i,j} \\
 & + \text{MONTH}_{i,j} + \text{HOUR}_{i,j}
 \end{aligned} \quad (1)$$

Equation (1) is applicable to the RF, GBRT, and DNN, where $\text{PM}_{2.5i,j}$ is the PM_{2.5} at time i at station j .

To prevent model parameters from being controlled by a large or small range of data and speed up the convergence rate of the model, the data must be normalized before starting the training process. Finally, the three optimal sub-models are linearly combined to achieve the final mixed model. To verify the model performance, this paper uses the “10-fold cross-validation” method (Adams et al., 2020). In this method, the data are split into 10 copies, 9 copies for training and 1 copy for verification; this process is repeated 10 times, and then the average of the 10 predictions is computed as the final result. Finally, the predicted value and the measured value are fitted linearly. At the same time, several indicators are used to evaluate the model, including the mean absolute error (MAE; when the predicted value and the true value are exactly equal to 0, that is, the perfect model; the larger the error, the greater the value), the root mean square error (RMSE; when the predicted value and the real value are completely consistent, this is equal to 0, that is, the perfect model; the larger the error, the greater the value), the slope of the fitting equation and the determination coefficient R^2 (the greater the value, the better the model fitting effect), the bias (Bias; the difference between the predicted values and the true values, so models with larger bias performed worse), and the GEB (generalization error of the bias; it is generally believed that bias should be expressed as a square when using the generalization error). The calculation formula of each indicator is shown as follows:

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}, \quad (2)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|, \quad (3)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}, \quad (4)$$

$$\text{Bias} = \frac{\sum_{i=1}^N \hat{y}_i - y_i}{N}, \quad (5)$$

$$\text{GEB} = \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{N}, \quad (6)$$

where \hat{y}_i represents the predicted value, y_i shows the true value, ss_{res} denotes the error between the regression data and the mean value, SS_{tot} represents the error between the real data and the mean value, and the mean value is the mean value of the true value.

The research process is illustrated in Fig. 2.

4 Results and discussion

4.1 Modeling results

According to the above steps, the mixed model RGD-LHMLM is obtained through modeling verification and is compared with the RF, GBRT, and DNN. The fitting and verification accuracy results of each model are shown in Table 1.

The PM_{2.5} inversion results of a single machine learning model show that the DNN has the best inversion performance, followed by the GBRT, and the RF has the worst performance. The expression of the mixing model obtained after linear mixing is as follows:

$$\text{PM}_{2.5\text{RGD-LHMLM}} = 0.25\text{PM}_{2.5\text{RF}} + 0.17\text{PM}_{2.5\text{GBRT}} + 0.62\text{PM}_{2.5\text{DNN}} - 2.13. \quad (7)$$

The weight coefficient of the DNN in the mixed model was the largest (0.62). The R^2 of RGD-LHMLM in the training set was 0.98, and the RMSE was only $4.39 \mu\text{g m}^{-3}$, indicating that the model had an excellent data fitting effect. Meanwhile, the generalization ability of the mixed model is also good, with R^2 of 0.84 and RMSE of $12.92 \mu\text{g m}^{-3}$ on the validation dataset. Among all the models, the deviation generalization error of the linear mixed model is also the lowest, indicating that the difference between the results obtained by this model and the real value is the lowest. Compared with the RF, GBRT, and DNN, the inversion performance of RGD-LHMLM is improved. In other words, the combination of multiple models can improve the robustness and generalization ability of the model (Wolpert, 1992). The linear fitting equation coefficients between the predicted and measured values in the training set and the verification set were 0.98 and 0.84, respectively, indicating that the prediction accuracy of the model reached a high level. The fitting curve between the model-predicted value and the real value is shown in Fig. 3. The RGD-LHMLM model has the smallest degree of data dispersion, and the slope of the fitting line reaches 0.84, indicating that 84 % of the prediction results are accurate, higher than in the three sub-models. The accuracy of the model decreased in the site-based validation, in

which the R^2 and RMSE values are 0.8 and $14.59 \mu\text{g m}^{-3}$, respectively.

4.2 Model performance analysis

4.2.1 Bias analysis of model

The average bias of the mixed model in different PM_{2.5} concentration ranges was analyzed, and the result is shown in Fig. 4. When the PM_{2.5} concentration is less than $60 \mu\text{g m}^{-3}$, the average bias of the model is less than 0. As the PM_{2.5} concentration increases, the model deviation gradually increases. In other words, when the PM_{2.5} concentration is small, the predicted value of the model will generally overestimate PM_{2.5}, and when the PM_{2.5} further increases, it will underestimate the PM_{2.5} concentration.

4.2.2 Performance analysis of monitoring station model

The spatial performance of the model was analyzed by measuring R^2 , RMSE, and MAE at the monitoring stations. According to Fig. 5, there are regional differences in the inversion performance of RGD-LHMLM. At all monitoring stations, the average R^2 was reported as 0.74, and R^2 was above 0.7 at more than 70 % of the stations, especially in the densely populated and industrially developed areas. The model prediction accuracy was reported as low ($R^2 < 0.6$) in Xinjiang, Tibet, Qinghai, western Sichuan, and a few other areas of Northeast China. The mean values of RMSE and MAE were reported as 11.4 and $8.01 \mu\text{g m}^{-3}$, respectively. In fact, the mean values of RMSE and MAE were below 20 and $15 \mu\text{g m}^{-3}$ in more than 95 % of stations, showing a low estimation error.

Based on the analysis of spatial differences in the RGD-LHMLM inversion performance, the following deductions can be made. First, the environmental monitoring stations in the central and eastern regions with better inversion performance were distributed densely, and there are many data available; therefore, the model had a satisfactory training effect. Moreover, data matching was lower in the western region than in other regions, something which resulted in model over-fitting and reduced accuracy (Zhang et al., 2018). Second, some areas of western and northeastern China are covered by snow and the Gobi Desert has high surface albedo. This reduces the accuracy of AOD obtained by satellite observation and introduces errors into model training. Finally, the Himawari-8 scanning range is limited, and the satellite observation data obtained in western China are limited in terms of quantity and accuracy. In general, RGD-LHMLM has a satisfactory spatial performance, especially in areas with high annual average concentrations of PM_{2.5}; therefore, it can have a good inversion effect.

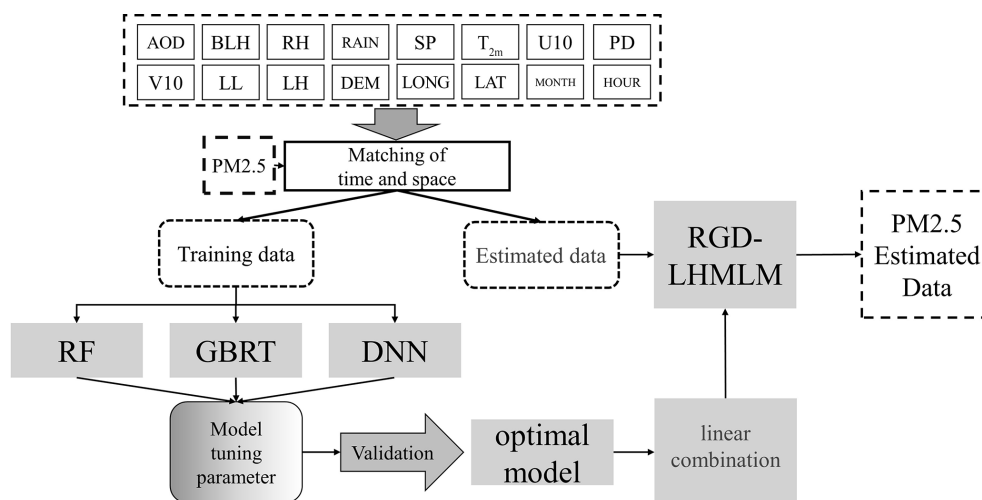


Figure 2. Schematic diagram of model.

Table 1. Comparison of model accuracy.

Model	Fitting				Validation			
	R^2	RMSE	MAE	GEB	R^2	RMSE	MAE	GEB
RF	0.95	6.99	4.05	114.19	0.79	14.89	9.33	208.97
GBRT	0.96	6.87	4.52	110.00	0.81	14.09	9.18	198.65
DNN	0.97	5.03	3.49	59.16	0.80	14.45	9.06	221.86
RGD-LHMLM	0.98	4.39	3.00	44.97	0.84	12.92	8.01	166.95

4.2.3 Timescale model performance analysis

Figure 6 shows the scatterplot fitted with the inversion results of the mixed model from 9:00–17:00 LT. The model R^2 ranged from 0.556 to 0.88 at different times. Except for 17:00 LT when the model had the worst performance, the model R^2 exceeded 0.7, indicating that the model had a good performance. The optimal performance time is 13:00 LT when R^2 is 0.88. According to the results, the hourly differences in model performance were significant.

Figure 7 shows the inversion performance results of the hybrid model collected from January to December 2019. The model performed the worst in the summer months – June, July, and August – when R^2 was reported as 0.73, 0.72, and 0.68, respectively; however, RMSE and MAE were only 9.37, 9.22, 8.26 and 6.59, 6.34, and 5.91 $\mu\text{g m}^{-3}$, respectively, due to the lower average concentration of PM_{2.5} in summer. Winter and autumn models gained better performance results with an average R^2 over 0.8. However, in contrast to those of summer, the estimation errors of these two seasons were relatively large, with average RMSE of 20.10 and 10.72 $\mu\text{g m}^{-3}$ and average MAE of 11.20 and 7.25 $\mu\text{g m}^{-3}$, respectively. The mean R^2 was 0.74, and the mean RMSE and MAE were 13.71 and 8.39 $\mu\text{g m}^{-3}$, respectively.

4.2.4 Feature importance analysis

The model performance differences were also analyzed to extract and rank the model features of the RF and GBRT based on the feature importance. The higher the feature importance, the greater the contribution of factors to the model. Figure 8 shows that AOD, boundary layer height, 2 m surface temperature, and relative humidity had the greatest effect on the mixed model performance out of all variable characteristic parameters. Accordingly, AOD is greatly affected by the fine particulate matter and is the main factor in the inversion of PM_{2.5}. Changes in the boundary layer height can affect the diffusion ability of the atmosphere. If the boundary layer height is low, the accumulation of pollutants will be caused. At the same time, the 2 m surface temperature has a great impact on the boundary layer height (Miao et al., 2018). Finally, higher rates of atmospheric humidity can improve the fine particulate matter accumulation.

As shown in Fig. 9, the correlation coefficients between the monthly mean values of important meteorological parameters (AOD, BLH, $T_{2\text{m}}$, and RH) and R^2 were also analyzed. According to the results of Table 2, the correlation coefficients between the meteorological parameters and PM_{2.5} were lower in summer. Furthermore, there are many rainy days and large cloud coverage, which is not conducive to

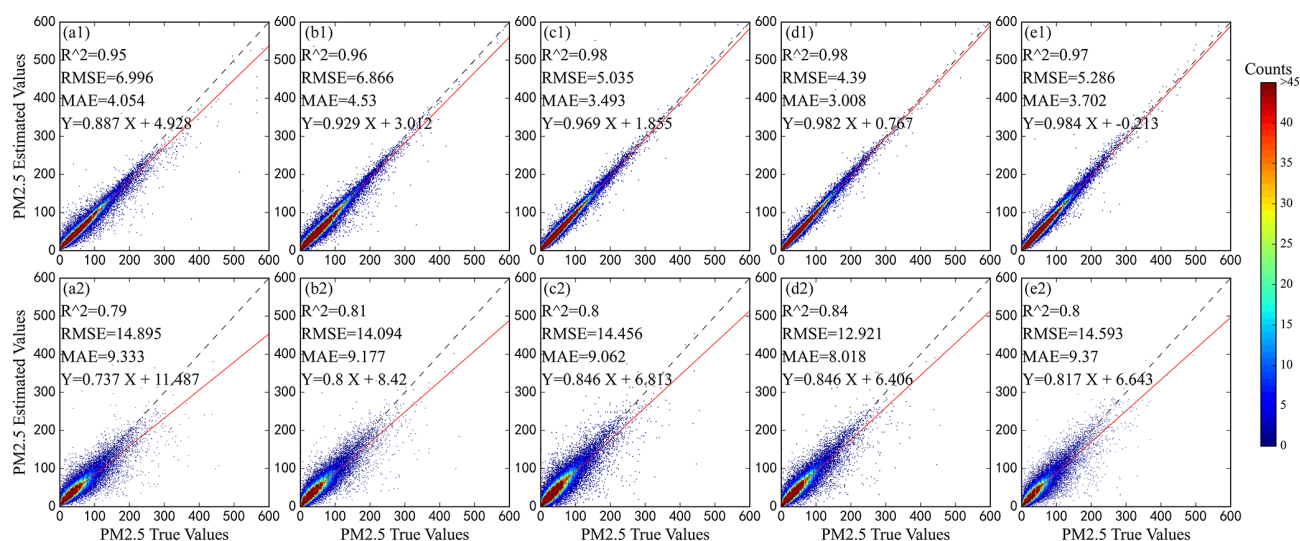


Figure 3. Accuracy of model fitting (the first row) and validation (the second row) (a: RF, b: GBRT, c: DNN, d: RGD-LHMLM (based on sample), e: RGD-LHMLM (based on site)). R^2 represents the determination coefficient; RMSE represents root mean square error; MAE represents mean absolute error; N represents the number of samples. The equation with terms Y and X represents the fitting relationship between the actual and estimated PM_{2.5} values. Dashed black line represents 1 : 1 line, and red line represents best-fit line from linear regression.

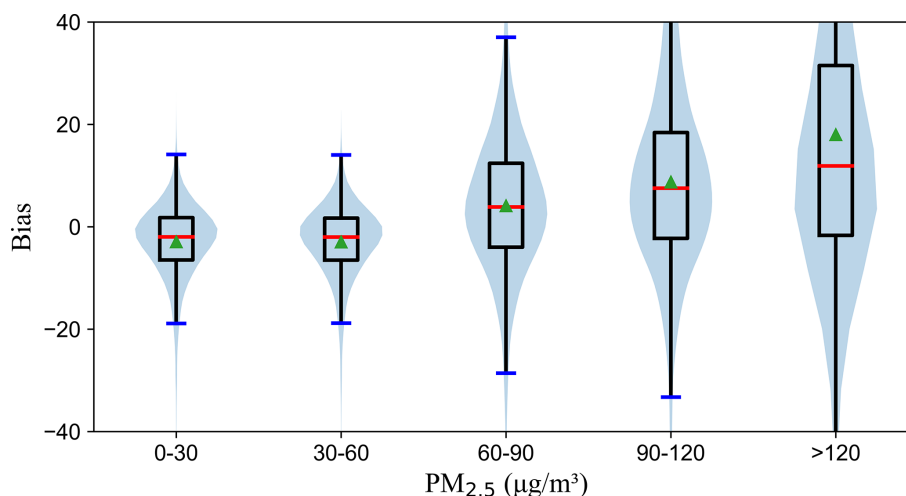


Figure 4. Boxplots of resulting bias (y axis) for different PM_{2.5} concentration ranges in $\mu\text{g m}^{-3}$ (x axis) (the green arrow symbol and dark blue and red marks represent the average bias, the median of bias, and the extrema of bias, respectively. Data density is represented by the light blue shading.)

satellite observation and decreases the accuracy of AOD data in summer. Therefore, the summer model performance is poor. There was a strong correlation between meteorological parameters and PM_{2.5} in autumn. There were also similar correlations between spring and winter; however, the winter model performed better. The reasons can be interpreted as follows. The winter temperature and boundary layer height are low, whereas the atmosphere is stable but not conducive to the diffusion of pollutants. Moreover, during the heating period in winter, pollutant emissions soar greatly and result in a sharp rise in the concentration of PM_{2.5}. The increased

pollution in winter ensures the quality and quantity of data, thereby improving the model performance effectively.

4.3 Temporal and spatial distribution characteristics of PM_{2.5} concentration in China

In terms of spatial distribution, Shandong, Henan, Jiangsu, and Anhui, as well as parts of Hubei and Hebei, were the most polluted areas in China in 2019, with an annual average PM_{2.5} concentration of $82.86 \mu\text{g m}^{-3}$. On the one hand, these areas are economically developed and densely popu-

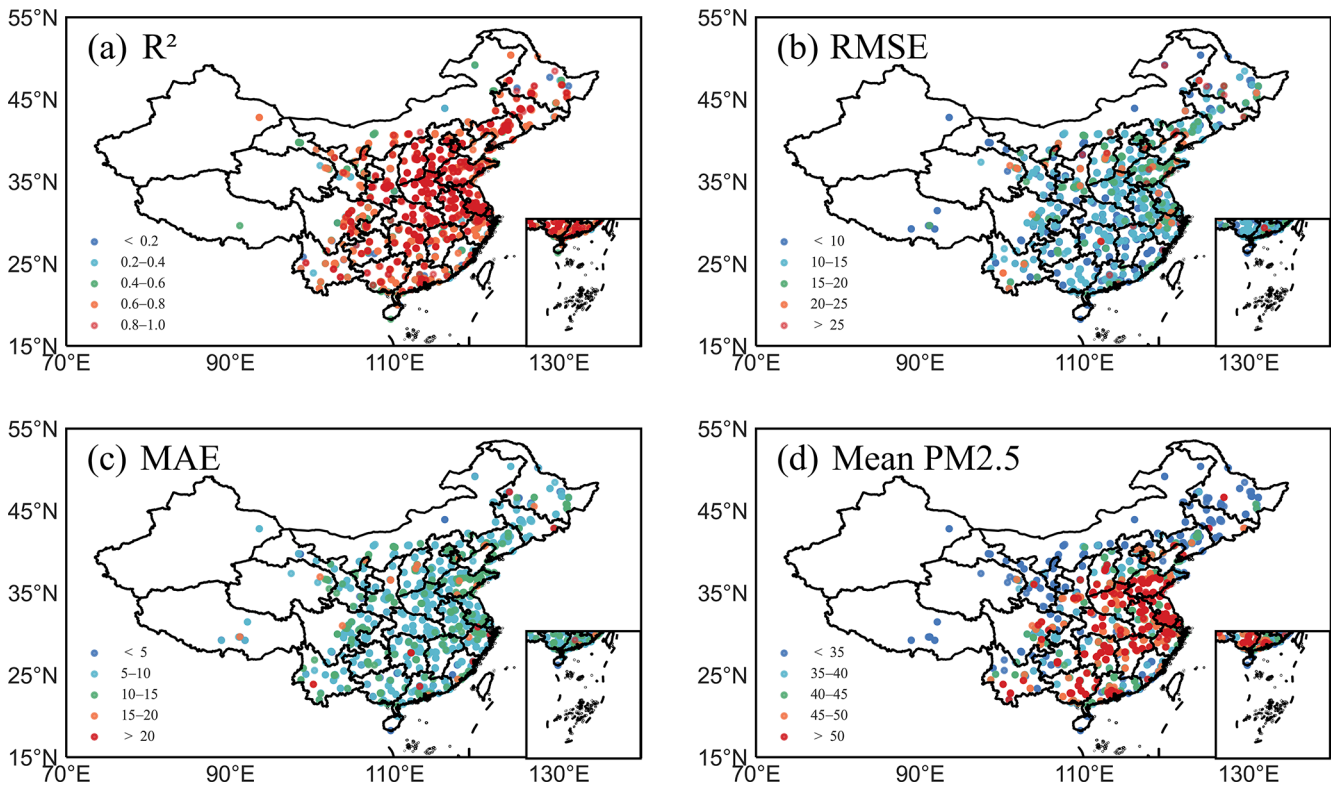


Figure 5. Spatial distributions of model precision in terms of (a) determination coefficient (R^2), (b) root mean square error (RMSE), (c) mean absolute error (MAE), and (d) mean PM_{2.5} concentration at each site in China. Colored circles represent different value ranges of statistical parameters shown.

Table 2. Correlation coefficient between meteorological parameters with PM_{2.5}.

Season	AOD	BLH	T_{2m}	RH
Spring	0.47	−0.33	0.12	0.36
Summer	0.42	−0.21	0.06	0.19
Autumn	0.38	−0.29	0.24	0.41
Winter	0.44	−0.33	0.12	0.35

lated, resulting in a large quantity of pollutant emissions. On the other hand, the barrier of the peripheral mountains (Taihang Mountains, the Qinling, and the southern hills) leads to the accumulation of pollutants that are difficult to diffuse. The Sichuan Basin is a rare area with a high PM_{2.5} value due to its unique topography (L. Zhang et al., 2019), with an annual average PM_{2.5} concentration of 64.69 $\mu\text{g m}^{-3}$. In addition, in Inner Mongolia, Qinghai, Tibet, and other places, the pollution level is low: the average annual PM_{2.5} concentration is less than 40 $\mu\text{g m}^{-3}$.

The temporal distribution of PM_{2.5} is shown in Fig. 10. The PM_{2.5} concentration began to rise from 09:00 LT, and peaked at 55.65 $\mu\text{g m}^{-3}$ between 10:00 and 11:00 LT every day. After that, it maintained a high concentration until

15:00 LT and then began to decrease. In the most polluted areas of China, the peak concentration of PM_{2.5} can reach 85.05 $\mu\text{g m}^{-3}$, while the peak in the less polluted areas is only about 40 $\mu\text{g m}^{-3}$. On a national scale, daily PM_{2.5} concentrations fluctuate slightly.

PM_{2.5} concentration in China varies significantly with the seasons. As shown in Fig. 11, PM_{2.5} concentration in winter is the highest, with an average value of 62.10 $\mu\text{g m}^{-3}$. January 2019 was the most polluted month in China, with the average PM_{2.5} concentration reaching 63.58 $\mu\text{g m}^{-3}$. The average PM_{2.5} concentration was 47.39 $\mu\text{g m}^{-3}$ in summer. The average concentration of PM_{2.5} in spring and autumn was 54.21 and 52.26 $\mu\text{g m}^{-3}$, respectively, indicating similar levels of pollution.

5 Conclusions

It is essential to collect the spatiotemporal evolution characteristics regarding the concentration of PM_{2.5} for air pollution prevention and containment. Based on the linear hybrid machine learning model, this paper used the AOD data of Himawari-8 to invert the concentration of PM_{2.5} in China and obtain its distribution characteristics. The model perfor-

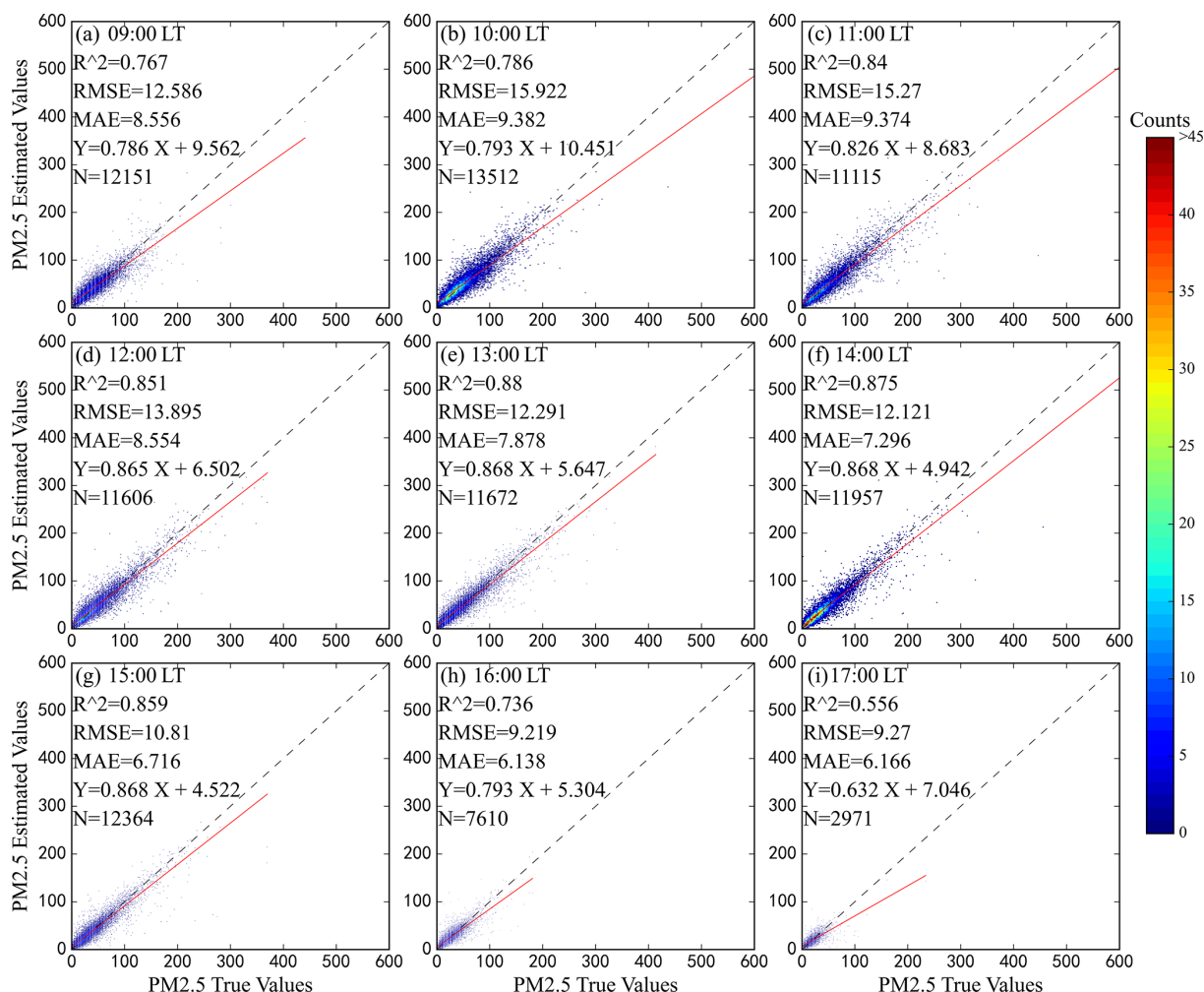


Figure 6. Density scatterplot of actual hourly PM_{2.5} values (x axis) and model-estimated values (y axis) in hourly PM_{2.5} estimates in China from (a) 09:00 LT to (i) 17:00 LT. R^2 represents the determination coefficient; RMSE represents root mean square error; MAE represents mean absolute error; N represents the number of samples. The equation with terms Y and X represents the fitting relationship between the actual and estimated PM_{2.5} values. Dashed black line represents 1 : 1 line, and red line represents best-fit line from linear regression.

mance and inversion results are analyzed and summarized below:

1. In the RGD-LHMLM obtained from linear fitting, the DNN accounted for the largest proportion with a weight coefficient of 0.62. The R^2 of RGD-LHMLM was 0.84, and its generalization ability was significantly better than that of a single model (DNN, 0.80; GBRT, 0.81; RF, 0.79). Moreover, RMSE and MAE were 12.92 and $8.01 \mu\text{g m}^{-3}$, respectively.
2. RGD-LHMLM was spatially stable, with $R^2 > 0.7$ in more than 70 % of sites as well as $\text{RMSE} < 20 \mu\text{g m}^{-3}$ and $\text{MAE} < 15 \mu\text{g m}^{-3}$ in more than 95 % of sites. These sites are mainly located in densely populated and industrially developed areas. The correlation difference between the inversion factor and PM_{2.5} in various seasons would lead to seasonal variations in the model per-

formance. In addition, the performance was the worst in summer with an average R^2 of 0.71; winter showed the best performance with an average R^2 of 0.84. The diurnal variation in the model inversion effect is also obvious, and the 11:00–14:00 LT model usually has better performance.

3. Changes in the spatiotemporal characteristics were obvious in the concentration of PM_{2.5} in China. In other words, North China and East China had the highest concentration of PM_{2.5} with an average annual concentration of $82.86 \mu\text{g m}^{-3}$, whereas Inner Mongolia, Qinghai, Tibet, and other regions had low pollution levels with an average annual concentration of PM_{2.5} below $40 \mu\text{g m}^{-3}$. In winter, the concentration of PM_{2.5} was higher with an average of $62.10 \mu\text{g m}^{-3}$, whereas the pollution was lighter in summer with an average con-

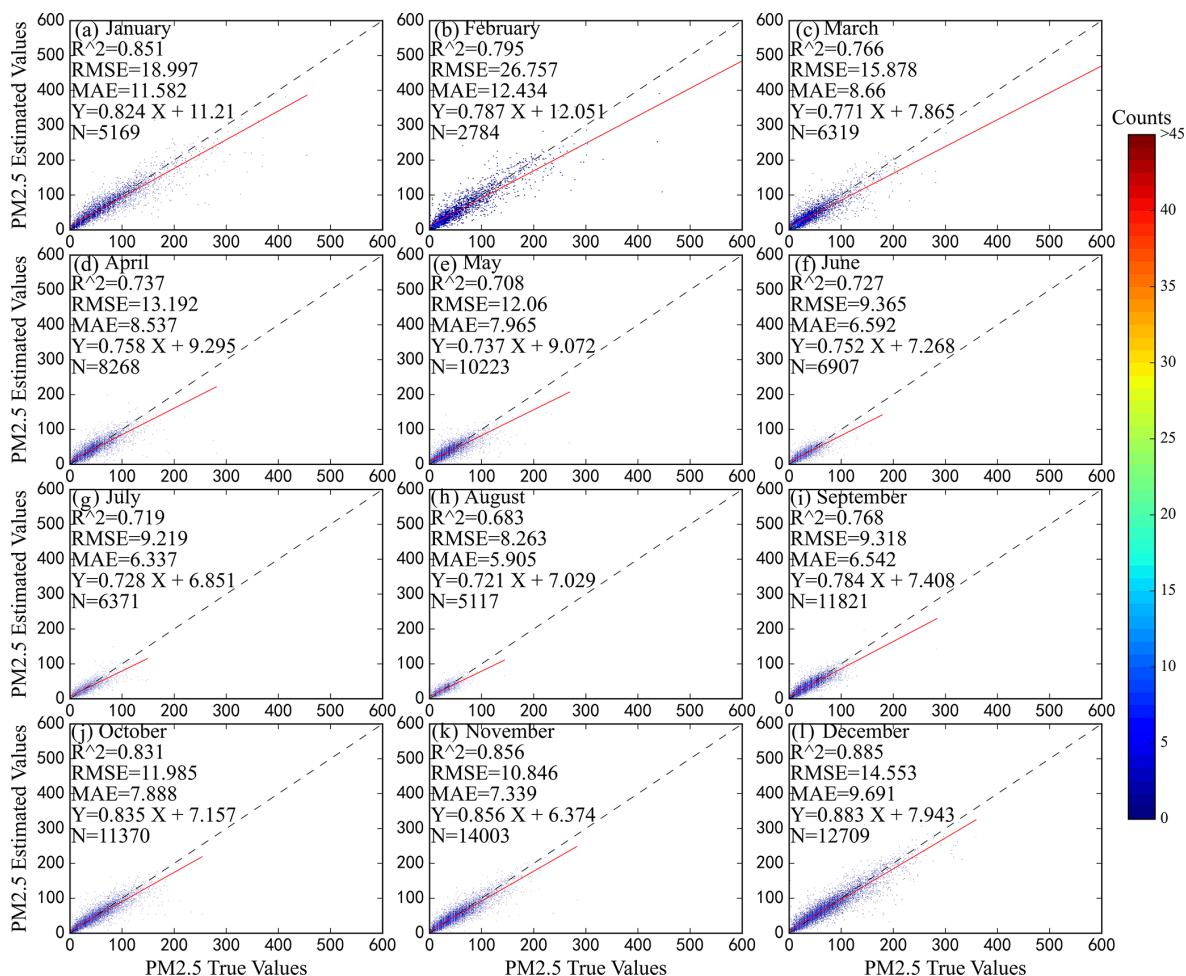


Figure 7. Same as Fig. 6 but for monthly PM_{2.5} estimates.

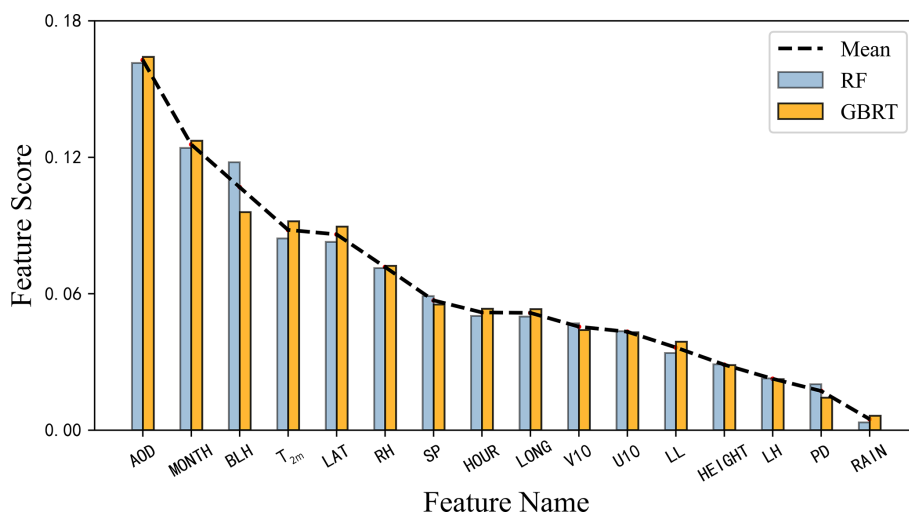


Figure 8. Score (y axis) for each model contributing feature factor (x axis) for the RF (blue) and GBRT (orange). Dashed line represents the mean values.

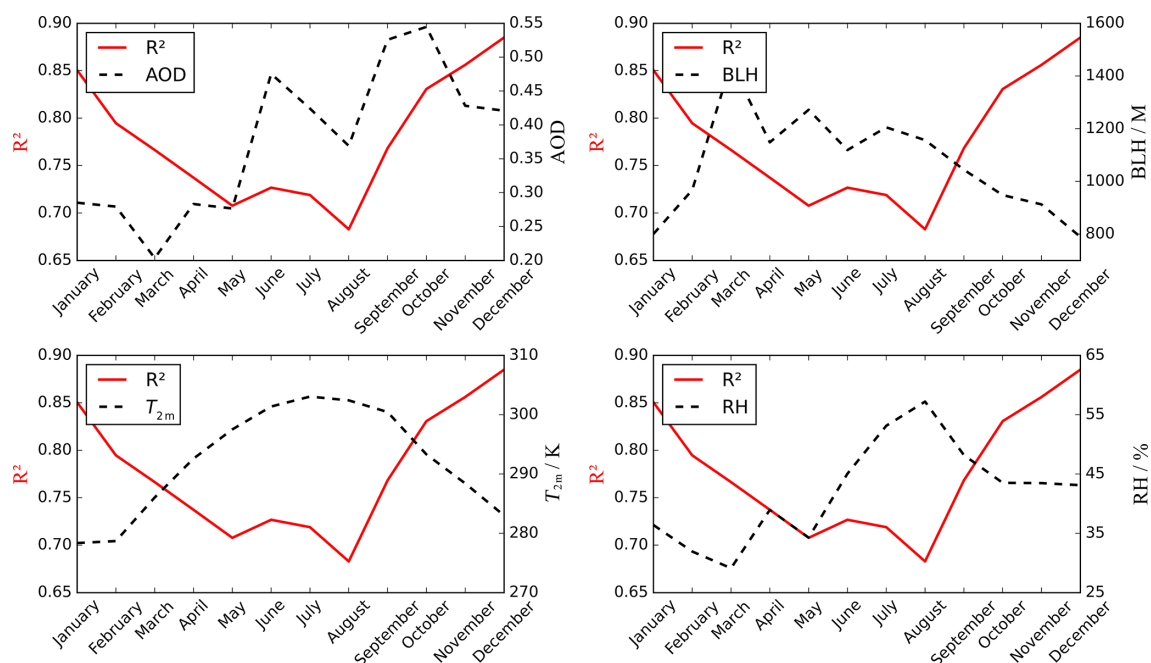


Figure 9. Annual variability (x axis) in monthly average of meteorological parameters AOD, BLH (m), T_{2m} (K), and RH (%) (right y axis) and R^2 (left y axis).

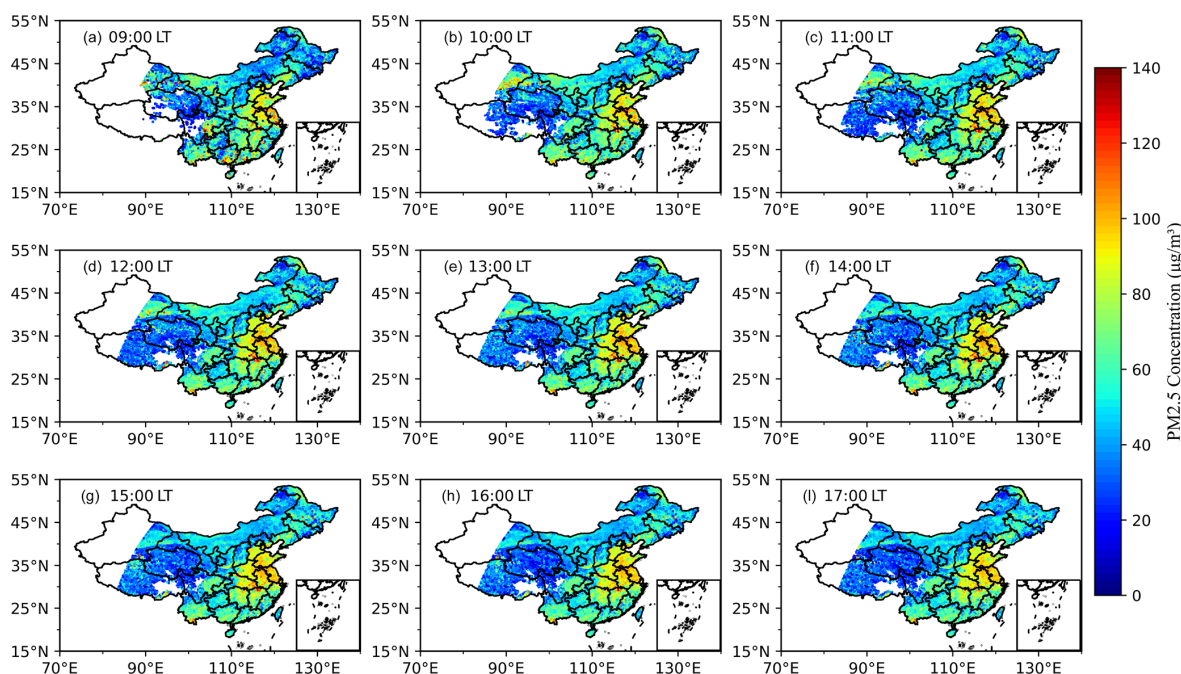


Figure 10. Hourly spatial distribution of PM_{2.5} concentration in China at different local times from (a) 09:00 LT to (i) 17:00 LT.

centration of PM_{2.5} being reported $47.39 \mu\text{g}/\text{m}^3$. In the most polluted areas, the peak concentration of PM_{2.5} could reach $85.05 \mu\text{g}/\text{m}^3$, but the daily PM_{2.5} concentration fluctuated slightly.

In conclusion, RGD-LHMLM can accurately measure the concentration of PM_{2.5} and demonstrate the seasonal evolution of pollutants. These results can help control the local pollution. This study also indicated that integrating multiple machine learning models improved the accuracy of fitting results effectively. For more accurate pollutant data, such mod-

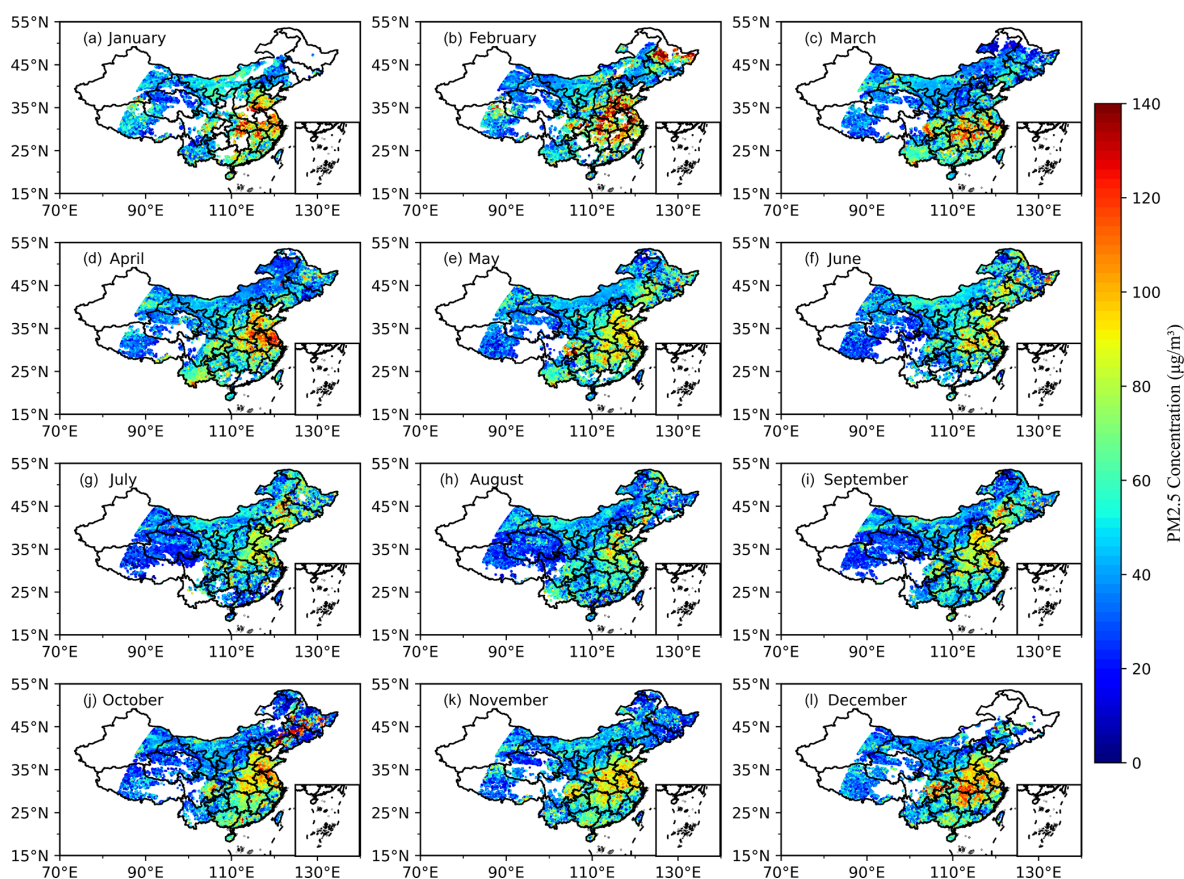


Figure 11. Same as Fig. 10 but for monthly spatial distribution.

els can be employed to fit the PM_{2.5} in the future with more parameters closely related to PM_{2.5}. However, there are some vacant values in the results of this study. There are also no data for some areas. Thus, other satellite data can be used in future studies to solve this problem.

Code and data availability. Datasets and code related to this paper can be requested from the corresponding author (chenbin@lzu.edu.cn). The PM_{2.5} data download address is <http://www.cnemc.cn> (CNEMC, 2021); Himawari-8 AOD data provided by the Japan Meteorological Agency can be downloaded from <https://www.eorc.jaxa.jp/ptree/> (JAXA, 2021); ERA5 meteorological data can be downloaded from the European Centre for Medium-Range Weather Forecasts (ECMWF) at <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=form> (ECMWF, 2021); the ground elevation SRTM3 data download address is <https://srtm.csi.cgiar.org/srtmdata/> (CGIAR Consortium for Spatial Information, 2021); NASA's Socioeconomic Data and Applications Center population density data download address is <http://sedac.ciesin.columbia.edu/data/collection/gpw-v4/documentation> (SEDAC, 2021).

Author contributions. BC proposed the content of the study. ZS performed data processing, model building, result analysis, and article writing. YH, LD, and TY checked the content of the article.

Competing interests. The authors declare that they have no conflict of interest.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. We thank the China National Environmental Monitoring Center, Japan Meteorological Agency, European Centre for Medium-Range Weather Forecasts, NASA, and the National Imagery and Mapping Agency of the US Department of Defense.

Financial support. This research has been supported by the National Key Research and Development Program of China (grant no. 2019YFA0606800), the National Natural Science Foundation of China (grant no. 41775021), and the Fundamental Research Funds for the Central Universities (grant no. lzujbky-2019-43).

Review statement. This paper was edited by Omar Torres and reviewed by three anonymous referees.

References

- Adams, M. D., Massey, F., Chastko, K., and Cupini, C.: Spatial modelling of particulate matter air pollution sensor measurements collected by community scientists while cycling, land use regression with spatial cross-validation, and applications of machine learning for data correction, *Atmos. Environ.*, 230, <https://doi.org/10.1016/j.atmosenv.2020.117479>, 2020.
- Apte, J. S., Marshall, J. D., Cohen, A. J., and Brauer, M.: Addressing Global Mortality from Ambient PM_{2.5}, *Environ. Sci. Technol.*, 49, 8057–8066, <https://doi.org/10.1021/acs.est.5b01236>, 2015.
- Bessho, K., Date, K., Hayashi, M., Ikeda, A., Imai, T., Inoue, H., Kumagai, Y., Miyakawa, T., Murata, H., Ohno, T., Okuyama, A., Oyama, R., Sasaki, Y., Shimazu, Y., Shimoji, K., Sumida, Y., Suzuki, M., Taniguchi, H., Tsuchiyama, H., Uesawa, D., Yokota, H., and Yoshida, R.: An Introduction to Himawari-8/9-Japan's New-Generation Geostationary Meteorological Satellites, *J. Meteorol. Soc. Jpn.*, 94, 151–183, <https://doi.org/10.2151/jmsj.2016-009>, 2016.
- Breiman, L.: Random forests, *Mach. Learn.*, 45, 5–32, <https://doi.org/10.1023/A:1010933404324>, 2001.
- CGIAR Consortium for Spatial Information: SRTM3, available at: <https://srtm.csi.cgiar.org/srtmdata/>, CGIAR [data set], last access: 1 July 2021.
- Chen, B. J., You, S. X., Ye, Y., Fu, Y. Y., Ye, Z. R., Deng, J. S., Wang, K., and Hong, Y.: An interpretable self-adaptive deep neural network for estimating daily spatially-continuous PM_{2.5} concentrations across China, *Sci. Total Environ.*, 768, 144724, <https://doi.org/10.1016/j.scitotenv.2020.144724>, 2021.
- Chen, J. P., Yin, J. H., Zang, L., Zhang, T. X., and Zhao, M. D.: Stacking machine learning model for estimating hourly PM_{2.5} in China based on Himawari 8 aerosol optical depth data, *Sci. Total Environ.*, 697, 134021, <https://doi.org/10.1016/j.scitotenv.2019.134021>, 2019.
- China: Ambient air quality standards, GB 3095-2012, Environmental Science Press, Beijing, China, available at: <http://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/dqhjbh/dqhjbz/201203/W020120410330232398521.pdf> (last access: 1 July 2021), 2012.
- China National Environmental Monitoring Centre (CNEMC): Homepage, available at: <http://www.cnemc.cn>, last access: 1 July 2021.
- Diederik, P. K. and Jimmy, B.: Adam: A Method for Stochastic Optimization, *arXiv [preprint]*, arXiv:1412.6980, 22 December 2014.
- ECMWF: ERA5, ECMWF [data set], available at: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=form>, last access: 1 July 2021.
- Emili, E., Popp, C., Petitta, M., Riffler, M., Wunderle, S., and Zebisch, M.: PM₁₀ remote sensing from geostationary SEVIRI and polar-orbiting MODIS sensors over the complex terrain of the European Alpine region, *Remote Sens. Environ.*, 114, 2485–2499, <https://doi.org/10.1016/j.rse.2010.05.024>, 2010.
- Engel-Cox, J. A., Holloman, C. H., Coutant, B. W., and Hoff, R. M.: Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality, *Atmos. Environ.*, 38, 2495–2509, <https://doi.org/10.1016/j.atmosenv.2004.01.039>, 2004.
- Fang, X., Zou, B., Liu, X. P., Sternberg, T., and Zhai, L.: Satellite-based ground PM_{2.5} estimation using timely structure adaptive modeling, *Remote Sens. Environ.*, 186, 152–163, <https://doi.org/10.1016/j.rse.2016.08.027>, 2016.
- Friedman, J. H.: Greedy function approximation: A gradient boosting machine, *Ann. Stat.*, 29, 1189–1232, <https://doi.org/10.1214/aos/1013203451>, 2001.
- Gao, M., Guttikunda, S. K., Carmichael, G. R., Wang, Y. S., Liu, Z. R., Stanier, C. O., Saide, P. E., and Yu, M.: Health impacts and economic losses assessment of the 2013 severe haze event in Beijing area, *Sci. Total Environ.*, 511, 553–561, <https://doi.org/10.1016/j.scitotenv.2015.01.005>, 2015.
- Gui, K., Che, H. Z., Wang, Y. Q., Wang, H., Zhang, L., Zhao, H. J., Zheng, Y., Sun, T. Z., and Zhang, X. Y.: Satellite-derived PM_{2.5} concentration trends over Eastern China from 1998 to 2016: Relationships to emissions and meteorological parameters, *Environ. Pollut.*, 247, 1125–1133, <https://doi.org/10.1016/j.envpol.2019.01.056>, 2019.
- Guo, B., Zhang, D. M., Pei, L., Su, Y., Wang, X. X., Bian, Y., Zhang, D. H., Yao, W. Q., Zhou, Z. X., and Guo, L. Y.: Estimating PM_{2.5} concentrations via random forest method using satellite, auxiliary, and ground-level station dataset at multiple temporal scales across China in 2017, *Sci. Total Environ.*, 778, 146288, <https://doi.org/10.1016/j.scitotenv.2021.146288>, 2021.
- Guo, J. P., Xia, F., Zhang, Y., Liu, H., Li, J., Lou, M. Y., He, J., Yan, Y., Wang, F., Min, M., and Zhai, P. M.: Impact of diurnal variability and meteorological factors on the PM_{2.5} – AOD relationship: Implications for PM_{2.5} remote sensing, *Environ. Pollut.*, 221, 94–104, <https://doi.org/10.1016/j.envpol.2016.11.043>, 2017.
- Han, Y., Wu, Y. H., Wang, T. J., Zhuang, B. L., Li, S., and Zhao, K.: Impacts of elevated-aerosol-layer and aerosol type on the correlation of AOD and particulate matter with ground-based and satellite measurements in Nanjing, southeast China, *Sci. Total Environ.*, 532, 195–207, <https://doi.org/10.1016/j.scitotenv.2015.05.136>, 2015.
- Hoff, R. M. and Christopher, S. A.: Remote Sensing of Particulate Pollution from Space: Have We Reached the Promised Land?, *J. Air Waste Manage.*, 59, 645–675, <https://doi.org/10.3155/1047-3289.59.6.645>, 2009.
- Hu, X. F., Waller, L. A., Al-Hamdan, M. Z., Crosson, W. L., Estes, M. G., Estes, S. M., Quattrochi, D. A., Sarnat, J. A., and Liu, Y.: Estimating ground-level PM_{2.5} concentrations in the southeastern US using geographically weighted regression, *Environ. Res.*, 121, 1–10, <https://doi.org/10.1016/j.envres.2012.11.003>, 2013.
- JAXA: Himawari-8 AOD, JAXA [data set], available at: <https://www.eorc.jaxa.jp/ptree/>, last access: 1 July 2021.
- Jiang, Y., Yang, K., Shao, C., Zhou, X., Zhao, L., Chen, Y., and Wu, H.: A downscaling approach for constructing high-resolution precipitation dataset over the Tibetan Plateau from ERA5 reanalysis, *Atmos. Res.*, 256, 105574, <https://doi.org/10.1016/j.atmosres.2021.105574>, 2021.

- Johnson, N. E., Bonczak, B., and Kontokosta, C. E.: Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment, *Atmos. Environ.*, 184, 9–16, <https://doi.org/10.1016/j.atmosenv.2018.04.019>, 2018.
- Lee, H. J., Coull, B. A., Bell, M. L., and Koutrakis, P.: Use of satellite-based aerosol optical depth and spatial clustering to predict ambient PM_{2.5} concentrations, *Environ. Res.*, 118, 8–15, <https://doi.org/10.1016/j.envres.2012.06.011>, 2012.
- Li, T. W., Shen, H. F., Zeng, C., Yuan, Q. Q., and Zhang, L. P.: Point-surface fusion of station measurements and satellite observations for mapping PM_{2.5} distribution in China: Methods and assessment, *Atmos. Environ.*, 152, 477–489, <https://doi.org/10.1016/j.atmosenv.2017.01.004>, 2017a.
- Li, T. W., Shen, H. F., Yuan, Q. Q., Zhang, X. C., and Zhang, L. P.: Estimating Ground-Level PM_{2.5} by Fusing Satellite and Station Observations: A Geo-Intelligent Deep Learning Approach, *Geophys. Res. Lett.*, 44, 11985–11993, <https://doi.org/10.1002/2017gl075710>, 2017b.
- Li, Z. Q., Zhang, Y., Shao, J., Li, B. S., Hong, J., Liu, D., Li, D. H., Wei, P., Li, W., Li, L., Zhang, F. X., Guo, J., Deng, Q., Wang, B. X., Cui, C. L., Zhang, W. C., Wang, Z. Z., Lv, Y., Xu, H., Chen, X. F., Li, L., and Qie, L. L.: Remote sensing of atmospheric particulate mass of dry PM_{2.5} near the ground: Method validation using ground-based measurements, *Remote Sens. Environ.*, 173, 59–68, <https://doi.org/10.1016/j.rse.2015.11.019>, 2016.
- Lim, C. H., Ryu, J., Choi, Y., Jeon, S. W., and Lee, W. K.: Understanding global PM_{2.5} concentrations and their drivers in recent decades (1998–2016), *Environ. Int.*, 144, 106011, <https://doi.org/10.1016/j.envint.2020.106011>, 2020.
- Liu, Y., Sarnat, J. A., Kilaru, A., Jacob, D. J., and Koutrakis, P.: 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>, 2005.
- Liu, Y., Cao, G. F., Zhao, N. Z., Mulligan, K., and Ye, X. Y.: 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>, 2018.
- Lv, B. L., Hu, Y. T., Chang, H. H., Russell, A. G., Cai, J., Xu, B., and Bai, Y. Q.: Daily estimation of ground-level PM_{2.5} concentrations at 4 km resolution over Beijing-Tianjin-Hebei by fusing MODIS AOD and ground observations, *Sci. Total Environ.*, 580, 235–244, <https://doi.org/10.1016/j.scitotenv.2016.12.049>, 2017.
- Mao, F. Y., Hong, J., Min, Q. L., Gong, W., Zang, L., and Yin, J. H.: Estimating hourly full-coverage PM_{2.5} over China based on TOA reflectance data from the Fengyun-4A satellite, *Environ. Pollut.*, 270, 116119, <https://doi.org/10.1016/j.envpol.2020.116119>, 2021.
- Miao, Y. C., Liu, S. H., Guo, J. P., Huang, S. X., Yan, Y., and Lou, M. Y.: Unraveling the relationships between boundary layer height and PM_{2.5} pollution in China based on four-year radiosonde measurements, *Environ. Pollut.*, 243, 1186–1195, <https://doi.org/10.1016/j.envpol.2018.09.070>, 2018.
- Pan, Z. X., Mao, F. Y., Wang, W., Zhu, B., Lu, X., and Gong, W.: Impacts of 3D Aerosol, Cloud, and Water Vapor Variations on the Recent Brightening during the South Asian Monsoon Season, *Remote Sens.-Basel*, 10, 651, <https://doi.org/10.3390/rs10040651>, 2018.
- Pun, V. C., Kazemiparkouhi, F., Manjourides, J., and Suh, H. H.: Long-Term PM_{2.5} Exposure and Respiratory, Cancer, and Cardiovascular Mortality in Older US Adults, *Am. J. Epidemiol.*, 186, 961–969, <https://doi.org/10.1093/aje/kwx166>, 2017.
- Qin, K., Wang, L. Y., Wu, L. X., Xu, J., Rao, L. L., Letu, H., Shi, T. W., and Wang, R. F.: A campaign for investigating aerosol optical properties during winter hazes over Shijiazhuang, China, *Atmos. Res.*, 198, 113–122, <https://doi.org/10.1016/j.atmosres.2017.08.018>, 2017.
- Schönlau, M.: Boosted regression (boosting): An introductory tutorial and a Stata plugin, *Stata J.*, 5, 330–354, <https://doi.org/10.1177/1536867x0500500304>, 2005.
- SEDAC: Gridded Population of the World (GPW), v4, available at: <http://sedac.ciesin.columbia.edu/data/collection/gpw-v4/documentation>, NASA [data set], last access: 1 July 2021.
- Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., de Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein, A., Viegi, G., Kloog, I., and Schwartz, J.: Estimation of daily PM₁₀ and PM_{2.5} concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model, *Environ. Int.*, 124, 170–179, <https://doi.org/10.1016/j.envint.2019.01.016>, 2019.
- Tian, J. and Chen, D. M.: A semi-empirical model for predicting hourly ground-level fine particulate matter (PM_{2.5}) concentration in southern Ontario from satellite remote sensing and ground-based meteorological measurements, *Remote Sens. Environ.*, 114, 221–229, <https://doi.org/10.1016/j.rse.2009.09.011>, 2010.
- Wang, W., Mao, F. Y., Du, L., Pan, Z. X., Gong, W., and Fang, S. H.: Deriving Hourly PM_{2.5} Concentrations from Himawari-8 AODs over Beijing-Tianjin-Hebei in China, *Remote Sens.-Basel*, 9, 858, <https://doi.org/10.3390/rs9080858>, 2017.
- Wang, X. H., Zhong, S. Y., Bian, X. D., and Yu, L. J.: Impact of 2015–2016 El Niño and 2017–2018 La Niña on PM_{2.5} concentrations across China, *Atmos. Environ.*, 208, 61–73, <https://doi.org/10.1016/j.atmosenv.2019.03.035>, 2019a.
- Wang, X. P. and Sun, W. B.: Meteorological parameters and gaseous pollutant concentrations as predictors of daily continuous PM_{2.5} concentrations using deep neural network in Beijing-Tianjin-Hebei, China, *Atmos. Environ.*, 211, 128–137, <https://doi.org/10.1016/j.atmosenv.2019.05.004>, 2019.
- Wang, X. Q., Wei, W., Cheng, S. Y., Yao, S., Zhang, H. Y., and Zhang, C.: Characteristics of PM_{2.5} and SNA components and meteorological factors impact on air pollution through 2013–2017 in Beijing, China, *Atmos. Pollut. Res.*, 10, 1976–1984, <https://doi.org/10.1016/j.apr.2019.09.004>, 2019b.
- Wei, J., Huang, W., Li, Z. Q., Xue, W. H., Peng, Y. R., Sun, L., and Cribb, M.: Estimating 1-km-resolution PM_{2.5} concentrations across China using the space-time random forest approach, *Remote Sens. Environ.*, 231, 111221, <https://doi.org/10.1016/j.rse.2019.111221>, 2019a.
- Wei, J., Li, Z., Sun, L., Peng, Y., Zhang, Z., Li, Z., Su, T., Feng, L., Cai, Z., and Wu, H.: Evaluation and uncertainty estimate of next-generation geostationary meteorological Himawari-8/AHI aerosol products, *Sci. Total Environ.*, 692, 879–891, <https://doi.org/10.1016/j.scitotenv.2019.07.326>, 2019b.
- Wei, J., Li, Z., Pinker, R. T., Wang, J., Sun, L., Xue, W., Li, R., and Cribb, M.: Himawari-8-derived diurnal variations in ground-level PM_{2.5} pollution across China using the fast space-

- time Light Gradient Boosting Machine (LightGBM), *Atmos. Chem. Phys.*, 21, 7863–7880, <https://doi.org/10.5194/acp-21-7863-2021>, 2021a.
- Wei, J., Li, Z. Q., Lyapustin, A., Sun, L., Peng, Y. R., Xue, W. H., Su, T. N., and Cribb, M.: Reconstructing 1-km-resolution high-quality PM_{2.5} data records from 2000 to 2018 in China: spatiotemporal variations and policy implications, *Remote Sens. Environ.*, 252, 112136, <https://doi.org/10.1016/j.rse.2020.112136>, 2021b.
- Wolpert, D. H.: Stacked Generalization, *Neural Networks*, 5, 241–259, [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1), 1992.
- Xu, J. H., Lindqvist, H., Liu, Q. F., Wang, K., and Wang, L.: Estimating the spatial and temporal variability of the ground-level NO₂ concentration in China during 2005–2019 based on satellite remote sensing, *Atmos. Pollut. Res.*, 12, 57–67, <https://doi.org/10.1016/j.apr.2020.10.008>, 2021.
- Yang, X. C., Jiang, L., Zhao, W. J., Xiong, Q. L., Zhao, W. H., and Yan, X.: Comparison of Ground-Based PM_{2.5} and PM₁₀ Concentrations in China, India, and the US, *Int. J. Env. Res. Pub. He.*, 15, 1382, <https://doi.org/10.3390/ijerph15071382>, 2018.
- Yesilkanat, C. M.: Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm, *Chaos Soliton. Fract.*, 140, 110210, <https://doi.org/10.1016/j.chaos.2020.110210>, 2020.
- Yin, J. H., Mao, F. Y., Zang, L., Chen, J. P., Lu, X., and Hong, J.: Retrieving PM_{2.5} with high spatio-temporal coverage by TOA reflectance of Himawari-8, *Atmos. Pollut. Res.*, 12, 14–20, <https://doi.org/10.1016/j.apr.2021.02.007>, 2021.
- Yoshida, M., Kikuchi, M., Nagao, T. M., Murakami, H., Nomaki, T., and Higurashi, A.: Common Retrieval of Aerosol Properties for Imaging Satellite Sensors, *J. Meteorol. Soc. Jpn. Ser. II*, 96B, 193–209, <https://doi.org/10.2151/jmsj.2018-039>, 2018.
- Yumimoto, K., Nagao, T. M., Kikuchi, M., Sekiyama, T. T., Murakami, H., Tanaka, T. Y., Ogi, A., Irie, H., Khatri, P., Okumura, H., Arai, K., Morino, I., Uchino, O., and Maki, T.: Aerosol data assimilation using data from Himawari-8, a next-generation geostationary meteorological satellite, *Geophys. Res. Lett.*, 43, 5886–5894, <https://doi.org/10.1002/2016gl069298>, 2016.
- Zang, L., Mao, F. Y., Guo, J. P., Gong, W., Wang, W., and Pan, Z. X.: Estimating hourly PM₁ concentrations from Himawari-8 aerosol optical depth in China, *Environ. Pollut.*, 241, 654–663, <https://doi.org/10.1016/j.envpol.2018.05.100>, 2018.
- Zhang, L., Guo, X. M., Zhao, T. L., Gong, S. L., Xu, X. D., Li, Y. Q., Luo, L., Gui, K., Wang, H. L., Zheng, Y., and Yin, X. F.: A modelling study of the terrain effects on haze pollution in the Sichuan Basin, *Atmos. Environ.*, 196, 77–85, <https://doi.org/10.1016/j.atmosenv.2018.10.007>, 2019.
- Zhang, T. H., Zhu, Z. M., Gong, W., Zhu, Z. R., Sun, K., Wang, L. C., Huang, Y. S., Mao, F. Y., Shen, H. F., Li, Z. W., and Xu, K.: Estimation of ultrahigh resolution PM_{2.5} concentrations in urban areas using 160 m Gaofen-1 AOD retrievals, *Remote Sens. Environ.*, 216, 91–104, <https://doi.org/10.1016/j.rse.2018.06.030>, 2018.
- Zhang, T. X., Zang, L., Wan, Y. C., Wang, W., and Zhang, Y.: Ground-level PM_{2.5} estimation over urban agglomerations in China with high spatiotemporal resolution based on Himawari-8, *Sci. Total Environ.*, 676, 535–544, <https://doi.org/10.1016/j.scitotenv.2019.04.299>, 2019.
- Zhang, Z., Wu, W., Fan, M., Tao, M., Wei, J., Jin, J., Tan, Y., and Wang, Q.: Validation of Himawari-8 aerosol optical depth retrievals over China, *Atmos. Environ.*, 199, 32–44, <https://doi.org/10.1016/j.atmosenv.2018.11.024>, 2019.
- Zheng, C., Zhao, C., Zhu, Y., Wang, Y., Shi, X., Wu, X., Chen, T., Wu, F., and Qiu, Y.: Analysis of influential factors for the relationship between PM_{2.5} and AOD in Beijing, *Atmos. Chem. Phys.*, 17, 13473–13489, <https://doi.org/10.5194/acp-17-13473-2017>, 2017.