Science of the Total Environment 805 (2022) 150338

Contents lists available at ScienceDirect

Science of the Total Environment

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

Obtaining vertical distribution of $PM_{2.5}$ from CALIOP data and machine learning algorithms

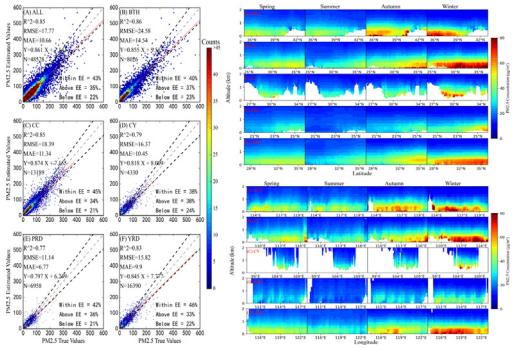
Bin Chen *, Zhihao Song, Feng Pan, Yue Huang

College of Atmospheric Science, Lanzhou University, Lanzhou 730000, China

HIGHLIGHTS

- AOD-$PM_{2.5}$ ET model has a better overall performance with an R^2 of 0.85.
- Correlation between optimal layer AOD and $PM_{2.5}$ exhibited regional differences.
- Optimal layer's weight was set to 1, other layers depended on proportion of AOD.
- $PM_{2.5}$ vertical concentration from 2015 to 2019 was obtained based on the ET model.
- $PM_{2.5}$ concentration showed a decreasing trend in China from 2015 to 2019.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 30 June 2021
Received in revised form 10 September 2021
Accepted 10 September 2021
Available online 15 September 2021

Editor: Anastasia Paschalidou

Keywords:
$PM_{2.5}$ vertical distribution
CALIOP
AOD
Machine learning

ABSTRACT

Aerosol optical depth (AOD) has been widely used to estimate the near-surface $PM_{2.5}$ (fine particulate matter with particle size less than 2.5 μm). However, the total-column AOD obtained by passive remote sensing instruments can neither distinguish the contribution of AOD in various altitude layers nor obtain the vertical $PM_{2.5}$ concentration. In this study, we compared several AOD-$PM_{2.5}$ models including Extra Trees (ET), Random Forest (RF), Deep Neural Network (DNN), and Gradient Boosting Regression Tree (GBRT), and analyzed the corresponding results using AOD of different altitudes and auxiliary data from the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP). The results indicate that the ET model performs best in terms of the model effectiveness and feature interpretation on the training dataset. We conclude that the feature importance of the bottom layer AOD is higher than that of the upper and total column AOD. The results showed that regional differences existed in the optimal height of the AOD-$PM_{2.5}$ correlation in study area. The results of cross-validation indicate that ET manages the most appealing overall performance with an R^2 (RMSE) of 0.85 (17.77 μg/m^3). Regarding the 729 sites involved in this study, 73% had $R^2 > 0.7$, and the region or season with higher AOD feature importance achieves better model performance. The results of the AOD-$PM_{2.5}$ model in each layer were corrected using the AOD weight, to obtained the $PM_{2.5}$ vertical concentrations from 2015 to 2019. The results highlight that the high $PM_{2.5}$ concentration area is primarily near the ground and decreases with height. Additionally, the $PM_{2.5}$ vertical concentration in Beijing-Tianjin-Hebei (-1.80 μg/m^3, $P < 0.001$), Central China (-1.62 μg/m^3, $P < 0.001$), and Pearl River Delta (-0.66 μg/m^3, $P < 0.001$) show an apparent downward trend. We believe that the vertical distribution analysis of $PM_{2.5}$ can provide meaningful information for studying air pollution.

* Corresponding author.
E-mail address: chenbin@lzu.edu.cn (B. Chen).

https://doi.org/10.1016/j.scitotenv.2021.150338

1. Introduction

Atmospheric $PM_{2.5}$ has a significant impact on the global climate and environmental change (Kampa and Castanas, 2008; Wilson et al., 2010). $PM_{2.5}$ is a hazardous atmospheric pollutant, and the long-term exposure to it may cause human diseases and even death through direct or indirect effects (Brauer et al., 2016; Qin et al., 2016; Wu et al., 2018; Chen et al., 2019c; Wei et al., 2019a). For a long time, the primary means of observation $PM_{2.5}$ were environmental monitoring stations. However, the quantity of available $PM_{2.5}$ data presented regional differences due to the uneven station distribution.

Several studies indicated a good correlation between $PM_{2.5}$ and aerosol Optical Depth (AOD) obtained by satellite observation (Chu et al., 2003; Paciorek et al., 2008; Zhang et al., 2021), which facilitated the satellite remote sensing AOD inversion method of $PM_{2.5}$ becoming the primary means to study ground-level fine particulate matter (Zhang and Li, 2015; Zhang et al., 2019). Currently, many studies related to the concentration of $PM_{2.5}$ have employed AOD data from the Moderate-resolution Imaging Spectroradiometer (MODIS) (Li et al., 2018; Zeydan and Wang, 2019; Zhao et al., 2020), the Visible Infrared Imaging Radiometer Suite (VIIRS) (Wu et al., 2016), the Goddard Earth Observing System Data (Buchard et al., 2016), and the geostationary meteorological satellite Himawari-8 (Chen et al., 2019b; Sun et al., 2021). These studies utilize various statistical methods, including Semi-empirical formula (Li et al., 2016b), multiple regression (Chelani, 2019), geographically weighted regression (Ma et al., 2014; Qin et al., 2018), machine learning (Ghahremanloo et al., 2021; Wei et al., 2021a), and neural networks (Paschalidou et al., 2011; Li et al., 2017). These satellite data and research methods obtained relatively accurate high-resolution $PM_{2.5}$ (Fan et al., 2020; Gui et al., 2021; Wei et al., 2021b; Zhong et al., 2021), which effectively compensated for the lack of $PM_{2.5}$ data in China. Current studies focus on the spatiotemporal variation of $PM_{2.5}$ concentration on the surface (Li et al., 2021b), and to the best of our knowledge, few studies have explored the vertical distribution in $PM_{2.5}$.

The Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP), onboard the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) satellite, can continuously observe atmospheric clouds and aerosols globally and extracts their optical signatures (Winker et al., 2007; Omar et al., 2009; Huang et al., 2015). CALIOP is a cloud-aerosol lidar with orthogonal polarization that provides the vertical distribution characteristics of high-resolution clouds and aerosols to effectively determine the cloud and the aerosol types (Chen et al., 2010; Rajeevan et al., 2013; Huang et al., 2018; Liu et al., 2019; Niu et al., 2019). The CALIOP data can be used to distinguish planetary boundary layers from free atmospheres (Liu et al., 2015; Zhu et al., 2018), and have also been applied to explore the fine particle concentration. For example, Ma et al. (2020) obtained the global distribution and diurnal variation of fine particulate matter from 2007 to 2016 using the CALIOP observation data and developed a mass concentration formula. Based on the volume mass modeling method, Toth et al. (2019) calculated the concentration of particulate matter in the United States from 2008 to 2009 using CALIOP data. CALIOP can also provide better AOD data quality, with an R^2 exceeding 0.9 compared with the ground-based AOD data from AERONET (Schuster et al., 2012; Kumar et al., 2018). In summary, we believe that it is feasible to invert the $PM_{2.5}$ data from the CALIOP observation data.

Inspired by the above content, this study established an AOD-$PM_{2.5}$ model based on the optimal layer AOD from CALIOP data and the hourly mean $PM_{2.5}$ concentration from ground observation, which were used to obtain the vertical PM2.5 distribution. Specifically, in this work, we discussed the difference between the AOD of different altitudes and total-column AOD for the $PM_{2.5}$ estimation. We then present the optimal correlation height between $PM_{2.5}$ and AOD in different regions. Finally, based on the Extra Trees (ET) model, we estimated the vertical $PM_{2.5}$ concentration in five study regions by utilizing the AOD and meteorological factors to ultimately derive the vertical spatial and temporal distribution of $PM_{2.5}$ from 2015 to 2019.

2. Data

2.1. Ground PM2.5 observation data

The ground-level $PM_{2.5}$ used in this study was obtained from the China Environmental Monitoring Center (CEMC) Air Quality Real-time Publicity System (Li et al., 2021a), which provides the hourly mean $PM_{2.5}$ concentration data ($\mu g/m^3$), calibrated and quality-controlled according to national standards GB 3095-2012 (China, 2012). Fig. 1 shows the spatial distribution of environmental monitoring stations in the study area, which are divided into five main research areas: Beijing-Tianjin-Hebei region (BTH, 79 stations), Central China region (CC, 209 stations), Cheng-Yu region (CY, 102 stations), Pearl River Delta region (PRD, 102 stations) and Yangtze River Delta region (YRD, 237 stations). The total number of stations is 729, and the time range of the data is from 2015 to 2019, providing five years of hourly averaged $PM_{2.5}$.

2.2. CALIOP observation datasets

CALIPSO, a polar-orbiting satellite, was launched on April 28, 2006. The onboard CALIOP observes cloud and aerosol particles at wavelengths of 532 and 1064 nm; it provides optical characteristics such as the particle depolarization ratio (degree of particle irregularity) and color ratio (particle size) in the vertical profile (Winker et al., 2013; Ghomashi and Khalesifard, 2020). The CALIOP Level-2 Aerosol Profiles (APRO) Version 4.20 data (including daytime and nighttime data) provides the aerosol extinction coefficient (*EX*) from the ground to the top of the atmosphere and generates AOD with a vertical resolution of 60 m and a horizontal resolution of 5 km (Solanki and Singh, 2014). In this study, we utilized the AOD, depolarization ratio, and color ratio data from the CALIOP APRO produced from 2015 to 2019 to estimate the $PM_{2.5}$.

2.3. Meteorological data

This work considered several meteorological factors, including the ozone concentration density (O3, per m^3), air pressure (P, hPa), air temperature (T, °C), relative humidity (RH, %), and wind speed (U, V, m/s). These meteorological data were derived from the MERRA-2 data product provided to the CALIPSO project by the Global modeling and Assimilation Office (GMAO). These data are ancillary to the CALIOP published product and are available directly from the Level 2 APRO product with resolutions identical to those of the CALIOP data. The height of the planetary boundary layer (BLH, m) was provided by ERA-5 data from the European Centre for Medium-Range Weather Forecasts (ECWMF).

2.4. Data matching

The AOD and meteorological factors were temporally and spatially matched with the ground station data. To match the satellite data with the station data, we averaged the satellite data within different radii (5, 10, 15, 20, 30, 50, 75, and 100 km) of a ground station to represent the satellite's AOD. Additionally, the $PM_{2.5}$ data obtained from the ground station within 30 min and centered on the CALIPSO transit time were collected as ground observations. The matching method for the meteorological data was the same as for the satellite data.

3. Methodology

3.1. Extra trees

An ensemble learning method ET (Geurts et al., 2006), was used and is similar to the random forest (RF) comprising multiple decision trees.

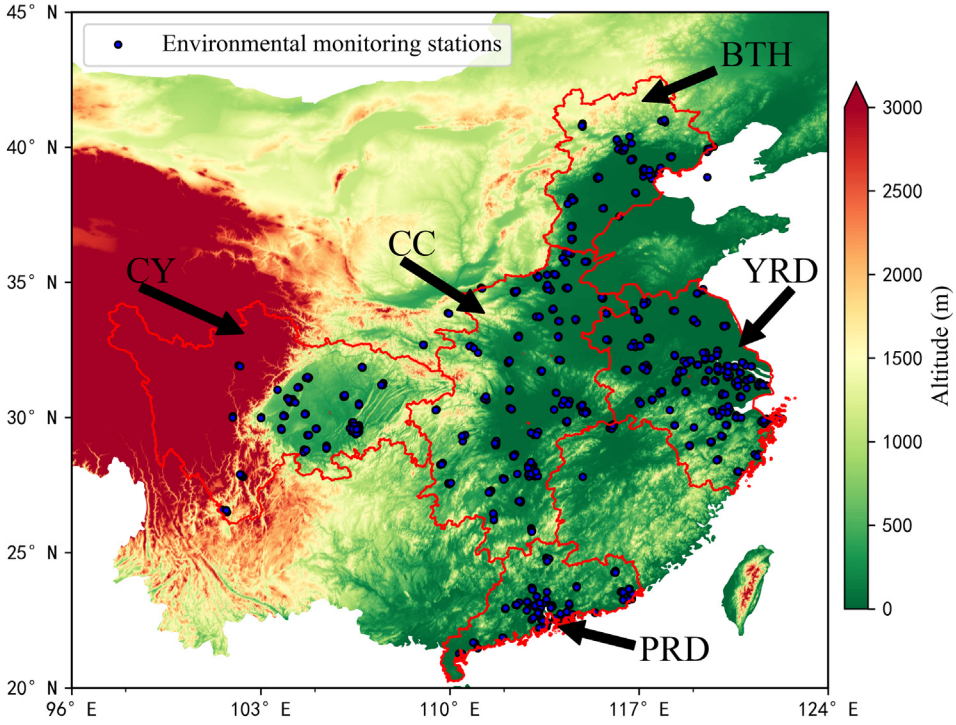


Fig. 1. Diagram of the study area.

However, the RF applies the bagging model and random puts back samples, while ET exploits all samples, and only the sample features are randomly selected. Additionally, the ET characteristics include the following:

(1) Random features: In a sample set D, each sample has M attributes, of which m are randomly selected satisfying the condition $m \ll M$.
(2) Splitting randomness: When each node in the decision tree needs to be split, one of the m attributes is randomly selected as the basis for splitting, and the attribute is randomly selected.

Once the samples and features are selected, the decision trees are constructed according to the optimal bifurcation attribute. The above steps are repeated until many decision trees are constructed to form the ET. Finally, the average classification or regression results obtained from all decision trees in the ET were used as the final outputs. Our study found that the ET model has the best fitting performance of AOD and PM2.5, and the model is also interpretable. In addition, in some studies, the strong data fitting ability of ET model has been proved (Ahmad et al., 2018; Qin et al., 2020; Zhang et al., 2022).

A schematic diagram of the ET is shown in Fig. 2, which explains the use of machine learning methods to derive $PM_{2.5}$.

3.2. Comparison models

Three most commonly used machine learning models including RF, Deep Neural Network (DNN), and Gradient Boosting Regression Trees (GBRT) are selected to compare with ET model. Details of these models can be found in the study of Song et al. (2021). RF and GBRT are ensemble learning methods, that are often used for data modeling (Chen et al., 2019a; Wei et al., 2019b; Gui et al., 2020; Chen et al., 2021). While DNN is a deep learning model, comprising an input layer, some hidden layers, and an output layer. The nonlinear transformation between each layer was conducted using the activation function (Wang and Sun, 2019; Lu et al., 2021).

3.3. Model validation

The model performance was evaluated using the 10-fold cross-validation scheme; the training data were divided into ten parts, nine of which were used as the training set and the remaining one as the validation set (Rodriguez et al., 2010). The error between the estimated and true values was evaluated using several evaluation indexes, including the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and bias. The formula for each evaluation index is as follows:

$$R^2 = 1 - \frac{ss_{res}}{SS_{tot}} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3}$$

$$Bias = \hat{y}_i - y_i \tag{4}$$

where $\hat{y}_i$ represents the estimated value, y_i the true value, ss_{res} the error between the regression data and mean value, SS_{tot} is the error between the actual data and the mean value, with "the mean value" referring to the mean value of the true value.

3.4. Feature importance

Additionally, we used the feature importance to evaluate the contribution of each factor to the model (total gains of splits that use the feature, not the physical contribution (Wei et al., 2021a)). The feature importance increases with the contribution of the feature to the

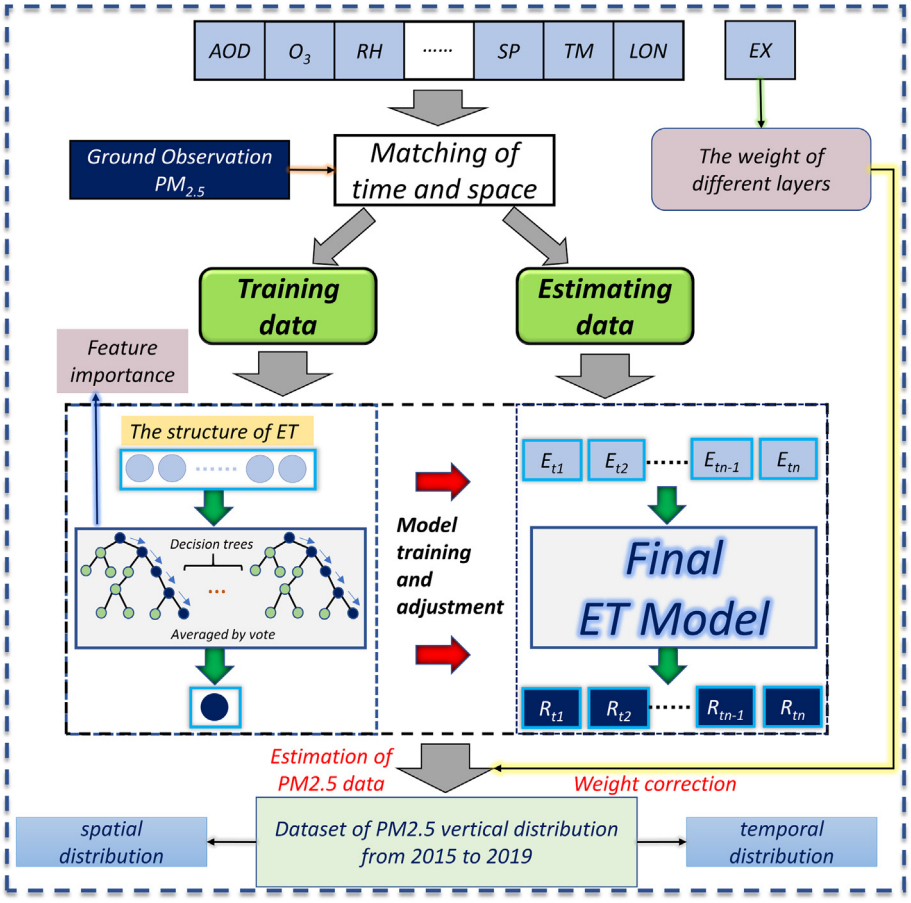


Fig. 2. Schematic diagram of the model (E_{t1} is the estimated data set at time t1, R_{t1} is the estimation result of $PM_{2.5}$ at time t1).

model. In general, we use the Gini index (GI) to calculate the feature importance (FI) (Calle and Urrea, 2011), which is expressed as:

$$GI_m = \sum_{k=1}^{k} p_{mk}(1-p_{mk}) = 1-\sum_{k=1}^{k} p_{mk}^2 \quad (5)$$

where GI_m represents the GI of node "m", and k the number of categories. Subsequently, P_{mk} denotes the proportion of category k in node m. The importance of one feature (X_j) in node m is represented that the Gini index changes before and after the node m branch. It is given by the following:

$$\mathrm{FI}_{jm} = GI_m - GI_l - GI_r \quad (6)$$

where, GI_l and GI_r represent the GI of the two new nodes after branching, respectively. For one feature (X_j) in an ET model with n decision trees ($i = 1, 2, \ldots, n$), its importance score can be expressed as:

$$FI_j = \sum_{i=1}^{n} \Delta FI_{ij} = \sum_{i=1}^{n} \sum_{m \in M} FI_{jm} \quad (7)$$

where ΔFI_{ij} represents the importance of X_j in the ith tree when the node of feature Xj in decision tree j belongs to set M. Finally, the normalization process is conducted on the importance scores obtained for each feature.

4. Results and discussion

4.1. Relationship between the accuracy of different models and the data matching radii

As presented in Table 1, the total number of matched data samples from 2015 to 2019 was obtained by matching the ground stations within different radii of the CALIOP trajectory (5, 10, 15, 20, 30, 50, 75,

Table 1
Comparison of inversion accuracy of different models.

			ET			RF			DNN			GBRT	
Radius	Sample size	R^2	RMSE	IM	R^2	RMSE	IM	R^2	RMSE	IM	R2	RMSE	IM
5 km	1235	0.75	23.90	0.18	0.71	25.50	0.19	0.74	24.13	\	0.70	25.81	0.23
10 km	2637	0.81	20.50	0.17	0.79	21.30	0.18	0.78	21.56	\	0.75	23.16	0.21
15 km	4085	0.82	19.80	0.18	0.80	20.80	0.19	0.81	20.22	\	0.77	22.23	0.23
20 km	5569	0.84	19.55	0.17	0.83	19.73	0.19	0.83	19.85	\	0.78	22.85	0.22
30 km	8571	0.84	18.97	0.16	0.82	20.13	0.18	0.84	18.69	\	0.75	23.44	0.21
50 km	16,275	0.85	17.81	0.15	0.83	18.62	0.17	0.84	18.02	\	0.78	21.61	0.20
75 km	25,486	0.85	17.34	0.14	0.84	17.72	0.15	0.85	17.21	\	0.76	22.09	0.17
100 km	35,767	0.86	16.73	0.13	0.84	17.70	0.14	0.85	17.00	\	0.75	22.46	0.16

IM: The feature importance of AOD.

and 100 km). We use the ET, RF, DNN, and GBRT models for the datasets to fit the total-column AOD and ground $PM_{2.5}$. Subsequently, we compared and verified the performance of each model on the test datasets. The results indicated that as the data matching radius increased, the number of samples increased, and the fitting effect of the model in the test dataset became gradually stronger; this indicated that the greater the number of samples, the better the model's fitting effect. However, the feature importance of AOD decreases with an increase in the matching radius, suggesting that the larger the matching radius of the station, the less representative the collected samples are. Therefore, considering the number of samples, the model fitting effect, and the feature importance of AOD, we choose a 50 km radius for the spatiotemporal data matching. Notably, the radius selected by others for the comparison of satellite and ground station data was also 50 km (Wu et al., 2014; Zheng et al., 2017; Mangla et al., 2020), which also conforms to the average aerosol travel velocity of 50 km/h (Ichoku et al., 2002). As the DNN is still a black-box type of operation, the importance of features cannot be realized. Thus, based on the model's fitting effect and feature interpretability, we model AOD-$PM_{2.5}$ based on the ET in this study.

4.2. Accuracy evaluation of $PM_{2.5}$ estimation by AOD of different altitudes and total-column AOD

The CALIOP data can provide AOD at different altitudes. Thus, we divided the AOD from the ground to an altitude of 2.16 km into nine segments with a resolution of 240 m, which were matched with the $PM_{2.5}$ data, and substituted into the ET. The results were then compared with the fitting result of the total column AOD and the corresponding results are presented in Table 2. Good results can be obtained when each layer's AOD and meteorological factors are fitted with the ground $PM_{2.5}$. Additionally, we concluded that the fitting effect of the bottom AOD was better than that of the total-column AOD. However, as the height increased, the importance score of AOD gradually decreased, and the importance of temperature, relative humidity, and O_3 gradually exceeded that of AOD.

Moreover, the feature importance of the total column AOD was lower than that of the temperature (T). Finally, through the interpretability of the feature importance and the model's fitting effect, we found it more reasonable to estimate $PM_{2.5}$ using the optimal layer (the layer with the highest feature importance of AOD) AOD, which has a higher feature importance score than the total-column AOD.

Fig. 3 illustrates the variation in the correlation coefficient between AOD (AOD of each layer retrieved by CALIOP) and $PM_{2.5}$ with altitude under different conditions. Fig. 3(A) shows that, the correlation coefficient between AOD and $PM_{2.5}$ changes with height, while BTH and YRD gradually decrease from the surface. However, CC, CY, and PRD exhibited a trend that initially increased and then decreased subsequently. The correlation coefficient between AOD and $PM_{2.5}$ is less than 0.2 in all regions at approximately 1 km from the ground, where the correlation coefficient was already low. When altitude reached 2 km, the correlation coefficient decreased to zero. According to the characteristics of the correlation change, the optimal layer in BTH and YRD is 0–0.24 km, while that in CC, CY, and PRD is 0.24–0.48 km.

Generally, the concentration of $PM_{2.5}$ is affected by human activities, such as industrial pollutant emissions, construction activities, and urban motor vehicles emissions (Yau et al., 2013; Li et al., 2016a; Peng et al., 2016; Wang et al., 2016). Meteorological factors, such as RH and BLH, are also conducive to formation the particulate matter (Zhang et al., 2015; Liao et al., 2017; Yin et al., 2017; You et al., 2017; Zheng et al., 2017). To explore the reasons for the regional differences in the correlation coefficient between AOD and $PM_{2.5}$ with altitude variation, we investigated the RH and the BLH, which significantly influences on the correlation between AOD and $PM_{2.5}$ (Qin et al., 2017; Gui et al., 2019). Specifically, we utilize Eqs. (8) and (9) (Guo et al., 2009; Chen et al., 2014; Wang et al., 2014; Kong et al., 2016) to eliminate the effect of humidity on AOD and obtain the corrected data and the BLH data to match the ground station data. Our calculations show that the stations' average BLH in the study area is 0.6 km. Then we discussed the influence of the BLH on the correlation coefficient between AOD and $PM_{2.5}$ in each region, considering that the value of the BLH is lower than or higher than the average BLH.

$$f_{(RH)} = \frac{1}{\left(1 - \frac{RH}{100}\right)} \tag{8}$$

$$AOD_{dry} = \frac{AOD}{f_{(RH)}} \tag{9}$$

where RH is the relative humidity, AOD is the original CALIOP data, AOD_{dry} is the AOD after humidity correction, and $f_{(RH)}$ is the humidity factor.

As presented in Fig. 3(B), the correlation between AOD and $PM_{2.5}$ after humidity correction is consistent in all regions, initially increasing and then decreasing with height, while the correlation coefficient increased to some extent. As illustrated in Fig. 2(C), when the BLH is lower than 600 m, the correlation in all regions decreased as height increases. Owing to the basin effect in the CY area, the correlation initially decreases as height increases and subsequently decreases. Fig. 3(D) shows that when the BLH is higher than 600 m, the correlation in all regions initially increases and then decreases with height; however, the correlation increases when the BLH is lower than 600 m. To summarize, the RH and BLH can explain the differences in the optimal layer in different regions to some extent. The BLH and RH were higher in CY, CC, and PRD, but the opposite was true for BTH and YRD (Liu et al., 2015; Li and Zha, 2018). Based on the AOD feature importance obtained and the correlation coefficient obtained, the data of the optimal layer in each region are employed as the input of model.

4.3. $PM_{2.5}$ model validation results

After the ET model's parameters are appropriately tuned, its performance was evaluated with the cross-validation is illustrated in Fig. 4. Overall, the performance of ET is reasonable, with R^2, RMSE, and MAE values of 0.85, 17.77 μg/m^3, and 10.66 μg/m^3, respectively. The results highlighted that BTH, CC, and YRD had a better training effect with an R^2 of 0.83–0.86, while CY and PRD had an R^2 of less than 0.80. Conversely, the RMSE values are 24.58 μg/m^3 in BTH, 11.14 μg/m^3 in PRD, and between 15.82 and 18.39 μg/m^3 in the other three regions. The validation results of independent datasets were shown in Fig. S1, and the model also performs well. These results confirm previous studies suggesting that these differences are related to the local meteorological conditions and pollution levels (Boyouk et al., 2010; Guo et al., 2017; Zhang et al., 2019; Xu et al., 2021).

Fig. 5 shows that the ET performs well during winter and autumn, achieving R^2 values of 0.86 and 0.84, with RMSE (MAE) values of

Table 2
The feature importance of each layer and the fitting R^2.

Height/km	AOD	O_3	T	RH	R^2
0.00–0.24	0.18	0.09	0.18	0.08	0.86
0.24–0.48	0.18	0.10	0.18	0.08	0.85
0.48–0.72	0.13	0.12	0.19	0.08	0.85
0.72–0.96	0.09	0.13	0.19	0.09	0.86
0.96–1.20	0.08	0.13	0.19	0.09	0.84
1.20–1.44	0.07	0.13	0.19	0.09	0.86
1.44–1.68	0.07	0.13	0.19	0.09	0.86
1.68–1.92	0.06	0.13	0.20	0.09	0.86
1.92–2.16	0.06	0.13	0.20	0.10	0.86
ALL	0.15	0.11	0.19	0.08	0.85

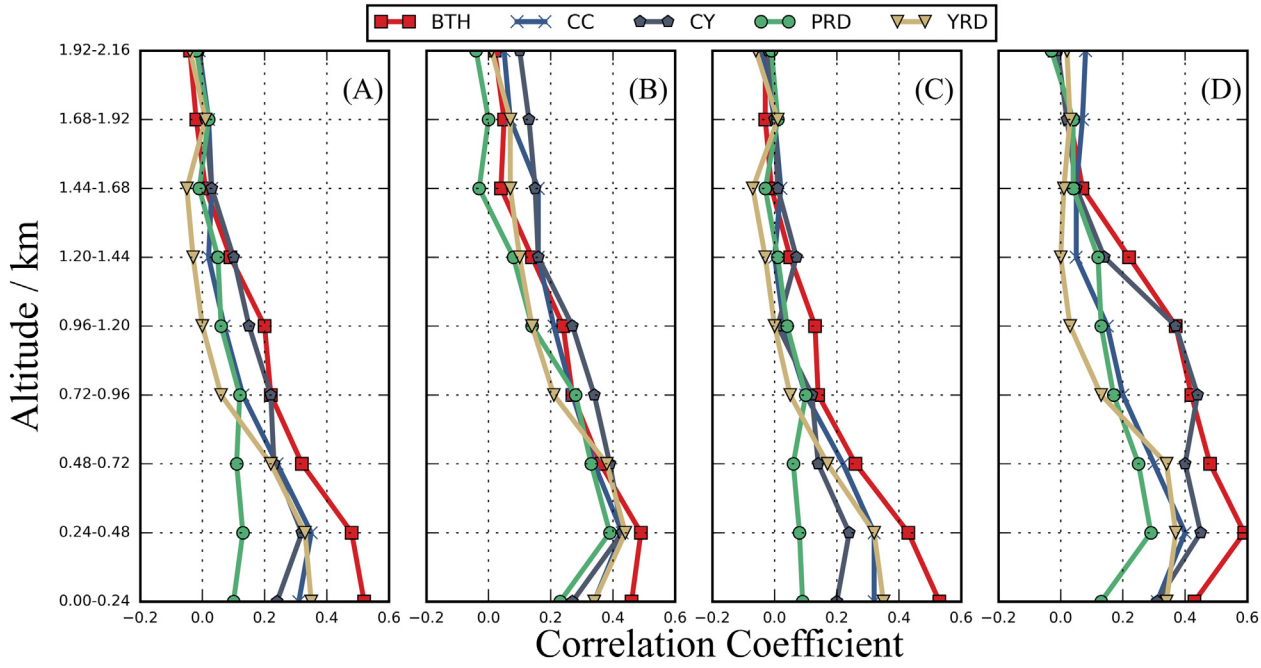


Fig. 3. Profile of the correlation coefficient between AOD at different altitudes and ground $PM_{2.5}$. Correlation coefficient profile of (A) original data, (B) AOD after humidity correction, (C) BLH lower than 600 m, and (D) BLH higher than 600 m.

24.03 and 16.13 μg/m^3 (14.41 and 10.01 μg/m^3), respectively. The model achieved poor performance during summer with an R^2 of only 0.7, while it performed well in spring with an R^2 of 0.78 and an RMSE (MAE) of 17.2 μg/m^3(10.59 μg/m^3). RMSE and MAE values are relatively low in summer (12.33 and 7.98 μg/m^3, respectively) due to lower $PM_{2.5}$ concentrations.

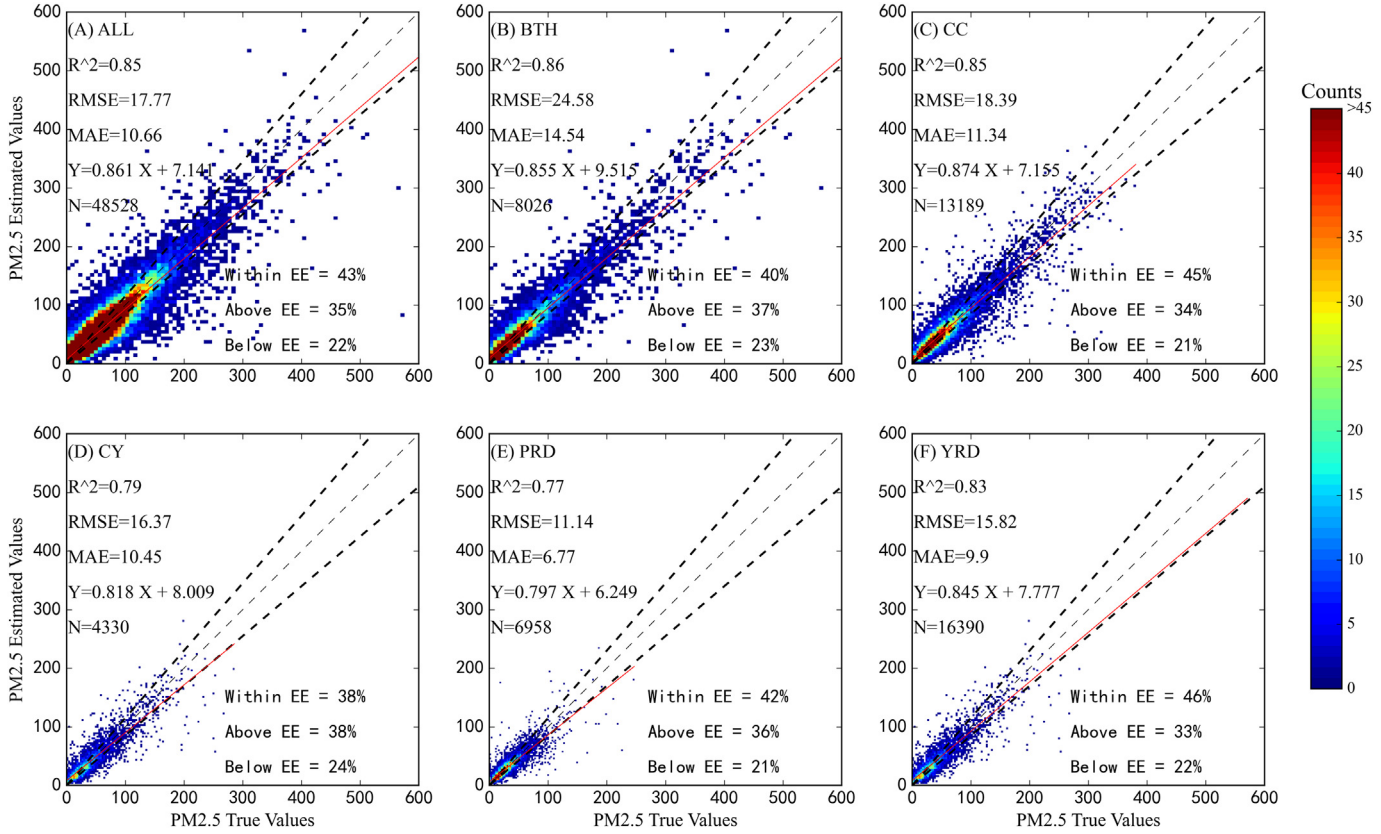


Fig. 4. Model accuracy in different regions (A) model performance in all regions and (B—F) model performances in BTH, CC, CY, PRD and YRD. R^2 represents the determination coefficient, RMSE represents root mean square error, MAE represents mean absolute error, N represents the number of samples. Equations Y and X represent the fitting relationship between the actual and estimated $PM_{2.5}$ values. The shallow black dashed line represents the 1:1 line, and the red line represents the best-fit line from the linear regression. Deep black dashed line represents the upper (1:1.5) and lower (1:0.85) expected error lines. EE presents the expected error; when the ratio of the estimated value to the true value is between 1.15 and 0.85, the error between them is called the expected error.

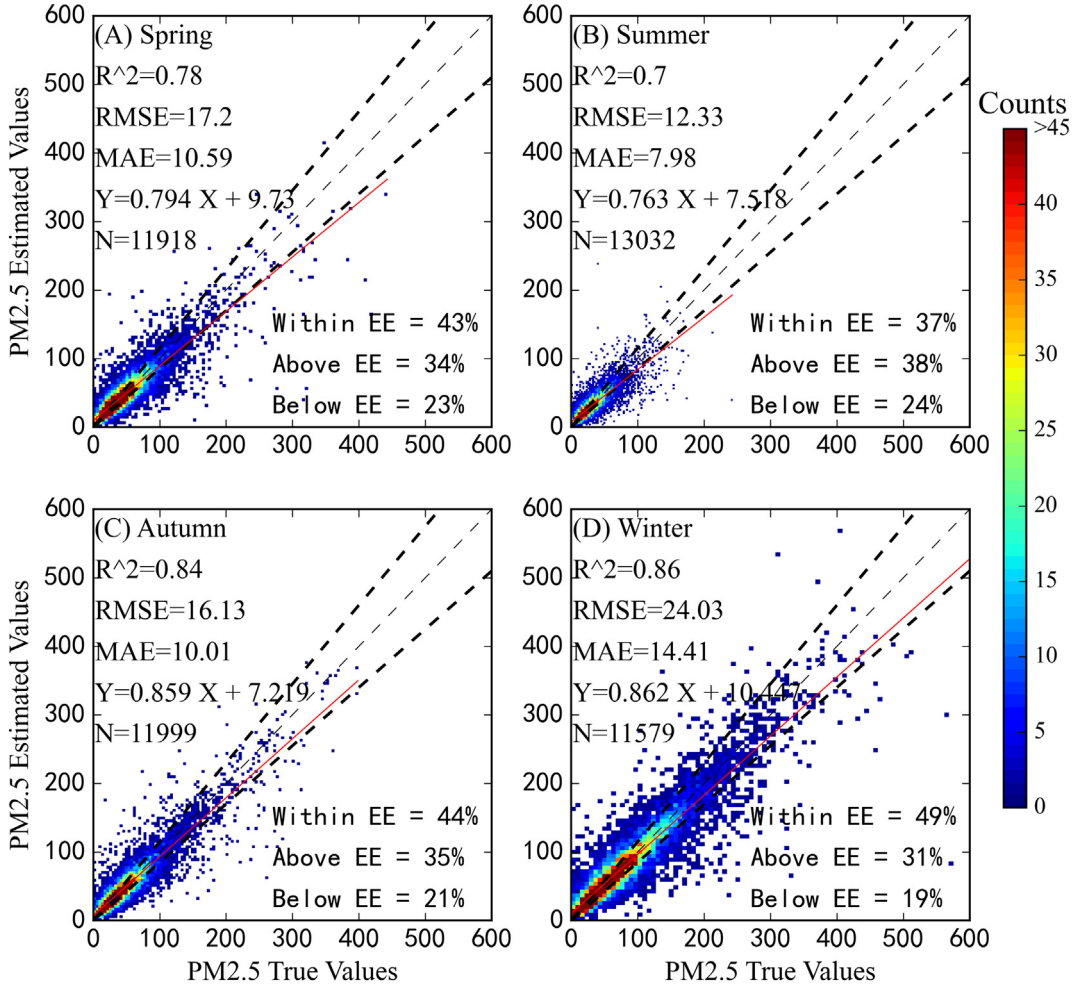


Fig. 5. Same as Fig. 4, but for seasonal $PM_{2.5}$ estimates. (A-D) Model performance in spring, summer, autumn, and winter.

The bias analysis of the ET was shown in Fig. 6, demonstrates that the average bias in each region is close to zero, and highlighting that the estimated model results for each study area are close to the true values. Specifically, when the $PM_{2.5}$ concentration is less than 60 μg/m^3, the model bias is slightly less than zero, indicating an underestimation. Accordingly, when the concentration of $PM_{2.5}$ is greater than 60 μg/m^3, the bias increases and exceeds zero, indicating that for high pollution levels, the estimated $PM_{2.5}$ value is less than the true values.

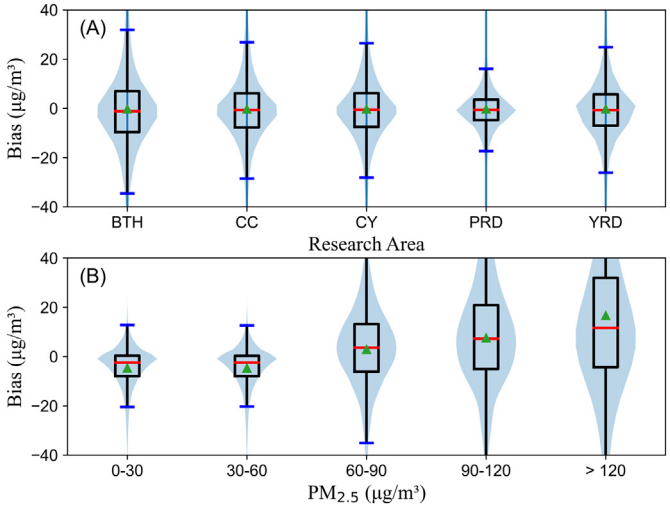


Fig. 6. Boxplots of resulting bias (y-axis) for different (A) $PM_{2.5}$ concentration ranges and (B) research areas in μg/m^3 (x-axis). The green arrow, dark blue mark, and red mark represent the average bias, the median of bias and the extremum of bias, respectively. Data density is represented by the light blue shading.

The importance scores of all features used in the ET model are shown in Fig. 7. The contributions of factors such as optimal layer AOD, ozone, air pressure, air temperature, depolarization ratio, color ratio, relative humidity, 10 m wind U and V components, longitude, and latitude were analyzed in different regions and seasons (Zhang et al., 2019). Seasonal analysis showed that the most significant contribution to the performance of ET was AOD, followed by relative humidity, air temperature, and ozone. The significant contribution of ozone is due to its strong negative correlation with $PM_{2.5}$ (Dong et al., 2021). From the perspective of regional characteristics, the AOD, air temperature, ozone, and relative humidity are the most important factors, with the zonal wind speed posing a more significant influence in the PDR. In general, the higher the feature importance of AOD, the higher the $PM_{2.5}$ concentration, and the better the model prediction result.

The performance of the model at different sites within the study area is illustrated in Fig. 8. Models perform well at most sites, as 73% of them show an $R^2 > 0.7$ and 9% show an R^2 less than 0.3. Moreover, the RMSE was less than 15 μg/m^3 in 64% of the sites and more than 25 μg/m^3 in 9%. The sites with better R^2 are mainly distributed in the urban agglomeration areas of each study area, while those with unsatisfactory R^2 values are mainly distributed in areas with complex topographic conditions, such as Western Sichuan. The RMSE metric was relatively low in most

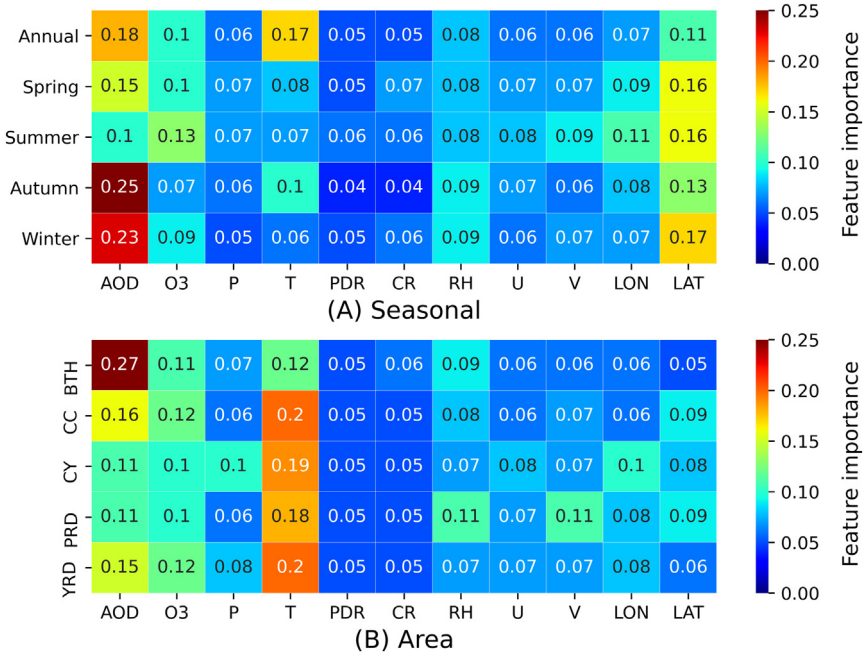


Fig. 7. The feature importance of ET in different (A) seasons and (B) areas. The color and number of each grid point on the panel represent the feature importance score in the ET model.

areas, opposing the high values in Northern China. As shown in Fig. S2, the same site distribution also appears in independent datasets validation. In general, the $PM_{2.5}$ concentration predicted by the model was consistent with the ground observation results.

To explore the performance differences of the model at various sites, the site data of $R^2 < 0.1$ (44 sites) and $R^2 > 0.9$ (213 sites) were input to the ET model for fitting to obtain the feature importance. The corresponding results are presented in Table 3, wherein the feature

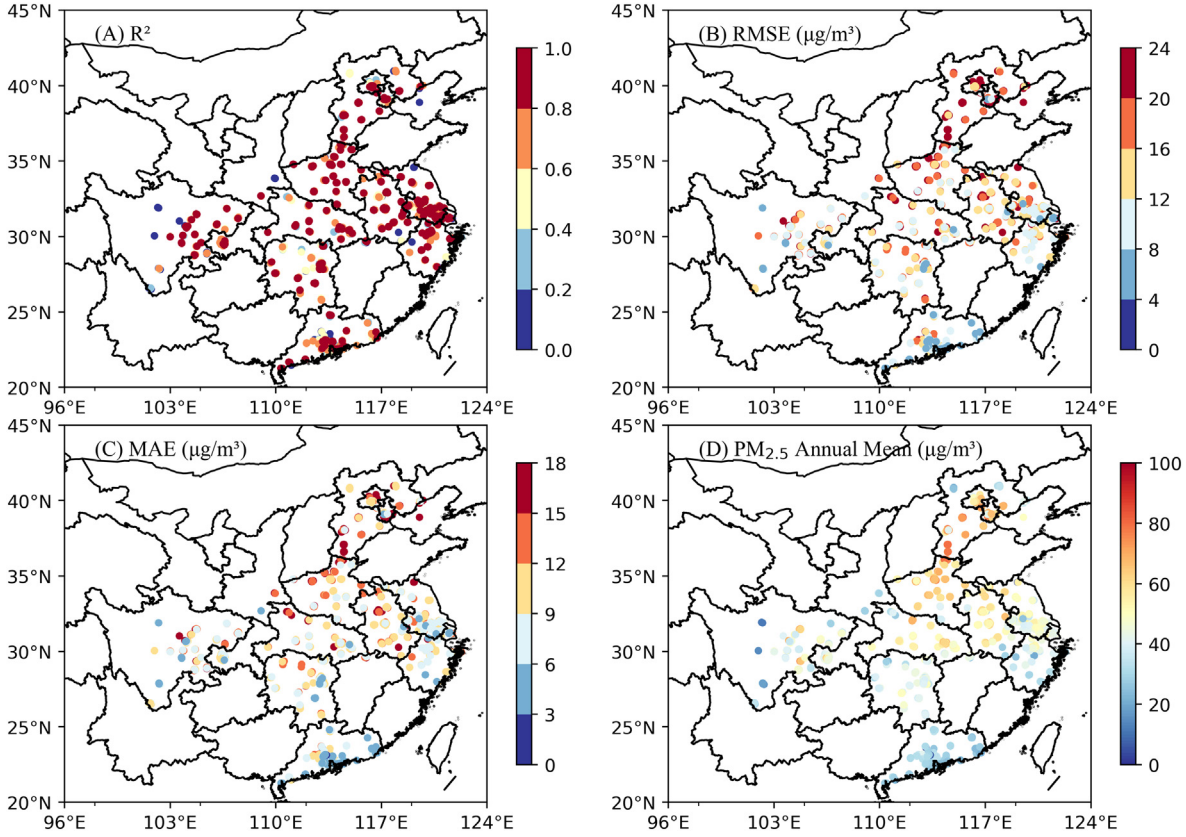


Fig. 8. Spatial distributions of the evaluation index in terms of (A) determination coefficient (R^2), (B) root mean square error (RMSE), (C) mean absolute error (MAE), and (D) mean $PM_{2.5}$ concentration at each site in China. Colored circles represent different value ranges of the presented statistical parameters.

Table 3
feature importance of site with $R^2 < 0.1$ and $R^2 > 0.9$.

	AOD	O3	P	T	PDR	CR	RH	U	V	LON	LAT
$R^2 < 0.1$	0.14	0.09	0.10	0.15	0.07	0.08	0.08	0.07	0.07	0.08	0.06
$R^2 > 0.9$	0.20	0.10	0.05	0.20	0.05	0.05	0.08	0.06	0.06	0.05	0.10

importance of AOD and temperature at $R^2 < 0.1$ are much less than those at $R^2 > 0.9$, while the feature importance of other factors, especially pressure, is higher than that at $R^2 > 0.9$. This suggests that the AOD and temperature, which affect the BLH (Chu et al., 2019), are crucial for constructing the $PM_{2.5}$ estimation model.

4.4. Estimating the concentration of $PM_{2.5}$ in the vertical direction

Fig. 9 illustrates the vertical profile of the extinction coefficients at 532 nm for the study area under consideration, as observed by CALIOP. Owing to the influence of the BLH on the diffusion of pollutants (Guo et al., 2016; Miao et al., 2020), aerosols in summer can be transported to the upper atmosphere. Thus, the extinction coefficient is generally low, with high values appearing only in a few regions. During winter, the diffusion height was lower and the near-surface aerosol accumulation was apparent. The corresponding high values in the BTH region are primarily distributed in Hebei and at a small high-value area near 39°N. The high-value areas in the CC are mainly distributed in the Northeastern part of the region with more population, while the aerosol emissions were lower in the Western and Southern parts of the region. Owing to the unique basin topography and the long-distance transport of surrounding pollutants in CY, aerosols accumulate and lead to severe pollution here (Zhao et al., 2019). There is an oblique structure in the zonal-height distribution at 23°N in the YRD, indicating that pollutants are transmitted, but the overall pollution in this area is relatively light. The extinction coefficient in the PRD is distributed in urban areas with higher values. So far, we have been using CALIOP AOD data during the daytime and at nighttime for analysis. As shown in Fig. 9, due to the influence of the sun at 532 nm during the daytime, there is some noise in the profiles of aerosol extinction coefficients at 532 nm. Therefore, from Figs. 10–13, we only use the nighttime data to analyze the vertical distribution of $PM_{2.5}$.

4.4.1. Weight calculation of each layer

By analyzing the distribution of the extinction coefficient, we concluded that the vertical structure of the AOD varies significantly in different regions and seasons. The model includes the contributions of meteorological features and AOD data. If the ground $PM_{2.5}$ estimation model is directly transplanted to the calculation of $PM_{2.5}$ at different heights, errors will occur owing to the influence of meteorological features. These errors can be expressed as the variation of $PM_{2.5}$ with height is not apparent, which is inconsistent with the vertical distribution characteristics of aerosols. Based on this consideration, the vertical distribution characteristics of AOD were used to provide the weight coefficients for each layer; namely, the weight of the optimal AOD-$PM_{2.5}$ correlation altitude layer in the study area was set to one, and the weights of the remaining layers were proportionally calculated.

4.4.2. Vertical distribution characteristics of $PM_{2.5}$

The CALIOP data was matched according to each file to generate sorted data, which were then input to the ET model, and the weight coefficient was used to correct the results. Subsequently, the latitude (longitude) -height distribution of $PM_{2.5}$ concentration were linearly interpolated before being extracted. Finally, we averaged the $PM_{2.5}$ data into latitude and longitude bins.

The corresponding latitude-altitude distribution of $PM_{2.5}$ in different seasons from 2015 to 2019, using nighttime data, are shown in Fig. 10, while the longitude-altitude distributions are shown in Fig. 11. Seasonal variations in $PM_{2.5}$ were consistent in all study areas. $PM_{2.5}$ concentration is significantly lower in summer, but pollutants can spread to higher elevations. This is mainly due to the intense weather conditions during this season, such as higher BLH and sufficient precipitation, facilitating the diffusion of pollutants. In winter, the pollutants are concentrated below 1 km, and the pollutant concentration is higher

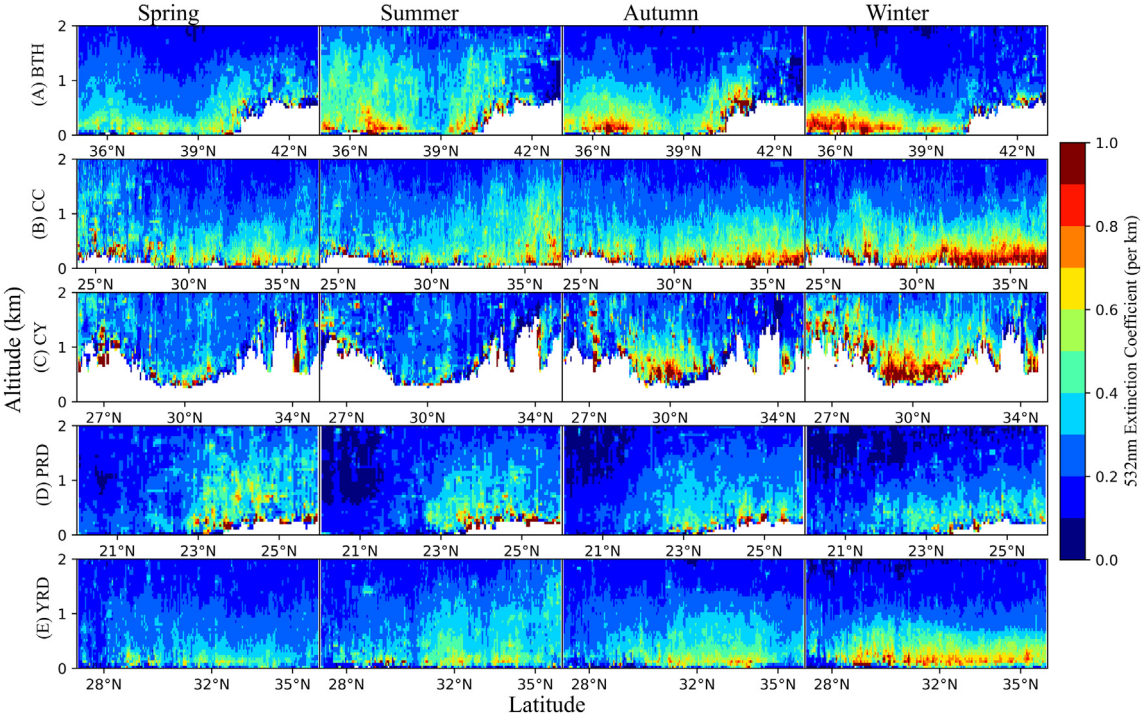


Fig. 9. Profiles of aerosol extinction coefficients at 532 nm in different seasons and areas. (A)-(E) The seasonal average of aerosol extinction coefficients at 532 nm in BTH, CC, CY, PRD, and YRD, respectively.

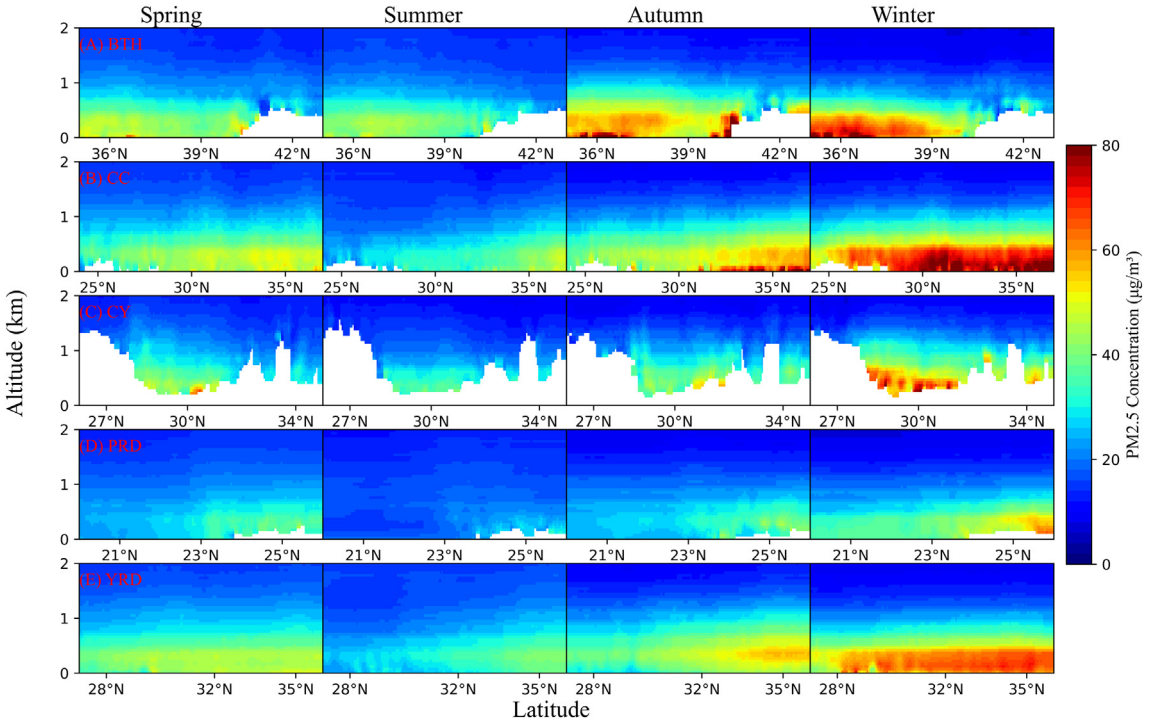


Fig. 10. Latitude and altitude profiles of the model estimated $PM_{2.5}$ in the study area. (A)-(E) The seasonal average of $PM_{2.5}$ concentration in BTH, CC, CY, PRD, and YRD, respectively.

than in the other three seasons. This is mainly caused by the meteorological conditions that are unconducive to the diffusion of pollutants due to the massive burning of fossil fuels in winter. Additionally, during winter, $PM_{2.5}$ concentrations of approximately 30 μg/m^3 were observed at distance of 1–2 km in some areas. In general, areas with high $PM_{2.5}$ concentrations are mainly concentrated near the ground and decrease rapidly with an increase in altitude.

BTH was the most polluted area among all study regions. Except for the Northern mountainous area, the annual ground $PM_{2.5}$ concentration in other areas is higher than 40 μg/m^3, and the high pollutant area (>80 μg/m^3) is mainly concentrated in the Hebei region. The $PM_{2.5}$ concentration in CC is lower than 40 μg/m^3 in summer and higher than 80 μg/m^3 in winter. In general, the pollutant concentration is mainly low in the South and West and high in the North and East. The

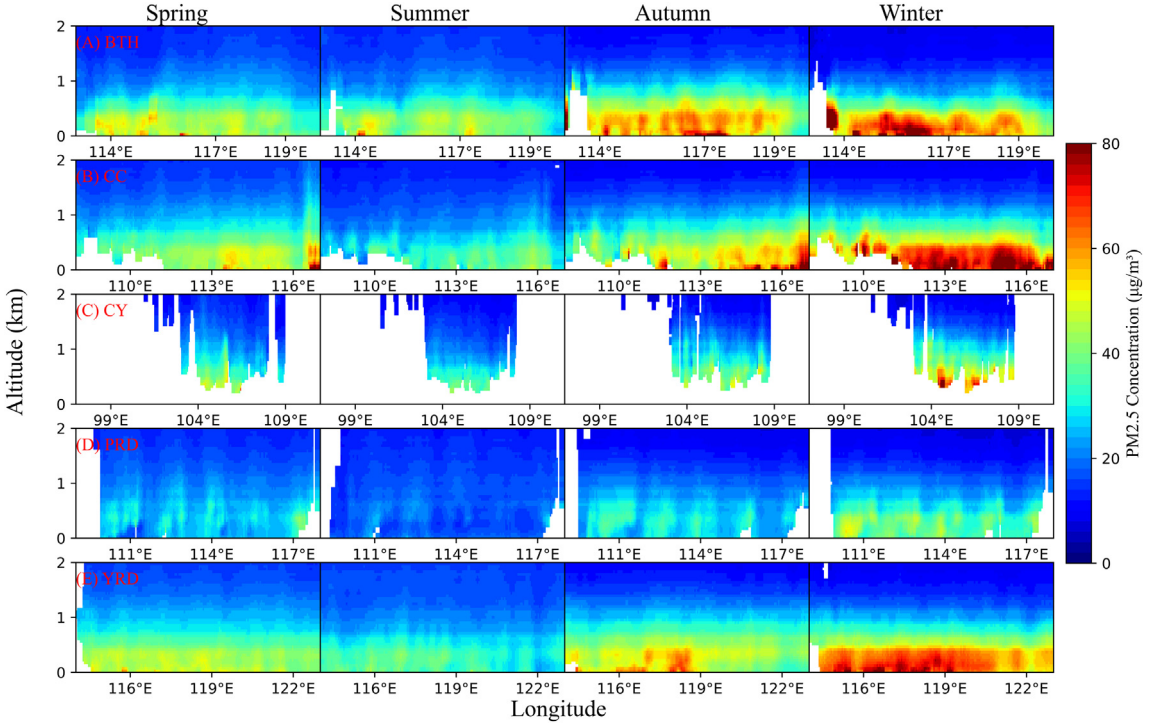


Fig. 11. Same as Fig. 10, but for longitude and altitude profiles.

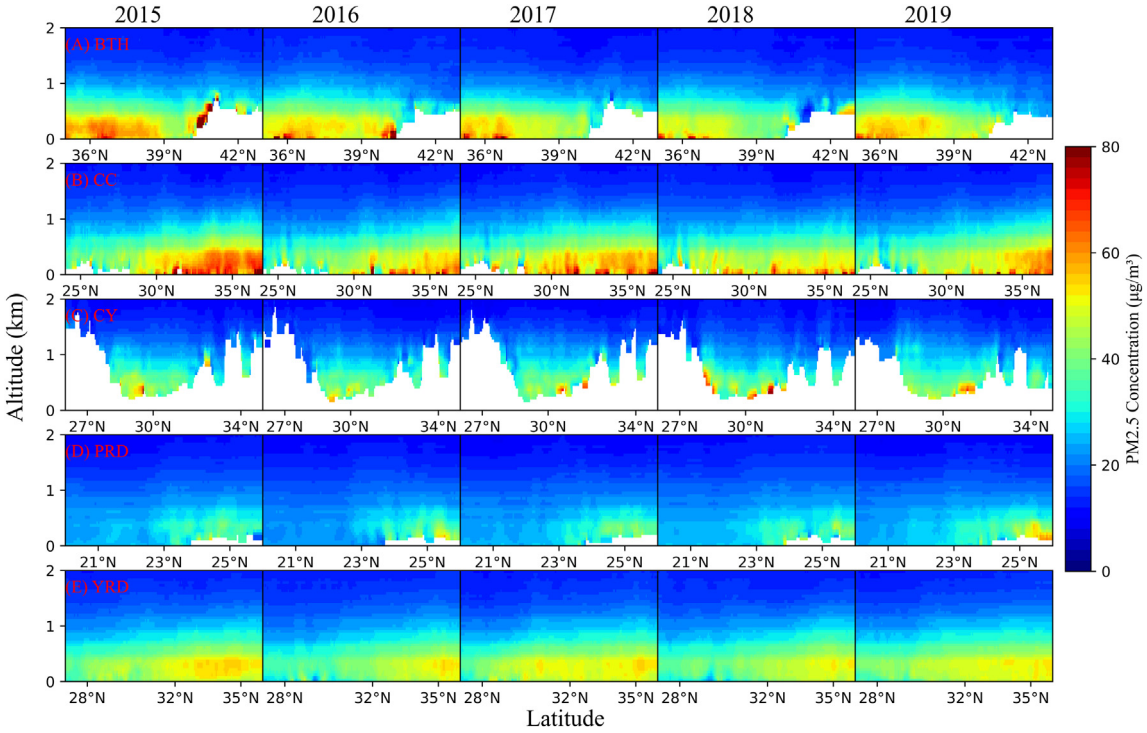


Fig. 12. Latitude and altitude profiles of the model estimated $PM_{2.5}$ concentrations in the study area from 2015 to 2019. (A-E) The annual average of $PM_{2.5}$ concentration in BTH, CC, CY, PRD, and YRD, respectively.

$PM_{2.5}$ concentration in CY and PRD was low in spring, summer, and autumn. However, in winter, the $PM_{2.5}$ concentration near the ground in the Sichuan Basin >60 μg/m^3, while that in PRD was less than 40 μg/m^3. Simultaneously, the highest level of upward diffusion of $PM_{2.5}$ in CY and PRD was lower than that in the other three regions. In the YRD area, the $PM_{2.5}$ concentration distribution during each season was relatively uniform and there was a high $PM_{2.5}$ concentration area in the middle area.

By comparing the latitude and longitude-altitude distributions, we found that the zonal vertical distribution of $PM_{2.5}$ was relatively continuous, while the longitudinal vertical distribution showed an apparent discontinuity. This is mainly because CALIPSO is a polar-

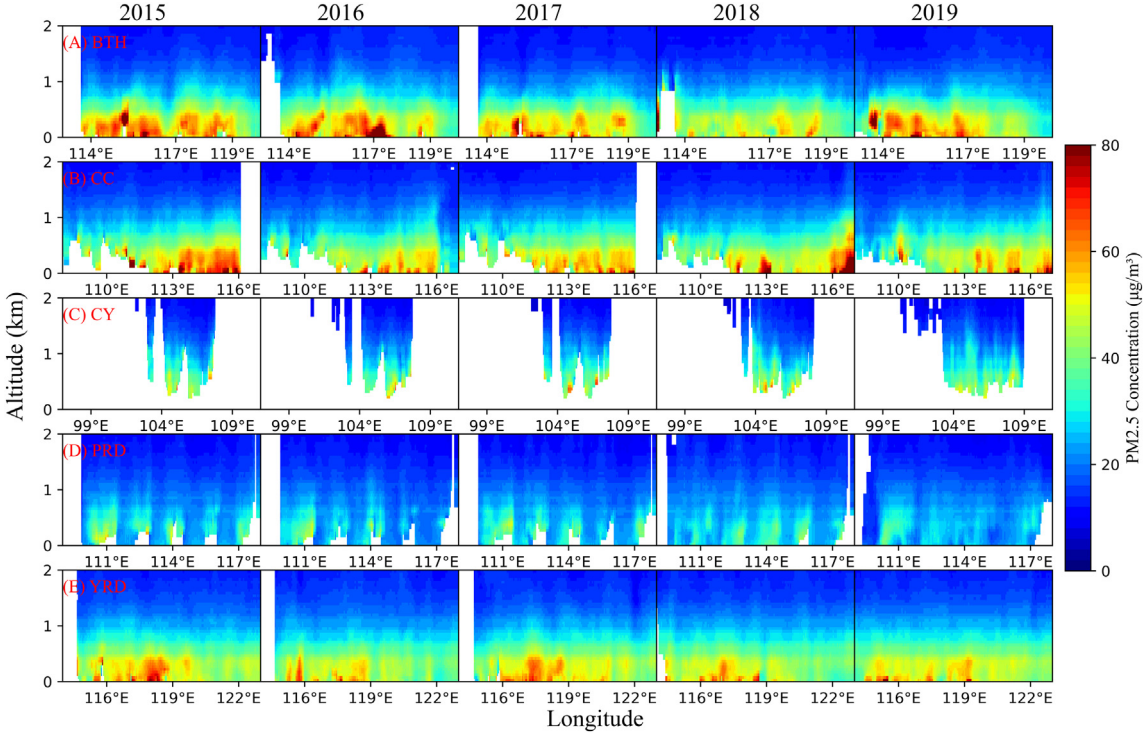


Fig. 13. Same as Fig. 12, but for longitude and altitude profiles.

orbiting satellite with a near-meridian orbit; thus, the region generating more $PM_{2.5}$ can be determined from the discontinuous distribution of the longitude-height profile. According to the distribution characteristics of $PM_{2.5}$, high-value areas in the meridian-height profile are primarily generated in urban areas, which tend to spread to the upper air. The $PM_{2.5}$ concentration in non-urban areas is significantly lower, and the highest $PM_{2.5}$ diffusion height is lower than that in urban areas.

4.4.3. Latitude (Longitude) - height profile variation trend of $PM_{2.5}$ concentration from 2015 to 2019

Figs. 12 and 13 show the average $PM_{2.5}$ concentration in the latitude (longitude) - altitude profile from 2015 to 2019. From the perspective of time change, BTH (-1.80 μg/m^3, $P < 0.001$), CC (-1.62 μg/m^3, $P < 0.001$), PRD (-0.66 μg/m^3, $P < 0.001$) exhibited a significant decline in $PM_{2.5}$ concentrations. $PM_{2.5}$ increased by 0.33 μg/m^3 in CY and decreased by 0.10 μg/m^3 in YRD, but the trends in these two regions are not statistically significant. This highlights that the $PM_{2.5}$ trend is closely related to the geographical location. From the perspective of regional changes, the high $PM_{2.5}$ value area in the study regions indicated certain changes. For example, $PM_{2.5}$ concentrations in the North and East of BTH gradually decreased, while those in the Southwest remained stable except during the year 2018. The main emission points in each region hardly changed, but the diffusion range and height of $PM_{2.5}$ decreased; overall, $PM_{2.5}$, in China showed a downward trend.

5. Conclusion

Estimation of $PM_{2.5}$ (fine particulate matter with particle size less than 2.5 μm) using AOD (aerosol optical depth) has long garnered attention (Wang and Christopher, 2003; Liu et al., 2004). However, most studies focusing on surface $PM_{2.5}$ estimation from the total-column AOD and ignoring the optical depth contributed by aerosols in the upper atmosphere (Liao et al., 2021). Most importantly, the vertical distribution of $PM_{2.5}$ has not been studied extensively. Utilizing observation instruments to obtain the vertical distribution of $PM_{2.5}$ is also limited by many factors. Therefore, obtaining $PM_{2.5}$ vertical data through satellite observation is very helpful for studying $PM_{2.5}$ vertical characteristics.

In this study, we developed a $PM_{2.5}$ estimation model based on machine learning. We calculated the $PM_{2.5}$ concentration using the optimal layer data from the CALIOP APRO (Cloud-Aerosol Lidar with Orthogonal Polarization Level-2 Aerosol Profiles V4.20) data and meteorological factors, combined with observational $PM_{2.5}$ data from the ground environmental monitoring stations. The conclusions are as follows:

(1) The optimal layer AOD contributes more to the model than the total-column AOD, and the RH and the BLH can affect the fitting relationship between AOD and $PM_{2.5}$.
(2) The ET model achieved the best fitting effect ($R^2 = 0.85$, RMSE decreased to 17.77 μg/m^3); regarding the seasonal effects, the model obtained the best fitting effect in autumn and winter (R^2 values of =0.84 and 0.86, respectively), followed by spring ($R^2 = 0.78$) and summer ($R^2 = 0.70$). The model's fitting effect R^2 exceeds 0.7 at 73% of the sites and is less than 0.3 at 9%. Through analyzing the model feature importance, we found that the greater the AOD importance is for the model, the higher the model performance.
(3) $PM_{2.5}$ concentrations in autumn and winter are higher than those in spring and summer, but $PM_{2.5}$ can be transmitted to higher altitude layers in spring and summer. The areas with high $PM_{2.5}$ concentrations are mainly near the ground and decrease with increasing height.
(4) By analyzing the vertical profile of $PM_{2.5}$ from 2015 to 2019, we concluded that $PM_{2.5}$ in China is generally reducing. Meteorological factors partially influence this change (Huang et al., 2021), but the pollution reduction measures introduced by the Chinese government are vital (Zhang et al., 2020).

Using CALIOP data, this work discussed the influence of AOD at different altitudes on the construction of the $PM_{2.5}$ model and analyzed the related reasons. The three-dimensional distribution of $PM_{2.5}$ in the seasonal and annual averages in the last five years were also analyzed. The study effectively improved the importance of AOD in the estimation of the $PM_{2.5}$ and increased the interpretability of $PM_{2.5}$, which was constructed by machine learning. Regarding the model's performance, we believed that the higher the importance of AOD and the higher the degree of pollution, the better the model effect. In our trials, we set the weight of the optimal layer to one, and the weights of the remaining layers were assigned according to the proportion of the aerosol extinction coefficient. These weight coefficients affected the model estimation results, and ultimately afforded to obtain the vertical PM2.5 concentration. The accurate representation of the vertical structure of $PM_{2.5}$, combined with the spatial distribution of $PM_{2.5}$, can help understand the transport process of such pollutants in the local and global atmosphere. Furthermore, it is of great significance for pollution prevention and control.

Given that CALIPSO's (the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation) orbital scan width is relatively narrow, a specific matching error between the station and satellite data affects the model's accuracy during the training and data testing. Additionally, we revealed that the influencing factors affect the concentration of $PM_{2.5}$ and that the fitting effect changes under different spatiotemporal characteristics. In the future, we will further analyze the influence of data errors of various factors on $PM_{2.5}$ inversion to improve the effectiveness of this study.

CRediT authorship contribution statement

Bin Chen: Conceptualization, Methodology, Writing – review & editing. **Zhihao Song:** Software, Methodology, Data curation, Writing – original draft. **Feng Pan:** Supervision. **Yue Huang:** Software, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work Supported by the National Key Research and Development Program of China (Grant number 2019YFA0606800), the National Natural Science Foundation of China (Grant 41775021), and the Fundamental Research Funds for the Central Universities (Grant lzujbky-2019-43). The CALIOP data are available from https://subset.larc.nasa.gov/calipso/. The air quality data were obtained from https://www.aqistudy.cn/historydata/. And the ERA-5 data are available from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview. The authors would like to thank China National Environmental Monitoring Center and European Centre for Medium-Range Weather Forecasts for their datasets.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2021.150338.

References

Ahmad, M.W., Reynolds, J., Rezgui, Y., 2018. Predictive modelling for solar thermal energy systems: a comparison of support vector regression, random forest, extra trees and regression trees. J. Clean. Prod. 203, 810–821. https://doi.org/10.1016/j.jclepro.2018.08.207.

Boyouk, N., Leon, J.F., Delbarre, H., Podvin, T., Deroo, C., 2010. Impact of the mixing boundary layer on the relationship between PM2.5 and aerosol optical thickness. Atmos. Environ. 44, 271–277. https://doi.org/10.1016/j.atmosenv.2009.06.053.

Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R.V., Dentener, F., van Dingenen, R., Estep, K., Amini, H., Apte, J.S., Balakrishnan, K., Barregard, L., Broday, D., Feigin, V., Ghosh, S., Hopke, P.K., Knibbs, L.D., Kokubo, Y., Liu, Y., Ma, S.F., Morawska, L., Sangrador, J.L.T., Shaddick, G., Anderson, H.R., Vos, T., Forouzanfar, M.H., Burnett, R.T., Cohen, A., 2016. Ambient air pollution exposure estimation for the global burden of disease 2013. Environ. Sci. Technol. 50, 79–88. https://doi.org/10.1021/acs.est.5b03709.

Buchard, V., da Silva, A.M., Randles, C.A., Colarco, P., Ferrare, R., Hair, J., Hostetler, C., Tackett, J., Winker, D., 2016. Evaluation of the surface PM2.5 in version 1 of the NASA MERRA aerosol reanalysis over the United States. Atmos. Environ. 125, 100–111. https://doi.org/10.1016/j.atmosenv.2015.11.004.

Calle, M.L., Urrea, V., 2011. Letter to the editor: stability of Random Forest importance measures. Brief. Bioinform. 12, 86–89. https://doi.org/10.1093/bib/bbq011.

Chelani, A.B., 2019. Estimating PM2.5 concentration from satellite derived aerosol optical depth and meteorological variables using a combination model. Atmos. Pollut. Res. 10, 847–857. https://doi.org/10.1016/j.apr.2018.12.013.

Chen, B., Huang, J., Minnis, P., Hu, Y., Yi, Y., Liu, Z., Zhang, D., Wang, X., 2010. Detection of dust aerosol by combining CALIPSO active Lidar and passive IIR measurements. Atmos. Chem. Phys. 10, 5359. https://doi.org/10.5194/acp-10-5359-2010.

Chen, J., Xin, J., An, J., Wang, Y., Liu, Z., Chao, N., Meng, Z., 2014. Observation of aerosol optical properties and particulate pollution at background station in the Pearl River Delta region. Atmos. Res. 143, 216–227. https://doi.org/10.1016/j.atmosres.2014.02.011.

Chen, D., Chang, N.J., Xiao, J.F., Zhou, Q.B., Wu, W.B., 2019a. Mapping dynamics of soil organic matter in croplands with MODIS data and machine learning algorithms. Sci. Total Environ. 669, 844–855. https://doi.org/10.1016/j.scitotenv.2019.03.151.

Chen, J.P., Yin, J.H., Zang, L., Zhang, T.X., Zhao, M.D., 2019b. Stacking machine learning model for estimating hourly PM2.5 in China based on Himawari 8 aerosol optical depth data. Sci. Total Environ. 697. https://doi.org/10.1016/j.scitotenv.2019.134021.

Chen, S., Li, D.C., Zhang, H.Y., Yu, D.K., Chen, R., Zhang, B., Tan, Y.F., Niu, Y., Duan, H.W., Mai, B.X., Chen, S.J., Yu, J.Z., Luan, T.G., Chen, L.P., Xing, X.M., Li, Q., Xiao, Y.M., Dong, G.H., Niu, Y.J., Aschner, M., Zhang, R., Zheng, Y.X., Chen, W., 2019c. The development of a cell-based model for the assessment of carcinogenic potential upon long-term PM2.5 exposure. Environ. Int. 131. https://doi.org/10.1016/j.envint.2019.104943.

Chen, B.J., You, S.X., Ye, Y., Fu, Y.Y., Ye, Z.R., Deng, J.S., Wang, K., Hong, Y., 2021. An interpretable self-adaptive deep neural network for estimating daily spatially-continuous PM2.5 concentrations across China. Sci. Total Environ. 768. https://doi.org/10.1016/j.scitotenv.2020.144724.

China, 2012. Ambient Air Quality Standards, GB 3095-2012. China Environmental Science Press, Beijing.

Chu, D.A., Kaufman, Y.J., Zibordi, G., Chern, J.D., Mao, J., Li, C.C., Holben, B.N., 2003. Global monitoring of air pollution over land from the earth observing system-Terra moderate resolution imaging spectroradiometer (MODIS). J. Geophys. Res.-Atmos. 108. https://doi.org/10.1029/2002jd003179.

Chu, Y.Q., Li, J., Li, C.C., Tan, W.S., Su, T.N., Li, J., 2019. Seasonal and diurnal variability of planetary boundary height in Beijing: intercomparison between MPL and WRF results. Atmos. Res. 227, 1–13. https://doi.org/10.1016/j.atmosres.2019.04.017.

Dong, L., Chen, B., Huang, Y., Song, Z.H., Yang, T.T., 2021. Analysis on the characteristics of air pollution in China during the COVID-19 outbreak. Atmosphere 12. https://doi.org/10.3390/atmos12020205.

Fan, W., Kai, Q., Cui, Y., Ding, L., Bilal, M., 2020. Estimation of hourly ground-level PM2.5 concentration based on Himawari-8 apparent reflectance. IEEE Trans. Geosci. Remote Sens., 1–10 https://doi.org/10.1109/TGRS.2020.2990791.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Ghahremanloo, M., Choi, Y., Sayeed, A., Salman, A.K., Pan, S., Amani, M., 2021. Estimating daily high-resolution PM2.5 concentrations over Texas: machine learning approach. Atmos. Environ. 247. https://doi.org/10.1016/j.atmosenv.2021.118209.

Ghomashi, F., Khalesifard, H.R., 2020. Investigation and characterization of atmospheric aerosols over the Urmia Lake using the satellite data and synoptic recordings. Atmos. Pollut. Res. 11, 2076–2086. https://doi.org/10.1016/j.apr.2020.08.020.

Gui, K., Che, H., Wang, Y., Wang, H., Zhang, L., Zhao, H., Zheng, Y., Sun, T., Zhang, X., 2019. Satellite-derived PM2.5 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.

Gui, K., Che, H., Zeng, Z., Wang, Y., Zhai, S., Wang, Z., Luo, M., Zhang, L., Liao, T., Zhao, H., Li, L., Zheng, Y., Zhang, X., 2020. Construction of a virtual PM2.5 observation network in China based on high-density surface meteorological observations using the extreme gradient boosting model. Environ. Int. 141, 105801. https://doi.org/10.1016/j.envint.2020.105801.

Gui, K., Che, H., Zheng, Y., Wang, Y., Zhang, L., Zhao, H., Li, L., Zhong, J., Yao, W., Zhang, X., 2021. Seasonal variability and trends in global type-segregated aerosol optical depth as revealed by MISR satellite observations. Sci. Total Environ. 787, 147543. https://doi.org/10.1016/j.scitotenv.2021.147543.

Guo, J.-P., Zhang, X.-Y., Che, H.-Z., Gong, S.-L., An, X., Cao, C.-X., Guang, J., Zhang, H., Wang, Y.-Q., Zhang, X.-C., Xue, M., Li, X.-W., 2009. Correlation between PM concentrations and aerosol optical depth in eastern China. Atmos. Environ. 43, 5876–5886. https://doi.org/10.1016/j.atmosenv.2009.08.026.

Guo, J.P., Miao, Y.C., Zhang, Y., Liu, H., Li, Z.Q., Zhang, W.C., He, J., Lou, M.Y., Yan, Y., Bian, L.G., Zhai, P., 2016. The climatology of planetary boundary layer height in China derived from radiosonde and reanalysis data. Atmos. Chem. Phys. 16, 13309–13319. https://doi.org/10.5194/acp-16-13309-2016.

Guo, J.P., Xia, F., Zhang, Y., Liu, H., Li, J., Lou, M.Y., He, J., Yan, Y., Wang, F., Min, M., Zhai, P.M., 2017. Impact of diurnal variability and meteorological factors on the PM2.5 - AOD relationship: implications for PM2.5 remote sensing. Environ. Pollut. 221, 94–104. https://doi.org/10.1016/j.envpol.2016.11.043.

Huang, J.P., Liu, J.J., Chen, B., Nasiri, S.L., 2015. Detection of anthropogenic dust using CALIPSO lidar measurements. Atmos. Chem. Phys. 15, 11653–11665. https://doi.org/10.5194/acp-15-11653-2015.

Huang, Z.W., Nee, J.B., Chiang, C.W., Zhang, S., Jin, H.C., Wang, W.C., Zhou, T., 2018. Real-time observations of dust-cloud interactions based on polarization and Raman Lidar measurements. Remote Sens. 10. https://doi.org/10.3390/rs10071017.

Huang, Y., Chen, B., Dong, L., Zhang, Z.J., 2021. Analysis of a dust weather process over East Asia in may 2019 based on satellite and ground-based lidar. Chin. J. Atmos. Sci. 45, 524–538. https://doi.org/10.3878/j.issn.1006-9895.2008.19249.

Ichoku, C., Chu, D.A., Mattoo, S., Kaufman, Y.J., Remer, L.A., Tanre, D., Slutsker, I., Holben, B.N., 2002. A spatio-temporal approach for global validation and analysis of MODIS aerosol products. Geophys. Res. Lett. 29. https://doi.org/10.1029/2001gl013206.

Kampa, M., Castanas, E., 2008. Human health effects of air pollution. Environ. Pollut. 151, 362–367. https://doi.org/10.1016/j.envpol.2007.06.012.

Kong, L., Xin, J., Zhang, W., Wang, Y., 2016. The empirical correlations between PM2.5, PM10 and AOD in the Beijing metropolitan region and the PM2.5, PM10 distributions retrieved by MODIS. Environ. Pollut. 216, 350–360. https://doi.org/10.1016/j.envpol.2016.05.085.

Kumar, A., Singh, N., Anshumali, Solanki, R., 2018. Evaluation and utilization of MODIS and CALIPSO aerosol retrievals over a complex terrain in himalaya. Remote Sens. Environ. 206, 139–155. https://doi.org/10.1016/j.rse.2017.12.019.

Li, L., Zha, Y., 2018. Mapping relative humidity, average and extreme temperature in hot summer over China. Sci. Total Environ. 615, 875–881. https://doi.org/10.1016/j.scitotenv.2017.10.022.

Li, G.D., Fang, C.L., Wang, S.J., Sun, S., 2016a. The effect of economic growth, urbanization, and industrialization on fine particulate matter (PM2.5) concentrations in China. Environ. Sci. Technol. 50, 11452–11459. https://doi.org/10.1021/acs.est.6b02562.

Li, Z., Zhang, Y., Shao, J., Li, B., Hong, J., Liu, D., Li, D., Wei, P., Li, W., Li, L., Zhang, F., Guo, J., Deng, Q., Wang, B., Cui, C., Zhang, W., Wang, Z., Lv, Y., Xu, H., Chen, X., Li, L., Qie, L., 2016b. Remote sensing of atmospheric particulate mass of dry PM2.5 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.

Li, T.W., Shen, H.F., Yuan, Q.Q., Zhang, X.C., Zhang, L.P., 2017. Estimating ground-level PM2.5 by fusing satellite and station observations: a geo-intelligent deep learning approach. Geophys. Res. Lett. 44, 11985–11993. https://doi.org/10.1002/2017gl075710.

Li, L.F., Zhang, J.H., Meng, X., Fang, Y., Ge, Y., Wang, J.F., Wang, C.Y., Wu, J., Kan, H.D., 2018. Estimation of PM2.5 concentrations at a high spatiotemporal resolution using constrained mixed-effect bagging models with MAIAC aerosol optical depth. Remote Sens. Environ. 217, 573–586. https://doi.org/10.1016/j.rse.2018.09.001.

Li, H., Yang, Y., Wang, H., Li, B., Wang, P., Li, J., Liao, H., 2021a. Constructing a spatiotemporally coherent long-term PM2.5 concentration dataset over China during 1980–2019 using a machine learning approach. Sci. Total Environ. 765, 144263. https://doi.org/10.1016/j.scitotenv.2020.144263.

Li, H.M., Yang, Y., Wang, H.L., Li, B.J., Wang, P.Y., Li, J.D., Liao, H., 2021b. Constructing a spatiotemporally coherent long-term PM2.5 concentration dataset over China during 1980–2019 using a machine learning approach. Sci. Total Environ. 765 (doi: 10.1016/j.scitotenv.2020.144263).

Liao, T.T., Wang, S., Ai, J., Gui, K., Duan, B.L., Zhao, Q., Zhang, X., Jiang, W.T., Sun, Y., 2017. Heavy pollution episodes, transport pathways and potential sources of PM2.5 during the winter of 2013 in Chengdu (China). Sci. Total Environ. 584, 1056–1065. https://doi.org/10.1016/j.scitotenv2017.01.160.

Liao, T.T., Gui, K., Li, Y.F., Wang, X.Y., Sun, Y., 2021. Seasonal distribution and vertical structure of different types of aerosols in Southwest China observed from CALIOP. Atmos. Environ. 246. https://doi.org/10.1016/j.atmosenv.2020.118145.

Liu, Y., Park, R.J., Jacob, D.J., Li, Q., Kilaru, V., Sarnat, J.A., 2004. Mapping annual mean ground-level PM2.5 concentrations using multiangle imaging spectroradiometer aerosol optical thickness over the contiguous United States. J. Geophys. Res. Atmos. 109 (doi:10.1029/2004JD005025).

Liu, J.J., Huang, J.P., Chen, B., Zhou, T., Yan, H.R., Jin, H.C., Huang, Z.W., Zhang, B.D., 2015. Comparisons of PBL heights derived from CALIPSO and ECMWF reanalysis data over China. J. Quant. Spectrosc. Radiat. Transf. 153, 102–112. https://doi.org/10.1016/j.jqsrt.2014.10.011.

Liu, Z.Y., Kar, J., Zeng, S., Tackett, J., Vaughan, M., Avery, M., Pelon, J., Getzewich, B., Lee, K.P., Magill, B., Omar, A., Lucker, P., Trepte, C., Winker, D., 2019. Discriminating between clouds and aerosols in the CALIOP version 4.1 data products. Atmos. Meas. Tech. 12, 703–734. https://doi.org/10.5194/amt-12-703-2019.

Lu, X.M., Wang, J.J., Yan, Y.T., Zhou, L.G., Ma, W.C., 2021. Estimating hourly PM2.5 concentrations using Himawari-8 AOD and a DBSCAN-modified deep learning model over the YRDUA, China. Atmos. Pollut. Res. 12, 183–192. https://doi.org/10.1016/j.apr.2020.10.020.

Ma, Z.W., Hu, X.F., Huang, L., Bi, J., Liu, Y., 2014. Estimating ground-level PM2.5 in China using satellite remote sensing. Environ. Sci. Technol. 48, 7436–7444. https://doi.org/10.1021/es5009399.

Ma, X.J., Huang, Z.W., Qi, S.Q., Huang, J.P., Zhang, S., Dong, Q.Q., Wang, X., 2020. Ten-year global particulate mass concentration derived from space-borne CALIPSO lidar observations. Sci. Total Environ. 721. https://doi.org/10.1016/j.scitotenv.2020.137699.

Mangla, R., Indu, J., Chakra, S.S., 2020. Inter-comparison of multi-satellites and Aeronet AOD over Indian Region. Atmos. Res. 240. https://doi.org/10.1016/j.atmosres.2020.104950.

Miao, Y., Che, H., Zhang, X., Liu, S., 2020. Relationship between summertime concurring PM2.5 and O3 pollution and boundary layer height differs between Beijing and

Shanghai, China. Environ. Pollut. 268, 115775. https://doi.org/10.1016/j.envpol.2020.115775.
Niu, H.W., Kang, S.C., Gao, W.N., Wang, Y.H., Paudyal, R., 2019. Vertical distribution of the Asian tropopause aerosols detected by CALIPSO. Environ. Pollut. 253, 207–220. https://doi.org/10.1016/j.envpol.2019.06.111.
Omar, A.H., Winker, D.M., Kittaka, C., Vaughan, M.A., Liu, Z.Y., Hu, Y.X., Trepte, C.R., Rogers, R.R., Ferrare, R.A., Lee, K.P., Kuehn, R.E., Hostetler, C.A., 2009. The CALIPSO automated aerosol classification and Lidar ratio selection algorithm. J. Atmos. Ocean. Technol. 26, 1994–2014. https://doi.org/10.1175/2009jtecha1231.1.
Paciorek, C.J., Liu, Y., Moreno-Macias, H., Kondragunta, S., 2008. Spatiotemporal associations between GOES aerosol optical depth retrievals and ground-level PM2.5. Environ. Sci. Technol. 42, 5800–5806. https://doi.org/10.1021/es703181j.
Paschalidou, A.K., Karakitsios, S., Kleanthous, S., Kassomenos, P.A., 2011. Forecasting hourly PM10 concentration in Cyprus through artificial neural networks and multiple regression models: implications to local environmental management. Environ. Sci. Pollut. Res. 18, 316–327. https://doi.org/10.1007/s11356-010-0375-2.
Peng, X., Shi, G.L., Zheng, J., Liu, J.Y., Shi, X.R., Xu, J., Feng, Y.C., 2016. Influence of quarry mining dust on PM2.5 in a city adjacent to a limestone quarry: seasonal characteristics and source contributions. Sci. Total Environ. 550, 940–949. https://doi.org/10.1016/j.scitotenv.2016.01.195.
Qin, K., Wu, L., Wong, M.S., Letu, H., Hu, M., Lang, H., Sheng, S., Teng, J., Xiao, X., Yuan, L., 2016. Trans-boundary aerosol transport during a winter haze episode in China revealed by ground-based Lidar and CALIPSO satellite. Atmos. Environ. 141, 20–29. https://doi.org/10.1016/j.atmosenv.2016.06.042.
Qin, K., Wang, L., Wu, L., Xu, J., Rao, L., Letu, H., Shi, T., Wang, R., 2017. 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.
Qin, K., Zou, J., Guo, J., Lu, M., Bilal, M., Zhang, K., Ma, F., Zhang, Y., 2018. Estimating PM1 concentrations from MODIS over Yangtze River Delta of China during 2014–2017. Atmos. Environ. 195, 149–158. https://doi.org/10.1016/j.atmosenv.2018.09.054.
Qin, K., Han, X., Li, D., Xu, J., Loyola, D., Xue, Y., Zhou, X., Li, D., Zhang, K., Yuan, L., 2020. Satellite-based estimation of surface NO2 concentrations over east-central China: a comparison of POMINO and OMNO2d data. Atmos. Environ. 224, 117322. https://doi.org/10.1016/j.atmosenv.2020.117322.
Rajeevan, M., Rohini, P., Kumar, K.N., Srinivasan, J., Unnikrishnan, C.K., 2013. A study of vertical cloud structure of the Indian summer monsoon using CloudSat data. Clim. Dyn. 40, 637–650. https://doi.org/10.1007/s00382-012-1374-4.
Rodriguez, J.D., Perez, A., Lozano, J.A., 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32, 569–575. https://doi.org/10.1109/Tpami.2009.187.
Schuster, G.L., Vaughan, M., MacDonnell, D., Su, W., Winker, D., Dubovik, O., Lapyonok, T., Trepte, C., 2012. Comparison of CALIPSO aerosol optical depth retrievals to AERONET measurements, and a climatology for the lidar ratio of dust. Atmos. Chem. Phys. 12, 7431–7452. https://doi.org/10.5194/acp-12-7431-2012.
Solanki, R., Singh, N., 2014. LiDAR observations of the vertical distribution of aerosols in free troposphere: comparison with CALIPSO level-2 data over the Central Himalayas. Atmos. Environ. 99, 227–238. https://doi.org/10.1016/j.atmosenv.2014.09.083.
Song, Z., Chen, B., Huang, Y., Dong, L., Yang, T., 2021. Estimation of PM2.5 concentration in China using linear hybrid machine learning model. Atmos. Meas. Tech. 14, 5333–5347. https://doi.org/10.5194/amt-14-5333-2021.
Sun, J., Gong, J.H., Zhou, J.P., 2021. Estimating hourly PM2.5 concentrations in Beijing with satellite aerosol optical depth and a random forest approach. Sci. Total Environ. 762. https://doi.org/10.1016/j.scitotenv.2020.144502.
Toth, T.D., Zhang, J.L., Reid, J.S., Vaughan, M.A., 2019. A bulk-mass-modeling-based method for retrieving particulate matter pollution using CALIOP observations. Atmos. Meas. Tech. 12, 1739–1754. https://doi.org/10.5194/amt-12-1739-2019.
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. https://doi.org/10.1029/2003GL018174.
Wang, X.P., Sun, W.B., 2019. Meteorological parameters and gaseous pollutant concentrations as predictors of daily continuous PM2.5 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.
Wang, Z., Chen, L., Tao, J., Liu, Y., Hu, X., Tao, M., 2014. An empirical method of RH correction for satellite estimation of ground-level PM concentrations. Atmos. Environ. 95, 71–81. https://doi.org/10.1016/j.atmosenv.2014.05.030.
Wang, Y.N., Jia, C.H., Tao, J., Zhang, L.M., Liang, X.X., Ma, J.M., Gao, H., Huang, T., Zhang, K., 2016. Chemical characterization and source apportionment of PM2.5 in a semi-arid and petrochemical-industrialized city, Northwest China. Sci. Total Environ. 573, 1031–1040. https://doi.org/10.1016/j.scitotenv.2016.08.179.
Wei, C.C., Lin, H.J., Lim, Y.P., Chen, C.S., Chang, C.Y., Lin, C.J., Chen, J.J.Y., Tien, P.T., Lin, C.L., Wan, L., 2019a. PM2.5 and NOx exposure promote myopia: clinical evidence and experimental proof. Environ. Pollut. 254. https://doi.org/10.1016/j.envpol.2019.113031.
Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., Cribb, M., 2019b. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. Remote Sens. Environ. 231, 111221. https://doi.org/10.1016/j.rse.2019.111221.
Wei, J., Li, Z., Pinker, R.T., Wang, J., Sun, L., Xue, W., Li, R., Cribb, M., 2021a. Himawari-8-derived diurnal variations in ground-level PM2.5 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.
Wei, J., Li, Z.Q., Lyapustin, A., Sun, L., Peng, Y.R., Xue, W.H., Su, T.N., Cribb, M., 2021b. Reconstructing 1-km-resolution high-quality PM2.5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. Remote Sens. Environ. 252 (doi:10.1016/j.rse.2020.112136).
Wilson, D.I., Piketh, S.J., Smirnov, A., Holben, B.N., Kuyper, B., 2010. Aerosol optical properties over the South Atlantic and Southern Ocean during the 140th cruise of the M/V SA agulhas. Atmos. Res. 98, 285–296. https://doi.org/10.1016/j.atmosres.2010.07.007.
Winker, D.M., Hunt, W.H., McGill, M.J., 2007. Initial performance assessment of CALIOP. Geophys. Res. Lett. 34. https://doi.org/10.1029/2007gl030135.
Winker, D.M., Tackett, J.L., Getzewich, B.J., Liu, Z., Vaughan, M.A., Rogers, R.R., 2013. The global 3-D distribution of tropospheric aerosols as characterized by CALIOP. Atmos. Chem. Phys. 13, 3345–3361. https://doi.org/10.5194/acp-13-3345-2013.
Wu, Y.H., Cordero, L., Gross, B., Moshary, F., Ahmed, S., 2014. Assessment of CALIPSO attenuated backscatter and aerosol retrievals with a combined ground-based multi-wavelength Lidar and sunphotometer measurement. Atmos. Environ. 84, 44–53. https://doi.org/10.1016/j.atmosenv.2013.11.016.
Wu, J., Yao, F., Li, W., Si, M., 2016. VIIRS-based remote sensing estimation of ground-level PM2.5 concentrations in Beijing–Tianjin–Hebei: a spatiotemporal statistical model. Remote Sens. Environ. 184, 316–328. https://doi.org/10.1016/j.rse.2016.07.015.
Wu, J.Z., Ge, D.D., Zhou, L.F., Hou, L.Y., Zhou, Y., Li, Q.Y., 2018. Effects of particulate matter on allergic respiratory diseases. Chronic Diseases and Translational Medicine. 4. https://doi.org/10.1016/j.cdtm.2018.04.001.
Xu, Q.Q., Chen, X.L., Yang, S.B., Tang, L.L., Dong, J.D., 2021. Spatiotemporal relationship between Himawari-8 hourly columnar aerosol optical depth (AOD) and ground-level PM2.5 mass concentration in mainland China. Sci. Total Environ. 765. https://doi.org/10.1016/j.scitotenv.2020.144241.
Yau, P.S., Lee, S.C., Cheng, Y., Huang, Y., Lai, S.C., Xu, X.H., 2013. Contribution of ship emissions to the fine particulate in the community near an international port in Hong Kong. Atmos. Res. 124, 61–72. https://doi.org/10.1016/j.atmosres.2012.12.009.
Yin, Z.C., Wang, H.J., Che, H.P., 2017. Understanding severe winter haze events in the North China plain in 2014: roles of climate anomalies. Atmos. Chem. Phys. 17, 1642–1652. https://doi.org/10.5194/acp-17-1641-2017.
You, T., Wu, R.G., Huang, G., Fan, G.Z., 2017. Regional meteorological patterns for heavy pollution events in Beijing. J. Meteorol. Res. 31, 597–611. https://doi.org/10.1007/s13351-017-6143-1.
Zeydan, O., Wang, Y.H., 2019. Using MODIS derived aerosol optical depth to estimate ground-level PM(2.5)concentrations over Turkey. Atmos. Pollut. Res. 10, 1565–1576. https://doi.org/10.1016/j.apr.2019.05.005.
Zhang, Y., Li, Z., 2015. Remote sensing of atmospheric fine particulate matter (PM2.5) mass concentration near the ground from satellite observation. Remote Sens. Environ. 160, 252–262. https://doi.org/10.1016/j.rse.2015.02.005.
Zhang, C., Ni, Z.W., Ni, L.P., 2015. Multifractal detrended cross-correlation analysis between PM2.5 and meteorological factors. Physica A 438, 114–123. https://doi.org/10.1016/j.physa.2015.06.039.
Zhang, T.X., Zang, L., Wan, Y.C., Wang, W., Zhang, Y., 2019. Ground-level PM2.5 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.
Zhang, X., Xu, X., Ding, Y., Liu, Y., Zhang, H., Wang, Y., Zhong, J., 2020. The impact of meteorological changes from 2013 to 2017 on PM2.5 mass reduction in key regions in China. Sci. China Earth Sci. 50, 483–500. https://doi.org/10.1360/N072018-00303.
Zhang, Y., Li, Z., Bai, K., Wei, Y., Xie, Y., Zhang, Y., Ou, Y., Cohen, J., Zhang, Y., Peng, Z., Zhang, X., Chen, C., Hong, J., Xu, H., Guang, J., Lv, Y., Li, K., Li, D., 2021. Satellite remote sensing of atmospheric particulate matter mass concentration: advances, challenges, and perspectives. Fundam. Res. 1, 240–258. https://doi.org/10.1016/j.fmre.2021.04.007.
Zhang, Q., Shi, R., Singh, V.P., Xu, C.-Y., Yu, H., Fan, K., Wu, Z., 2022. Droughts across China: drought factors, prediction and impacts. Sci. Total Environ. 803, 150018. https://doi.org/10.1016/j.scitotenv.2021.150018.
Zhao, S.P., Yu, Y., Qin, D.H., Yin, D.Y., Dong, L.X., He, J.J., 2019. Analyses of regional pollution and transportation of PM2.5 and ozone in the city clusters of Sichuan Basin, China. Atmos. Pollut. Res. 10, 374–385. https://doi.org/10.1016/j.apr.2018.08.014.
Zhao, C., Wang, Q., Ban, J., Liu, Z.R., Zhang, Y.Y., Ma, R.M., Li, S.S., Li, T.T., 2020. Estimating the daily PM2.5 concentration in the Beijing-Tianjin-Hebei region using a random forest model with a 0.01° ×0.01° spatial resolution. Environ. Int. 134. https://doi.org/10.1016/j.envint.2019.105297.
Zheng, C.W., Zhao, C.F., Zhu, Y.N., Wang, Y., Shi, X.Q., Wu, X.L., Chen, T.M., Wu, F., Qiu, Y.M., 2017. Analysis of influential factors for the relationship between PM2.5 and AOD in Beijing. Atmos. Chem. Phys. 17, 13473–13489. https://doi.org/10.5194/acp-17-13473-2017.
Zhong, J., Zhang, X., Gui, K., Wang, Y., Che, H., Shen, X., Zhang, L., Zhang, Y., Sun, J., Zhang, W., 2021. Robust prediction of hourly PM2.5 from meteorological data using LightGBM. Natl. Sci. Rev. https://doi.org/10.1093/nsr/nwaa307.
Zhu, Z.M., Zhang, M., Huang, Y.S., Zhu, B., Han, G., Zhang, T.H., Liu, B.M., 2018. Characteristics of the planetary boundary layer above Wuhan, China based on CALIPSO. Atmos. Res. 214, 204–212. https://doi.org/10.1016/j.atmosres.2018.07.024.