An optimized semi-empirical physical approach for satellite-based $PM_{2.5}$ retrieval: embedding machine learning to simulate complex physical parameters

Caiyi Jin [a], Qiangqiang Yuan [a, c, d, *], Tongwen Li [b, *], Yuan Wang [a], Liangpei Zhang [c, e]

[a] School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China.
[b] School of Geospatial Engineering and Science, Sun Yat-Sen University, Zhuhai 519082, China
[c] The Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China.
[d] The Key Laboratory of Geospace Environment and Geodesy, Ministry of Education, Wuhan University, Wuhan 430079, China.
[e] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China.

* Corresponding author.
E-mail address: yqiang86@gmail.com, litw8@mail.sysu.edu.cn

ABSTRACT

Satellite remote sensing of $PM_{2.5}$ mass concentration has become one of the most popular atmospheric research aspects, resulting in the development of different models. Among them, the semi-empirical physical approach constructs the transformation relationship between the aerosol optical depth (AOD) and $PM_{2.5}$ based on the optical properties of particles, which has strong physical significance. Also, it performs the $PM_{2.5}$ retrieval independently of the ground stations. However, due to the complex physical relationship, the physical parameters in the semi-empirical approach are difficult to calculate accurately, resulting in relatively limited accuracy. To achieve the optimization effect, this study proposes a method of embedding machine learning into a semi-physical empirical model (RF-PMRS). Specifically, based on the theory of the physical $PM_{2.5}$ remote sensing approach (PMRS), the complex parameter (VE_f, a columnar volume-to-extinction ratio of fine particles) is simulated by the random forest model (RF). Also, a fine mode fraction product with higher quality is applied to make up for the insufficient coverage of satellite products. Experiments in North China show that the surface $PM_{2.5}$ concentration derived by RF-PMRS has an average annual value of 57.92 μg/m³ versus the ground value of 60.23 μg/m³. Compared with the original

method, RMSE decreases by 39.95 μg/m³, and the relative deviation reduces by 44.87%. Moreover, validation at two AERONET sites presents a time series change closer to the true values, with an R of about 0.80. This study is also a preliminary attempt to combine model-driven and data-driven models, laying a foundation for further atmospheric research on optimization methods.

Keywords: $PM_{2.5}$; Physical approach; Machine learning; Volume-to-extinction ratio; Fine mode fraction

1. Introduction

Epidemiological studies have indicated that $PM_{2.5}$ (fine particulate matter with an aerodynamic equivalent diameter no greater than 2.5 μm) can adversely affect human health, such as increasing the risk of diabetes and respiratory diseases (Bowe et al., 2018; Pope III et al., 2002; Xu et al., 2013), and accurate surface $PM_{2.5}$ concentration is the basis of air pollution-health related research. Satellite remote sensing has the advantages of high resolution and global coverage (Ma et al., 2014; Wu et al., 2020; He et al., 2022), including variables strongly associated with $PM_{2.5}$ such as aerosol optical depth (AOD). Therefore, it has become a mainstream method for fine particle estimation (Zhang et al., 2021).

There are mainly three satellite-based ways of retrieving $PM_{2.5.}$ 1) Chemical transport models-based method. It calculates a scaling factor η between AOD and $PM_{2.5}$ simulated by atmospheric chemical transport models (CTM) (Lyu et al., 2022; Xiao et al., 2022) and then transfers the proportional relationship to satellite AOD data when calculating surface $PM_{2.5}$ concentration (Geng et al., 2015; Van Donkelaar et al., 2006). However, the assumption of a constant factor between simulated and observed values has large spatiotemporal limitations. 2) Univariate/Multivariate regression. This kind of data-driven method establishes a statistical model between AOD, auxiliary variables, and ground $PM_{2.5}$ observations. Machine learning is a common tool for such regression methods due to its powerful nonlinear fitting ability between multiple variables (Irrgang et al., 2021). But the regression algorithms in machine learning are affected by the distribution and density of ground stations (Gupta and Christopher, 2009; Li et al., 2017). 3) Semi-empirical physical approach. Taking the physical theory as the basis, surface $PM_{2.5}$ is derived through an empirical formula constructed from AOD and some PM-related key parameters, including an important empirical parameter related to the optical properties (S). The process steps are explicit and independent of ground station observations. Meanwhile, this approach has stronger physical interpretability than the previous two methods with a large space for optimization.

Due to the complexity of the physical parameters, many studies have optimized the semi-empirical physical approach. Based on 355nm-band radar observations, Raut and Chazette (2009) introduced a specific extinction cross-section to simplify the expression of S, and $PM_{2.5}$ concentration was estimated. Kokhanovsky et al. (2009) constructed a particle-effective radius model, which can obtain the particle concentrations throughout the atmospheric column. Furthermore, Zhang and Li (2015) proposed the physical $PM_{2.5}$ remote sensing method (PMRS). It replaced S by defining a volume-to-extinction ratio of fine particles (VE_f) and used a quadratic polynomial of fine mode fraction (FMF) to simulate VE_f, showing certain advantages (Li et al., 2016; Zhang et al., 2020).

However, the above semi-physical empirical models have some shortcomings. Firstly, the satellite data used in the models are blocked by clouds and fog in some areas, thus high-coverage and high-precision products need to be excavated and applied; secondly, there are still large uncertainties in estimating physical parameters (such as a simple polynomial fit to S in the PMRS method) and their expressions need to be improved. To date, machine learning (ML) has developed rapidly (He et al., 2021). It can detect complex nonlinear relationships of multiple data and model their interaction (Yuan et al., 2020; Lee et al.,2022). This provides an idea for improving the accuracy of physical parameter acquisition, so as to estimate high-precision $PM_{2.5}$ through semi-physical empirical models.

According to this idea, our study proposes an optimized semi-empirical physical model (RF-PMRS) based on the PMRS theory, which attempts to explore the possibility of combining physical models and ML. To be specific, we creatively embed ML (the random forest model) into the PMRS method to simulate the physical parameter (i.e.,

VE_f) derived from FMF and related variables, thus optimizing the previous polynomial expression. Besides, to further improve the $PM_{2.5}$ retrieval accuracy, the physical-deep learning FMF (Phy-DL FMF) dataset generated by a hybrid retrieval algorithm of ML and physical mechanisms is introduced. Ultimately, we comprehensively validate the performance of the $PM_{2.5}$ obtained by our optimized approach.

The remained part of our article is as follows. Section 2 describes the experimental datasets. Section 3 illustrates the specific derivation process of the proposed method. Section 4 analyzes the evaluation results. Some supporting experiments are discussed in Section 5. And the final part provides the conclusion.

2. Data

2.1. AERONET data

The Aerosol Robotic Network (AERONET) is a federation of ground-based sun-sky radiometer networks, providing worldwide remote sensing aerosol data for more than 25 years (Holben et al., 1998). Until now, the Version 3 dataset has been released (Giles et al., 2017). Due to its high quality, the data from AERONET have been regarded as theoretical true values to evaluate satellite-based products in related studies (Chen et al., 2020; Gao et al., 2016; Wang et al., 2019). AOD, FMF, and Volume Size Distribution products with Level 2.0 (quality-assured) are applied to calculate the true values of the physical parameters, and then to implement our modeling purpose (not involved in $PM_{2.5}$ calculations). A total of 9 AERONET sites corresponding to four typical aerosol types participate in the training. Table 1 shows the specific information.

Table 1. Data information of 9 AERONET sites classified by aerosol types. Location indicates the latitude and longitude, where '-' means the south latitude and west longitude. Two sites in bold fonts participate in the $PM_{2.5}$ validation experiment.

Aerosol Type	Site	Location (LAT, LON)	Training period	Isolated-validation period
Urban–industrial	**Beijing**	**39.98°, 116.38°**	2001-2017	2018-2019
	Beijing-CAMS	**39.93°, 116.32°**	2012-2017	2018-2019
	XiangHe	39.75°, 116.96°	2004-2017	/
	Ascension Island	-7.98°, -14.41°	2010-2017	2018-2019

	Capo Verde	16.73°, -22.94°	2010-2017	2018
Biomass burning	CUIABA MIRANDA	-15.73°, -56.07°	2010-2017	2018-2019
Desert dust	GSFC	38.99°, -76.84°	2010-2017	2018-2019
	Mexico City	19.33°, -99.18°	2010-2017	/
Oceanic	Solar Village	24.91°, 46.40°	2010-2013	/

2.2. MODIS AOD

MCD19A2, the Moderate-resolution Imaging Spectroradiometer (MODIS) C6 Level-2 gridded (L2G) land AOD product (Lyapustin and Wang, 2015), is selected in this study. It is derived from the Multi-Angle Implementation of the Atmospheric Correction (MAIAC) algorithm, which can improve the accuracy in cloud detection and aerosol retrieval (Lyapustin et al., 2011). Besides, this new advanced algorithm jointly combines MODIS Terra and Aqua into a single sensor (Lyapustin et al., 2014). The product is produced daily with a 1km resolution, including aerosol parameters such as 470nm/550nm AOD, quality assurance (QA), and uncertainty factors.

The processing of MCD19A2 data (HDF format) is mainly divided into five steps: AOD/QA band extraction, best quality AOD selection, Terra/Aqua data synthesis, missing information reconstruction, and mosaic. Finally, the daily AOD distribution in GeoTiff format is obtained.

2.3. Phy-DL FMF dataset

The original global land FMF products have poor data integrity and low accuracy. To enhance their reliability, Yan et al. (2022) have released a satellite-based dataset called Phy-DL FMF, which integrates physical and deep learning methods. Specifically, it selects the FMF data obtained by a physical method (i.e., Look-Up-Table-based Spectral Deconvolution Algorithm, LUT-SDA) as the optimization target (Yan et al., 2017). Then it combines the Phy-based FMF into a deep-learning model along with multiple auxiliary data such as satellite observations for the final Phy-DL results. Note that the process is trained with AERONET data as the ground truth. The product has a spatial resolution of 1° and covers from 2001 to 2020 (daily scale). In the comparison experiment against the ground FMF, Phy-DL FMF shows a higher accuracy (R = 0.78, RMSE = 0.100) than MODIS FMF (R = 0.37, RMSE = 0.282) (Yan et al., 2022).

2.4. Meteorological data

The meteorological data are obtained from the ERA5 dataset, including the values of planetary boundary layer height (PBLH) and relative humidity (RH). As the fifth-generation reanalysis product released by the European Center for Medium-Range Weather Forecasts (ECMWF), ERA5 provides atmospheric data at 0.25° every hour based on the data assimilation principle (Hersbach et al., 2018). It should be noted that RH is not archived directly in ERA5, thus should be calculated by 2m temperature T and dew point temperature T_d (referred to ERA-Interim: documentation).

$$RH = 100 \times \frac{e_s(T_d)}{e_s(T)} \tag{1}$$

Here, $e_s(t)$ represents the saturation vapor pressure related to a Celsius temperature t (Simmons et al., 1999).

$$e_s(t) = 6.112 \times \exp\left(\frac{17.67 \times t}{t + 243.5}\right) \tag{2}$$

2.5. Ground $PM_{2.5}$ measurements

The North China Region (NC) is chosen as the main experimental validation area for the final $PM_{2.5}$ calculations. The near-surface hourly $PM_{2.5}$ values are obtained from the China National Environmental Monitoring Center (CNEMC). Nowadays, over 1600 ground-based monitors are working continuously and a total of 232 stations (in 2017) participate in this work. Fig. 1 displays the site distributions of the NC region.

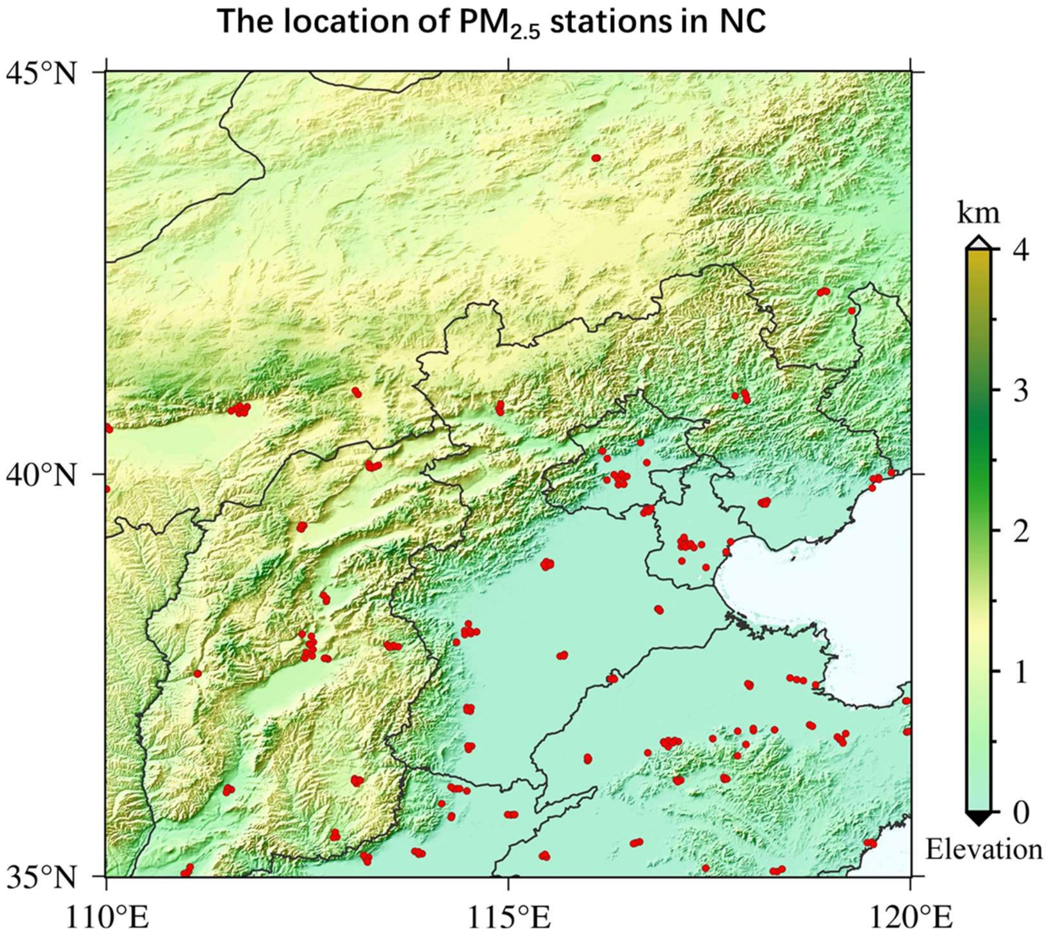

Fig. 1. The location of $PM_{2.5}$ ground monitoring stations in the NC region (35°-45°N, 110°-120°E). The red points represent the $PM_{2.5}$ stations.

3. Methods

Based on the basic physical properties of atmospheric aerosols, the semi-physical empirical approach starts from the integration of PM mass concentration and AOD. Then it combines several key factors related to $PM_{2.5}$, to derive the in situ $PM_{2.5}$ concentration through multiple remote sensing variables (Koelemeijer et al., 2006). The overall empirical relationship can be represented as:

$$PM_{2.5} = AOD \frac{\rho}{H \cdot f(RH)} S \tag{3}$$

where ρ denotes the particle density and H denotes the atmospheric boundary layer height. $f(RH)$ represents the hygroscopic growth factor related to relative humidity (RH). S is an optical characteristic parameter that should be simulated.

3.1. PMRS method

3.1.1. The expression of VEf

To illustrate S more precisely, PMRS defines the columnar volume-to-extinction ratio of fine particles (i.e., VE_f), which can be regarded as the basis of our optimization method. So equation (3) is transformed into:

$$PM_{2.5} = AOD\frac{\rho}{H \cdot f(RH)}VE_f \tag{4}$$

Related to particle size, aerosol extinction, and other properties, VE_f can be expressed as:

$$VE_f = \frac{V_{f,column}}{AOD_f} \tag{5}$$

$$AOD_f = AOD \cdot FMF \tag{6}$$

Here, AOD_f is the fine particle AOD and FMF is the fine mode fraction. $V_{f,column}$ can be expressed by the vertical integral of particle volume size distributions (PVSD) within a certain aerodynamic diameter range:

$$V_{f,column} = \int_0^{D_{p,c}} V(D_p)dD_p \tag{7}$$

$D_{p,c}$ represents the cutting diameter, and the empirical value of 2.0 μm is chosen based on previous literature (Hand and Kreidenweis, 2002; Hänel and Thudium, 1977). And $V(D_p)$ represents the PVSD corresponding to the geometric equivalent diameter (D_p).

3.1.2. Specific process and limitations

The PMRS method is developed from equation (4). Based on satellite AOD, the near-surface $PM_{2.5}$ can be obtained through multi-step transformation. Fig. 2(a) shows its specific process. Each arrow refers to a step, respectively: size cutting (output: AOD_f), volume visualization (output: $V_{f,column}$), bottom isolation (output: V_f, fine particle volume near the ground), particle drying (output: $V_{f,dry}$, dry V_f) and $PM_{2.5}$ weighting. The overall expression is as follows:

$$PM_{2.5} = AOD\frac{FMF \cdot VE_f \cdot \rho_{f,dry}}{PBLH \cdot f_0(RH)} \tag{8}$$

$$f_0(RH) = \left(1 - \frac{RH}{100}\right)^{-1} \tag{9}$$

where FMF denotes the fine mode fraction, $\rho_{f,dry}$ denotes the dry mass density of $PM_{2.5}$, and $PBLH$ represents the planet boundary layer height. $f_0(RH)$ represents the approximation of $f(RH)$ in equation (4), as expressed in equation (9). Considering the aerosol types in different regions, PMRS fits $\mathrm{VE_f}$ to a quadratic polynomial relation of FMF (Zhang and Li, 2015):

$$VE_f = 0.2887FMF^2 - 0.4663FMF + 0.356 \quad (0.1 \le FMF \le 1.0) \tag{10}$$

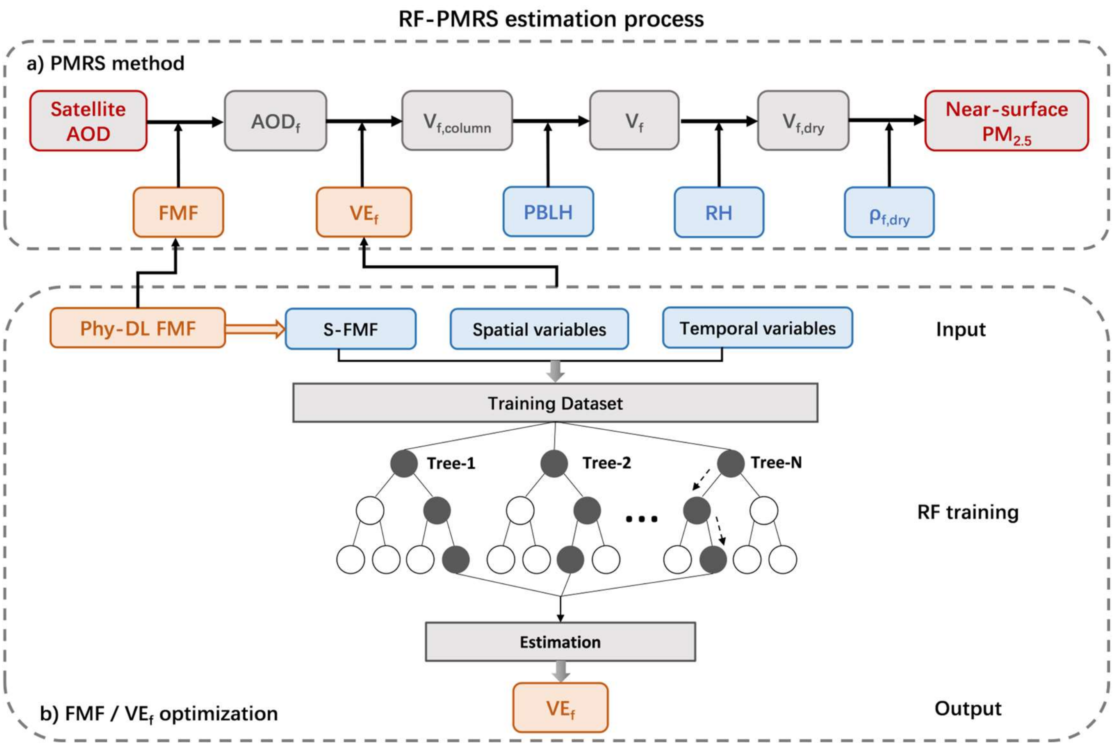

Fig. 2. Surface $PM_{2.5}$ estimation flow of RF-PMRS. a) The five steps of the PMRS method. Gray boxes are the intermediate outputs, blue boxes are the input data, and orange ones denote the variables to be optimized. b) The specific optimization of RF-PMRS: FMF dataset replacement and $\mathrm{VE_f}$ simulation by RF model.

PMRS has strong physical significance, the calculation steps are well-defined and site-independent. Zhang and Li (2015) tested the performance of PMRS on 15 stations, and the validation results had an uncertainty of 34%. Compared with the ground value of Jinhua city in China, a 31.3% relative error was generated in Li et al. (2016). Besides, Zhang et al. (2020) applied it to the $PM_{2.5}$ change analysis and prediction experiments in China over 20 years. However, there may be a more complex nonlinear relationship between VE_f with FMF, not just a simple quadratic formula. Since VE_f is related to the aerosol type, adding other spatiotemporal variables may optimize the fitting process. Additionally, high-quality FMF data is the basic guarantee for the estimated $PM_{2.5}$ quality. In a word, to further improve the physical method, a better nonlinear model between VE_f and related variables from reliable datasets needs to be explored.

3.2. Optimization method: RF-PMRS

Therefore, to overcome the above disadvantages, an optimized method called RF-PMRS is proposed. Fig. 2(b) shows the process of our method, while optimizations for FMF and VE_f are described separately below.

1) FMF dataset selection

We introduce the Phy-DL FMF dataset into the PMRS method to improve the accuracy of size-cutting results. In terms of performance, it exhibits higher accuracy and wider space-time coverage than satellite products (Yan, 2021). See the data section for details.

2) VE_f simulation based on ML

The main idea is to establish an ML model between the VE_f truth obtained from multiple AERONET sites and related variables, thus improving the subsequent VE_f-simulation accuracy (Fig. 3).

Step 1 VE_f calculation

The VE_f true values are calculated concerning equations (5)-(7). Due to the spatiotemporal variability of different aerosol types, we calculate the VE_f values at 9 AERONET stations around the world (Table 1) to train a universal model. The first step in Fig. 3 shows their distribution characteristics. Among them, Beijing and Beijing-CAMS sites are highlighted since they participate in the subsequent point validation experiment.

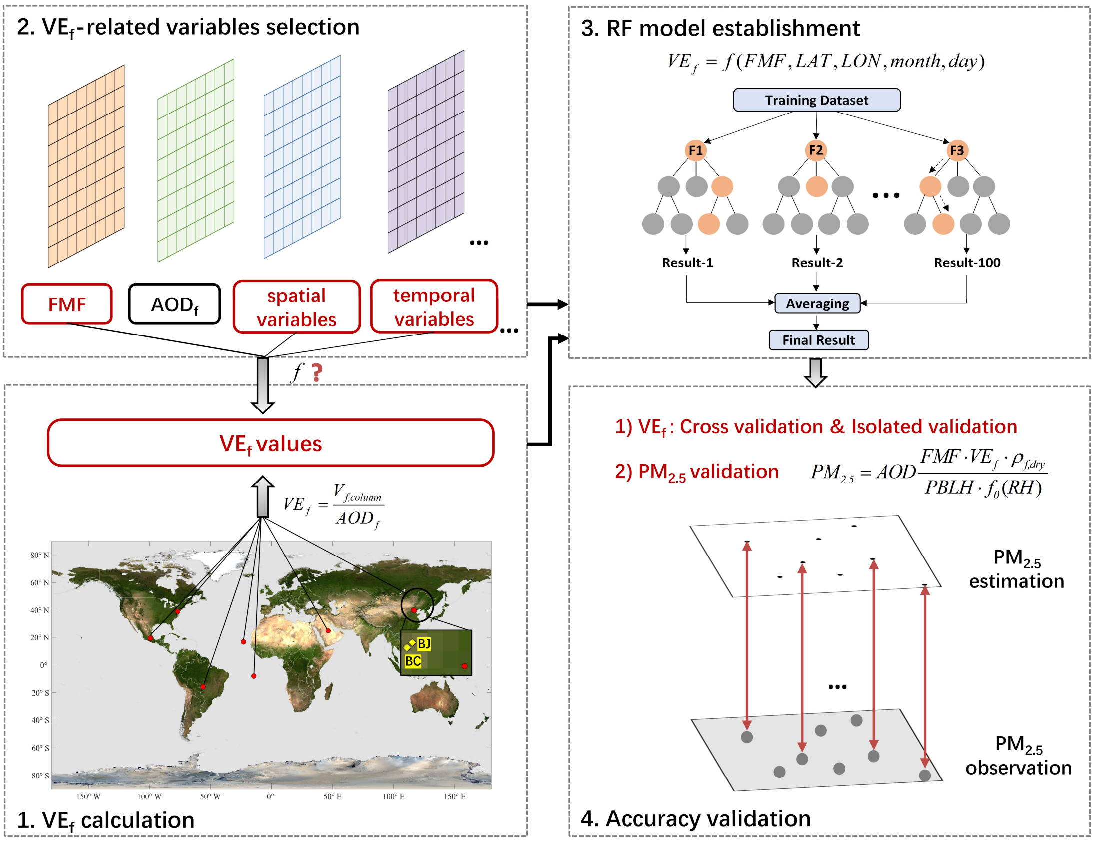

Fig. 3. Specific steps for simulating VE_f based on ML in our RF-PMRS method. The map used in step 1 is from NASA Visible Earth (https://visibleearth.nasa.gov/images/57752/blue-marble-land-surface-shallow-water-and-shaded-topography). The red points in step 1 represent the distribution of the 9 AERONET sites and the two yellow quadrangles in the zoom-in view highlight the Beijing (BJ) and Beijing-CAMS (BC) sites.

Step 2 VE_f-related variables selection

According to the theory, FMF is selected as the most important modeling variable. Previous studies have also shown that the FMF-VE_f relationship has a good single-value correspondence, which is not affected by AOD. Compared with AOD_f and $V_{f,column}$, FMF is a better indicator for estimation (Zhang and Li, 2015). In addition, considering the spatiotemporal heterogeneity of VE_f, the latitude, longitude (LAT, LON), and data time (month, day) of each site are added to the training.

Step 3 RF model establishment

From step 2, VE_f can be expressed as:

$$VE_f = f(FMF, LAT, LON, month, day) \tag{11}$$

We optimize VE_f expression based on random forest (RF). RF is made up of multiple decision trees that can build high-accuracy models based on fewer variables (Ho, 1995; Yang et al., 2020). This ensemble ML method randomly samples the training dataset to form multiple subsets and random combinations of features are selected in node splitting (Belgiu and Drăguţ, 2016). The specific process is to 1) generate training subsets, 2) build an optimal model, and 3) calculate the result (Fig. 3 shows its flowchart). Note that the station FMF values (S-FMF) from AERONET sites are used when training.

Step 4 Accuracy validation

The VE_f estimation is also based on equation (11), where f is the optimal relationship after RF parameter adjustment, and Phy-DL FMF is applied to realize the extension of model results from point to surface. 10-fold cross-validation (CV) (Rodriguez et al., 2009) and isolated-validation (IV) are used to evaluate model performance (For details of the validation methods, see Appendix A1).

3) $PM_{2.5}$ value estimation and evaluation

Then, we calculate $PM_{2.5}$ according to the corresponding process (equation (8)). The variables (in sections 2.2 to 2.4) are spatially matched to ground sites at their respective resolutions. And based on UTC, the $PM_{2.5}$ validation is conducted on a daily scale in 2017. Because of the effective quantity of the AERONET public dataset and MODIS data, we choose 2017 as the representative year. Note that we select the measured empirical value of $\rho_{f,dry}$ (i.e., 1.5 g/cm^3) for the NC region from Gao et al. (2007).

The statistical indicators used in the evaluation include correlation coefficient (R), mean bias (MB), relative mean bias (RMB), root mean square error (RMSE), and mean absolute error (MAE). In addition, relative predictive error (RPE) is added to validate the accuracy of the RF-based VE_f model. See Appendix A2 for the specific information on these indicators.

4. Experiment results

Three main experiments are conducted to verify the proposed RF-PMRS method, and the specific information is shown in Table 2.

Table 2. A brief information summary of the experiments conducted in our study.

Experiment	Object	Region	Period	Time scale
Model performance for training VE_f	VE_f	Global scale (Nine AERONET sites)	CV: Training period in Table 1 IV: Isolated-validation period in Table 1 (See Appendix A1)	Daily
Accuracy evaluation of PMRS/RF-PMRS	$PM_{2.5}$	Two AERONET Sites: Beijing, Beijing-CAMS	2017	Daily
Generalization performance of RF-PMRS	$PM_{2.5}$	North China region	2017	Daily

4.1. RF model performance for training VE_f

The simulation model of VE_f is trained based on the data in Table 1. Specifically, the 10-fold CV result is used to determine the optimal combination of parameters for the model, and see Appendix A3 for the adjustment of the model parameters. Considering that the completeness of the training data will optimize the generalization performance of the model, the experiment fine-tunes the model based on all the original datasets (the training period of Table 1) under the optimal parameters, then the final RF model is constructed. This is also the most common method for ML model construction. Next, the IV experiment provides independent time validation of the final model.

Table 3 shows the CV and IV results to respectively demonstrate the internal and external accuracy of the final RF model. It can be seen that RF can capture the complex relationship between VE_f and related variables well. R is as high as 0.974 (0.975), RMSE and MAE are both small, and RPE is around 30%, which suggests the desired estimation accuracy. Overall, the CV results represent the great performance of the RF model for extracting information, that is, the relationship of multi-source data to VE_f. In the meantime, the statistical results in CV and IV experiments are similar, indicating that the RF model has no obvious overfitting phenomenon.

Table 3. Performance statistics of the RF model for training VE_f. N represents the number of data, and VE_f has no unit.

	R	RMSE	RPE	MAE	N
Cross-validation（CV）	0.974	0.076	32.9%	0.034	6463
Isolated-validation（IV）	0.975	0.067	29.8%	0.037	814

4.2. Accuracy evaluation of PMRS/RF-PMRS at AERONET stations

The purpose of RF-PMRS is to construct an optimal model from the obtained point matching data pairs, and generalize it to the space-time continuous surface data for VE_f derivation. In the subsequent experiments in sections 4.2 and 4.3, the VE_f values are obtained by introducing the Phy-DL FMF dataset (surface data) to the final RF model. At the same time, the Phy-DL FMF data is also applied to the $PM_{2.5}$ calculation process (FMF variable in Equation 8) for a wide range of $PM_{2.5}$ concentration.

Then, the experiment compares $PM_{2.5}$ results of PMRS and RF-PMRS at Beijing (BJ) and Beijing-CAMS (BC) AERONET sites in 2017. Here, RF-PMRS simulates VE_f based on RF, and replaces the polynomial of the PMRS method. Note that the results of the two sites are compared with their respective nearest ground $PM_{2.5}$ stations (distances of 3.64 km and 3.91 km, respectively, in line with the representative range of ground stations in previous studies (Shi et al., 2018)). Fig. 4 displays the time series of $PM_{2.5}$ values for different models at two sites. The blue line fits the red line better than the gray one, confirming that the $PM_{2.5}$ results of RF-PMRS are closer to the true values. Within the range of the black circles at positions 1 and 2, the variation of RF-PMRS results has better consistency with the ground truth, while the PMRS results show dislocation and excessive growth. The overall performance of the RF-PMRS estimations can signify the effectiveness of our proposed method framework. As observed in the red boxes at positions 3 and 4, both models have a certain degree of deviation, which is found to be consistent with the time regularity of the AOD high values. Meanwhile, Fig. B1 (in Appendix B) plots the bias time series between PMRS/RF-PMRS and in-situ values. As can be seen, the bias of the optimization method (RF-PMRS) is stably distributed around zero, which greatly reduces the numerical uncertainty. And it is worth noting that our method has well mitigated the apparent overestimation of the original model (PMRS) in the case of above-normal aerosol loadings. Furthermore, the average $PM_{2.5}$ values from ground stations, PMRS, and RF-PMRS are compared. As for the two sites, the RF-PMRS results are satisfactory. As depicted in Fig. 5, the RF-PMRS and station mean values are close, with a difference of 4.82 μg/m³ (BJ) and 2.73 μg/m³ (BC), suggesting a good estimation. Nevertheless, the PMRS results have deviations greater than 40 μg/m³, and overestimation exists at both sites. It can be inferred that, in our proposed method, the optimization of VE_f can greatly improve the $PM_{2.5}$ estimation accuracy.

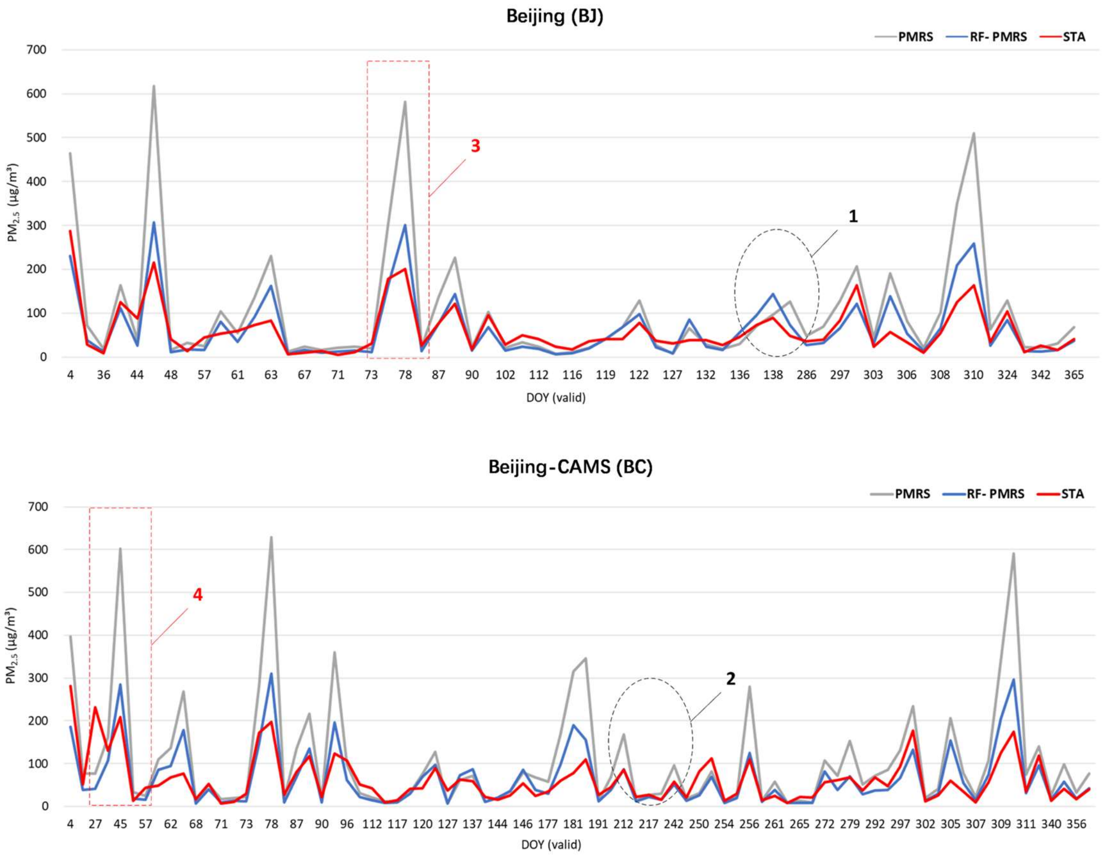

Fig. 4. Three $PM_{2.5}$ time series at the Beijing (BJ) and Beijing-CAMS (BC) sites under their respective DOYs in 2017. Here, DOY (valid) means the day of the year with valid AOD, FMF, and other $PM_{2.5}$-related data. Grey, blue, and red lines represent $PM_{2.5}$ values of PMRS, RF-PMRS, and stations (STA), respectively. The red boxes and black circles select a specific period for analysis.

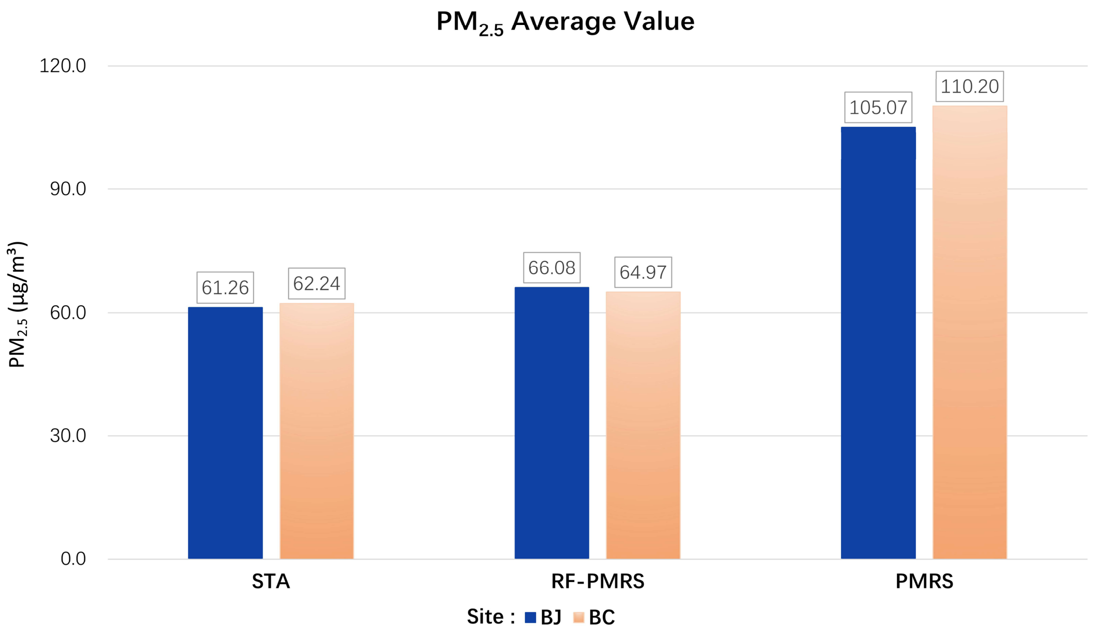

Fig. 5. Annual average $PM_{2.5}$ values from stations (left), RF-PMRS (middle), and PMRS model (right) at the BJ and BC sites.

To visually compare the optimization effect, Fig. 6 plots the $PM_{2.5}$ bias distribution patterns for two methods. From the boxplot, the average $PM_{2.5}$ bias of RF-PMRS is close to zero (less than 5 $\mu g/m^3$), which is greatly lower than that of PMRS. Besides, PMRS $PM_{2.5}$ has a larger deviation range, which manifests in two aspects. One is the maximum bias, specifically, it has exceeded 100 $\mu g/m^3$ at the BC site. The other is the overall distribution of the data bias, the BJ site ones are mostly distributed below zero, indicating an obvious overestimation. As for RF-PMRS, the above circumstances are not obviously reflected in it. In addition, as can be seen from the indicators, RMSE and MAE of RF-PMRS $PM_{2.5}$ decrease by about half in comparison with PMRS. And the experiment has confirmed that the RF-PMRS $PM_{2.5}$ values have a strong linear relationship with the ground truth at both sites, with R around 0.8 (0.82 at BJ and 0.78 at BC). Such a large optimization effect is attributed to the VE_f expression replacement to the fitted RF model.

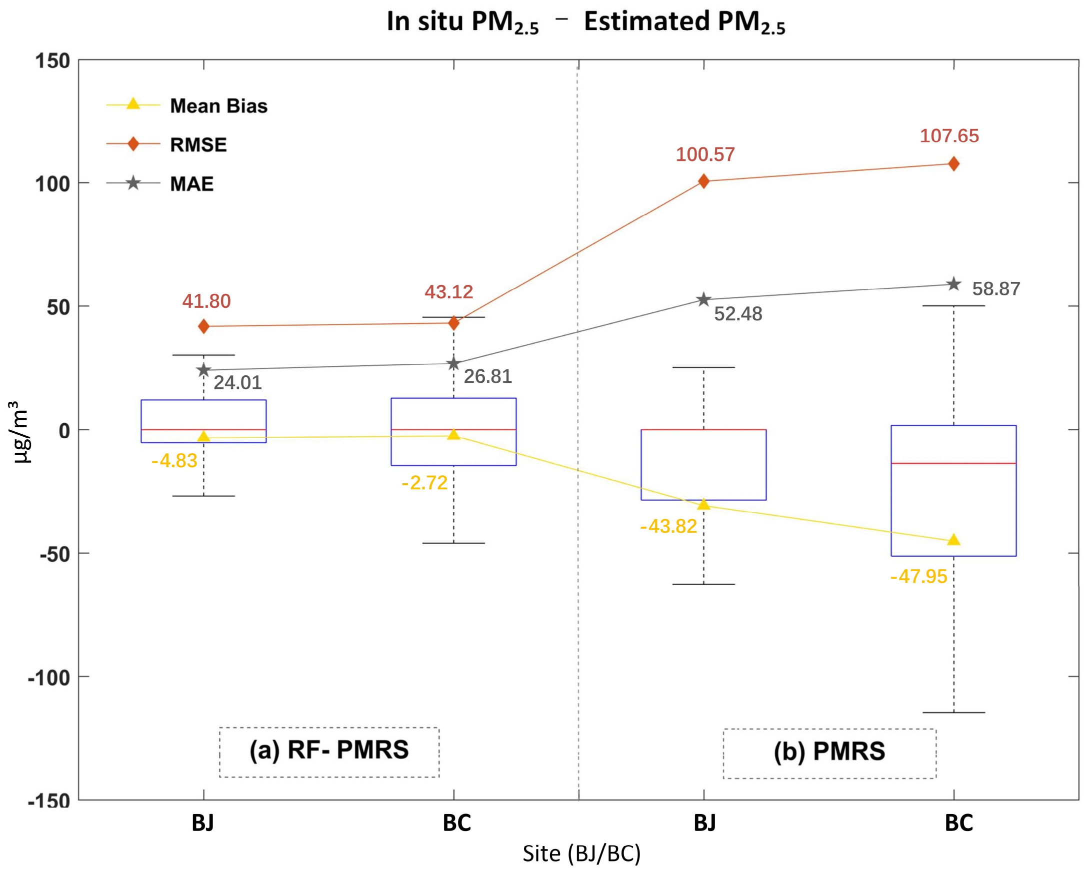

Fig. 6. Boxplots of RF-PMRS (a) and PMRS (b) $PM_{2.5}$ bias at the BJ and BC sites. The upper (lower) black line of each box represents the largest (smallest) value, the blue upper (lower) border represents the upper (lower) quartile, and the red line denotes the median. Besides, the yellow, orange, and gray symbols are the MB, RMSE, and MAE of the corresponding $PM_{2.5}$ concentration.

4.3. Generalization performance of RF-PMRS

Then, we estimate $PM_{2.5}$ based on PMRS and RF-PMRS within North China in 2017 (Fig. 1 exhibits the distribution pattern of the validation stations). Table 4 shows the accuracy statistics. It can be seen that RF-PMRS greatly reduces the bias (about 44.87%), with MB of about 2.31 μg/m³. Similar to the results at the sites, the RF-PMRS method can derive $PM_{2.5}$ concentration with practically no overestimation (underestimation). Although there is not much difference in R values of the two models (R of RF-PMRS is only improved by 0.01), RMSE and MAE of which decrease by about 39.96 μg/m³ and 18.86 μg/m³, respectively. As a result, the optimized method deserves to be considered excellent.

Table 4. Validation results of PMRS and RF-PMRS $PM_{2.5}$ in North China.

Method	R	MB (μg/m³)	RMB (%)	RMSE (μg/m³)	MAE (μg/m³)
PMRS	0.69	-29.34	48.71%	79.98	44.72
RF-PMRS	0.70	2.31	3.84%	40.02	25.86

Meanwhile, $PM_{2.5}$ scatterplots are presented below. As depicted in Fig. 7, there are sufficient estimated samples (28305) in the NC region, which guarantees the credibility of our validation results. In general, the RF-PMRS $PM_{2.5}$ values are distributed around the 1:1 reference line evenly, with a slightly higher R of 0.70 compared to that of the original method. And the slope of the linear fitting relationship reaches 0.82, which indicates that the proposed method greatly reduces the overestimation of PMRS with a linear slope of 1.46. Although the overall performance of the RF-PMRS estimations maintains an excellent level, defects do remain. To be specific, in areas with high $PM_{2.5}$ concentration (especially greater than 150 μg/m³), RF-PMRS results exist a slight underestimation. It may be caused by the relatively small number of high-value $PM_{2.5}$ points (only 1319 out of 28305), which is difficult to adequately reflect the fitting effect of the method.

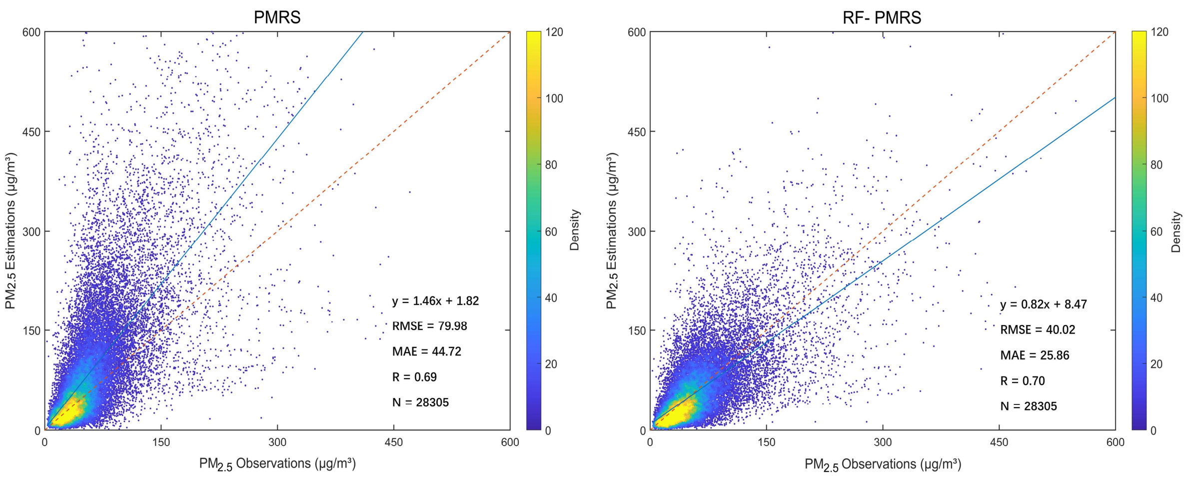

Fig. 7. Validation scatterplots of $PM_{2.5}$ results from PMRS (left) and RF-PMRS (right). Red dashed lines are 1:1 reference lines, and blue solid lines stand for the linear fits. The right legends show the point densities (frequency) represented by different colors.

As for RF-PMRS, the deviation is reduced to a large extent, so the probability density function maps based on the bias of PMRS and RF-PMRS are further drawn. Fig. 8 visualizes the probability densities within different bias ranges. In terms of distribution characteristics, the overall bias of RF-PMRS from the zero value (black solid line) is small. About the curve shape, it is high and narrow, manifesting that the bias has a lower standard deviation (STD) and is more prone to appear around the mean. However, PRMS shows a more discrete distribution pattern, and there are many outliers outside the range of greater than 600 μg/m³. Simultaneously, as can be concluded from the three boxes, within the bias range of ±20 μg/m³ and ±40 μg/m³, the data numbers of RF-PMRS results increase by 8.32% and 12.81%, respectively. Outside the range of ±100 μg/m³, the number decreases by 9.10%. Therefore, as far as the accuracy is concerned, RF-PMRS results have lower bias and better stability.

In addition to the above general performance comparison in Section 4.3, Fig. 9 presents the annual average RMSE spatial distribution of PMRS and RF-PMRS $PM_{2.5}$ at NC stations. The two methods show a large deviation in the middle and southeast, and the RMSE map of PMRS has more red points. However, RF-PMRS can weaken this phenomenon very well since its RMSE representative colors are generally light. In particular, the proportion of dark red sites (RMSE greater than 60 μg/m³) decreases from 65.44% (PMRS) to 4.15% (RF-PMRS). In the areas where the ground stations are clustered, the deviation also reduces significantly.

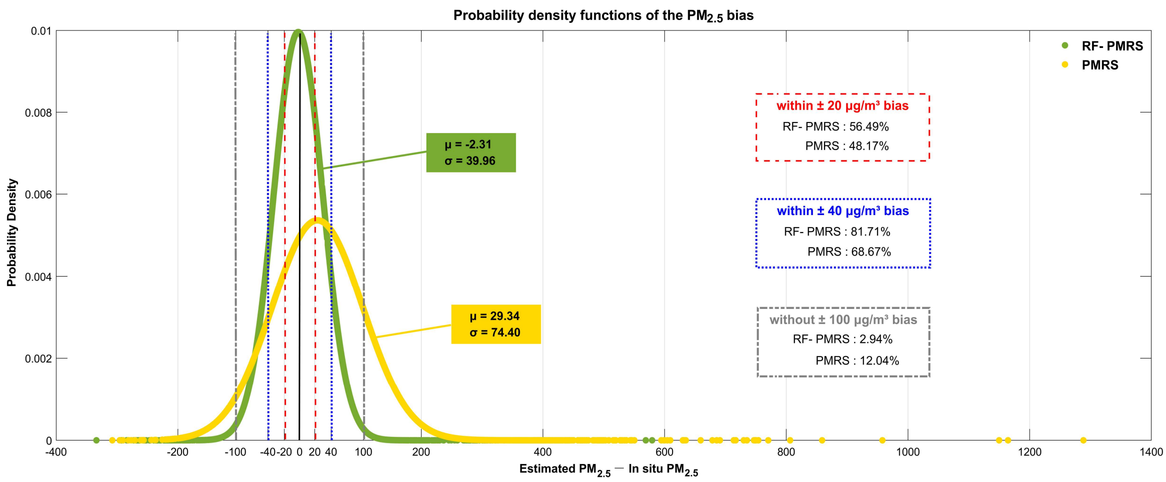

Fig. 8. Probability density functions of PMRS (yellow) and RF-PMRS (green) $PM_{2.5}$ bias. The red,

blue and grey dotted lines indicate the bias boundaries of ±20 μg/m³, ±40 μg/m³, and ±100 μg/m³, respectively. μ and σ represent the mean value and standard deviation of each data.

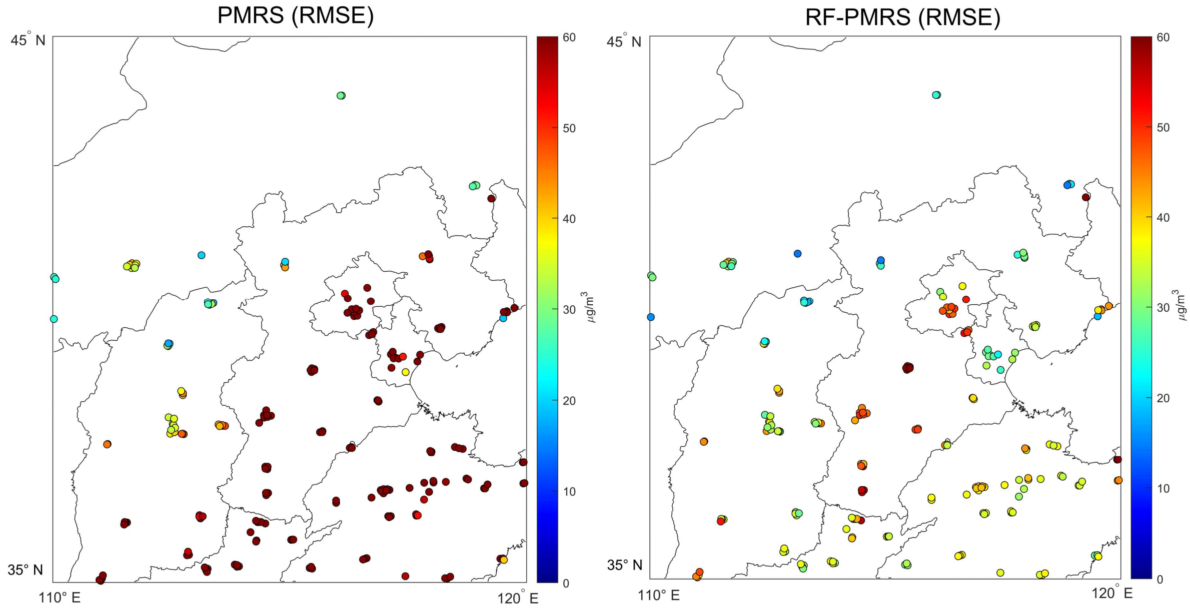

Fig. 9. RMSE of the year-average $PM_{2.5}$ concentration values between different models and ground stations (left: PMRS $PM_{2.5}$, right: RF-PMRS $PM_{2.5}$). Note that the top red of the RMSE legend indicates RMSE values equal to or greater than 60 μg/m³.

In a word, the above analysis demonstrates that compared with the simple quadratic polynomial relationship (equation (10)), the established RF model in RF-PMRS can more accurately capture the relationship between VE_f and multiple variables, thereby improving the $PM_{2.5}$ estimation accuracy.

5. Discussion

5.1. Accuracy comparison of PMRS using MODIS/Phy-DL FMF

To confirm the superiority of the Phy-DL FMF data adopted in our method framework, the experiment takes the BJ and BC sites as examples (in 2017), and then compares the $PM_{2.5}$ accuracy and the number of effective days calculated by PMRS based on different FMF. Table 5 presents the overall day-level results. Here, ‘DOY’ means the day of the year and ‘valid’ means that all variables related to the $PM_{2.5}$ calculation are valid. As can be seen, after the FMF replacement, the valid DOY turns out to become more (an increase of 113 days), which illustrates that the number of effective $PM_{2.5}$ concentrations has gone up by about 5 times. Moreover, the accuracy has been significantly enhanced, with R increased by about 0.30, RMSE and MAE decreased by 26.14% and 16.47% accordingly. On the whole, Phy-DL FMF contributes to the improvement of PMRS results, signifying the first step optimization of the proposed RF-PMRS method is effective.

Table 5. Validation results of the PMRS method using different FMF data. The valid DOY refers to the number of days that the AOD, FMF, and other data are not missing when calculating $PM_{2.5}$. Note that since the valid days of the two schemes are different, the MB and RMB are not compared.

	Valid DOY	R	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)
PMRS with MODIS FMF	30	0.38	63.01	35.64
PMRS with Phy-DL FMF	143	0.68	46.54	29.77

5.2. Performance compared with other ML models

Different machine learning models are suitable for diverse research data, and decision tree (DT) models can better fit experiments with fewer variables, such as this study. For comparison, except for RF, the Extremely Randomized Tree (ERT) (Geurts et al., 2006) and Gradient Boosting Decision Tree (GBDT) (Friedman, 2001) models have also been established. The results of training VE_f based on the above three DT models are presented in Table 6 and Table 7. By contrast, RF performs best in CV and IV experiments, as indicated by the multiple accuracy indicators. Although ERT and GBDT models are comparable to RF in some indicators, there exists a certain degree of overfitting in the above two models, which is manifested in that their IV results are clearly worse than their respective CV ones. Thus, the RF model is applied to our study.

Table 6. Cross-validation results in comparison of the decision tree models for training VE_f. N represents the number of data, and VE_f has no unit.

CV results					
	R	**RMSE**	**RPE**	**MAE**	**N**
RF	0.974	0.076	0.330	0.034	6463
ERT	0.972	0.079	0.343	0.035	
GBDT	0.973	0.078	0.339	0.036	

Table 7. Isolated-validation results in comparison of the decision tree models for training VE_f. The indicators are the same as those in Table 6.

IV results					
	R	RMSE	RPE	MAE	N
RF	0.975	0.067	0.299	0.037	814
ERT	0.967	0.076	0.340	0.042	
GBDT	0.969	0.074	0.331	0.040	

5.3. Feature importance of the embedded RF model

Additionally, the feature importance of RF is calculated to evaluate the contribution of model predictors to VE_f simulation. Fig. B2 (in Appendix B) shows the results by normalization (taking 100 as the total). Without a doubt, FMF accounts for the largest proportion, about 76.4%, which is consistent with the analysis when selecting the VE_f-related variables (see Section 3.2). The contribution of spatiotemporal variables is about 1/3 of FMF, which indirectly affirms the credibility of RF feature learning. Also, it provides a basis for further uncertainty optimization of VE_f and $PM_{2.5}$ accuracy.

5.4. Advantages and disadvantages

5.4.1. Advantages of the RF-PMRS method

From the perspective of model parameter optimization, this paper embeds RF to replace the subprocess parameter of the semi-empirical physical model. As a result, the proposed method, RF-PMRS, reduces the uncertainty of the complex physical parameter (i.e., VE_f) based on the estimation steps of strong physical significance, and realizes the coupling of machine learning and model mechanism. The proposed method does not rely on the $PM_{2.5}$ values of ground stations and is not affected by the station density and distribution mode, which can estimate the $PM_{2.5}$ concentration independently.

Meanwhile, as for the method, we construct the VE_f model based on RF using high-precision point data and extend it to surface data for $PM_{2.5}$ estimations. The experimental results demonstrate the overall performance of the model (Section 4.1) and its applicability in North China (Sections 4.2 to 4.3), showing that the method has certain universality from point scale to surface scale.

1) The overall performance of the model is high. We use the ground data of 9 AERONET sites around the world to train the RF model and simulate the VE_f values, the site distribution is relatively uniform and the amount of training data is sufficient. Table 1 shows a total of 6463 data matching pairs in the training period, which is enough to establish a credible RF model. Table 3 results show that in IV experiments, the accuracy of the model is well and can be generalized in different periods. For VE_f, the model shows both high internal accuracy (CV) and external accuracy (IV), so it can be generalized in regions with different aerosol types.

2) In the subsequent $PM_{2.5}$ estimation, the model displays high applicability in North China. From the perspective of model construction, the four aerosol types are the classification basis of the training data, and comprehensive modeling can improve the generalization performance. Also, the addition of spatiotemporal variables can increase the model applicability in North China. On the other hand, the number of stations used in an area does not determine the regional accuracy of the established model, which can be derived from our results. Compared with the $PM_{2.5}$ ground measurements in the NC region, the relative deviation of the RF-PMRS $PM_{2.5}$ is only 2.31 μg/m³, which confirms that RF can represent the relationships within North China.

5.4.2. Limitations on the scope of validation region

However, there are still some shortcomings, mainly manifested in the scope of the validation region. Due to limited experimental data, we only conduct experiments in North China (the main aerosol type is urban-industrial). The main reasons are: 1) insufficient $\rho_{f,dry}$ value. As the empirical value in the semi-physical empirical model, the $\rho_{f,dry}$ value is often obtained by field measurement and induction. The insufficient $\rho_{f,dry}$ values hinder the derivation of $PM_{2.5}$ in other regions and more research results are needed; 2) disclosure limits on global $PM_{2.5}$ ground measurements. Accurate and sufficient in-situ $PM_{2.5}$ values are the basic guarantee for the verification of estimated $PM_{2.5}$ results; 3) fewer public AERONET sites. Therefore, only BJ and BC sites in North China are used for representative point-scale validation.

5.4.3. Data differences and uncertainty analysis

In the RF-PMRS method, the VE_f model constructed by high-precision site data is generalized to surface data for validation, and the data types involved are as follows.

1) AERONET AOD vs. MODIS AOD

Two types of AOD are used for different experimental steps, among which AERONET AOD is applied to calculate the true values of VE_f for establishing the RF simulation model. And the RF model construction is a step of $PM_{2.5}$ estimation (as VE_f variable in equation (8)). MODIS AOD is satellite AOD data, which is the most commonly used remote sensing data for large-scale retrieval of $PM_{2.5}$. It is an important variable for $PM_{2.5}$ estimation in RF-PMRS (as AOD variable in equation (8)). Thus, there is no error in the $PM_{2.5}$ calculation caused by AOD category replacement.

As for uncertainty, AERONET AOD provides truth values for calculating VE_f, which theoretically has negligible uncertainty, and the simulation accuracy of VE_f represents its influence on estimating $PM_{2.5}$ to a certain extent. And it is generally considered that MODIS AOD has guaranteed quality and sufficient accuracy to be used directly.

2) S-FMF vs. Phy-DL FMF

S-FMF is obtained directly from the AERONET monitoring sites and is one of the variables of the RF model (as FMF variable in equation (11)). In the point-to-surface extension, Phy-DL FMF is introduced into the RF model to replace S-FMF, and the 2017 VE_f values are obtained. The basis of the above replacement is that the accuracy of Phy-DL FMF is relatively consistent with that of S-FMF (Yan et al., 2022). Besides, Phy-DL FMF data is applied to the $PM_{2.5}$ estimation steps (as FMF variable in equation (8)) for a wider range of validation experiments. The results show that the $PM_{2.5}$ concentration estimated by RF-PMRS has high accuracy, proving the credibility of Phy-DL FMF.

3) FMF uncertainty

Different surface data sources may affect the $PM_{2.5}$ results, introducing some uncertainty. Section 5.1 compares the $PM_{2.5}$ accuracy using two FMF data in 2017. The data missing time for MODIS FMF and Phy-DL FMF in North China are different, which can be found in the statistics on their respective available days (refer to valid DOY). There are far more valid days based on Phy-DL FMF than MODIS FMF (143 and 31 days), demonstrating the superiority of Phy-DL FMF. Although the specific validation time of two FMF varies, the overall accuracy of the $PM_{2.5}$ estimation (which can be regarded as the average accuracy over the year) shows that the Phy-DL FMF increases R to 0.68 (MODIS FMF: 0.38) with low uncertainty.

4) $\rho_{f,dry}$ uncertainty

As introduced earlier, the $\rho_{f,dry}$ value is often obtained by field measurement. In our study, we select 1.5 g/cm^3 as the $\rho_{f,dry}$ value for North China. There are certain variations in the empirical values of different regions, and there will be errors (uncertainty) between the values in Beijing and other places in the NC region. However, our experimental area is not large, and we use 1.5 g/cm^3 to represent $\rho_{f,dry}$ of the whole region, which has been applied in previous articles (Zhang and Li, 2015; Li et al., 2016).

5) Uncertainty between variable resolutions

In most experiments, the lowest resolution of all data will be taken as the unified resolution when obtaining data values. The different data may lose some spatial details during the upsampling/downsampling process, which brings uncertainty to the estimation results. In RF-PMRS method, there is no such uncertainty problem. We set 1° as the unified spatial unit, and take the longitude and latitude of each cell's center as the reference longitude and latitude. The variables in the data section are spatially matched to ground sites at their respective resolutions and the space-time matching method has been described in the method section. So, all kinds of data uncertainties only exist in their instrument measurement or statistical release.

Overall, RF-PMRS shows excellent estimation performance in North China, and the accuracy of surface $PM_{2.5}$ estimation based on remote sensing data is guaranteed. Next, with the improvement of related experimental data, we will verify our proposed method in a broader range and continuously optimize it from all aspects.

6. Conclusion

Among various satellite remote sensing methods for $PM_{2.5}$ retrieval, the semi-empirical physical approach has strong physical significance and clear calculation steps and derives the $PM_{2.5}$ mass concentration independently of in situ observations. However, the parameters with the meaning of optical properties are difficult to express, which need to be optimized. Hence, the study proposes a method (RF-PMRS) that embeds machine learning in a physical model to obtain surface $PM_{2.5}$: 1) Based on the PMRS method and select the Phy-DL FMF product with a combined mechanism; 2) Use the RF model to fit the parameter VE_f, rather than a simple quadratic polynomial. In the point-to-surface validation, RF-PMRS shows great optimized performance. Experiments at two AERONET sites show that R reaches up to 0.8. And in North China, RMSE decreases by 39.95 μg/m³ with a 44.87% reduction in relative deviation. In the future, we will further explore the combination of atmospheric mechanism and machine learning, then research the $PM_{2.5}$ retrieval methods with physical meaning and higher accuracy.

Appendix A: Supplementary description

A1. 10-fold cross-validation and isolated-validation

The sample-based 10-fold cross-validation method is applied to tune the model parameters and test the internal accuracy of our model. The original dataset is randomly divided into ten parts, nine of which are used as the training set for model fitting, and the remaining one is used for prediction, then the cross-validation process is repeated ten rounds until each data has been used as the test set.

At the same time, when verifying the RF-based VE_f model, the dataset in the period that did not participate in the training in Table 1 is used for isolated-validation.

A2. Statistical indicators

$$R = \frac{\sum_{i=1}^{m}(y_i - \overline{y})\sum_{i=1}^{m}(f_i - \overline{f})}{\sqrt{\sum_{i=1}^{m}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{m}(f_i - \overline{f})^2}}$$

$$MB = \overline{y} - \overline{f}$$

$$RMB = \mathrm{abs}\left(\frac{\bar{y}-\bar{f}}{\bar{y}}\right)$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - f_i\right)^2}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - f_i\right|$$

$$RPE = \frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - f_i\right)^2}}{\bar{y}}$$

where m is the total number of observations, i is the number of measurements, y_i is the i-th observation, f_i is the corresponding estimation result. And $\bar{y}$ and $\bar{f}$ are the averages of all observations and estimates, respectively.

A3. Parameter adjustments of the RF model

The four parameters of RF are adjusted, that is the correlation coefficient r changes with (a) the number of trees, (b) maximum depth, (c) maximum number of features when splitting, (d) minimum number of split samples. Experiments show that the maximum depth varies greatly in a small range. To prevent overfitting, the four parameters of RF are adjusted to 60, 10, 2, and 8. It can ensure high accuracy while improving training efficiency.

Appendix B: Figures

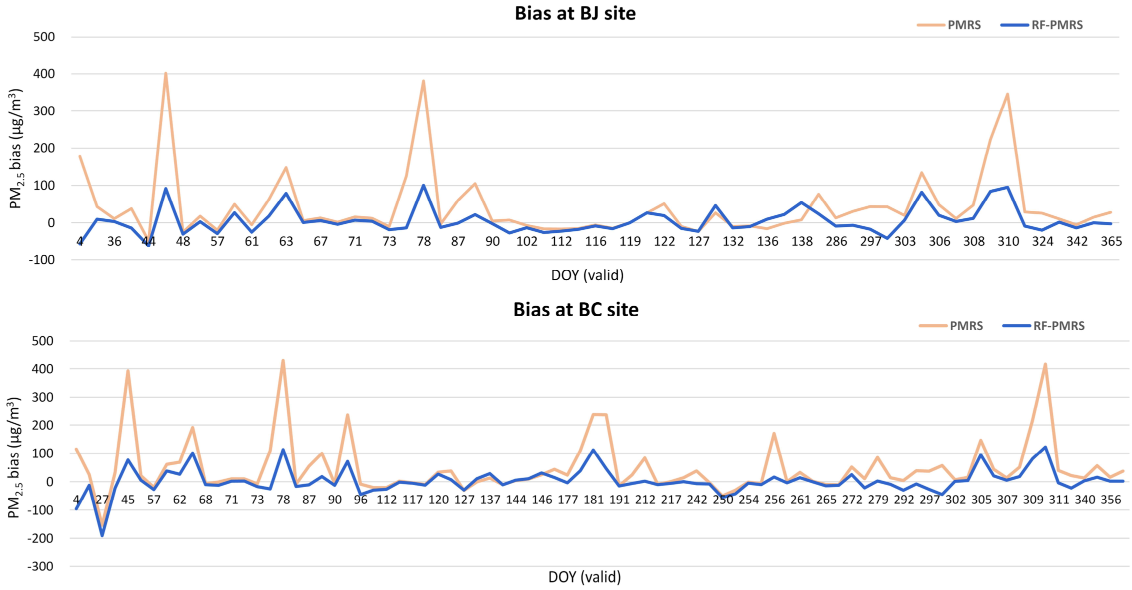

Fig. B1. The time series of PMRS/RF-PMRS $PM_{2.5}$ bias at the Beijing and Beijing-CAMS sites under their respective DOYs in 2017. The orange line represents the bias between the $PM_{2.5}$ values of PMRS and stations, while the blue one indicates the $PM_{2.5}$ difference between RF-PMRS and stations.

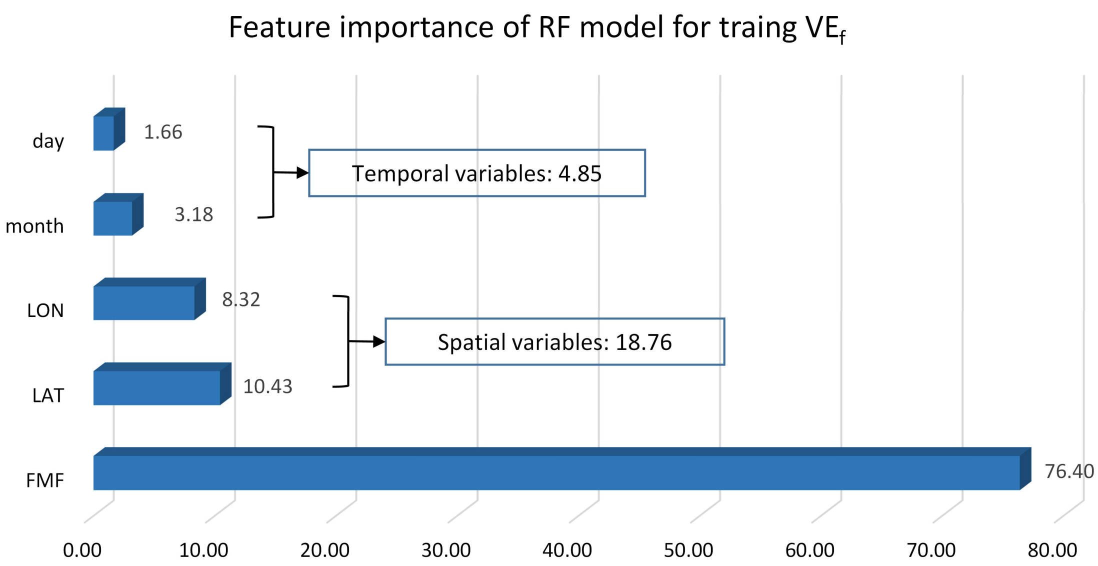

Fig. B2. The predictor importance results (normalized) of the RF model for training VE_f.

Code and data availability

All relevant codes as well as the intermediate data of this work are archived at https://doi.org/10.5281/zenodo.7183822 (Jin, 2022). The MCD19A2 data can be downloaded on https://ladsweb.modaps.eosdis.nasa.gov (last access: 30-09-2022) (Lyapustin and Wang, 2015). Detailed information about the Phy-DL FMF dataset can be found at https://doi.org/10.5281/zenodo.5105617 (Yan, 2021). Meteorological data used in this work are obtained at https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels (last access: 30-09-2022) (Hersbach et al., 2018). AERONET data was downloaded from https://aeronet.gsfc.nasa.gov/ (last access: 30-09-2022) (Giles et al., 2019).

Author contributions

Caiyi Jin: Data curation, Methodology, Formal analysis, Writing - original draft. **Qiangqiang Yuan:** Conceptualization, Supervision, Project administration, Writing - review and editing. **Tongwen Li:** Resources, Methodology, Writing - review and editing, Formal analysis. **Yuan Wang:** Methodology, Validation, Writing - review and editing. **Liangpei Zhang:** Supervision, Writing - review and editing.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgments

We gratefully acknowledge the Atmosphere Archive and Distribution System (LAADS), the ECMWF, the AERONET project, and the CNEMC for respectively providing the MODIS products, the meteorological data, the ground aerosol data, and the surface $PM_{2.5}$ concentration. We also thank other institutions which provide related data in this work.

Financial support

This research was funded in part by the National Natural Science Foundation of China (No. 41922008 and No. 42201359), the Hubei Science Foundation for Distinguished Young Scholars (No. 2020CFA051) and the Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515010492).

References

Belgiu, M., and Drăguţ, L.: Random forest in remote sensing: A review of applications and future directions, ISPRS J. Photogramm. Remote Sens., 114, 24-31, https://doi.org/10.1016/j.isprsjprs.2016.01.011, 2016.

Bowe, B., Xie, Y., Li, T., Yan, Y., Xian, H., and Al-Aly, Z.: The 2016 global and national burden of diabetes mellitus attributable to PM2.5 air pollution, Lancet Planet. Health, 2, e301-e312, https://doi.org/10.1016/S2542-5196(18)30140-2, 2018.

Chen, X., de Leeuw, G., Arola, A., Liu, S., Liu, Y., Li, Z., and Zhang, K.: Joint retrieval of the aerosol fine mode fraction and optical depth using MODIS spectral reflectance over northern and eastern China: Artificial neural network method, Remote Sens Environ, 249, 112006, https://doi.org/10.1016/j.rse.2020.112006, 2020.

Friedman, J.H.: Greedy function approximation: a gradient boosting machine, Ann Stat, 29(5), 1189–1232, http://www.jstor.org/stable/2699986, 2001.

Gao, J., Zhou, Y., Wang, J., Wang, T., and Wang, W.X.: Inter-comparison of WPSTM-TEOMTM-MOUDITM and investigation on particle density, Huan Jing Ke Xue, 28, 1929-1934, https://doi.org/10.3321/j.issn:0250-3301.2007.09.005, 2007.

Gao, L., Li, J., Chen, L., Zhang, L., and Heidinger, A.K.: Retrieval and validation of atmospheric aerosol optical depth from AVHRR over China, IEEE Trans Geosci Remote Sens, 54, 6280-6291, https://doi.org/10.1109/TGRS.2016.2574756, 2016.

Geng, G., Zhang, Q., Martin, R.V., van Donkelaar, A., Huo, H., Che, H., Lin, J., and He, K.: Estimating long-term PM2.5 concentrations in China using satellite-based aerosol optical depth and a chemical transport model, Remote Sens Environ, 166, 262-270, https://doi.org/10.1016/j.rse.2015.05.016, 2015.

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

Giles, D.M., Holben, B.N., Eck, T.F., Smirnov, A., Sinyuk, A., Schafer, J., Sorokin, M.G., and Slutsker, I.: Aerosol robotic network (AERONET) version 3 aerosol optical depth and inversion products, in: American Geophysical Union (AGU) 98th Fall Meeting Abstracts, New Orleans, America, 11-15 December 2017, A11O-01, 2017.

Giles, D. M., Sinyuk, A., Sorokin, M. G., Schafer, J. S., Smirnov, A., Slutsker, I., Eck, T. F., Holben, B. N., Lewis, J. R., Campbell, J. R., Welton, E. J., Korkin, S. V., and Lyapustin, A. I.: Advancements in the Aerosol Robotic Network (AERONET) Version 3 database - automated near-real-time quality control algorithm with improved cloud screening for Sun photometer aerosol optical depth (AOD) measurements, Atmos Meas Tech, 12, 169–209, https://doi.org/10.5194/amt-12-169-2019, 2019.

Gupta, P., and Christopher, S.A.: Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach, J. Geophys. Res. Atmos., 114, D14205, https://doi.org/10.1029/2008JD011496, 2009.

Hand, J.L., and Kreidenweis, S.M.: A new method for retrieving particle refractive index and effective density from aerosol size distribution data, Aerosol Sci Technol, 36, 1012-1026, https://doi.org/10.1080/02786820290092276, 2002.

Hänel, G., and Thudium, J.: Mean bulk densities of samples of dry atmospheric aerosol particles: A summary of measured data, Pure Appl. Geophys., 115, 799-803, https://doi.org/10.1007/BF00881211, 1977.

He, J., Yuan, Q., Li, J., and Zhang, L.: PoNet: A universal physical optimization-based spectral super-resolution network for arbitrary multispectral images, Inform Fusion, 80, 205-225, https://doi.org/10.1016/j.inffus.2021.10.016, 2022.

He, J., Li, J., Yuan, Q., Shen, H., and Zhang, L.: Spectral Response Function-Guided Deep Optimization-Driven Network for Spectral Super-Resolution, IEEE Trans Neural Netw. Learn. Syst., PP(99), 1-15, https://doi.org/10.1109/TNNLS.2021.3056181, 2021.

Ho, T.: Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14-16 August 1995, pp. 278-282 vol.1, http://doi.org/10.1109/ICDAR.1995.598994, 1995.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N.: ERA5 hourly data on single levels from 1979 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], (Accessed on 30-09-2022), https://doi.org/10.24381/cds.adbb2d47, 2018.

Holben, B.N., Eck, T.F., Slutsker, I., Tanré, D., Buis, J.P., Setzer, A., Vermote, E., Reagan, J.A., Kaufman, Y.J., Nakajima, T., Lavenu, F., Jankowiak, I., and Smirnov, A.: AERONET — A federated instrument network and data archive for aerosol characterization, Remote Sens Environ, 66, 1-16, https://doi.org/10.1016/S0034-4257(98)00031-5, 1998.

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E.A., Kadow, C., Staneva, J., and Saynisch-Wagner, J.: Towards neural Earth system modelling by integrating artificial intelligence in Earth system science, Nat. Mach. Intell., 3, 667-674, https://doi.org/10.1038/s42256-021-00374-3, 2021.

Jin, C.: An optimized semi-empirical physical approach for satellite-based PM2.5 retrieval: using random forest model to simulate the complex parameter, Zenodo [code], https://doi.org/10.5281/zenodo.7183822, 2022.

Koelemeijer, R.B.A., Homan, C.D., and Matthijsen, J.: Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe, Atmospheric Environ., 40, 5304-5315, https://doi.org/10.1016/j.atmosenv.2006.04.044, 2006.

Kokhanovsky, A.A., Prikhach, A.S., Katsev, I.L., and Zege, E.P.: Determination of particulate matter vertical columns using satellite observations, Atmos Meas Tech, 2, 327-335, https://doi.org/10.5194/amt-2-327-2009, 2009.

Lee, J.-B., Lee, J.-B., Koo, Y.-S., Kwon, H.-Y., Choi, M.-H., Park, H.-J., and Lee, D.-G.: Development of a deep neural network for predicting 6 h average PM2.5 concentrations up to 2 subsequent days using various training data, Geosci. Model Dev., 15, 3797–3813, https://doi.org/10.5194/gmd-15-3797-2022, 2022.

Li, T., Shen, H., Zeng, C., Yuan, Q., and Zhang, L.: Point-surface fusion of station measurements and satellite observations for mapping PM2.5 distribution in China: Methods and assessment, Atmospheric Environ., 152, 477-489, https://doi.org/10.1016/j.atmosenv.2017.01.004, 2017.

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., and Qie, L.: 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, 2016.

Lyapustin, A., Wang, Y., Laszlo, I., Kahn, R., Korkin, S., Remer, L., Levy, R., and Reid, J.S.: Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm, J. Geophys. Res. Atmos., 116, D03211, https://doi.org/10.1029/2010JD014986, 2011.

Lyapustin, A., Wang, Y., Xiong, X., Meister, G., Platnick, S., Levy, R., Franz, B., Korkin, S., Hilker, T., Tucker, J., Hall, F., Sellers, P., Wu, A., and Angal, A.: Scientific impact of MODIS C5 calibration degradation and C6+ improvements, Atmos Meas Tech, 7, 4353-4365, https://doi.org/10.5194/amt-7-4353-2014, 2014.

Lyapustin, A., andWang, Y.: MCD19A2 MODIS/Terra+Aqua Aerosol Optical Thickness Daily L2G Global 1km SIN Grid, NASA LP DAAC [data set], (Accessed on 30-09-2022), http://doi.org/10.5067/MODIS/MCD19A2.006, 2015.

Lyu, B., Huang, R., Wang, X., Wang, W., and Hu, Y.: Deep-learning spatial principles from deterministic chemical transport models for chemical reanalysis: an application in China for PM2.5, Geosci. Model Dev., 15, 1583–1594, https://doi.org/10.5194/gmd-15-1583-2022, 2022.

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

Pope III, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., and Thurston, G.D.: Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution, JAMA, 287, 1132-1141, https://doi.org/10.1001/jama.287.9.1132, 2002.

Raut, J., and Chazette, P.: Assessment of vertically-resolved PM10 from mobile lidar observations, Atmospheric Chem. Phys., 9, 8617-8638, https://doi.org/10.5194/acp-9-8617-2009, 2009.

Rodriguez, J.D., Perez, A., and Lozano, J.A.: 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, 2009.

Shi, X., Zhao, C., Jiang, J.H., Wang, C., Yang, X., and Yung, Y.L.: Spatial representativeness of PM2.5 concentrations obtained using observations from network stations, J. Geophys. Res. Atmos., 123, 3145-3158, https://doi.org/10.1002/2017JD027913, 2018.

Simmons, A.J., Untch, A., Jakob, C., Kållberg, P., and Undén, P.: Stratospheric water vapour and tropical tropopause temperatures in ECMWF analyses and multi-year simulations, Q J R Meteorol Soc, 125, 353-386, https://doi.org/10.1002/qj.49712555318, 1999.

Van Donkelaar, A., Martin, R.V., and Park, R.J.: Estimating ground-level PM2. 5 using aerosol optical depth determined from satellite remote sensing, J. Geophys. Res. Atmos., 111, D21201, https://doi.org/10.1029/2005JD006996, 2006.

Wang, Y., Yuan, Q., Li, T., Shen, H., Zheng, L., and Zhang, L.: Evaluation and comparison of MODIS Collection 6.1 aerosol optical depth against AERONET over regions in China with multifarious underlying surfaces, Atmospheric Environ., 200, 280-301, https://doi.org/10.1016/j.atmosenv.2018.12.023, 2019.

Wu, X., Wang, Y., He, S., and Wu, Z.: PM2.5/PM10 ratio prediction based on a long short-term memory neural network in Wuhan, China, Geosci. Model Dev., 13, 1499–1511, https://doi.org/10.5194/gmd-13-1499-2020, 2020.

Xiao, Y., Wang, Y., Yuan, Q., He, J., and Zhang, L.: Generating a long-term (2003−2020) hourly 0.25° global PM2.5 dataset via spatiotemporal downscaling of CAMS with deep learning (DeepCAMS), Sci. Total Environ., 848, 157747, https://doi.org/10.1016/j.scitotenv.2022.157747, 2022.

Xu, P., Chen, Y., and Ye, X.: Haze, air pollution, and health in China, Lancet, 382, 2067, https://doi.org/10.1016/S0140-6736(13)62693-8, 2013.

Yan, X., Zang, Z., Li, Z., Luo, N., Zuo, C., Jiang, Y., Li, D., Guo, Y., Zhao, W., Shi, W., and Cribb, M.: A global land aerosol fine-mode fraction dataset (2001--2020) retrieved from MODIS using hybrid physical and deep learning approaches, Earth Syst. Sci. Data, 14, 1193-1213, https://doi.org/10.5194/essd-14-1193-2022, 2022.

Yan, X., Li, Z., Shi, W., Luo, N., Wu, T., and Zhao, W.: An improved algorithm for retrieving the fine-mode fraction of aerosol optical thickness, part 1: Algorithm development, Remote Sensing of Environment, 192, 87-97, https://doi.org/10.1016/j.rse.2017.02.005, 2017.

Yan, X.: Physical and deep learning retrieved fine mode fraction (Phy-DL FMF), Zenodo [data set], (Accessed on 30-09-2022), https://doi.org/10.5281/zenodo.5105617, 2021.

Yang, Q., Yuan, Q., Li, T., and Yue, L.: Mapping PM2.5 concentration at high resolution using a cascade random forest based downscaling model: Evaluation and application, J. Clean. Prod., 277, 123887, https://doi.org/10.1016/j.jclepro.2020.123887, 2020.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., and Zhang, L.: Deep learning in environmental remote sensing: Achievements and challenges, Remote Sens Environ, 241, 111716, https://doi.org/10.1016/j.rse.2020.111716, 2020.

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., and Li, D.: Satellite remote sensing of atmospheric particulate matter mass concentration: Advances, challenges, and perspectives, Fundamental Research, 1, 240-258, https://doi.org/10.1016/j.fmre.2021.04.007, 2021.

Zhang, Y., Li, Z., Chang, W., Zhang, Y., de Leeuw, G., and Schauer, J.J.: Satellite observations of PM2.5 changes and driving factors based forecasting over China 2000−2025, Remote Sens., 12(16), 2518, https://doi.org/10.3390/rs12162518, 2020.

Zhang, Y., and Li, Z.: 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, 2015.