IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 1, 2019, accepted March 22, 2019, date of publication April 3, 2019, date of current version April 15, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2908975*

# High Spatial Resolution $PM_{2.5}$ Retrieval Using MODIS and Ground Observation Station Data Based on Ensemble Random Forest

**XUNLAI CHEN[1,2], HUI LI[1,2], SHUTING ZHANG[1,2], YUANZHAO CHEN[1,2], AND QI FAN[3]**

[1]Shenzhen Meteorological Bureau, Shenzhen 518040, China
[2]Shenzhen Key Laboratory of Severe Weather in South China, Shenzhen 518040, China
[3]Department of Atmospheric Science, Sun Yat-sen University, Guangzhou 510275, China

Corresponding author: Yuanzhao Chen (943508839@qq.com)

This work was supported in part by the Guangdong Meteorological Bureau Science and Technology Project under Grant GRMC2018Z06, in part by the China Meteorological Administration Forecaster Special Research Project under Grant CMAYBY2019081, in part by the Science and Technology Planning Project of Guangdong Province under Grant 2019B020208016, and in part by the National Natural Science Foundation under Grant 91544102.

**ABSTRACT** Limited by the number of ground observation stations, $PM_{2.5}$ retrieval from the remote sensing data is an effective complement to conventional ground observations and is a current research hotspot. The general principle behind the remote sensing retrieval of $PM_{2.5}$ is to first retrieve the aerosol optical depth (AOD) and calculate the $PM_{2.5}$ via the AOD-based statistical relationships. This method is likely to cause error propagation, which leads to instability in the retrieval model. In this paper, we propose a $PM_{2.5}$ remote sensing retrieval method via an ensemble random forest machine learning method to directly establish the relationship between the moderate-resolution imaging spectroradiometer (MODIS) images and ground observational $PM_{2.5}$ to avoid retrieval errors from the atmospheric aerosol optical depths and obtain $PM_{2.5}$ retrieval results with higher precision and spatial resolution. The proposed method first uses a random forest to train and validate the MODIS images and ground observation station $PM_{2.5}$ data; then, an optimal multi-model group, according to the determination coefficient R-square ($R^2$) index, is selected. Finally, the optimal multi-model group is used on the whole MODIS image to obtain the $PM_{2.5}$ retrieval result for the whole area. In an attempt to use machine learning technology to retrieve $PM_{2.5}$, the experiments selected a substantial amount of MODIS image data during four seasons in Guangdong Province for validation and compared three performance indicators ($R^2$, RMSE, and correlation coefficient (CC) to verify the superiority of the proposed algorithm.

**INDEX TERMS** Ensemble random forest, machine learning, remote sensing based $PM_{2.5}$ retrieval, Kriging interpolation, aerosol optical depth.

## I. INTRODUCTION

Increasing trends in industrialization and urbanization have led to severe environmental pollution, which further increases public attention and concern regarding the air quality index. Currently, fine particle matter ($PM_{2.5}$) is used as a crucial indicator to measure air quality. Thus, a scientific method to monitor the distributions and concentrations of $PM_{2.5}$ is significant for exploring its physical and chemical characteristics, discovering the reasons that contribute to the emergence of haze, and proposing possible measures to enable efficient air protection.

The state-of-the-art measures for $PM_{2.5}$ monitoring mainly include ground monitoring and satellite-based remote sensing monitoring [1]. It is widely agreed that ground monitoring can produce accurate results through field investigations. However, ground monitoring is impossible to generate massive temporally updated data with respect to $PM_{2.5}$ over large-scale areas due to a series of challenges involving high economic costs, a limited number of monitoring stations, and insufficient labour productivity [2], [3]. Satellite-based remote sensing monitoring can offset some limitations of traditional ground monitoring, which provides the capability

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

to conveniently acquire data and an extensive monitoring range. The current methods for $PM_{2.5}$ retrieval mainly consist of three sequential steps: 1) conducting the retrieval of atmospheric aerosol optical depths (AODs); 2) establishing the relationship between atmospheric AODs and field-measured $PM_{2.5}$; and 3) estimating $PM_{2.5}$ values over an area where monitoring stations are not available based on the relationship generated by the statistical analysis. Many scholars around the world have conducted a large number of studies using this type of approach.

Tian *et al.* developed a semi-empirical model to predict the hourly concentrations of ground-level $PM_{2.5}$ that coincided with a satellite overpass at a regional scale, which corrected AOD data from moderate-resolution imaging spectroradiometer (MODIS) by assimilating the parameters characterizing the boundary layer and further adjusting the corrected values according to meteorological conditions near the ground [4]. Lee et al. proposed a new methodology to calibrate AOD data obtained from MODIS, which was used to predict daily ground $PM_{2.5}$ concentrations in the New England region [5]. Liu et al. proposed an empirical model based on the regression between daily $PM_{2.5}$ concentrations and AOD values from the multiangle imaging spectroradiometer (MISR). The experimental results using data from the eastern United States during 2001 showed that the empirical model explained 48% of the variability in $PM_{2.5}$ concentrations [6]. Donkelaar *et al.* estimated the ground-level $PM_{2.5}$ concentrations from January 2001 to October 2002 using space-based measurements from MODIS and Multi-angle Imaging Spectroradiometer (MISR) satellite instruments and additional information from a global chemical transport model (GEOS-CHEM) [7]. Song et al used the MODIS C5 AOD products to retrieve $PM_{2.5}$ concentrations in the Pearl River Delta region [8]. Li *et al.* selected and investigated two intense winter haze events in Beijing that occurred in 2011 and 2012 using ground-based remote sensing measurements and the CIMEL CE318 sun-sky radiometer [9]. Chu *et al.* retrieved $PM_{2.5}$ concentrations from an area north of Taiwan [10]. However, errors occurred during the retrieval of AOD, which were further enhanced when building the statistical relationship between AOD and ground truth $PM_{2.5}$. Moreover, the errors suggested that AOD retrieval would influence the accuracy of $PM_{2.5}$ estimation.

The basic idea behind the atmospheric AOD retrieval includes the following aspects: assuming different aerosol modes and observing conditions to calculate the correlations among the optical thickness of atmospheric aerosols, the hemispherical reflectivity in the atmospheric boundary layer (ABL), the atmospheric reflectivity values, the satellite zenith angle and solar zenith angle. Based on the above relationship, a lookup table is created to obtain the AOD through a dynamic aerosol model. Based on the assumption that the top of the atmosphere is parallel to the surface and no clouds exist, Griggs and Rao et.al discovered the correlation between AOD and infrared and visible wavebands through simulation via an atmospheric radiation transmission model [11], [12]. Levy *et al.* integrated the aerosol information of the middle infrared waveband into the retrieval process and updated the previous lookup table to access the retrieval relationship between AOD and atmospheric aerosol modes [13], [14]. In addition to MODIS data, other remote sensing data have also been used to support the retrieval of optical thicknesses of atmospheric aerosols. Holben *et al.* [15] retrieved the AOD in the Mali Sahel region using Advanced Very High-Resolution Radiometer (AVHRR) data with an error of approximately 0.1. Isakov *et al.* [16] used Airborne Visible Infrared Imaging Spectrometer (AVIRIS) data to estimate the AOD in Oklahoma and Rapid City, South Dakota. The aforementioned retrieval methods have achieved good results in practical application, but due to the use of auxiliary data, the calculation accuracy is difficult to control. In addition, the AOD data itself have errors in the retrieval process; hence, the above methods cannot avoid error propagation [17]. Therefore, determining how to reduce error propagation and obtain a higher retrieval accuracy has been a hot topic in recent research [18], [19].

The overall processes of $PM_{2.5}$ retrieval based on remote sensing images generally consists of retrieving AODs, building a statistical relationship between AOD and $PM_{2.5}$, and using the developed relationship to predict $PM_{2.5}$ values in areas where ground-observed data are not available. The above mentioned processes are designed based on the assumption that there is a stable relationship between AOD and $PM_{2.5}$. Hence, the high accurate AOD is the key factor of the above processes. Thus, a majority of existing studies have focused on increasing the accuracy of AOD retrieval via improved estimation of surface reflectance and assumptions of aerosol models, as well as importing several corrections, adding auxiliary data, and integrating numerical prediction models.

However, several challenges still remain for the above processes:

(1) Stability in the relationship between AOD and $PM_{2.5}$. This is the basic assumption for retrieving $PM_{2.5}$ from AOD. A number of studies have proven that although a statistical relationship can be discovered between AOD and $PM_{2.5}$, this relationship varies across different areas and times. Thus, the stability of this relationship at a specific location and time is critical to $PM_{2.5}$ retrieval.

(2) Error propagation process. Although fine physical models can increase the accuracy of AOD retrieval and support the accurate retrieval of $PM_{2.5}$, these models cannot avoid error propagation. At present, there is no one physical model can exactly simulate the atmospheric motion. Hence, the error will occur when inventing any atmospheric parameter. More parameters are invented, more errors will propagate and accumulate together. The errors included in the AOD retrieval will have influence on the result from $PM_{2.5}$ retrieval. Thus, error propagation may reduce the accuracy of $PM_{2.5}$ retrieval in some regions. More auxiliary data are used, more uncertainties will be introduced. Currently, many researchers use daily average values, monthly average values,

seasonal average values and annual average values to study the relationship between AOD and $PM_{2.5}$, which could eliminate the uncertainty caused by error propagation to a certain degree. However, the influence of error propagation might be much greater at a specific time scale, for example, the hourly scale.

(3) Usability of the model. Although operations, including importing multiple corrections, adding auxiliary data and integrating numerical models, can increase the accuracy of AOD retrieval, importing more parameters increases the probability of producing more uncertainties, which weakens the usability of a model. Thus, since all the introduced auxiliary data are strongly related with the local environmental conditions, a model that integrates auxiliary data can fit for a study area, but may be incompetent for another area. That is, it may be difficult to transfer a fitted model to another region.

To address the above challenges, in this paper, we try another approach based on the following considerations: since the remote sensing derived parameters will introduce uncertainty and propagate errors, and the original image actually contains the atmospheric information when imaging (although many atmospheric components are mixed together), maybe we can establish the relationship between original spectral information and $PM_{2.5}$ so as to bypass the error propagation process using machine learning technology. Therefore, this paper proposes a $PM_{2.5}$ retrieval approach by integrating ensemble random forest machine learning and MODIS imagery. Compared with the other machine learning technologies, such as support vector machine [20], relevance vector machine [21], sparse representation [19] and so on, the random forest usually could provide better performance, and it could deal with very high dimensional data without feature selection [22], [23]. Based on MODIS remote sensing images, the proposed approach tries to establish a relationship between satellite imagery and $PM_{2.5}$ with the machine learning technique directly. Although the basic idea of the proposed method is simple, and lacks some strict physical meanings, the experimental results show that the proposed method can produce better retrieval results with a higher spatial resolution. As a simple approach and an attempt to address the previous mentioned challenges, there are much rooms for further improvement.

The remainder of this paper is organized as follows. Section 2 reviews the AOD retrieval algorithms of MODIS. Section 3 discusses the proposed ensemble random forest for $PM_{2.5}$ retrieval using MODIS data. Section 4 describes the experimental results with real images and compares the proposed results with the existing AOD-based results. Section 5 provides the discussion and last section is the conclusion.

## II. MODIS-BASED AOD RETRIEVAL

Terra and Aqua are two substantial satellites of the Earth Observation System (EOS). The MODIS, which is utilized in these two satellites, provides products with 36 spectral bands ranging from 0.4 $\mu$m to 14 $\mu$m and spatial resolutions ranging from 250 m to 1 km. The scanning width of MODIS is 2330 km. Thus, the advantages of MODIS, such as high spatial and temporal resolutions, multi-channels, and broad coverage, make it popularly to retrieve AOD.

After a number of improvements, the latest MODIS AOD was updated to the C6.1 version. The C5 version of MODIS AOD, which was released in 2008, provides a dark target (DT) algorithm and a deep blue (DB) algorithm to retrieve dark target areas and light target areas, respectively. It offers AOD products with a 10 km spatial resolution, and the integration of the DT and DB retrieval results is not available. In 2012, the C6 version of MODIS AOD supported the integration of the results generated from the DT and DB algorithms. Additionally, the spatial resolution of the C6.1 product generated by DT of the C6.1 version reached 3 km.

### A. DARK TARGET (DT) ALGORITHM

The C6 version of the DT algorithm [24] shares the same principles as the C5 version [13] and inherits from the original Kaufman implementation. The algorithm consists of the following steps: (1) Based on images with a spatial resolution of 500 m, pixels with an atmospheric reflectance (TOA) between 0.01 and 0.25 are selected. A $20 * 20$ pixel calculation window box is set and the inappropriate surface pixels such as clouds, desert, snow, ice, inland water, and so on are removed. (2) Some of the brightest and darkest pixels are discarded. Specifically, the 20% darkest and 50% brightest pixels from the calculation window box are discarded. Then more than 50 pixels from the remaining 120 pixels for the highest quality guaranteed aerosol retrieval ($QAF = 3$) are chosen. For $QAF = 0$, 1 and 2 retrieval, at least 12, 20 and 30 pixels should be selected. (3) Based on the assumed spectral/directional relationship, the short-wave infrared TOA reflectance is related to the surface reflectance in the visible bands of 470 nm and 650 nm. The fine-mode and the coarse mode (dust) dominated aerosol model are weighted by matching the average TOA reflectance over these bands, thereby applying this hypothetical relationship to the total AOD retrieval.

Recently, the aerosol product of the C6.1 algorithm has been released, and compared with the C6 retrieval, the following improvements have been made: For Land regions, if there are more than 50% coastal pixels or 20% of water pixels in $10 \times 10$ km box, the quality of retrievals is degraded to zero. When the percentage of urban area is larger than 20%, the algorithm for Aerosol retrieval over land surface is modified using a revised surface characterization using MYD09 spectral surface reflectance product.

### B. DEEP BLUE (DB) ALGORITHM

Hsu *et al.* improved and extended the original DB algorithm to form a new DB algorithm, i.e., the MODIS C6 version [25]. The new DB algorithm retrieves the cloudless and snowless pixels on $1 * 1$ km spatial resolution image, calculates the surface reflectance on the 0.412, 0.470, and 0.650 $\mu$m channels,

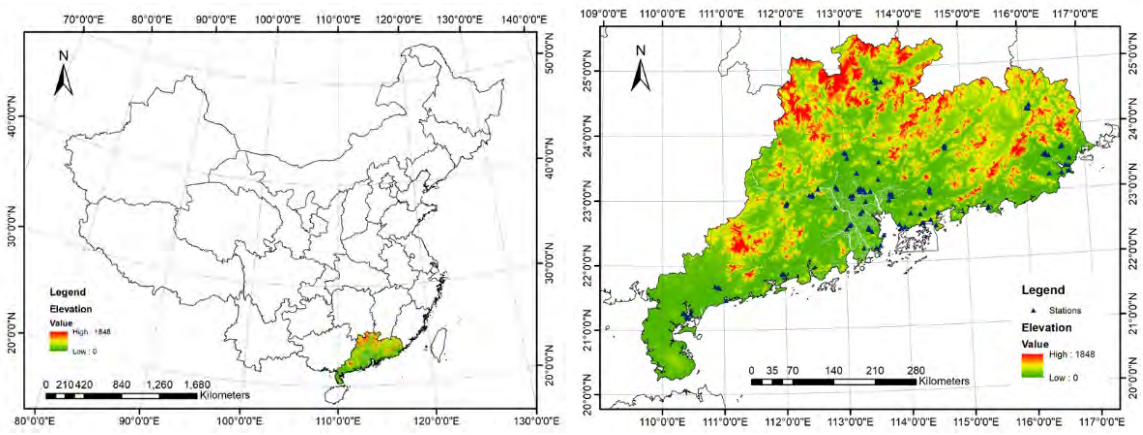


**FIGURE 1.** Study area.

and then re-aggregates to a resolution of $10 * 10$ km. At the same time, in addition to the surface of the vegetation, it is also retrieved on the surface of bright cities and deserts. Surface reflectance calculation depends on location, season and land cover type. There are three options available: (1) Surface reflectance database based on location, season, scattering angle and normalized difference vegetation index; (2) an empirical model derived from a bidirectional reflection distribution function for a particular region and season; (3) a spectral/direction relationship determined by the type of surface classification. The choice of these options depends on the TOA reflectivity in the 2.1 $\mu$m band and NDVI. For the hypothetical aerosol optical model, AOD was separately retrieved on the 470 nm, 412 nm, and 650 nm bands and then combined to determine the AOD at 550 nm.

For the newly released C6.1 version, the basic principle is consistent with C6, but further improvements have been made on the C6 version, including: (1) The internal smoke detection masks are improved to distinguish true cloud contamination. (2) The surface reflectance modelling for heterogeneous terrain surface is improved, and the QA tests to identify artefact pixels is also improved. (3) New surface reflectance models are developed for elevated terrain types to remove the systematic biases. (4) The assumed aerosol optical models in some areas are updated based on biases identified from the C6 validation work.

### C. MERGED DT/DB

Due to differences in assumptions and algorithms, the results of DT and DB algorithms have large differences in AOD coverage and spatial resolution. Therefore, a merged strategy is provided in the merged scientific data set (SDS), which combines the advantages of DT and DB to make new products which can increase the spatial coverage of AOD retrieval while ensuring good retrieval quality. The basic strategy of integration is to use MODIS's NDVI climate dataset (MYD13C2 dataset) to generate NDVI data with a spatial resolution of 0.25° per month for one year. Based on this, the merged AOD product is generated in three cases. (1) When the NDVI is less than or equal to 0.2, the merged SDS is filled with DB data; (2) When the NDVI is greater than or equal to 0.3, it is filled with DT data; (3) When the NDVI is between 0.2 and 0.3, the surface is at the transition zone between arid and vegetation, at this time, a product with a higher QA value is selected. If the QA values of both DT and DB are equal to 3, the average of the two is used. For marine areas, it is populated directly using the ocean algorithm. The new C6.1 version has the same integration strategy as the C6 version, but due to the update of the DT and DB algorithms in C6.1, the C6.1 merged data is different from C6.

## III. PROPOSED METHOD

### A. STUDY AREA

Guangdong Province is located in the southernmost region of mainland China. It borders Fujian to the east, Jiangxi and Hunan to the north, Guangxi to the west, and the South China Sea to the south. The east and west sides of the Pearl River estuary are bordered by Hong Kong and the Macao Special Administrative Region, respectively. The southwestern Leizhou Peninsula faces Hainan Province across the Qiongzhou Strait. The entire area is located between 20°13'-25°31'N and 109°39'-117°19'E. It spans approximately 800 km from east to west and approximately 600 km from north to south. The land area of the province is 179,800 square km. Guangdong Province belongs to the East Asian monsoon region; from north to south are the tropical regions of central Asia and South Asia, which have tropical climates and are some of the most abundant areas of light, heat and water resources in China. The Pearl River Delta region, which is primarily in Guangzhou Province, is one of the regions with the fastest urbanization in China, and the accompanying air pollution problems are also prominent. Figure 1 illustrates the study area. The triangular points represent the 102 observation stations, which offer hourly $PM_{2.5}$ ground truth data.

### B. RANDOM FOREST

The random forest (RF) is an extended version of the parallel ensemble learning method. A RF is based on a decision tree and introduces random attribute selection into the training process of the decision tree. It uses the bootstrap resampling technique (i.e., bootstrapping) to generate new training sample sets by randomly selecting $k$ samples via the replacement of $N$ in the original training set, which generates $k$ classification trees in the combined random forest. The classification results are determined based on the scores received from the votes given by all classification trees. Essentially, a random forest is an improved decision tree algorithm. Multiple decision trees are combined together. The creation of each tree depends on an independently selected sample. Every tree in the forest has the same distribution. The classification error depends on the capability of every tree and the correlations among them. During the feature selection procedure, every node is split randomly to obtain errors under different conditions. The number of selected features is determined by the estimation error, classification capacity and relativity. Although the classification ability of a single tree may be low, after randomly generating many decision trees, a test sample could be assigned to the most expected class via the statistical results of the classification for every tree. Particularly, when dividing the attributes, traditional decision trees select an optimal attribute from the attribute set of a current node (assuming that there are $d$ attributes). For each node in the original decision tree, the RF randomly selects a subset with $k$ attributes from the attribute set of one node; then, it selects an optimal attribute for division. The parameter$k$controls the degree of randomness, which generally equals $\log_2$d.

There are two main ideas in a RF:

(1) Bagging: A total of $K$ training sample sets ($T_k$, $k = 1, 2, \ldots, K$), which are of the same size, are randomly selected from the original sample set $X$ using random sampling with a replacement strategy. Approximately 37% of the samples are not selected each time. Then, a corresponding decision tree is created based on each training sample set.

(2) Featured subspace: When splitting each node in a decision tree, a featured subset is randomly selected from all features (generally, the number of selected features is $\log_2(M) + 1$, where M represents the total number of features); then, an optimal feature is selected from this featured subset to split the node.

While building each decision tree, the process of random selection for the training sample set and the feature subset is independent. It is a sequence comprising the independent and identical distribution of random variables. The training process in a RF is a process that trains each individual decision tree. Since the training process for each decision tree is mutually independent, parallel processing is suitable for substantially improving the efficiency of model generation. Then, a RF is generated by combining $K$ trained decision trees.

### C. THE PROPOSED APPROACH

The proposed method uses the RF machine learning algorithm to build the relationship between $PM_{2.5}$ and MODIS satellite images directly without retrieving the AOD. The following presents the details of each step.

#### 1) SAMPLE SELECTION

There are 102 $PM_{2.5}$ ground observation stations in the study area. The $PM_{2.5}$ datasets from these 102 stations are randomly assigned as training samples, validation samples, and test samples at a ratio of 4:4:3. First, we randomly select 32 stations as test samples and randomly divide the remaining stations into 40 training samples and 30 validation samples. Moreover, the influence of clouds is considered when generating the training samples (i.e., by overlapping a cloud mask on the MODIS dataset); if a cloud exists in a point over an observation station, the pixel encompassing this point in the MODIS image is not be selected as a training sample. The structure of the training data from the $i$-th observation station is as follows if there are no clouds over the station:

$$\langle x_1, x_2, \ldots, x_{16}, x_{17}, \ldots, x_{38}, x_{39}, \ldots x_{60}\rangle_i, y_i \qquad (1)$$

where $x_1$-$x_{16}$ represent the 16 emissivity wavebands, $x_{17}$-$x_{38}$ represent the 22 radiance wavebands, and $x_{39}$-$x_{60}$ represent the 22 reflectance wavebands. $y_i$ denotes the $PM_{2.5}$ ground truth data at the $i$-th observation station.

To improve the robust predictivity of the model, we enabled some enhanced pre-processing on the training data. The new training samples include not only the pixel corresponding to the observation station but also all pixels in the $5 * 5$ neighborhood of the station. Then, new training samples are created, including the pixels and their corresponding $PM_{2.5}$ ground truth data after interpolation. This enhancement is reasonable due to the first law of geography: ''all attributed values on a geographic surface are related to each other, but closer values are more strongly related than more distant ones'' [26]. Hence, the data within a $5 * 5$ neighborhood of an observation station have great confidence when treated as ground truth data. Thus, there are up to 25 training samples at one observation station which is not covered by clouds.

Otherwise, only pixels corresponding to an observation station are selected as the validation and test samples. Thus, there are up to 30 validation samples and 32 test samples when the condition that clouds don't exist is met.

#### 2) MODEL TRAINING

Generally, evenly distributed training samples are the most representative. In addition to the first randomly selected 32 test samples, the remaining 70 are randomly divided into 40 training samples and 30 validation samples. To enhance the even distribution of training samples, 150 different sets of training and validation samples are randomly generated. Then, based on 150 sets of training and evaluation samples, 150 training models are generated using the RF algorithm. The performance of each model is evaluated based on

the determination coefficient defined as R-squared ($R^2$) and refers to the ratio between the sum of squares of dependent variable $X$ and the sum of squares of independent variable $Y$. Thus, the value of $R^2$ determines the relative degree of closeness. When $R^2$ is close to 1, the determined equation has a higher reliability. Otherwise, when $R^2$ is closer to 0, the determined equation has lower liability. In this paper, $R^2$, root mean square error (RMSE) and correlation coefficient (CC) are computed by the following equations,

$$\begin{cases} R_j^2 = \dfrac{\sum_{i=1}^{N_j} (y'_{j,i} - \overline{y_j})^2}{\sum_{i=1}^{N_j} (y_{j,i} - \overline{y_j})^2} \\ RMSE_j = \sqrt{\sum_{i=1}^{N_j} \dfrac{(y'_{j,i} - y_{j,i})^2}{N_j - 1}} \\ CC_j = \dfrac{\sum_{i=1}^{N_j} (y'_{j,i} - \overline{y_j})(y_{j,i} - \overline{y_j})}{\sqrt{\sum_{i=1}^{N_j} (y'_{j,i} - \overline{y_j})^2} \sqrt{\sum_{i=1}^{N_j} (y_{j,i} - \overline{y_j})^2}} \end{cases} \tag{2}$$

where j represents the number of training sets (in this paper, j = 150), $N_j$ represents the number of valid validation samples in the $j$-th model, $y'_{j,i}$ and $y_{j,i}$ represent the predicted and ground truth values, respectively, $\overline{y_j}$ represents the average of the ground truth data.

Then, a histogram of all $R^2$ values in the 150 models is obtained using a 0.1 interval. Based on the interval with the highest frequency, all models corresponding to the $R^2$ values in this interval are selected to build an optimized multi-model group.

#### 3) MODEL TESTING

Every model in the previous multi-model group is used to test the samples, and the corresponding prediction is obtained. Then, the weighted integration of the prediction result for each model is applied based on its $R^2$ value using the following weighted process to obtain the final prediction values:

$$PM(i) = \sum_{k=1}^{n} w_k \times PM(i)_k = \sum_{k=1}^{n} \frac{R_k^2}{\sum_{p=1}^{n} R_p^2} \times PM(i)_k \tag{3}$$

where $n$ represents the number of models in the optimal multi-model group, $PM(i)_k$ represents the prediction value of the $i$-th test sample generated by the $k$-th model. $R_k^2$ refers to the $R^2$ value of the $k$-th model. The $R^2$ value, RMSE, CC and other indices are calculated to validate the performance of the trained model.

#### 4) MODEL APPLICATION

Every model in the previous multi-model group is used to identify the cloudless areas in the entire MODIS image, and the corresponding prediction is obtained using the similar weighted integration:

$$PM(i,j) = \sum_{k=1}^{n} w_k \times PM(i,j)_k = \sum_{k=1}^{n} \frac{R_k^2}{\sum_{p=1}^{n} R_p^2} \times PM(i,j)_k \tag{4}$$

where $n$ represents the number of models in the optimal multi-model group, $PM(i,j)_k$ represents the prediction value of a pixel located in $(i,j)$ generated by the $k$-th model. $R_k^2$ refers to the $R^2$ value of the $k$-th model. $PM$ represents the final retrieved result with the same spatial resolution as the original MODIS image (1 km), which is much higher than that of the AOD (3 km).

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

### A. DATASET

The MODIS datasets covering the study area are used to evaluate the proposed method for $PM_{2.5}$ retrieval. The dataset is collected from NASA (https://ladsweb.modaps.eosdis.nasa.gov/). The MODIS datasets provide 16 emissivity bands, 22 radiance bands and 22 reflectance bands. The compared data derive from the AOD products in the MODIS datasets with a 3 km resolution, which is accessible from http://modis-atmos.gsfc.nasa.gov/products.html and applies the C6.1 version product of the DT algorithms. Moreover, Weka [27] is used to implement the RF machine learning algorithm, which is available through the given link (https://www.cs.waikato.ac.nz/ml/weka/index.html).

Due to the characteristics of the atmosphere and environment in Guangdong Province, the MODIS datasets are generally covered by clouds, which results in missing data in the AOD products. Thus, this paper applies kriging interpolation to designate values for the missing data. Although a number of approaches have been proposed for AOD retrieval, these approaches generally import auxiliary data and specific operations to improve retrieval accuracy. To avoid the influences from other factors, we only use the AOD products at the highest spatial resolution (C6.1 version, 3 km) to retrieve $PM_{2.5}$ with a classic linear regression:

$$\mathrm{Y} = x_i^T \alpha + \beta_i, \quad \mathrm{i} = 1, 2, \ldots\ldots \mathrm{n} \tag{5}$$

where Y represents the dependent variable, $x_i$ represents the $i$-th explanatory variable, $\alpha$ is the estimating coefficient, and $\beta_i$ is a constant.

Since the study area is usually covered by clouds, we select data that have few clouds in 2015 and 2016. The experimental ground observational data for $PM_{2.5}$ derives from 102 environmental monitoring stations in Guangdong Province. Thirty-two stations are randomly selected for testing, and the remaining 70 stations are randomly divided into 40 training stations and 30 verification stations. At the same time, the coefficient of determination ($R^2$) and the root mean square error (RMSE) are used as evaluation indicators to compare the retrieval effects.

### B. SPATIAL-TEMPORAL MATCHING DURING PRE-PROCESSING FOR AOD-BASED RETRIEVAL

As mentioned previously, the satellite dataset used in this paper is MOD021KM, which provides 16 emissivity bands, 22 radiance bands and 22 reflectance bands at a 1 km resolution. The AOD dataset is the MODIS version C6.1 DT products at a 3 km resolution. The $PM_{2.5}$ ground truth datasets

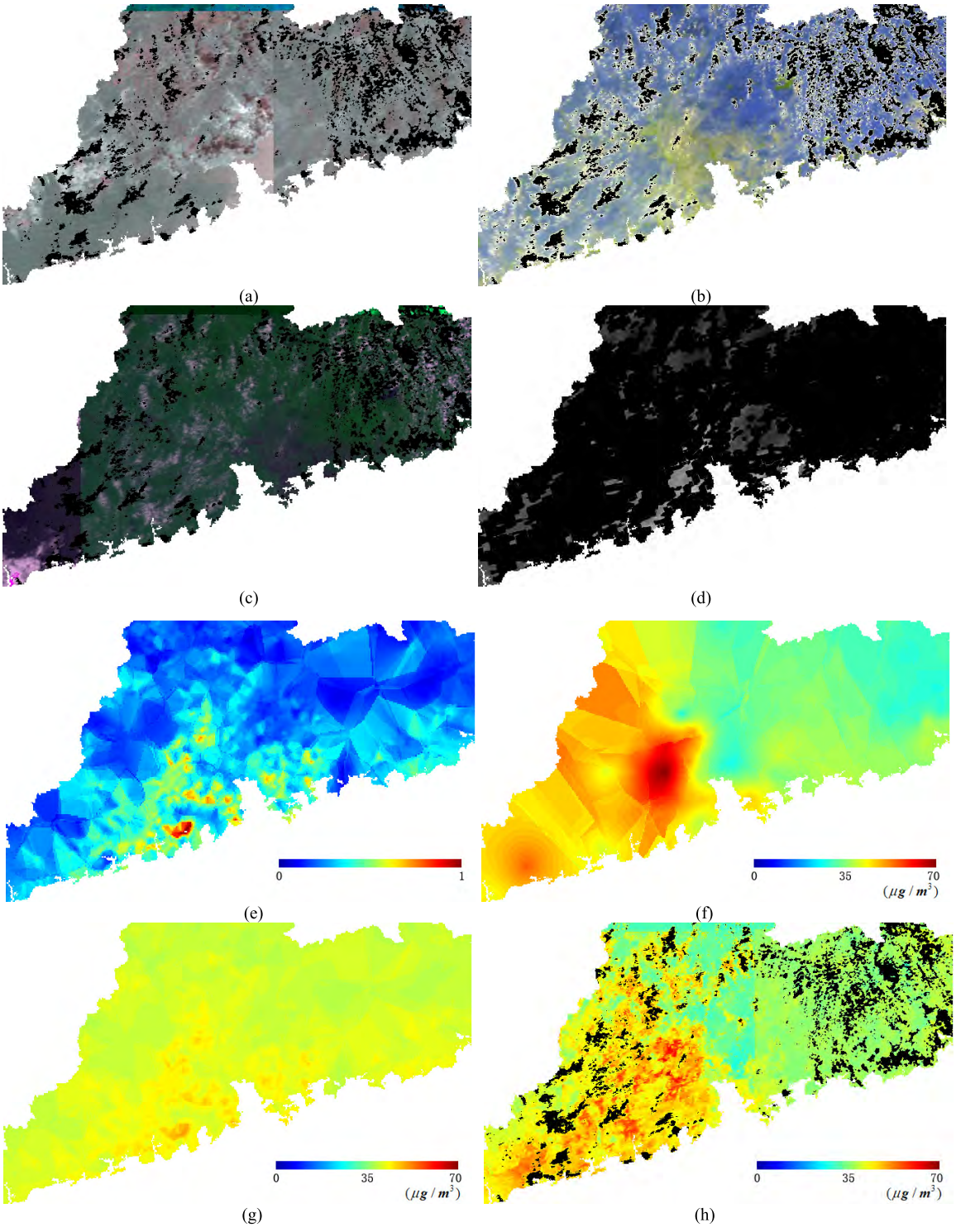


**FIGURE 2.** Experimental result on 2015.08.25. (a) Original emissivity data composed with 1, 2 and 3 wavebands. (b) Original reflectance data composed with 1, 2 and 3 wavebands. (c) Original radiance data composed with 4, 5 and 6 wavebands. (d) Original AOD data with lacking data in black. (e) AOD Kriging interpolation data. (f) $PM_{2.5}$ observational data without clouds. (g) $PM_{2.5}$ data retrieved by AOD. (h) $PM_{2.5}$ data retrieved by RF.

comprise hourly observation data from 102 observation stations.

The time when the MODIS Terra satellite starts to pass over the study area is 10:30 am. To unify the data generation time, we select the average value of the $PM_{2.5}$ observational data achieved from 10 am to 11 am as the ground truth data.

The spatial resolution of the AOD dataset and MODIS images are 3 km and 1 km, respectively. Thus, spatial

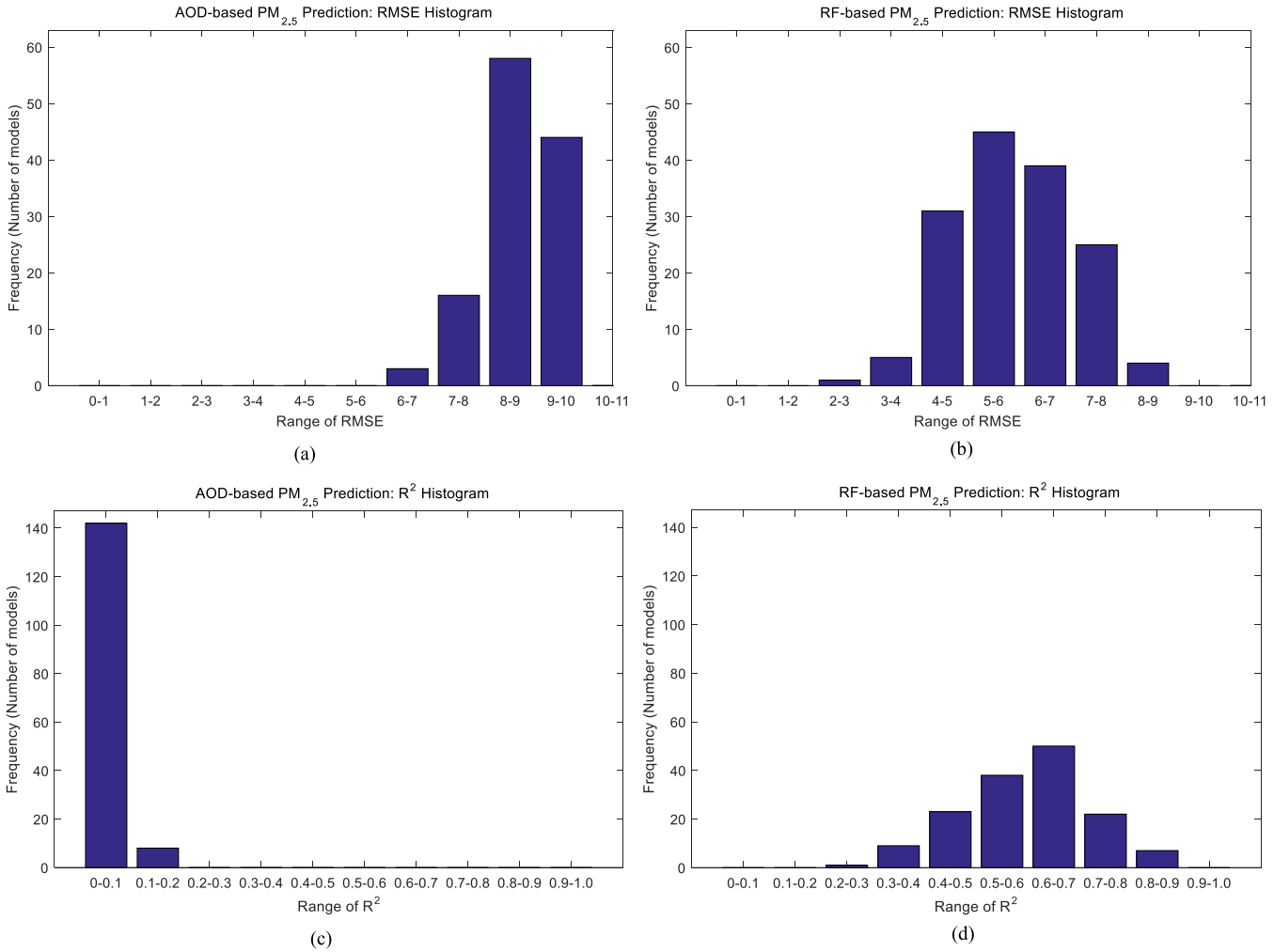


**FIGURE 3.** **Experimental results on 2015.08.25. (a) RMSE histogram distribution of AOD-based retrieval in training and validating procedure. (b) RMSE histogram distribution of RF-based retrieval in training and validating procedure. (c) $R^2$ histogram distribution of AOD-based retrieval in training and validating procedure. (d) $R^2$ histogram distribution of RF-based retrieval in training and validating procedure.**

matching of these two datasets is required based on geographical coordinates. To display the spatial distribution of the ground truth $PM_{2.5}$, we convert the $PM_{2.5}$ observational data from all 102 stations into a new dataset with a 1 km spatial resolution via kriging interpolation. Moreover, AOD datasets usually lack data due to the influence of clouds and other factors. Thus, we also use kriging interpolation to regenerate missing data.

### C. RETRIEVAL RESULTS AND COMPARISONS ON 2015.08.25

Figure 2 shows the MOD021KM data taken on 8/25/2015. The images visualize emissivity and radiance by combining wavebands No. 1, 2 and 3 and reflectivity by combining wavebands No. 4, 5 and 6. We use the cloud detection function in the MODIS dataset to create masks for the cloud pixels, which are shown as black regions in Figures 2(a), (b), (c) and (h). Moreover, there are many data lost in the AOD product. A large black block can be observed in Figure 2(d). Thus, we apply kriging to interpolate the original AOD dataset, and the interpolation results are shown as a pseudo-colored map in Figure 2(e), which also shows clear blocking effects. The original $PM_{2.5}$ dataset is the point data. We apply kriging interpolation to convert the point data into area data, which are shown in Figure 2(f). Figure 2(g) shows the $PM_{2.5}$ retrieval results generated by the linear regression on AOD data after the kriging interpolation, and Figure 2(h) shows the proposed result. The redder the color, the greater the concentration of $PM_{2.5}$; the bluer the color, the lower the concentration of $PM_{2.5}$.

From Fig. 2(f) we can see that the central and southwest areas have high $PM_{2.5}$ concentrations, and the east areas have lower $PM_{2.5}$ concentrations. The retrieval result based on the AOD data (Fig. 2(g)) is much different than that from the $PM_{2.5}$ ground observation data, which is mainly due to the missing data in the AOD dataset. Otherwise, the results generated by our proposed approach are similar to the ground observation data in general, with higher $PM_{2.5}$ concentrations

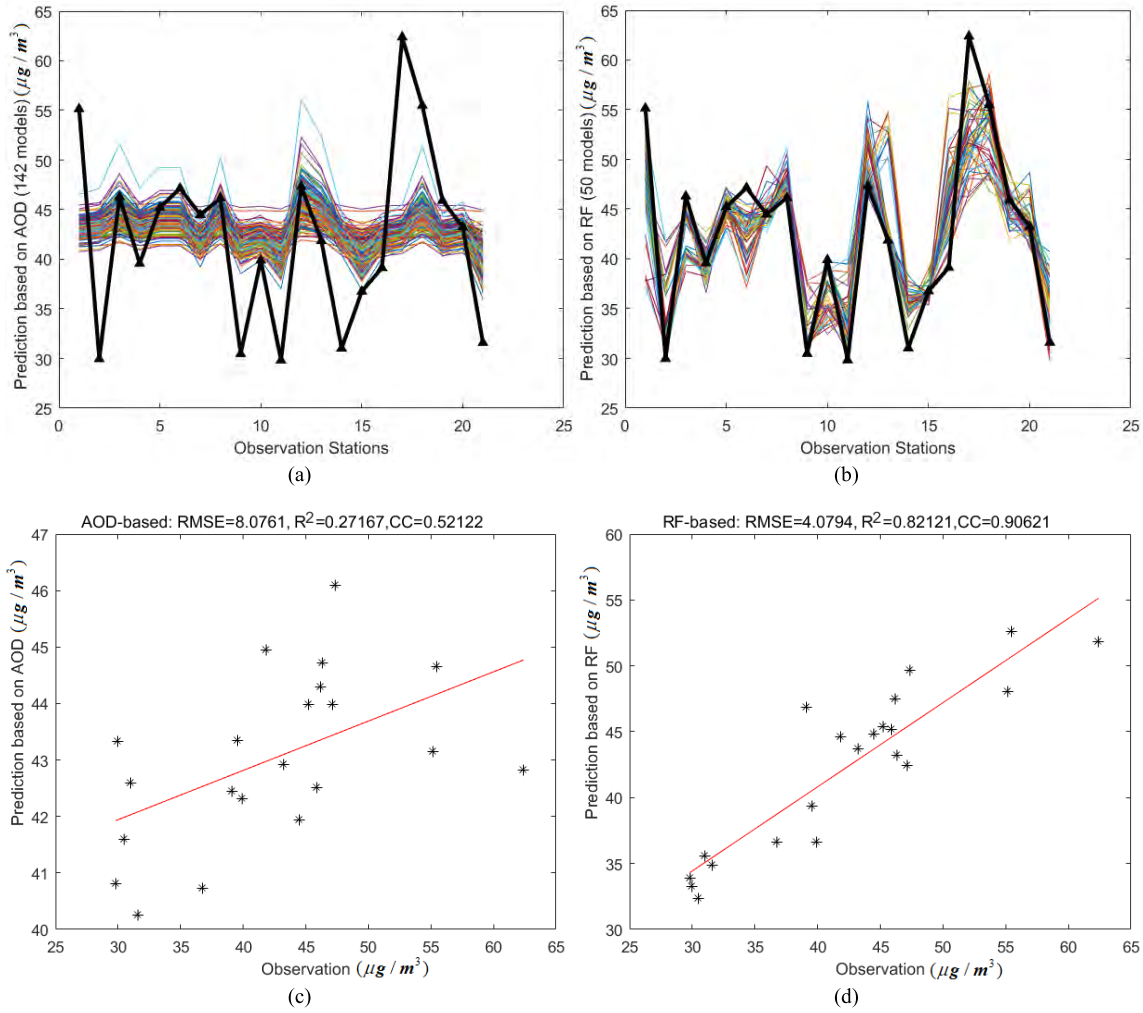


**FIGURE 4.** Experimental results on 2015.08.25. (a) Prediction results of AOD optimal multi-models group. (b) Prediction results of RF optimal multi-models group. (c) AOD-based prediction scatter and statistical result. (d) RF-based prediction scatter and statistical result.

in the central and southwest areas and lower $PM_{2.5}$ concentrations in the eastern areas.

Figure 3 shows a histogram distribution of the RMSE and $R^2$ for 150 models. The horizontal axis represents the range of values and the interval of the RMSE histogram (0-10 for range and 1 for interval), and those of the $R^2$ histogram (0-1 for range and 0.1 for interval), respectively. The vertical axis represents the frequencies of RMSE and $R^2$ located at specific intervals (i.e., the number of models).

From the histogram representing AOD retrieval, a majority of RMSE and $R^2$ values are distributed in ranges of 8-9 and 0-0.1, respectively. Otherwise, based on the histogram distribution of RF retrieval, a majority of RMSE and $R^2$ values are distributed in ranges of 5-6 and 0.6-0.7, respectively. This proves that the RF retrieval outperforms the AOD retrieval. Since the training and validation data for 150 models are randomly selected from 70 stations, when the observation stations are not distributed evenly, higher RMSE values and lower $R^2$ values might be observed. Thus, we select the models with the highest frequency as those in the optimized multi-model group. A total of 142 models are selected from the AOD model in the interval 0-0.1, and 50 models are selected from the RF model in the interval 0.6-0.7.

Due to the influence of cloud coverage, the data of only 21 test observation stations are available. Figure 4 shows the statistical results of 21 observation stations. Figures 4(a) and (b) show the prediction results of 21 observation stations generated by the AOD model and RF model, respectively. The thick black line denotes the ground truth values, and the colored thin lines refer to the results generated by the every selected AOD and RF model. Figures 4(c) and (d) shows the prediction results in a scatterplot, where the RMSE value, $R^2$ value and CC of the 21 observation stations are generated by the optimized AOD multi-model group and optimized RF multi-model group, respectively.

From Figs. 4(a) and (b), the prediction results generated by the AOD optimal multi-model group have significant differences from those of the ground truth data, while the

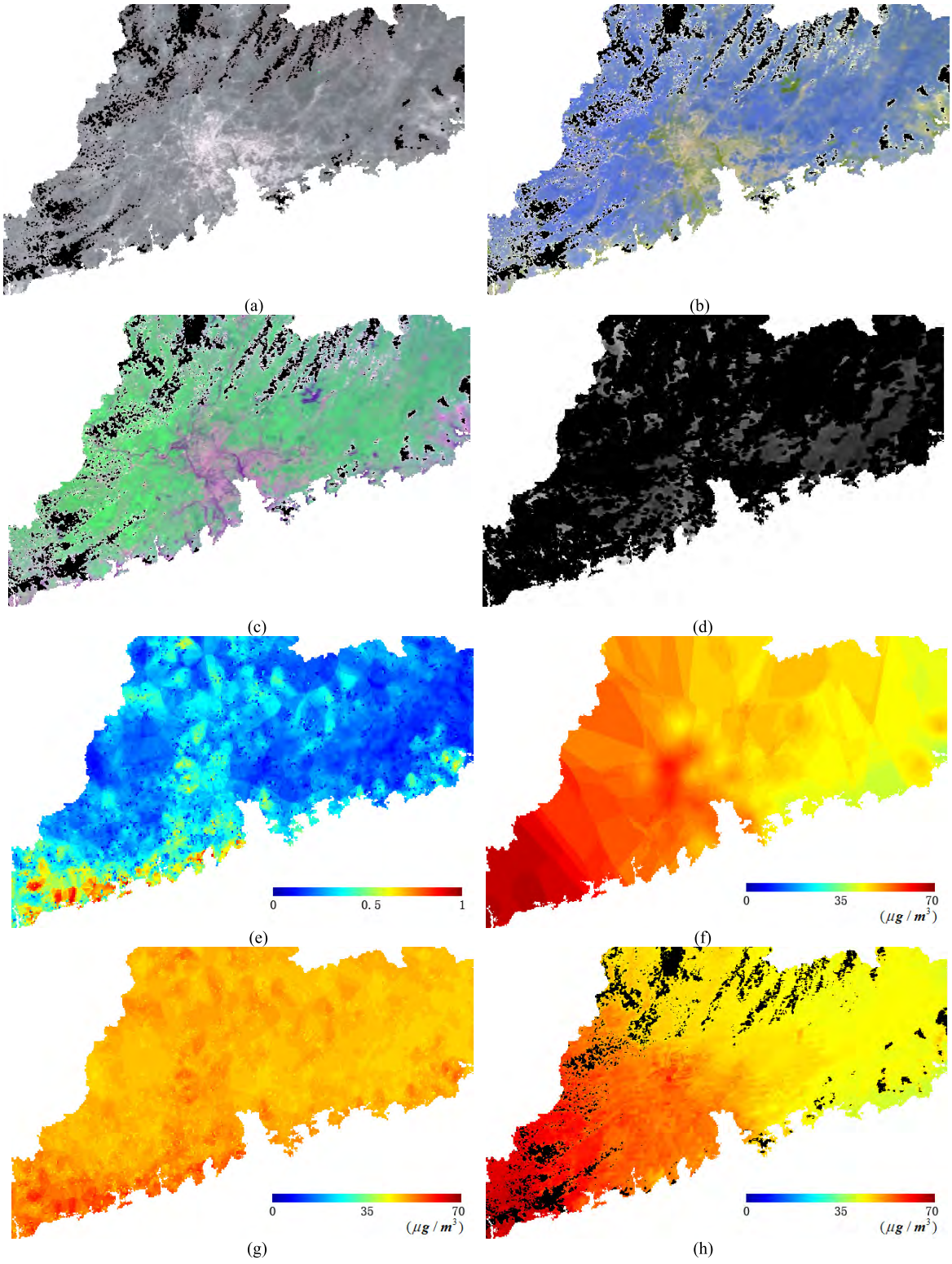


**FIGURE 5.** Experimental result on 2015.08.26. (a) Original emissivity data composed with 1, 2 and 3 wavebands. (b) Original reflectance data composed with 1, 2 and 3 wavebands. (c) Original radiance data composed with 4, 5 and 6 wavebands. (d) Original AOD data with lacking data in black. (e) AOD Kriging interpolation data. (f) $PM_{2.5}$ observational data without clouds. (g) $PM_{2.5}$ data retrieved by AOD. (h) $PM_{2.5}$ data retrieved by RF.

RF-based prediction results are similar to the ground truth data. From Figures 4(c) and (d), after the weighted integration, the RMSE, $R^2$ and CC by the RF-based approach are approximately 4, 0.82 and 0.9, respectively, which shows a strong correlation. These indicators prove that the RF-based approach outperforms the AOD-based method. Moreover, it

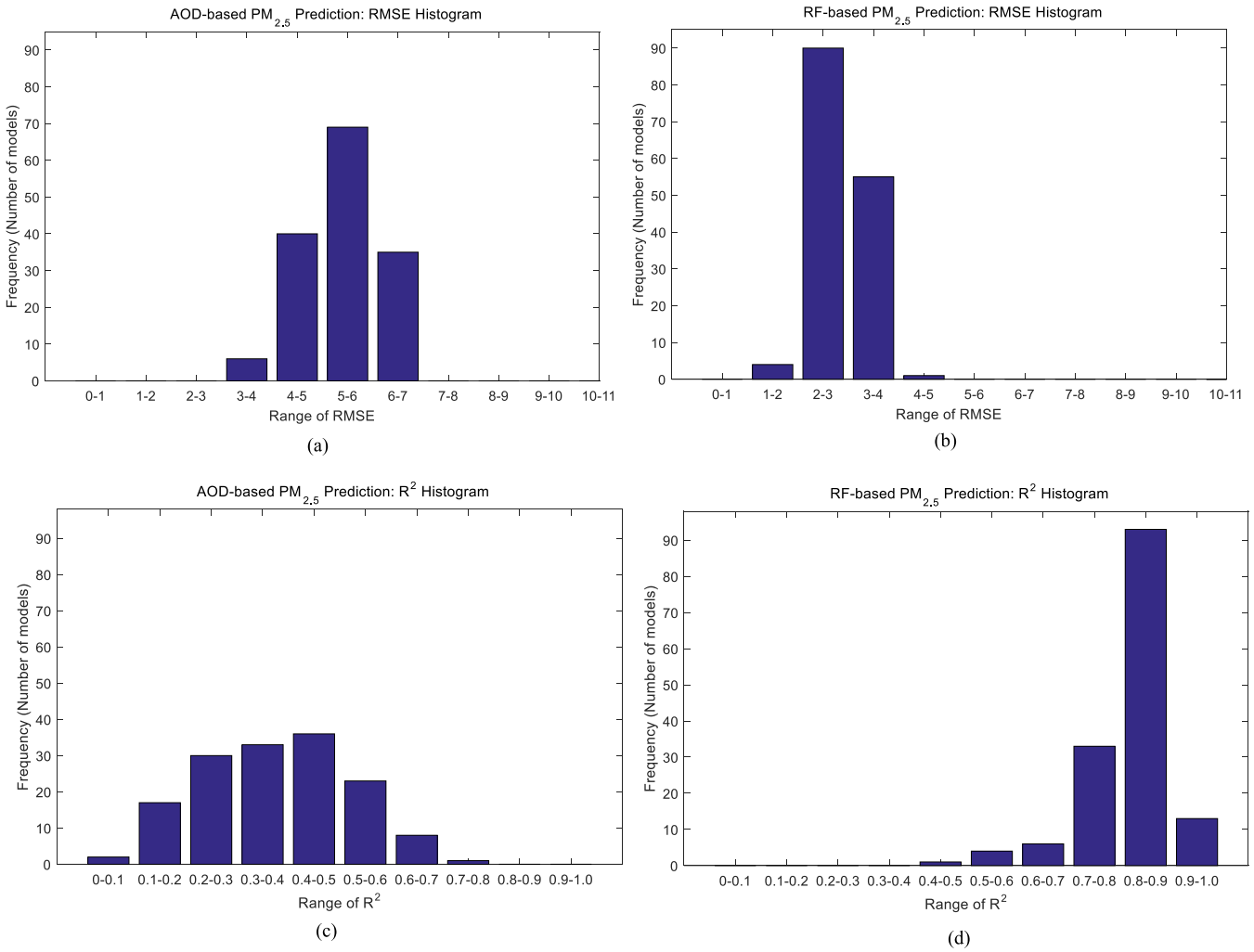


**FIGURE 6.** Experimental results on 2015.08.26. (a) RMSE histogram distribution of AOD-based retrieval in training and validating procedure. (b) RMSE histogram distribution of RF-based retrieval in training and validating procedure. (c) $R^2$ histogram distribution of AOD-based retrieval in training and validating procedure. (d) $R^2$ histogram distribution of RF-based retrieval in training and validating procedure.

shows that missing AOD data have a significantly negative influence on the $PM_{2.5}$ retrieval.

### D. RETRIEVAL RESULTS AND COMPARISONS ON 2015.08.26

Figure 5 shows the experimental results from another dataset with fewer clouds. The black areas represent masks that are provided by the cloud detection product in the MODIS dataset. Although the degree of missing AOD data is slightly smaller than the previous one, many missing data are still observed in Fig. 5(d), which also produces blocking effects after the kriging interpolation in Fig. 5(e). Fig. 5(f) represents the kriging interpolated ground observational values, and Figs. 5(g) and (h) show the retrieval results by the AOD method and our proposed method, respectively.

Based on the kriging interpolated ground observational results, it can be seen that the concentrations of $PM_{2.5}$ in the central and southwest regions are relatively large, and those in the east are relatively small. The results of the AOD retrieval are obviously large for the whole region, especially in the eastern region. However, the proposed RF-based method still shows obvious consistency with the ground observations. Because the clouds are mainly concentrated over mountains in the study area, 32 stations are randomly selected.

Figure 6 shows a histogram of the RMSE and $R^2$ for the validation samples in 150 models via the AOD and RF-based approaches. The representations of horizontal and vertical axis are the same as that of Fig.3. The RMSE is distributed across the 0-11 range at an interval of 1, and $R^2$ is distributed across the 0-1 range at an interval of 0.1.

Based on the RMSE results in Fig. 6, the RMSEs of the 150 AOD-based retrieval models are concentrated between 5-6, while those via the RF model are concentrate between 2-3. Based on the $R^2$ results, the AOD retrieval results are relatively concentrated from 0.4-0.5, and the RF-based results are clearly concentrated from 0.8-0.9. From the above results, it can be seen that most models via the RF-based method in this paper can obtain good prediction results regarding the validation set, while the AOD-based method has a relatively weak prediction capability.

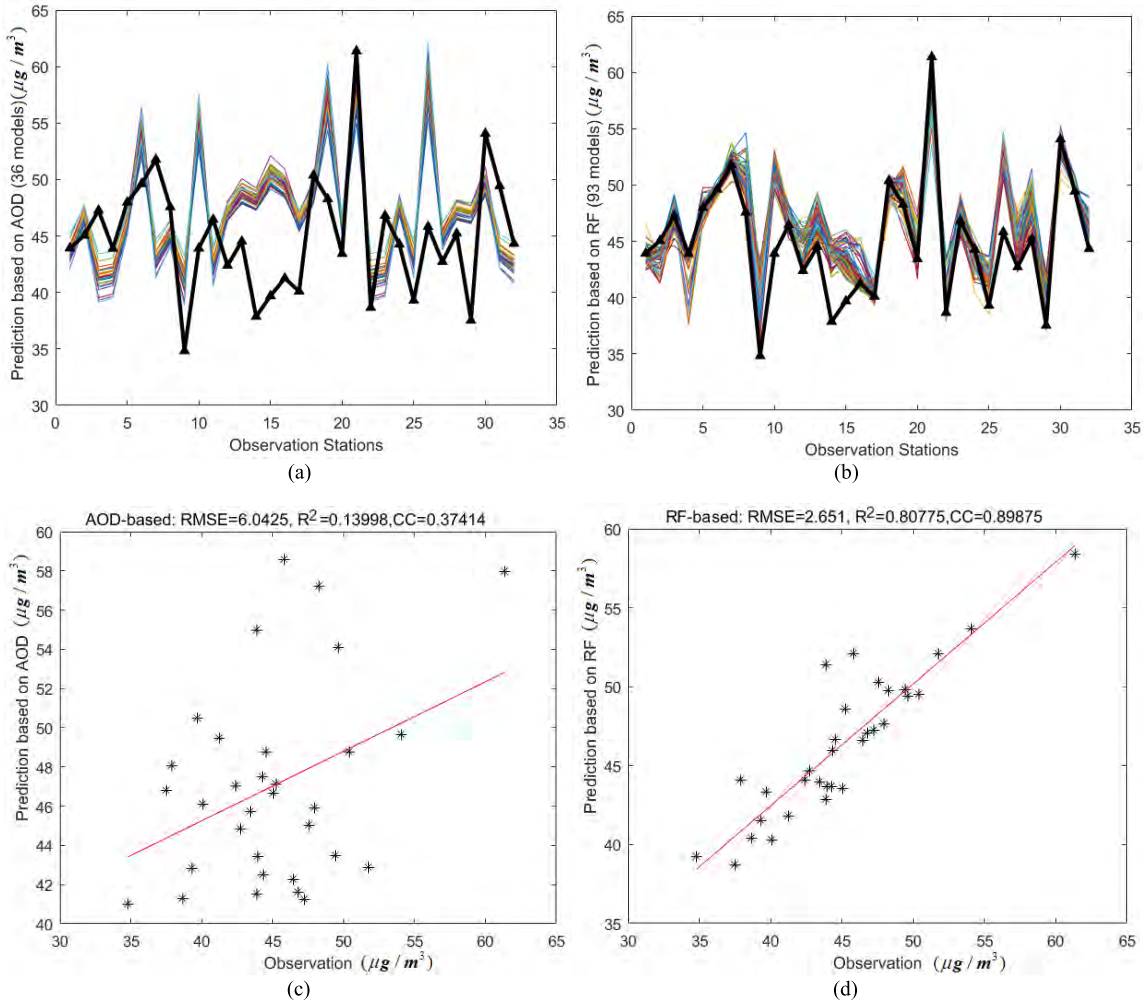


**FIGURE 7.** Experimental results on 2015.08.26. (a) Prediction results of AOD optimal multi-models group. (b) Prediction results of RF optimal multi-models group. (c) AOD-based prediction scatter and statistical result. (d) RF-based prediction scatter and statistical result.

Figure 7 shows the performances of the optimal multi-model group of the AOD and RF methods on the tested samples. The AOD optimal multi-model group has 36 members, while that of the RF has 93 members, indicating that the RF multi-model group has a better stability performance.

From Figs. 7(a) and (b), it can be seen that there are significant differences between the effects of the AOD-based prediction results and the ground observational data on the test samples. However, the trend in the RF-based method in this paper is consistent with that of the ground observational data, and there are only certain differences among a few stations. From the statistical indicators in the prediction results, the RMSE of the AOD-based method exceeds 6, the $R^2$ is only 0.14, and the CC is only 0.37. However, the statistical performance of the RF-based method in this paper is much better than that of the AOD-based method, as it shows a clear linear relationship.

### *E. MORE EXPERIMENTAL RESULTS*

This subsection provides the experimental results based on other datasets examined on 2015.04.15, 2015.04.17, 2015.08.08, 2015.10.15, 2015.10.17, 2015.12.20, 2016.02.6, 2016.02.9 and 2016.03.20, and data from each of the four seasons are included. Figure 8 only shows the experimental results of the datasets examined on 2015.10.15 and 2016.03.20.

From Fig. 8, the prediction results generated by the AOD optimal multi-model group are significantly different from those of the ground observational data, but the RF-based prediction results are similar to those of the ground observational data. Moreover, the results produced by our proposed method have a higher spatial resolution. Table 1 lists the RMSE and $R^2$ indices for the retrieval results from different datasets.

From the above table, we can see that the $R^2$ value of the proposed method is larger than that of the AOD method, and the RMSE is lower, indicating that the proposed method has obvious advantages. Cloud coverage and missing AOD data are two main reasons for the variations in $R^2$ and RMSE. There are a number of AOD data that are missing in the study area, and the data resulting from kriging interpolation cannot precisely represent the true distribution of AOD values.

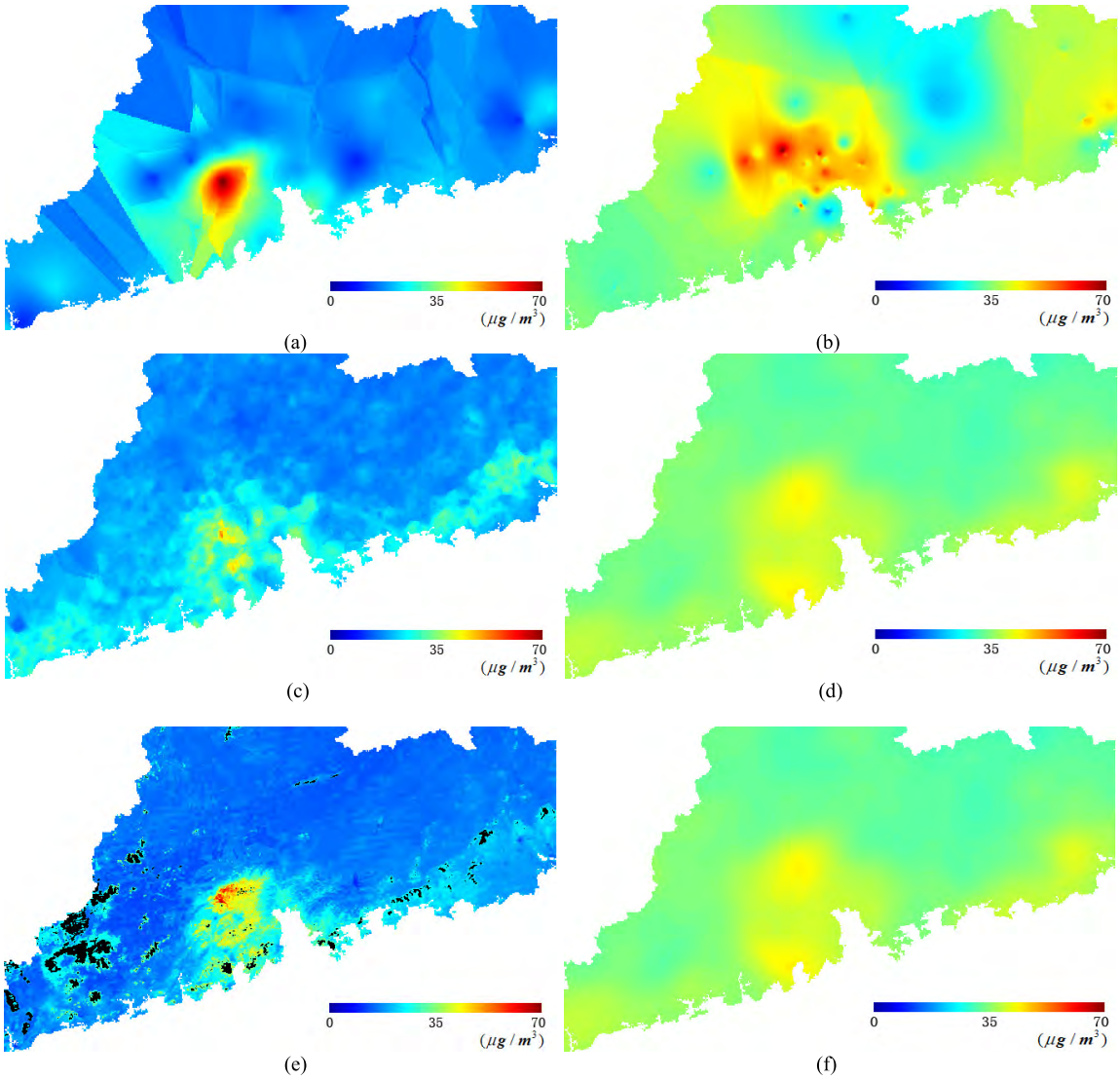


**FIGURE 8.** More retrieval results. (a)(c)(e) are the Kriging interpolated $PM_{2.5}$ ground observational data, AOD-based retrieval result and RF-based retrieval result on 2015.10.15. (b)(d)(f) are the Kriging interpolated $PM_{2.5}$ ground observational data, AOD-based retrieval result and RF-based retrieval result on 2016.3.20.

**TABLE 1.** Assessment results for more experiments.

| | $R^2$ | | RMSE | |
|---|---|---|---|---|
| | RF | AOD | RF | AOD |
| 2015.4.15 | 0.8792 | 0.1701 | 3.6193 | 7.4939 |
| 2015.4.17 | 0.8237 | 0.2113 | 3.9745 | 8.2316 |
| 2015.8.8 | 0.9013 | 0.1939 | 2.9452 | 4.7128 |
| 2015.10.15 | 0.8832 | 0.2338 | 3.1972 | 8.3832 |
| 2015.10.17 | 0.8721 | 0.2093 | 4.0138 | 7.9651 |
| 2015.12.20 | 0.8194 | 0.1973 | $_{2.5}$982 | 7.1395 |
| 2016.2.6 | 0.8891 | 0.2574 | 3.2867 | 6.8374 |
| 2016.2.9 | 0.8015 | 0.0916 | 3.6345 | 9.1374 |
| 2016.3.20 | 0.8284 | 0.2521 | 3.1134 | 7.7362 |

## V. DISCUSSION

This paper considers the error propagation process in the existing $PM_{2.5}$ retrieval method, and proposes a direct retrieval method of $PM_{2.5}$ based on random forest machine learning. It tries to skip the error propagation process and directly learn $PM_{2.5}$ characteristics from the original image to enable high-precision retrieval of $PM_{2.5}$. The experimental results show that the method is more accurate than the AOD product based retrieval. However, as a basic and simplest attempt, there are still some issues to be explored in this paper.

The ground $PM_{2.5}$ monitoring station is relatively rare. Therefore, in order to obtain the $PM_{2.5}$ concentration distribution on the surface, the Kriging interpolation technique is used. The $PM_{2.5}$ concentration after Kriging interpolation is not only used to establish the relationship between AOD and $PM_{2.5}$ in the conventional method, but also used in the enhanced pre-processing of the training samples in the proposed method. The choice of interpolation method will have a certain impact on the final retrieval results. At the same time, the terrain is also one of the factors affecting the $PM_{2.5}$ concentration distribution. In this paper, the interpolation process does not consider the terrain. For one reason,

the physical influence mechanism of terrain on $PM_{2.5}$ is not so clear [28]–[30], which may increase the uncertainty of $PM_{2.5}$ concentration distribution after interpolation; for another reason, it is limited by spatial resolution, i.e., at 3 km or 1 km spatial resolution, the terrain influence may not be as big. The goal of this paper is an attempt, so only the most conventional Kriging interpolation is used. In the later research, we will study the influence of terrain on $PM_{2.5}$ retrieval to further improve the retrieval accuracy.

As mentioned above, most of the existing $PM_{2.5}$ retrieval methods firstly use the atmosphere transmission model, accurately invert AOD, and then establish the relationship between AOD and $PM_{2.5}$. There are a lot of research results, introducing a large number of methods using physical or chemical principles, and various auxiliary data to improve the accuracy of AOD retrieval. In the process, introducing an additional data will increase uncertainty at the same time. These methods are based on certain reasonable physical or chemical assumptions, reflecting to some extent the true transmission process of the atmosphere. In fact, the atmosphere transmission process is very complex. There is currently no physical or chemical model which can accurately simulate the effects of various components of the atmosphere and various meteorological conditions on $PM_{2.5}$ distribution. Any approximation process is error-prone. In the imaging process of remote sensing images, the effects of various components of the surface, atmosphere and meteorological conditions are fixed in the spectral information of the pixels. Although the influence of each surface and atmospheric component cannot be accurately distinguished from the spectral information of a single pixel, from a holistic point of view, the spectral information of the pixel also reflects the influence of the surface and atmosphere to a certain extent, and the $PM_{2.5}$ of different concentrations is correspondingly reflected in the spectral information. Based on this assumption, this paper uses machine learning techniques to directly learn the relationship between surface and atmosphere and $PM_{2.5}$ from spectral information to attempt to retrieval $PM_{2.5}$. Although the physical or chemical mechanism of the method is not clear, the results show that the method is still feasible. At the same time, the use of spectral information directly in this paper is still too simple. Further pre-processing of the spectral information, for example, extracting more specific features, may be beneficial for improving $PM_{2.5}$ retrieval accuracy.

The atmospheric environment is complex and unpredictable, especially in the coastal areas of South China, where atmospheric convection is very frequent and atmospheric instability effect is obvious. Under this condition, trying to build a generic generalized model to retrieval the daily $PM_{2.5}$ concentration is obviously very difficult. Based on this consideration, the machine learning-based method proposed in this paper, for each day's $PM_{2.5}$ concentration retrieval, uses only the image of the day and the corresponding observed $PM_{2.5}$ concentration for training. The trained model is only valid for the day. The data of some observation stations are trained, and the model obtained by the training is used to generate the $PM_{2.5}$ concentration of other areas without the stations, thereby realizing the $PM_{2.5}$ concentration retrieval on a wide range. Through this kind of thinking, the changes in the daily atmospheric environment do not have much influence on the method proposed in this paper, because the daily training data will change with the changes of the atmospheric environment, and the trained model can also adapt to the daily atmospheric instability effect. However, finding a more generalized model is still the focus of future research. Using more powerful machine learning methods, for example, deep learning, to learn the laws of atmospheric change may be one of the ways to improve the generalization ability of the model.

## VI. CONCLUSION

Improvements on the MODIS AOD algorithm have produced more precise AOD products. However, these methods cannot avoid errors during the process of AOD retrieval. This paper integrates an ensemble RF machine learning algorithm and remote sensing images to directly build a relationship between remote sensing data and ground observation $PM_{2.5}$ data, which effectively reduces the error propagation. The experimental results at a 3 km resolution via the AOD products in MODIS and the $PM_{2.5}$ data from 102 observation stations in Guangdong Province indicate that our proposed approach can better produce $PM_{2.5}$ retrieval results. In addition, the spatial resolution of $PM_{2.5}$ retrieval result derived from AOD products in MODIS is 3 km, while the proposed result could achieve 1 km spatial resolution, which is the same as the spatial resolution of MODIS image itself. This work attempts to directly establish the relationship between satellite images and ground truth $PM_{2.5}$ using ensemble machine learning technology, and the experimental results have proven the effectiveness of the proposed approach, although the proposed method is simple. Therefore, future work will focus on improving the robustness of our proposed approach by extending the size of the study area and using more datasets. The integration of other machine learning methods to build new retrieval models will also be a key point in future works.

## REFERENCES

[1] C. Lin, Y. Li, Z. Yuan, A. K. H. Lau, C. Li, and J. C. H. Fung, "Using satellite remote sensing data to estimate the high-resolution distribution of ground-level $PM_{2.5}$," *Remote Sens. Environ.*, vol. 156, pp. 117–128, Jan. 2015.

[2] W. Gong, Y. Huang, T. Zhang, Z. Zhu, Y. Ji, and H. Xiang, "Impact and suggestion of column-to-surface vertical correction scheme on the relationship between satellite AOD and ground-level $PM_{2.5}$ in china," *Remote Sens.*, vol. 9, no. 10, p. 1038, 2017.

[3] Z. Chen *et al.*, "Examining the influence of crop residue burning on local $PM_{2.5}$ concentrations in heilongjiang province using ground observation and remote sensing data," *Remote Sens.*, vol. 9, no. 10, p. 971, 2017.

[4] J. Tian and D. Chen, "A semi-empirical model for predicting hourly ground-level fine particulate matter $PM_{2.5}$ concentration in southern Ontario from satellite remote sensing and ground-based meteorological measurements," *Remote Sens. Environ.*, vol. 114, pp. 221–229, Feb. 2010.

[5] H. J. Lee, Y. Liu, B. A. Coull, J. Schwartz, and P. Koutrakis, "A novel calibration approach of MODIS AOD data to predict $PM_{2.5}$ concentrations," *Atmos. Chem. Phys.*, vol. 11, no. 15, pp. 1–12, 2011.

[6] Y. Liu, J. Sarnat, V. Kilaru, D. Jacob, and P. Koutrakis, "Estimating ground-level $PM_{2.5}$ in the eastern United States using satellite remote sensing," *Environ. Sci. Technol.*, vol. 39, no. 9, pp. 3269–3278, May 2005.

[7] A. van Donkelaar, R. V. Martin, and R. J. Park, "Estimating ground-level $PM_{2.5}$ using aerosol optical depth determined from satellite remote sensing," *J. Geophys. Res. Atmos.*, vol. 111, pp. 1–10, Nov. 2006.

[8] W. Song, H. Jia, J. Huang, and Y. Zhanga, "A satellite-based geographically weighted regression model for regional $PM_{2.5}$, estimation over the pearl river delta region in China," *Remote Sens. Environ.*, vol. 154, pp. 1–7, Nov. 2014.

[9] Z. Li *et al.*, "Aerosol physical and chemical properties retrieved from ground-based remote sensing measurements during heavy haze days in beijing winter," *Atmos. Chem. Phys.*, vol. 13, pp. 10171–10183, Oct. 2013.

[10] D. A. Chu *et al.*, "Interpreting aerosol lider profiles to better estimate surface $PM_{2.5}$, for columnar AOD measurements," *Atmos. Environ.*, vol. 79, no. 11, pp. 172–187, Nov. 2013.

[11] M. Griggs, "Measurements of atmospheric aerosol optical thickness over water using ERTS-1 data," *J. Air Pollut. Control Assoc.*, vol. 25, no. 6, pp. 622–626, 1975.

[12] C. Rao, L. Stowe, E. Clain, J. Snapper, "Development and application of aerosols remote sensing with AVHRR data from the NOAA satellites," in *Aerosols and Climate*, P. V. Hobbs and M. P. McCormickedited, Eds. Hampton, VA, USA: Deepak, 1988, pp. 69–79.

[13] R. C. Levy, L. A. Remer, S. Mattoo, E. F. Vermote, and Y. J. Kaufman, "Second-generation operational algorithm: Retrieval of aerosol properties over land from inversion of moderate resolution imaging spectroradiometer spectral reflectance," *J. Geophys. Res. Atmos.*, vol. 112, no. D13, 2007, Art. no. D13211.

[14] R. C. Levy, L. A. Munchak, S. Mattoo, F. Patadia, L. A. Remer, and R. E. Holz, "Towards a long-term global aerosol optical depth record: Applying a consistent aerosol retrieval algorithm to MODIS and VIIRS-observed reflectance," *Atmos. Meas. Tech.*, vol. 10, no. 8, pp. 4083–4110, 2015.

[15] B. Holben, E. Vermote, Y. J. Kaufman, D. Tanre, and V. Kalb, "Aerosol retrieval over land from AVHRR data-application for atmospheric correction," *IEEE Trans. Geosci. Remote Sens.*, vol. 30, no. 2, pp. 212–222, Mar. 1992.

[16] V. Y. Isakov, R. E. Feind, O. B. Vasilyev, and R. M. Welch, "Retrieval of aerosol spectral optical thickness from AVIRIS data," *Int. J. Remote Sens.*, vol. 17, no. 11, pp. 2165–2184, 1996.

[17] Y. Shi, T. Matsunaga, Y. Yamaguchi, Z. Li, X. Gu, and X. Chen, "Long-term trends and sptial patterns of satellite-retrieved $PM_{2.5}$ concentrations in south and southeast Asia from 1999 to 2014," *Sci. Total Environ.*, vol. 615, pp. 177–186, Feb. 2018.

[18] C.-R. Jung, B.-F. Hwang, and W.-T. Chen, "Incorporating long-term satellite-based aerosol optical depth, localized land use data, and meteorological variables to estimate ground-level $PM_{2.5}$ concentrations in Taiwan from 2005 to 2015," *Environ. Pollut.*, vol. 237, pp. 1000–1010, Jun. 2018.

[19] Z. Shao, L. Zhang, and L. Wang, "Stacked sparse autoencoder modeling using the synergy of airborne LiDAR and satellite optical and SAR data to map forest above-ground biomass," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 12, pp. 5569–5582, Dec. 2017.

[20] Z. Shao, L. Zhang, X. Zhou, and L. Ding, "A novel hierarchical semisupervised SVM for classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 9, pp. 1609–1613, 2014.

[21] J. Liu *et al.*, "Semantic classification for hyperspectral image by integrating distance measurement and relevance vector machine," *MultiMedia Syst.*, vol. 23, no. 1, pp. 95–104, Feb. 2017.

[22] H. Fu *et al.*, "Cloud detection for FY meteorology satellite based on ensemble thresholds and random forests approach," *Remote Sens.*, vol. 11, no. 1, p. 44, 2019.

[23] Z. Chunhui, G. Bing, Z. Lejun, and W. Xiaoqing, "Classification of Hyperspectral Imagery based on spectral gradient, SVM and spatial random forest," *Infr. Phys. Technol.*, vol. 95, pp. 61–69, Dec. 2018.

[24] C. R. Levy *et al.*, "The collection 6 MODIS aerosol products over land and ocean," *Atmos. Meas. Tech.*, vol. 6, pp. 2989–3034, Nov. 2013.

[25] C. Hsu *et al.*, "Enhanced deep blue aerosol retrieval algorithm: The second generation," *J. Geophys. Res. Atmos.*, vol. 118, no. 16, pp. 9296–9315, 2013.

[26] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Econ. Geography*, vol. 46, pp. 234–240, Jun. 1970.

[27] F. Eibe, A. Mark, and H. Ian, *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington, MA, USA: Morgan Kaufmann, 2016.

[28] W. Xiaoyan, R. E. Dickinson, L. Su, C. Zhou, and K. Wang, "$PM_{2.5}$ pollution in china and how it has been exacerbated by terrain and meteorological conditions," *Bull. Amer. Meteorol. Soc.*, vol. 99, no. 1, pp. 105–120, Jan. 2018.

[29] B. Tunno *et al.*, "Spatial patterning in $PM_{2.5}$ constituents under an inversion-focused sampling design across an urban area of complex terrain," *J. Expo. Sci. Environ. Epidemiol.*, vol. 26, no. 4, pp. 385–396, 2016.

[30] P. E. Saide *et al.* "Forecasting urban PM10 and PM2.5 pollution episodes in very stable nocturnal conditions and complex terrain using WRF-Chem CO tracer model," *Atmos. Environ.*, vol. 45, no. 16, pp. 2769–2780, May 2011.


**XUNLAI CHEN** received the Ph.D. degree from Sun Yat-sen University, in 2007, and the B.S. degree from Nanjing University, in 2002. He was a Senior Engineer of meteorology with the Shenzhen Meteorological Bureau. His research interests include numerical weather prediction, nowcasting technique, and numerical simulation of air pollution.

**HUI LI** received the M.S. degree from The Hongkong University of Science and Technology, in 2005, and the B.S. degree from the Nanjing Institute of Meteorology, in 2000. She was an Engineer of meteorology with the Shenzhen Meteorological Bureau. Her research interests include typhoon forecast technique and post-processing of numerical model.

**SHUTING ZHANG** received the master's degree in atmospheric physics and atmospheric environment from Sun Yat-sen University, in 2016. She was an Assistant Engineer and Weather Forecaster with Shenzhen Meteorological Bureau. Her research interests include numerical weather prediction, air pollution nowcasting technique, and numerical simulation.

**YUANZHAO CHEN** received the degree from Chengdu Meteorological College, in 1994. He was a Professorate Senior Engineer, Chief Forecaster of the Shenzhen Meteorological Bureau. His research interests include short-term nowcasting and technology development.

**QI FAN** received the B.S. and Ph.D. degrees from Sun Yat-sen University, in 1998 and 2003, respectively. She was a Professor with the Department of Atmospheric Science, Sun Yat-sen University. Her research interests include air pollution modeling and aerosol chemistry.