Estimating PM_{2.5} Concentrations in Contiguous Eastern Coastal Zone of China Using MODIS AOD and a Two-Stage Random Forest Model

LIJUAN YANG,^a HANQIU XU,^{b,c,d,e} AND SHAODE YU^{f,g}

^a Department of Surveying and Mapping Engineering, Minjiang University, Fuzhou, China

^b College of Environment and Safety Engineering, Fuzhou University, Fuzhou, China

^c Institute of Remote Sensing Information Engineering, Fuzhou University, Fuzhou, China

^d Key Laboratory of Spatial Data Mining and Information Sharing, Ministry of Education, Fuzhou University, Fuzhou, China

^e Fujian Provincial Key Laboratory of Remote Sensing of Soil Erosion, Fuzhou University, Fuzhou, China

^f College of Information and Communication Engineering, Communication University of China, Beijing, China

^g Key Laboratory of Convergent Media and Intelligent Technology, Communication University of China, Ministry of Education, Beijing, China

(Manuscript received 5 January 2021, in final form 13 October 2021)

ABSTRACT: The coarse Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol optical depth (AOD) product (spatial resolution: 3 km) retrieved by the dark-target algorithm always generates the missing values when being adopted to estimate the ground-level PM_{2.5} concentrations. In this study, we developed a two-stage random forest using MODIS 3-km AOD to obtain the PM_{2.5} concentrations with full coverage in a contiguous coastal developed region, i.e., Yangtze River delta–Fujian–Pearl River delta (YRD–FJ–PRD) region of China. A first-stage random forest–integrated six meteorological fields was employed to predict the missing values of AOD product, and the combined AOD (i.e., random forest–derived AOD and MODIS 3-km AOD) incorporated with other ancillary variables were developed for predicting PM_{2.5} concentrations within a second-stage random forest model. The results showed that the first-stage random forest could explain 94% of the AOD variability over YRD–FJ–PRD region, and we achieved a site-based cross validation (CV) R^2 of 0.87 and a time-based CV R^2 of 0.85. The full-coverage PM_{2.5} concentrations illustrated a spatial pattern with annual-mean PM_{2.5} of 46, 40, and 35 $\mu\text{g m}^{-3}$ in YRD, PRD, and FJ, respectively, sharing the same trend with previous studies. Our results indicated that the proposed two-stage random forest model could be effectively used for PM_{2.5} estimation in different areas.

KEYWORDS: Model evaluation/performance; Nonlinear models; Air pollution; Air quality; Software; Machine learning

1. Introduction

During the past decades, environmental issues including air pollution, soil erosion, and land desertification have aroused the strong attention of the scientists and the public in China (Laurent et al. 2005; Dai et al. 2018; Li et al. 2019). The atmospheric environment has been deteriorating with high concentrations of fine particulate matters suspended in the air due to the extreme urban expansion. As the main pollutant of haze in most of regions, PM_{2.5} (i.e., fine particles with aerodynamic diameter less than 2.5 μm) has been reported to associate with adverse human health, such as asthma (Prunicki et al. 2018; Wang et al. 2019) and mortality (Rojas-Rueda et al. 2019). It is necessary and of great significance to carry out the PM_{2.5} estimation in the mainland cities of China. The traditional ground measurement method using PM_{2.5} monitoring network can provide the accurate spatial and temporal data;

however, it cannot be used for deriving PM_{2.5} concentrations where the monitors are unavailable (Liu et al. 2005; Yang et al. 2018). The aerosol optical depth (AOD) derived from satellites [e.g., Moderate Resolution Imaging Spectroradiometer (MODIS), Multi-Angle Implementation of Atmospheric Correction (MAIAC)] have been widely utilized for predicting PM_{2.5} concentrations in different scales (Zeng et al. 2018; Goldberg et al. 2019; Mhawish et al. 2019; Sathe et al. 2019).

Most previous studies have demonstrated that AOD are nonlinearly correlated with PM_{2.5} concentrations, as the AOD reflects the integral of the extinction coefficient of the total atmospheric column (Gupta et al. 2006; Wang et al. 2010; Yang et al. 2018). The advanced statistical models have then been developed to improve the predictability of PM_{2.5} concentrations by introducing meteorological parameters and land-use variables. For example, Hu et al. (2013) employed a geographically weighted regression model (GWR) along with MODIS AOD and meteorological parameters for PM_{2.5} estimation in the United States, and the GWR-derived PM_{2.5} presented a high correlation with the measured PM_{2.5} ($R^2 = 0.82$). He and Huang (2018) expanded the GWR model by considering the temporal variation of AOD–PM_{2.5} relationship, and proposed a geographically and temporally weighted regression (GTWR) model consisting of five meteorological parameters to predict the surface-level PM_{2.5} in mainland China. Their results showed that the R^2 derived from GTWR model reached 0.80. Yang et al. (2018) proposed a linear

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JTECH-D-20-0214.s1>.

Corresponding author: Hanqiu Xu, hxu@fzu.edu.cn

DOI: 10.1175/JTECH-D-20-0214.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

FIG. 1. Location of the YRD-FJ-PRD region.

mixed effects (LME) model incorporating planetary boundary layer height, relative humidity, and road density to estimate the $PM_{2.5}$ in Fuzhou city of China, and an overall R^2 of 0.81 was achieved. These advanced models have also been demonstrated to exhibit high predictability in $PM_{2.5}$ estimation in large to global scale (Ma et al. 2016; Yang et al. 2020).

Machine learning algorithms incorporated with multiple variables were also employed for $PM_{2.5}$ estimation in different scales. For example, Gupta and Christopher (2009) used an artificial neural network (ANN) framework to reduce the error of $PM_{2.5}$ estimation introduced by satellite AOD, and the ANN-derived $PM_{2.5}$ showed 61% agreement with the observed values. Mehdi-pour et al. (2018) compared three algorithms including decision tree (DT), Bayesian network (BN), and support vector machine (SVM) for $PM_{2.5}$ estimation in Tehran, Iran, and they found that SVM performed the best among the three. Mirroring advanced statistical models, the autocorrelation between independent variables should be validated before model fitting, and the results of above-mentioned machine learning algorithms are hard to interpret, as the importance of each variable could not be presented in most of statistical models, which would limit their applications in other study areas. Random forest, on the contrary, is an ensemble method which introduces the bagging and bootstrap resampling algorithm to improve the model accuracy and has already been demonstrated as a significant method for $PM_{2.5}$ predictions in North America (Hu et al. 2017; Brokamp et al. 2018).

In this study, we examined the random forest for predicting $PM_{2.5}$ concentrations in the contiguous eastern coastal region of China, i.e., Yangtze River delta–Fujian Province–Pearl River delta (hereinafter referred as YRD-FJ-PRD region), by using MODIS AOD with spatial resolution of 3 km (i.e., MODIS 3-km AOD). It is noted that MODIS 3-km AOD is unavailable in urban areas with high reflectance (e.g., building areas), since it is retrieved by using dark target algorithm.

Although our previous research has demonstrated that the top-of-atmosphere (TOA) reflectance could be used for $PM_{2.5}$ estimation in YRD region, the AOD is still the most important indicator for atmospheric pollution in a region (Yang et al. 2020). To this end, we developed a first-stage random forest to achieve full-covered AOD distribution by using six meteorological parameters. A second-stage random forest was further employed for $PM_{2.5}$ estimation by using integrated AOD (i.e., random forest-derived AOD and MODIS 3-km AOD) and the meteorological and landform parameters. As far as we are concerned, this is the first study that developed a two-stage random forest using integrated AOD and the auxiliary variables to achieve the $PM_{2.5}$ concentrations with full coverage in a contiguous coastal zone of China.

2. Materials and methods

a. Study area

The contiguous eastern region of China (20.19°–35.17°N, 109.62°–122.97°E) consisting of 55 adjoining cities which are mainly distributed in Jiangsu, Zhejiang, Fujian, and Guangdong Province (i.e., YRD-FJ-PRD) was defined as our study case (Fig. 1). The YRD-FJ-PRD region has a wide range of altitudes (0–2153 m) and represents two different levels of air pollution in China. The air quality report released by the Department of Environmental Protection of China showed that YRD and PRD region are the most air-polluted developed urban agglomerations in China, which has also been validated in previous studies (Li et al. 2015; Li et al. 2018; Yang et al. 2020). Fujian Province, located between YRD to the north and PRD to the south, is a developed eastern coastal area predominated by mountainous and hilly regions. The long-term annual-mean $PM_{2.5}$ concentrations were much lower in Fujian Province, compared with YRD and PRD.

TABLE 1. Summary of the meteorological fields used in this study.

Field	Temporal resolution	Variables	Abbreviation	Units	Used in model	Vertical structure
A1 ^a	1-h time average	Planetary boundary layer height	PBLH	m	First stage	Horizontal only
		2-m air temperature	T2M	K		
		10-m air temperature	T10M	K		
		10-m eastward wind	U10M	K		
		10-m northward wind	V10M	m s ⁻¹		
		Surface skin temperature	TS	m s ⁻¹	Second stage	
A3dyn ^b	3-h time average	Relative humidity	RH	%	First stage	Surface pressure ^c
		Eastward component of wind	E component	m s ⁻¹	Second stage	
		Northward component of wind	N component	m s ⁻¹	Second stage	
I3 ^d	3-h instantaneous	Surface pressure	PS	hPa	Second stage	Surface pressure

^a A1 means the time-averaged 1-hourly, two-dimensional, single-level variables on the native resolution grid.

^b A3dyn field means the time-averaged 3-hourly, three-dimensional, assimilated state variables on the native level.

^c Surface pressure defines the values of the A3dyn and I3 fields near the surface.

^d I3 field means the instantaneous, three-dimensional, assimilated state variables on the native level.

b. Modeling data

1) PM_{2.5} OBSERVATIONS AND SATELLITE DATA

The PM_{2.5} observations for 2018 (i.e., 1 January to 31 December 2018) from 297 monitoring stations in 55 cities of YRD-FJ-PRD were collected from the website of Provincial Environmental Protection Department (<http://106.37.208.233:20035/>). The PM_{2.5} concentrations with less than 0 $\mu\text{g m}^{-3}$ (i.e., recorded as “damaged”) were removed before data integration.

The MODIS 3-km AOD product covering YRD-FJ-PRD region (from 1 January to 31 December 2018) was collected from NASA’s *Terra* satellite (<https://ladsweb.modaps.eosdis.nasa.gov/search/>). We extracted the AOD at 0.55- μm wavelength from MODIS 3-km AOD product by using IDL for data modeling in this study.

2) METEOROLOGICAL FIELDS

The meteorological fields from Goddard Earth Observing System Data Assimilation System GEOS-5 Forward Processing (GEOS-5 FP) were downloaded in this paper (<http://rain.ucis.dal.ca/>). There are three different fields including A1, A3dyn, and I3 with spatial resolution of $0.25^\circ \times 0.3125^\circ$ in GEOS-5 FP product. The temporal resolutions are 1 h time average, 3-h time average, and 3-h instantaneous for A1, A3dyn, and I3 field, respectively. Ten parameters, i.e., planetary boundary layer height (PBLH), surface skin temperature (TS), 2-m air temperature (T2M), 10-m air temperature (T10M), 10-m eastward wind (U10M), 10-m northward wind (V10M), relative humidity (RH), eastward component of wind (E component), northward component of wind (N component), and surface pressure (PS) were used in this study. Details of the meteorological fields are shown in Table 1.

3) LAND-USE DATA

We downloaded the MODIS monthly vegetation index (MVI) representing vegetation cover from NASA’s MOD13A3 product in this study (<https://ladsweb.modaps.eosdis.nasa.gov/>). The spatial resolution of MVI is 1 km. We also obtained the

shape file of major roadways (e.g., national highway and express way) from Baidu StreetMap for YRD-FJ-PRD.

4) DATA INTEGRATION

A $3 \text{ km} \times 3 \text{ km}$ grid covering YRD-FJ-PRD region was created for data integration. The hourly PM_{2.5} concentrations were averaged for each monitoring station. The meteorological parameters and the MVI were resampled into $3 \text{ km} \times 3 \text{ km}$ grid cell by utilizing nearest neighbor interpolation algorithm. We defined the ratio of the total length of each $3 \text{ km} \times 3 \text{ km}$ grid cell to 9 km^2 as the road density, i.e., the density was proportional to the total length of roads on each grid cell. All the independent variables were finally extracted according to the geolocation of 297 monitoring stations. The data integration was processed in Python and IDL.

c. Model development and validation

Random forests regression is an ensemble machine learning method which has widely been used in the fields of medical science and biology due to its high accuracy and ability to handle large features with small samples (L.-L. Zhao et al. 2019; Grinberg et al. 2020). It uses two concepts of bootstrap resampling and random subset feature selection to reduce the correlation between the regression trees (Khosravi and Alavi-panah 2019). In this study, we developed a two-stage random forest to estimate the PM_{2.5} concentrations in YRD-FJ-PRD region of China in 2018. Four procedures were included in our model development as described below:

- 1) The MODIS 3-km AOD product is retrieved by using the atmospheric radiation transfer model. This transfer model not only makes the AOD retrieval progress complex, but also generates the missing values in some urban areas with relatively high reflectance. Our previous research has demonstrated that the random forest could be employed for estimating PM_{2.5} concentrations by using TOA reflectance; however, the AOD, instead of TOA reflectance, was the most important indicator for explaining the

PM_{2.5} variability. C. Zhao et al. (2019) developed a random forest model using T2M, T10M, PBLH, and other meteorological parameters to estimate the AOD in Beijing–Tianjin–Hebei region, and their results showed that the random forest could explain 92% variability of AOD. In view of this, a first-stage random forest incorporating six meteorological parameters (i.e., PBLH, RH, U10M, V10M, T2M, and T10M) was developed for predicting the missing AOD values in YRD–FJ–PRD region.

- 2) The AOD derived from the first-stage random forest (hereinafter referred as RF-derived AOD) along with the MODIS 3-km AOD were integrated as an independent variable (hereinafter referred as integrated AOD). A second-stage random forest was further developed to predict the PM_{2.5} concentrations by using the integrated AOD and other six auxiliary variables (i.e., TS, E-component, N-component, PS, MVI, and road density).
- 3) Both random forests were fine-tuned using library “randomForest” and package “caret” in R language to reduce the overfitting. Two important factors [i.e., the number of predictors tried for each node (m_{try}) and the number of decision trees in each random forest (n_{tree})], represented as the best performance of the random forests, were finally obtained for the two-stage random forests.
- 4) The tenfold cross validation (CV) method was carried out to assess the random forests. The dataset was randomly split into 10 segments, of which 9 segments were employed for model training, and the left was utilized to make the predictions (Yang et al. 2020). This process was repeated for 10 times, and the averaged CV result was used as the final accuracy. Two different types of CV methods were conducted in this study: (i) we first partitioned the dataset by monitoring sites distributed in the entire study area, namely, site-based CV method, to assess the spatial prediction accuracy of the random forests, and (ii) a time-based CV method with the dataset partitioned throughout the whole time period was employed to obtain the temporal prediction accuracy of the model. The coefficient of determination (R^2) and the root-mean-square error (RMSE) were used to evaluate the correlation of the estimated and observed PM_{2.5} concentrations.

The two-stage random forests can be simply described as Eqs. (1) and (2), and the framework of our model is presented in Fig. 2:

$$AOD_{ij} = RF(PBLH_{ij}, RH_{ij}, U10M_{ij}, V10M_{ij}, T2M_{ij}, T10M_{ij}), \quad (1)$$

$$PM_{2.5ij} = RF(int_AOD_{ij}, TS_{ij}, E_component_{ij}, N_component_{ij}, PS_{ij}, MVI, road_density), \quad (2)$$

where PBLH_{ij}, RH_{ij}, U10M_{ij}, V10M_{ij}, T2M_{ij}, T10M_{ij}, TS_{ij}, E-component_{ij}, N-component_{ij}, and PS_{ij} are the PBLH, RH, U10M, V10M, T2M, T10M, TS, E-component, N-component,

and PS, respectively, at station i on day j ; AOD_{ij}, int_AOD_{ij}, and PM_{2.5ij} are the measured AOD, integrated AOD, and PM_{2.5}, respectively, at station i on day j . MVI and road_density are the vegetation index and road density, respectively; RF defines the random forest.

3. Results

a. Modeling and validation

The daily AOD imageries at 0.55 μm from 1 January to 31 December 2018 were extracted from MODIS 3-km AOD products, and a total of 111 days with the spatial coverage of available AOD exceeding 40% was collected. The dataset with 29873 corresponding PM_{2.5} observations was finally obtained for YRD–FJ–PRD region during the entire study period. Table S1 in the online supplemental material gives the descriptive statistics of the variables used for modeling. Among three separated regions (i.e., YRD, FJ, and PRD), YRD region exhibited the highest values of daily PM_{2.5} concentrations ($\sim 377 \mu\text{g m}^{-3}$) and AOD (~ 2.19), while the relatively lower PM_{2.5} ($\sim 1.00 \mu\text{g m}^{-3}$) and AOD (~ 0.0004) were found located in the mountainous areas of FJ region. The PM_{2.5} concentrations of the main cities (e.g., Suzhou, Nanjing, and Wuxi) were much higher in the seasons of winter and spring than those in summer and autumn. Figure S1 illustrates the correlation of AOD and PM_{2.5} concentrations in three separated regions and four seasons. The PM_{2.5} concentrations were not highly correlated with the AOD in three regions (~ 0.30). The AOD presented correlation (r) of ~ 0.46 , 0.40 , 0.31 , and 0.30 with PM_{2.5} concentrations in summer, autumn, spring, and winter, respectively. The MVI showed a negative correlation ($r = -0.34$) with the PM_{2.5} concentrations, while the road density was positively correlated with PM_{2.5} ($r = 0.51$) (Fig. S2).

By tuning the random forest, we finally defined m_{try} as 3 and n_{tree} as 500 with smallest out-of-bag (OOB) error for both models. Six meteorological parameters including PBLH, RH, U10M, V10M, T2M, and T10M were employed to predict AOD in the first-stage random forest. The results showed that the RF-derived AOD presented a relatively high R^2 (~ 0.94) with MODIS 3-km AOD, and the PBLH was the most significant variable for explaining the AOD variability. The second-stage random forest was further trained by using the integrated AOD (i.e., RF-derived AOD and MODIS 3-km AOD) and other six auxiliary variables. Figure 3 illustrates the fitting and validation results of the second-stage random forest. We finally achieved the fitted R^2 of 0.91 and the RMSE of $9.07 \mu\text{g m}^{-3}$ between predicted and observed PM_{2.5} concentrations, which was relatively lower than that of the model using AOD–PM_{2.5} measurements ($R^2 = 0.92$, RMSE = $8.76 \mu\text{g m}^{-3}$). The validation results showed that the R^2 values for site-based and time-based CV method were 0.87 (RMSE = $12.09 \mu\text{g m}^{-3}$) and 0.85 (RMSE = $12.09 \mu\text{g m}^{-3}$), respectively, which indicated that more spatial PM_{2.5} variability was explained than temporal variability.

FIG. 2. The framework of the two-stage random forest.

The importance assessment carried out in the random forest model showed that AOD, PBLH, and PS were the most three important variables for PM_{2.5} variability in YRD-FJ-PRD, as their %IncMSE [i.e., increase in mean squared error (%)] values were much higher than other variables (Fig. 4). The predictive power of the two-stage random forest was also validated in three regions (Table 2). The results showed that the predicted PM_{2.5} agreed well with PM_{2.5} observations, as the fitted R^2 reached 0.93, 0.85, and 0.89 in YRD, FJ, and PRD, respectively. The RMSE was relatively higher ($\sim 10.09 \mu\text{g m}^{-3}$) in YRD than those in PRD ($\sim 7.96 \mu\text{g m}^{-3}$) and FJ ($\sim 7.63 \mu\text{g m}^{-3}$). The overall R^2 derived from two CV methods exhibited the same trend with that from model fitting, showing higher values in YRD than those in PRD and FJ. The model performance was further validated for the subdatasets divided by four seasons, and we obtained the fitted R^2 of 0.91, 0.86, 0.92, and 0.92 in winter, spring, summer, and autumn, respectively (Table 2). The RMSEs were much lower in summer ($\sim 6.11 \mu\text{g m}^{-3}$) and

autumn ($\sim 7.84 \mu\text{g m}^{-3}$) than those in spring ($\sim 8.49 \mu\text{g m}^{-3}$) and winter ($\sim 12.53 \mu\text{g m}^{-3}$).

b. Estimation of PM_{2.5} concentrations

Figure 5 illustrates the distribution of seasonal/annual mean PM_{2.5} concentrations derived from the two-stage random forest model for YRD-FJ-PRD in 2018. The annual-mean PM_{2.5} concentrations ranged from 29 to $64 \mu\text{g m}^{-3}$ with relatively high values ($>46 \mu\text{g m}^{-3}$) in most of YRD region and much lower values ($<37 \mu\text{g m}^{-3}$) in FJ Province. The concentrations of PM_{2.5} were 46.32, 38.80, and $36.15 \mu\text{g m}^{-3}$ in winter, spring, and autumn, respectively. The average PM_{2.5} ($\sim 30.16 \mu\text{g m}^{-3}$) was nearly 35% lower in summer compared with that in winter. Among three separated regions, YRD was the most polluted region with much higher seasonal-mean PM_{2.5} than PRD and FJ. Specifically, the developed and more-populated cities of YRD region, i.e., Suzhou, Wuxi, and Suqian, reached high city-based average PM_{2.5} values of 71.38, 68.91, and $68.82 \mu\text{g m}^{-3}$ in winter,

FIG. 3. The results of the two-stage random forest for (a) model fitting, (b) site-based CV, (c) time-based CV, and (d) three separated regions (black solid line: one-to-one line; dotted line: best-fit line).

respectively. These three cities also ranked the top five most polluted regions with high $PM_{2.5}$ concentrations in spring and autumn. PRD region showed relatively lower concentrations of $PM_{2.5}$ than YRD for four seasons, and three developed cities (i.e., Guangzhou, Dongguan, and Foshan) exhibited much higher $PM_{2.5}$ values than other coastal cities in PRD region. Our result was highly consistent with the findings of previous studies carried out in YRD and PRD region (Song et al. 2014; Ma et al. 2016). The annual- and seasonal- mean $PM_{2.5}$ concentrations were the lowest in FJ Province in 2018.

Three important meteorological variables including PBLH, PS, and RH were used for explaining the variation of seasonal $PM_{2.5}$ distribution in YRD–FJ–PRD region. The PBLH values were much deeper in warm days (i.e., summer and autumn) than those in cold days (i.e., winter and spring), while the average PS exhibited low values in warm days and

relatively high values in cold days. The low PBLH with high PS under moist environment (e.g., average RH = 61%) would make the aerosol particles accumulate rapidly (Yao et al. 2019), and therefore generate the high $PM_{2.5}$ concentrations in winter ($\sim 46.32 \mu\text{g m}^{-3}$) and spring ($\sim 38.80 \mu\text{g m}^{-3}$). Conversely, the deeper PBLH was favorable for the diffusion of the air pollutants in summer ($\sim 30.16 \mu\text{g m}^{-3}$). The unique characteristics of the meteorological fields were also responsible for the nonlinear correlation of AOD and $PM_{2.5}$ in the study area (Song et al. 2014). The $PM_{2.5}$ concentrations measured at surface level were relatively high in cold days, while the corresponding AOD values defined as the integral of the extinction coefficient of the total atmospheric column were much affected by the PBLH and RH. We found that the PBLH-and-RH-corrected AOD showed much higher correlation with $PM_{2.5}$ concentrations when compared with the original AOD.

FIG. 4. The importance of the random forest using modeling dataset (top) with available AOD measurements and (bottom) with integrated AOD [%IncMSE defines the increase in mean squared error (%)].

4. Discussion

In this paper, we developed a two-stage random forest integrating 14 ancillary parameters for predicting $PM_{2.5}$ in contiguous eastern coastal region of China. Some benefits of our method can be concluded as follows. First of all, the integrated AOD (RF-derived AOD and MODIS AOD) instead of MODIS AOD was used as the main independent variable. A total of 111 days with the spatial coverage of available AOD exceeding 40% was collected before data modeling, and 29873 $PM_{2.5}$ observations were finally obtained during the entire study period. However, only 33% of the dataset had AOD– $PM_{2.5}$ matchups, and 67% had $PM_{2.5}$ observations but not AOD available. Thus, we trained a random forest using 33% of the dataset with AOD– $PM_{2.5}$ matchups, and an overall R^2 of 0.92 with RMSE of $8.76 \mu g m^{-3}$ was obtained for model fitting. As the AOD was much affected by the

meteorological conditions, such as planetary boundary layer height and relative humidity, the AOD– $PM_{2.5}$ relationship cannot be simply described by using linear or multiple regression models. Therefore, we developed a random forest model incorporating PBLH, RH, PS, TS, U10M, and V10M to obtain the full-covered AOD distribution, and our results showed that the RF-derived AOD was highly correlated with MODIS 3-km AOD ($R^2 = 0.94$). Several other meteorological variables from GEOS-FP (e.g., total precipitation and albedo) were also added in the first-stage random forest, but the determined R^2 of the model was not improved significantly. Previous studies that have used random forests integrated with single AOD product (Vu et al. 2019), or multiple AOD products from different sensors (Hu et al. 2017), also achieved significant performance with R^2 of 0.7–0.8; however, it is still difficult to reach a 100% coverage of AOD distribution. Our previous research demonstrated that the TOA reflectance from MODIS level-1B product with spatial resolution of 1 km could be successfully employed to obtain the full-covered $PM_{2.5}$ concentrations in YRD region (Yang et al. 2020). We achieved a comparable result with determined R^2 values of 0.92 and 0.88 for site-based CV and time-based CV, respectively. However, it should be noted that although the TOA presented a high potential for estimating $PM_{2.5}$ concentrations in YRD region, the AOD is still the most important indicator for atmospheric pollution. Our previous research showed that the PBLH and RH, rather than the TOA reflectance at three bands, explained more variability of the $PM_{2.5}$ concentrations in YRD region, while the MODIS AOD was demonstrated as the most important indicator which accounted for the $PM_{2.5}$ variability in this study, followed by PBLH and RH. Our model reported a relatively higher R^2 of 0.91 using a volume of 29873 samples, compared with previous study conducted by Shen et al. (2018) which presented an overall R^2 of 0.87 using 4244 samples.

Another strength of our research is that we combined three regions which represent different levels of air quality in China as our study case. Compared with previous studies that have focused on $PM_{2.5}$ concentrations in YRD or PRD, respectively, our research presented the full-covered $PM_{2.5}$ concentrations in eastern coastal zone of China. Our results indicated that PBLH, RH, and PS were the most important meteorological predictors for explaining the $PM_{2.5}$ variability in YRD–FJ–PRD region, which has also been validated in previous studies carried out in several regions (Brokamp et al. 2018; Vu et al. 2019). Besides the meteorological parameters, two land-use variables (i.e., MVI and road density) also showed reasonable correlation with $PM_{2.5}$ concentrations, as the higher $PM_{2.5}$ concentrations mainly occurred in areas with lower MVI and higher road density. We also applied our model to estimate the $PM_{2.5}$ concentrations in YRD–FJ–PRD region from 1 January to 30 June 2019 and from 1 January to 30 June 2020, respectively, to validate the skill and stability of our proposed model, and the relatively low R^2 values of 0.65 and 0.45 were achieved for 2019 and 2020, respectively. The reason for the insignificant performance of our model in other periods was that 1) we only used 1-yr data (~2018) for

TABLE 2. Model performance for different regions and seasons.

	Model fitting		Site-based CV		Time-based CV	
	R^2	RMSE ($\mu\text{g m}^{-3}$)	R^2	RMSE ($\mu\text{g m}^{-3}$)	R^2	RMSE ($\mu\text{g m}^{-3}$)
Region						
YRD	0.93	10.09	0.90	12.21	0.88	14.00
FJ	0.85	7.63	0.80	8.89	0.75	9.78
PRD	0.89	7.96	0.85	9.11	0.83	10.19
Season						
Spring	0.91	8.49	0.87	9.89	0.81	11.57
Summer	0.86	6.11	0.82	7.05	0.75	8.33
Autumn	0.92	7.84	0.88	8.01	0.83	8.34
Winter	0.92	12.53	0.90	14.21	0.89	16.72

modeling, the model would present an unstable predictability for $\text{PM}_{2.5}$ estimation in 2019 and 2020; and 2) the low performance of the model in 2020 was mainly due to the substantial reduction of human activities during the COVID-19 period, which greatly affect the $\text{PM}_{2.5}$ distribution in YRD–FJ–PRD region.

Finally, it is likely to apply our proposed two-stage random forest to other coastal regions with the meteorological and landform condition similar to our study case. The $\text{PM}_{2.5}$ concentrations could be effectively predicted by using the two-

stage random forest incorporated with full-covered auxiliary variables in areas where the $\text{PM}_{2.5}$ measurements are missing or sparse. It should be noted that the daily MODIS AOD imagery with spatial coverage less than 40% was discarded before data matching; thus, the performance of the first-stage random forest would be affected when applied in regions with insufficient AOD samples. Besides the meteorological fields and land-use variables, the human induced factors (e.g., population density and industrial emission sources) are also the

FIG. 5. Full-covered seasonal- and annual-mean $\text{PM}_{2.5}$ in YRD–FJ–PRD.

important variables, which account for the local air pollution and could be included in the estimation model for enhancing the model performance (especially during the COVID-19 period). Specifically, the large industrial emission sources, such as power plants and steel plants, would greatly affect the particulate matter concentrations emitted into the atmosphere. Future research will focus on how the human induced factors affect the regional atmospheric environment and enhance the stability of our model in different periods and regions.

5. Conclusions

Our research is unique that we extended the widely used study case of YRD and PRD to a contiguous eastern coastal zone, i.e., YRD-FJ-PRD, which could highlight the large spatial variation of $PM_{2.5}$ in three separated regions. Our method showed a high predictability for $PM_{2.5}$ estimation with full spatial coverage, and the retrieved $PM_{2.5}$ concentrations presented high values in most of regions in YRD and relatively low values in FJ Province. Comparisons of the different models such as random forest and GWR model applied in a same region were also carried out, and our random forest model outperformed other statistical models with lower RMSE. Our results supported that the two-stage model driven by MODIS 3-km AOD and other ancillary information could be used for estimating $PM_{2.5}$ concentrations with relatively high accuracy.

Acknowledgments. The authors thank NASA for the use of MODIS and GEOS-FP data. This work was supported by the Initial Scientific Research Fund of Talents in Minjiang University (MJY20001), Fujian Provincial Natural Science Foundation program (2021J05204), and the project of Fujian Educational Bureau (2021JAT200435). The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The dataset on which this paper is based is too large to be retained or publicly archived with available resources. Documentation and methods used to support this study are available from the first author (email: 2611@mju.edu.cn) at Minjiang University.

REFERENCES

- Brokamp, C., R. Jandarov, M. Hossain, and P. Ryan, 2018: Predicting daily urban fine particulate matter concentrations using a random forest model. *Environ. Sci. Technol.*, **52**, 4173–4179, <https://doi.org/10.1021/acs.est.7b05381>.
- Dai, Q. H., X. D. Peng, P. J. Wang, C. L. Li, and H. B. Shao, 2018: Surface erosion and underground leakage of yellow soil on slopes in karst regions of southwest China. *Land Degrad. Dev.*, **29**, 2438–2448, <https://doi.org/10.1002/ldr.2960>.
- Goldberg, D. L., P. Gupta, K. Wang, C. Jena, Y. Zhang, Z. F. Lu, and D. G. Streets, 2019: Using gap-filled MAIAC AOD and WRF-Chem to estimate daily $PM_{2.5}$ concentrations at 1 km resolution in the eastern United States. *Atmos. Environ.*, **199**, 443–452, <https://doi.org/10.1016/j.atmosenv.2018.11.049>.
- Grinberg, N. F., O. I. Orhobor, and R. D. King, 2020: An evaluation of machine-learning for predicting phenotype: Studies in yeast, rice, and wheat. *Mach. Learn.*, **109**, 251–277, <https://doi.org/10.1007/s10994-019-05848-5>.
- Gupta, P., and S. A. Christopher, 2009: Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. *J. Geophys. Res.*, **114**, D20205, <https://doi.org/10.1029/2008JD011497>.
- , —, J. Wang, R. Gehrig, Y. Lee, and N. Kumar, 2006: Satellite remote sensing of particulate matter and air quality assessment over global cities. *Atmos. Environ.*, **40**, 5880–5892, <https://doi.org/10.1016/j.atmosenv.2006.03.016>.
- He, Q. Q., and B. Huang, 2018: Satellite-based mapping of daily high-resolution ground $PM_{2.5}$ in China via space-time regression modeling. *Remote Sens. Environ.*, **206**, 72–83, <https://doi.org/10.1016/j.rse.2017.12.018>.
- Hu, X., and Coauthors, 2013: Estimating ground-level $PM_{2.5}$ concentrations in the southeastern U.S. using geographically weighted regression. *Environ. Res.*, **121**, 1–10, <https://doi.org/10.1016/j.envres.2012.11.003>.
- , J. H. Belle, X. Meng, A. Wildani, L. A. Waller, M. J. Strickland, and Y. Liu, 2017: Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach. *Environ. Sci. Technol.*, **51**, 6936–6944, <https://doi.org/10.1021/acs.est.7b01210>.
- Khosravi, I., and S. K. Alavipanah, 2019: A random forest-based framework for crop mapping using temporal, spectral, textural and polarimetric observations. *Int. J. Remote Sens.*, **40**, 7221–7251, <https://doi.org/10.1080/01431161.2019.1601285>.
- Laurent, B., B. Marticorena, G. Bergametti, P. Chazette, F. Maignan, and C. Schmechtig, 2005: Simulation of the mineral dust emission frequencies from desert areas of China and Mongolia using an aerodynamic roughness length map derived from the POLDER/ADEOS 1 surface products. *J. Geophys. Res.*, **110**, D18S04, <https://doi.org/10.1029/2004JD005013>.
- Li, L. J., B. Z. Chen, Y. H. Zhang, Y. Z. Zhao, Y. Xian, G. Xu, H. F. Zhang, and L. F. Guo, 2018: Retrieval of daily $PM_{2.5}$ concentrations using nonlinear methods: A case study of the Beijing-Tianjin-Hebei region, China. *Remote Sens.*, **10**, 2006, <https://doi.org/10.3390/rs10122006>.
- Li, L. L., J. X. Yang, and Y. P. Wang, 2015: Retrieval of high-resolution atmospheric particulate matter concentrations from satellite-based aerosol optical thickness over the Pearl River delta area, China. *Remote Sens.*, **7**, 7914–7937, <https://doi.org/10.3390/rs70607914>.
- Li, Z. B., D. P. Roy, H. K. K. Zhang, E. F. Vermote, and H. Y. Huang, 2019: Evaluation of Landsat-8 and Sentinel-2A aerosol optical depth retrievals across Chinese cities and implications for medium spatial resolution urban aerosol monitoring. *Remote Sens.*, **11**, 122, <https://doi.org/10.3390/rs11020122>.
- Liu, Y., J. A. Sarnat, A. Kilaru, D. J. Jacob, and P. Koutrakis, 2005: Estimating ground-level $PM_{2.5}$ in the eastern United States using satellite remote sensing. *Environ. Sci. Technol.*, **39**, 3269–3278, <https://doi.org/10.1021/es049352m>.
- Ma, Z. W., Y. Liu, Q. Y. Zhao, M. M. Liu, Y. C. Zhou, and J. Bi, 2016: Satellite-derived high resolution $PM_{2.5}$ concentrations in Yangtze River delta region of China using improved linear mixed effects model. *Atmos. Environ.*, **133**, 156–164, <https://doi.org/10.1016/j.atmosenv.2016.03.040>.
- Mehdipour, V., D. S. Stevenson, M. Memarianfar, and P. Sihag, 2018: Comparing different methods for statistical modeling of

- particulate matter in Tehran, Iran. *Air Qual. Atmos. Health*, **11**, 1155–1165, <https://doi.org/10.1007/s11869-018-0615-z>.
- Mhawish, A., T. Banerjee, M. Sorek-Hamer, A. Lyapustin, D. M. Broday, and R. Chatfield, 2019: Comparison and evaluation of MODIS Multi-Angle Implementation of Atmospheric Correction (MAIAC) aerosol product over South Asia. *Remote Sens. Environ.*, **224**, 12–28, <https://doi.org/10.1016/j.rse.2019.01.033>.
- Prunicki, M., and Coauthors, 2018: Exposure to NO₂, CO, and PM_{2.5} is linked to regional DNA methylation differences in asthma. *Clin. Epigenetics*, **10**, 2, <https://doi.org/10.1186/s13148-017-0433-4>.
- Rojas-Rueda, D., M. Vrijheid, O. Robinson, A. G. Marit, R. Grazeviciene, R. Slama, and M. Nieuwenhuijsen, 2019: Environmental burden of childhood disease in Europe. *Int. J. Environ. Res. Public Health*, **16**, 1084, <https://doi.org/10.3390/ijerph16061084>.
- Sathe, Y., S. Kulkarni, P. Gupta, A. Kaginalkar, S. Islam, and P. Gargava, 2019: Application of Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol optical depth (AOD) and Weather Research Forecasting (WRF) Model meteorological data for assessment of fine particulate matter (PM_{2.5}) over India. *Atmos. Pollut. Res.*, **10**, 418–434, <https://doi.org/10.1016/j.apr.2018.08.016>.
- Shen, H. F., T. W. Li, Q. Q. Yuan, and L. P. Zhang, 2018: Estimating regional ground-level PM_{2.5} directly from satellite top-of-atmosphere reflectance using deep belief networks. *J. Geophys. Res. Atmos.*, **123**, 13 875–13 886, <https://doi.org/10.1029/2018JD028759>.
- Song, W. Z., H. F. Jia, J. F. Huang, and Y. Y. Zhang, 2014: A satellite-based geographically weighted regression model for regional PM_{2.5} estimation over the Pearl River delta region in China. *Remote Sens. Environ.*, **154**, 1–7, <https://doi.org/10.1016/j.rse.2014.08.008>.
- Vu, B. N., and Coauthors, 2019: Developing an advanced PM_{2.5} exposure model in Lima, Peru. *Remote Sens.*, **11**, 641, <https://doi.org/10.3390/rs11060641>.
- Wang, Z. F., L. F. Chen, J. H. Tao, Y. Zhang, and L. Su, 2010: Satellite-based estimation of regional particulate matter (PM) in Beijing using vertical-and-RH correcting method. *Remote Sens. Environ.*, **114**, 50–63, <https://doi.org/10.1016/j.rse.2009.08.009>.
- Wang, Z. T., S. L. Gao, J. F. Xie, and R. J. Li, 2019: Identification of multiple dysregulated metabolic pathways by GC-MS-based profiling of lung tissue in mice with PM_{2.5}-induced asthma. *Chemosphere*, **220**, 1–10, <https://doi.org/10.1016/j.chemosphere.2018.12.092>.
- Yang, L. J., H. Q. Xu, and Z. F. Jin, 2018: Estimating spatial variability of ground-level PM_{2.5} based on a satellite-derived aerosol optical depth product: Fuzhou, China. *Atmos. Pollut. Res.*, **9**, 1194–1203, <https://doi.org/10.1016/j.apr.2018.05.007>.
- , —, and S. D. Yu, 2020: Estimating PM_{2.5} concentrations in Yangtze River delta region of China using random forest model and the top-of-atmosphere reflectance. *J. Environ. Manage.*, **272**, 111061, <https://doi.org/10.1016/j.jenvman.2020.111061>.
- Yao, F., J. S. Wu, W. F. Li, and J. Peng, 2019: A spatially structured adaptive two-stage model for retrieving ground-level PM_{2.5} concentrations from VIIRS AOD in China. *ISPRS J. Photogramm. Remote Sens.*, **151**, 263–276, <https://doi.org/10.1016/j.isprsjprs.2019.03.011>.
- Zeng, Q. L., and Coauthors, 2018: Estimation of ground-level PM_{2.5} concentrations in the major urban areas of Chongqing by using FY-3C/MERSI. *Atmosphere*, **9**, 3, <https://doi.org/10.3390/atmos9010003>.
- Zhao, C., Z. R. Liu, Q. Wang, J. Ban, N. X. Chen, and T. T. Li, 2019: High-resolution daily AOD estimated to full coverage using the random forest model approach in the Beijing-Tianjin-Hebei region. *Atmos. Environ.*, **203**, 70–78, <https://doi.org/10.1016/j.atmosenv.2019.01.045>.
- Zhao, L. L., X. J. Qiu, W. B. Wang, R. M. Li, and D. S. Wang, 2019: NMR metabolomics and random forests models to identify potential plasma biomarkers of blood stasis syndrome with coronary heart disease patients. *Front. Physiol.*, **10**, 1109, <https://doi.org/10.3389/fphys.2019.01109>.