Environmental Research 216 (2023) 114465

Contents lists available at ScienceDirect

Environmental Research

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

# Evaluation of different machine learning approaches and aerosol optical depth in $PM_{2.5}$ prediction

Hamed Karimian [a], Yaqian Li [a], Youliang Chen [a,b,*], Zhaoru Wang [c]

[a] School of Civil and Surveying & Mapping Engineering, Jiangxi University of Science and Technology, Ganzhou, 341000, China
[b] School of Geosciences and Info Physics, Central South University, Changsha, China
[c] School of Resources and Environmental Engineering, Jiangxi University of Science and Technology, Ganzhou, 341000, China

ARTICLE INFO

*Keywords:*
Air pollution
Himawari-8
AOD
Wavelet transform
Remote sensing

ABSTRACT

Atmospheric Aerosol Optical Depth (AOD), derived from polar-orbiting satellites, has shown potential in $PM_{2.5}$ predictions. However, this important source of data suffers from low temporal resolution. Recently, geostationary satellites provide AOD data in high temporal and spatial resolution. However, the feasibility of these data in $PM_{2.5}$ prediction needs further study. In this paper, we analyzed the impact of AOD derived from Himawari-8 in $PM_{2.5}$ predictions. Moreover, by combining wavelet, machine learning techniques, and minimum redundancy maximum relevance (mRMR), a novel hybrid model was proposed. The results showed that AOD missing rate over Yangtze River Delta region is the highest in Nanjing, Hefei, and Maanshan. In addition, missing rates are the lowest in winter and summer (~80%). Moreover, we found that considering AOD, as an auxiliary variable in the model, could not improve the accuracy of $PM_{2.5}$ predictions, and in some cases decreased it slightly. In comparison with other models, our proposed hybrid model showed higher prediction accuracy, $R^2$ is improved by 11.64% on average, and root mean square error, mean absolute error, and mean absolute percentage error is reduced by 26.82%, 27.24%, and 29.88% respectively. This research provides a general overview of the availability of Himawari-8 AOD data and its feasibility in $PM_{2.5}$ predictions. In addition, it evaluates different machine learning approaches in $PM_{2.5}$ predictions. Our proposed framework can be used in other regions to predict different air pollutants concentrations and can be used as an aid for air pollution controlling programs.

## 1. Introduction

In recent years, the remarkable growth in the urbanization, civilization and industrialization has crucially resulted in environmental pollution problems and concerns such as water pollution, air pollution, soil pollution, noise pollution, plastic pollution, light pollution, deforestation, ozone depletion, radioactive waste and so on (Akhoondi et al., 2021; Doan et al., 2021; Mansoorianfar et al., 2022; Tran et al., 2022a). Consequently, scholars in different fields such as metals and steel, mining, chemical, textile, construction, gas and oil, marine, biomaterials, pharmaceutical, and etc., have investigated methods to model, prevent and diminish of various pollutants (Bijad et al., 2021; Hojjati-Najafabadi et al., 2022; Nguyen et al., 2021; Rao et al., 2022; Tran et al., 2022b). Fine particulate matter ($PM_{2.5}$) has become a serious environmental issue in China, especially in economically well-developed regions. $PM_{2.5}$ can penetrate inside our lungs and enter the blood circulatory system. Studies have shown that long-term exposure to $PM_{2.5}$ is closely related to the occurrence of many epidemic diseases (Khalili et al., 2018; Siyuan et al., 2018). Therefore, accurate prediction of $PM_{2.5}$ concentrations is of great significance for air pollution control and sustainable development (Karimian et al., 2012).

Generally speaking, methods for predicting $PM_{2.5}$ concentrations can be categorized into two major types: simulation-based methods and data mining-based methods (Karimian et al., 2019b). In simulation-based methods, global or regional chemical transport models are used to simulate emission, aerosol properties (e.g., aerosol type and size), and chemical transformation of air pollution. However, this approach is associated with numerical model uncertainties and difficulties in parametrizing aerosol emissions due to the lack of data (Karimian et al., 2016). Data-mining based method uses statistical or machine learning techniques to find patterns between explanatory variables and dependent variables (Karimian et al., 2019a). By assuming the linear relationship between $PM_{2.5}$ and predictors, models like multiple linear regression (Li et al., 2020), geographically weighted regression

* Corresponding author. School of Civil and Surveying & Mapping Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China.
*E-mail address:* 9120010023@jxust.edu.cn (Y. Chen).

https://doi.org/10.1016/j.envres.2022.114465
Received 1 May 2022; Received in revised form 11 September 2022; Accepted 27 September 2022
Available online 12 October 2022

(Karimian et al., 2017), land use regression (Shi et al., 2020) and Geodetector model (Chen et al., 2022b) have been developed. However, sometimes the linear assumption may not reflect the actual correlation between explanatory and response variables.

Machine learning-based approaches can get the optimal solution to a problem in a short time and can solve high-dimensional problems (Karimian et al., 2022; Sujatha et al., 2021a). Moreover, they have shown their feasibility to model non-linear systems such as in waste water treatment (Natarajan et al., 2021), chemical extraction (Sujatha et al., 2022) and in food recycling industry (Sujatha et al., 2021b). They have also become a powerful tool to predict various air pollutant concentrations such as Ozone (Mo et al., 2021) and $PM_{2.5}$ (Wu et al., 2018). Mo et al. (2020) and Dai et al. (2017) combined support vector machine (SVM) and particle swarm optimization (PSO) algorithms to establish a hybrid prediction model to predict ozone and $PM_{2.5}$ concentrations, respectively. The authors claimed that the machine learning approach is capable of forecasting ground $PM_{2.5}$ concentrations. Zhan et al. (2017) developed a Geographically-Weighted Gradient Boosting Machine (GW-GBM) model by building spatial smoothing kernels to weigh the loss function. This addresses the spatial non-stationary relationship between $PM_{2.5}$ concentrations and predictors. Some works have used time series and deep neural networks such as recursive neural networks (Biancofiore et al., 2017) and convolutional neural networks (Luo et al., 2020), to capture temporal and spatial dependencies in $PM_{2.5}$ forecasts. Moreover, some scholars have developed hybrid models and claimed robust performances in air pollution prediction (Su et al., 2019; Wang et al., 2019). Qi et al. (2019) proposed a novel hybrid Graph Convolutional networks and Long Short-Term Memory model (GC-LSTM) to improve the $PM_{2.5}$ spatiotemporal forecasting accuracy in China. Experimental results showed that the hybrid model could achieve the best performance for predictions, compared with state-of-the-art methods. Niu et al. (2016) applied complementary ensemble empirical mode decomposition with support vector regression optimized by a grey wolf optimizer to predict the daily average concentrations of $PM_{2.5}$. The empirical study indicates that the proposed hybrid model is remarkably superior to other benchmark models for its higher prediction accuracy. A hybrid-wavelet model for the forecasting of hourly $PM_{2.5}$ was presented by Wang et al. (2019). The authors concluded that the generalization ability of hybrid models is more robust than those of single Artificial Intelligence (AI) models, and proved the potential of wavelet transformation in air pollution forecasts.

Although ground level monitoring stations can report $PM_{2.5}$ concentrations with high accuracy and temporal resolution, their spatial resolution is coarse and it is not appropriate for exposure analysis and epidemiological studies. Atmospheric aerosol optical depth (AOD) is one of the remote sensing products that has been widely used as a complementary source of data, beside ground monitoring data, to produce the surface distribution of $PM_{2.5}$ (Shi et al., 2018; Sun et al., 2019). However, because of its low temporal resolution, it has been seldom used in times series modeling. The growth in the availability of AOD derived from geostationary satellites has provided the chance for data mining researchers to use this source of data as an explanatory variable. In comparison with polar-orbiting satellites, geostationary satellites provide AOD in higher temporal and spatial resolution. The Advanced Himawari Imager (AHI) onboard Himawari −8 and −9 is the geostationary satellite that provides the full disk image of the Earth every 10 min. The location of satellites, at 140.7°E, enables complete coverage of East and South Asia. Yoshida et al. (2018) proposed an optimal estimation method for aerosol retrieval over land. The AHI level-3 datasets provide hourly AOD at 500 nm and with 5 km × 5 km spatial resolution. Lim et al. (2018) reported a good match between Himawari AOD and AERONET, and the best correlation was observed in summer ($r = 0.93$). Jiang et al. (2019) reported a good correlation ($R = 0.92$) between MODIS Deep Blue AOD and that of Himawari. Moreover, the best compatibility with AERONET was observed over East Asia.

In this study, we perform deep analysis to examine the role of Himawari AOD in $PM_{2.5}$ predictions. Besides, we propose and evaluate a novel hybrid model to forecast $PM_{2.5}$ concentrations. Through that we investigate the role of wavelet technique in improving the performance of different models. In addition, the feasibility of the minimum redundancy maximum relevance (mRMR) technique, for the optimal selection of input features, is examined in combination with various techniques and models.

## 2. Data and methods

### 2.1. Study area

The Yangtze River Delta region (YRD) (29.33°~32.57°N, 115.77°~123.42°E) is located in the lower reaches of the Yangtze, and it plays a pivotal role in China's economic development. Because the YRD is one of the areas with relatively severe pollution, this article takes it as the research area to analyze the availability of AOD data at 137 sites. Moreover, to verify the prediction accuracy of our proposed hybrid model, four representative sites with the largest amount of available data were chosen: Hefei (Binhu New Area: 117.29°E), Hangzhou (Xiasha: 119.74°E, 30.09°N), Nanjing (Pukou: 118.62°E, 32.07°N) and Shanghai (Pudong New Area: 121.52°E, 31.22°N). The specific geographic location and the distribution of air quality monitoring stations are shown in Fig. 1.

### 2.2. Satellite data

The geostationary satellite data involved in this research are Himawari-8 AOD (level-3), the wavelength is 0.64 μm , and they are provided by Japan Meteorological Agency (JMA). To extract AOD over each air pollution monitoring station, we used the Python netcdf4 library. As AOD is reported in 500 nm which is in visible band, it is only reported during day time and in the existence of sun. Therefore, the AOD is available only for daytime hours.

### 2.3. Meteorological data

In oppose to ground-based monitoring stations that suffers from low spatial resolution (coarse distribution) and cannot monitor the variation of weather condition across large area (e.g. city), model-based data provide surface distribution of different meteorological factors with high spatial resolution (Karimian et al., 2019b). Therefore, We collected hourly meteorological data from European Centre for Medium-Range Weather Forecasts (ECMWF) (http://www.ecmwf.int/) from January 1, 2016 to December 31, 2017. These data include 2-m temperature (T), Boundary layer height (BLH), Relative humidity (RH) 10-m wind east and north components (U, V), surface-level pressure (P), and forecasted albedo (FA). This centre provides data in high spatial resolution (0.1° × 0.1°) and up to 10 days forecast.

### 2.4. Ground-level pollutants data

We collected hourly ground-based air pollutants data ($PM_{2.5}$, $PM_{10}$, $SO_2$, $NO_2$, CO, $O_3$) from the real-time air quality release system of China National Environmental Monitoring Centre for the period of our study. Table 1 provides a statistical summary and features of input data.

### 2.5. Wavelet transform

Wavelet transform (WT) decomposes the original time series data of $PM_{2.5}$ concentrations into components of different frequencies and therefore can provide more detailed change information (Fang et al., 2022). It can be divided into two types: discrete and continuous. The factors used in discrete wavelet transform processing are called discrete data, while these factors in continuous wavelet transform are called continuous data. The former is mainly used for discrete operations such

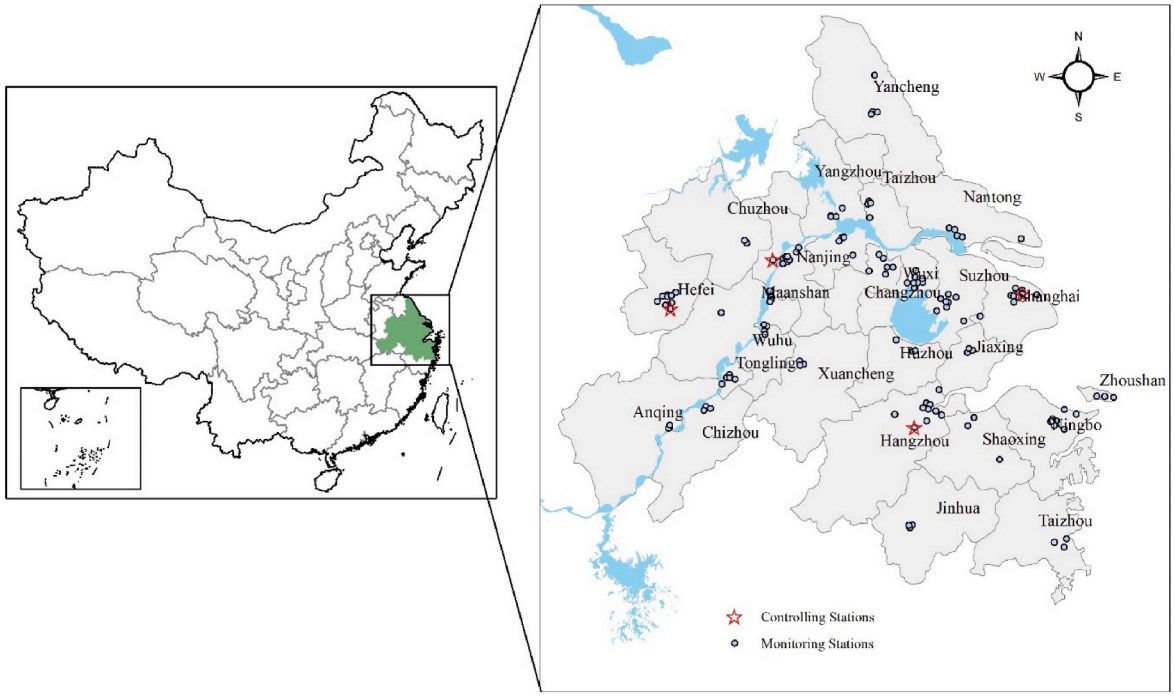


**Fig. 1.** YRD area and distribution of air quality monitoring stations (circle) and controlling stations (asterisk).

**Table 1**
Statistical summary of input data. Mean values are shown with one standard deviation.

| VARIABLE | UNIT | MEAN | RANGE |
|---|---|---|---|
| $PM_{2.5}$ | $\mu g/m^3$ | 38.01 ± 30.25 | [1265] |
| $O_3$ | $\mu g/m^3$ | 76.91 ± 48.87 | [1353] |
| $SO_2$ | $\mu g/m^3$ | 12.59 ± 6.83 | [2,97] |
| $NO_2$ | $\mu g/m^3$ | 44.33 ± 27.04 | [2212] |
| $PM_{10}$ | $\mu g/m^3$ | 56.99 ± 36.08 | [3531] |
| CO | $mg/m^3$ | 0.746 ± 0.36 | [0.001,4.5] |
| AOD | – | 0.413 ± 0.270 | [0.056,1.841] |
| T | K | 290.65 ± 8.02 | [267.35,305.89] |
| BLH | m | 512.82 ± 335.81 | [13.15,2228.02] |
| RH | % | 80 ± 12 | [36,99] |
| U | m/s | −1.71 ± 3.31 | [-11.65,11.59] |
| V | m/s | −0.36 ± 4.87 | [-15.33,11.20] |
| P | KPa | 101.61 ± 0.89 | [99.73,104.08] |
| FA | – | 0.06 ± 6.07E-06 | [0.05998,0.06001] |

as denoising, and the latter is used for time series change analysis (Yang et al., 2017).

Suppose $\psi(t)$ is a square-integrable function, that is, $\psi(t) \in L^2(R)$. If the function can satisfy Eq. (1), then $\psi(t)$ is the wavelet basis function. The wavelet transformation can be derived through Eq. (2).

$$\int_{-\infty}^{\infty} \frac{\psi(\omega)^2}{\omega} d\omega < \infty \tag{1}$$

$$\omega_{a,b}(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) dt \tag{2}$$

In the case of a continuity function $f(t)$, the continuous wavelet transforms expression is:

$$W_f(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \psi^*\left(\frac{t-b}{a}\right) dt \tag{3}$$

where $a$ represents the scale factor, $b$ is the translation factor, and $\psi^*$ denotes the complex conjugate.

The number of wavelet decomposition levels affects the model efficiency. Moreover, if the number of decomposition levels is too high, the linear part after decomposition will significantly differ from the original features. Therefore, the number of optimum decomposition levels is derived through Eq. (4).

$$Smooth(M) = \frac{\sum_{i=1}^{N-1} (a_M(i+1) - a_M(i))^2}{\sum_{i=1}^{N-1} (X(i+1) - X(i))^2} \tag{4}$$

In the above equation, $N$ represents the number of samples, $M$ represents the number of decomposition levels, $X$ is the $PM_{2.5}$ original data set, and $a_M$ is the approximate sequence. By setting a threshold (T), the number of decomposition levels is determined if Smooth(M) ≤ T. It is recommended to set the threshold T to 0.005 (Su et al., 2019). Through Eq. (4), the decomposition level was set to 6. Moreover, as db5 can provide the smallest variability of $PM_{2.5}$ at different levels (Sun et al., 2019), it was selected as the wavelet function. Finally, different models are used to estimate the decomposed high- and low-frequency information effectively.

### 2.6. Minimum redundancy maximum relevance

Using highly correlated explanatory variables reduces the applicability of models. Some scholars (Juhos et al., 2008; Zhang et al., 2017) have used correlation coefficient or principal component analysis to select independent variables. However, the application of the correlation coefficient method is based on the independent setting of each variable. In addition, principal component analysis can only deal with linear problems. Therefore, the mRMR algorithm based on mutual information was used for feature selection. The core of this algorithm is to maximize the relevancy between independent and target variables while minimizing the redundancy between features (Ju and He, 2018) The mutual information of variables X and Y is defined as:

$$I(X,Y) = \iint \rho(X,Y) \log \frac{\rho(X,Y)}{\rho(X)\rho(Y)} dXdY \tag{5}$$

where $\rho(X)$ and $\rho(Y)$ are the probability density functions of $X$ and $Y$,

respectively, and $\rho(X,Y)$ is the joint probability density function. The core of the mRMR algorithm can be expressed by max relevance (Eq. (6)) and min redundancy (Eq. (7)).

$$\begin{cases} \max D(S,p) \\ D = \frac{1}{n}\sum_{i=1}^{n} I(x_i,p) \end{cases} \tag{6}$$

$$\begin{cases} \min R(S) \\ R = \frac{1}{C_n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} I(x_i,y_i) \end{cases} \tag{7}$$

In the above, $S$ represents the feature subset, $n$ is the number of features, $I(x_i,p)$ represents the mutual information between features, and $PM_{2.5}$, $p$ represents the target feature, and $I(x_i,y_i)$ means the mutual information between features. Generally, by combining Eq. (6) and Eq. (7), the final maximum relevance and minimum redundancy judgment conditions are obtained in Eq. (8).

$$\begin{cases} \max \varphi(D,R) \\ \varphi(D,R) = D - R \end{cases} \tag{8}$$

### 2.7. Random forest

Random forest (Liaw et al., 2002) (RF) is a classifier ensemble algorithm, which is not prone to overfitting. It can be used for classification and regression problems. In this method, decision trees are built based on bootstrapped sample techniques, a random sampling technique with replacement. Moreover, to avoid splitting the trees only rely the strongest predictors, each time a split in a tree happened an arbitrary number of predictors is considered. For regression, the random forest algorithm draws a bootstrap sample and then feeds a tree to this sample. These two stages are repeated several times, based on the selected number of trees. To predict a new record, run the record down each tree, and at each time a prediction is computed. The final prediction for a new record is the average of these B individual predictions (Eq. (9)).

$$RF = \frac{1}{B}\sum_{1}^{B}\left(\sum_{J=1}^{J}\overline{y}\bullet I(x\in R_J)\right) \tag{9}$$

### 2.8. Gradient Boosting Regression Tree

The estimation result of the Gradient Boosting Regression Tree (GBRT) model is affected by the value of the learning rate (*lr*) and the number of regression trees (*M*). When the learning rate is fixed, as the number of regression trees increases, the model prediction accuracy improves. However, after *M* reaches a certain value, the accuracy will not increase significantly. When the value of *M* is fixed and *lr* increases, the model's accuracy is improved, but if *lr* is large, it is easy to ignore the global optimal solution. In the case of small *r*, a locally optimal solution will appear. This study uses the grid search method to determine the optimal parameters.

### 2.9. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) is an extension of the gradient boosting algorithm. The gradient boosting algorithm is a machine learning-based technique used for regression and classification problems. It generates a prediction model in the form of an integration of multiple weak prediction models.

By assuming that XGBoost generates $k$ trees, for any input $x$, each tree has an output $f_i x$, then the output of the XGBoost model $F(x) = \sum_{i=1}^{k} f_i(x)$. The cost function of XGBoost includes a regularization term to reduce overfitting. The XGBoost algorithm has the advantages of fast running speed, high accuracy, and does not overfit easily. The algorithm has good universality and performs well in many applications. Therefore, this research utilized it for $PM_{2.5}$ predictions.

### 2.10. The hybrid forecasting framework

The first step in our proposed framework is using wavelet transform (WT) to decompose the original $PM_{2.5}$ concentration sequence into an approximate sequence (A6) and six wavelets transform sequences (D1, D2, D3, D4, D5, D6). A6 reflects the general trend of the data. D1-D6 are the detailed sequences reflecting the slight fluctuations of the sequence, and each sub-sequence is used as the model output. In the second step, mRMR is used to select the optimal feature subset as the input of each subsequence. Finally, the sub-sequence prediction results are integrated to obtain the predicted $PM_{2.5}$ concentrations. The wavelet decomposition and feature selection part are run on the Matlab platform, and the machine learning model construction and parameter selection part are completed in Python. Fig. 2 illustrates the schematic of our proposed hybrid model. It is worth mentioning that our models were trained using 60% of the data and 20% of data were used for validation. The remaining 20% has been used for testing.

### 2.11. Evaluation criteria

To evaluate the performances of different models in this study, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) were computed (Chen et al., 2022a). These evaluation indicators measure the closeness of forecasts to the observed values over the test dataset (20% of data). Moreover, to analyze the prediction strength of models, the coefficient of determination ($R^2$) was also considered(Sujatha et al., 2021b).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{y}_i - y_i| \tag{10}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}{n}} \tag{11}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\widehat{y}_i - y_i}{y_i}\right| \times 100\% \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\overline{y}_i - y_i)^2} \tag{13}$$

In the above, $n$ is the number of samples used for evaluation, $y_i$ is the actual observation value, and $\widehat{y}_i$ is the predicted value.

## 3. Results and discussions

### 3.1. Analysis of the availability of AOD

Based on the hourly AOD data of Himawari-8 from 8:00 to 18:00, we calculated the average percentage of missing AOD data at each station in the YRD from 2016 to 2017. As shown in Fig. 3, the distribution of AOD missing rates shows significant differences. The areas with high missing AOD data (85%~100%) are located in the western inland such as Hefei, Nanjing, Zhenjiang, Yangzhou, and Maanshan. This may be due to the high satellite zenith angles and excessive cloud cover in the western regions. Moreover, miscalculations in AOD retrieval may happen in urban areas with stronger surface reflectance that causing wrongly discarding of correct data by different filters. The areas with higher AOD availability are mainly located in southeast coastal regions such as Zhoushan and Taizhou. In terms of seasons, the AOD missing rate in the whole region is the highest in spring (88%), followed by autumn (86%),

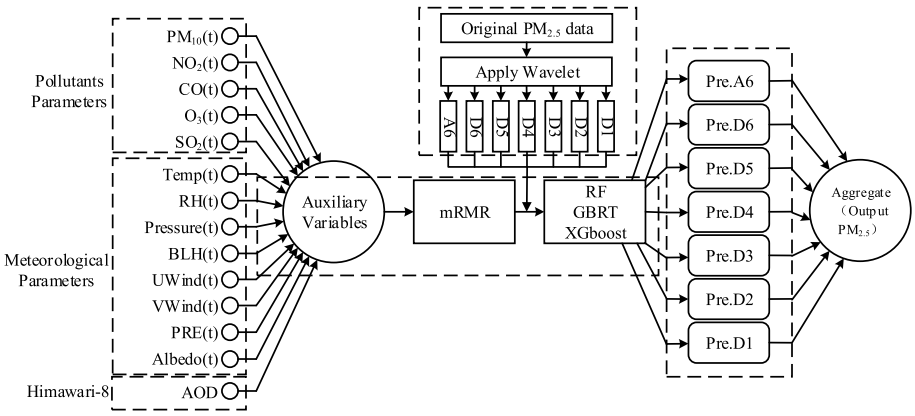


**Fig. 2.** The basic structure of the hybrid forecasting framework.

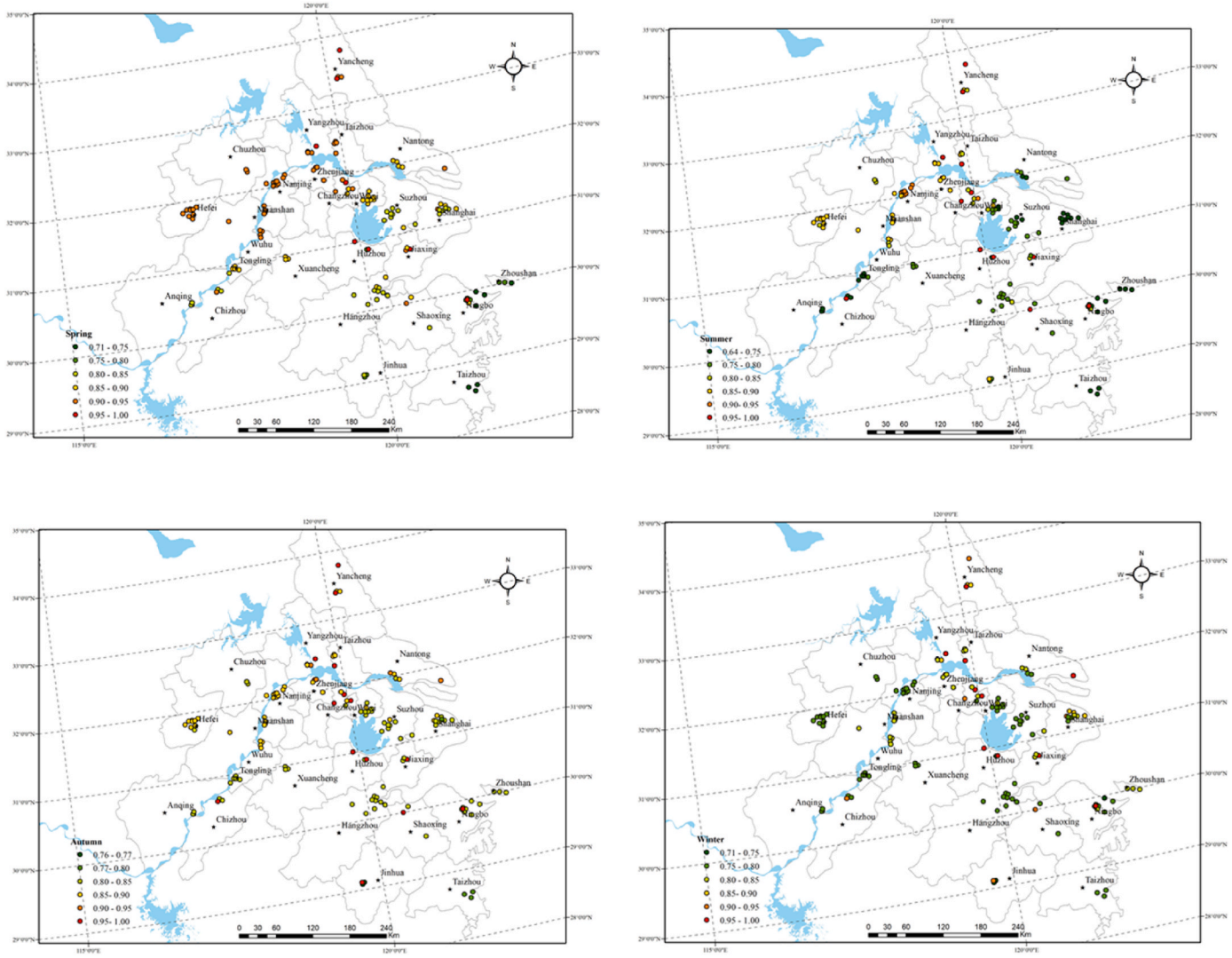


**Fig. 3.** Missing rate of AOD in different seasons.

and similar in summer and winter (81%). As the wet seasons in YRD are spring and autumn, cloud cover can be considered as one of the main reasons for high missing AOD data in these seasons. The lower rate of missing data in winter may be due to the less cloud cover than in wet seasons. In summer, the study area has dense vegetation, fewer clouds, and good light conditions that are favorable for retrieving AOD. As the radiation transmission between the satellite and the ground station always maintains an oblique path, and the simulation error of the radiation transmission model is greater when it is tilted, this can be considered as another reason for the high rate of missing data.

### 3.2. Evaluating the impact of AOD on prediction accuracy

According to the discussion in the previous section, we selected four representative sites (Pudong, Pukou, Xiasha, Binhu) that had the lowest rate of missing AOD. To evaluate the role of Himawari AOD in $PM_{2.5}$ predictions, we considered three scenarios. In the first scenario, we input the whole dataset (the quantity is 32164) to the RF model but excluded AOD data. In the second scenario, we used those observations which have AOD (the quantity is 5046). As there are few observations in our dataset that have AOD, and the size of the input dataset has a direct effect on model performance, in the third scenario, we input the same dataset as in the second step but excluded AOD from it.

The forecasting results are shown in Table 2. Overall, the RMSE, MAE, and MAPE of the same observations with AOD (scenario 2) are greater than those without AOD (scenario 3) and over the whole dataset (scenario 1). The RMSE, MAE, and MAPE of observations with AOD are 14.43%, 15.15%, and 15.92% higher than that of scenario 1. In general, the $R^2$ of scenario 1 is the largest, the error is the smallest, and the prediction accuracy is the highest. Therefore, adding AOD as an explanatory variable in the model will not improve the $PM_{2.5}$ prediction accuracy. Therefore, AOD is not used in our hybrid model to forecast the $PM_{2.5}$ concentrations. This result is inconsistent with the one done by Wei et al. (2017) over Beijing Tianjin and Hebei in which the authors claimed that AOD from Himawari-8 is useful for $PM_{2.5}$ predictions. This may be due to the large missing AOD data, over our study area that makes the role of AOD insignificant. Our findings are compatible with Zamani Joharestani et al. (2019) in which the authors claimed that the AOD derived from MODIS could not improve their model performance for $PM_{2.5}$ predictions.

### 3.3. Comparative analysis of prediction results of multiple models

This study aims to forecast $PM_{2.5}$ concentrations. However, the high variability of the $PM_{2.5}$ time series makes the accurate prediction a challenging task. By using a wavelet, the original time series is decomposed into several low variability sub-series. Applying our proposed model to each of these sub-series and then summing up the results may improve the model's performance. The decomposition of the $PM_{2.5}$ time series in Pukou from 2016 to 2017 is shown in Fig. 4.

**Table 2**
The performance of RF with and without AOD.

| Site | Index | Over the whole dataset (Scenario 1) | Observations with AOD (Scenario 2) | Same observations without AOD (Scenario 3) |
|---|---|---|---|---|
| Pudong New Area (1149A) | RMSE | **8.42** | 9.18 | 8.57 |
| | MAE | **5.70** | 6.29 | 5.80 |
| | $R^2$ | **0.88** | 0.88 | 0.88 |
| | MAPE | **25.51** | 28.81 | 27.24 |
| Pukou (1157A) | RMSE | **11.02** | 12.61 | 12.31 |
| | MAE | **7.59** | 8.75 | 8.70 |
| | $R^2$ | **0.91** | 0.91 | 0.91 |
| | MAPE | **19.61** | 22.73 | 22.67 |
| Xiasha (1226A) | RMSE | **8.53** | 9.19 | 9.43 |
| | MAE | **5.27** | 5.89 | 6.05 |
| | $R^2$ | **0.92** | 0.89 | 0.89 |
| | MAPE | **12.36** | 13.05 | 13.30 |
| Binhu New Area (1278A) | RMSE | **14.35** | 14.93 | 14.91 |
| | MAE | **9.16** | 10.46 | 10.38 |
| | $R^2$ | **0.86** | 0.77 | 0.77 |
| | MAPE | **22.26** | 23.49 | 23.17 |
| Overall | RMSE | **10.58** | 11.47 | 11.30 |
| | MAE | **6.93** | 7.85 | 7.73 |
| | $R^2$ | **0.90** | 0.86 | 0.86 |
| | MAPE | **19.93** | 22.02 | 21.84 |

#### 3.3.1. Prediction of $PM_{2.5}$ concentrations based on machine learning and hybrid model

The optimal features selected by the mRMR method were used as the inputs of our proposed hybrid model. Fig. 5 shows the scatter plots based on different models in Pudong. The blues are plain models, the yellows are WT + Plain models, and the reds are WT + Plain + mRMR models. There is a reasonable agreement between the ground truth and the predicted values using the WT + Plain + mRMR model, and our hybrid model can explain 90% ($R^2 = 0.90$) of the variability in observed concentrations. For different models, the slopes of linear regression are less than 1, and intercepts are positive. We can infer the tendency of different models to underestimate and overestimate high and low concentrations, respectively.

To further analyze the prediction accuracy of models, Fig. 6 shows the results of RMSE, MAE, MAPE, and $R^2$. As can be seen, the prediction accuracy of the nonlinear model is better than the multiple linear regression model in all four stations. Moreover, the prediction accuracy of GBRT as the plain model is lower than RF and XGboost models. All four WT + Plain models in evaluation indicators are better than plain models. Compared with the plain model, $R^2$ has increased by 10.68% in the WT + Plain model, and RMSE, MAE, and MAPE have decreased by 23.93%, 22.80%, and 24.66%, respectively. This shows that using a wavelet can discover more detailed information, and therefore can effectively improve the accuracy of $PM_{2.5}$ predictions. Except for Binhu, where the WT + Plain is the optimal model, for the other three sites, the WT + Plain + mRMR got the highest prediction accuracy than other models. Therefore, it can be inferred that the optimal subset selected by the mRMR algorithm is effective. It reduces the dimensionality of the input data while improving the prediction accuracy. From a spatial view, the accuracy of all models built at the four sites is different. The model performed the best in Xiasha, where $R^2$ of the WT + Plain + mRMR is greater than 0.9, and RMSE <10. Therefore, it is necessary to establish a prediction model for different regions individually.

## 4. Conclusions

In contrast with polar-orbiting satellites that provide AOD with the low temporal resolution, AOD from geostationary satellites has a high temporal resolution (hourly data). This paper studied the influence of geostationary AOD on $PM_{2.5}$ predictions and proposed a hybrid model to improve the accuracy of $PM_{2.5}$ predictions effectively. We found that the missing rate of Himawari-8 AOD in the Yangtze River Delta is noticeable and varied spatially and temporally. The areas with highest missing rate are located in western regions such as Nanjing, Hefei, and Maanshan. The missing rate is the highest in spring (88%) and the lowest in summer (81%) and winter (81%). This lowest rate can be explained by dense vegetation and less cloud cover in summer and winter in the study area. In investigating the role of AOD in $PM_{2.5}$ predictions we found that adding AOD as an auxiliary variable does not improve the prediction accuracy. Comparing various models performance, machine learning prediction accuracy is higher than the multiple linear regression model. The $R^2$ of the machine learning model (RF, GBRT, XGboost) is 14.49% higher than the multiple linear regression on average, and RMSE, MAE, and MAPE are decreased by 18.74%, 17.95%, 16.92%. Moreover, considering wavelet decomposition (WT + Plain models) can effectively improve the $PM_{2.5}$ prediction accuracy ($R^2$: 0.89, RMSE: 10.91 μg/m$^3$, MAE: 7.46 μg/m$^3$, MAPE: 21% vs $R^2$: 0.80, RMSE:14.40 μg/m$^3$, MAE: 9.68 μg/m$^3$, MAPE: 27%, respectively). Overall, the WT + XGboost + mRMR model proposed in this study performed the best among all prediction models. The findings and methods of this study can be used in other regions and in future air pollution studies.

## Credit author statement

Conceptualization & Verification, Hamed Karimian and Youliang Chen; Methodology, Yaqian Li & Hamed Karimian; Software, Yaqian Li;

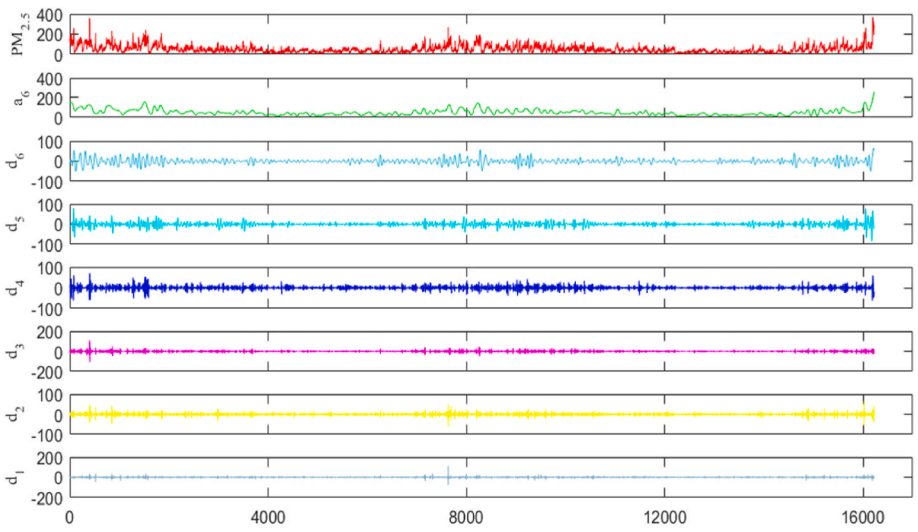


**Fig. 4.** Decomposition diagram of $PM_{2.5}$ time series using wavelet in Pukou. d1-d6 represent the wavelet coefficients at different levels (i.e. the detailed components) and a6 represents the coarse approximation signal.

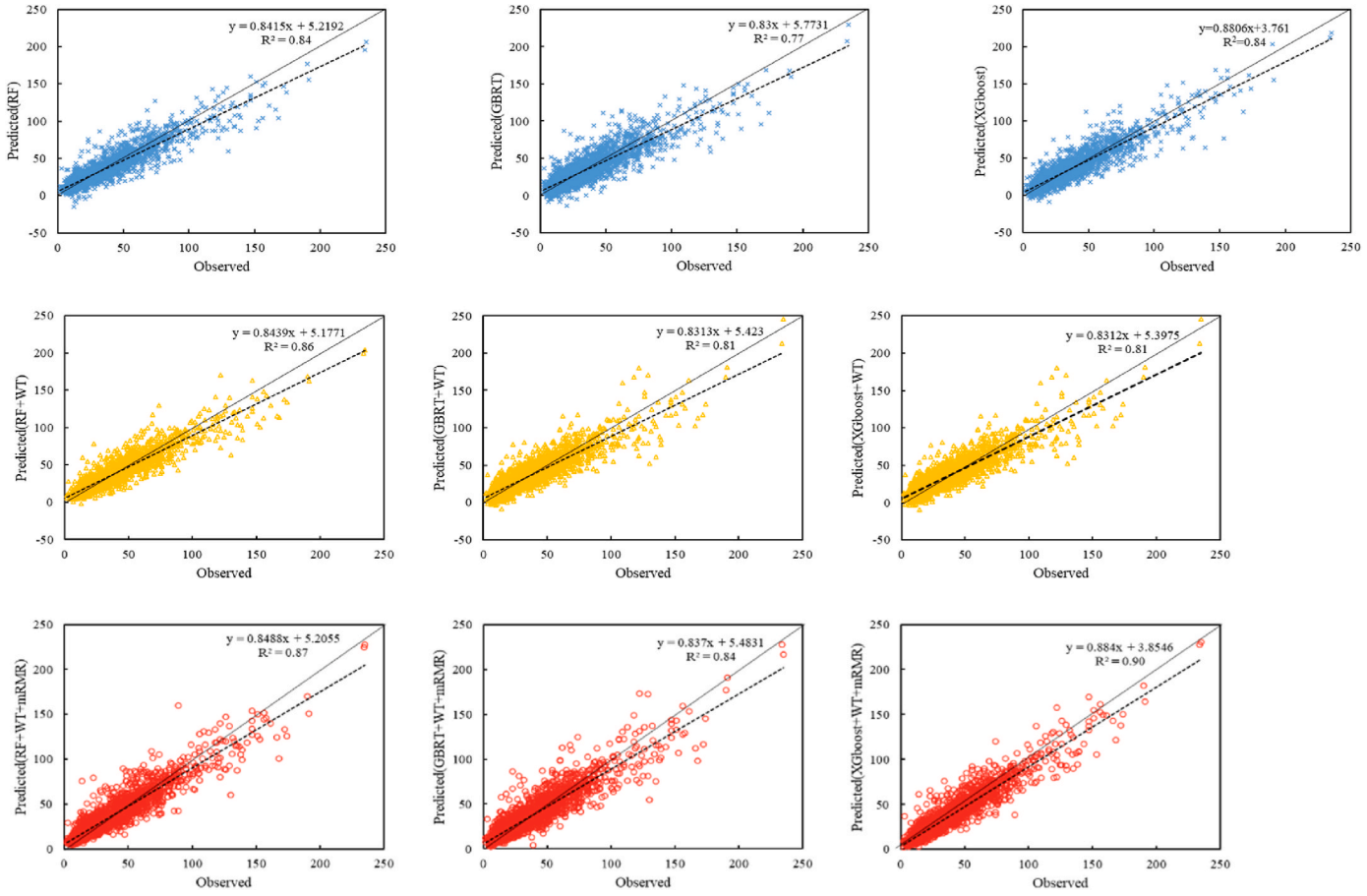


**Fig. 5.** Scatter plot of observed and predicted values of $PM_{2.5}$ mass concentration. Note: The dotted line indicates the degree of fit, and the solid line indicates the perfect fit, ie y = x.

Validation, Yaqian Li & Hamed Karimain; Formal Analysis, Yaqian Li; Writing-Original Draft Preparation, Yaqian Li; Writing-Review & Editing, Hamed Karimian; Visualization, Zhaoru Wang; Supervision, Hamed Karimian &Youliang Chen.

## Declaration of competing interest

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

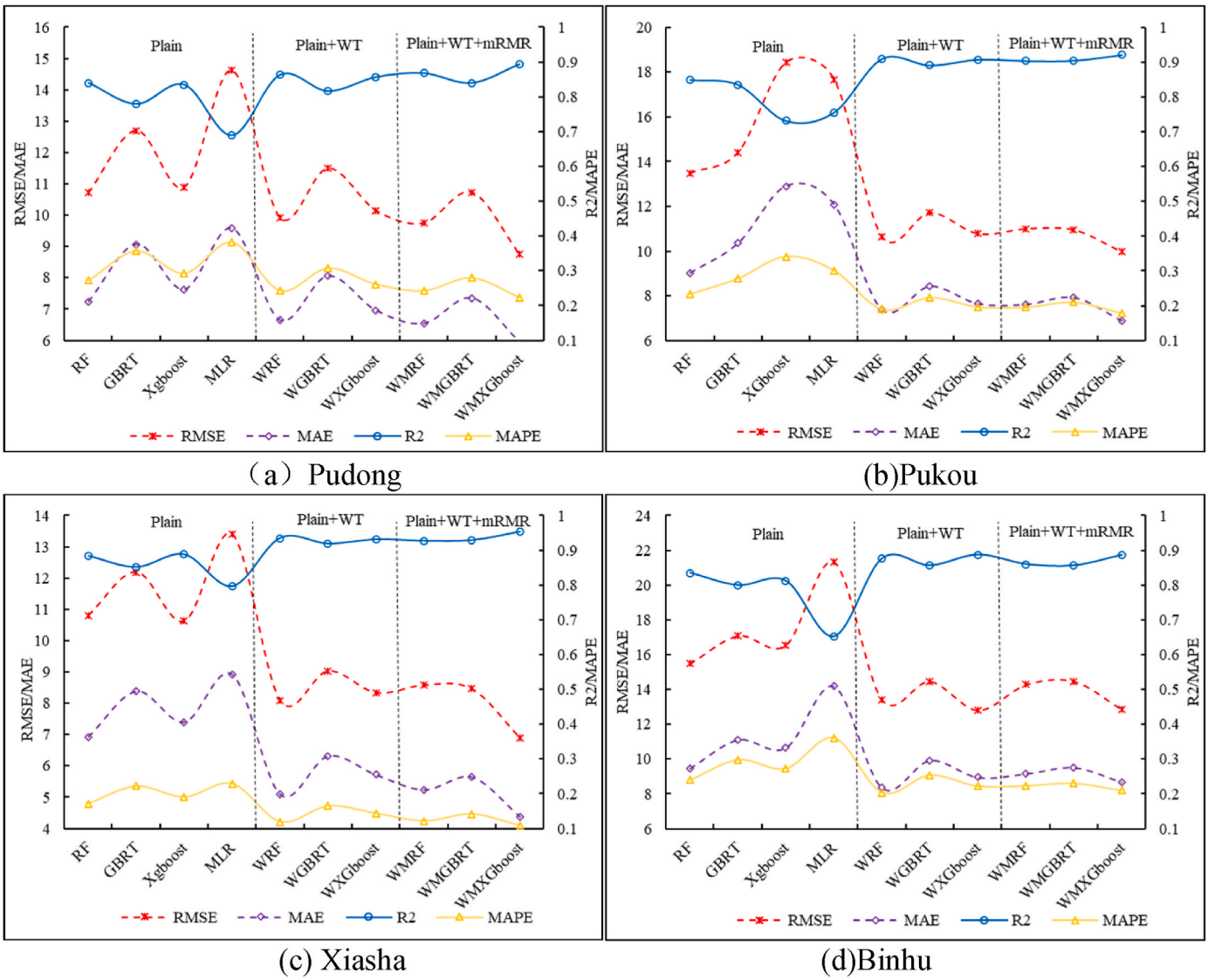


**Fig. 6.** Comparison of various model evaluation indicators.

## Data availability

Data will be made available on request.

## Acknowledgment:

The authors are grateful for the National Natural Science Foundation of China [No. 42261072] for providing financial support for this research.

## References

Akhoondi, A., et al., 2021. Advances in metal-based vanadate compound photocatalysts: synthesis, properties and applications. Synthesis and Sintering 1, 151–168.
Biancofiore, F., et al., 2017. Recursive neural network model for analysis and forecast of PM10 and PM2.5. Atmos. Pollut. Res. 8, 652–659.
Bijad, M., et al., 2021. An overview of modified sensors with focus on electrochemical sensing of sulfite in food samples. Eurasian Chemical Communications 3, 116–138.
Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System.
Chen, Y., et al., 2022a. The relationship between air quality and MODIS aerosol optical depth in major cities of the Yangtze River Delta. Chemosphere, 136301.
Chen, Y., et al., 2022b. Spatio-temporal variation of ozone pollution risk and its influencing factors in China based on Geodetector and Geospatial models. Chemosphere 302, 134843.
Dai, L., et al., 2017. Dynamic forecasting model of short-term PM2.5 concentration based on machine learning. J. Comput. Appl. 37, 3057–3063.
Doan, V.D., et al., 2021. Comparative study on adsorption of cationic and anionic dyes by nanomagnetite supported on biochar derived from Eichhornia crassipes and Phragmites australis stems. Environ. Nanotechnol. Monit. Manag. 16, 100569.
Fang, S., et al., 2022. DESA: a novel hybrid decomposing-ensemble and spatiotemporal attention model for PM2.5 forecasting. Environ. Sci. Pollut. Res. Int. 29, 54150–54166.
Hojjati-Najafabadi, A., et al., 2022. Magnetic-MXene-based nanocomposites for water and wastewater treatment: a review. J. Water Proc. Eng. 47, 102696.
Jiang, T., et al., 2019. Himawari-8/AHI and MODIS aerosol optical depths in China: evaluation and comparison. Rem. Sens. 11.
Ju, Z., He, J.J., 2018. Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection. Anal. Biochem. 550, 1–7.
Juhos, I., et al., 2008. Forecasting of traffic origin NO and NO2 concentrations by support vector machines and neural networks using principal component analysis. Simulat. Model. Pract. Theor. 16, 1488–1502.
Karimian, H., et al., 2012. Assessing urban sustainable development in isfahan. Appl. Mech. Mater. 253–255, 244–248.
Karimian, H., et al., 2019a. Spatio-temporal variation of wind influence on distribution of fine particulate matter and its precursor gases. Atmos. Pollut. Res. 10, 53–64.
Karimian, H., et al., 2016. An improved method for monitoring fine particulate matter mass concentrations via satellite remote sensing. Aerosol Air Qual. Res. 16, 1081–1092.
Karimian, H., et al., 2017. Daily estimation of fine particulate matter mass concentration through satellite based aerosol optical depth. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences IV-4/W2, 175–181.
Karimian, H., et al., 2019b. Evaluation of different machine learning approaches to forecasting PM2.5 mass concentrations. Aerosol Air Qual. Res. 19, 1400–1410.
Karimian, H., et al., 2022. Landscape ecological risk assessment and driving factor analysis in Dongjiang river watershed. Chemosphere 307, 135835.

Khalili, R., et al., 2018. Early-life exposure to PM2.5 and risk of acute asthma clinical encounters among children in Massachusetts: a case-crossover analysis. Environ. Health 17, 25.
Li, Y., et al., 2020. Spatiotemporal analysis of air quality and its relationship with meteorological factors in the Yangtze River Delta. J. Elem. 25, 1059–1075.
Liaw, A., et al., 2002. Classification and Regression with Random Forest, vol. 23. R News.
Lim, H., et al., 2018. AHI/Himawari-8 yonsei aerosol retrieval (YAER): algorithm, validation and merged products. Rem. Sens. 10.
Luo, Z., et al., 2020. PM2.5 concentration estimation using convolutional neural network and gradient boosting machine. J. Environ. Sci. 98, 85–93.
Mansoorianfar, M., et al., 2022. Recent progress on adsorption of cadmium ions from water systems using metal-organic frameworks (MOFs) as an efficient class of porous materials. Environ. Res. 214, 114113.
Mo, Y., et al., 2020. A Novel Framework for Daily Forecasting of Ozone Mass Concentrations Based on Cycle Reservoir with Regular Jumps Neural Networks, vol. 220. Atmospheric Environment.
Mo, Y., et al., 2021. Daily spatiotemporal prediction of surface ozone at the national level in China: an improvement of CAMS ozone product. Atmos. Pollut. Res. 12, 391–402.
Natarajan, R., et al., 2021. Petroleum refinery wastewater treatment using rGo-biochar composite - parametric studies and neural network modeling. Desalination Water Treat. 233, 62–69.
Nguyen, H.-T.T., et al., 2021. Microwave-assisted solvothermal synthesis of bimetallic metal-organic framework for efficient photodegradation of organic dyes. Mater. Chem. Phys. 272, 125040.
Niu, M., et al., 2016. A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2.5 concentration forecasting. Atmos. Environ. 134, 168–180.
Qi, Y., et al., 2019. A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory. Sci. Total Environ. 664, 1–10.
Rao, L., et al., 2022. Lotus seedpods biochar decorated molybdenum disulfide for portable, flexible, outdoor and inexpensive sensing of hyperin. Chemosphere 301, 134595.
Shi, T., et al., 2020. Land use regression modelling of PM2.5 spatial variations in different seasons in urban areas. Ence of the Total Environment 743, 140744.
Shi, Y., et al., 2018. Improving satellite aerosol optical Depth-PM2.5 correlations using land use regression with microscale geographic predictors in a high-density urban context. Atmos. Environ. 190, 23–34.
Siyuan, et al., 2018. Effects of PM2.5 and O3 on human health at a suburban area of beijing, China. J. Environ. Protect. 9.
Su, X., et al., 2019. Support vector machine regression forecasting of O3 concentrations based on wavelet transformation. China Environ. Sci. 39, 3719–3726.
Sujatha, S., et al., 2021a. Extraction of nickel using a green emulsion liquid membrane–Process intensification, parameter optimization and artificial neural network modeling. Chemical Engineering and Processing-Process Intensification 165, 108444.
Sujatha, S., et al., 2022. Parameter Screening, Optimization and Artificial Neural Network Modeling of Cadmium Extraction from Aqueous Solution Using Green Emulsion Liquid Membrane, vol. 25. Environmental Technology & Innovation.
Sujatha, S., et al., 2021b. Conversion of waste cooking oil into value-added emulsion liquid membrane for enhanced extraction of lead: performance evaluation and optimization. Chemosphere 284, 131385.
Sun, J., et al., 2019. Investigating the PM2.5 mass concentration growth processes during 2013-2016 in Beijing and Shanghai. Chemosphere 221, 452–463.
Tran, V.A., et al., 2022a. Metal-organic framework for lithium and sodium-ion batteries: progress and perspectivez. Fuel 319, 123856.
Tran, V.A., et al., 2022b. Metal-organic-framework-derived metals and metal compounds as electrocatalysts for oxygen evolution reaction: a review. Int. J. Hydrogen Energy 47, 19590–19608.
Wang, P., et al., 2019. A hybrid-wavelet model applied for forecasting PM2.5 concentrations in Taiyuan city, China. Atmos. Pollut. Res. 10, 1884–1894.
Wei, W., et al., 2017. Deriving hourly PM2.5 concentrations from himawari-8 AODs over beijing–tianjin–hebei in China. Rem. Sens. 9, 858.
Wu, C., et al., 2018. PM2. 5 concentration prediction using convolutional neural networks. Sci. Surv. Mapp. 43, 68–75.
Yang, Z., et al., 2017. Electricity price forecasting by a hybrid model, combining wavelet transform, ARMA and kernel-based extreme learning machine methods. Appl. Energy 190, 291–305.
Yoshida, M., et al., 2018. Common retrieval of aerosol properties for imaging satellite sensors. J. Meteorol. Soc. Jpn. 96.
Zamani Joharestani, M., et al., 2019. PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere 10.
Zhan, Y., et al., 2017. Spatiotemporal prediction of continuous daily PM2.5 concentrations across China using a spatially explicit machine learning algorithm. Atmos. Environ. 155, 129–139.
Zhang, Y.W., et al., 2017. PM2.5 Prediction Model of BP Neural Network Based on Pearson Correlation Index. Journal of Qingdao University.