Environ Monit Assess (2024) 196:487
https://doi.org/10.1007/s10661-024-12644-9

RESEARCH

Predicting ambient $PM_{2.5}$ concentrations via time series models in Anhui Province, China

Ahmad Hasnain · Muhammad Zaffar Hashmi · Sohaib Khan · Uzair Aslam Bhatti · Xiangqiang Min · Yin Yue · Yufeng He · Geng Wei

Received: 27 January 2024 / Accepted: 19 April 2024 / Published online: 30 April 2024

Abstract Due to rapid expansion in the global economy and industrialization, $PM_{2.5}$ (particles smaller than 2.5 μm in aerodynamic diameter) pollution has become a key environmental issue. The public health and social development directly affected by high $PM_{2.5}$ levels. In this paper, ambient $PM_{2.5}$ concentrations along with meteorological data are forecasted using time series models, including random forest (RF), prophet forecasting model (PFM), and autoregressive integrated moving average (ARIMA) in Anhui province, China. The results indicate that the RF model outperformed the PFM and ARIMA in the prediction of $PM_{2.5}$ concentrations, with cross-validation coefficients of determination R^2, RMSE, and MAE values of 0.83, 10.39 μg/m^3, and 6.83 μg/m^3, respectively. PFM achieved the average results ($R^2=0.71$, RMSE = 13.90 μg/m^3, and MAE = 9.05 μg/m^3), while the predicted results by ARIMA are comparatively poorer ($R^2=0.64$, RMSE = 15.85 μg/m^3, and MAE = 10.59 μg/m^3) than RF and PFM. These findings reveal that the RF model is the most effective method for predicting $PM_{2.5}$ and can be applied to other regions for new findings.

A. Hasnain
Department of Atmospheric and Oceanic Sciences,
Institute of Atmospheric Sciences, Fudan University,
Shanghai 200438, China

M. Z. Hashmi
Institute of Molecular Biology and Biotechnology, The
University of Lahore, Lahore, Pakistan

M. Z. Hashmi
Department of Civil and Environmental Engineering,
Michigan State University 1449 Engineering Research,
East Lansing, MI 48823, USA

M. Z. Hashmi
Department of Environmental Health, Health Services
Academy, Islamabad, Pakistan

S. Khan · X. Min
School of Geography, Nanjing Normal University,
Nanjing 210023, China
e-mail: msohaib.k122@gmail.com

U. A. Bhatti (✉)
School of Information and Communication Engineering,
Hainan University, Haikou, China
e-mail: uzair@hainanu.edu.cn

Y. Yue
Xinjiang Key Laboratory of Oasis Ecology, College
of Geography and Remote Sensing Science, Xinjiang
University, Urumqi, China

Y. He
Key Laboratory of Poyang Lake Wetland and Watershed
Research, Ministry of Education, Jiangxi Normal
University, Nanchang 330022, China

G. Wei
School of Surveying and Geoinformation Engineering,
East China University of Technology, Nanchang 330013,
China

Keywords $PM_{2.5}$ · Random forest · Prophet forecasting model · ARIMA · Time series models

Introduction

People in industrial and manufacturing civilizations are willing to sacrifice the environment to pursue economic progress and development. Future generations' interests are directly damaged by this condition (Bhatti et al., 2023). These days, environmental tightness and economic growth are neither simple nor easy games. According to Shakya et al. (2023), there will be negative and detrimental effects on social development during the initial period of uniform adjustment and production volume reduction. However, in the long run, the economy will grow healthily and effectively because of the relaxation of environmental protection requirements (Shakya et al., 2023).

$PM_{2.5}$, also known as fine particulate matter, is a serious issue left over from the sightless quest of commercial and economic progress. Compared with PM_{10}, $PM_{2.5}$ has more toxic and harmful effects, which can enhance the noxious substances in the air and persist for a long time in the body (Bilal et al., 2021; Hasnain et al., 2022; Zhu et al., 2019). $PM_{2.5}$ causes several diseases such as immune diseases, cardiovascular diseases, respiratory diseases, and tumors (Liu and Sun, 2019; Wu et al., 2023). In recent years, air pollution has attracted wide attention by people. $PM_{2.5}$ has also received vast interest due to its adverse health impacts (He and Huang, 2018). Scholars and researchers have also begun related work. If $PM_{2.5}$ concentration is predicted, the status of air quality can be acknowledged in advance. This helps to control air pollution and plan accordingly (Hasnain et al., 2023; Yang et al., 2024). The major sources of $PM_{2.5}$ are power plants, industrial production, construction activities, and automobile exhaust emissions. These sources contain toxic and poisonous substances such as heavy metals (Ghasempour et al., 2021; Drewil and Al-Bahadili, 2022). Due to the long-range sources of $PM_{2.5}$, it is difficult to detect the primary source, which poses a constant task to its prediction (Guo et al., 2017). Today, artificial intelligence is widely used to generate a large amount of real-time data in modern cities. The major challenge is how to use these low informative and massive data to execute smart city operation monitoring and help the effective process of the city in the new era (Liu et al., 2018).

The time series prediction method is a common and well-known method, which is widely used by scholars and researchers in many fields. It is also used to predict $PM_{2.5}$ concentrations (Lee et al., 2020; Wu et al., 2023). The concentrations of $PM_{2.5}$ were compared with meteorological variables and other contaminants. The Auto-Regressive Integrated Moving Average (ARIMA) was used in the prediction of $PM_{2.5}$ concentrations. However, the model showed low performance due to fewer time series consideration (Zhang et al., 2018).

In recent years, machine learning methods have been widely used in many fields and areas (Wei et al., 2021). Several algorithms have been studied such as support vector and decision trees (Shen et al., 2020). Some researchers and scholars have used them in the prediction of $PM_{2.5}$ concentrations. Chuang et al. (2011) estimated the rise in air pollution using generalized linear mixed models. Lee et al. (2012) used ARIMA model for investigating the future air quality. Song et al. (2014) estimated regional ground-level $PM_{2.5}$ using a geographically weighted regression method. They found that the model was able to elucidate 73.8% of the variability in the concentration of $PM_{2.5}$. Wang et al. (2017) developed a new hybrid Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model to merge several prediction algorithms of support vector machine (SVM) and ARIMA. The literature (Silva et al., 2001) proposes nonparametric and multivariate adaptive regression methods to predict the concentrations of PM_{10} and $PM_{2.5}$ in Santiago. Huang et al. (2018) used a random forest model for $PM_{2.5}$ prediction in Hebei and Shandong and found a strong relationship between $PM_{2.5}$ and chronic diseases (Huang et al., 2018).

There are also several studies that use many approaches for prediction (Cekim, 2020; Akdi et al., 2020; Wu et al., 2021; Dong et al., 2021; Han et al., 2021; Hasnain et al., 2023; Maciąg et al., 2023). Lu et al. (2021) estimated $PM_{2.5}$ concentrations using random forest, support vector regression (SVR), and artificial neural network (ANN) methods in the Yangtze River Delta. They predicted $PM_{2.5}$ concentrations using a hybrid model based on deep learning approaches. The literature (Guo et al., 2018) uses the coupled Lagrangian particle diffusion model system

(FLEXPART-WRF) to predict and measure the estimation of $PM_{2.5}$ concentrations in Xuzhou, China. The authors of this study presented an inverse method to improve and increase the production calculation and mixture ratio estimation of $PM_{2.5}$. Chelani (2018) developed a combined method to estimate the concentrations of $PM_{2.5}$ from environmental variables and aerosol optical depths. Moisan et al. (2018) presented a method based on dynamic multilinear equation to forecast $PM_{2.5}$ in Santiago, Chile. The literature (Fang et al., 2022) proposes a hybrid decomposing-ensemble and spatiotemporal attention (DESA) method for $PM_{2.5}$ prediction. Qiao et al. (2022) developed an air quality forecasting model based on random forest and ant colony algorithm combined with back-propagation neural network (IACA-BPNN) to predict $PM_{2.5}$ and O_3 concentrations in Chengdu city. Feng et al. (2015) predicted the daily average concentrations of $PM_{2.5}$ using a new hybrid model coupled with wavelet transform and trajectory analysis. Zeng et al. (2020) proposed a generalized additive model to forecast $PM_{2.5}$ concentrations combined with meteorological parameters in Chengdu, China. Their results indicate that the model captured 73.9% of the variability in the daily average $PM_{2.5}$ concentrations.

From an air pollution perspective, the air pollution in China is different from the world's air pollution. China is a largest developing country in the world and due to rapid development in industries and transportation, many areas and regions in the country has experienced heavy pollution in recent years. In this study, three time series models including random forest, prophet forecasting model, and ARIMA were used to predict and examine the concentrations of $PM_{2.5}$ for the most polluted areas in China. These models were also used to investigate and forecast the short-term $PM_{2.5}$ concentrations for all the cities of Anhui. This study's main contributions are that it shows the spatial pattern of air quality and offers time-dependent pollution forecasts. It differs from the other research in that it employed multiple forecasting models to predict the concentrations of $PM_{2.5}$ and then compared the results. The paper is organized as follows: In the "Methodology" section, we defined and explained the three methods, data sources and the model's performance metrics used in this study; in the "Results and discussion" section, the results of the fitted models and spatial distribution of $PM_{2.5}$ are presented. "Conclusion" section presents and discusses the conclusion of this paper.

Methodology

Random forest method

With multiple classification and regression tree (CART) integrations, the random forest is a new model (Breiman, 2001; Brokamp et al., 2018). CART consists of three unique qualities. Initially, several trees are created in the original dataset using a bootstrap sample, and then a single tree is developed in CART using all the raw data. Second, the model employs an optimal version to segment the tree nodes. To segment the tree nodes, CART chooses the best option from each predictor. Ultimately, the model's fully developed trees aid in its ability to forecast (Liu et al., 2018). Three training parameters make up the model: max_features (the number of features for the best split; by default, max_features = n_features); min_samples_lea (the minimum sample number for a leaf node; one is the default value); and n_estimators (the number of trees in the forest based on a bootstrap observation sample). Based on the out-of-bag (OOB) calibration error rate, the two crucial parameters (n_estimators and max_features) were optimized and determined to estimate $PM_{2.5}$.

Prophet forecasting method

The prophet forecasting model, developed by Facebook, is a powerful tool for time series analysis, and it takes a short time to fit the model. The model uses the following formula:

$$y(t) = g(t) + s(t) + h(t) + \epsilon t \quad (1)$$

where *y(t)* is the predicted data; *g(t)* and *s(t)* represent seasonality; *h(t)* is the holiday outliers; and ϵ_t is the unexpected error. There are several parameters of the model and the model type can be expected as linear or logistic. The linear model has no maximum or minimum limit set, while the highest and lowest values are specified in the logistic model. The prophet forecasting model takes a Bayesian-based curve fitting method to predict and smooth time series data, which is one of the most distinctive features of the model (Taylor & Letham, 2017). Change points are important parameters in the model and the explicit values of change points can be stated; the model showed the best performance with higher change points. To evaluate the number of change points, the model plots a large value, and then it uses L1 regularization to select few points to use. L1 regularization was used to avoid overfitting. The following equation denotes L1 regularization,

$$L(x,y) \equiv \sum_{i=1}^{n} (y_{i-} h_\theta(x_i))^2 + \lambda \sum_{\mathrm{i=1}}^{\mathrm{n}} |\theta_i| \tag{2}$$

where x and y represent the coordinates of the change points.

$\sum_{i=1}^{n}(y_i - h_\theta(x_i))^2$ denotes the change between original and predicted value squared. To avoid overfitting, $\lambda\sum_{\mathrm{i=1}}^{\mathrm{n}}|\theta_i|$ is used to sustain the balance in weights, where λ indicates how much the weights are disciplined and penalized. The model determines the value of λ based on the estimator's number.

ARIMA method

The ARIMA model contains the autoregressive (AR) and moving average (MA) models with a difference (integration) term. The model was introduced by Box and Jenkins (1976). The seasonal ARIMA can be defined as ARIMA$(p,d,q)(P,D,Q)_s$ where P an p represent the seasonal and non-seasonal degrees of the AR model, Q and q denote the seasonal and non-seasonal degrees of the MA model, and D and d are the seasonal and non-seasonal degrees of difference respectively, where s denotes the seasonal frequency (Anggraeni et al., 2015). The ARIMA model uses the following formula:

$$\Upsilon_s^D \Upsilon^d \Lambda_P(B^s)\lambda_p(B)Y_t = \Pi_Q(B^s)\pi_q(B)\xi_t \tag{3}$$

where Y_t and ξ_t indicate the time and error series, $\Lambda_P(B^s)$ and $\Pi_P(B^s)$ represent the seasonal autoregressive and moving average polynomials, $\lambda_p(B)$ and $\pi_p(B)$ denote the non-seasonal autoregressive and moving average polynomials, $\Upsilon_s^D = (1 - B^s)^D$ indicate the seasonal, and $\Upsilon^d = (1 - B)^d$ indicate the non-seasonal machinists, respectively. Here, the lag operator $(B^i Y_t = Y_{t-i})$ is B. The series should be stationary for determining the superlative ARIMA model. The difference operations determine the levels of differencing for d and D. Moreover, the values of P, Q, p, and q are selected as the optimum model (Athanasopoulos et al., 2011). Finally, the model is carried out a white noise test to determine whether the residuals will be generally dispersed (Molina et al., 2018). Hyndman and Khandakar (2008) discussed in detail the steps of the ARIMA model.

Data sources

Anhui Province, the provincial administrative region of China, is located in the middle and east (between 114°54′–119°37′ E, 29°41′–34°38′ N) (Fig. 1). Hefei, the capital of Anhui Province, is located in the Yangtze River Delta region. The province is bordered by Jiangsu in the east, Hubei and Henan in the west, Shandong in the north, Jiangxi in the south, and Zhejiang in the southeast. According to the latest census data, Anhui has a great population, an approximately 61.03 million, ranking 9th in the country. The province has diverse and complex landforms, with plains, hills, and mountains. Anhui is subjugated by highlands and mountains, spanning the three most important water systems of the Yangtze River, the Xia'an River, and the Huai River, with several lakes. Anhui Province has rapidly industrialized, especially the districts surrounding its capital, Hefei, and other large cities, including as Wuhu and Ma'anshan. Due to the expansion of industry, manufacturers are now emitting more particulate matter, sulfur dioxide, nitrogen oxides, heavy metals, and sulfur dioxide. As a result, variations in air quality patterns exist throughout different regions, which will eventually aid in our ability to study more effectively. In recent years, the province has seen rapid growth and development in industry and manufacturing sectors. This has led to extreme and severe air pollution issues, especially in the capital city of Anhui Province, Hefei.

The daily average data of $PM_{2.5}$ were collected through 68 monitoring stations ranging from 1 January 2018 to 31 December 2023 along with five meteorological parameters including temperature (TEMP),

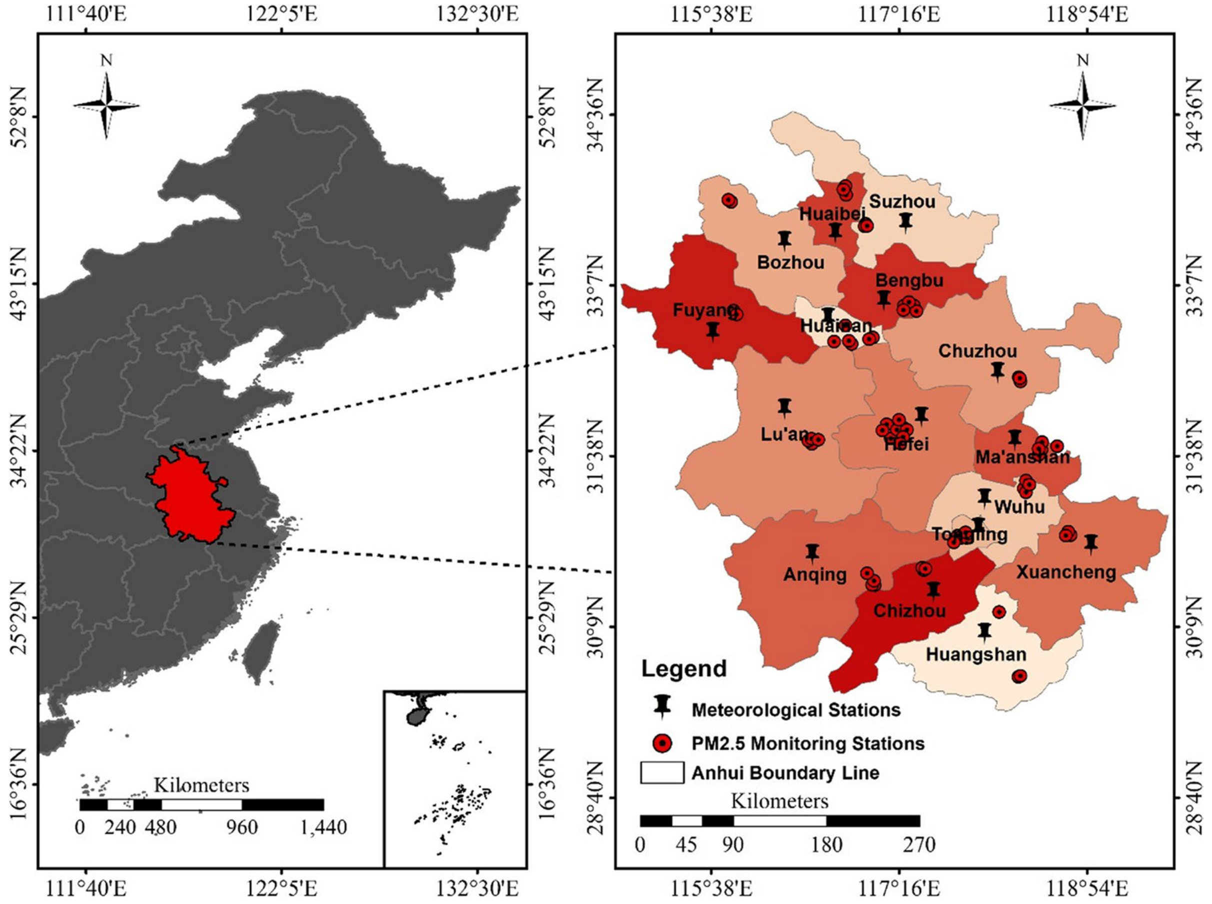


Fig. 1 Location of the $PM_{2.5}$ and meteorological monitoring sites in Anhui

relative humidity (RH), wind speed (WS), wind direction (WD), and precipitation (PCPN) through 16 monitoring stations for the same window of time to build the models (Table 1, Fig. 1). $PM_{2.5}$ data were downloaded from the China Weather Website Platform (CNEMC, 2019), while the meteorological data were retrieved from the NASA meteorological data service (https://power.larc.nasa.gov). The basic statistics for the meteorological and $PM_{2.5}$ data are presented in Table 1. In general, approximately 80% and 20% data are considered, respectively, as training and test data. In this paper, we divided the data into three sets, entirely, 3-year dataset and yearly to evaluate the forecast accuracy of the three models. The actual and predicted values were also compared at the municipal level in this work.

Model performance metrics

In this work, we used the three statistical metrics to evaluate the performance of the models, which are determination coefficient R^2, root mean squared error (RMSE), and mean absolute error (MAE). These metrics defined as

$$R^2 = \sum_{i=1}^{n}(P_i - \overline{M})^2 / \sum_{i=1}^{n}(M_i - \overline{M})^2 \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|M_i - P_i\right|^2} \tag{5}$$

Table 1 Basic statistics for the $PM_{2.5}$ and meteorological parameters during the study period

Pollutant	2018			2019			2020			2021			2022			2023		
	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min
$PM_{2.5}$ (μg/m^3)	177.15	48.57	11.57	152.23	45.94	7.68	130.93	38.40	8.92	130.21	35.08	5.66	141.08	34.69	6.71	194.88	35.65	6.54
TEMP (°C)	31.37	16.38	−6.10	32.22	16.78	−1.27	30.13	16.42	−5.44	30.02	16.88	−5.36	30.17	16.62	−4.45	31.12	16.76	−5.21
RH (%)	94.50	77.23	42.90	92.81	71.87	41.85	94.34	78.54	52.10	93.51	77.87	50.75	92.55	75.85	49.38	93.61	76.34	47.62
WS (m/s)	9.01	3.26	1.34	7.51	3.05	1.31	9.21	3.15	1.33	8.41	3.32	1.47	8.55	3.22	1.41	8.13	3.16	1.44
WD (°)	306.39	151.43	23.19	320.36	145.46	24.87	318.21	156.36	32.89	341.43	162.42	22.67	332.53	158.23	26.53	324.83	152.43	28.63
PCPN (mm)	79.28	3.84	0	42.16	2.76	0	63.53	5.41	0	38.21	3.86	0	44.53	4.13	0	52.54	3.94	0

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|M_i - P_i\right| \tag{6}$$

where M and P are the observed and predicted values and n denotes the number of samples. The smaller values of RMSE, MAE, and values of R^2 closest to one indicate that the prediction accuracy of the model is higher.

Results

RF performance

The cross-validation (CV), determination coefficient R^2, RMSE, and MAE were used to estimate the model's performance. The predicted results of the three models are presented in Table 2. The results indicate that the RF model outperformed the PFM and ARIMA in the prediction of $PM_{2.5}$ in Anhui Province. RF predicts with an overall CV R^2 of 0.83, RMSE value of 10.39 μg/m^3, and MAE value of 6.83 μg/m^3, respectively (Fig. 2). In a 3-year dataset prediction, by RF, the values of R^2, RMSE, and MAE were 0.81, 9.96 μg/m^3, and 6.99 μg/m^3, respectively (Fig. 2). Figure 5 is showing the yearly comparison between the actual and predicted $PM_{2.5}$ of the three models. Compared with an overall CV, the value of R^2 is slightly lower, while the value of RMSE is better in the second time frame. In a yearly prediction, the predicted values of R^2, RMSE, and MAE were 0.80, 10.54 μg/m^3, and 7.60 μg/m^3, respectively. It should be noted that the predicted value of R^2 (0.83) is greater in the entire dataset than that of the 3 years' time frame and a yearly prediction, but the values of RMSE and MAE are poorer in this period. Moreover, RF has the poorer R^2, RMSE, and MAE values compared with an overall CV or the half dataset prediction. Figure 6 shows the comparison results for the 16 cities of Anhui.

PFM performance

Compared with the RF model, PFM showed relatively a poorer performance in the prediction of $PM_{2.5}$ concentrations in Anhui. The results indicate that PFM predicted with an overall CV R^2, RMSE, and MAE values of 0.71, 13.90 μg/m^3, and 9.05 μg/m^3 (Fig. 3).

Table 2 The performance of different models for all the cities of Anhui

City	RF			PFM			ARIMA			Bhatti et al. (2021)			Hasnain et al. (2022)		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Hefei	0.82	9.76	7.81	0.76	11.97	9.51	0.63	15.05	11.73	0.60	14.87	12.22	0.56	19.13	16.93
Suzhou	0.79	12.56	10.31	0.77	16.40	14.69	0.69	17.97	13.26	0.58	19.22	16.64	0.55	21.43	18.36
Huaibei	0.81	15.15	13.19	0.81	15.19	13.25	0.77	16.69	13.51	0.63	18.45	16.21	0.61	19.05	17.63
Bozhou	0.83	13.60	11.07	0.82	15.52	13.81	0.75	18.59	14.50	0.61	18.58	17.82	0.51	20.43	17.42
Fuyang	0.80	14.46	12.57	0.82	16.67	14.21	0.72	20.72	15.90	0.55	20.77	16.35	0.49	23.43	18.42
Bengbu	0.88	9.18	6.90	0.75	12.77	10.06	0.51	17.86	13.81	0.67	16.24	13.03	0.42	21.64	16.61
Huainan	0.81	12.67	10.43	0.81	14.15	11.88	0.45	24.21	18.69	0.46	17.35	14.52	0.33	28.53	23.52
Chuzhou	0.81	10.40	8.73	0.79	11.08	8.95	0.54	16.09	12.29	0.57	15.19	13.73	0.44	21.64	17.29
Luan	0.85	10.84	9.17	0.81	12.14	9.80	0.57	17.69	13.48	0.52	16.59	12.53	0.39	21.75	18.53
Anqing	0.85	10.93	8.92	0.75	13.94	11.40	0.60	17.66	12.79	0.61	17.99	13.11	0.48	20.64	17.47
Tongling	0.82	9.72	8.53	0.59	14.41	12.13	0.41	17.33	13.34	0.49	19.21	14.52	0.34	21.63	18.74
Wuhu	0.85	10.60	9.16	0.67	14.75	12.19	0.17	24.39	17.65	0.34	18.65	15.13	0.27	28.42	24.75
Maanshan	0.85	9.46	8.09	0.56	15.57	12.69	0.39	18.71	14.14	0.41	19.11	15.14	0.31	22.74	19.53
Xuancheng	0.74	10.60	9.12	0.52	15.29	12.71	0.33	21.37	17.64	0.48	18.34	15.52	0.29	24.63	21.36
Chizhou	0.80	10.91	9.24	0.73	13.05	11.23	0.53	16.96	13.28	0.33	17.14	14.46	0.43	20.63	17.53
Huangshan	0.77	6.52	5.38	0.46	12.01	10.09	0.40	15.66	13.61	0.39	16.56	13.11	0.36	20.62	17.42

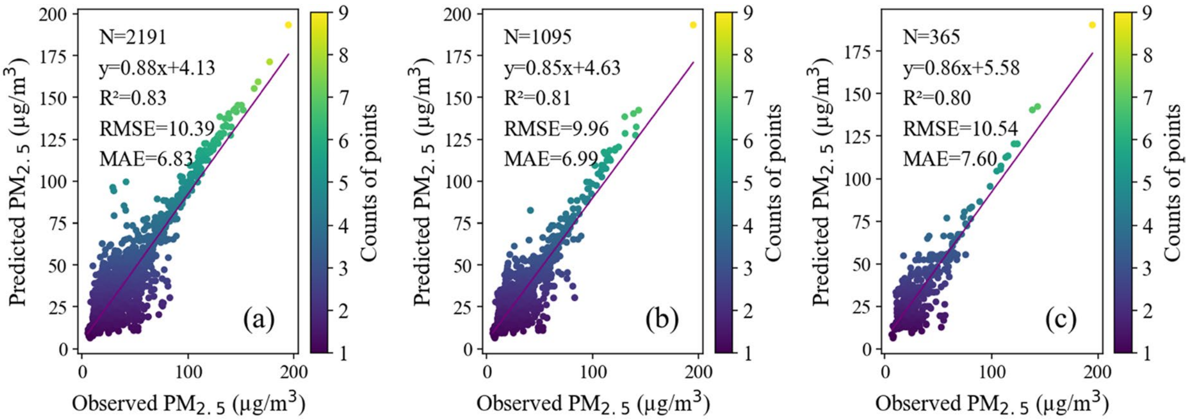


Fig. 2 Validation between predicted and actual $PM_{2.5}$ concentration by random forest model; **a** overall CV, **b** 3-year dataset, and **b** yearly prediction

Table 2 and Fig. 7 present the performance of the PFM model at the municipal level. In the 3-year period prediction, the model predicts with R^2 value of 0.70, RMSE value of 12.83 μg/m^3, and MAE value of 8.39 μg/m^3. The cross-validation R^2 slightly decreases in this period, while the values of RMSE and MAE are demonstrating higher performance compared with an overall CV. Previously, Shen et al. (2020) used the PFM model for predicting air pollution and compared with the said study, the performance of the current work is higher. Moreover, the R^2, RMSE, and MAE values between the actual and predicted $PM_{2.5}$ are 0.74, 12.51 μg/m^3, and 8.69 μg/m^3, respectively, in a yearly prediction (Fig. 3). PFM achieved the best performance during this window of time. A small difference in the values of R^2, RMSE, and MAE can be seen between the 3-year period prediction and a yearly.

ARIMA performance

Figure 4 shows the validation between the actual and predicted $PM_{2.5}$ by the ARIMA model. The results indicate that by ARIMA, an overall CV R^2 is 0.64, which is lower than that of RF and PFM, while the values of RMSE and MAE are 15.85 μg/m^3 and 10.59 μg/m^3, respectively. The predicted results for all the cities of Anhui are listed in

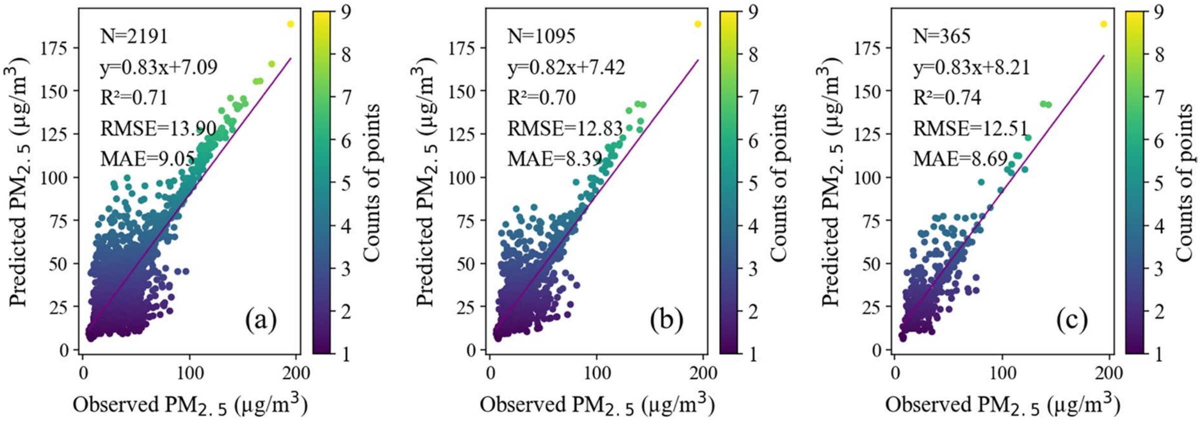


Fig. 3 Validation between predicted and actual $PM_{2.5}$ concentration by prophet forecasting model; **a** overall CV, **b** 3-year dataset, and **b** yearly prediction

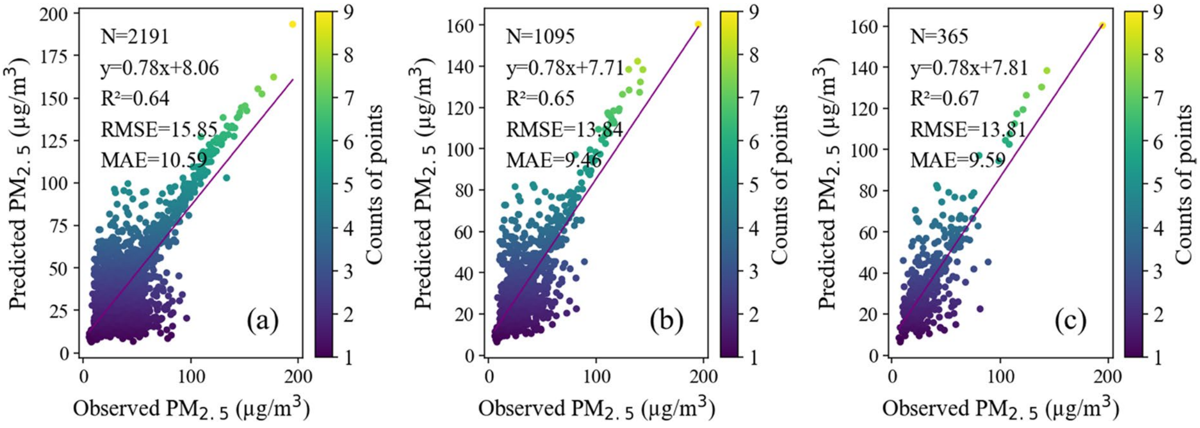


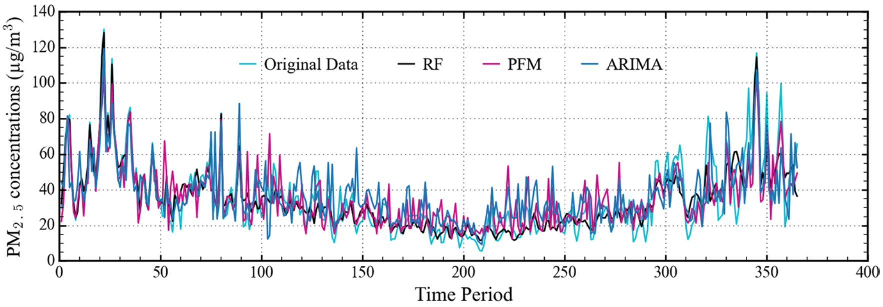
Fig. 4 Validation between predicted and actual PM$_{2.5}$ concentration by ARIMA model; **a** overall CV, **b** 3-year dataset, and **b** yearly prediction

Table 2. Compared with an overall CV, ARIMA showed a higher performance in half dataset prediction. The predicted R^2, RMSE, and MAE values for ARIMA are 0.65, 13.84 μg/m^3, and 9.46 μg/m^3, respectively, in the corresponding period (Figs. 4, (8). A slight increase can be seen in the value of R^2, while the difference between the values of RMSE and MAE is larger compared with an overall CV. Moreover, ARIMA achieved a better performance in yearly prediction compared with the overall CV and 3-year period prediction (Fig. 4). The results indicate that the R^2, RMSE, and MAE values between the actual and predicted PM$_{2.5}$ are 0.67, 13.81 μg/m^3, and 9.59 μg/m^3, respectively. It can be noted that there is a slight difference in the values of RMSE and MAE between the 3-year period prediction and the yearly prediction. ARIMA has the best R^2 and RMSE values in a yearly prediction, while it has a relatively lower MAE value in this period.

Comparisons of the models

Here, we compare the results of the three models. As shown in Fig. 2, the performance of the RF model is higher than that of the PFM and ARIMA models. For example, by RF, an overall CV R^2 is 0.83, while these values are 0.71 and 0.64, respectively, for PFM and ARIMA (Figs. 3, 4). Similarly, the values of RMSE and MAE are also demonstrating better results of RF than that of PFM and ARIMA. Figure 5 is showing the comparison between the actual and predicted PM$_{2.5}$ for the three models. Lu et al. (2021) developed random forest (RF), support vector regression (SVR), and artificial neural network (ANN) for predicting PM$_{2.5}$ concentrations in the Yangtze River Delta region. Their results showed that by RF, the value of cross-validation R^2 was 0.77, by SVR it was 0.703, while by AAN the predicted value of R^2 was 0.702. Ye (2019) presented an ARIMA-PFM model in the prediction of PM concentrations. Another study

Fig. 5 Comparison between actual and predicted PM$_{2.5}$ concentration using different models



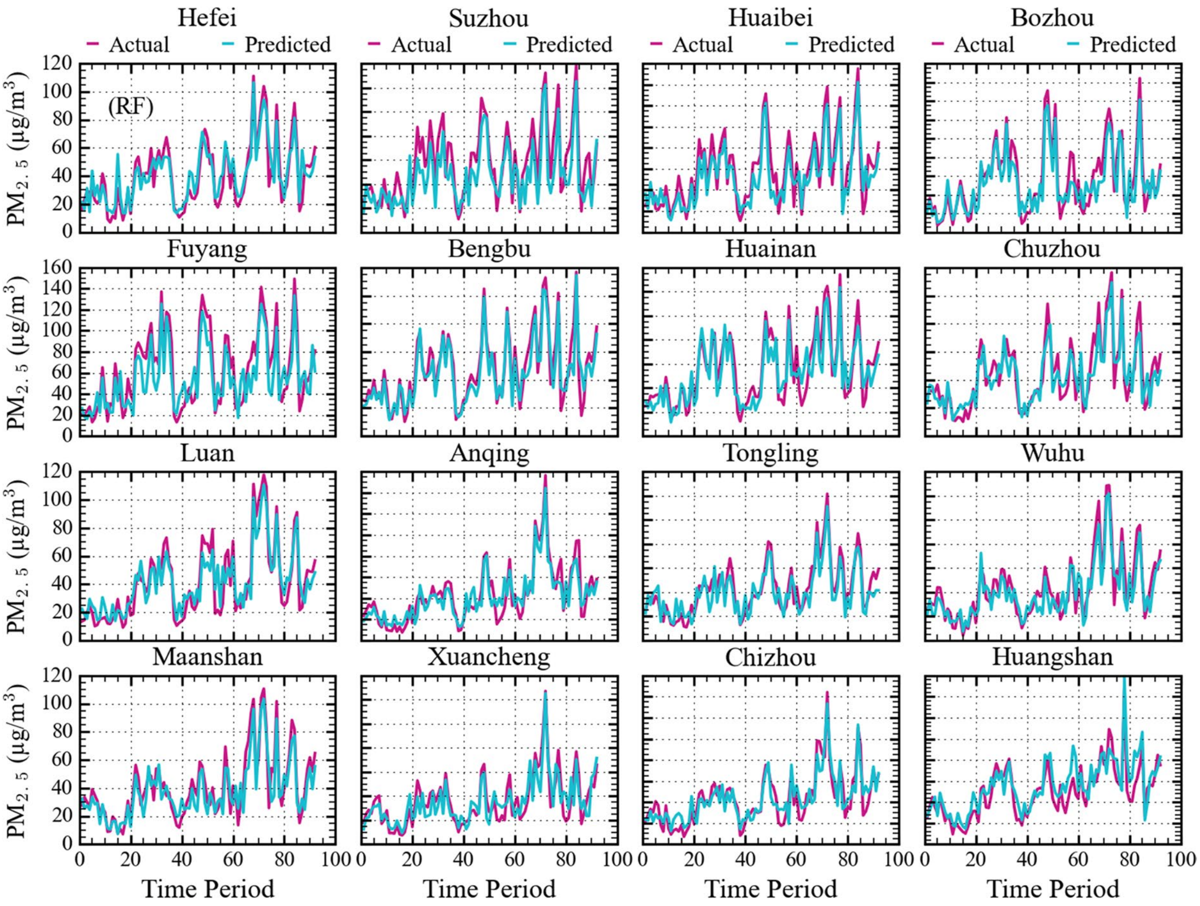


Fig. 6 Comparison between actual and predicted $PM_{2.5}$ using random forest model

documented by Chang et al. (2020) presented a deep learning approach in the prediction of air pollution. However, compared with the said studies, our models are showing higher performance.

The prediction results for all the cities of Anhui are presented in Table 2. The R^2 values by RF for all the cities of Anhui are ranged from 0.74 to 88. RF has the best R^2 value for the Bengbu city, while it has the worst R^2 value for Xuancheng (Table 2, Fig. 6). The RMSE values ranged from 6.52 to 15.15 μg/m^3, while the MAE values ranged from 5.38 to 13.19 μg/m^3 for the RF model. It should be noted that RF predicts the best RMSE value for the Huangshan city, while the worst for the Huaibei city. Although the performance of the RF model differed slightly for all the cities of Anhui, the overall RF's stability was good. Small fluctuations can be seen in the values of R^2, RMSE, and MAE.

PFM is also showing the best performance in the prediction of $PM_{2.5}$. However, compared with the RF model, PFM has relatively poor performance. By PFM, the R^2, RMSE, and MAE values for all the cities of Anhui were 0.46–82, 11.08–16.67 μg/m^3, and 8.95–14.69 μg/m^3, respectively (Table 2). By PFM, the fluctuations in the values of R^2 were greater than those of the RF model. PFM achieved the best prediction results for Hefei, Chuzhou, and Luan with RMSE and MAE values, while it predicts the best R^2 values for Bozhou and Fuyang. Compared with the RF and PFM models, ARIMA has low accuracy (Table 2). The results indicate that by ARIMA, the R^2 between the actual and predicted $PM_{2.5}$ values ranged from 0.17 to 0.77, RMSE ranged from 15.05 to 24.39 μg/m^3, and MAE values ranged from 11.73 to 18.69 μg/m^3 for all the cities of Anhui (Table 2). ARIMA has the best R^2 and

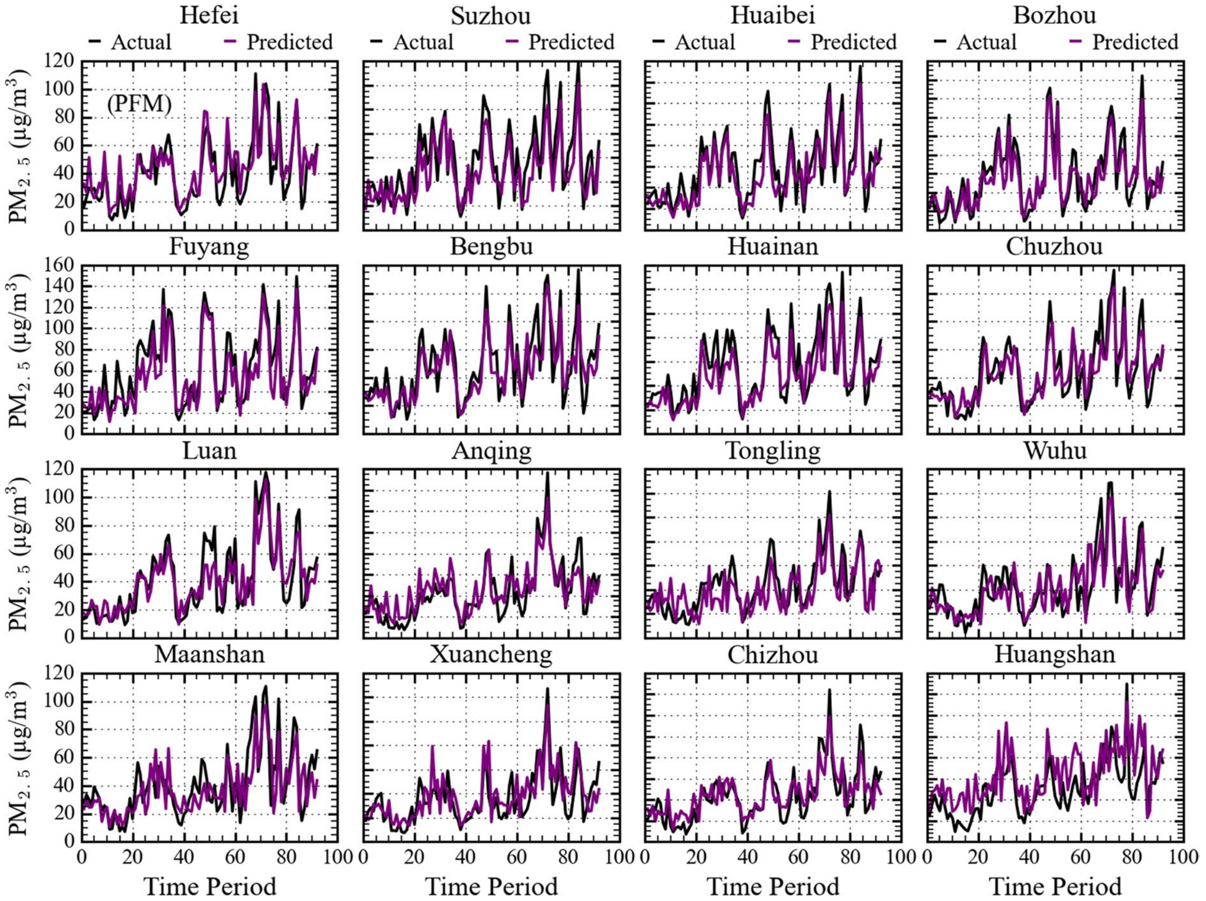


Fig. 7 Comparison between actual and predicted $PM_{2.5}$ using prophet forecasting model

RMSE values for Huaibei and Hefei respectively, while it has the worst R^2 and RMSE values for Wuhu and Huainan, respectively. Overall, the comparison analysis indicates that the RF model outperformed the PFM and ARIMA in the prediction of $PM_{2.5}$ concentrations (Fig. 5).

The present models' performance was also evaluated against the results of previous research, as reported by Bhatti et al. (2021) and Hasnain et al. (2022) (Table 2). Our models exhibit high accuracy when compared to the previously mentioned investigations, as indicated by the findings. All three of the chosen models performed better than the comparable approaches, which had poor performance (Table 2) (Figs. 7 and 8).

Spatial distributions of $PM_{2.5}$

The annual concentrations of $PM_{2.5}$ are shown in Fig. 9. The concentrations of ambient $PM_{2.5}$ were higher in the central and northern parts of Anhui, especially in Hefei, Suzhou Lu'an, Chuzhou, Bengbu, Huaibei, Bozhou, Fuyang, and Huainan, while the southern areas had lower concentrations (Fig. 9). The areas with higher concentrations are mainly concentrated in the industrial and economically developed areas. The annual average concentration of $PM_{2.5}$ from 2018 to 2023 was 39.72 μg/m^3. According to the obtained results, the concentration levels of $PM_{2.5}$ continuously decreased during the study period, while a slight increase was observed in 2023 (Table 1). The

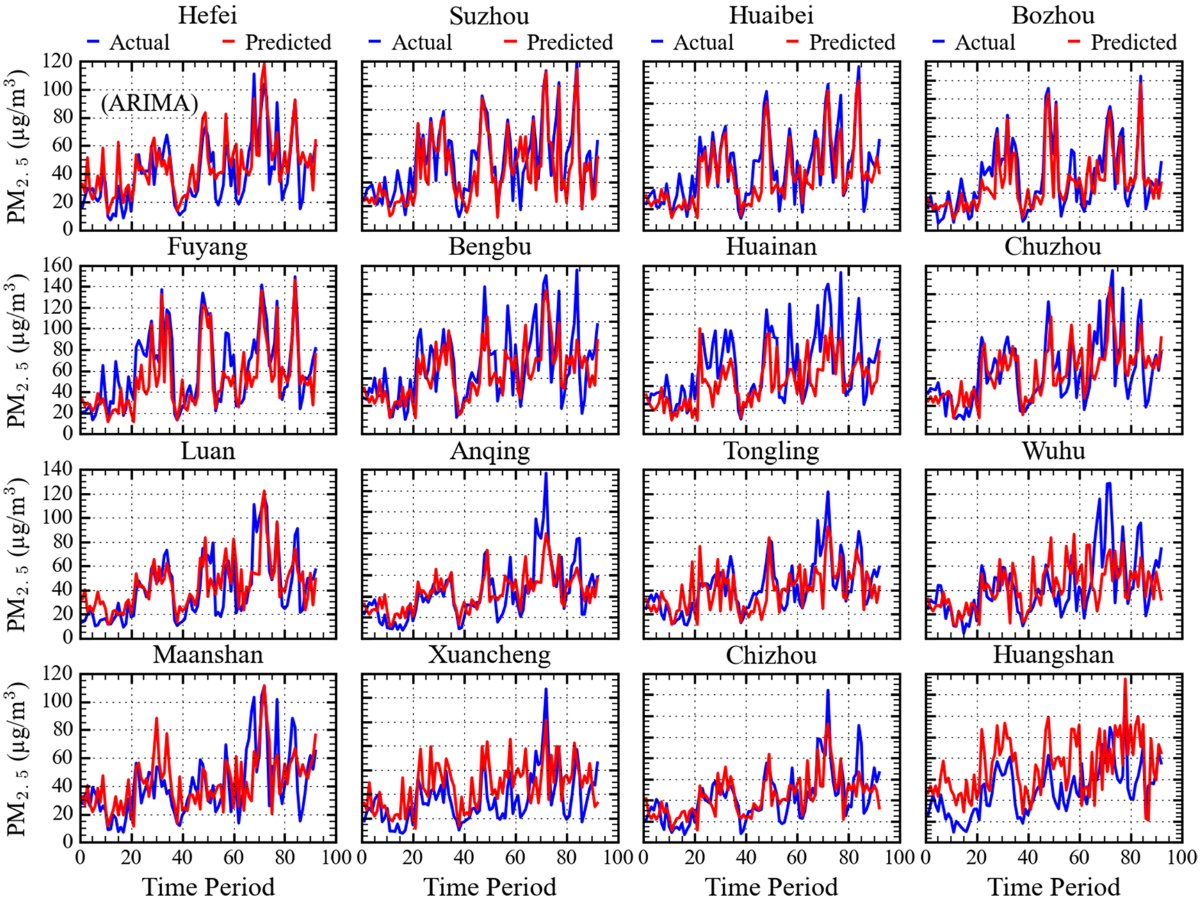


Fig. 8 Comparison between actual and predicted $PM_{2.5}$ using ARIMA model

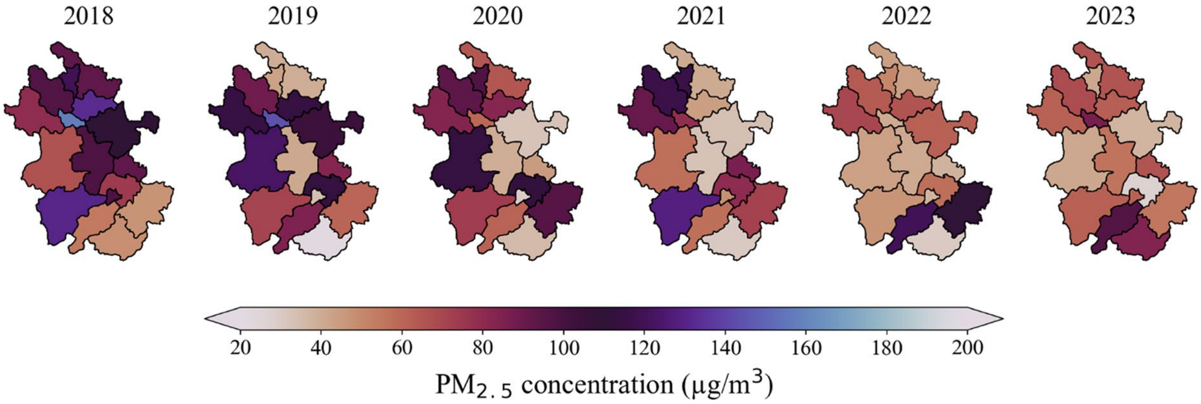


Fig. 9 Spatial distributions of $PM_{2.5}$ concentrations in Anhui

reduction in the levels of $PM_{2.5}$ was due to the strict control measures implemented by the government of China (Hasnain et al., 2023).

Discussion

The findings from this study underscore the effectiveness of the Random Forest (RF) model over the Prophet Forecasting Model (PFM) and Autoregressive Integrated Moving Average (ARIMA) in forecasting $PM_{2.5}$ concentrations in Anhui Province, China. The superior performance of the RF model, as indicated by its higher R^2 value and lower RMSE and MAE scores, suggests that it can more accurately capture the complex relationships and patterns inherent in environmental data influenced by various factors, including meteorological conditions and industrial activities.

The moderate performance of the PFM and the relatively poorer outcomes of the ARIMA model highlight the challenges and limitations associated with applying time series analysis to environmental data. The variability and complexity of such data might not be fully accounted for by models like ARIMA, which are typically more suited to linear time series data without complex interactions.

Given the critical importance of accurately forecasting $PM_{2.5}$ levels due to their significant health and social implications, the results of this study advocate for the adoption of more sophisticated machine learning techniques like RF in environmental monitoring and policymaking. Such approaches can enhance the precision of pollution forecasts, thereby facilitating more effective public health interventions and environmental management strategies.

Conclusions

In the current study, we used three time series models including random forest, prophet forecasting, and ARIMA to predict ambient $PM_{2.5}$ concentrations in Anhui Province. The results indicate that the RF model outperformed the PFM and ARIMA in the prediction of ambient $PM_{2.5}$ concentrations. The predicted results at the municipal level also showed the efficiency of the RF model. The performance of PFM was relatively poorer than that of RF. Compared with the RF and PFM methods, ARIMA showed low performance. Moreover, the concentration levels of ambient $PM_{2.5}$ decreased from 2018 to 2022, while a slight increase was seen in 2023, in Anhui. The present study concludes that the RF model is the most effective and powerful method for predicting ambient $PM_{2.5}$ concentrations and it can be applied to other regions for new findings.

Acknowledgements The authors are thankful to the kind and precious suggestions of Prof. Dr. Muhammad Zaffar Hashmi.

Author contribution Ahmad Hasnain: conceptualization; methodology; data curation; formal analysis; writing—original draft; writing—review and editing; validation; visualization. Muhammad Zaffar Hashmi: supervision, conceptualization, resources, writing—review and editing. Sohaib Khan: validation, investigation, data curation, writing—review and editing. Uzair Aslam Bhatti: supervision, investigation, conceptualization, data curation, writing—review and editing. Xiangqiang Min: data curation, formal analysis, writing—review and editing. Yin Yue: data curation, validation, writing—review and editing. Yufeng He: data curation, writing—review and editing. Geng Wei: data curation; writing—review and editing; validation. All authors read and approved the final manuscript.

Data availability Not applicable.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent to publish Not applicable.

Competing interests The authors declare no competing interests.

References

Akdi, Y., Okkaoglu, Y., Golveren, E., & Yucel, M. E. (2020). Estimation and forecasting of PM10 air pollution in Ankara via time series and harmonic regressions. *International Journal of Environmental Science and Technology, 17*, 3677–3690. https://doi.org/10.1007/s13762-020-02705-0

Anggraeni, W., Vinarti, R. A., & Kurniawati, Y. D. (2015). Performance comparisons between arima and arimax method in moslem kids clothes demand forecasting: Case study. *Procedia Computer Science, 72*, 630–637.

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting, 27*, 822–844.

Bhatti, U. A., Yan, Y., Zhou, M., Ali, S., Hussain, A., Qingsong, H., et al. (2021). Time series analysis and forecasting of air pollution particulate matter (PM2.5): An SARIMA and factor analysis approach. *IEEE Access, 9*, 41019–41031. https://doi.org/10.1109/access.2021.3060744

Bhatti, U. A., Marjan, S., Wahid, A., Syam, M. S., Huang, M., Tang, H., & Hasnain, A. (2023). The effects of socioeconomic factors on particulate matter concentration in China's: New evidence from spatial econometric model. *Journal of Cleaner Production, 417*, 137969. https://doi.org/10.1016/j.jclepro.2023.137969

Bilal, M., Mhawish, A., Nichol, J. E., Qiu, Z., Nazeer, M., Ali, M. A., et al. (2021). Air pollution scenario over Pakistan: characterization and ranking of extremely polluted cities using long-term concentrations of aerosols and trace gases. *Remote Sensing of Environment, 264*, 112617. https://doi.org/10.1016/j.rse.2021.112617

Box, G., & Jenkins, G. (1976). *Time series analysis: Forecasting and control*. Holden-Day.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Brokamp, C., Jandarov, R., Hossain, M., & Ryan, P. (2018). Predicting daily urban fine particulate matter concentrations using a random forest model. *Environmental Science and Technology, 52*, 4173–4179.

Cekim, H. O. (2020). Forecasting PM10 concentrations using time series models: A case of the most polluted cities in Turkey. *Environmental Science and Pollution Research, 27*, 25612–25624. https://doi.org/10.1007/s11356-020-08164-x

Chang, Y. S., Abimannan, S., Chiao, S. T., Lin, C. Y., & Huang, Y. P. (2020). An ensemble learning based hybrid model and framework for air pollution forecasting. *Environmental Science and Pollution Research, 27*, 38155–38168. https://doi.org/10.1007/s11356-020-09855-1

Chelani, A. B. (2018). Estimating PM2.5 concentration from satellite derived aerosol optical depth and meteorological variables using a combination model. *Atmospheric Pollution Research*

Chuang, Y. H., Mazumdar, S., Park, T., Tang, G., Arena, V. C., & Nicolich, M. J. (2011). Generalized linear mixed models in time series studies of air pollution. *Atmospheric Pollution Research, 2*, 428–435.

CNEMC (2019). China national environmental monitoring centre. http://www.cnemc.cn/. Accessed 8 Aug 2019.

Dong, Y., Zhang, C., Niu, M., Wang, S., & Sun, S. (2021). Air pollution forecasting with multivariate interval decomposition ensemble approach. *Atmospheric Pollution Research, 12*, 101230. https://doi.org/10.1016/j.apr.2021.101230

Drewil, G. I., & Al-Bahadili, R. J. (2022). Air pollution prediction using LSTM deep learning and metaheuristics algorithms. *Measurement Sensors, 24*, 100546. https://doi.org/10.1016/j.measen.2022.100546

Fang, S., Li, Q., Karimian, H., Liu, H., & Mo, Y. (2022). DESA: A novel hybrid decomposing-ensemble and spatiotemporal attention model for PM2.5 forecasting. *Environmental Science and Pollution Research, 29*, 54150–54166.

Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., & Wang, J. (2015). Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment, 107*, 118–128.

Ghasempour, F., Sekertekin, A., & Kutoglu, S. H. (2021). Google Earth Engine based spatio-temporal analysis of air pollutants before and during the first wave COVID-19 outbreak over Turkey via remote sensing. *Journal of Cleaner Production, 319*, 128599.

Guo, Y., Tang, Q., Gong, D. Y., & Zhang, Z. (2017). Estimating ground-level PM2.5 concentrations in Beijing using a satellite-based geographically and temporally weighted regression model. *Remote Sensing of Environment, 198*, 140–149.

Guo, L., et al. (2018). Improving PM2.5 forecasting and emission estimation based on the Bayesian optimization method and the coupled FLEXPART-WRF model. *Atmosphere, 9*, 428.

Han, Y., Lam, J. C. K., Li, V. O., & Reiner, D. (2021). A Bayesian LSTM model to evaluate the effects of air pollution control regulations in Beijing, China. *Environmental Science & Policy, 11*, 26–34. https://doi.org/10.1016/j.envsci.2020.10.004

Hasnain, A., Sheng, Y., Hashmi, M. Z., Bhatti, U. A., Hussain, A., Hameed, M., Marjan, S., Bazai, S. U., Hossain, M. A., Sahabuddin, M., Wagan, R. A., & Zha, Y. (2022). Time series analysis and forecasting of air pollutants based on prophet forecasting model in Jiangsu Province, China. *Frontiers in Environmental Science, 10*, 945628. https://doi.org/10.3389/fenvs.2022.945628

Hasnain, A., Sheng, Y., Hashmi, M. Z., Bhatti, U. A., Ahmed, Z., & Zha, Y. (2023). Assessing the ambient air quality patterns associated to the COVID-19 outbreak in the Yangtze River Delta: A random forest approach. *Chemosphere, 314*, 137638. https://doi.org/10.1016/j.chemosphere.2022.137638

He, Q., & Huang, B. (2018). Satellite-based mapping of daily high-resolution ground PM2.5 in China via space-time regression modeling. *Remote Sensing of Environment, 206*, 72–83. https://doi.org/10.1016/j.rse.2017.12.018

Huang, K., Xiao, Q., Meng, X., Geng, G., Wang, Y., Lyapustin, A., Gu, D., & Liu, Y. (2018). Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China plain. *Environmental Pollution, 242*, 675–683.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast Package for R. *The Journal of Statistical Software, 27*, 1–22.

Lee, M. H., Rahman, N. H. A., Latif, M. T., Nor, M. E., & Kamisan, N. A. B. (2012). Seasonal ARIMA for forecasting air pollution index: A case study. *American Journal of Applied Sciences, 9*, 570–578.

Lee, M., Lin, L., Chen, C. Y., Tsao, Y., et al. (2020). Forecasting air quality in Taiwan by using machine learning. *Science and Reports, 10*, 4153. https://doi.org/10.1038/s41598-020-61151-7

Liu, D., & Sun, K. (2019). Short-term PM2.5 forecasting based on CEEMD-RF in five cities of China. *Environmental Science and Pollution Research, 26*, 32790–32803. https://doi.org/10.1007/s11356-019-06339-9

Liu, Y., Cao, G., Zhao, N., Mulligan, K., & Ye, X. (2018). Improve ground-level PM2.5 concentration mapping using a random forests-based geostatistical approach. *Environmental Pollution, 235*, 272–282.

Lu, D., Mao, W., Zheng, L., Xiao, W., Zhang, L., & Wei, J. (2021). Ambient PM2.5 estimates and variations during COVID-19 Pandemic in the Yangtze River delta using machine learning and big data. *Remote Sens, 13*, 1423. https://doi.org/10.3390/rs13081423

Maciąg, P. S., Bembenik, R., Piekarzewicz, A., et al. (2023). Effective air pollution prediction by combining time series decomposition with stacking and bagging ensembles of evolving spiking neural networks. *Environ Model Soft, 170*, 105851. https://doi.org/10.1016/j.envsoft.2023.105851

Moisan, S., Herrera, R., & Clements, A. (2018). A dynamic multiple equation approach for forecasting PM2.5 pollution in Santiago. *Chile. Int J Forecast, 34*, 566–581.

Molina, L. L., Angon, E., Garcıa, A., Moralejo, R. H., Caballero-Villalobos, J., & Perea, J. (2018). Time series analysis of bovine venereal diseases in La Pampa, Argentina. *PloS one, 13*, 1–17.

Qiao, D. W., Yao, J., Zhang, J. W., Li, X. L., Mi, T., & Zeng, W. (2022). Short-term air quality forecasting model based on hybrid RF-IACABPNN algorithm. *Environmental Science and Pollution Research, 29*, 39164–39181. https://doi.org/10.1007/s11356-021-18355-9

Shakya, D., Deshpande, V., Goyal, M. K., & Agarwal, M. (2023). PM2.5 air pollution prediction through deep learning using meteorological, vehicular, and emission data: A case study of New Delhi India. *Journal of Cleaner Production, 427*, 139278. https://doi.org/10.1016/j.jclepro.2023.139278

Shang, Z., Deng, T., He, J., & Duan, X. (2019). A novel model for hourly PM2.5 concentration prediction based on CART and EELM. *Science of the Total Environment, 651*, 3043–3052.

Shen, J., Valagolam, D., & McCalla, S. (2020). Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants (PM2.5, PM10, O3, NO2, SO2, CO) in Seoul. *South Korea. PeerJ, 8*, e9961. https://doi.org/10.7717/peerj.9961

Silva, C., Perez, P., & Trier, A. (2001). Statistical modelling and prediction of atmospheric pollution by particulate material: Two nonparametric approaches. *Environmetrics, 12*(2), 147–159.

Song, W., Jia, H., Huang, J., & Zhang, Y. (2014). A satellite-based geographically weighted regression model for regional PM2.5 estimation over the Pearl River Delta region in China. *Remote Sensing of Environment, 154*, 1–7.

Taylor, S. J., & Letham, B. (2017). Forecasting at scale. Am. *Statistician, 72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Wang, P., Zhang, H., Qin, Z., & Zhang, G. (2017). A novel hybrid-Garch model based on ARIMA and SVM for PM2.5 concentrations forecasting. *Atmospheric Pollution Research, 8*, 850–860.

Wei, J., Li, Z., Pinker, R. T., Sun, L., et al. (2021). Himawari-8-derived diurnal variations of ground-level PM2.5 pollution across China using a fast space-time Light Gradient Boosting Machine. *Atmospheric Chemistry and Physics*. https://doi.org/10.5194/acp-2020-1277

Wu, J., Wang, Y., Liang, J., & Yao, F. (2021). Exploring common factors influencing PM2.5 and O3 concentrations in the Pearl River Delta: Tradeoffs and synergies. *Environmental Pollution, 285*, 117138. https://doi.org/10.1016/j.envpol.2021.117138

Wu, F., Min, P., Jin, Y., Zhang, K., Liu, H., & Zhao, J. (2023). A novel hybrid model for hourly PM2.5 prediction considering air pollution factors, meteorological parameters and GNSS-ZTD. *Environmental Modelling & Software, 167*, 105780.

Yang, W., Wu, Q., Li, J., Chen, X., et al. (2024). Predictions of air quality and challenges for eliminating air pollution during the 2022 Olympic Winter Games. *Atmospheric Research, 300*, 107225. https://doi.org/10.1016/j.atmosres.2024.107225

Ye, Z. (2019). Air pollutants prediction in Shenzhen based on Arima and prophet method. *E3S Web of Conferences, 136*, 05001. https://doi.org/10.1051/e3sconf/201913605001

Zeng, Y., Jaffe, D. A., Qiao, X., Miao, Y., & Tang, Y. (2020). Prediction of potentially high PM2.5 concentrations in Chengdu, China. *Aerosol and Air Quality Research, 20*, 956–965. https://doi.org/10.4209/aaqr.2019.11.0586

Zhang, L., Lin, J., Qiu, R., Hu, X., Zhang, H., Chen, Q., Tan, H., Lin, D., & Wang, J. (2018). Trend analysis and forecast of PM2.5 in Fuzhou, China using the ARIMA model. *Ecological Indicators, 95*, 702–710.

Zhu, J., Lee, R. W., Twum, C., & Wei, Y. (2019). Exposure to ambient PM2.5 during pregnancy and preterm birth in metropolitan areas of the state of Georgia. *Environmental Science and Pollution Research, 26*, 2492–2500.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.