scientific reports

OPEN

Air quality prediction models based on meteorological factors and real-time data of industrial waste gas

Ying Liu[1]✉, Peiyu Wang[2], Yong Li[1], Lixia Wen[1] & Xiaochao Deng[1]

With the rapid economic growth, air quality continues to decline. High-intensity pollution emissions and unfavorable weather conditions are the key factors for the formation and development of air heavy pollution processes. Given that research into air quality prediction generally ignore pollutant emission information, in this paper, the random forest supervised learning algorithm is used to construct an air quality prediction model for Zhangdian District with industrial waste gas daily emissions and meteorological factors as variables. The training data include the air quality index (AQI) values, meteorological factors and industrial waste gas daily emission of Zhangdian District from 1st January 2017 to 30th November 2019. The data from 1st to 31th December 2019 is used as the test set to assess the model. The performance of the model is analysed and compared with the backpropagation (BP) neural network, decision tree, and least squares support vector machine (LSSVM) function, which has better overall prediction performance with an RMSE of 22.91 and an MAE of 15.80. Based on meteorological forecasts and expected air quality, a daily emission limit for industrial waste gas can be obtained using model inversion. From 1st to 31th December 2019, if the industrial waste gas daily emission in this area were decreased from 6048.5 million cubic meters of waste gas to 5687.5 million cubic meters, and the daily air quality would be maintained at a good level. This paper deeply explores the dynamic relationship between waste gas daily emissions of industrial enterprises, meteorological factors, and air quality. The meteorological conditions are fully utilized to dynamically adjust the exhaust gas emissions of key polluting enterprises. It not only ensures that the regional air quality is in good condition, but also promotes the in-depth optimization of the procedures of regional industrial enterprises, and reduces the conflict between environmental protection and economic development.

Air quality is a critical issue related to people's health and livelihoods, and one of the obstacles to regional economic development and social progress. In addition to air quality monitoring and management, air quality forecasting during periods of polluted weather has also become a focus of environmental management. Especially during major events and heavy pollution emergencies, timely and accurate air quality prediction and pollution source analysis can provide a decision-making basis for management departments. If the exhaust gas emissions of enterprises can be determined according to the requirements of regional ambient air indicators and meteorological conditions, and then it could guide enterprises to adjust production processes accordingly. Air pollution caused by unfavorable meteorological factors can be effectively avoided, and enterprises can expand the production of heavy pollution processes when the weather conditions are favorable. Based on air quality prediction and pollution source analysis, it is of great practical significance to make full use of meteorological conditions to coordinate the relationship between air quality and regional development.

Some scholars at home and abroad have conducted qualitative analysis research on factors affecting air quality from the perspective of the environment, society, and economic activity, considering various factors such as waste incineration, vehicle exhaust emissions, population growth, coal combustion, industrial waste gas discharge and industrial flue gas dust. These studies confirmed that air pollution results from environmental degradation that has been majorly generated from urban population growth, industrial activities, and road fleet[1]. Industrial waste

[1]Departments of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 610000, China. [2]IT Electronics Eleventh Design & Research Institute Scientific and Technological Engineering Corporation Limited, Chengdu 610000, China. ✉email: 642823770@qq.com

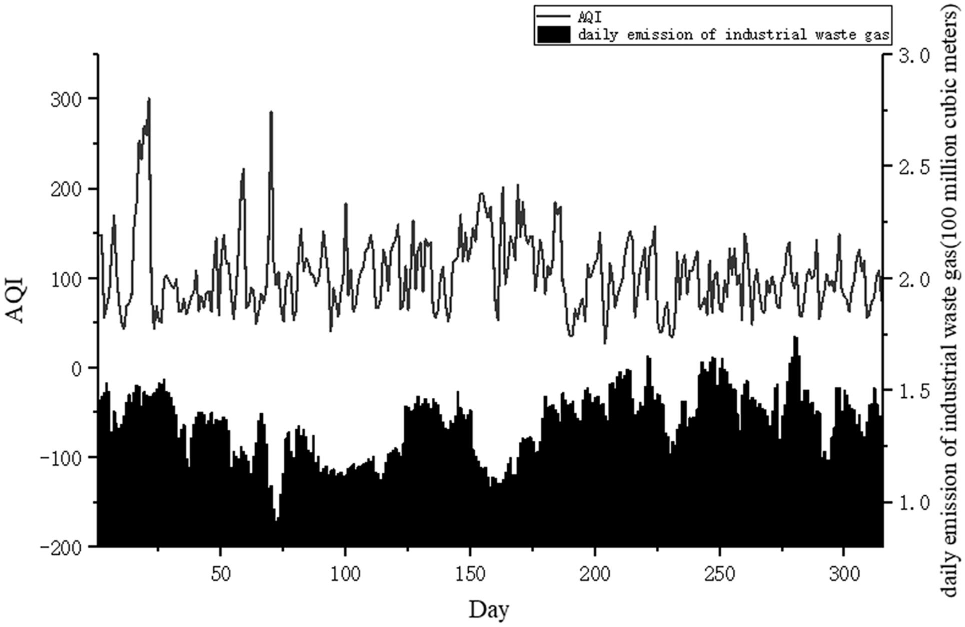


Figure 1. AQI and industrial waste gas emission statistics in Zhangdian District.

gas discharge is the main cause of air pollution in developing countries[2]. Thermal power plants and manufacturing industries are the largest sources of urban air pollution[3,4]. Other studies have selected meteorological factors such as average temperature, relative humidity, visibility, wind force scale, sun exposure, and wind direction to research the correlation between these meteorological factors and air quality. Research results indicated that average temperature, relative humidity, visibility, and wind force scale are the principal factors that affect air quality[5]. The variation in pollutant emissions affects an area within a hundred-kilometers radius from the source, depending also on local meteorological and geomorphological conditions[6].

There is also quantitative analysis of air quality based on pollutant transport, and diffusion processes. Currently, there are two main types of air quality prediction model: mechanism models and non-mechanism models. Mechanism models involve complex physical and chemical processes, which all possess great uncertainty. They require the establishment of a relatively complete emission source inventory, accurate meteorology fields, and related models of physical and chemical processes, such as pollutant transport and diffusion. Non-mechanism models, represented by statistical models and machine learning models, do not require complex pollutant boundary fields or meteorological boundary fields, nor do they need the investigation of complex mechanism processes generated by the results. This approach can determine the trend of pollution at a certain stage only by the extraction of data characteristics. Compared with mechanism models, non-mechanism models are more convenient and practical. The most commonly applied classical statistical methods mainly include linear and nonlinear models[7], multiple regression equation[8], time series[9], etc. Some conventional machine learning methods that are widely used include support vector machines[10], decision trees[11], Bayesian networks[12], artificial neural networks[13], backpropagation (BP) neural networks[14], etc. With the continuous development of artificial intelligence, deep machine learning models has been successfully implemented to forecast air quality using time series air pollutant and meteorological datasets with excellent performances[15].

Some scholars also look forward to the research on air quality prediction, pointing out that the existing research on the impact of industrial waste gas emissions on air quality is qualitative analysis, and the air quality prediction research ignores the emission information of pollution sources[1–3]. Some extensive studies can be further conducted to gasses emission estimating and its impact on the surrounding environments[16]. Urban air pollution mainly comes from industry, transportation and daily life. Industrial waste gas discharge are the largest sources of urban air pollution[17]. Traffic and household emissions are relatively stable and can be regarded as constant, with little impact on fluctuations in air quality. From the daily emissions data of industrial waste gas in Zhangdian District in 2018, it can be seen that the daily emissions of industrial waste gas fluctuate greatly (Fig. 1). Air quality prediction results will inevitably be inaccurate if pollutant variable emissions are not taken into account. Individual studies use source emission inventories, which treated industrial pollutant emissions as constant[18,19]. The emission inventory of pollution sources is compiled based on the base year. According to the technical guidelines for the compilation of air pollutant emission inventory of various industries, it is mainly calculated by the emission coefficient method. The estimation is difficult to be accurate, and generally lags by 2–3 years. The data is constant and cannot be updated dynamically, and cannot reflect the impact of real-time changes in emissions from pollution sources.

Because the existing air quality prediction research ignores the real-time emission effect of industrial pollutants in the model establishment, and cannot establish a quantitative correlation between air quality and industrial pollution sources, it cannot expand the application value of air quality prediction. This study uses machine learning algorithms to an air quality forecast model by considering real-time industrial waste gas emissions and meteorological factors as variables. The current weather forecast time frame (15 days) is considered a period. During this period, according to the weather forecast, the daily emission limit of industrial pollution is determined by model inversion. By increasing or reducing the output of polluting processes or sections within the enterprise and balancing the intensity of pollution emissions, it not only ensures that the regional air environment quality remains good, but also meets the company's supporting production requirements. It not only ensures that the environmental quality meets the standard, but also meets the normal operation of the enterprise. Regarding the selection of model algorithms, the random forest algorithm has several advantages compared with other machine learning algorithms. Firstly, the random forest algorithm can evaluate the importance of input variables and accurately predict output variables[20] Besides, it has good anti-noise ability and does not easily fall into the problem of overfitting[21]. Finally, the random forest algorithm is suitable for modelling high-dimensional data and has strong adaptability to data sets[22]. However, a key problem with the random forest algorithm is that parameters cannot be accurately optimised. In this paper, we use the "RandomizedSearchCV" and "GridSearchCV" functions to solve this issue and realise the precise optimisation of parameters. Next, the BP neural network, decision tree, and least squares support vector machine (LSSVM) are used to compare their model performance with the random forest algorithm. To eliminate long-term cumulative systematic errors caused by factors such as inter-annual fluctuations in the number of motor vehicles, a multi-step sliding window method (using the first 365 days of data to predict the next day's AQI) was adopted for the training set. By continuously incorporating measured data of air quality, meteorological conditions and industrial exhaust emissions into the training set and updating the training set in real time, the impact of long-term changes in traffic emissions on air quality can be reflected.

Study area and data

Zhangdian District is located in the middle of Zibo City, Shandong Province. It is located in the junction of the Shandong Zhongshan Mountains and the North Shandong Plain. It belongs to the warm temperate monsoon type semi-dry and semi-humid continental climate. Zibo City is one of the five traditional architectural ceramics production areas in China, and its architectural ceramics enterprises are mainly located in Zhangdian District. Zhangdian District has a total of 60 key industrial enterprises above designated size, including 41 non-metallic mineral products (37 building ceramics enterprises and 4 cement production enterprises), 12 chemical products manufacturing enterprises, 3 non-ferrous metal smelting and processing enterprises, and 4 other enterprises. For reference, a relief of the research area is shown in Fig. 2.

The input variables are meteorological factors and daily emissions of industrial waste gases, while the output variable is the AQI (Air Quality Index). According to the "Ambient Air Quality Index (AQI) Technical Regulations (Trial)" (HJ 633-2012), the air quality index is divided into 0–50, 51–100, 101–150, 151–200, 201–300 and greater than 300 the six levels, corresponding to the six levels of air quality (excellent, good, light pollution, moderate pollution, heavy pollution and serious pollution). Data sources are shown in Table 1. Measured data of the AQI and the daily emissions of industrial waste gases of major polluters were obtained from Zhangdian District Bureau of Ecology and Environment. Meteorological factors (precipitation, air temperature, relative humidity, wind scale, air pressure, total sunshine intensity and precipitation) were taken from the WheatA-Big Data on Agricultural Meteorology.

Methods

Establishment of the random forest model.

The random forest algorithm is a classification and regression algorithm that integrates multiple decision trees through ensemble learning[23]. First, the random forest algorithm uses the decision tree as the basic random forest classifier. Then, the second random forest classifier bagging method is used to generate the training data set and a random subspace is used to establish the classification of each strategic decision tree. The third random forest classifier randomly selects some attributes, then divides and combines the optimal attributes of each tree. The introduction of double randomisation makes it difficult for the random forest to fall into overfitting. Besides, there is diversity among classifiers, so the random forest has superior classification and regression performance[24].

The AQI prediction model is obtained by fitting training samples. The random forest modelling process is as follows:

(1) Define the AQI prediction training set, $X_i \to Y_i$, $(i = 298)$. Here, Y_i is the real value in the random forest prediction model, which is mapped to the measured AQI value of the *i*th sample in the data. Besides, X_i represents meteorological factors and industrial waste gas emissions of the *i*th sample in the data. The established feature vector, $\{I_{i1}, I_{i2}, \ldots I_{in}\} \to X_i$ represents the *i*th sample to the nth impact factor.

(2) Based on the training set, establish a single regression decision tree. Through the eigenvector X and its corresponding real value Y in the training sample, search for the splitting variables and splitting values. The regression decision tree divides the whole vector space into M partitions $\{R_1, R_2, \ldots R_m\}$. Any partition can be mapped to model C_m, and the vector can be divided into two parts by the value of a feature. The expression is:

$$R_1(j,s) = \{(I|I_j \le s)\}, \tag{1}$$

$$R_2(j,s) = \{(I|I_j > s)\}. \tag{2}$$

In the above equations, j represents an impact factor and s signifies the value when splitting. The objective function of the vector space split variable and split value search is:

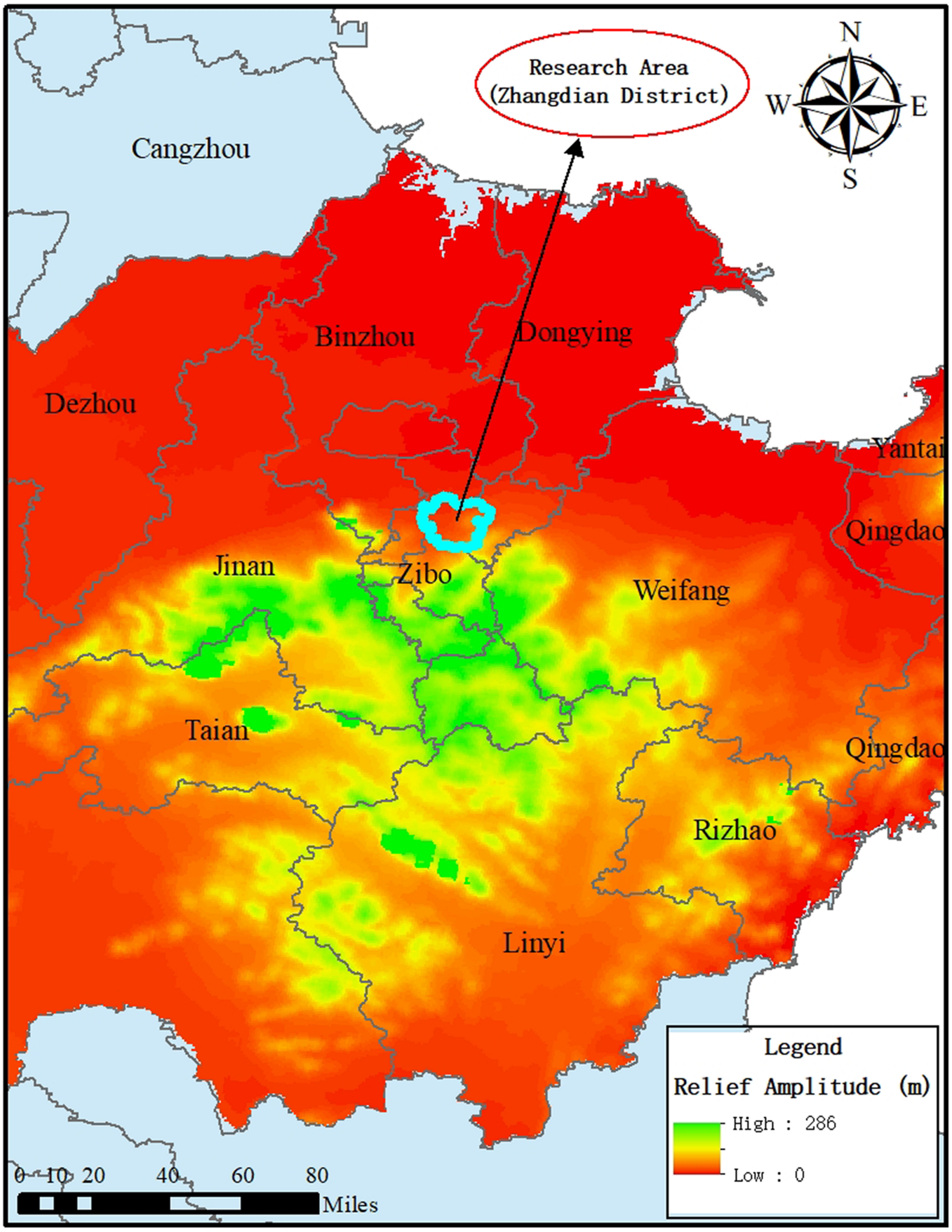


Figure 2. Relief amplitude of research area. The map was generated with ArcGIS10.2 (https://www.esri.com/en-us/arcgis/products/develop-with-arcgis/overview).

Data	Data source
AQI	Zhangdian District Bureau of Ecology and Environment
Meteorological factors	WheatA—Big Data on Agricultural Meteorology
Daily emissions of industrial waste gas	Environmental Statistics Yearbook of Zhangdian District

Table 1. Data sources.

$$z : \min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right]. \tag{3}$$

Here, z is the minimum variance of the measured AQI value, y_i represents the measured value of AQI in the ith sample, x_i is the eigenvector of the ith sample, while c_1 and c_2 denote the mean value of the measured AQI values in the first and second parts.

(3) Construct a complete random forest model on the basis of a single decision tree, where the generated model is a multiple nonlinear regression analysis model. The predicted value of the AQI is the average value of all the predicted values of the decision trees.

Since the random forest algorithm cannot accurately find its optimal parameters, in this paper, the model is enhanced through the "RandomizedSearchCV" and "GridSearchCV" functions to find its optimal parameters. Among them, RandomizedSearchCV is used to obtain the best parameters by randomly selecting parameter values and performing assigned times parameter combinations within the assigned parameter range; GridSearchCV is used to obtain the best parameters by exhaustively running through the given parameter values; CV is used for cross -validation, as well as parameter adjustment. Typically, RandomizedSearchCV is used first to obtain the optimal solution with a high probability of parameters, and then GridSearchCV is used to fine-tune the parameters within a certain floating range to obtain the optimal combination of parameters. "Finding Parameters" in the above figure is what RandomizedSearchCV and GridSearchCV need to do, which is to find the optimal combination of parameters. The specific process is described in Figs. 3 and 4, as follows.

Importance evaluation of variables. The importance evaluation of variables is a vital part of the random forest algorithm. It can evaluate the influence of input variables on output variables by using the mean square residual reduction in the decision-making process of the random forest. It is the result of continuous analysis and optimisation in the training process of the random forest. Based on various permutations, the mean-square residual reduction (%IncMSE) can be used to measure the influence of corresponding independent variables and is the standard for variable importance scoring[25]. The following is the calculation method of the mean square residual:

(1) Establish a regression tree for each training data set and then use this model to predict the OOB (out of bag) error. The mean square residual of b OOBs can be obtained: MSE_1, $MSE_2, \ldots MSE_b$.

(2) The number of variables selected by the self-help method in the random forest is random. Each variable X_i can be randomly transposed across b OOB datasets. This creates a new set of OOB tests. When the random forest regression model is used to predict the new test set, the mean square residual of the OOB after random replacement can be obtained. The matrix is as follows:

$$\begin{matrix} MSE_{11} & \cdots & MSE_{1b} \\ \cdots & \cdots & \cdots \\ MSE_{k1} & \cdots & MSE_{kb} \end{matrix}. \tag{4}$$

(3) Next, subtract from line of the equation. Then divide the mean by the standard error to obtain the mean square residual of variable, i.e., the variable importance score. The equation is expressed as follows:

$$VIM_i(MSE) = \left(\frac{1}{b} \sum_{j=1}^{b} \left(MSE_j - MSE_{ij} \right) \right) / S_E, (1 \leq i \leq k). \tag{5}$$

Evaluation of model prediction accuracy. In this study, the root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) were used for comparison between the measured and modelled AQI values[26,27]. These values can be determined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{k} \left(\widehat{y_i} - y_i \right)^2}{n}}, \tag{6}$$

$$MAE = \frac{1}{k} \sum_{i=1}^{k} \left| \widehat{y_i} - y_i \right|, \tag{7}$$

$$R^2 = \frac{\sum_{i=1}^{k} \left(\widehat{y_i} - \overline{y} \right)^2}{\sum_{i=1}^{k} \left(y_i - \overline{y} \right)^2}. \tag{8}$$

In the above equations, $\widehat{y_i}$ represents the AQI forecast of the ith sample, y_i is the measured AQI value of the ith sample, $\overline{y}$ denotes the average measured AQI value in all samples, and k is the sample size of the corresponding sample (k = 298).

Results

AQI and variation trend analysis. The graphs in Fig. 5 illustrate how meteorological factors, daily industrial waste gas emissions, and AQI varied in Zhangdian District from 1st January 2017 to 31th December 2019. It can be seen that the period with the largest variations in AQI was from December to March, as there are multiple

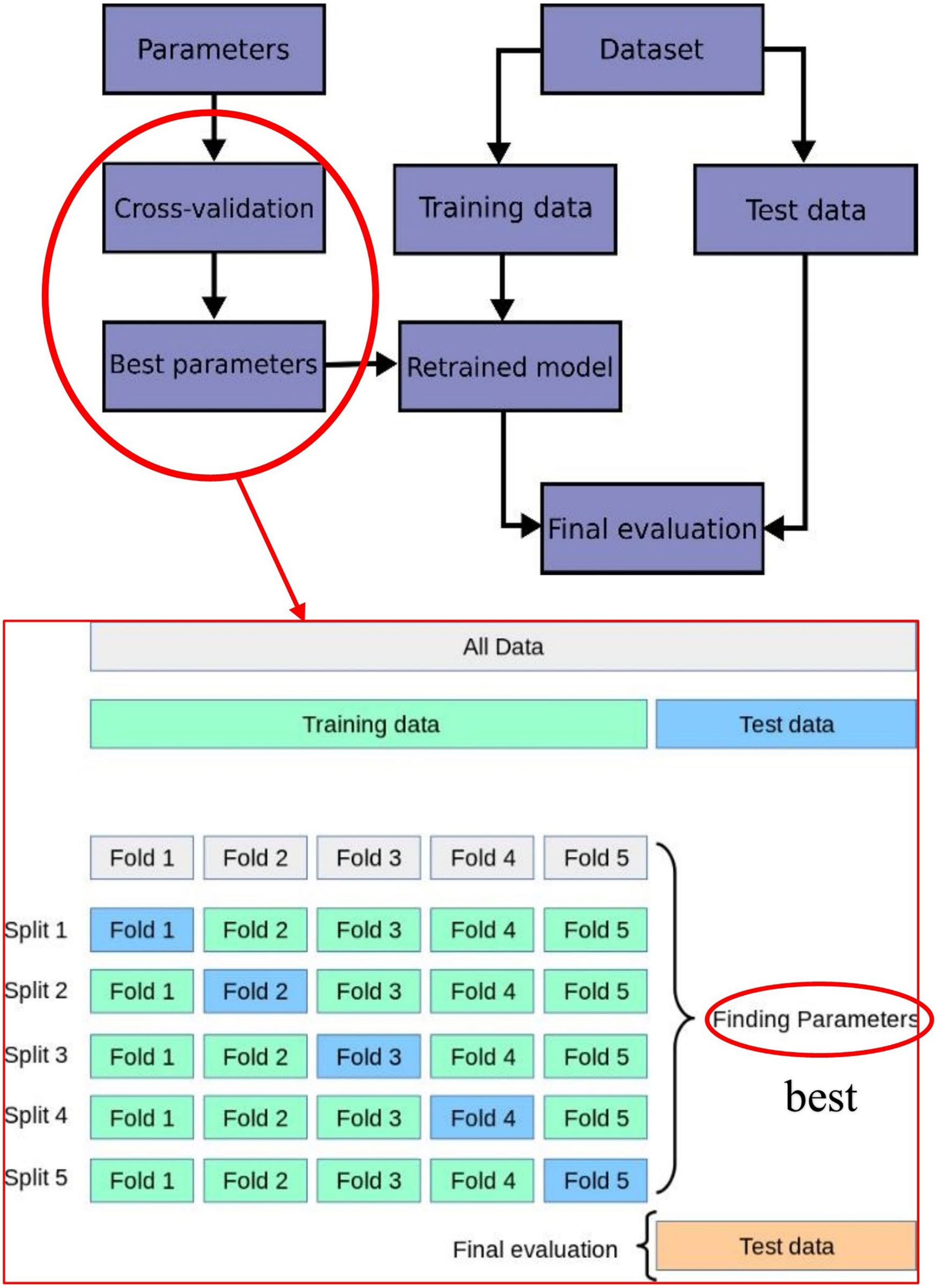


Figure 3. Schematic diagram of finding the optimal parameters of random forest model.

peaks during this time. The minimum value of AQI during these three months was 13 while the maximum AQI value was 313. Between June and August, there were also significant variations in the AQI. In the other months, the range of change was relatively low, with the AQI remaining around 90. The relative humidity fluctuated greatly from February to May, with an average of 46.2%. However, between July and August, the relative humidity only varied slightly, with an average value of 73.7%. The average temperature in February was the lowest, then from March to August it rose slowly, while from August to October it gradually decreased. The wind scale was

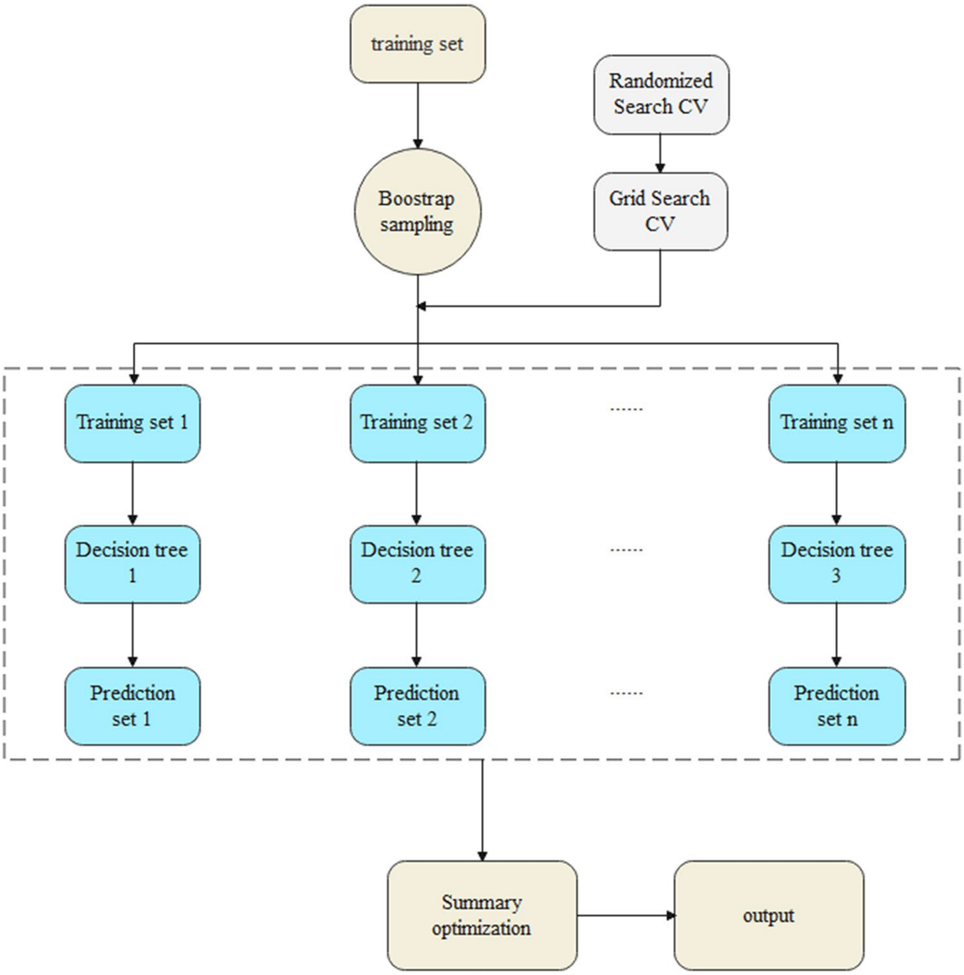


Figure 4. Flow chart of random forest model.

relatively stable, although in March and April the wind scale was more erratic. In the other months, the wind scale was generally category 1 or 2. Visibility varied greatly throughout the study period. The average value was about 12.5 km, while the maximum value was 29.5 km and the minimum value was 1.0 km. The average pressure in July was the lowest with an average of 98.7 kPa then from August to December it rose slowly, while from December to July it gradually decreased. Total sunshine intensity varied greatly throughout the study period. The average value was about 15.85 J/m^2, while the maximum value was 28.59 J/m^2 and the minimum value was 0.66 J/m^2. The precipitation fluctuated greatly from February to May, with an average of 46.2%. Finally, the average daily emissions of industrial waste gas over the whole study period were 153 million cubic meters, while the maximum value was 270 million cubic meters and the minimum value was 100 million cubic meters. Average daily emissions of industrial waste gas in 2019 were 33 million cubic meters and 85 million cubic meters more than in 2018 and 2017, respectively, but the AQI annual average in 2019 was lower than both 2018 and 2017. Because Shandong Province implemented several air pollutant emission standards ("Emission standard of air pollutants for building materials industry", Effective January 1, 2019) ("Emission standard of air pollutants for industrial furnace and kiln", Effective June 1, 2019) in 2019, stricter pollutant emission concentration limits were implemented.

AQI and variation correlation analysis. To verify that meteorological factors and industrial waste gas emissions affect air quality, we conducted a correlation analysis of AQI, meteorological factors, and industrial waste gas emissions in this paper, with the results presented in Table 2. Results indicate that industrial waste gas emissions were positively correlated with AQI, while visibility were negatively correlated with AQI. A rise in industrial waste gas emissions leads to an increase in AQI and the deterioration of air quality. As the amount of particulate matter in the air increases, it leads to the occurrence of haze and reduces visibility. There is a negative correlation between AQI and precipitation in the year and most seasons. This is because raindrops in the cloud can absorb and absorb pollutant particles, and at the same time, rainwater can wash and wash pollutants, resulting in lower pollutant concentrations, improved air quality, and lower AQI values. The correlation in winter is not obvious, which may be due to less precipitation in winter and uneven spatial and temporal distribution. There is a positively correlation between AQI and air temperature in the year and most seasons. From the sea-

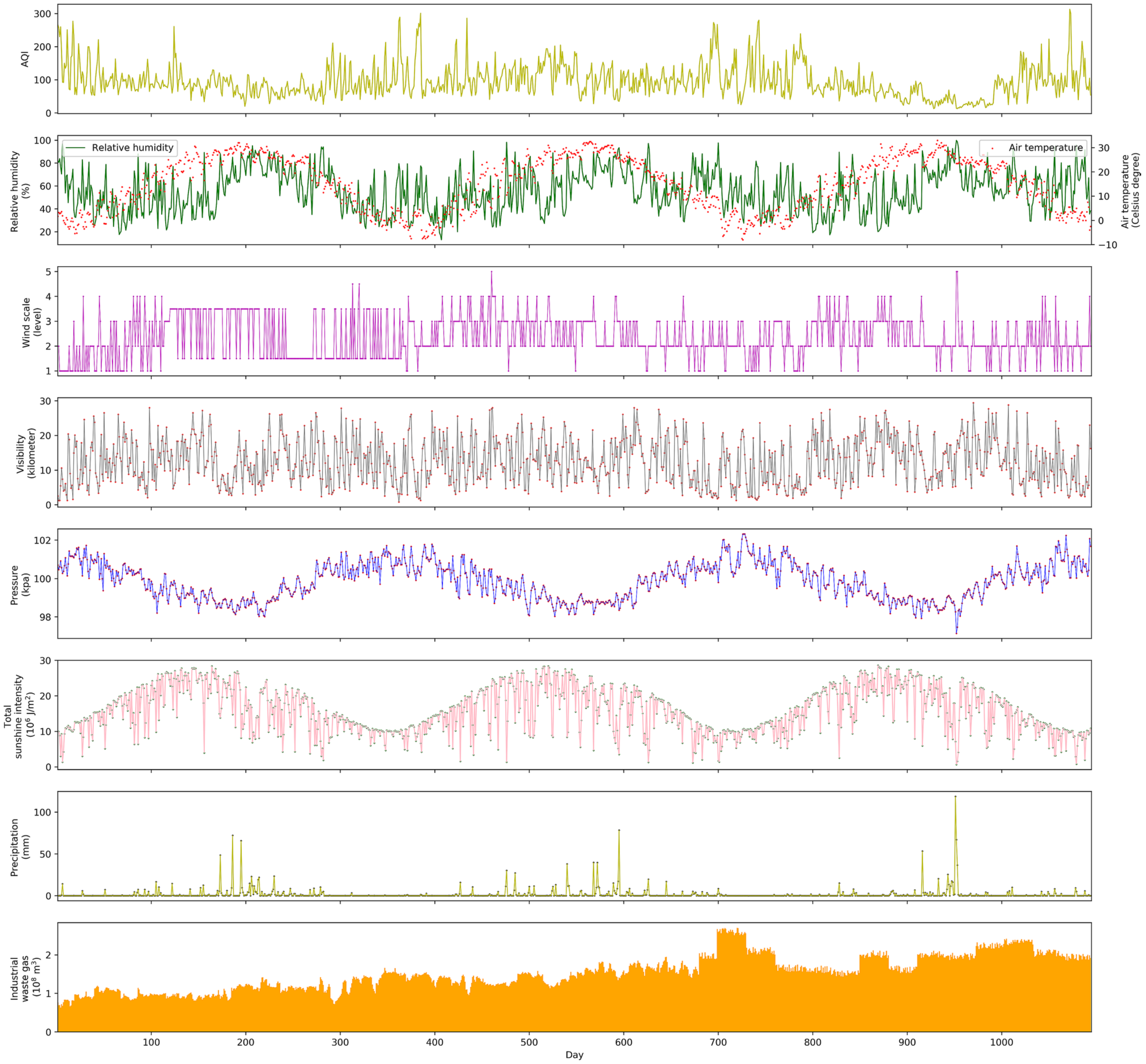


Figure 5. Trends of meteorological factors, industrial waste gas emissions, and AQI.

Season	Air temperature	Wind scale	Visibility	Air pressure	Total sunshine intensity	Relative humidity	Precipitation	Industrial waste gas emissions
Spring	0.034	−0.084	−0.513	−0.042	0.010	−0.018	−0.252	0.224
Summer	0.227	0.047	−0.189	−0.025	0.347	−0.363	−0.426	0.587
Autumn	−0.292	−0.063	−0.610	0.210	−0.248	0.068	−0.34	0.252
Winter	0.456	−0.208	−0.646	−0.353	−0.164	0.498	0.129	0.315
Year	0.354	−0.165	−0.526	0.293	−0.215	−0.012	−0.326	0.374

Table 2. Correlation between seasonal and annual AQI and meteorological elements from 2017 to 2019.

sonal scale, there is no obvious correlation between AQI and air temperature in spring. AQI in summer is significantly positively correlated with air temperature, which may have a certain relationship with the activity of cold and warm air masses, because when warm air masses pass through, the temperature will increase and a large amount of pollutants will accumulate. When the cold air passes through, it will reduce the temperature and often accompanied by wind, which is conducive to the diffusion of pollutants. The activities of cold and warm masses often occur frequently in summer. In autumn, atmospheric turbulence activities will intensify with the increase

Parameter	Value
max_depth	60
max_features	5
min_samples_leaf	2
min_samples_split	5
n_estimators	1400

Table 3. Optimal parameters of the random forest model.

Date	Measured AQI values	Predicted AQI values
1th Dec.	87	92
2th Dec.	62	67
3th Dec.	79	85
4th Dec.	112	99
5th Dec.	81	91
6th Dec.	79	90
…	…	…
26th Dec.	92	99
27th Dec.	71	86
28th Dec.	80	102
29th Dec.	69	137
30th Dec.	105	98
31th Dec.	54	61

Table 4. AQI prediction results of random forest model.

of air temperature, which will dilute and diffuse pollutants in the vertical direction of the lower layer, and further lead to the decrease of AQI. While rises in temperature can cause the temperature inversion phenomenon in winter, and exacerbating the air pollution problem. Ye's analysis of Fairbanks confirmed that air temperature and AQI were positively correlated, while visibility was negatively correlated with AQI[28]. Guo studied the correlation between meteorological factors and AQI and also verified that there was a positive correlation between temperature and AQI[29]. These are consistent with the correlation analysis results obtained in this paper.

Results indicate that the correlation of AQI with other meteorological elements (relative humidity, wind level, Air pressure, Total sunshine intensity and precipitation) is not the same on different time scales, because these meteorological elements vary greatly on different time scales. Taking relative humidity as an example, different scholars have studied the relationship between urban air pollution characteristics and meteorological conditions, and found that some cities have a positive correlation between pollutant concentrations and relative humidity[30–33], and some cities have a negative correlation with relative humidity[28,34–38]. In Zhangdian District, there are different correlations between AQI and relative humidity in different seasons. There is an obvious positive correlation in winter, a negative correlation in summer, and no correlation in spring and autumn. Under low humidity conditions, the growth of condensation nuclei in the atmosphere aggravates pollution, and under high humidity conditions, it will have a scavenging effect on pollutants due to deposition[39]. On the other hand, relative humidity is negatively correlated with AQI. The reason may be that when the relative humidity is low, it is often accompanied by strong winds, which is easy to cause sand and dust weather and make the air quality worse. It can be seen that relative humidity is not the dominant factor affecting the development of pollution, and comprehensive judgments need to be combined with pollution emissions, meteorological conditions, and chemical processes.

Results of the random forest model.

The data samples selected in this paper include meteorological factors (average temperature, wind scale, relative humidity, and visibility), industrial waste gas emissions, and AQI in Zhangdian District. In this study, we obtained a total of 1095 sets of data, among which 1064 sets of data were used as the training data for the AQI prediction model. The final 31 sets were used as test data to verify the model. The prediction process of the random forest model was implemented using the Python programming platform. In the Python program, we used the "RandomizedSearchCV" function to approximate the random forest algorithm parameters. Then, the "GridSearchCV" function was used to accurately search the parameters of the random forest. The optimal parameters that we obtained are presented in Table 3.

The random forest model was established after searching the optimal parameters of the random forest. The last 31 sets of original data were used as samples for prediction. The AQI prediction results are displayed in Table 4, which shows that the predicted AQI values are similar to the measured values, indicating that the predicted results are accurate.

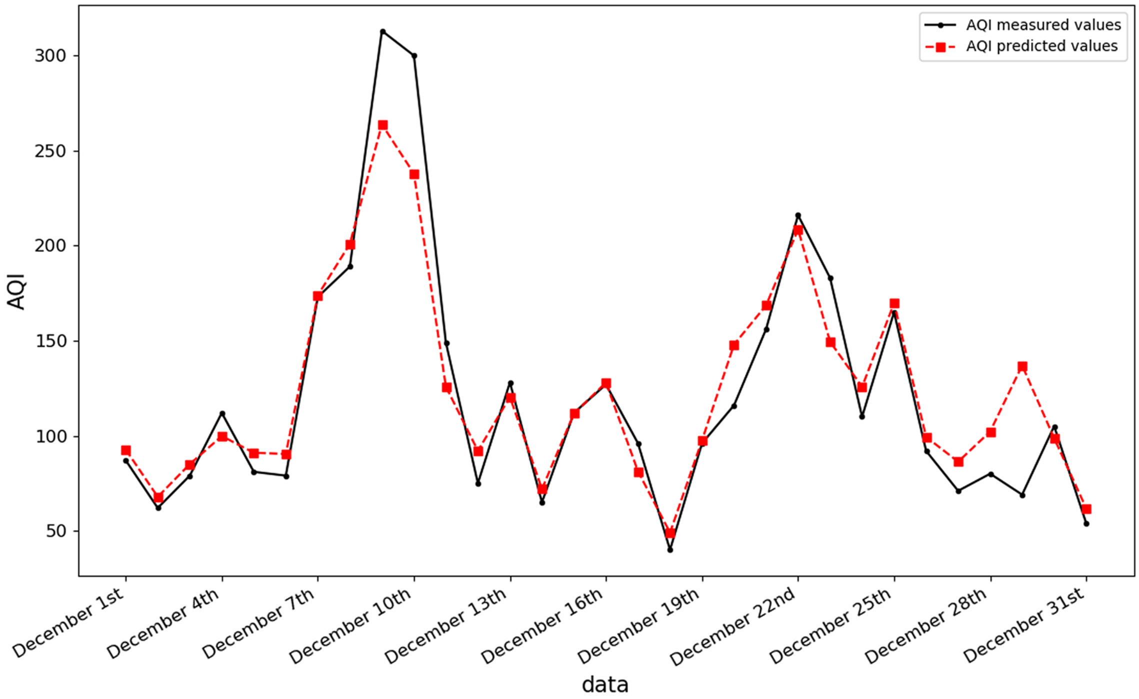


Figure 6. Comparison of modelled and measured AQI values.

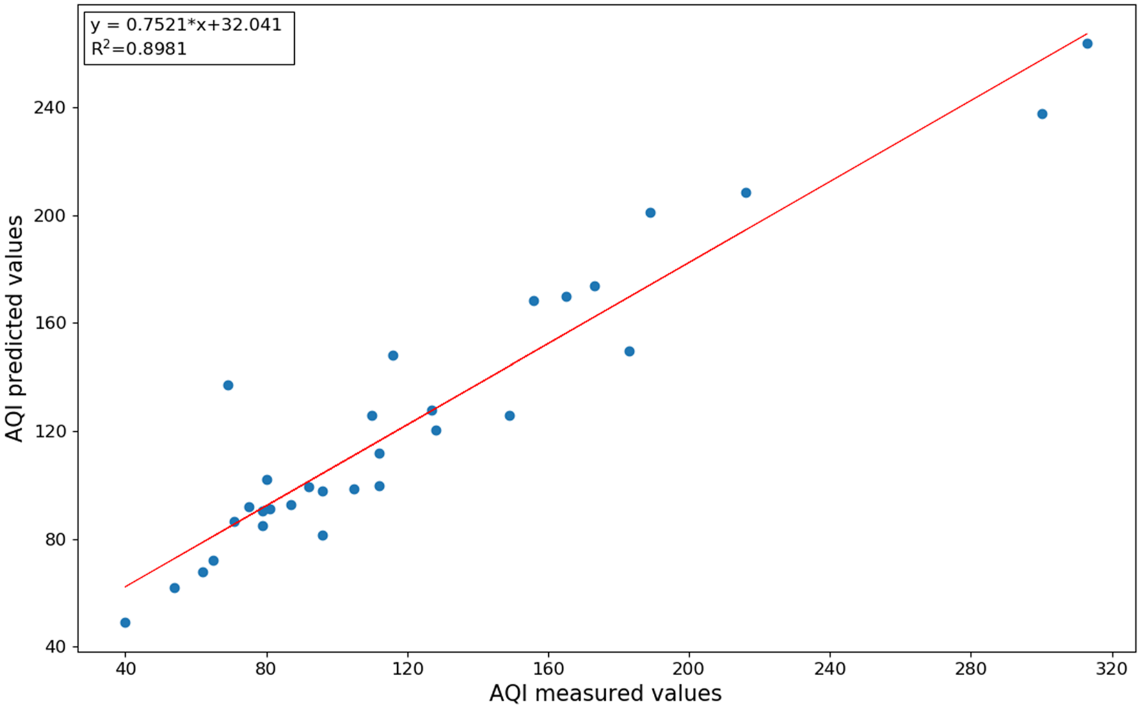
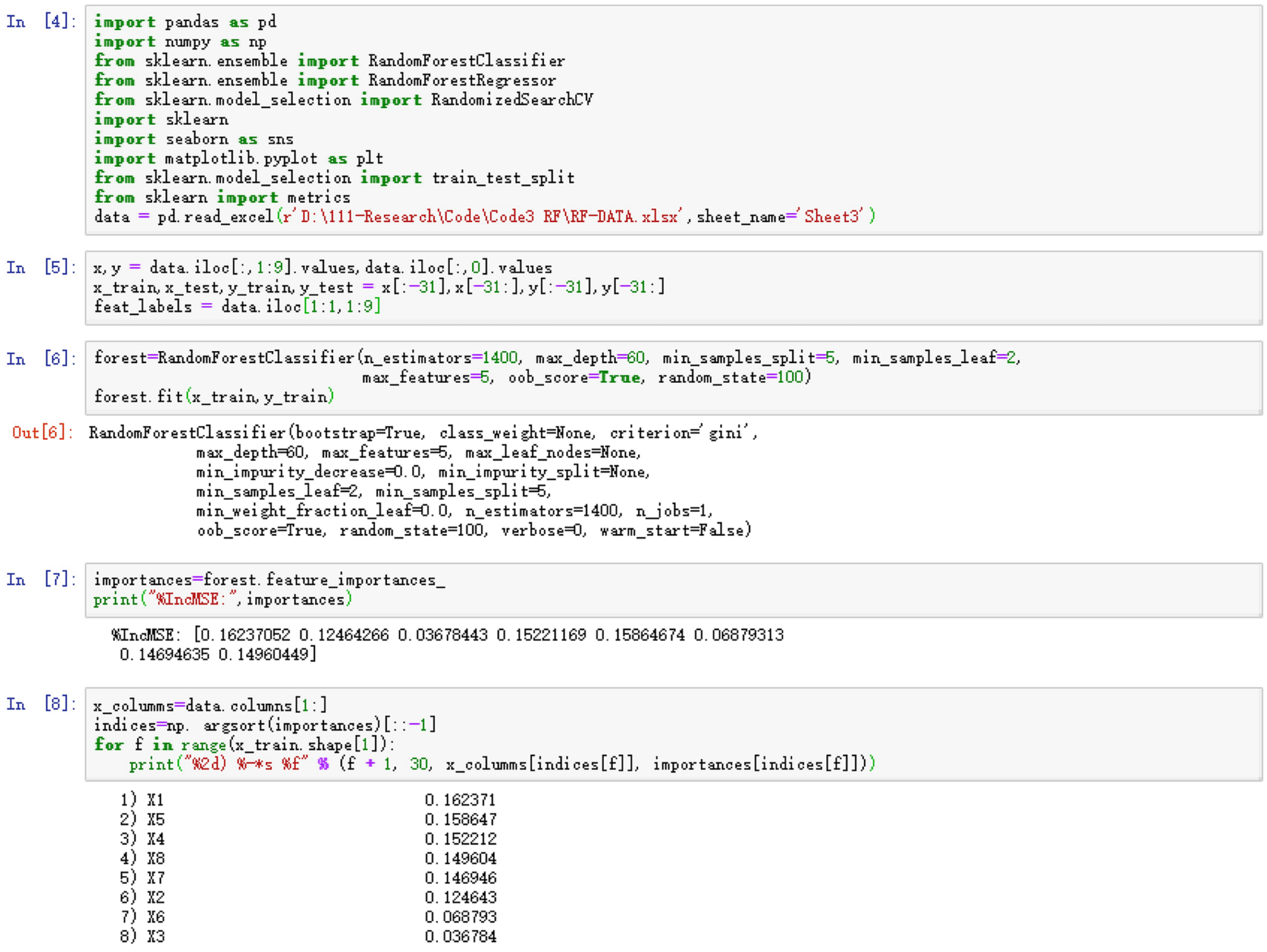


Figure 7. Linear fitting of predicted and measured AQI values.

The AQI predicted by the random forest model was compared with the measured AQI. It can be seen from Fig. 6 that the trend of the predicted and measured AQI is fundamentally the same. Figure 7 illustrates that the R2 value is 0.90, and the scatter points are precisely distributed at both ends of the line, indicating that the linear fitting is accurate. We can conclude that in this region it is effective to use meteorological factors and daily emissions of industrial waste gases to predict the AQI.

Variable importance evaluation. In this study, we used Python to calculate the mean square residual (%IncMSE) in the random forest algorithm and determine the importance of each input variable. The Python



```
In [4]: import pandas as pd
        import numpy as np
        from sklearn.ensemble import RandomForestClassifier
        from sklearn.ensemble import RandomForestRegressor
        from sklearn.model_selection import RandomizedSearchCV
        import sklearn
        import seaborn as sns
        import matplotlib.pyplot as plt
        from sklearn.model_selection import train_test_split
        from sklearn import metrics
        data = pd.read_excel(r'D:\111-Research\Code\Code3 RF\RF-DATA.xlsx',sheet_name='Sheet3')

In [5]: x,y = data.iloc[:,1:9].values,data.iloc[:,0].values
        x_train,x_test,y_train,y_test = x[:-31],x[-31:],y[:-31],y[-31:]
        feat_labels = data.iloc[1:1,1:9]

In [6]: forest=RandomForestClassifier(n_estimators=1400, max_depth=60, min_samples_split=5, min_samples_leaf=2,
                                      max_features=5, oob_score=True, random_state=100)
        forest.fit(x_train,y_train)

Out[6]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                    max_depth=60, max_features=5, max_leaf_nodes=None,
                    min_impurity_decrease=0.0, min_impurity_split=None,
                    min_samples_leaf=2, min_samples_split=5,
                    min_weight_fraction_leaf=0.0, n_estimators=1400, n_jobs=1,
                    oob_score=True, random_state=100, verbose=0, warm_start=False)

In [7]: importances=forest.feature_importances_
        print("%IncMSE:",importances)

          %IncMSE: [0.16237052 0.12464266 0.03678443 0.15221169 0.15864674 0.06879313
           0.14694635 0.14960449]

In [8]: x_columns=data.columns[1:]
        indices=np. argsort(importances)[::-1]
        for f in range(x_train.shape[1]):
            print("%2d) %-*s %f" % (f + 1, 30, x_columns[indices[f]], importances[indices[f]]))

         1) X1                                   0.162371
         2) X5                                   0.158647
         3) X4                                   0.152212
         4) X8                                   0.149604
         5) X7                                   0.146946
         6) X2                                   0.124643
         7) X6                                   0.068793
         8) X3                                   0.036784
```

Figure 8. Python code for importance evaluation calculations.

Input variable	%IncMSE
X1	0.162371
X5	0.158647
X4	0.152212
X8	0.149604
X7	0.146946
X2	0.124643
X6	0.068793
X3	0.036784

Table 5. Tanking of importance of variable.

code and parameters are presented in Fig. 8. A larger mean square residual reduction value indicates that the input variable has a larger influence on the output variable. As shown in Table 5, industrial waste gas (X1) was the greatest variable affecting AQI, followed by visibility (X5), relative humidity (X4), total sunshine intensity (X8), air pressure (X7), air temperature (X2) and precipitation (X6). The mean square residual value of the wind scale (X1) is the smallest, indicating that the influence of the wind on AQI is negligible compared with the other variables.

Model prediction accuracy evaluation. By comparing the random forest algorithm with other machine learning algorithms, we can verify the applicability of the random forest algorithm for air quality prediction in Zhangdian District. In this paper, four kinds of machine learning algorithms were used to predict AQI, and their results were compared to ascertain the most appropriate machine learning algorithm. The RMSE, MAE, and R^2 measures were used to evaluate the prediction accuracy of the four machine learning algorithms[28]. For these

Model	RMSE	MAE	R^2
Random forest	22.91	15.80	0.90
BP neural network	26.72	17.53	0.81
Decision tree	29.85	18.11	0.76
LSSVM	26.29	17.37	0.80

Table 6. Model prediction accuracy evaluation.

Air pollution Date	Predicted AQI value	Daily emissions of industrial waste gas (10^8 m^3)	Daily emissions of industrial waste gas to reach target AQI of 100 (10^8 m^3)	Difference (10^8 m^3)
1th Dec.	92	2.027	2.188	− 0.161
2th Dec.	67	1.972	2.907	− 0.935
3th Dec.	85	1.960	2.304	− 0.344
4th Dec.	99	1.946	1.950	− 0.004
5th Dec.	91	1.868	2.052	− 0.184
6th Dec.	90	2.029	2.245	− 0.216
7th Dec.	174	1.855	1.067	0.788
8th Dec.	201	1.966	0.978	0.988
9th Dec.	263	1.919	0.728	1.191
10th Dec.	237	1.897	0.797	1.1
11th Dec.	125	2.029	1.613	0.416
12th Dec.	92	2.028	2.203	− 0.175
13th Dec.	120	1.997	1.658	0.339
14th Dec.	72	1.891	2.628	− 0.737
15th Dec.	112	1.912	1.712	0.2
16th Dec.	128	2.017	1.576	0.441
17th Dec.	81	1.869	2.301	− 0.432
18th Dec.	49	1.857	3.789	− 1.932
19th Dec.	98	2.011	2.060	− 0.049
20th Dec.	148	1.987	1.343	0.644
21st Dec.	169	1.980	1.175	0.805
22nd Dec.	209	1.885	0.904	0.981
23rd Dec.	149	2.023	1.353	0.67
24th Dec.	126	1.960	1.559	0.401
25th Dec.	169	1.955	1.152	0.803
26th Dec.	99	1.884	1.897	− 0.013
27th Dec.	86	1.967	2.279	− 0.312
28th Dec.	102	2.004	1.960	0.044
29th Dec.	137	1.880	1.372	0.508
30th Dec.	98	1.988	2.015	− 0.027
31th Dec.	61	1.922	3.110	− 1.188

Table 7. Target industrial emissions at AQI of 100.

algorithms, lower RMSE and MAE values indicate higher prediction accuracy, while the closer the R^2 value is to 1, the more accurate the prediction is. The results presented in Table 6 confirm that the prediction accuracy of the random forest model is better than the other three machine learning models, indicating that the random forest model is the most suitable algorithm for the AQI prediction model of Zhangdian District.

Control of industrial exhaust emissions based on target AQI.

It can be seen from Table 7 that the measured AQI value of this region on 9th December was 263 (heavy pollution), and the industrial waste gas emission on that day was 191.9 million m^3. The modelled results show that if the daily industrial waste gas emissions were controlled at 72.8 million m^3, the air quality of the day could reach an acceptable level (AQI = 100). Conversely, on 18th December, the measured AQI value of the region was 49. The local meteorological conditions were favourable on this day, so the production time of high-polluting manufacturing processes load could be appropriately increased, and the daily industrial waste gas emissions could be increased by 378.9 million m^3.

It can also be seen from Table 6 that the air quality in this region was poor in December 2019. There were 4 days of heavy air pollution, 3 days of moderate air pollution and 9 days of mild air pollution. According to the rationality of meteorological conditions, the air quality in this area could be maintained in good condition (AQI < 100) by increasing or reducing the industrial exhaust emission. It can also be seen from Table 6 that the total allowable exhaust emission in this area would be decreased by 361 million m^3 compared with the actual emission in December 2019. The production capacity of enterprises would be decreased, but it would be better than the direct shutdown. According to Zibo City's Emergency Plan for Heavy Pollution Weather (implemented in 2021), if the air quality index is greater than 200, these 60 key enterprises will directly stop work and production.

There are a large number of ceramic factories in this region, and there are two main sources of exhaust gases in the production of ceramics. The first is dust from crushing, screening, granulation, and spray drying in the manufacture of preformed moulds, glaze materials, and colouring materials. The second is high-temperature flue gas containing $S0_2$ and smoke produced in the operation of various kiln firing equipment. Due to the different operating times of each process in the different factories vary, the collective operational load and pollution load of the processes are not balanced. This leads to great fluctuations in the daily emissions of industrial waste gas. By reducing the scale of "firing" processes and appropriately increasing the level of "raw material preparation" or "moulding" in periods of adverse meteorological conditions, the daily emissions of industrial waste gas can be reduced to ensure that the local environmental air quality is maintained at an acceptable level. On the other hand, increasing the operation of "firing" processes in favourable weather can balance the requirements of enterprises, allowing them to reach production targets. Given this, factories could reasonably adjust their production processes depending on the coming meteorological conditions, especially adverse meteorological conditions, to ensure that the regional environmental air quality is preserved in an optimal state.

Feasibility analyze of enterprise process adjustment. Because the production process of the enterprise has the characteristics of multi-section cooperation, multi-machine parallel, and random "fluctuation" and nonlinear interaction between unit sections, the production process network presents great complexity and uncertainty. Production scheduling optimization research has always been a research hotspot. But the current research mainly focuses on the aspects of profit maximization[40], time constraints[41], capital constraints[42], resource constraints[43], energy constraints[44], and production equipment constraints[45]. This research provides a new idea for the optimization of production scheduling in industrial enterprises.

The operation time of each production section within the enterprise is different, and the load is not balanced, so the sections that run every day are also different. The production process of some heavy air pollution industries (surface coating, pharmaceuticals, packaging and printing, building materials production, etc.) has certain discrete and intermittent sections, such as magnetic pole smears in motor manufacturers, purification in pharmaceutical companies, and burning in architectural ceramics companies. into the waiting section. These polluting sections have the characteristics of discontinuous intermittent, and the operation time is flexible and adjustable.

Adjustment of polluting processes or sections in the enterprise: (1) Verify the contribution index or scale model of each polluting process or section of the enterprise to the overall emission of the enterprise; (2) Insert the model index into the ERP (Enterprise Resource Planning) self-made parts material scheduling module to convert the process capability; (3) the process capability is adjusted by bringing the environmental prediction index in one cycle into the process capability calculation.

Conclusions

In this study, a random forest model is used to construct an air quality prediction model in Zhangdian District based on the real-time dynamic emission effect of industrial waste gas-meteorological conditions, and to quantify the impact of industrial waste gas on air quality in the region. Using this model, the daily emission limit of industrial pollution can be determined according to the weather forecast inversion, and the air pollution risk caused by unfavorable meteorological factors can be effectively avoided by adjusting the production capacity of the internal production process of the enterprise. This research actively responds to the "Fourteenth Five-Year Plan for National Economic and Social Development of Zhangdian District and the Outline of Vision 2035": by promoting the implementation of typical production scenarios, empowering actions, focusing on digital industrial applications, using cloud computing, big data and other new-generation information technologies, and guidelines for building a new industrialized strong city in the country. It provides a new idea for Zhangdian District's "14th Five-Year Plan" to achieve an average annual growth rate of regional GDP of more than 7% and the harmonious development of industry and environment.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 14 February 2022; Accepted: 16 May 2022
Published online: 03 June 2022

References

1. Tella, A. & Balogun, A. L. GIS-based air quality modelling: Spatial prediction of PM10 for Selangor State, Malaysia using machine learning algorithms. *Environ. Sci. Pollut. Res.* https://doi.org/10.1007/s11356-021-16150-0 (2021).
2. Zhu, Z. P. *et al.* The impact of meteorological conditions on air quality index under different urbanization gradients: A case from Taipei. *Environ. Dev. Sustain.* **23**(3), 3994–4010 (2021).

3. Xiao, J. N. *et al.* Spatiotemporal distribution pattern of ambient air pollution and its correlation with meteorological factors in Xiamen City. *Acta Sci. Circum.* **36**(9), 3363–3371 (2016).
4. Michanowicz, D. R. *et al.* A hybrid land use regression/AERMOD model for predicting intra-urban variation in PM2.5. *Atmos. Environ.* **131**, 307–315 (2016).
5. Guo, J. Q. & Feng, Z. K. Study on spatial temporal distribution characteristics of air quality index in Beijing and its correlation with local meteorological conditions. *Discr. Dyn. Nat. Soc.* https://doi.org/10.1155/2019/1462034 (2019).
6. Carnevale, C. *et al.* Assessing the economic and environmental sustainability of a regional air quality plan. *Sustainability* **10**(10), 3568 (2018).
7. Amanollahi, J. & Ausati, S. Validation of linear, nonlinear, and hybrid models for predicting particulate matter concentration in Tehran, Iran. *Theor. Appl. Climatol.* **140**, 709–717 (2020).
8. Abdullah, S. *et al.* Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia. *Atmosphere* **11**(3), 14 (2020).
9. Cekim, H. O. Forecasting PM10 concentrations using time series models: A case of the most polluted cities in Turkey. *Environ. Sci. Pollut. Res.* **27**, 25612–25624 (2020).
10. Nieto, P. J. G., Combarro, E. F., Diaz, J. J. D. & Montanes, E. A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study. *Appl. Math. Comput.* **219**(17), 8923–8937 (2013).
11. Wang, Y. N. & Kong, T. Air quality predictive modeling based on an improved decision tree in a weather-smart grid. *IEEE Access.* **7**, 172892–172901 (2019).
12. Naili, M., Bourahla, M., Naili, M. & Tari, A. Stability-based dynamic Bayesian network method for dynamic data mining. *Eng. Appl. Artif. Intell.* **77**, 283–310 (2019).
13. Goulier, L., Paas, B., Ehrnsperger, L. & Klemm, O. Modelling of urban air pollutant concentrations with artificial neural networks using novel input variables. *Int. J. Environ. Res. Public Health* **17**(6), 2025 (2020).
14. Huang, Y., Xiang, Y. X., Zhao, R. X. & Cheng, Z. Air quality prediction using improved PSO-BP neural network. *IEEE Access.* **8**, 99346–99353 (2020).
15. Xu, W. X. *et al.* Understanding the spatial-temporal patterns and influential factors on air quality index: The case of North China. *Int. J. Environ. Res. Public Health* **16**(16), 23 (2019).
16. Nur'atiah, Z., Lee, W. E., Ali, N. A. & Marlinda, A. M. A systematic literature review of deep learning neural network for time series air quality forecasting. *Environ. Sci. Pollut. Res.* **29**(4), 4958–4990 (2022).
17. Zhao, M. L., Liu, F. Y., Song, Y. J. & Geng, J. B. Impact of Air pollution regulation and technological investment on sustainable development of green economy in Eastern China: Empirical analysis with panel data approach. *Sustainability* **12**(8), 3073 (2020).
18. Sun, S. Q., Wang, S. G., Luo, B., Du, Y. S. & Zhang, W. Air pollution forecast in winter based on machine learning method in Chengdu. *J. Meteorol. Environ.* **36**(2), 98–104 (2020).
19. Shang, Z. W., Kang, Y. Z., Du, H. & Wang, S. G. Study on the relationship between air pollution and meteorological conditions in Beijing and their forecasting. *J. Lanzhou Univ. Nat. Sci.* **56**(3), 380–387 (2020).
20. Zhang, H. *et al.* Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model. *Sci. Total Environ.* **592**, 704–713 (2017).
21. Chen, W. *et al.* A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *CATENA* **151**, 147–160 (2017).
22. Speiser, J. L., Miller, M. E., Tooze, J. & Ip, E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* **134**, 93–101 (2019).
23. Jeung, M. *et al.* Evaluation of random forest and regression tree methods for estimation of mass first flush ratio in urban catchments. *J. Hydrol.* **575**, 1099–1110 (2019).
24. Wang, H., Sun, J. X., Sun, J. B. & Wang, J. L. Using random forests to select optimal input variables for short-term wind speed forecasting models. *Energies.* https://doi.org/10.3390/en10101522 (2017).
25. Gregorutti, B., Michel, B. & Saint-Pierre, P. Correlation and variable importance in random forests. *Stat. Comput.* **27**(3), 659–678 (2017).
26. Piepho, H. P. A coefficient of determination (R-2) for generalized linear mixed models. *Biom. J.* **61**(4), 860–872 (2019).
27. Willmott, C. J. & Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **30**(1), 79–82 (2005).
28. Ye, L. X. & Wang, Y. G. Long-term air quality study in Fairbanks, Alaska: Air pollutant temporal variations, correlations, and PM2.5 source apportionment. *Atmoshere* **11**(11), 19 (2020).
29. Guo, Q. G. H. *et al.* Air pollution forecasting using artificial and wavelet neural networks with meteorological conditions. *Aerosol Air Qual. Res.* **20**(6), 1429–1439 (2020).
30. Cai, Z. Y. *et al.* Improvement of environmental model prediction based on inversion and aerosol assimilatin. *Environ. Sci. (Beijing).* https://doi.org/10.13227/j.hjkx.202109263 (2021).
31. Wang, W., Cheng, X. Y., Hu, C., Xia, S. H. & Wang, T. Spatio-temporal distribution characteristics of PM2.5 and air quality evaluation in urban street canyons: Take Changhuai Street in Hefei as an example. *Ecol. Environ. Sci.* **30**(11), 2157–2164 (2021).
32. Yin, X. M. *et al.* Effect analysis of meteorological conditions on air quality during the winter COVID-19 lockdown in Beijing. *China Environ. Sci. (Chin. Ed.)* **41**(05), 1985–1994 (2021).
33. Liu, F. L. & Liao, J. J. Spatial-temporal distribution characteristics and influencing factors of air quality in urban cluster along middle reach of Yangtze River. *Environ. Sci. Technol. (Wuhan)* **44**(10), 172–186 (2021).
34. Zhang, H. *et al.* Characteristics of primary pollutants of air quality and their relationships with meteorological conditions in Heyuan. *J. Meteorol. Environ.* **38**(01), 40–47 (2022).
35. Zhou, M. G., Yang, Y., Sun, Y., Zhang, F. Y. & Li, Y. H. Spatio-temporal characteristics of air quality and influencing factors in Shandong Province from 2016 to 2020. *Environ. Sci.* https://doi.org/10.13227/j.hjkx.202109020 (2021).
36. Qin, Z. F., Liao, H., Chen, L., Zhu, J. & Qian, J. Fenwei plain air quality and the dominant meteorological parameters for its daily and interannual variations. *Chin. J. Atmos. Sci.* **45**(06), 1273–1291 (2021).
37. Gu, X. *Study on PM2.5 Pollution Characteristics and Regional Transport in Jingzhou City in Recent Years* (Nanjing University of Information Science & Technology, 2021).
38. Guo, L. *Spatial-temporal Distribution Characteristics and Influencing Factors of Air Quality in Hubei Province from 2015 to 2019* (Nanjing University of Information Science & Technology, 2021).
39. Zhu, H. R., Liu, H. N., Zhang, H. L. & Yin, C. J. Characteristics of air quality and its relationship with meteorological factors in Harbin. *J. Meteorol. Environ.* **34**(1), 53–58 (2019).
40. Liang, Q. Y. *Steelmaking Scheduling and Energy Optimization of Steel Enterprises Based on Process Network Simulation* (2021).
41. Gu, W. D., Song, L. G. & Li, Z. X. Research and application of die & mold shop scheduling for the considering bottleneck process outsourcing. *Die Mould Manuf.* **21**(08), 5–9 (2021).
42. Wang, J. M., Li, Y. L., Liu, Z. W. & Liu, J. S. Evolutionary algorithm with precise neighborhood structure for flexible workshop scheduling. *J. Tongji Univ. Nat. Sci.* **49**(03), 440–448 (2021).
43. Liu, D. Genetic algorithm based machining scheduling optimization of key bottleneck process of customized high-end underground equipment. *Manuf. Autom.* **42**(05), 151–156 (2020).

44. Zhu, Y. C. *et al.* Response strategy research of adjustable load demand in composite material industry's production process. *Power Demand Side Manage.* **24**(01), 63–67 (2022).
45. Xie, Z. Q., Zhou, W. & Yu, Z. R. Integrated scheduling algorithm for dynamic adjustment of equipment maintenance start time. *J. Mech. Eng. Chin. Ed.* **57**(04), 240–246 (2021).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant No. 51779211, Grant No. 51209178), and the Sichuan Science and Technology Program (Grant No. 2019YJ0233). We deeply appreciate the editors' and anonymous reviewers' significant comments and suggestions for adding to the quality of the study.

Author contributions

Y.L. contributed to the conception of the study, selected the methodology and wrote and main manuscript text. P.W. wrote the manuscript text, contributed significantly to experiment and prepared Figs. 3, 4, 5, 6, 7 and 8 and Tables 2, 3, 4, 5, 6 and 7. Y.L. contributed to the writing-reviewing and editing. L.W. contributed to the data preprocessing and prepared the Figs. 1, 2 and Table 1. X.D. performed the data analyses and parameter optimization. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

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