scientific reports

Check for updates

OPEN

Calibration of miniature air quality detector monitoring data with PCA–RVM–NAR combination model

Bing Liu[1]✉ & Yirui Zhang[2]

The development of miniature air quality detectors makes it possible for humans to monitor air quality in real time and grid. However, the accuracy of measuring pollutants by miniature air quality detectors needs to be improved. In this paper, the PCA–RVM–NAR combined model is proposed to calibrate the measurement accuracy of the miniature air quality detector. First, correlation analysis is used to find out the main factors affecting pollutant concentrations. Second, principal component analysis is used to reduce the dimensionality of these main factors and extract their main information. Thirdly, taking the extracted principal components as independent variables and the observed values of pollutant concentrations as dependent variables, a PCA–RVM model is established by the relevance vector machine. Finally, the nonlinear autoregressive neural network is used to correct the error and finally complete the establishment of the PCA–RVM–NAR model. Root mean square error, goodness of fit, mean absolute error and relative mean absolute percent error are used to compare the calibration effect of PCA–RVM–NAR model and other commonly used models such as multiple linear regression model, support vector machine, multilayer perceptron neural network and nonlinear autoregressive models with exogenous input. The results show that, no matter which pollutant, the PCA–RVM–NAR model achieves better calibration results than other models in the four indicators. Using this model to correct the data of the miniature air quality detector can improve its accuracy by 77.8–93.9%.

Certain air pollutants, such as $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, O_3 ("two dusts and four gases") can affect human health and cause respiratory diseases and cardiovascular diseases[1–3]. According to statistics, more than 3 million people die worldwide due to air quality problems every year[4,5]. Therefore, obtaining air pollutant concentration information is very necessary to control air pollution and prevent health problems caused by air pollution.

Air quality monitoring platform. Many large cities in developed countries have established some air quality monitoring stations (national control points) in order to obtain information on the concentration of air pollutants. The concentrations of pollutants monitored by these air quality monitoring stations are relatively accurate. However, due to the high cost of establishing monitoring stations and high maintenance costs, the deployment of monitoring stations is relatively sparse. Another disadvantage of national control point monitoring is that the release of data is delayed, making it difficult to monitor the concentration of air pollutants in the entire region in real time. The development of miniature air quality detectors effectively overcomes these shortcomings of reference monitoring stations. The miniature air quality detector has low production and maintenance costs and is easy to install, so it can realize grid deployment and control of specific areas. For this specific areas where the miniature air quality detector is installed for the convenience of monitoring, this paper calls them self-built points. Another advantage of the miniature air quality detector is that it is easy to read the readings, so it can realize real-time monitoring of the concentration of air pollutants[6,7]. In addition, while monitoring the concentration of air pollutants, it can also monitor meteorological parameters such as temperature, humidity, wind speed, air pressure, and precipitation in the region.

Electrochemical sensors are one of the core components of many miniature air quality detectors. It works by reacting with the measured gas and producing an electrical signal proportional to the gas concentration. The

[1]Public Foundational Courses Department, Nanjing Vocational University of Industry Technology, Nanjing 210023, China. [2]School of Intelligent Manufacturing, Sanmenxia Polytechnic, Sanmenxia 472000, China. ✉email: Liub1@niit.edu.cn

gas reacts with the sensor through the tiny capillary-shaped openings and reaches the electrode surface, so that an appropriate amount of gas reacts with the sensing electrode to form a sufficient electrical signal, and finally achieve the purpose of monitoring. The miniature air quality detector will have zero drift or span drift after a period of use. In addition, unconventional pollutants in the air, weather factors, etc. will also cause errors in the measurement of the miniature air quality detector[8]. Therefore, it is very meaningful to establish a pollutant concentration prediction model to calibrate the self-built point data.

Introduction to air quality prediction model. At present, many researchers have studied air quality prediction models. The main research methods are divided into two categories: chemical mechanism prediction and statistical model prediction. The chemical mechanism prediction is to quantitatively describe the changes of atmospheric pollutants in a certain area by using the numerical method of atmospheric dynamics and comprehensively considering the atmospheric physical and chemical mechanism[9,10]. Chemical mechanism prediction has the advantages of multi-scale and openness, but the main disadvantage is that the uncertainty of pollutant emission sources is large, the calculation time is long, and the prediction accuracy is not high. Statistical model forecasting first uses statistical methods to screen out meteorological factors that are strongly correlated with air pollution concentrations, and then uses statistical models to establish quantitative relationships between them and air pollution concentrations. Statistical model forecasting has the advantages of simplicity and economy, good forecast timeliness and accuracy, so it is widely used in air quality forecasting.

Traditional statistical forecasting models include Multiple Linear Regression (MLR) model[11–13], grey models[14], hidden Markov models[15,16], time series models[17,18] and so on. These traditional models are simple in structure, strong in interpretability and short in operation time, and are often used in air quality forecasting in recent years. Suriano et al. designed and developed the SentinAir system for field evaluation of sensor performance. In order to evaluate the system function and capability, indoor and outdoor experiments were performed independently. Linear regression (LR) and multiple linear regression models were used to calibrate the ten sensor data. The results show that the calibration effect of the MLR model is better than that of the LR model because it allows the quantification of the interfering effects of temperature, relative humidity and other gases[19]. However, the factors affecting air quality are complex, and it is difficult for these models to accurately reflect the nonlinear relationship between various factors and air quality. With the rise of big data and artificial intelligence, artificial neural networks[20–22] have also been used to predict air quality. Arsic et al. used multiple regression analysis and artificial neural network to predict ground-level ozone concentrations in the close vicinity of the city of Zrenjanin (Serbia). The comparison results show that the artificial neural network has a better effect in monitoring the ozone concentration than the multiple linear regression model[23].

Although the prediction effect of artificial neural network is good, neural network usually requires more data than traditional machine learning algorithms, and the output results are difficult to interpret. Random forest algorithm[24–26] is also commonly used to predict air quality in recent years, but random forest is prone to overfitting in some noisy regression or classification problems. Support Vector Machine (SVM) can cleverly solve small sample, high-dimensional, nonlinear problems, and it follows the principle of structural risk minimization. Suarez Sanchez et al. used 2006–2008 experimental data on air pollutants to create a highly nonlinear model of the air quality in the Aviles urban nucleus (Spain) based on SVM techniques[27]. Liu et al. successfully predicted the concentration of air pollutants in Nanjing with the help of support vector regression machine, and calibrated the measurement data of the miniature air quality detector[28].

However, for the air quality prediction problem, the support vector machine model also has certain shortcomings. First, as the dimension of training samples increases, the model prediction time is prolonged, which seriously restricts the timeliness of the model. Second, there are many parameters in the principle of the support vector regression machine[29,30]. In addition to the kernel function parameters, the penalty factor C and the radius of the insensitive loss area ε will have a greater impact on the accuracy of the model, and it is difficult to establish a high-precision air quality prediction model. To address these issues, a Bayesian framework-based sparse probabilistic learning model relevance vector machine (RVM) is introduced in this paper to predict air quality. The relevance vector machine uses the active correlation decision theory to realize the sparseness of the model, which greatly reduces the amount of calculation, and the time of model prediction is better controlled. In addition, some model parameters can obtain the optimal solution through self-adaptive iteration, and there are few adjustment parameters, which is convenient for model optimization.

The main work of this paper is to find out the influencing factors affecting the concentration of six types of air pollutants through correlation analysis, and then use Principal Component Analysis (PCA) to extract the main information in these influencing factors. Then, these main information are used as input, the concentration of pollutants in the air is used as output, and the air quality prediction model is established with the help of relevance vector machine. Finally, the prediction residuals are corrected by the Nonlinear Autoregressive (NAR) neural network to further improve the prediction accuracy of the model. We call this combined model the PCA–RVM–NAR combined model. In practical applications, this model has achieved good results in air quality prediction, and it can provide a reference model for air quality prediction and data calibration of miniature air quality detectors.

Material and methods

Data source and preprocessing. The emergence and development of miniature air quality detectors provide the possibility for grid and real-time monitoring of air quality. However, its measurement is affected by many factors, so the measurement data will have errors. This paper uses a statistical model to calibrate it. A total of two sets of data (http://www.mcm.edu.cn/html_cn/node/b0ae8510b9ec0cc0deb2266d2de19ecb.html) are used in this study to establish the calibration model of the miniature air quality detectors. The first set of data

Input variable	Ranges	Mean	Standard deviation	Skewness	Kurtosis
$PM_{2.5}$ (μg/m^3)	1 to 216.883	64.127	37.328	0.988	0.701
PM_{10} (μg/m^3)	2 to 443.25	102.391	65.267	1.476	2.862
CO (mg/m^3)	0.05 to 3.895	0.863	0.452	1.463	3.136
NO_2 (μg/m^3)	0.947 to 157.136	45.209	28.403	0.653	− 0.259
SO_2 (μg/m^3)	1 to 651.3	19.397	18.723	12.781	342.11
O_3 (μg/m^3)	0.579 to 259	61.586	40.941	1.091	2.035
Wind speed (m/s)	0.133 to 2.387	0.7	0.346	0.862	0.748
Pressure (Pa)	996.871 to 1039.8	1018.8	8.889	− 0.093	− 0.599
Precipitation (mm/m^2)	0 to 312.1	132.084	87.004	0.245	− 0.728
Temperature (°C)	− 3.882 to 37.944	11.882	8.603	0.625	− 0.399
Humidity (rh%)	10.667 to 100	68.903	21.931	− 0.487	− 0.756

Table 1. Descriptive statistics of pollutant concentrations and meteorological parameters measured by national control point and self-built point after pretreatment.

comes from an air quality monitoring station in Nanjing, which contains 4200 sets of data and is considered accurate in this paper. It recorded the hourly concentration of two dusts and four gases from November 14, 2018 to June 11, 2019. The second set of data is provided by a miniature air quality detector juxtaposed with the air quality monitoring station. It contains 234,717 sets of data, and the interval between each set of data is no more than five minutes. The second set of data includes not only the concentration of two dusts and four gases, but also five meteorological parameters such as temperature, humidity, wind speed, air pressure, and precipitation.

Data preprocessing is the first step in establishing the data correction model of the miniature air quality detector. Data that is more than 3 times the mean value of the left and right nearest neighbors is regarded as an outlier and eliminated in this paper. Then average the measured values of the self-built point within an hour to compare with the data of the national control point, and delete some data that cannot correspond to the self-built point and the national control point. After data preprocessing, a total of 4135 sets of corresponding data are obtained, and Table 1 shows them.

Data exploratory analysis. Exploratory analysis of data can give us a deeper understanding of the interrelationships between variables. In order to more intuitively reflect the relationship between the national control point measurement data and the self-built point measurement data, we average the measurement data on a daily basis and conduct visual analysis[6,9]. It can be seen from Fig. 1 that no matter what kind of pollutants it is, the general trend of the self-built point data and the national control point data is the same, but there are also certain errors. The difference between $PM_{2.5}$ and PM_{10} is relatively small, indicating that the miniature air quality detector has high accuracy in measuring the concentrations of these two types of pollutants. The errors of NO_2 and O_3 in the previous period are relatively large, and the errors in the latter period are relatively small. It may be that the climate has a great influence on the concentration of these two pollutants measured by the miniature air quality detector. The measurement errors of CO and SO_2 are large, indicating that it is difficult to monitor the concentrations of these two pollutants with a miniature air quality detector.

Figure 2 is a line chart of the changes of five meteorological parameters with time. It can be seen that there is abundant rain in this area, the daily average temperature is relatively mild, the daily average air pressure is stable between 1000–1050 Pa, and the air humidity and wind speed change more obviously. Further discussion and analysis are needed to find the relationship between meteorological parameters and the concentrations of the six types of pollutants.

The measurement error of the miniature air quality detector may have a certain relationship with the meteorological parameters, and there are obvious differences in the meteorological parameters in different seasons. We have drawn a boxplot[31] of the six categories of pollutants by season as shown in Fig. 3. It can be seen that the concentrations of $PM_{2.5}$, PM_{10}, CO, and SO_2 are the highest in autumn and winter. The main reason is that the temperature in autumn and winter is lower, and it is difficult for the lower air and upper air to generate convection, resulting in slower diffusion of pollutants. The reason for the high concentration of O_3 in summer is the strong solar radiation and high temperature in summer, which is easy to cause photochemical smog and secondary ozone production. The slightly higher NO_2 concentration in spring may be related to thunderstorms. In addition, the errors between the measured and actual values of the six types of pollutants have obvious differences in the four seasons, indicating that meteorological parameters will affect the measurement of the miniature air quality detector.

Correlation analysis. The concentration of pollutants in the air is an important criterion for evaluating air quality. Different geographical environments have different influence factors on the concentration of air pollutants. The Pearson correlation coefficient is used in this paper to screen the main factors affecting air quality[25,32]. Equation (1) is its expression, where x_i is the value of the first variable, y_i is the value of the second variable, $\overline{x}$ is the mean of x, $\overline{y}$ is the mean of y, and n represents the number of samples. The value range of the Pearson correlation coefficient is [− 1, 1], and the larger its absolute value, the stronger the correlation between the two variables.

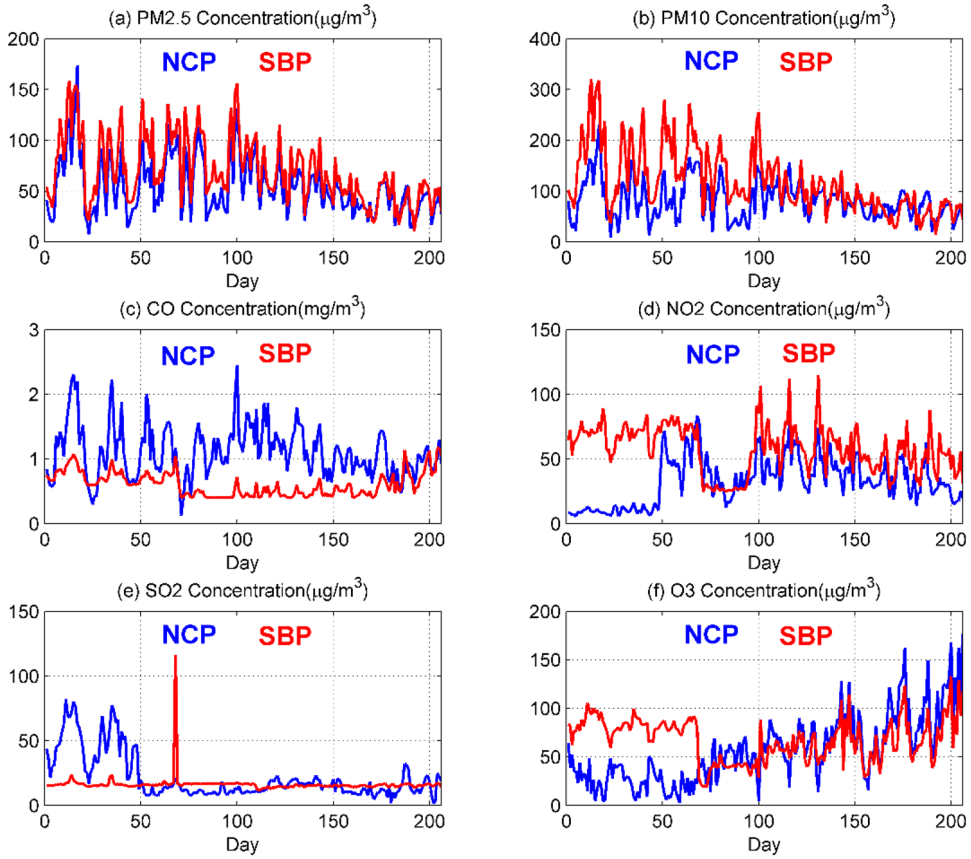

Figure 1. Comparison of daily average data of six types of pollutants at national control point (NCP) and self-built point (SBP). Figures are generated using Matlab (Version R2016a, https://www.mathworks.com/) [Software].

It can be seen from Table 2 that under the premise of the significant level 0.05, except for NO_2 concentration and temperature, the other variables are significantly correlated with each other. The positive correlation between $PM_{2.5}$ concentration and PM_{10} concentration is the highest, and the correlation coefficient is 0.89, indicating that they have the same trend of change. The negative correlation between temperature and air pressure is the highest, and the correlation coefficient is −0.85, indicating that there is a reverse trend between them.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1}$$

Establishment of sensor calibration model

Introduction to basic principles. The relevance vector machine is a sparse probability model similar to the support vector machine proposed by Tipping in 2000. It is a new supervised learning method. The model combines theories such as Markov's, Bayes's principle and maximum likelihood. Due to the high sparsity of the algorithm and the structure based on probabilistic learning, RVM can enable us to obtain high prediction accuracy. In addition, compared with the support vector machine, it greatly reduces the number of kernel functions involved in the prediction calculation and reduces the prediction calculation time. RVM also has the advantages of probabilistic prediction, automatic parameter setting and arbitrary use of kernel functions[33–35].

$$t_n = y(x_n;\ \omega) = \sum_{n=1}^{N} \omega_n k(x,\ x_n) + \varepsilon_n \tag{2}$$

$$p(t_n|x_n) = N\left(t_n|y(x_n),\ \sigma^2\right) \tag{3}$$

$$p\left(t|\omega,\ \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\|t - \Phi\omega\|^2\right\} \tag{4}$$

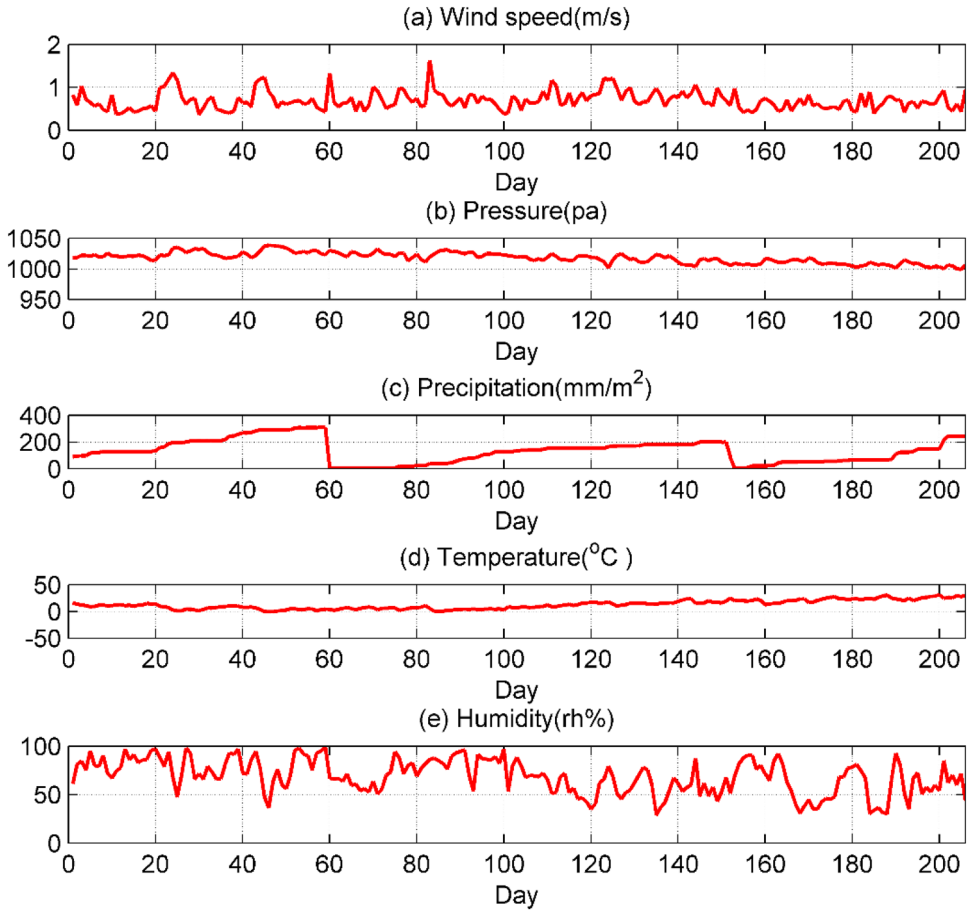

Figure 2. Variation of the daily average value of five meteorological parameters.

$$p(\omega|\alpha) = \prod_{n=0}^{N} N\left(\omega_n|0,\ a_n^{-1}\right) \tag{5}$$

$$p(t_*|t) = \int p\left(t_*|\omega,\ \alpha,\ \sigma^2\right)p\left(\omega,\ \alpha,\ \sigma^2|t\right) \times d\omega d\alpha d\sigma^2 \tag{6}$$

The relevance vector machine is not constrained by the Mercer condition when selecting the kernel function, it can achieve binary classification and probability output, and the running speed is fast. Let the training data samples be $\{x_n,\ t_n|n = 1,\ 2, \ldots,\ N\}$, where x_n is the input value,t_n is the output value, N is the number of data samples, Eq. (2) is the expression of the regression model, where $k(x,\ x_n)$ is the kernel function, $\omega = \{\omega_n\}_{n=0}^{N}$ is the weight value of each input quantity, ε_n is the data noise and obeys the Gaussian distribution, $\varepsilon_n \sim N(0,\ \sigma^2)$, σ^2 is an unknown quantity. Thus, the Eq. (3) that satisfies the Gaussian distribution is obtained, where t_n is related to $y(x_n)$ and σ^2, and t_n is independent of each other. Equation (4) is the likelihood function of the training sample set, where $t = \{t_1,\ t_2, \ldots,\ t_N\}^T$,$\omega = \{\omega_0,\ \omega_1, \ldots,\ \omega_N\}^T$,$\Phi = [\varphi(x_1),\ \varphi(x_2), \ldots,\ \varphi(x_N)]^T$ is an $N \times (N+1)$ matrix, and the expression of each column in the matrix is $\varphi(x_n) = [1,\ k(x_n,\ x_1),\ k(x_n,\ x_2), \ldots,\ k(x_n,\ x_N)]^T$. The hyperparameter $\alpha = \{\alpha_0,\ \alpha_1, \ldots,\ \alpha_N\}^T$ is introduced to solve ω and σ^2 in Eq. (4), ω_n satisfies the Gaussian distribution, and its expression is Eq. (5). Equation (6) is the expression of the input value x_* and the output value t_* of the prediction data set. According to the Bayesian and Markov properties and Eq. (6), Eq. (7) can be obtained by simultaneous simplification, where Eqs. (8)–(10) represent the covariance and weight mean.

$$p\left(\omega|t,\ \alpha,\ \sigma^2\right) = (2\pi)^{-\frac{N+1}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\omega-\mu)^T\Sigma^{-1}(\omega-\mu)\right\} \tag{7}$$

$$\Sigma = (\sigma^2\Phi^T\Phi + A)^{-1} \tag{8}$$

$$\mu = \sigma^{-2}\sum\Phi^T t \tag{9}$$

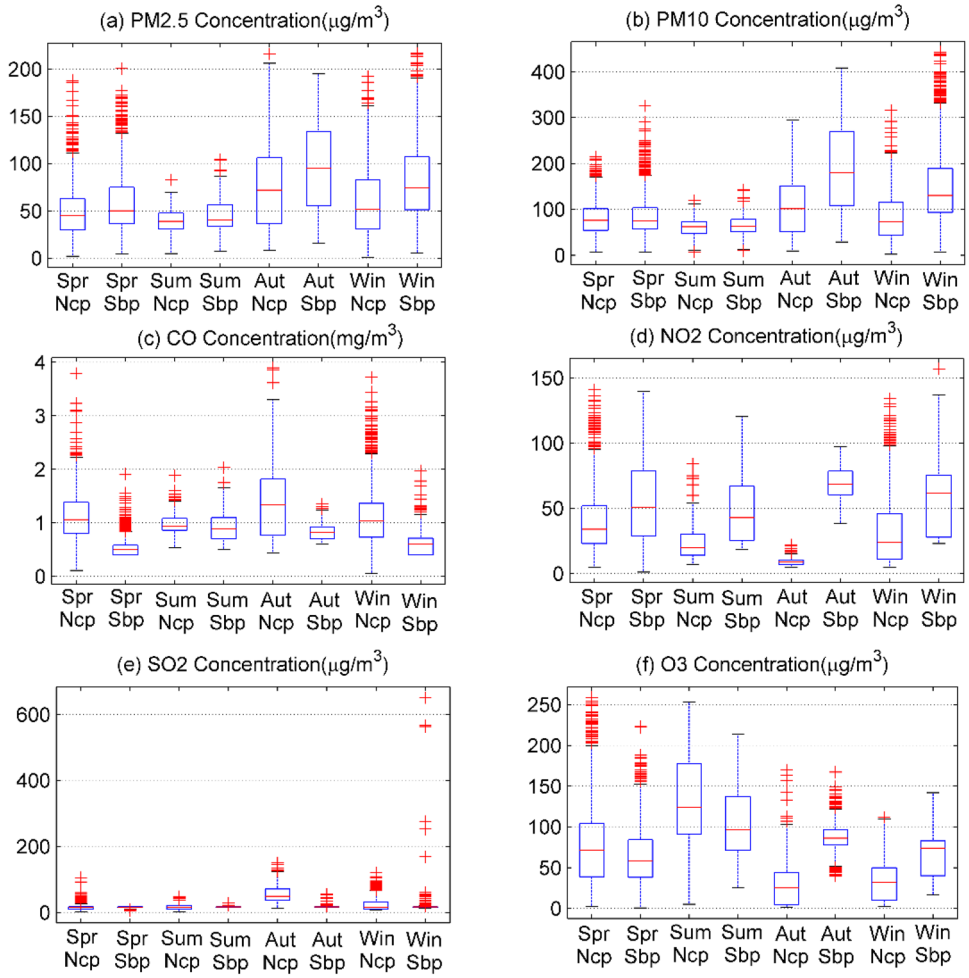

Figure 3. Comparing the concentration of six types of pollutants at national control sites (Ncp) and self-built sites (Sbp) on a seasonal basis.

Variable	$PM_{2.5}$	PM_{10}	CO	NO_2	SO_2	O_3	Wind speed	Pressure	Precipitation	Temperature	Humidity
$PM_{2.5}$	1.00	0.89*	0.66*	0.26*	0.29*	−0.26*	−0.23*	0.89*	−0.70*	−0.16*	0.18*
PM_{10}		1.00	0.63*	0.34*	0.35*	−0.19*	−0.18*	0.38*	−0.10*	−0.03*	−0.09*
CO			1.00	0.30*	0.31*	−0.27*	−0.31*	−0.07*	0.08*	−0.05*	0.22*
NO_2				1.00	−0.34*	−0.26*	−0.36*	−0.10*	−0.14*	−0.02	−0.11*
SO_2					1.00	−0.28*	−0.19*	0.19*	0.27*	−0.10*	0.11*
O_3						1.00	0.39*	−0.45*	−0.12*	0.68*	−0.62*
Wind speed							1.00	0.09*	0.06*	0.07*	−0.32*
Pressure								1.00	0.23*	−0.85*	0.15*
Precipitation									1.00	−0.14*	0.86*
Temperature										1.00	−0.49*
Humidity											1.00

Table 2. Pearson linear correlation coefficient between the concentrations of six types of air pollutants measured at national control point and five meteorological parameters measured at self-built point (Band * indicates significant correlation at a significant level of 0.05).

$$A = diag(a_0,\ a_1, \ldots,\ a_N) \tag{10}$$

Equation (11) can be obtained after the maximum likelihood function is simplified. Find the partial derivative of α and σ^2 in Eq. (11), and let them be 0 to establish two equations. After simplification, Eqs. (12)–(13) can be obtained, where $\gamma_n = 1 - \alpha_n \Sigma_{nn}$, Σ_{nn} is the element of row n and column n of Σ. α and σ^2 are obtained through

the update iteration of Eqs. (12)–(13). At the same time, the weight posterior mean μ and the covariance matrix Σ change continuously until the convergence condition or the maximum number of iterations is satisfied. In the iterative process, new optimal solutions α_{MP} and σ^2_{MP} will be obtained, and most of the weights will approach 0, and the corresponding basis functions will be ignored, which reflects the sparsity of the RVM model, and other weights will approach a constant, and the corresponding basis functions are called relevance vectors. The expected value y_* and the noise variance σ^2 (Eqs. (14)–(15)) can be obtained by predicting the relationship between the input value x_* and the output value t_* of the data set (Eq. (6)), where x_* is the sample to be predicted, y_* is the mean of the output value t_*.

$$p(t|\alpha,\ \sigma^2) = \int p(t|\omega,\ \sigma^2)p(\omega|\alpha)d\omega = (2\pi)^{-\frac{N}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[\frac{1}{2}\times(\omega-\mu)^T\Sigma^{-1}(\omega-\mu)\right] \tag{11}$$

$$\alpha_n^{new} = \frac{\gamma_n}{\mu_n^2} \tag{12}$$

$$(\sigma^2)^{new} = \frac{\|t-\Phi\mu\|^2}{N-\sum_n^N \gamma_n} \tag{13}$$

$$y_* = \mu^T\varphi(x_*) \tag{14}$$

$$\sigma_*^2 = \sigma_{MP}^2 + \varphi(x_*)^T\Sigma\varphi(x_*) \tag{15}$$

PCA–RVM model construction. Air quality is affected by a variety of factors, and the relationship between the influencing factors is intricate. The variables input to the model have a great relationship with the accuracy of prediction. According to the previous correlation analysis, it can be seen that the pollutant concentration measured by the miniature air quality detector and the five meteorological parameters are significantly related to the air quality, so they all have a certain impact on the air quality. In addition, since the input variables also affect each other, if all variables are directly input into the relevance vector machine, some repetitive information will be input into the model, which not only makes the training time of the model longer, but also makes the model generalization ability deteriorates.

Principal component analysis is a method of data dimensionality reduction and denoising. It converts a series of components that are originally related in the system into several uncorrelated components through orthogonal transformation, and this group of components after conversion is called the principal component. Then, according to the contribution of each component to the data system, the principal components are recombined to highlight the hidden features in the original data to construct a mapping matrix, and then the original data is transformed by the mapping matrix to achieve the purpose of denoising[28]. The process of principal component analysis is generally as follows: (i) Standardize the original data; (ii) Calculate the correlation coefficient matrix R; (iii) Calculate the eigenvalues and eigenvectors; (iv) Select $p\,(p \le m)$ principal components and calculate the comprehensive evaluation value. In this paper, the principle of extracting the number of principal components is that the cumulative contribution rate exceeds 99%.

Figure 4 shows the principal component contribution rate and the principal component cumulative contribution rate after dimension reduction by principal component analysis. It can be seen that the contribution rate of the first principal component reaches 29.2%, and the contribution rate of the second, third and fourth principal components also exceeds 10% respectively, and the cumulative contribution rate of the first four principal components exceeds 70%. In addition, it can be seen from the broken line in the figure that the cumulative contribution rate of the first 9 principal components has exceeded 99%, which is in line with the principle of the number of extracted principal components. It shows that PCA is effective for dimensionality reduction of air quality data, and can provide more reliable input for subsequent prediction.

After the principal component dimension reduction is performed on the original data, the first 9 principal components after dimension reduction are used as input independent variables, and the predicted values of six types of pollutant concentrations are used as output variables, and the relevance vector machine is used to build the air quality prediction model. This combined model is called the PCA–RVM model in this paper. Since the construction process of the six types of pollutant prediction models is similar, we take $PM_{2.5}$ concentration as an example, and other pollutant concentration prediction models can be obtained similarly.

We randomly divided 4135 groups of data, 3000 groups are selected as the training set, and the other 1135 groups are selected as the test set, and used Matlab2016a for modeling. For the training of the RVM model, according to the RVM regression principle, it can be seen that the hyperparameter α and the noise σ^2 are not sensitive to the initial value, and the optimal value can be obtained by iterative adaptation. The kernel function of the relevance vector machine uses the Gaussian kernel function, because the Gaussian kernel function can obtain a very smooth estimation[36,37]. The value of the model kernel function width γ is obtained by the grid optimization method, the optimization interval is [0.5, 10], and the step size is 0.5. Equation (16) is the expression of Root Mean Square Error (RMSE), where y_i represents the target value, w_i represents the model predicted value. In this paper, the RMSE between the target value of the sample training set and the model predicted value is used as the objective function for optimization. During the training process, for each parameter value, we train the

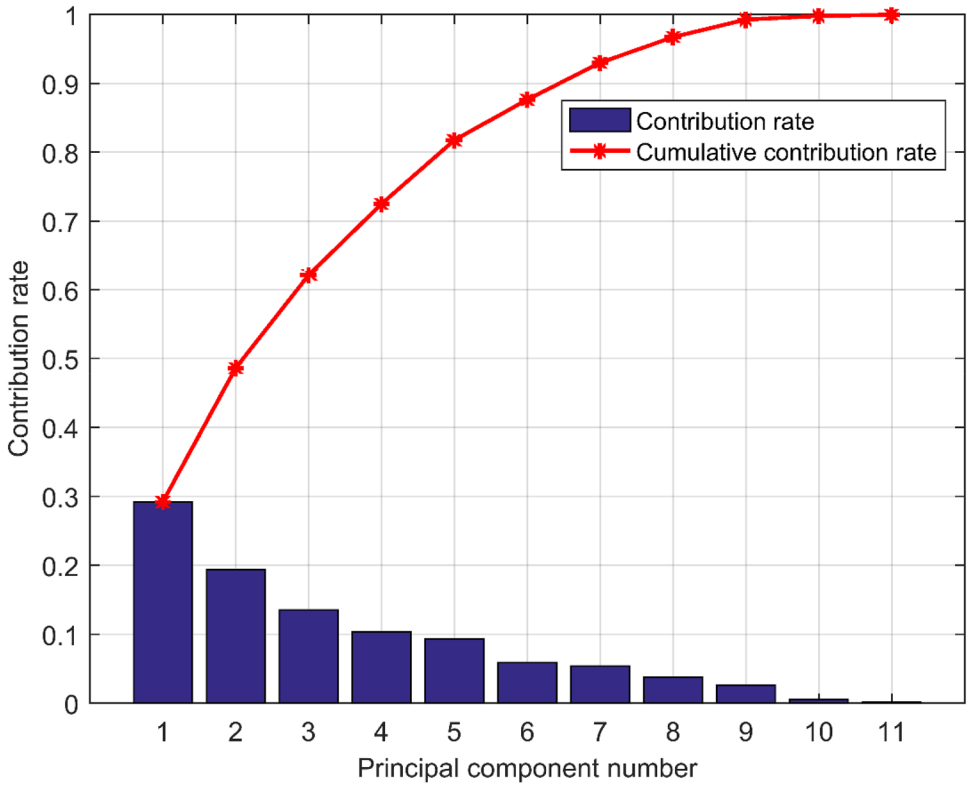

Figure 4. The principal component contribution rate and the cumulative contribution rate of the self-built point measurement data.

model 10 times, and average the output values of the 10 training times as the final output of the model. Through empirical analysis, when $\gamma = 1.5$ is the optimal value, the PCA–RVM air quality prediction model is established.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - w_i)^2} \tag{16}$$

PCA–RVM–NAR model construction. The PCA–RVM model can be used to calibrate the miniature air quality detector data. It can be seen from Fig. 5 that the residual of the $PM_{2.5}$'s PCA–RVM model is greatly improved compared to the residual of the self-built point, whether it is the training set or the test set. In the training set, the residual of the model is concentrated in $[-10, 10]$, and the absolute value of the maximum residual is 32.06 μg/m^3, while the residual of the self-built point is concentrated in $[-40, 20]$, and the absolute value of the maximum residual is 110.44 μg/m^3. In the test set, the residual of the model is concentrated in $[-20, 20]$, and the absolute value of the maximum residual is 67.2 μg/m^3, while the residual of the self-built point is concentrated in $[-50, 25]$, and the absolute value of the maximum residual is 90 μg/m^3. The PCA–RVM model performs well in both the training set and the test set, indicating that the generalization ability of the model is good.

Although the $PM_{2.5}$ concentration prediction effect of the PCA–RVM model is good, a set of time series residual data is obtained, and some residuals in the model are still high. Autoregressive integrated moving average model and NAR neural network model are commonly used to deal with time series data. This paper uses a NAR neural network to further mine the residual information.

The NAR neural network belongs to the dynamic neural network and can be expressed by Eq. (17), where $y(t)$ is the output value at the current moment, $y(t-1)$, $y(t-2), \ldots,$ $y(t-d)$ are the output value at the historical moment, and d is the delay order. NAR neural network consists of input layer, hidden layer and output layer[38]. For the selection of the number of neurons in the hidden layer and the order of input delay, we also use grid optimization to optimize in $[5, 15] \times [1, 5]$. The training function of the NAR neural network adopts the default Levenberg–Marquardt (LM) algorithm in the Neural Net Time Series in Matlab. The core idea of the LM algorithm is to use the Jacobian matrix to replace the solution of the positive definite matrix in the gradient learning algorithm to optimize the operation efficiency of the training network. For the objective function, RMSE is also chosen as the objective function, and the final output is also obtained by averaging 10 times of training. After optimization, it is found that the optimal value is when the number of neurons in the hidden layer is 9 and the delay order is 3. The structure of the NAR neural network is shown in Fig. 6, where w is the weight of the neural network model, and b is the threshold of the neural network model. The PCA–RVM–NAR air quality prediction model has now been constructed.

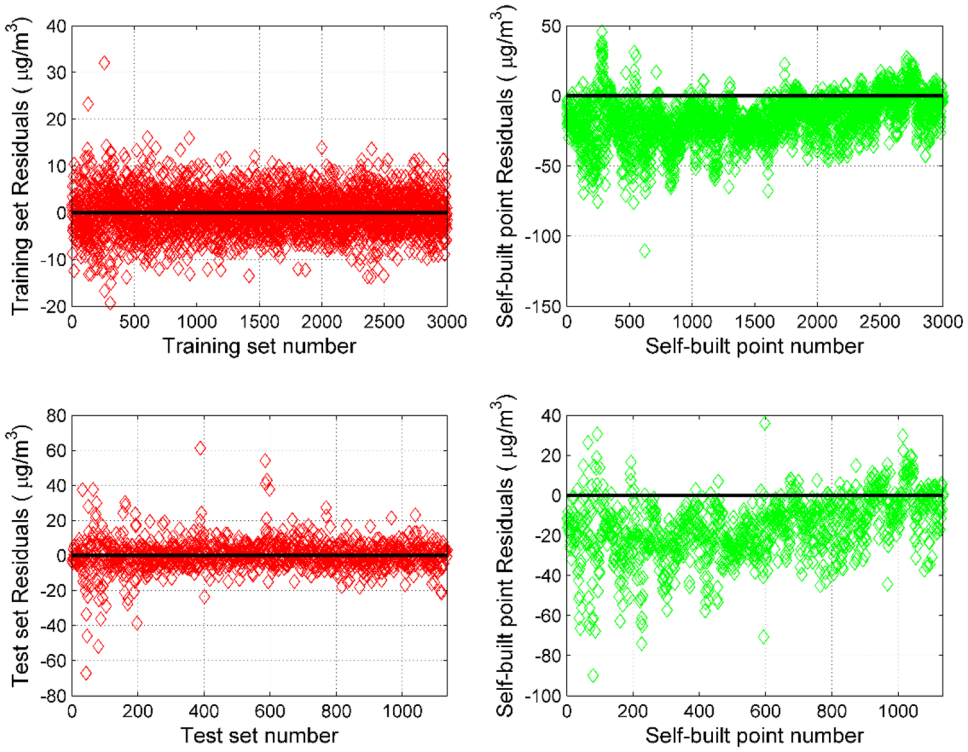

Figure 5. Comparison of $PM_{2.5}$'s PCA–RVM model residuals and self-built point residuals. The comparison of the training set is on the left, and the comparison of the test set is on the right.

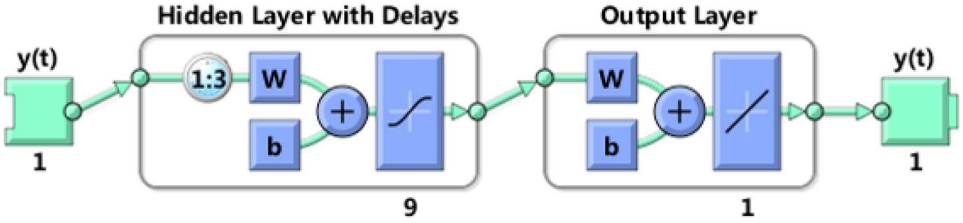

Figure 6. The frame structure of the PCA–RVM–NAR model, where the input is the residual of the PCA–RVM model. This network has 1 inputs, 1 hidden layer with 9 hidden neurons, 3 input delay orders, and 1 linear output layer leading to 1 output.

$$y(t) = f(y(t-1),\ y(t-2), \ldots,\ y(t-d)) \tag{17}$$

Figure 7 shows the measured value of $PM_{2.5}$ concentration at the national control point and the predicted value of PCA–RVM–NAR combined model. It can be seen that the change trend of the two is consistent, and the correlation coefficient between the measured value of the national control point and the predicted value of the PCA–RVM–NAR model is greater than 0.95 in both the training set and the test set. Both models in the training set and the test set passed the significance test at the significance level of 0.01. The regression coefficients in the two regression models are also close to 1, indicating that the PCA–RVM–NAR model is more accurate in $PM_{2.5}$ concentration prediction.

Residual analysis is also a necessary step in statistical modeling[12,15]. It can be seen from the residual analysis diagram in Fig. 8 that most of the residuals of the PCA–RVM–NAR model are concentrated in [− 10, 10], and the residuals are evenly distributed near the zero point. The absolute values of residuals at the 172nd and 1481st sample points are larger than 50 μg/m^3. We checked the corresponding data, and the $PM_{2.5}$ concentration measured at the national control point has changed greatly at this moment, indicating that the measurement residual of the model will increase when the pollutant concentration changes rapidly. In order to better display the residual characteristics of the model, this paper deletes these two points and draws a residual histogram. From the histogram we can see that the residuals are roughly normally distributed. A total of 3981 sets of data residuals are located in [− 10,10], exceeding 96.2%, and only 27 sets of residuals whose absolute value exceeds

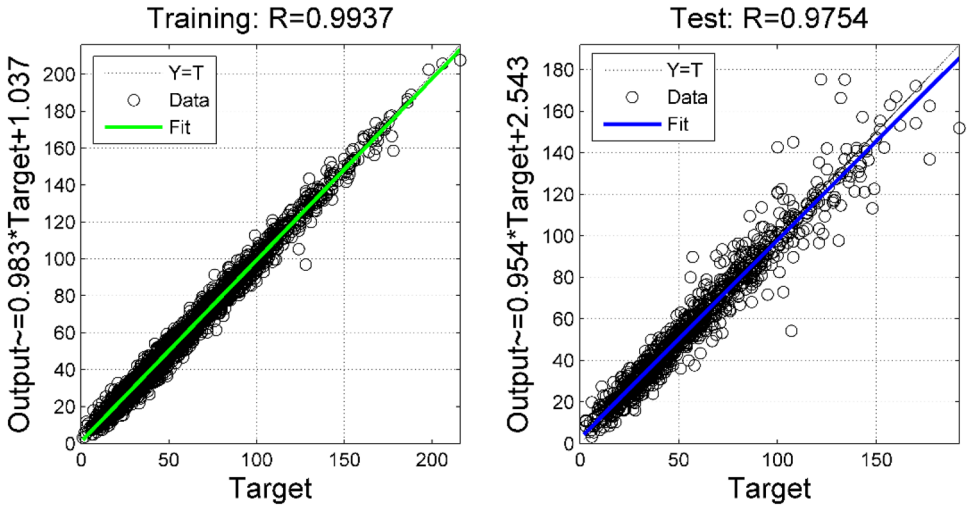

Figure 7. The prediction effect of $PM_{2.5}$'s PCA–RVM–NAR model on the training set and test set.

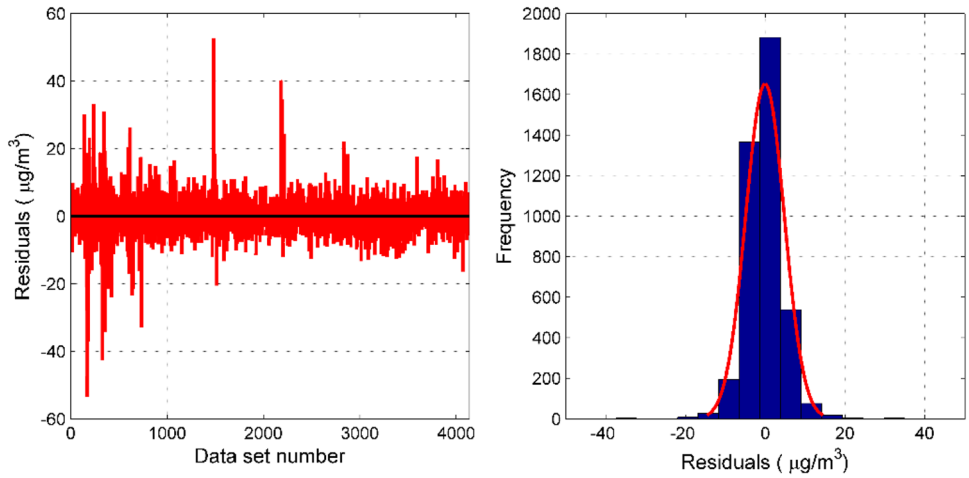

Figure 8. Residual test of $PM_{2.5}$'s PCA–RVM–NAR model. The residual plot of the PCA–RVM–NAR model is seen on the left. The histogram of the residuals is seen on the right.

20, do not exceed 0.5% of the total. In addition, 91.3% of the data prediction residuals are within 20%, and 73.3% of the data prediction residuals are within 10%.

Discussion

The PCA–RVM–NAR combination model can calibrate the $PM_{2.5}$ measurement concentration of the miniature air quality detector, and has achieved good results. In addition, multiple linear regression model, Support Vector Regression machine (SVR), Multilayer Perceptron neural networks (MLP) and Nonlinear Autoregressive models with Exogenous Inputs (NARX) can also calibrate the $PM_{2.5}$ measurement concentration of the miniature air quality detector[39–41]. In order to visually compare the calibration effects of various models, Taylor diagram is used in this paper to compare them.

Taylor diagram is a visual chart that can simultaneously represent three indicators of correlation coefficient, standard deviation and centered root mean square difference. The scatter points in the Taylor diagram represent different models, the radial line represents the correlation coefficient, the horizontal and vertical axes represent the standard deviation, and the dashed line represents the centered root mean square difference. Equation (1), Eqs. (18)–(19) are their expressions, where y_i represents the true value, w_i represents the model predicted value, $\overline{y}$ represents the mean value of y, and $\overline{w}$ represents the mean value of w. Taylor diagram can compare the relationship between model indicators from multiple perspectives and dimensions. It can be seen from Fig. 9 that the distance between the self-built point and the observation point (national control point) is the farthest, indicating that the $PM_{2.5}$ measurement accuracy of the self-built point is the lowest, and the measurement value of the self-built point needs to be calibrated. Multiple linear regression model, multilayer perceptron neural network and NARX neural network can calibrate the $PM_{2.5}$ measurement accuracy of self-built point, but the calibration

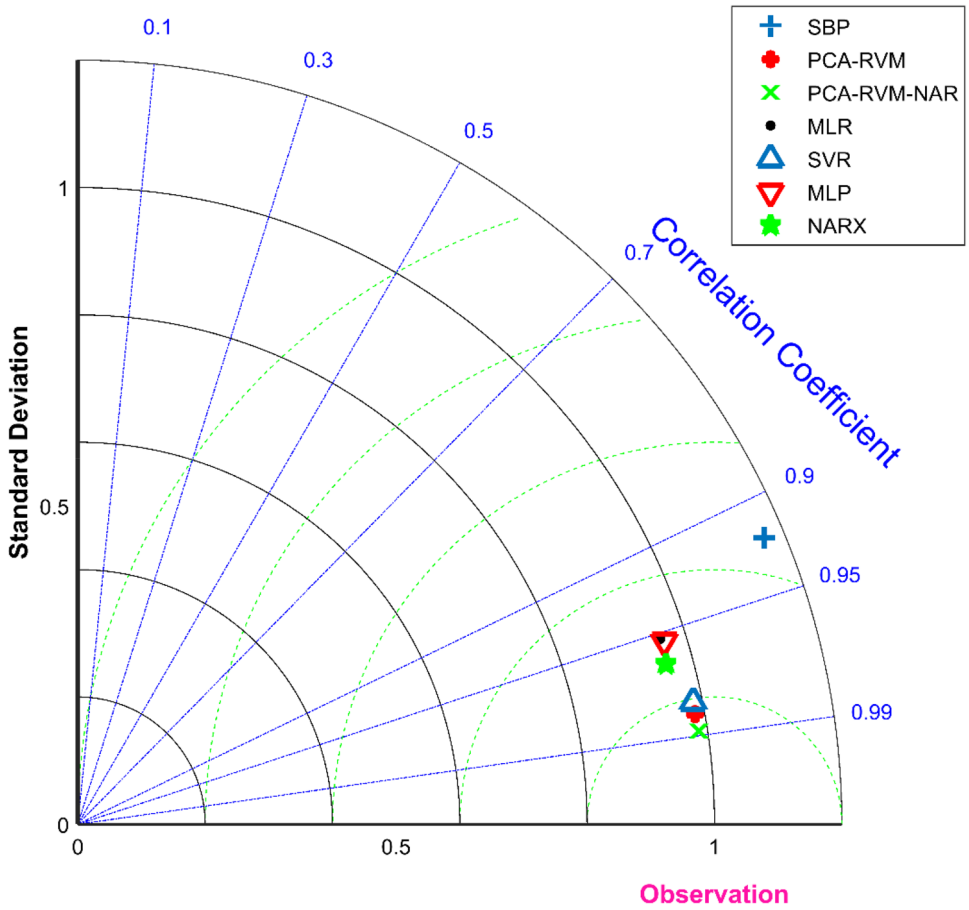

Figure 9. Taylor diagrams of the predicted $PM_{2.5}$ concentration for the six models and the measured value of the self-built point, where SBP represents the self-built point.

accuracy needs to be improved. The calibration effect of the support vector machine and the PCA–RVM model is better, but in general, the PCA–RVM–NAR combined model given in this paper performs the best in the calibration of $PM_{2.5}$ measurement accuracy.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(w_i - \overline{w})^2} \tag{18}$$

$$E' = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(y_i - \overline{y}) - (w_i - \overline{w})]^2} \tag{19}$$

In order to comprehensively compare the accuracy of the PCA–RVM–NAR model with other commonly used air quality prediction models, four commonly used indicators are used to compare the models in this paper[32,39]. These four indicators include Root Mean Square Error, Goodness of fit (R^2), Mean Absolute Error (MAE) and relative Mean Absolute Percent Error (MAPE). Equation (16), Eqs. (20)–(22) are their expressions, where y_i represents the measured values of six types of pollutants in the national control point, and w_i represents the predicted values of various prediction models. The comparison of each indicator of two dusts and four gases is shown in Tables 3, 4, 5 and 6. It can be seen that the error of the self-built point is not only the largest in the $PM_{2.5}$ measurement concentration, but also the largest in other pollutants. It should be noted that the R^2 of some pollutants is negative, which is caused by the large measurement error of the self-built point. This indicator is eliminated when the calculation model improves the measurement accuracy. The support vector regression machine is obviously better than the MLR, MLP and NARX models in each evaluation index value, which shows that the SVR is more suitable for the calibration of the monitoring data of the miniature air quality detector. The performance of correlation vector machine is better than that of support vector regression machine in each evaluation indicator, and the PCA–RVM–NAR model proposed in this paper has the best performance in four indicators of six pollutants. The PCA–RVM–NAR model has the lowest improvement in the measurement accuracy of the

Input variable	Self-built point	PCA–RVM	PCA–RVM–NAR	MLR	SVR	MLP	NARX
$PM_{2.5}$	22.436	5.873	4.970	10.149	8.649	10.777	8.800
PM_{10}	66.263	10.605	7.740	20.050	11.656	19.126	13.911
CO	0.679	0.131	0.085	0.344	0.175	0.304	0.158
NO_2	37.183	6.597	5.049	16.653	7.725	13.216	8.081
SO_2	26.24	4.018	2.843	15.305	4.116	9.984	5.104
O_3	45.673	8.669	6.627	21.451	11.304	18.603	12.477

Table 3. The RMSE of self-built point and various air quality prediction models, in which national control point is used as comparison object.

Input variable	Self-built point	PCA–RVM	PCA–RVM–NAR	MLR	SVR	MLP	NARX
$PM_{2.5}$	0.551	0.969	0.978	0.908	0.933	0.907	0.931
PM_{10}	− 1.076	0.947	0.972	0.810	0.938	0.827	0.909
CO	− 0.929	0.929	0.970	0.506	0.872	0.708	0.895
NO_2	− 1.333	0.927	0.957	0.532	0.899	0.752	0.890
SO_2	− 0.726	0.960	0.980	0.413	0.958	0.786	0.935
O_3	0.094	0.967	0.981	0.800	0.945	0.864	0.932

Table 4. The R^2 of self-built point and various air quality prediction models, in which national control point is used as comparison object.

Input variable	Self-built point	PCA–RVM	PCA–RVM–NAR	MLR	SVR	MLP	NARX
$PM_{2.5}$	18.181	4.032	3.430	7.042	5.821	7.763	6.070
PM_{10}	50.151	6.958	4.877	13.689	7.080	13.184	9.218
CO	0.549	0.089	0.058	0.263	0.110	0.237	0.100
NO_2	29.838	4.189	3.144	12.641	4.658	9.991	4.924
SO_2	12.867	2.090	1.459	10.206	2.116	7.246	2.684
O_3	36.63	5.702	4.266	16.582	7.647	14.396	7.948

Table 5. The MAE of self-built point and various air quality prediction models, in which national control point is used as comparison object.

Input variable	Self-built point	PCA–RVM	PCA–RVM–NAR	MLR	SVR	MLP	NARX
$PM_{2.5}$	0.447	0.104	0.083	0.166	0.133	0.185	0.151
PM_{10}	0.887	0.118	0.086	0.221	0.107	0.210	0.147
CO	0.478	0.097	0.058	0.319	0.112	0.283	0.096
NO_2	2.129	0.173	0.130	0.639	0.170	0.471	0.1816
SO_2	0.685	0.134	0.090	0.741	0.131	0.530	0.161
O_3	4.322	0.294	0.290	1.261	0.373	1.002	0.428

Table 6. The MAPE of self-built point and various air quality prediction models, in which national control point is used as comparison object.

miniature air quality detector is the RMSE of $PM_{2.5}$. The measurement accuracy of this detector improves of the 77.8% considering the self-built point (RMSE = 22.436) and the PCA–RVM–NAR model (RMSE = 4.97). The PCA–RVM–NAR model has the highest improvement in the measurement accuracy of the miniature air quality detector is the MAPE of NO_2. The measurement accuracy of this detector improves of the 93.9% considering the self-built point (MAPE = 2.129) and the PCA–RVM–NAR model (MAPE = 0.13).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - w_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{20}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - w_i\right| \tag{21}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - w_i}{y_i}\right| \tag{22}$$

Conclusions

Air quality is related to the quality of human life[3,4]. The main pollutants affecting air quality are $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2 and O_3. Real-time monitoring of pollutant concentrations is of great help for the government and relevant departments to take corresponding measures to pollution sources. The development of miniature air quality detectors is very helpful for human beings to monitor air quality in real time and grid. However, due to various reasons, the measurement accuracy of the miniature air quality detector needs to be improved. The PCA–RVM–NAR model proposed in this study successfully improved the measurement accuracy of the miniature air quality detector by 77.8–93.9%. In addition, the PCA–RVM–NAR model performs very well on both the training set and the test set, indicating that it has a strong generalization ability. It uses a total of 4135 sets of data, and the data of four seasons are covered in the model, which also shows that the model has good stability. However, air quality is affected by many factors. The PCA–RVM–NAR model does not consider other external factors when it is established. Future work can try to introduce more external factors to improve the accuracy of the model. In addition, the climate in different regions is different, and the suitability of the model in other regions also needs further verification.

Data availability

The data that support the findings of this study are available from the corresponding author B.L. upon reasonable request.

Received: 21 February 2022; Accepted: 25 May 2022
Published online: 04 June 2022

References

1. Corrigan, A. E., Becker, M. M., Neas, L. M., Cascio, W. E. & Rappold, A. G. Fine particulate matters: The impact of air quality standards on cardiovascular mortality. *Environ. Res.* **161**, 364–369 (2018).
2. Poloniecki, J. D., Atkinson, R. W., Deleon, A. P. & Anderson, H. R. Daily time series for cardiovascular hospital admissions and previous day's air pollution in London. *UK. Occup. Environ. Med.* **54**, 535–540 (1997).
3. Qiu, H. *et al.* Differential effects of fine and coarse particles on daily emergency cardiovascular hospitalizations in Hong Kong. *Atmos. Environ.* **64**, 296–302 (2013).
4. Brauer, M. *et al.* Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. *Environ. Sci. Technol.* **46**, 652–660 (2012).
5. Akimoto, H. Global air quality and pollution. *Science* **302**, 1716–1719 (2004).
6. Cordero, J. M., Borge, R. & Narros, A. Using statistical methods to carry out in field calibrations of low cost air quality sensors. *Sens. Actuators B Chem.* **267**, 245–254 (2018).
7. Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M. & Bonavitacola, F. Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide. *Sens. Actuator B Chem.* **215**, 249–257 (2015).
8. Castell, N. *et al.* Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?. *Environ. Int.* **99**, 293–302 (2017).
9. Liu, Q., Liu, Y., Yang, Z., Zhang, T. & Zhong, Z. Daily variations of chemical properties in airborne particulate matter during a high pollution winter episode in Beijing. *Acta Sci. Circumst.* **34**, 12–18 (2014).
10. Lu, C. *et al.* Chemical composition of fog water in Nanjing area of China and its related fog microphysics. *Atmos. Res.* **97**, 47–69 (2010).
11. Huang, Z. & Zhang, R. Efficient estimation of adaptive varying-coefficient partially linear regression model. *Stat. Probab. Lett.* **79**, 943–952 (2009).
12. Tai, A. P. K., Mickley, L. J. & Jacob, D. J. Correlations between fine particulate matter ($PM_{2.5}$) and meteorological variables in the United States: Implications for the sensitivity of $PM_{2.5}$ to climate change. *Atmos. Environ.* **44**, 3976–3984 (2010).
13. Ayers, G. P. Comment on regression analysis of air quality data. *Atmos. Environ.* **35**, 2423–2425 (2001).
14. Dun, M., Xu, Z., Chen, Y. & Wu, L. Short-term air quality prediction based on fractional grey linear regression and support vector machine. *Math. Probl. Eng.* **2020**, 1–13 (2020).
15. Sun, W. *et al.* Prediction of 24-hour-average $PM_{2.5}$ concentrations using a hidden Markov model with different emission distributions in Northern California. *Sci. Total Environ.* **443**, 93–103 (2013).
16. Oettl, D., Almbauer, R. A., Sturm, P. J. & Pretterhofer, G. Dispersion modelling of air pollution caused by road traffic using a Markov chain–Monte Carlo model. *Stoch. Environ. Res. Risk A* **17**, 58–75 (2003).
17. Dong, M. *et al.* $PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining. *Expert Syst. Appl.* **36**, 9046–9055 (2009).
18. Elangasinghe, M. A., Singhal, N., Dirks, K. N., Salmond, J. A. & Samarasinghe, S. Complex time series analysis of PM_{10} and $PM_{2.5}$ for a coastal site using artificial neural network modelling and k-means clustering. *Atmos. Environ.* **94**, 106–116 (2014).
19. Suriano, D., Cassano, G. & Penza, M. Design and development of a flexible, plug-and-play, cost-effective tool for on-field evaluation of gas sensors. *J. Sensors* **2020**, 1–20 (2020).
20. Wang, Z., Feng, J., Fu, Q. & Gao, S. Quality control of online monitoring data of air pollutants using artificial neural networks. *Air Qual. Atmos. Health* **12**, 1189–1196 (2019).
21. Kyriakidis, I., Karatzas, K., Kukkonen, J., Papadourakis, G. & Ware, A. Evaluation and analysis of artificial neural networks and decision trees in forecasting of common air quality index in Thessaloniki, Greece. *Eng. Intell. Syst.* **2**, 111–124 (2013).
22. Liu, B., Zhao, Q., Jin, Y., Shen, J. & Li, C. Application of combined model of stepwise regression analysis and artificial neural network in data calibration of miniature air quality detector. *Sci. Rep. UK* **11**, 1–12 (2021).

23. Arsic, M., Mihajlovic, I., Nikolic, D., Zivkovic, Z. & Panic, M. Prediction of ozone concentration in ambient air using multilinear regression and the artificial neural networks methods. *Ozone Sci. Eng.* **42**, 79–88 (2019).
24. Zimmerman, N. *et al.* A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmos. Meas. Tech.* **11**, 291–313 (2018).
25. Liu, B., Yu, W., Wang, Y., Lv, Q. & Li, C. Research on data correction method of micro air quality detector based on combination of partial least squares and random forest regression. *IEEE Access* **9**, 99143–99154 (2021).
26. Yu, R., Yang, Y., Yang, L., Han, G. & Oguti, M. RAQ—A random forest approach for predicting air quality in urban sensing systems. *Sensors* **16**, 86–104 (2016).
27. Suarez Sanchez, A., Garcia Nieto, P. J., Riesgo Fernandez, P., Del Coz Diaz, J. J. & Iglesias Rodriguez, F. J. Application of an SVM-based regression model to the air quality study at local scale in the Aviles urban area (Spain). *Math. Comput. Model.* **54**, 1453–1466 (2011).
28. Liu, B., Jin, Y. & Li, C. Analysis and prediction of air quality in Nanjing from autumn 2018 to summer 2019 using PCR-SVR-ARMA combined model. *Sci. Rep. UK* **11**, 1–14 (2021).
29. Ortiz-Garcia, E. G., Salcedo-Sanz, S., Perez-Bellido, A. M., Portilla-Figueras, J. A. & Prieto, L. Prediction of hourly O_3 concentrations using support vector regression algorithms. *Atmos. Environ.* **44**, 4481–4488 (2010).
30. Deo, R. C., Wen, X. & Qi, F. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. *Appl. Energy* **168**, 568–593 (2016).
31. Wang, X. & Lu, W. Seasonal variation of air pollution index: Hong Kong case study. *Chemosphere* **63**, 1261–1272 (2006).
32. Liu, B., Tan, X., Jin, Y. & Li, C. Application of RR-XGBoost combined model in data calibration of micro air quality detector. *Sci. Rep. UK* **11**, 1–14 (2021).
33. Li, T. Z., Pan, Q. & Dias, D. Active learning relevant vector machine for reliability analysis. *Appl. Math. Model.* **89**, 381–399 (2021).
34. Olson, D. A., Riedel, T. P., Offenberg, J. H., Lewandowski, M. & Kleindienst, T. E. Quantifying wintertime O_3 and NO_X formation with relevance vector machines. *Atmos. Environ.* **11**, 1–8 (2021).
35. Tipping, M. E. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1**, 211–244 (2001).
36. Liu, X., Chen, X., Li, J., Zhou, X. & Chen, Y. Facies identification based on multikernel relevance vector machine. *IEEE Trans. Geosci. Remote* **10**, 7269–7282 (2020).
37. Qin, W., Liu, F., Tong, M. & Li, Z. A distributed ensemble of relevance vector machines for large-scale data sets on spark. *Soft Comput.* **10**, 7119–7130 (2021).
38. Khojasteh, D. N., Goudarzi, G., Taghizadeh-Mehrjardi, R., Asumadu-Sakyi, A. B. & Fehresti-Sani, M. Long-term effects of outdoor air pollution on mortality and morbidity-prediction using nonlinear autoregressive and artificial neural networks models. *Atmos. Pollut. Res.* **2**, 46–56 (2020).
39. Liu, B. *et al.* A data calibration method for micro air quality detectors based on a LASSO regression and NARX neural network combined model. *Sci. Rep. UK* **11**, 1–12 (2021).
40. Karagulian, F., Barbiere, M., Kotsev, A., Spinelle, L. & Borowiak, A. Review of the performance of low-cost sensors for air quality monitoring. *Atmosphere* **9**, 506 (2019).
41. Samia, A., Kaouther, N. & Abdelwahed, T. A hybrid ARIMA and artificial neural networks model to forecast air quality in urban areas: Case of Tunisia. *Adv. Mater.* **518**, 2969–2979 (2012).

Acknowledgements

This work was supported by the key scientific research project in Nanjing Vocational University of Industry Technology (No. YK17-10-02).

Author contributions

B.L. wrote the main manuscript text, and Y.Z. was responsible for data processing and model verification.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.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.