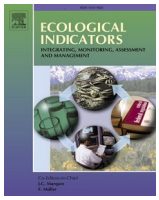
Ecological Indicators 144 (2022) 109416

Contents lists available at ScienceDirect

Ecological Indicators

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



Original Articles

A statistical learning framework for spatial-temporal feature selection and application to air quality index forecasting

Zixi Zhao [a], Jinran Wu [b], Fengjing Cai [a,*], Shaotong Zhang [c], You-Gan Wang [d]

[a] College of Mathematics and Physics, Wenzhou University, Wenzhou 325035, PR China
[b] School of Mathematical Sciences, Queensland University of Technology, Brisbane 4001, Australia
[c] Frontiers Science Center for Deep Ocean Multispheres and Earth System, Key Lab of Submarine Geosciences and Prospecting Techniques, MOE and College of Marine Geosciences, Ocean University of China, Qingdao 266100, PR China
[d] The Institute for Learning Sciences and Teacher Education, Australian Catholic University, Brisbane 4000, Australia

ARTICLE INFO

Keywords:
AQI forecasting
Statistical learning
Feature selection
Spatial auto-correlation

ABSTRACT

Accurate air quality index (AQI) forecasting makes a difference to public health, local economic development, and ecological environment. As a typical geographical datum, the spatial autocorrelation (SAC) of the AQI is often ignored, which may violate the assumptions of some models, such as machine learning which requires variables to be independent and identically distributed. Considering the strong SAC of the AQI, this study proposes a novel statistical learning framework integrating SAC variables, feature selection, and support vector regression (SVR) for AQI prediction in which correlation analysis and time series analysis are used to extract the spatial-temporal features. In addition, the historical AQI series of the target site is adjusted by using trigonometric regression to eliminate the non-stationarity. To further improve prediction accuracy, a feature selection method combining reinforcement learning with a heuristic algorithm is adopted. To demonstrate the effectiveness of our proposed framework, we select the AQI data of 34 cities from the Yangtze River Delta, which is one of the most polluted areas in eastern China, and focus on the three largest cities, Nanjing, Hangzhou, and Shanghai. We compared the proposed framework with several baselines, and the experiment illustrates that the forecasting accuracy of the proposed framework is significantly better than the baselines at all selected key sites that can provide accurate predictions for air quality.

1. Introduction

Air pollution frequently occurs in the Yangtze River Delta in China (Hao et al., 2018). Studies have shown that air pollution has a negative impact on residents' physical and mental health (Glencross et al., 2020; Jans et al., 2018), social ecological environment (Xi et al., 2020), and national economic development (Li and Peng, 2016). To scientifically issue warnings about air quality, China adopted the air quality index (AQI) as a new air quality evaluation standard in March 2012. The AQI is a dimensionless index that comprehensively reflects the concentrations of PM2.5, PM10, CO, NO_2, SO_2, and O_3. With foreknowledge of air quality standards (GB3095-2012), the government can plan emission reduction actions and transport network scheduling in advance, while residents can intuitively understand the air quality of the day and take corresponding preventive measures. Therefore, accurate prediction of the AQI is of great significance both at the social and the individual levels.

1.1. Literature review

In AQI forecasting fields, deterministic methods, traditional statistical methods, and machine learning are three common methods (Liu et al., 2020). Deterministic methods do not involve random processes and statistical theories that are based on the theory of aerodynamics, atmospheric physics, and atmospheric chemistry, which adopt mathematical methods to build models (Zannetti, 2013). The most representative models are the Community Multi-scale Air Quality model (Yang et al., 2019; Pino-Cortés et al., 2022), Weather Research and Forecasting model (Tan et al., 2017; Sati and Mohan, 2021), and Nested Air Quality Prediction Modeling System (Kong et al., 2021). Nevertheless,

* Corresponding author.
E-mail addresses: 20461026005@stu.wzu.edu.cn (Z. Zhao), j73.wu@qut.edu.au (J. Wu), caifj7704@wzu.edu.cn (F. Cai), shaotong.zhang@ouc.edu.cn (S. Zhang), you-gan.wang@acu.edu.au (Y.-G. Wang).

https://doi.org/10.1016/j.ecolind.2022.109416
Received 25 April 2022; Received in revised form 26 August 2022; Accepted 2 September 2022
Available online 18 September 2022

deterministic methods involve many fields of knowledge that require the comprehensive abilities of model users (Ma et al., 2019). Moreover, the model needs to assume that the discharge process of air pollution is stable and continuous, which is not consistent with the real situation (Stern et al., 2008). In fact, the discharge of pollution has a strong randomness. Given these limitations, some scholars have started to use statistical methods to deal with air quality prediction.

Compared with deterministic methods, statistical methods are more convenient to use because they are driven by data (Delavar et al., 2019; Callens et al., 2021). Among the traditional statistical methods, multiple linear regression (MLR) and time series analysis are widely used in the field of air quality prediction. For example, Stadlober et al. (2008) used MLR to predict the next day PM10 concentration of the western Alpe-Adria-Region, which supported the government in making decisions on traffic control. Gocheva-Ilieva et al. (2014) used autoregressive integrated moving average model for short-term prediction of air pollutant concentration in Bulgaria, which has low complexity but performs well. Koo et al. (2020) determined that the fuzzy time series models performed well in the accuracy of forecasting results and computing time when predicting the air pollution index of Kuala Lumpur. However, it is worth noting that most statistical methods assume that the relationship between independent and dependent variables is linear. This is inconsistent with the non-linearity of air quality data. Therefore, the predictive performance of these methods is limited.

Because of the high nonlinearity of air quality data, it is hard to build a predicting model (Brunelli et al., 2008). Machine learning has become popular in recent years because of its powerful ability to deal with complex nonlinear problems. Studies show that support vector machine (SVM) (Ma and Cheng, 2017; Li et al., 2017), long short-term memory (LSTM) (Wang and Song, 2018; Qunli and Lin, 2019), and neural network (Maleki et al., 2019; Qiao et al., 2020) perform well in predicting air quality. Alimissis et al. (2018) found that compared with MLR, artificial neural network has higher prediction accuracy under the condition of limited air quality network density. Li et al. (2020) proposed a new dynamic ensemble forecasting system based on machine learning to forecast the AQI, which generates accurate air quality forecasting. As for the choice of machine learning algorithms, many scholars prefer SVM to predict air quality because of its flexibility and scalability. Ketu and Mishra (2021) proposed an SVM classification algorithm based on extensible kernels to analyze air pollution events in India. Liu et al. (2019) used SVM for PM2.5 class prediction, proving that SVM obtained more convincing results than other deep learning methods. In particular, Drucker et al. (1996) proposed the support vector regression (SVR) model, which is a variant of SVM that is suitable for dealing with regression problems more effectively. Robert Kurniawan et al. (2022) set up an SVR with Harris Hawks optimization model to monitor changes of the ozone concentrations, and their experiment shows that SVR attains high accuracy and stable performance. Considering the excellent performance of SVR in dealing with nonlinear regression problems, this paper uses it as a forecasting model.

Air quality forecasting is a typical spatio-temporal prediction problem (Ge et al., 2021). According to Tobler's First Law of Geography, near things are more related to each other (Tobler, 1970). However, most of the mentioned studies on air quality prediction choose historical pollutant concentration and meteorological data as independent variables; they do not take the spatial data of the neighboring sites into consideration. Studies have proven that considering the spatial characteristics of the AQI may optimize the performance of the model. For example, Liu and Yang (2021) used spatial correlation analysis to select sites that are most correlated with the target site and then set up a spatial multi-resolution ensemble AQI predicting model. Phruksahiran (2021) proposed a geographically weighted forecasting method that combines machine learning algorithms to enhance the accuracy of AQI prediction. Given the importance of spatio-temporal characteristics in predicting the AQI, it is necessary to introduce them into the predicting model.

1.2. The motivation

According to the above literature review, it is meaningful to consider spatio-temporal characteristics when predicting the AQI. However, there are still some inadequacies in the existing research: (1) In the spatio-temporal analysis, researchers only consider the continuous temporal subsets of historical data but ignore the influence of cycle characteristics and individual lags. (2) Many prediction models incorporate a feature-reducing method, but they just use filters or wrappers to select features; a combination of these two elements is not considered to improve efficiency. (3) In order to achieve high precision, many existing studies use deep learning. The models constructed are complex in structure and place high requirements on hardware.

1.3. The contribution

To fill the gap, this study constructs a novel spatio-temporal model for daily AQI forecasting. The main contributions of this study are as follows:

- In this paper, we established a forecasting model for the AQI considering spatio-temporal effects, using a spatial correlation function and time series analysis to construct spatial auto-correlation variables of target sites. The spatial auto-correlation variables can effectively model the influence of spatial patterns. For the selection of an optimal spatial correlation function, this paper puts forward a model evaluation criterion to judge. The proposed model also considered the spatial correlation between sites, the lag effect of individuals, the relationship between air pollutants, and the cycle characteristics.
- A feature selection method combining reinforcement learning with a heuristic algorithm is selected to eliminate unimportant variables in the feature set, which can avoid the over-fitting phenomenon of prediction models that results from too many variables. In this paper, RR, MLR and SVR are selected as basic algorithms to predict the data respectively. According to the selected model performance evaluation indices, SVR is determined as the optimal forecasting algorithm for its lowest prediction error.
- Three target sites were selected in the Yangtze River Delta, and the actual data of 34 air quality monitoring sites were used to verify the correctness of the proposed model. In addition, it was compared with several baselines. In this study, three experimental steps were carried out to verify that each component of the proposed hybrid model can effectively enhance the accuracy of AQI prediction.

1.4. The structure of the paper

The rest of this paper is organized as follows: Section 2 introduces the basic theories of SVR and QBSO. The spatial autocorrelation and the new combined AQI forecasting model are presented in Section 3. Section 4 shows the study area, model evaluation criterion, and forecasting results analysis. Finally, the conclusion is discussed in Section 5. Moreover, the nomenclature of the paper is given in Table 1.

2. The preliminaries

2.1. Support vector regression

Support vector regression is an extension of SVM, which is more suitable for regression analysis (Drucker et al., 1996). The basic idea of SVR is to minimize the error by adjusting the hyper-plane to maximize the interval between two decision boundaries (Parbat and Chakraborty, 2020). The generalized equation for the hyper-plane is represented as follows:

$$f(x, \omega) = \omega x + b, \tag{1}$$

Table 1
The nomenclature.

Symbol	Definition	Symbol	Definition
ω	Normal vector of the hyper-plane in SVR.	$\rho_{i,j}$	Pearson correlation coefficient between two sites.
b	Intercept at $x = 0$ in SVR.	$D_{i,j}$	Euclidean distance between two sites.
C	Penalty parameter in SVR.	$y_i(t)$	Historical AQI series of the i-th site.
ϵ	Tolerance deviation in SVR.	$\overline{y_i(t)}$	Mean value of historical AQI series of the i-th site.
N	Number of observations in training set.	θ	Spatial correlation weights.
M	Number of samples in the testing set.	U_j	Residual sequence of time series of the j-th site.
n	Number of selected sites in this study.	R_t	Stationary series of AQI.
α_i	Lagrange multiplier to get the solution of SVR.	$\widehat{R_t}$	Prediction using the lag features of sites.
α_i^*	Lagrange multiplier to get the solution of SVR.	S_t	Seasonality of AQI.
Acc	Classification accuracy returned by KNN in QBSO.	T_y	Yearly tendency.
C_t	Condition set of bees in QBSO.	T_s	Seasonal tendency.
A_t	Action set of bees in QBSO.	T_m	Monthly tendency.
M_t	Reward set of bees in QBSO.	T_w	Weekly tendency.
λ	Learning rate in QBSO.	p	Lag order of the stationary series R_t.
γ	Discount parameter in QBSO.	$\hat{y}$	Prediction series of AQI.

where ω is the normal vector of the hyper-plane and b is the intercept at $x = 0$. ϵ represents the tolerance deviation, which is a manually set empirical value. The loss of all the samples falling into the interval band is not calculated, that is, only the support vector will affect its function model. The support vectors are the points closest to the hyper-plane. The coefficients can be estimated by minimizing the risk function $R(C)$, which is expressed as follows:

$$R(C) = C\sum_{i=1}^{N} L(y_i, f(x_i, \omega)) + \frac{1}{2}\|\omega\|^2. \tag{2}$$

The insensitive Laplacian loss function is defined as follows:

$$L(r) = \begin{cases} |r| - \epsilon, & |r| \leqslant \epsilon, \\ 0, & \text{else}, \end{cases} \tag{3}$$

with residuals $r = y - f$.

To obtain the solution of the original problem, the Lagrange multiplier algorithm can be used. The equation transformed with the multiplier is as follows:

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*)(x_i \cdot x) + b, \tag{4}$$

where N is the number of observations in the training set, and α_i, α_i^* are Lagrange multipliers. The input vector x_i is called the support vector, which provides its corresponding coefficients $(\alpha_i - \alpha_i^*) \neq 0$. Support vectors represent the entire support vector function because they cover most of the information for the training set.

Nonlinear problems are often difficult to solve. It is necessary to define a suitable kernel function to transform nonlinear problems into linear problems (Brereton and Lloyd, 2010). The choice of kernel function is the key to the success of SVR. The commonly used kernel functions are polynomial kernel, linear kernel, Gaussian kernel, and sigmoid kernel (Patle and Chouhan, 2013). Among all the available kernel functions, the Gaussian kernel function (Radial Basis Function, RBF) is the most widely used kernel function, which is well suited to dealing with nonlinear problems (Gopi et al., 2020).

2.2. Feature selection: Q-learning based bee swarm optimization

Before performing machine learning tasks, feature selection is usually used to eliminate irrelevant variables. Sadeg et al. (2019) proposed a hybrid meta-heuristic algorithm, which combines a reinforcement learning algorithm (Q-Learning) with a bee swarm optimization meta-heuristic algorithm (BSO) for feature selection, making the algorithm more efficient and adaptive. Q-learning (Kumar et al., 2020) is an algorithm that makes decisions through rewards and punishments and does not need to specify the process of task completion. The target bee only needs to perform actions according to its own state, and then obtain feedback rewards or punishments from the environment. In fact, in the BSO algorithm (Djenouri et al., 2018; Djenouri et al., 2019), when a bee is distributed a solution, it will execute a classical local search first, then evaluate all solutions around it and finally return the best one. If the local search executed by the bees is replaced by Q-Learning, each bee is regarded as an agent gathering useful information during the search and benefiting from the experience of other bees. In this way, bees only need to search the path with high reward or low punishment so that the efficiency of the original algorithm greatly improves. This algorithm uses K-Nearest Neighbor (KNN) as the basic classifier. The classification accuracy returned by KNN is calculated as follows:

$$Acc = \frac{\text{True positive} + \text{True negative}}{\text{Total number of sample}}, \tag{5}$$

with the value of reward r_t:

$$\begin{cases} r_t \leftarrow Acc(c_{t+1}) & \text{if } Acc(c_t) < Acc(c_{t+1}), \\ r_t \leftarrow Acc(c_{t+1}) - Acc(c_t) & \text{if } Acc(c_t) > Acc(c_{t+1}), \\ r_t \leftarrow \frac{1}{2} * Acc(c_{t+1}) & \text{if } num(c_t) > num(c_{t+1}), \\ r_t \leftarrow -\frac{1}{2} * Acc(c_{t+1}) & \text{if } num(c_t) < num(c_{t+1}), \end{cases} \tag{6}$$

where $A_t = \{a_{t_1}, a_{t_2}, \ldots, a_{t_n}\}$ is the set of possible actions of the current condition c_t; c_{t+1} is the next condition after selecting an action from A_t; $Acc(c_t)$ is the classification accuracy using the feature subset obtained from c_t; and $num(c_t)$ is the size of the feature subset obtained from c_t.

3. The proposed method

3.1. Spatial auto-correlation

Spatial auto-correlation (SAC) refers to the degree to which an object is similar in time and space to other nearby objects (Legendre, 1993). "Spatial" represents the spatial effect between sites, and "autocorrelation" describes the impact of individual lag. In practical scenarios, the most common type is positive spatial auto-correlation, where a property in a neighboring region has a similar changing tendency (Griffith, 2011). SAC is often measured to avoid violating certain fundamental statistical assumptions of certain statistical methods (Lichstein et al., 2002). For example, machine learning requiring variables to be independent and identically distributed. Violating assumptions will affect the performance of the model.

The degree to which sampling sites are correlated is influenced by the two-dimensional Euclidean distance between the two sites (Behrens et al., 2018). Thus, SAC can be thought of as a two-dimensional generalization of temporal auto-correlation in which the correlation (ρ) between two sites is inversely proportional to the Euclidean distance between the sites. To construct the SAC variables of the AQI, the Euclidean distance and Pearson correlation coefficient (Benesty et al., 2009) between each two sites need to be calculated first. The calculation

formulas are as follows:

$$D_{i,j} = \sqrt{\left(Lat_i - Lat_j\right)^2 + \left(Lon_i - Lon_j\right)^2}, \tag{7}$$

and

$$\rho_{i,j} = \frac{\sum_{t=1}^{N}\left(y_i(t) - \overline{y}_i(t)\right)\left(y_j(t) - \overline{y}_j(t)\right)}{\sqrt{\sum_{t=1}^{N}\left(y_i(t) - \overline{y}_i(t)\right)^2} \times \sqrt{\sum_{t=1}^{N}\left(y_j(t) - \overline{y}_j(t)\right)^2}}, \tag{8}$$

where $D_{i,j}$ and $\rho_{i,j}$ denote the distance and correlation coefficient between sites, respectively; Lat_i, Lon_i denote the latitude and longitude of the i-th site; N is the number of observations in the training set, $y_i(t)$ is the historical AQI series in the i-th site, and $\overline{y_i(t)}$ is the mean value of the historical AQI series in the i-th site.

If we want to quantify the spatio-temporal impact between sites, we can construct SAC variables by modelling in some form of spatial dependence correlation structure. The reasons for SAC are varied, and so are the manifestations in which sampling sites can be spatially correlated. The SAC variable of the i-th site can be calculated as follows:

$$X_{SAC_i} = \sum_{j=1}^{n} \theta_{i,j} U_j, \tag{9}$$

where θ is the weights calculated by the spatial correlation function; n is the number of selected sites; and U_j is the residual sequence of the time series of the j-th site. There are five commonly used spatial correlation functions (Cressie, 2015):

(1) Exponential: $\theta = e^{-\rho D}$;
(2) Gaussian: $\theta = e^{-(\rho D)^2}$;
(3) Linear: $\theta = 1 - \left(1 - \frac{\rho}{D}\right)\mathbb{I}(\rho < D)$;
(4) Quadratic: $\theta = \frac{1}{1+(\rho D)^2}$; and
(5) Spherical: $\theta = 1 - \left(1 - 1.5\frac{\rho}{D} + 0.5\left(\frac{\rho}{D}\right)^3\right)\mathbb{I}(\rho < D)$.

Fig. 1 shows the change of spatial correlation with distance when the value of ρ is set to 0.5, 0.8, and 1. It can be seen that the variation trend of each spatial correlation function is different, so the selection of the appropriate spatial correlation function is also the key to improve the accuracy of the prediction.

3.2. Proposed AQI forecasting model

The general framework of the proposed AQI forecasting model is given in Fig. 2, and this section gives a detailed introduction of the modeling process.

3.2.1. Seasonal adjustment

To eliminate the influence of seasonality on historical data, the AQI series of each site needs seasonal adjustment. In this study, we used a trigonometric function to extract the complex seasonality in the time series. To extract the seasonality comprehensively, we considered the yearly, seasonal, monthly, and weekly cycles. The function of the seasonal adjustment is as follows:

$$\begin{aligned} S_t = \; & b_0 + e_1 \sin\frac{2\pi t}{T_y} + e_2 \cos\frac{2\pi t}{T_y} + e_3 \sin\frac{2\pi t}{T_s} + e_4 \cos\frac{2\pi t}{T_s} + \\ & e_5 \sin\frac{2\pi t}{T_m} + e_6 \cos\frac{2\pi t}{T_m} + e_7 \sin\frac{2\pi t}{T_w} + e_8 \cos\frac{2\pi t}{T_w}, \end{aligned} \tag{10}$$

where b_0 and e are the intercept and coefficients of the equation; T_y, T_s, T_m, T_w are the yearly, seasonal, monthly, and weekly cycles of the time series, respectively; and t is the time to get the observation. Then, the stationary series can be obtained as follows:

$$R_t = Y_t - S_t, \tag{11}$$

where Y_t is the original time series. The remaining information is stored in the new series R_t, which is the stationary series we need.

3.2.2. SAC variables

To extract valid information reasonably, Pearson's test on the stationary series R_t is adopted to select monitoring sites that have a positive impact on the AQI in the target site. The correlation coefficients estimate the degree of spatial correlation of the AQI series between the target site

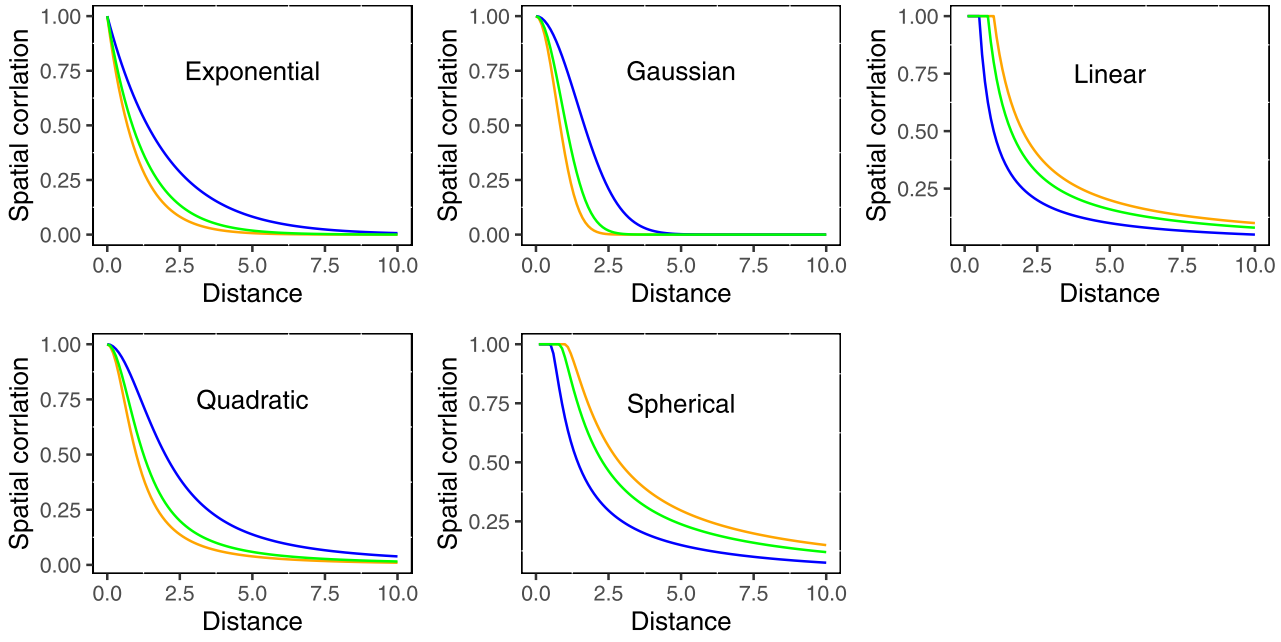


Fig. 1. The variation of spatial correlation with distance and the value of ρ. The orange line denotes $\rho = 1.0$; the green line denotes $\rho = 0.8$; the blue line denotes $\rho = 0.5$.

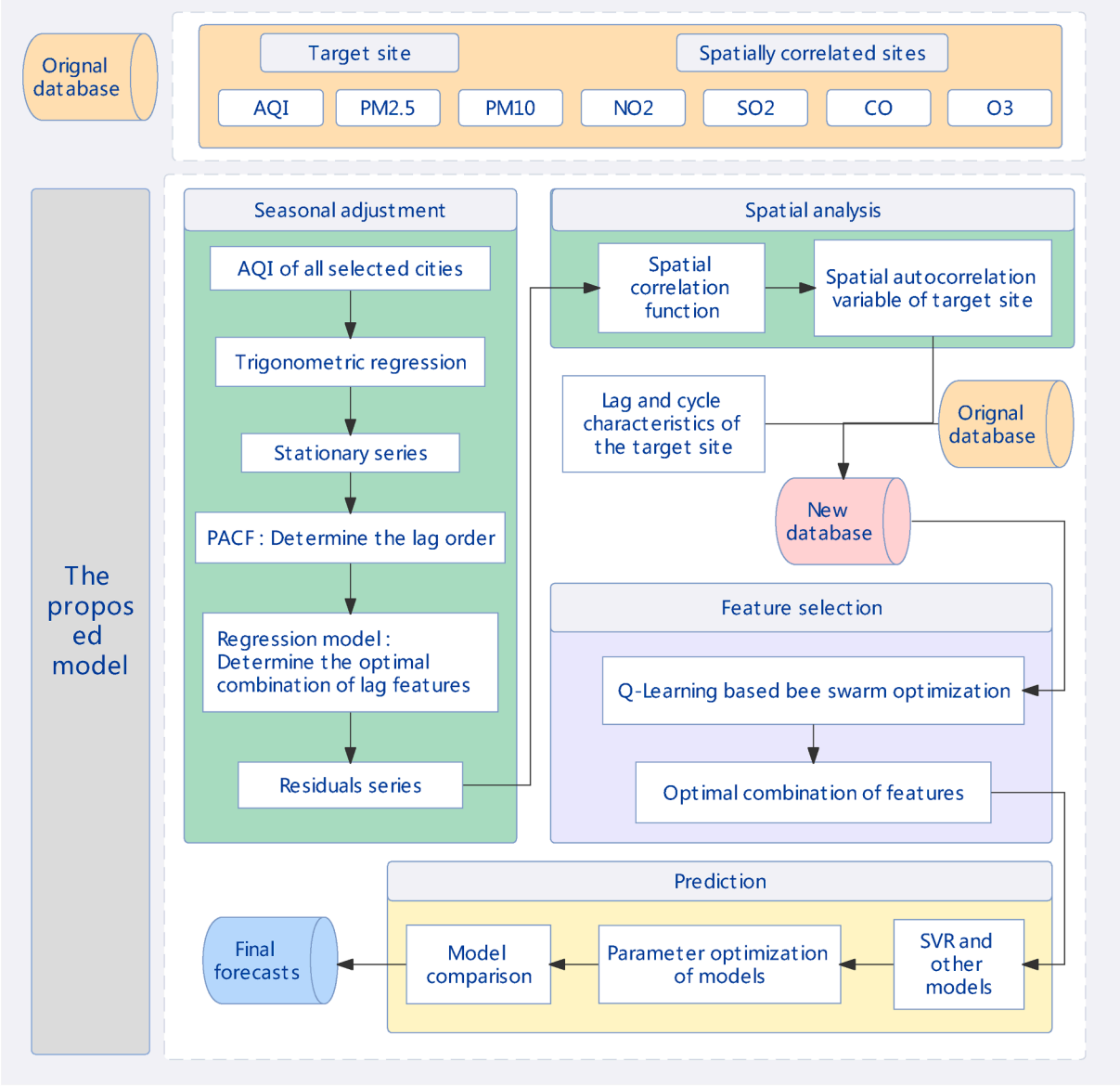


Fig. 2. Framework of the proposed AQI forecasting model.

and other sites. The closer the coefficient is to 1, the higher the correlation between the two sites. In this study, the sites with correlation coefficients above 0.7 are selected as spatially correlated sites. The selected sites are expected to express the spatial information of the AQI. Their historical data are added as a part of the training set to improve the accuracy of the target site's AQI forecasting.

As for the temporality of the AQI, previous studies have determined that the AQI of the current moment of a site has a certain correlation with the past moment (Liu and Chen, 2020). Thus, we based the PACF graph to determine the lag order p of the stationary series R_t:

$$\widehat{R}_t = f\left(R_{t-1}, R_{t-2}, \cdots, R_{t-p}\right). \tag{12}$$

The regression model f is used to produce the best combination of different time lag features; $\widehat{R}_t$ is the prediction using the lag features of the sites. The residual sequence of each selected site is calculated via the following:

$$U = R_t - \widehat{R}_t. \tag{13}$$

With the residuals, the SAC variables of target sites can be calculated as Eq. 9. After these preparations are completed, the final feature set consists of two parts: the first part is the target site's own features, including the AQI, six pollutant concentrations, the SAC variable, lag, and cycle characteristics; the second part is the six pollutant concentrations of the spatially correlated sites.

3.2.3. *Feature selection*

A hybrid meta-heuristic feature selection is adopted to search for an optimal combination of these features that combines Q-Learning and BSO. The algorithm can effectively reduce the dimension of data. The action in this algorithm consists of adding or removing a feature from the current feature set. Reward $m(c, a)$ is calculated by regarding the classification accuracy by KNN as the main standard and the size of the feature set as a second standard. To narrow the search space, only the actions in A_t that contain the maximum similarities between the current solution and the best global solution are adopted.

The choice of parameter values has a crucial impact on the dimensionality reduction. Here are some key parameters: γ is a discount parameter. If γ is close to 0, the bee tends to choose the current rewards. If it is close to 1, the bee tends to consider the future reward. *flip* is an empirical parameter. The value of *flip* affects the bees' searching

efficiency because it measures the distance between the current solution and the solution determining the search area. In fact, the smaller the value of *flip* is, the better it is to search for solutions, while the larger the value is, the bees are constantly expanding the search area which may cause the algorithm to converge to a local optimum. The process of QBSO can be seen in Fig. 3.

3.2.4. The forecasting model

MLR, RR, and SVR are all good choices when dealing with regression problems. In this paper, MLR and RR are used as benchmark models to verify the prediction performance of SVR. These models can be used independently or as a substitute for a component of the framework.

Algorithm: Q-Learning based Bee Swarm Optimization

Input: The original feature set

Output: The best solution

Require: Conditions $C = \{c_1, \ldots, c_n\}$, Actions $A = \{a_1, \ldots, a_n\}$, Reward $m(c, a)$

for $c \in C$ do

  Initialize the reward $m(c, a)$ to 0

end

while Q *is not converged* **do**

  Start in condition $c \in C$

  while c *is not terminal* **do**

    Use the next state c'

    then update the reward $m(c, a)$ as follows:

    $m(c,a) \leftarrow (1-\alpha) \cdot m(c,a) + \alpha(m(c,a) + \gamma \cdot \max_{a'} m(c', a'))$

    $c \leftarrow c'$

  end

  return $Q = m(c, a)$

end

Assign a solution from condition C to each bee, the solution is determined by Q.

for *each bee* **do**

  Perform a local search

  if *The corresponding feature is selected* **then**

    The position vector is set to 1

  else

    0

  end

end

return The position vector (The best solution).

Fig. 3. The pseudo code of QBSO.

1. Multiple linear regression (MLR) (Valentini et al., 2021): MLR is an extension of linear regression, which is commonly used to deal with multi-factor time series predicting problems.
2. Ridge regression (RR) (McDonald, 2009): RR is an improved least squares estimation method that introduces the bias of the least squares method and part of the information to make the estimation of the regression coefficients more consistent with the actual case and more reliable.

Some details of MLR and RR can be found in A and B, respectively.

The features obtained are put into three forecasting models, respectively. Considering the nonlinear relationship between air data, RBF is chosen as the kernel function of SVR. For parameters involved in RR and SVR, cross-validation is used to make decisions. The three target sites are processed in the same way to obtain forecasting results. The complete process of prediction is shown in Fig. 4.

4. Case study

4.1. Data collection

The Yangtze River Delta (YRD) is one of the most developed and fastest growing economic development regions in China. By the end of 2020, the YRD had a population of 235 million, covering an area of about 358,000 km^2, and the resident population was as high as 60%. However, this rapid economic development has caused serious air pollution. In recent years, urban air quality has deteriorated, and visibility is decreasing because of petrochemical combustion and pollution emissions (Ma et al., 2019). Now, the YRD has become one of the most polluted areas in eastern China. In this study, Shanghai, Hangzhou, and Nanjing are selected as the target sites to verify the effectiveness of the proposed model. Fig. 5 shows a brief introduction and the locations of these three sites.

This study selects daily AQI data from 34 air quality monitoring sites in the YRD collected from January 1st, 2019 to December 31st, 2020. Moreover, the calculation formula for the AQI is given as follows:

$$I = \frac{I_{\text{high}} - I_{\text{low}}}{C_{\text{high}} - C_{\text{low}}}\left(C_p - C_{\text{low}}\right) + I_{\text{low}}, \tag{14}$$

where I is the AQI output value, C_p is the pollutant concentration as input value, C_{low} is the concentration limit less than or equal to C_p, C_{high} is the concentration limit greater than or equal to C_p, I_{low} corresponds to the exponential limit of C_{low}, and I_{high} corresponds to the exponential limit of C_{high}. $C_{\text{low}}, C_{\text{high}}, I_{\text{high}}$ and I_{low} are all known constants. Based on Eq. 14, the value of the AQI is only related to the concentration of pollutants, which is not affected by meteorological factors. Therefore, in this study, only six major pollutants are added in the training set as auxiliary data, including the concentrations of PM2.5, PM10, CO, NO_2, SO_2, and O_3. The data were obtained from the air quality online monitoring and analysis platform (https://www.aqistudy.cn/).

Table 2 lists the statistical indicators of the experimental data and Fig. 6 shows the daily AQI at three target sites from 2019–01-01 to 2019–03-01, both of which show that the three target sites have different statistical characteristics.

4.2. Evaluation criterion

In this study, the model with the best prediction effect will be selected according to three commonly used evaluation indices in prediction problems, including RMSE (Root Mean Square Error), NSE (Nash–Sutcliffe efficiency coefficient), and MAPE (Mean Absolute Percentage Error). Similarly, the selection of the best SAC variable will be judged according to these three indices. The definitions of each index are as follows:(1) **RMSE**: It measures the deviation between the observed value and the true value, which is often used to measure the outcomes predicted by machine learning models. It is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2}. \tag{15}$$

(2) **NSE**: If its value is close to 1, the model quality is good and the model credibility is high. The closer its value is to 0, the more reliable the overall result is, but the simulation error is large. If the value is much less than 0, the model is not credible. It is defined as follows:

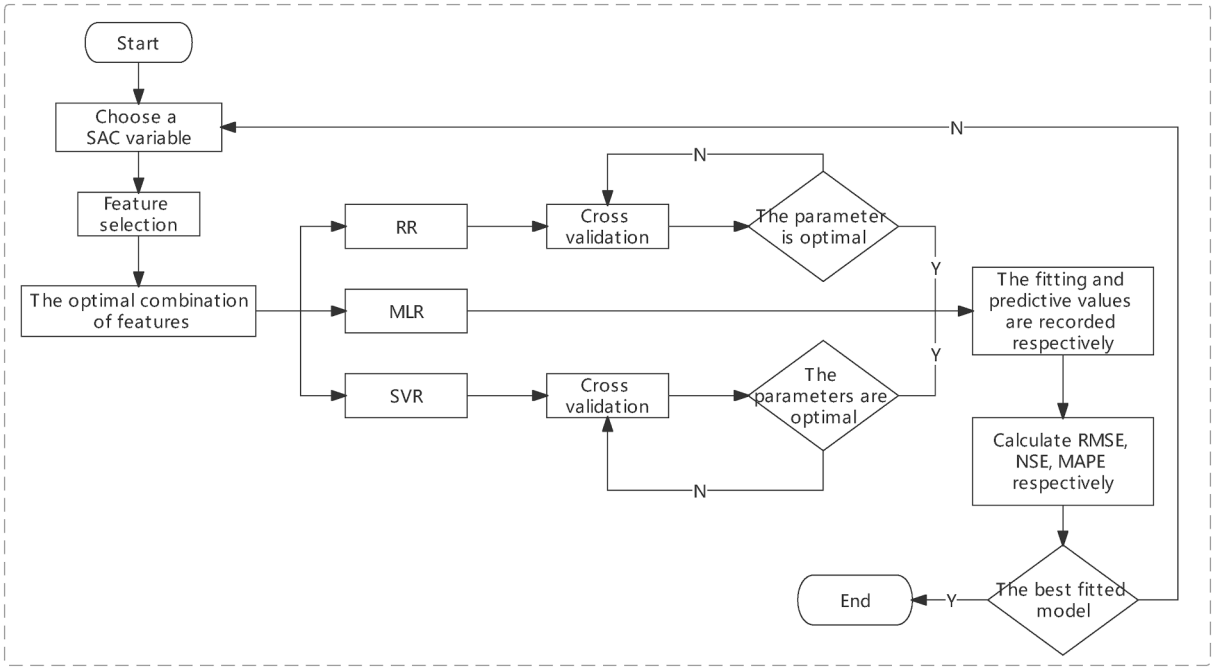


Fig. 4. The complete process of prediction.

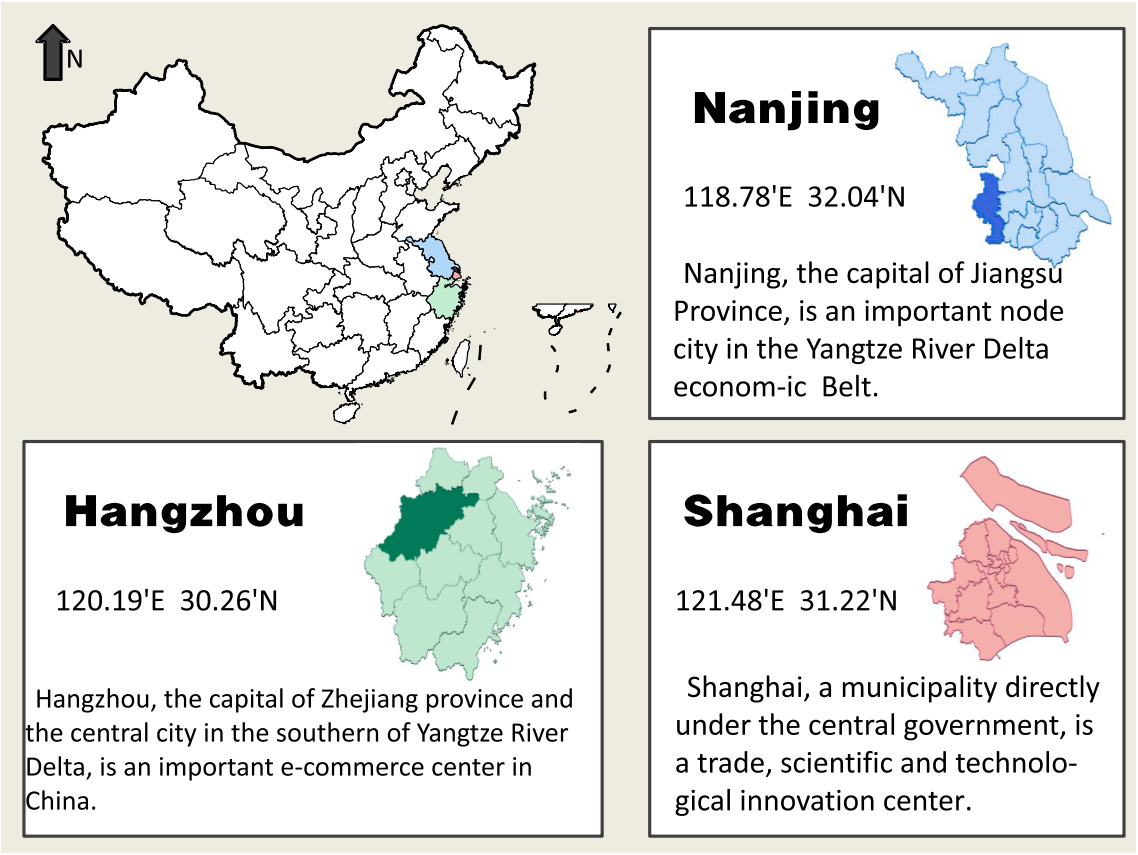


Fig. 5. Brief introduction to three cities in Yangtze River Delta.

Table 2
Statistical indicators of air pollution data in three target sites.

		Type of data						
Target site	Statistical indicator	AQI	PM2.5	PM10	CO	NO_2	SO_2	O_3
Nanjing	Minimum	17	5	11	0.3	10	3	8
	Maximum	222	145	365	1.6	103	24	244
	Mean	79.11	35.77	64.09	0.79	38.77	8.46	104.4
	Standard deviation	31.99	22.66	35.69	0.21	16.04	3.43	48.83
Shanghai	Minimum	20	6	7	0.3	6	4	14
	Maximum	206	156	212	0.6	115	16	274
	Mean	69.91	32.95	44.66	0.65	39.07	6.63	95.63
	Standard deviation	29.06	21.13	24.85	0.2	16.58	1.97	40.47
Hangzhou	Minimum	22	6	8	0.3	7	3	4
	Maximum	179	135	209	1.7	89	17	236
	Mean	73.87	33.97	60.79	0.746	39.72	6.29	94.86
	Standard deviation	29.16	19.82	32.19	0.19	15.32	1.83	50.73

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{M} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{M} (y_i - \overline{y_i})^2}. \tag{16}$$

(3) **MAPE**: The smaller the MAPE, the better the model. It is defined as follows:

$$\mathrm{MAPE} = \frac{100\%}{M} \sum_{i=1}^{M} \left| \frac{y_i - \widehat{y_i}}{y_i} \right|. \tag{17}$$

In these three formulas, M represents the number of samples in the testing set; y_i is the actual AQI series, representing the observed value of the i-th sample, and $\overline{y_i}$ denotes the total average of the actual values; $\widehat{y_i}$ is the forecasting result, representing the predicted value of the i-th sample.

4.3. The experimental results

The experiment in this study includes three main objectives: 1) to verify the influence of the SAC variable on the model; 2) to prove the effect of feature selection by QBSO; and 3) to confirm the influence of the SAC variable and QBSO synchronously. To validate these goals, some models are adopted in this paper, which are listed in Table 3. We classify these models into two categories: benchmark models and

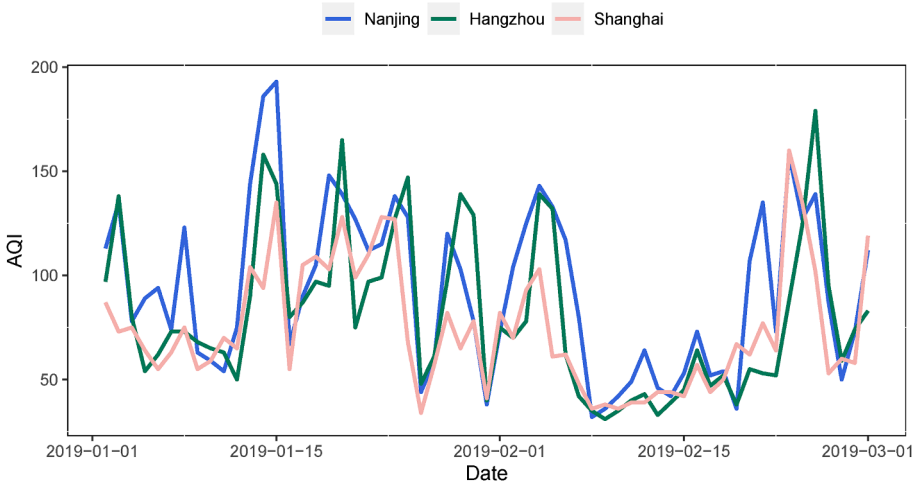


Fig. 6. The trends of air quality at three target sites from 2019-01-01 to 2019-03-01.

Table 3
All models adopted in the experiment.

Model	Abbreviation	Definition
Benchmark Models	MLR	Multiple Linear Regression
	RR	Ridge Regression
	SVR	Support Vector Regression
	SAC-MLR	MLR with one Spatial Auto-correlation variable
	SAC-RR	RR with one Spatial Auto-correlation variable
	SAC-SVR	SVR with one Spatial Auto-correlation variable
	QBSO-MLR	MLR with Q-Learning Based Bee Swarm Optimization
	QBSO-RR	RR with Q-Learning Based Bee Swarm Optimization
	QBSO-SVR	SVR with Q-Learning Based Bee Swarm Optimization
	SAC-QBSO-MLR	MLR with QBSO adding one SAC variable
	SAC-QBSO-RR	RR with QBSO adding one SAC variable
Proposed Model	SAC-QBSO-SVR	SVR with QBSO adding one SAC variable

proposed model. The data of the three target sites contain 731 daily observations. We use the cross-validation method to prevent over-fitting. The 1st - 585th data are the training set used to train forecasting models. The 586th - 731st data are the testing set used to evaluate the final forecasting performance.

4.3.1. *The result of the seasonal adjustment*

For air quality data, the common tendency types are yearly, seasonal, monthly, and weekly trends. Fig. 7 shows these four trends respectively by using the first 30 days of historical AQI data from the three target sites. The fifth picture for each city is the stationary series after seasonal adjustment.

4.3.2. *The construction of the SAC variable*

The proposed model uses the Pearson correlation coefficient to measure the spatial correlation of the AQI between the target site and other sites. Sites with a correlation coefficient greater than 0.7 are selected as spatially correlated sites. Fig. 8 shows the correlation coefficients among some cities in the YRD. Their geographical distribution is shown in Fig. 9. Table 4 lists the spatially correlated sites of each target site.

The lag order of each station was determined according to the PACF diagram, and the optimal combination of lag features and residual sequence were obtained by linear regression. Table 5 shows the lag features affecting the targeted sites and their regression coefficients.

The spatial correlation of each site is different. To find the optimal spatial correlation function of the target site, we calculated five SAC variables and introduced them into the model respectively. Taking the model without any SAC variables as a comparison, for the convenience of visualization, Fig. 10 only shows the model results of the exponential and spherical SAC variables. The complete results of the three target sites are shown in Table 6. As shown in Fig. 10, after adding an SAC variable, the models provide lower RMSE and higher NSE. Especially, the values of NSE are all close to 1, indicating that the models with an SAC variable have better predictive ability. For Nanjing, the introduction of the linear SAC variables can greatly reduce the RMSE from 237.296 to 7.981 of MLR. For Hangzhou, when adding the Gaussian SAC variable, the MAPE of SVR is reduced from 11.974 to 2.895. These results indicate that the SAC variables contain more related information on the AQI. For Hangzhou and Shanghai, when the forecasting models are MLR and RR, the MAPE value is not ideal compared with the model without any SAC variables. This is because MLR and RR do not handle extreme values well. Table 2 describes that there is a big difference between the maximum and minimum of air quality data in each city. Moreover, air quality data fluctuate greatly over time, which also leads to MLR and RR failing to provide robust predictions. In Fig. 10, we can also see that different spatial functions have different effects on the model. Nanjing and Shanghai are more suitable for the spherical spatial variable, and Hangzhou is suitable for the Gaussian spatial variable. In this study, for the selection of the best spatial correlation function of each site, the method we adopted is to observe the performance of the model when adding a corresponding SAC variable, and determine the best one with RMSE, NSE, and MAPE.

4.3.3. *Feature selection by QBSO*

Feature selection can effectively improve the efficiency of the model. Thus, QBSO is adopted in our experiment to eliminate irrelevant variables and avoid over-fitting. In our experiment, we manually tune the parameters to find the optimal parameter values for classification accuracy and running time, setting the learning rate λ and the discount parameter γ at 0.9 and 0.1, respectively. The value of *flip* is 5. Table 7 lists the number of the original features and selected features, respectively, while also containing the accuracy of the optimal solution and average time to evaluate a solution. As shown in Table 7, QBSO can quickly judge the correctness of a solution. For the three data sets selected in this study, QBSO achieved good classification results. To verify the effectiveness of QBSO, we compared it with unfiltered data

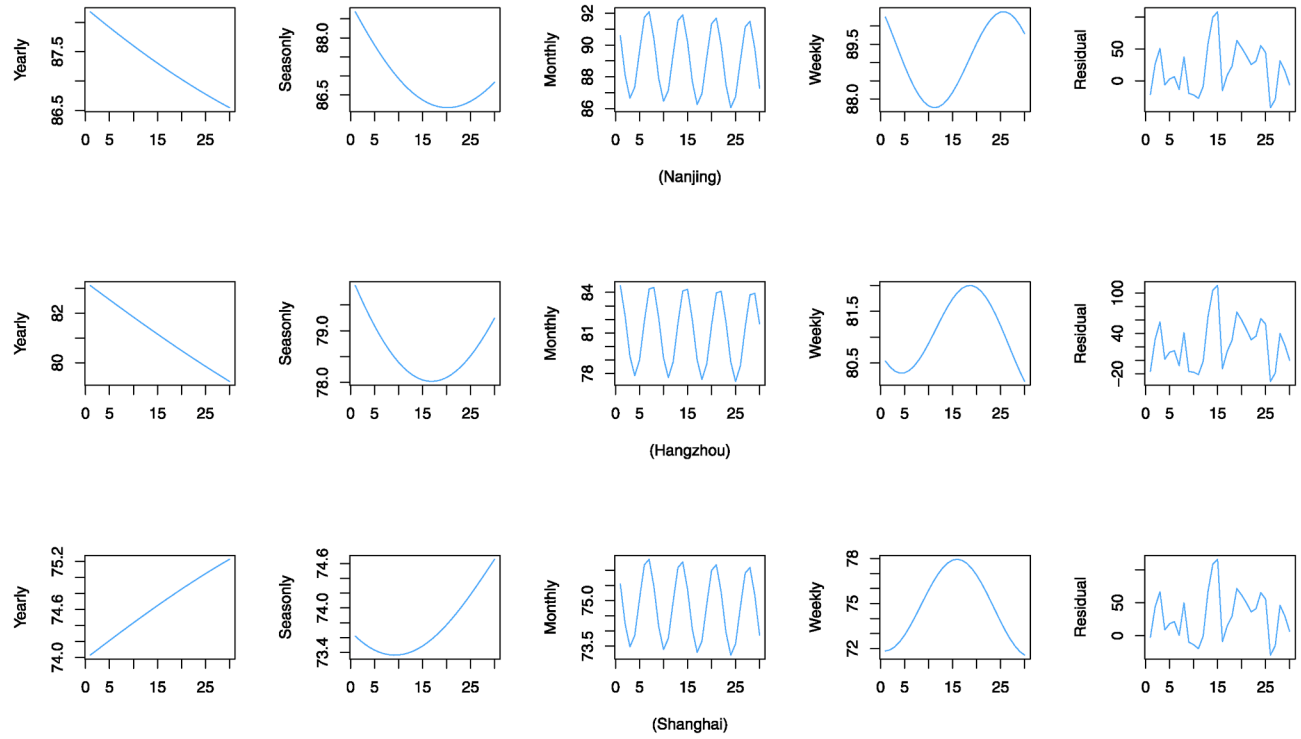


Fig. 7. Different tendencies and residual series of historical AQI of the three target sites.

sets. Fig. 11 shows the influence of QBSO on the model directly. Table 8 gives the specific values of the three evaluation indexes of each model. We can see that when QBSO is applied to the data set, SVR has the lowest MAPE and RMSE compared with other models in the three sites, and the improvement of NSE value is obvious. For Hangzhou and Shanghai, QBSO can reduce the MAPE of MLR and RR, but for RMSE, it has no significant decrease. About these two cities, QBSO is more suitable to be used in combination with SVR, because the nonlinear correlation of the data is also considered when QBSO filters variables, while MLR and RR cannot handle nonlinear relations well.

In addition, QBSO has the advantage of ranking the importance of individual features. Finding the most influential factor on the target site can help the government to make decisions about pollution control. Table 7 lists the top three most influential features of the target sites.

4.3.4. *The final forecasting model*

In this section, we proved the effectiveness of introducing SAC variables and QBSO simultaneously into the model. Fig. 12 shows each model's performance, and the complete data are in Table 9. It can be seen from Fig. 12 that the proposed SAC-QBSO-SVR framework performed best for all selected target sites. Under the condition of using the same data set, although the RMSE and NSE values of MLR and RR are also relatively satisfying, their MAPE values are very high, while the MAPE values of SVR are relatively low, which indicates that SVR is a more suitable forecasting model for this study. As for the framework constructed in this study, the selection of SAC variables was determined based on the RMSE, NSE, and MAPE of the model, and each site had its own optimal spatial calculation function. It is impossible to determine a single optimal spatial function for all sites because the spatial relevance of each site is affected by different factors. Fig. 13 shows the first 30 days' prediction from the test set of each model. The proposed model chose the optimal spatial correlation function of each target site, respectively. We find that the predicted values of the proposed model are closest to the actual values. Overall, the model established in this paper has high prediction accuracy and stable performance, without large error fluctuations, and it can accurately predict the AQI values of the sites.

4.4. *General discussion*

Summarizing the experimental results above, each component of the proposed framework, including spatio-temporal analysis, feature selection, and prediction, are essential to guarantee excellent performance. Taking Nanjing city as an example, compared with the original prediction set, the performance of MLR, RR, and SVR improved after any SAC variables was added, but the spherical SAC variable was the best one. As for the three forecasting models, SVR changed most significantly: its RMSE and MAPE decreased by 73.2% and 79.1%, respectively, and NSE increased by 59.5%. QBSO was used for dimensionality reduction, and the three evaluation indices of SVR were all improved by about 30%. The proposed SAC-QBSO-SVR model can optimized RMSE, NSE, and MAPE by 77.6%, 59.5%, and 81.4%, respectively, in our test dataset compared with other models. The proposed SAC-QBSO-SVR considers the spatial autocorrelation of different monitoring sites, which can significantly improve air quality prediction.

Besides monitoring air quality changes, the proposed model can also help develop feasible pollution management measures. The addition of SAC variable significantly improved the prediction accuracy, which confirmed that the air quality in the YRD region had spatial correlation characteristics. Therefore, cross-regional cooperation pollution control is essential to achieve the desired level of pollution management. We suggest imposing air pollutants emission limits on the major regional businesses and prohibit the construction or expansion of heavy polluting enterprises such as steel, nonferrous metals and chemical industries in urban areas and their suburbs. QBSO algorithm can rank the importance of features. As can be seen from Table 7, for the AQI of the three sites, the main influential pollutants are PM2.5, NO_2 and O_3, which are primarily caused by coal, oil combustion and automobile exhaust emission. Hence, we propose to adjust the energy structure, promoting the development of clean energy such as solar, wind and hydropower, and advocate the efficient use of coal. At the individual level, residents are urged to use green public transport as much as possible. These suggestions are not

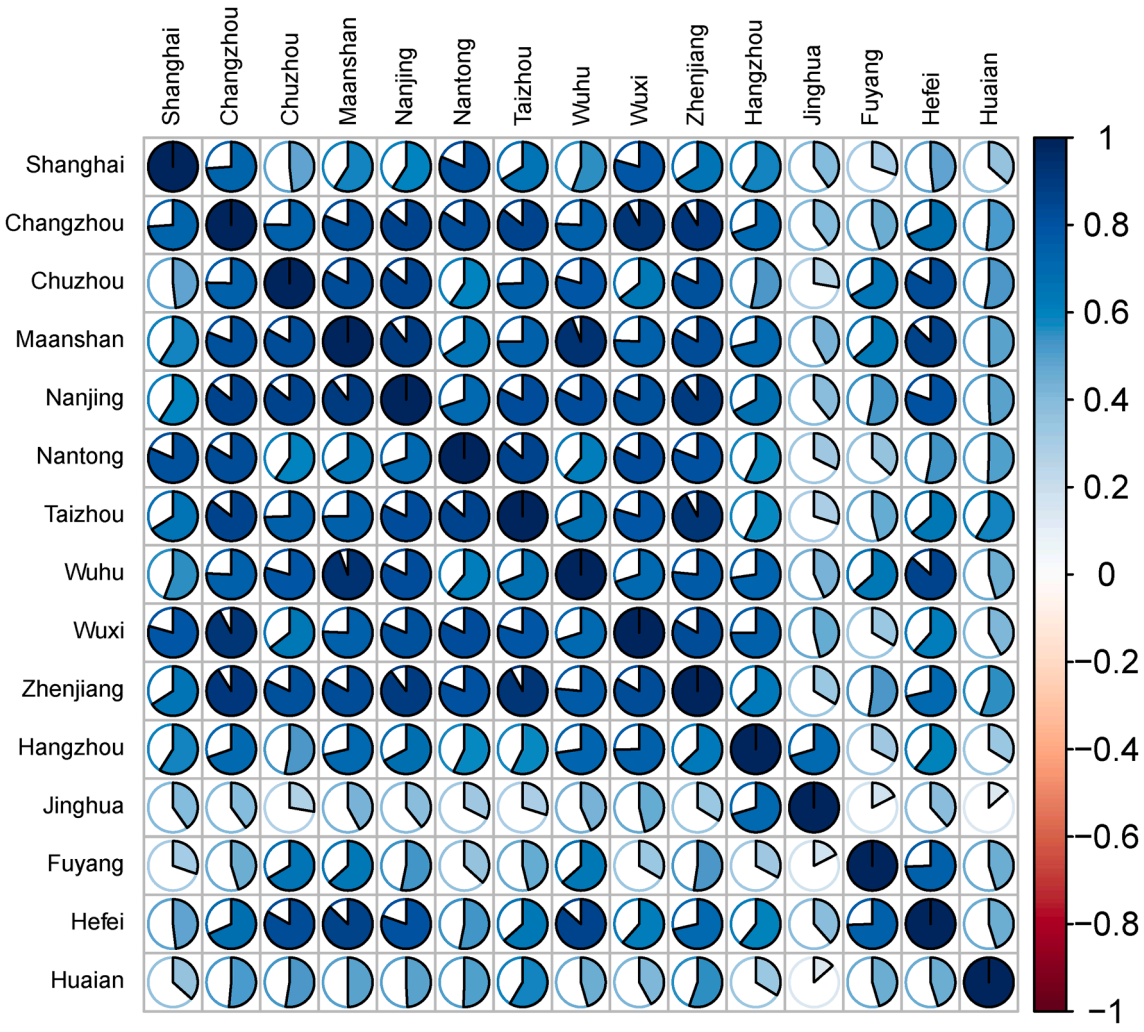


Fig. 8. Pearson correlation coefficient between part of cities in the Yangtze River Delta region.

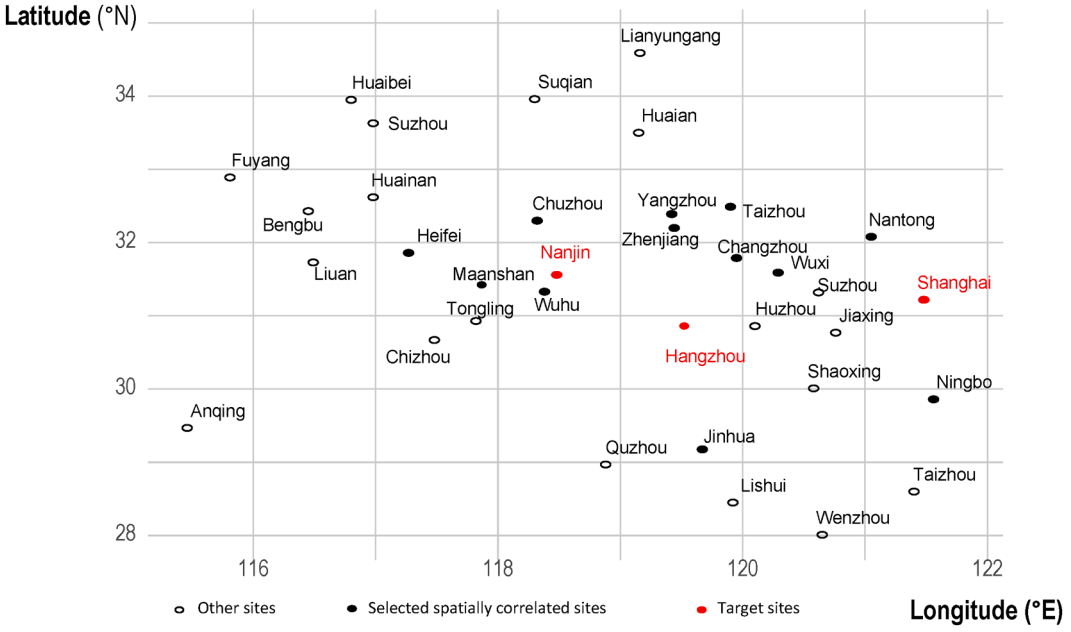


Fig. 9. Geographical distribution of the selected spatially correlated sites.

Table 4
The selected spatially correlated sites and their Pearson correlation coefficients between target sites.

Target sites	Spatially corrlated sites					
Nanjing	city	Zhenjiang	Maanshan	Yangzhou	Changzhou	Chuzhou
	ρ	0.896	0.894	0.870	0.859	0.857
	city	Taizhou	Wuhu	Wuxi	Hefei	
	ρ	0.822	0.822	0.813	0.804	
Hangzhou	city	Wuxi	Ningbo	Wuhu	Maanshan	Jinghua
	ρ	0.748	0.735	0.727	0.713	0.704
Shanghai	city	Nantong	Wuxi	Changzhou	Ningbo	
	ρ	0.815	0.793	0.739	0.714	

Table 5
Regression coefficients of lag features, their standard errors and t-value for target sites.

		Estimate	Std. Error	t-value
Nanjing	Intercept	−0.028	1.002	−0.028
	Lag1	0.525	0.037	14.185
	Lag2	−0.055	0.037	−1.491
Hangzhou	Intercept	0.008	0.882	0.009
	Lag1	0.539	0.031	17.276
Shanghai	Intercept	−0.009	−0.010	0.992
	Lag1	0.578	0.037	15.680
	Lag2	−0.099	0.037	−2.680

only applicable to the region discussed in this article, but also to other parts of China to combat air pollution.

5. Conclusions

This study put forward a novel AQI prediction model based on spatio-temporal effect. The model included spatial auto-correlation analysis and feature selection. Compared with the traditional statistical AQI forecasting model, the proposed SAC-QBSO-SVR considered the spatial correlation and lag effect of the selected sites concurrently. In addition, by using machine learning algorithms, the proposed model has lower complexity and satisfying performance. The spatial auto-correlation variables were constructed by using the residual sequence of the sites and were introduced into the feature set with other features of the target site and their strong spatially correlated sites, to achieve accurate and stable prediction of the AQI. After the predicted values are obtained, based on the air quality standards (GB3095-2012), residents can take corresponding preventive measures; the government can also prepare for bad weather and issue warnings in advance. In addition, by comparing it with several baseline models, it was found that the proposed model has higher prediction accuracy and lower predicting error, taking RMSE, NSE, and MAPE as the evaluation criteria. The main results of this study are as follows:

- The addition of AQI, PM2.5, PM10, CO, NO_2, SO_2, and O_3 concentrations in neighboring stations can contribute to the spatiality of the data and significantly improve the prediction accuracy of the model.
- The construction of spatial SAC variables can reflect the spatial and lag effects of AQI concurrently.
- The model combines a QBSO feature selection algorithm with SVR, which can not only effectively reduce the feature dimension, but also better deal with the nonlinear relationship between variables to obtain higher accuracy.

In fact, there is also spatial heterogeneity between the AQI and atmospheric pollutant concentrations (Yang et al., 2018), that is, the

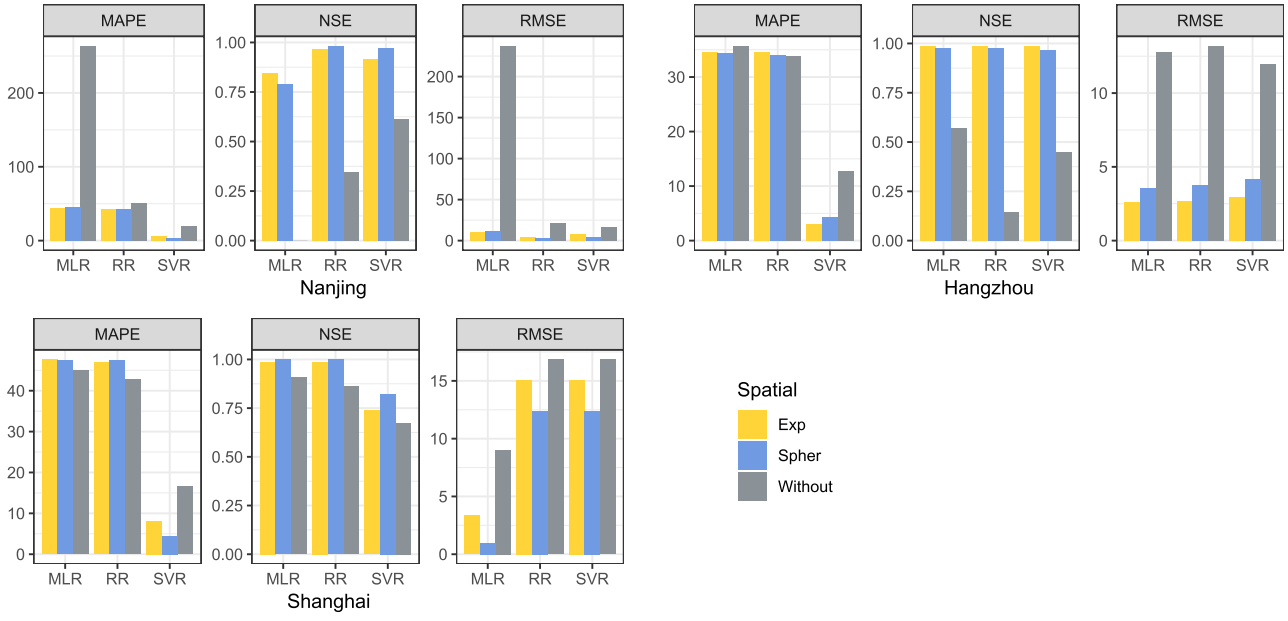


Fig. 10. The influence of SAC variables on prediction results. To facilitate visualization, the negative NSE value of Nanjing is set to 0.

Table 6
The influence of SAC variables on prediction results.

		Nanjing			Hangzhou			Shanghai		
		RMSE	NSE	MAPE	RMSE	NSE	MAPE	RMSE	NSE	MAPE
Without SAC variable	MLR	237.296	−80.395	263.542	12.782	0.569	35.686	9.010	0.907	44.903
	RR	21.296	0.345	50.438	13.191	0.147	33.805	10.986	0.861	42.780
	SVR	16.323	0.615	19.441	11.974	0.448	12.676	16.839	0.674	16.524
With SAC variable	Exp-MLR	10.435	0.843	44.615	2.629	0.987	34.484	3.346	0.987	47.604
	Gau-MLR	11.613	0.805	45.424	2.568	0.988	34.505	3.363	0.987	47.587
	Qua-MLR	10.240	0.848	44.561	2.925	0.984	34.458	3.683	0.984	47.602
	Spher-MLR	12.068	0.789	45.914	3.590	0.976	34.409	0.956	0.999	47.383
	Lin-MLR	7.981	0.908	43.365	4.090	0.969	34.564	4.070	0.734	47.643
	Exp-RR	4.699	0.968	42.645	2.705	0.986	34.616	3.518	0.986	46.892
	Gau-RR	4.808	0.967	42.626	2.770	0.986	34.758	3.542	0.986	46.857
	Qua-RR	4.932	0.965	42.607	2.997	0.983	34.585	3.902	0.982	46.808
	Spher-RR	3.493	0.982	42.956	3.748	0.974	34.038	0.956	0.999	47.383
	Lin-RR	4.862	0.966	42.822	4.111	0.968	34.272	4.089	0.693	47.137
	Exp-SVR	7.528	0.918	6.510	2.921	0.984	3.022	15.076	0.738	8.153
	Gau-SVR	7.881	0.910	6.756	**2.895**	**0.984**	**2.935**	15.155	0.736	8.086
	Qua-SVR	8.313	0.900	7.085	3.241	0.980	3.391	15.690	0.717	8.665
	Spher-SVR	**4.369**	**0.972**	**4.056**	4.158	0.968	4.242	**12.396**	**0.823**	**4.428**
	Lin-SVR	7.883	0.910	7.080	5.559	0.942	5.218	14.597	0.569	9.026

Table 7
The results of QBSO.

Target sites	Nanjing	Hangzhou	Shanghai
Number of the original features	74	51	47
Number of the selected features	32	23	22
Accuracy	0.89	0.86	0.90
Average time to evaluate a solution	0.043 s	0.037 s	0.042 s
The top three most influenting features	NO_2 of Nanjing O_3 of Maanshan PM2.5 of Wuhu	lag1 AQI of Hangzhou NO_2 of Wuxi SO_2 of Ningbo	lag1 AQI of Shanghai O_3 of Ningbo PM2.5 of Shanghai

relationship between the AQI value and atmospheric pollutant concentration and other related variables changes with the change of spatial distribution. Therefore, future research should properly consider spatio-temporal heterogeneity, not just spatial autocorrelation. Due to the availability of the data, other factors affecting concentration, such as temperature, humidity, wind direction, and human activities, are not considered in this paper. In future work, these factors can be considered to further improve the performance of the model. In addition, the experiments in this study were based on data from the Yangtze River Delta of China in 2019–2020, and the results obtained may only be suitable for the study area. In the research of air pollution, the proposed model is not only suitable for AQI data, but also for other pollutants concentration, such as PM2.5, SO_2 and so on. Additionally, the proposed model can be employed to other environmental fields of applications, like water quality, hydrology, and geology, and potentially achieve great forecasting accuracy.

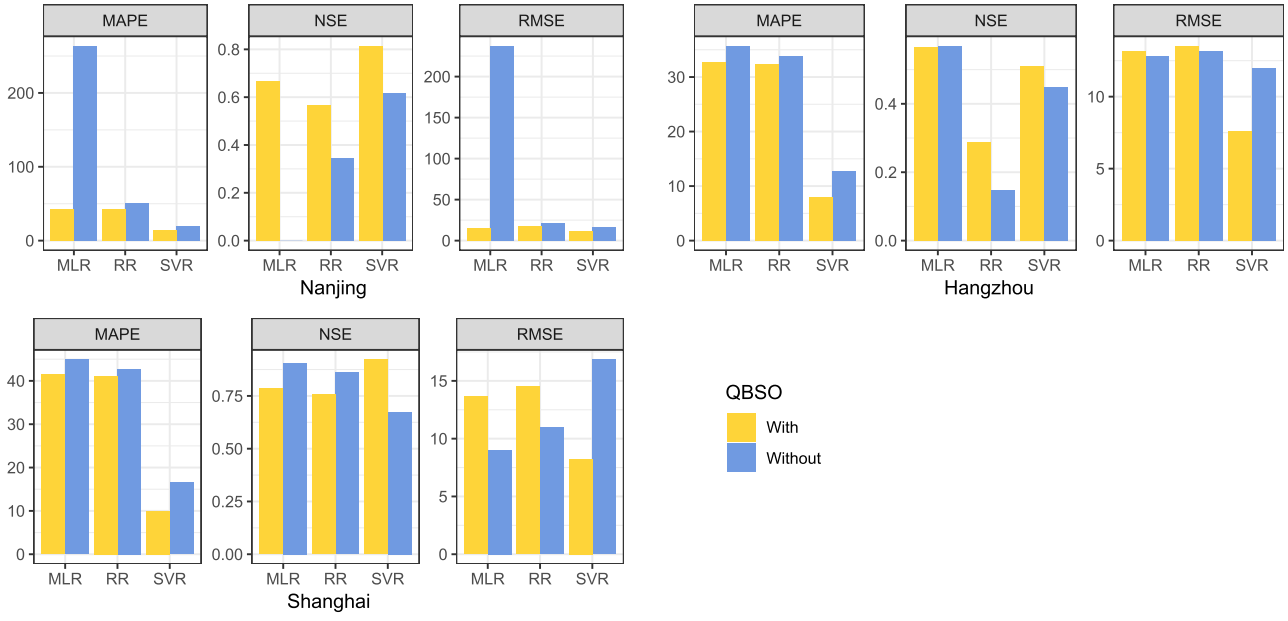


Fig. 11. The impact of QBSO on prediction results. To facilitate visualization, the negative NSE value of Nanjing is set to 0.

Table 8
The influence of QBSO on prediction results.

		Nanjing			Hangzhou			Shanghai		
		RMSE	NSE	MAPE	RMSE	NSE	MAPE	RMSE	NSE	MAPE
Without QBSO	MLR	237.296	−80.395	263.542	12.782	0.569	35.686	9.010	0.907	44.903
	RR	21.296	0.345	50.438	13.191	0.147	33.805	10.986	0.861	42.780
	SVR	16.323	0.615	19.441	11.974	0.448	12.676	16.839	0.674	16.524
With QBSO	MLR	15.244	0.668	42.563	13.176	0.565	32.767	13.649	0.786	41.431
	RR	17.372	0.568	42.510	13.517	0.287	32.408	14.510	0.758	41.162
	SVR	11.407	0.814	14.108	7.583	0.509	7.973	8.229	0.922	9.968

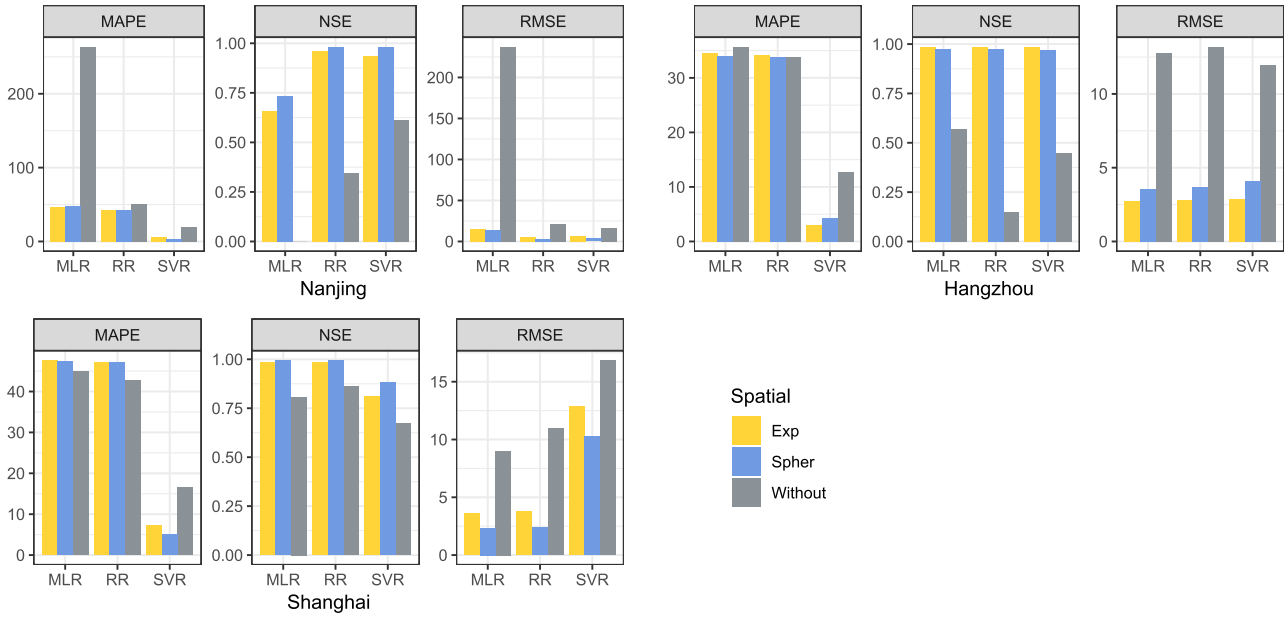


Fig. 12. The influence of SAC variables and QBSO on prediction results. To facilitate visualization, the negative NSE value of Nanjing is set as 0.

Table 9
The influence of SAC variable and QBSO on prediction results.

		Nanjing			Hangzhou			Shanghai		
	Model	RMSE	NSE	MAPE	RMSE	NSE	MAPE	RMSE	NSE	MAPE
Without SAC variable and QBSO	MLR	237.296	−80.395	263.542	12.782	0.569	35.686	9.010	0.807	44.903
	RR	21.296	0.345	50.438	13.191	0.147	33.805	10.986	0.861	42.780
	SVR	16.323	0.615	19.441	11.974	0.448	12.676	16.839	0.674	16.524
With SAC variable and QBSO	Exp-QBSO-MLR	15.327	0.660	47.245	2.748	0.986	34.56	3.592	0.985	47.595
	Gau-QBSO-MLR	17.48	0.660	48.969	2.712	0.986	34.299	3.611	0.985	47.568
	Qua-QBSO-MLR	15.974	0.631	47.559	3.086	0.982	34.157	3.907	0.982	47.609
	Spher-QBSO-MLR	13.614	0.732	47.480	3.585	0.976	33.959	2.346	0.994	47.320
	Lin-QBSO-MLR	13.85	0.723	45.428	4.312	0.965	34.065	4.450	0.977	47.788
	Exp-QBSO-RR	5.091	0.963	42.686	2.800	0.985	34.151	3.761	0.984	47.085
	Gau-QBSO-RR	5.223	0.961	42.677	2.741	0.986	34.241	3.801	0.983	47.034
	Qua-QBSO-RR	5.371	0.958	42.640	3.143	0.982	34.069	4.111	0.981	47.031
	Spher-QBSO-RR	3.574	0.982	43.069	3.708	0.974	33.803	2.396	0.993	47.119
	Lin-QBSO-RR	5.187	0.961	42.795	4.519	0.962	33.866	4.511	0.977	47.269
	Exp-QBSO-SVR	6.627	0.937	5.650	**2.905**	**0.984**	**3.084**	12.849	0.810	7.151
	Gau-QBSO-SVR	6.733	0.934	5.798	2.932	0.984	3.140	12.945	0.807	7.114
	Qua-QBSO-SVR	7.354	0.922	6.116	3.272	0.980	3.490	13.301	0.796	7.634
	Spher-QBSO-SVR	**3.642**	**0.981**	**3.620**	4.098	0.969	4.226	**10.233**	**0.880**	**5.087**
	Lin-QBSO-SVR	6.653	0.936	6.163	5.574	0.942	5.174	12.187	0.829	8.042

CRediT authorship contribution statement

Zixi Zhao: Software, Visualization, Formal analysis, Writing - original draft. **Jinran Wu:** Visualization, Formal analysis, Writing - review & editing, Data curation. **Fengjing Cai:** Investigation, Project administration. **Shaotong Zhang:** Writing - review & editing. **You-Gan Wang:** Supervision, Project administration, Investigation, Writing - review & editing.

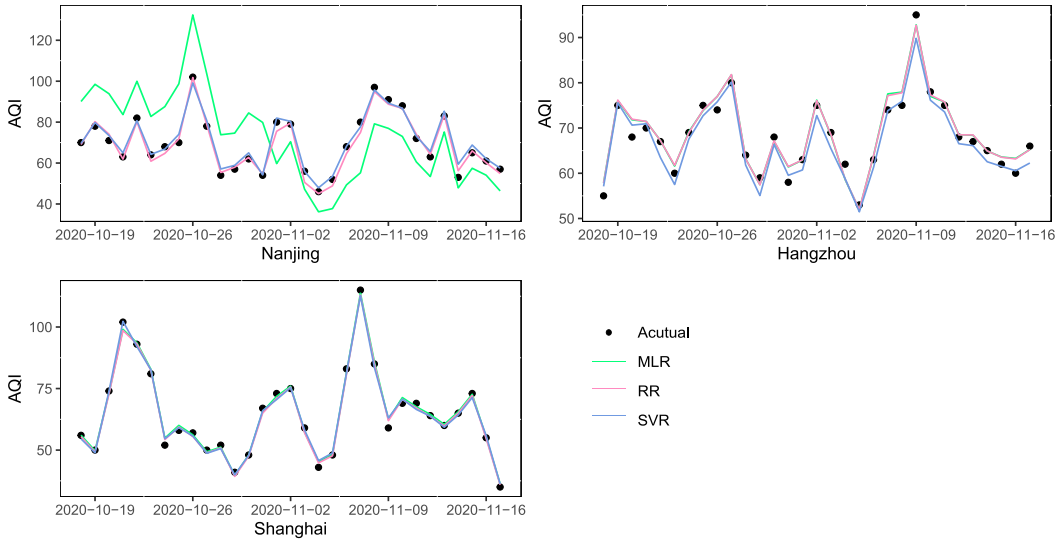


Fig. 13. The prediction of models when SAC variables and QBSO are added into model concurrently.

Declaration of Competing Interest

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

Acknowledgements

This work was supported in part by the Australian Research Council Discovery Project (DP160104292), Zhejiang Provincial Natural Science Foundation of China (Grant No: LY19A010014), and the Science and Technology Innovation Activity Plan for University Students in Zhejiang Province (Grant No: 2021R429049). Also, this work was supported by the Natural Science Foundation of Shandong Province (Grant No: ZR2019BD009).

Appendix A. Multiple linear regression

Multiple linear regression (MLR) is a kind of linear regression that is mainly used to deal with multi-factor problems. In fact, using the optimal combination of multiple independent variables to estimate or forecast the dependent variable is more effective and practical than using a single independent variable (Uyanık and Güler, 2013). Therefore, MLR is more commonly used than linear regression, which is used extensively in econometrics and time series analysis. The equation of MLR is as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon, \tag{A.1}$$

and the estimation of β can be solved by a matrix operation:

$$\widehat{\beta} = (X^T X)^{-1} X^T y = \left(\sum x_i x_i^T\right)^{-1} \left(\sum x_i y_i\right) \tag{A.2}$$

where $i = 1, 2, \cdots, n$, n is the number of observations; y_i is the dependent variable; x_i is the explanatory variable; β_0 is the intercept; β_p is the slope coefficients for each explanatory variable; and ϵ represents the model's error term.

Appendix B. Ridge regression

Ridge regression (RR) is a possible method to solve the imprecision of least square estimation when multiple collinearities exist in the linear regression model. Different from the unbiased estimation of linear regression, the advantage of RR lies in its unbiased estimation, which tends to shrink some coefficients towards 0. Therefore, it can alleviate the problems of multicollinearity and overfitting (McDonald, 2009).

Unlike MLR, the estimation of β is calculated as follows:

$$\widehat{\beta}_{\text{ridge}} = \left(X^T X + k I_n\right)^{-1} X^T Y \tag{B.1}$$

where I_n is the $n \times n$identity matrix and $k > 0$.

References

Hao, Yu, Peng, Hui, Temulun, T., Liu, Li-Qun, Mao, Jie, Lu, Zhi-Nan, Chen, Hao, 2018. How harmful is air pollution to economic development? new evidence from pm2. 5 concentrations of chinese cities. J. Clean. Prod. , 172: 743–757, 2018.

Glencross, Drew A., Ho, Tzer-Ren, Camina, Nuria, Hawrylowicz, Catherine M., Pfeffer, Paul E., 2020. Air pollution and its effects on the immune system. Free Radical Biol. Med. 151, 56–68.

Jans, Jenny, Johansson, Per, Nilsson, J. Peter, 2018. Economic status, air quality, and child health: Evidence from inversion episodes. J. Health Econ. 61, 220–232.

Xi, Lu, Zhang, Shaojun, Xing, Jia, Wang, Yunjie, Chen, Wenhui, Ding, Dian, Ye, Wu, Wang, Shuxiao, Duan, Lei, Hao, Jiming, 2020. Progress of air pollution control in china and its challenges and opportunities in the ecological civilization era. Engineering 6 (12), 1423–1431.

Li, Q., Peng, C.H., 2016. The stock market effect of air pollution: evidence from china. Appl. Econ. 48 (36), 3442–3461.

Liu, Hui, Yin, Shi, Chen, Chao, Duan, Zhu, 2020. Data multi-scale decomposition strategies for air pollution forecasting: A comprehensive review. J. Clean. Prod. 277, 124023.

Zannetti, Paolo, 2013. Air pollution modeling: theories, computational methods and available software. Springer Science & Business Media.

Yang, Xiaochun, Wu, Qizhong, Zhao, Rong, Cheng, Huaqiong, He, Huijuan, Ma, Qian, Wang, Lanning, Luo, Hui, 2019. New method for evaluating winter air quality: Pm2. 5 assessment using community multi-scale air quality modeling (cmaq) in xi'an. Atmosp. Environ., 211: 18–28, 2019.

Pino-Cortés, Ernesto, Carrasco, Samuel, Acosta, Jonathan, de Almeida Albuquerque, Taciana Toledo, Pedruzzi, Rizzieri, Díaz-Robles, Luis A, 2022. An evaluation of the photochemical air quality modeling using cmaq in the industrial area of quintero-puchuncavi-concon, chile. Atmosp. Pollut. Res., 13 (3): 101336.

Tan, Jiani, Zhang, Yan, Ma, Weichun, Qi, Yu, Wang, Qian, Qingyan, Fu, Zhou, Bin, Chen, Jianmin, Chen, Limin, 2017. Evaluation and potential improvements of wrf/cmaq in simulating multi-levels air pollution in megacity shanghai, china. Stoch. Env. Res. Risk Assess. 31 (10), 2513–2526.

Sati, Ankur P., Mohan, Manju, 2021. Impact of increase in urban sprawls representing five decades on summer-time air quality based on wrf-chem model simulations over central-national capital region, india. Atmospheric. Pollut. Res. 12 (2), 404–416.

Kong, Lei, Tang, Xiao, Zhu, Jiang, Wang, Zifa, Li, Jianjun, Huangjian, Wu, Qizhong, Wu, Chen, Huansheng, Zhu, Lili, Wang, Wei, et al., 2021. A 6-year-long (2013–2018) high-resolution air quality reanalysis dataset in china based on the assimilation of surface observations from cnemc. Earth Syst. Sci. Data 13 (2), 529–570.

Ma, Jun, Cheng, Jack C.P., Lin, Changqing, Tan, Yi, Zhang, Jingcheng, 2019. Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. Atmos. Environ. 214, 116885.

Stern, R., Builtjes, P., Schaap, M., Timmermans, R., Vautard, R., Hodzic, A., Memmesheimer, M., Feldmann, H., Renner, E., Wolke, R., et al., 2008. A model inter-comparison study focussing on episodes with elevated pm10 concentrations. Atmos. Environ. 42 (19), 4567–4588.

Delavar, Mahmoud Reza, Gholami, Amin, Shiran, Gholam Reza, Rashidi, Yousef, Nakhaeizadeh, Gholam Reza, Fedra, Kurt, Afshar, Smaeil Hatefi, 2019. A novel method for improving air pollution prediction based on machine learning approaches: a case study applied to the capital city of tehran. ISPRS International Journal of Geo-Information, 8 (2): 99, 2019.

Callens, Aurelien, Wang, You-Gan, Liya, Fu, Liquet, Benoit, 2021. Robust estimation procedure for autoregressive models with heterogeneity. Environ. Model. Assess. 26 (3), 313–323.

Stadlober, Ernst, Hörmann, Siegfried, Pfeiler, Brigitte, 2008. Quality and performance of a pm10 daily forecasting model. Atmos. Environ. 42 (6), 1098–1109.

Gocheva-Ilieva, Snezhana Georgieva, Ivanov, Atanas Valev, Voynikova, Desislava Stoyanova, Boyadzhiev, Doychin Todorov, 2014. Time series analysis and forecasting for air pollution in small urban area: an sarima and factor analysis approach. Stochastic Environmental Research And Risk Assessment 28 (4), 1045–1060.

Koo, Jian Wei, Wong, Shin Wee, Selvachandran, Ganeshsree, Long, Hoang Viet, Son, Le Hoang, 2020. Prediction of air pollution index in kuala lumpur using fuzzy time series and statistical models. Air Qual., Atmosp. Health, 13 (1): 77–88, 2020.

Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., Vitabile, S., 2008. Three hours ahead prevision of so2 pollutant concentration using an elman neural based forecaster. Build. Environ. 43 (3), 304–314.

Ma, Jun, Cheng, Jack C.P., 2017. Identification of the numerical patterns behind the leading counties in the us local green building markets using data mining. J. Clean. Prod. 151, 406–418.

Li, Weide, Kong, Demeng, Jinran, Wu, 2017. A new hybrid model fpa-svm considering cointegration for particular matter concentration forecasting: a case study of kunming and yuxi, china. Computat. Intell. Neurosci. 2017.

Wang, Junshan, Song, Guojie, 2018. A deep spatial-temporal ensemble model for air quality prediction. Neurocomputing 314, 198–206.

Qunli, Wu, Lin, Huaxing, 2019. Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and lstm neural network. Sustain. Cities Soc. 50, 101657.

Maleki, Heidar, Sorooshian, Armin, Goudarzi, Gholamreza, Baboli, Zeynab, Birgani, Yaser Tahmasebi, Rahmati, Mojtaba, 2019. Air pollution prediction by using an artificial neural network model. Clean Technol. Environ. Policy 21 (6), 1341–1352.

Qiao, Junfei, He, Zengzeng, Shengli, Du, 2020. Prediction of pm2.5 concentration based on weighted bagging and image contrast-sensitive features. Stoch. Env. Res. Risk Assess. 34 (3), 561–573.

Alimissis, A., Philippopoulos, K., Tzanis, C.G., Deligiorgi, D., 2018. Spatial estimation of urban air pollution with the use of artificial neural network models. Atmos. Environ. 191, 205–213.

Li, Hongmin, Wang, Jianzhou, Yang, Hufang, 2020. A novel dynamic ensemble air quality index forecasting system. Atmosp. Pollut. Res. 11 (8), 1258–1270.

Ketu, Shwet, Mishra, Pramod Kumar, 2021. Scalable kernel-based svm classification algorithm on imbalance air quality data for proficient healthcare. Compl. Intell. Syst. 7 (5), 2597–2615.

Liu, Wei, Guo, Geng, Chen, Fuji, Chen, Yihui, 2019. Meteorological pattern analysis assisted daily pm2. 5 grades prediction using svm optimized by pso algorithm. Atmospheric. Pollut. Res. 10 (5), 1482–1491.

Drucker, Harris, Burges, Christopher J., Kaufman, Linda, Smola, Alex, Vapnik, Vladimir, 1996. Support vector regression machines. Adv. Neural Inform. Process. Syst. 9, 155–161.

Robert Kurniawan, I., Setiawan, Nyoman, Caraka, Rezzy Eko, Nasution, Bahrul Ilmi, 2022. Using harris hawk optimization towards support vector regression to ozone prediction. Stoch. Env. Res. Risk Assess. 36 (2), 429–449.

Ge, Liang, Kunyan, Wu, Zeng, Yi, Chang, Feng, Wang, Yaqian, Li, Siyu, 2021. Multi-scale spatiotemporal graph convolution network for air quality prediction. Appl. Intell. 51 (6), 3491–3505.

Tobler, Waldo R., 1970. A computer movie simulating urban growth in the detroit region. Econ. Geogr. 46 (sup1), 234–240.

Liu, Hui, Yang, Rui, 2021. A spatial multi-resolution multi-objective data-driven ensemble model for multi-step air quality index forecasting based on real-time decomposition. Comput. Ind. 125, 103387.

Phruksahiran, Narathep, 2021. Improvement of air quality index prediction using geographically weighted predictor methodology. Urban Climate 38, 100890.

Parbat, Debanjan, Chakraborty, Monisha, 2020. A python based support vector regression model for prediction of covid19 cases in india. Chaos Solitons Fractals 138, 109942.

Brereton, Richard G., Lloyd, Gavin R., 2010. Support vector machines for classification and regression. Analyst 135 (2), 230–267.

Patle, Arti, Chouhan, Deepak Singh, 2013. Svm kernel functions for classification. In 2013 International Conference on Advances in Technology and Engineering (ICATE), pages 1–9. IEEE, 2013.

Gopi, Arepalli Peda, Jyothi, R., Lakshman Narayana, V., Satya Sandeep, K., 2020. Classification of tweets data based on polarity using improved rbf kernel of svm. Int. J. Inform. Technol. 1–16.

Sadeg, Souhila, Hamdad, Leila, Remache, Amine Riad, Karech, Mehdi Nedjmeddine, Benatchba, Karima, Habbas, Zineb, 2019. Qbso-fs: A reinforcement learning based bee swarm optimization metaheuristic for feature selection. In International Work-Conference on Artificial Neural Networks, pages 785–796. Springer, 2019.

Kumar, Aviral, Zhou, Aurick, Tucker, George, Levine, Sergey, 2020. Conservative q-learning for offline reinforcement learning. Adv. Neural Inform. Process. Syst. 33, 1179–1191.

Djenouri, Youcef, Belhadi, Asma, Belkebir, Riadh, 2018. Bees swarm optimization guided by data mining techniques for document information retrieval. Expert Syst. Appl. 94, 126–136.

Djenouri, Youcef, Djenouri, Djamel, Belhadi, Asma, Fournier-Viger, Philippe, Chun-Wei Lin, Jerry, Bendjoudi, Ahcene, 2019. Exploiting gpu parallelism in improving bees swarm optimization for mining big transactional databases. Inform. Sci., 496: 326–342, 2019.

Legendre, Pierre, 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74 (6), 1659–1673.

Griffith, Daniel A., 2011. Positive spatial autocorrelation impacts on attribute variable frequency distributions. Chilean J. Stat. 2 (2), 3–28.

Lichstein, Jeremy W., Simons, Theodore R., Shriner, Susan A., Franzreb, Kathleen E., 2002. Spatial autocorrelation and autoregressive models in ecology. Ecol. Monogr. 72 (3), 445–463.

Behrens, Thorsten, Schmidt, Karsten, Viscarra, Raphael A., Rossel, Philipp Gries, Scholten, Thomas, MacMillan, Robert A., 2018. Spatial modelling with euclidean distance fields and machine learning. Eur. J. Soil Sci. 69 (5), 757–770.

Benesty, Jacob, Chen, Jingdong, Huang, Yiteng, Cohen, Israel, 2009. Pearson correlation coefficient. In: Noise Reduction in Speech Processing. Springer, pp. 1–4.

Cressie, Noel, 2015. Statistics for spatial data. John Wiley & Sons.

Liu, Hui, Chen, Chao, 2020. Spatial air quality index prediction model based on decomposition, adaptive boosting, and three-stage feature selection: A case study in china. J. Clean. Prod. 265, 121777.

Valentini, Marlon, dos Santos, Gabriel Borges, Vieira, Bruno Muller, 2021. Multiple linear regression analysis (mlr) applied for modeling a new wqi equation for monitoring the water quality of mirim lagoon, in the state of rio grande do sul–brazil. SN Appl. Sci., 3 (1): 1–11, 2021.

McDonald, Gary C., 2009. Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics 1 (1), 93–100.

Ma, Tao, Duan, Fengkui, Kebin He, Yu, Qin, Dan Tong, Geng, Guannan, Liu, Xuyan, Li, Hui, Yang, Shuo, Ye, Siqi, et al., 2019. Air pollution characteristics and their relationship with emissions and meteorology in the yangtze river delta region during 2014–2016. J. Environ. Sci. 83, 8–20.

Yang, Wentao, Deng, Min, Feng, Xu, Wang, Hang, 2018. Prediction of hourly pm2. 5 using a space-time support vector regression model. Atmos. Environ. 181, 12–19.

Uyanık, Gülden Kaya, Güler, Neşe, 2013. A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences, 106: 234–240.