TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/22 pp99–111
DOI: 10.26599/TST.2022.9010060
Volume 29, Number 1, February 2024

A Hybrid Air Quality Prediction Model Based on Empirical Mode Decomposition

Yuxuan Cao, Difei Zhang, Shaoqi Ding, Weiyi Zhong, and Chao Yan*

Abstract: Air pollution is a severe environmental problem in urban areas. Accurate air quality prediction can help governments and individuals make proper decisions to cope with potential air pollution. As a classic time series forecasting model, the AutoRegressive Integrated Moving Average (ARIMA) has been widely adopted in air quality prediction. However, because of the volatility of air quality and the lack of additional context information, i.e., the spatial relationships among monitor stations, traditional ARIMA models suffer from unstable prediction performance. Though some deep networks can achieve higher accuracy, a mass of training data, heavy computing, and time cost are required. In this paper, we propose a hybrid model to simultaneously predict seven air pollution indicators from multiple monitoring stations. The proposed model consists of three components: (1) an extended ARIMA to predict matrix series of multiple air quality indicators from several adjacent monitoring stations; (2) the Empirical Mode Decomposition (EMD) to decompose the air quality time series data into multiple smooth sub-series; and (3) the truncated Singular Value Decomposition (SVD) to compress and denoise the expanded matrix. Experimental results on the public dataset show that our proposed model outperforms the state-of-art air quality forecasting models in both accuracy and time cost.

Key words: air quality prediction; Empirical Mode Decomposition (EMD); Singular Value Decomposition (SVD); AutoRegressive Integrated Moving Average (ARIMA)

1 Introduction

Air pollution is the contamination of pollutants that threaten human health and the planet as a whole[1]. It is becoming severe in urban areas with the expansion of urbanization. An estimated seven million people are killed by exposure to outdoor environments with air pollution every year[2]. Therefore, air pollution has become a major environmental health problem in the world, especially in the low- and middle-income countries.

In order to provide accurate measurements of air quality and protect residents from the dangers of ambient air pollution, many air quality monitoring stations have been established in different areas of major cities. Those monitoring stations collect and release real-time air quality data, which includes concentrations of SO_2, NO_2, CO, O_3, $PM_{2.5}$, PM_{10}, and Air Quality Index (AQI) calculated by the aforementioned concentrations. Besides monitoring the current air quality, it is essential to predict future data in next hours or days. Accurate

• Yuxuan Cao, Shaoqi Ding, and Weiyi Zhong are with School of Computer Science, Qufu Normal University, Rizhao 276826, China. E-mail: yuxuancao@qfnu.edu.cn; dsq981230@163.com; weiyizhong@qfnu.edu.cn.

• Difei Zhang is with School of Mathematical Sciences, Qufu Normal University, Qufu 273165, China. E-mail: dreamofcloudsss@gmail.com.

• Chao Yan is with School of Computer Science, Qufu Normal University, Rizhao 270826, China, and also with the College of Economic and Management, Shandong University of Science and Technology, Qingdao 250307, China. E-mail: yanchao@qfnu.edu.cn.

* To whom correspondence should be addressed.

Manuscript received: 2022-09-18; revised: 2022-11-15; accepted: 2022-11-26

and reliable air quality prediction can provide valuable information for residents' daily travel, the government's environmental management, and air pollution prevention and control[3].

Air quality prediction is usually viewed as a time series forecasting problem. With the rise of artificial intelligence[4, 5], existing time series forecasting models, including regression models[6], machine learning models[7], and deep learning models[8–10], have been applied in air quality prediction.

However, there are still many challenges for air quality prediction. First, because air quality is influenced by many factors, such as weather conditions, spatial features, even electronic devices[11], and so on, it is impossible to model all the factors. Second, air quality varies significantly at different times and locations. It is difficult to achieve high accuracy if we do not take the spatial and temporal features into account. Third, because of the vulnerability of the network and monitoring devices, the collected air quality data may be incomplete[12]. In another word, there is a lot of noise existing in the collected data, which may influence the accuracy of prediction models.

In view of the aforementioned challenges, we propose a hybrid model which integrates the extended AutoRegressive Integrated Moving Average (ARIMA) model with the Empirical Mode Decomposition (EMD)[13] and Singular Value Decomposition (SVD)[14] to predict air quality data in the next hours. In this study, we first extend the classic ARIMA model to cope with matrix time series forecasting. Then, we use EMD to decompose the non-stationary air quality series data into multiple smooth sub-series and merge them into a new matrix series. Finally, we compress and denoise the matrix using truncate SVD, and feed the matrix series into an extended ARIMA model. Concretely, the main contributions of this study are as follows:

(1) We extend the classical ARIMA model to support matrix time series forecasting, which can predict all seven air quality indicators from multiple monitoring stations simultaneously.

(2) To better handle the fluctuating data, we decompose original non-stationary air quality data into smooth sub-series by the empirical mode decomposition method.

(3) We use truncated SVD to compress air quality series to remove the noisy data and capture correlations among air pollutants and neighbor stations.

(4) We conducted experiments on a real-world air quality dataset. Experimental results show that our proposed model can improve the accuracy and reduce the computational cost compared with the state-of-art air quality prediction models.

The remainder of this paper is organized as follows. We present recent related research on air quality prediction in Section 2. Section 3 introduces our method in detail and presents the formulation process. The experimental evaluation is given in Section 4, and Section 5 concludes our work.

2 Related Work

Air quality prediction is a hot research topic in recent years, and many prediction methods have been proposed. With the development of artificial intelligence and big data technology[15], machine learning and deep models have also attracted attention in various fields[16–18]. For example, the random forest model[19], support vector regression[20], and long short-term memory[21] are used in air quality prediction. Traditional models are still widely used due to their simplicity and efficiency.

2.1 Air quality prediction based on traditional models

There have been a number of statistical methods widely used in air quality prediction, e.g., the classical ARIMA model and linear regression. Zhang et al.[22] proposed a new hybrid prediction model using a nonlinear autoregressive (namely NARX) network and ARMA to solve the deterministic part and error part of meteorological data, respectively. The results of experiments conducted on the data of NO_2, SO_2, $PM_{2.5}$, and O_3 in Beijing show that the model can be effectively used for the short-term prediction of air pollutant concentrations. Wang et al.[23] proposed a hybrid Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) model based on ARIMA and Support Vector Machine (SVM) to model panel data of air pollution and meteorological factors, where the SVM was used to learn nonlinear information that cannot be captured by ARIMA, and the GARCH model was used to deal with the conditional heteroskedasticity. Experiments on $PM_{2.5}$ from six monitoring stations in Shenzhen showed that the hybrid model outperforms the individual models and was a reliable and effective model for predicting $PM_{2.5}$. Ding et al.[24] employed an ARIMA-GARCH model to make short-term forecasts of subway patronage. The models of ARIMA and GARCH extracte the deterministic volatility component of the passenger flow seperately, and the results of prediction on three stations showed that the performance was superior to that of the conventional model. Moisan et al.[25] proposed a linear model based on dynamic multiple linear equations, which is still competitive in predicting $PM_{2.5}$ concentrations compared to nonlinear models such as neural networks, and does not suffer from overfitting and even surpasses the performance of nonlinear models. Du et al.[26] developed an end-to-end hybrid deep air quality prediction framework for learning the spatial and temporal characteristics of air quality time series data from multiple monitoring stations for $PM_{2.5}$ prediction.

In addition, attention-based neural networks and graph convolutional network based[27] air quality prediction methods also have received extensive attention in recent years.

2.2 Air quality prediction based on data decomposition

As the air quality data often changes sharply and the traditional prediction models cannot fully utilize the information of different frequencies of the data, some methods based on data decomposition are proposed[28]. Fan et al.[29] proposed a hybrid prediction model based on wavelet decomposition, using the wavelet decomposition technique to decompose the original air quality time series into high and low frequency components, and then used Long Short-Term Memory (LSTM) and ARIMA to predict those two components seperately. The experimental results showed a large improvement with respect to the single model. Based on the data decomposition method, Jin et al.[30] decomposed the original $PM_{2.5}$ data into trend part, period part, and residual part, and then used ARIMA and Gated Recurrent Unit (GRU) models for the prediction of the three parts seperately. Experiments on Beijing $PM_{2.5}$ data indicated that the proposed prediction model can effectively improve the accuracy of long-term prediction. Altıntaş and Davidson[31] combined the improved EMD with SVR and proposed a hybrid prediction model for short-term high-speed traffic flow prediction. The experimental results showed that the prediction accuracy of the model with EMD was effectively improved compared with the single SVR model. A hybrid prediction model SD-EMD-LSTM was proposed by Zheng et al.[32] to predict electrical load. The model clusters the power data of historical days, which are similar to the forecast days, by the K-means algorithm to determine the training data. Then the model decomposes the data by EMD and applies LSTM to each decomposed sub-series for forecasting. Jin et al.[33] decomposed the historical $PM_{2.5}$ data through EMD. All components were divided into three groups, the high, medium, and low frequencies, through the convolutional neural network CNN, and then the problem of different numbers of components of EMD was solved. The GRU was then used to predict each of the three frequencies. Experiments were carried out on $PM_{2.5}$ data, and the results showed the effectiveness of the method. Huang et al.[34] input all the smooth sequences of EMD into GRU for training and prediction separately, while adding meteorological features to GRU. Compared with the single GRU model, the performance of predicting $PM_{2.5}$ was significantly improved.

Wang et al.[35] decomposed streamflow data by EMD and Ensemble EMD (EEMD), and then built ARIMA models for each sub-series separately for prediction, and the experiments showed the effectiveness of EMD/EEMD-ARIMA prediction models for long-term runoff prediction. Fatema et al.[36] proposed a hybrid forecasting method, combining EMD, ARIMA, and Monte Carlo Simulation (MCS), which could learn linear and nonlinear behavior well for medical tourism forecasting.

However, these works simply obtain multiple relatively smooth components by applying EMD to the raw series. The potential relationships between different monitoring stations are not considered in their studies. In addition, these methods can only predict one pollutant concentration. To predict other pollutants, new predictors are needed to be developed.

3 Methodology

To overcome the aforementioned shortcomings of existing air quality prediction models, we propose a novel approach named SE-ARIMA to predict all pollutants from multiple stations at one time. The workflow of our method is shown in Fig. 1, where IMF1, IMF2, and IMF3 represent the sub-series of EMD, which are the intrinsic mode functions. First, we treat the seven air quality indicators from multiple stations at the current time slot as a matrix. Therefore, historical data in recent T time slots are treated as a matrix series. Secondly, we use EMD to decompose each element series in the matrix series into three smooth sub-series. Thirdly, we use truncated SVD to remove noise in the data, then fed

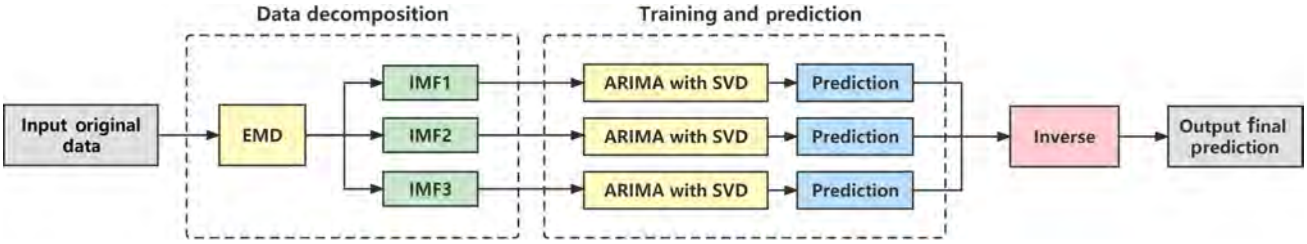


Fig. 1 Flow chart of the SE-ARIMA model for prediction.

the compressed matrix into the extended ARIMA model to make predictions for the three series. Finally, we adopt inverse EMD, i.e., summing the three sub-signals corresponding to each air quality indicator to obtain the final results.

3.1 Preliminaries

3.1.1 ARIMA

As a classical model, ARIMA is extensively used in time series forecasting. Since it was proposed [37], many variants have been presented.

Let y_t denote the value of the real data of the time series at time t, the AutoRegressive model (AR) can be expressed as

$$y_t = \mu + \sum_{i=1}^{p} \alpha_i y_{t-i} + \varepsilon_t \tag{1}$$

the Moving Average model (MA) can be expressed as

$$y_t = \mu + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} + \varepsilon_t \tag{2}$$

and the ARMA model can be expressed as

$$y_t = \mu + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} + \varepsilon_t \tag{3}$$

where μ is a constant, ε_t denotes random error, the mean of ε_t is zero and the variance is constant. α_i and β_i are the parameters of AR and MA. p and q are the numbers of AR and MA terms.

The ARIMA model can perform differential processing on time series data and convert non-stationary data into stationary data. Let d denote the order of the difference, then the d-order difference of the original sequence y_t can be represented as $\Delta^d y_t$. The ARIMA can be represented as

$$\Delta^d y_t = \mu + \sum_{i=1}^{p} \alpha_i \Delta^d y_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} + \varepsilon_i \tag{4}$$

3.1.2 EMD

Empirical mode decomposition, proposed by Huang et al.[38], is an efficient algorithm for dealing with nonlinear and non-stationary time-series data adaptively. It can decompose a non-stationary series into multiple smooth IMFs and a residual series based on the local characteristics of the data. Figure 2 shows the multiple IMFs and a residual (namely RES) obtained from the empirical modal decomposition of 600 hours of $PM_{2.5}$ data.

As shown in Fig. 3, since the characteristics and distributions of different series data are different, the number of IMFs obtained by EMD are different. That will result in the inconsistent size of our matrix series, which cannot be modeled by ARIMA.

The problem of the different numbers of components can be fixed by limiting the number of decompositions. We set it to stop after decomposing three times to ensure that the size of all decomposed matrices remains the same, all expanding to three times their initial size, and the series of decomposed matrices through EMD can be

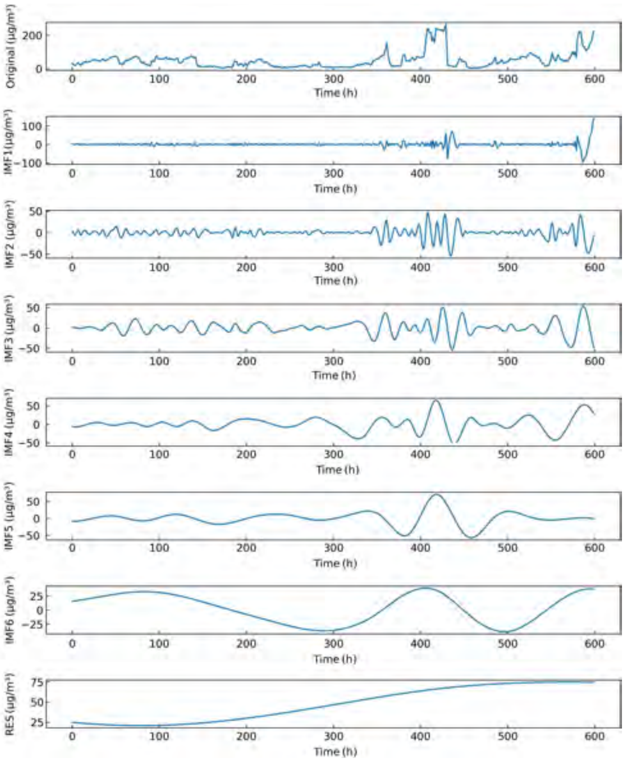
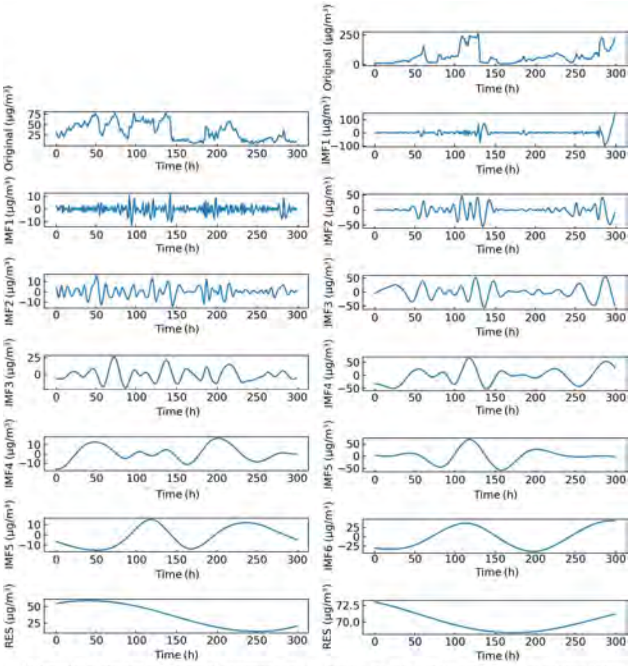


Fig. 2 Results of empirical mode decomposition of $PM_{2.5}$.



(a) Results of the first 300 hours of decomposition (b) Results of the second 300 hours of decomposition

Fig. 3 Results of decomposed PM$_{2.5}$ series in different time periods.

modeled and predicted by ARIMA.

In addition, the reality is that the number of components obtained by EMD for some sequences is less than three, and even does not meet the initial conditions of EMD (the results obtained by EMD for these sequences is still the original sequences). For this case, we add one or two series of 0 values, so that all series satisfy the requirement of three components. However, this approach is equivalent to introducing noisy data, and the truncated SVD which is introduced can solve this problem.

3.1.3 Truncated SVD

The truncated SVD provides optimal low-rank matrix approximation for the original matrix, which can be used to reduce matrix dimension, extract the main feature of data, and remove data noise[14]. For a given arbitrary matrix Z with m rows and n columns, the singular value decomposition of the matrix Z is shown below:

$$Z = P \Sigma V^{\mathrm{T}},\quad \text{s.t.,}\ P P^{\mathrm{T}} = I,\ V V^{\mathrm{T}} = I \tag{5}$$

where the diagonal matrix Σ is a singular value matrix. The elements of the diagonal of matrix Σ are the singular values of Z and are arranged in the order from largest to smallest. P and V are unitary matrices, each row and column of which represents the left and right singular vectors of the matrix Z, respectively. The first r columns of V are selected to construct the $n \times r$ feature matrix, where $r \ll \min(m, n)$, then the columns of the original air-quality matrix Z can be compressed.

3.2 ARIMA with EMD and truncated SVD

Let $\chi = \{\chi_1, \chi_2, \ldots, \chi_t, \ldots, \chi_T\}$ denote the original matrix sequence of air quality from 1 to T moments, which can be considered as a time series, where χ_t denotes the air quality matrix. Three new matrix sequences χ^{emd} can be obtained through decomposition of χ by EMD. For each $\chi^{\text{emd}} = \{\chi_1^{\text{emd}}, \chi_2^{\text{emd}}, \ldots, \chi_t^{\text{emd}}, \ldots, \chi_T^{\text{emd}}\}$, the d-order difference $\Delta^d \chi^{\text{emd}}$ of χ^{emd} can be denoted as

$$\Delta^d \chi^{\text{emd}} = \left\{\Delta^d \chi_{d+1}^{\text{emd}}, \Delta^d \chi_{d+2}^{\text{emd}}, \ldots, \Delta^d \chi_T^{\text{emd}}\right\} \tag{6}$$

Compressing $\Delta^d \chi^{\text{emd}}$ by truncated SVD, the compressed matrix $\Delta^d K_t$ can be denoted as

$$\Delta^d K_t = \Delta^d \chi_t^{\text{emd}} V_{\text{reduce}},\quad \text{s.t.,}\ V_{\text{reduce}} V_{\text{reduce}}^{\mathrm{T}} = I \tag{7}$$

where V_{reduce} denotes the truncated matrix V. Then $\Delta^d \chi^{\text{emd}}$ can be restituted by the product of $\Delta^d K_t$ and $V_{\text{reduce}}^{\mathrm{T}}$,

$$\Delta^d \hat{\chi}_t^{\text{emd}} = \Delta^d K_t V_{\text{reduce}}^{\mathrm{T}} \tag{8}$$

Our first optimization objective is to minimize the error betweenand $\Delta^d \chi_t^{\text{emd}}$ and $\Delta^d \hat{\chi}_t^{\text{emd}}$.

Furthermore, the compressed air quality matrix sequence is modeled by the extended ARIMA model. Then, the improved ARIMA model is expressed as

$$\Delta^d K_t = \sum_{i=1}^{p} \alpha_i \Delta^d K_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} + \varepsilon_t \tag{9}$$

Our second optimization objective is to minimize ε_t to zero. Therefore, we define the objective function as

$$\min_{\substack{\{\Delta^d K_t,\\ V_{\text{reduce}}\varepsilon_{t-i},\\ \alpha_i, \beta_i\}}} \sum_{t=s+1}^{T} \left(\left\| \Delta^d K_t - \Delta^d \chi_t^{\text{emd}} V_{\text{reduce}} \right\|_{\mathrm{F}}^2 + \left\| \Delta^d K_t - \sum_{i=1}^{p} \alpha_i \Delta^d K_{t-i} - \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} \right\|_{\mathrm{F}}^2 \right) \tag{10}$$

where $s = p + d + q$ is the minimum length of the time slot. For the two parts of Formula (10), we consider them to be equally important, so the weights are the same.

The above objective function can be minimized by the augmented Lagrangian method. First, we determine V_{reduce}, ε_{t-i}, α_i, and β_i. Computing the partial derivation of Formula (10) with respect to $\Delta^d K_t$, then equalizing it to zero. The update formula is as follows:

$$\Delta^d K_t = \frac{1}{2}\left(\Delta^d \chi_t^{\text{emd}} V_{\text{reduce}} + \sum_{i=1}^{p} \alpha_i \Delta^d K_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}\right) \quad (11)$$

Formula (10) with respect to V_{reduce} is

$$\min_{\{V_{\text{reduce}}\}} \sum_{t=s+1}^{T} \left(\left\| \Delta^d K_t - \Delta^d \chi_t^{\text{emd}} V_{\text{reduce}} \right\|_{\text{F}}^2 \right) \quad (12)$$

This is the problem equivalent to the orthogonal procrustes. The global optimal solution of Formula (12) is

$$\sum_{t=s+1}^{T} \left(\Delta^d K_t^{\text{T}} \Delta^d \chi_t^{\text{emd}} \right) = M \Sigma N^{\text{T}} \quad (13)$$

where M and N are the left and right singular vectors of the singular value decomposition of $\sum_{t=s+1}^{T} \left(\Delta^d K_t^{\text{T}} \Delta^d \chi_t^{\text{emd}} \right)$, respectively.

We use the Yule-Walker method for estimating the parameters α_i and β_i of AR and MA part. Computing the partial derivation of Formula (10) with respect to ε_{t-i}, then equalizing it to zero. The update formula of ε_{t-i} is given as

$$\varepsilon_{t-i} = \frac{\sum_{t=s+1}^{T} \left(\Delta^d K_t - \sum_{i=1}^{p} \alpha_i \Delta^d K_{t-i} + \sum_{k \neq i}^{q} \beta_k \varepsilon_{t-k} \right)}{(s+1-T)\beta_i} \quad (14)$$

The pseudocode of the training process for our method is detailed in Algorithm 1.

3.3 Prediction

We compute the compression matrix for the prediction at moment $T+1$ by using the following equation:

$$\Delta^d \hat{K}_{T+1} = \sum_{i=1}^{p} \alpha_i \Delta^d K_{T+1-i} + \sum_{i=1}^{q} \beta_1 \varepsilon_{T+1-i} + \varepsilon_t \quad (15)$$

Then $\Delta^d \hat{\chi}_{T+1}^{\text{emd}}$ is obtained by Eq. (8), and then the inverse d-order difference and inverse EMD are performed on $\Delta^d \hat{\chi}_{T+1}^{\text{emd}}$ to obtain the final prediction matrix $\hat{\chi}_{T+1}$.

The pseudocode of the prediction process is shown in Algorithm 2.

4 Experiment

In this section, our proposed SE-ARIMA model is evaluated on a public dataset. The details of experimental settings and results are presented below.

Algorithm 1 Training of the SE-ARIMA model

```
Input: $\chi \in \mathbf{R}^{M \times N \times T}$, $p, d, q$, and $r$
1: Decompose $\chi$ into three new matrix series $\chi^{\text{emd}}$;
2: for each $\chi^{\text{emd}}$ do
3:   Compute the $d$-order difference for each $\chi^{\text{emd}}$ and obtain
     $\Delta^d \chi_{d+1}^{\text{emd}}, \Delta^d \chi_{d+2}^{\text{emd}}, \ldots, \Delta^d \chi_T^{\text{emd}}$;
4:   Initialize random errors $\epsilon_{t-q}, \epsilon_{t-q+1}, \ldots, \epsilon_{t-1}$;
5:   Initialize $n \times r$ factor matrix $V_{\text{reduce}}$ ;
6:   for $t = p+d+q, p+d+q+1, \ldots, T$ do
7:     Compute the $\Delta^d K_t$ by Eq. (7);
8:     Estimate parameters $\alpha_1, \alpha_2, \ldots, \alpha_p$ and $\beta_1, \beta_2, \ldots, \beta_q$
       of ARIMA by Yule-Walker equations;
9:     Update $\Delta^d K_t$ by Eq. (11);
10:    Compute $M$ and $N$ by Eq. (13);
11:    Update $V_{\text{reduce}} = NM^{\text{T}}$;
12:    for $j = 1, 2, \ldots, q$ do
13:      Update $\epsilon_{t-j}$ by Eq. (14);
14:    end for
15:  end for
16:  Repeat Steps 6–13 until convergence;
17: end for
Output: $V_{\text{reduce}}$ , $\epsilon, \alpha_1, \alpha_2, \ldots, \alpha_p, \beta_1, \beta_2, \ldots, \beta_q$
```

Algorithm 2 Prediction of the SE-ARIMA model

```
Input: $V_{\text{reduce}}$ , $\alpha_1, \alpha_2, \ldots, \alpha_p, \beta_1, \beta_2, \ldots, \beta_q, \epsilon, \Delta^d K_{T-p+1}$,
       $\Delta^d K_{T-p+2}, \ldots, \Delta^d K_T$
1: for each $\Delta^d K$ do
2:   Compute $\Delta^d \hat{K}_{T+1}$ by Eq. (15);
3:   Compute $\Delta^d \hat{\chi}_{T+1}^{\text{emd}}$ by Eq. (8);
4:   Obtain the prediction matrix $\Delta^d \hat{\chi}_{T+1}^{\text{emd}}$ by inverse $d$-order
     difference;
5: end for
6: Obtain the final prediction value $\hat{\chi}_{T+1}$ by summing three
   $\Delta^d \hat{\chi}_{T+1}^{\text{emd}}$
Output: $\hat{\chi}_{T+1}$
```

4.1 Experimental setting

4.1.1 Dataset

We evaluate the SE-ARIMA through experiments on a real-world dataset where the air quality data is from the website of the Beijing Environmental Protection Monitoring Center, including the public dataset of historical air quality of 35 air quality monitoring stations in Beijing. The air quality data includes hourly data on the AQI and six concentrations of $PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3, and CO. We choose hourly data of seven air quality indicators in Beijing in January 2020 in this work. Due to most of the missing data in the Botanical Garden monitoring station, we directly eliminate the data of this monitoring station and select the data of the remaining 34 monitoring stations as the data for the final experiment.

We directly fill the missing data with the average value of the rest of the stations at that moment.

4.1.2 Experimental environment

All experiments are run on a personal computer (Intel i7-10750H CPU, NVIDIA GeForce GTX 1650 and 8 GB RAM). The experimental environment is Windows 10 operating system with Python3.7, and the python extension package for deep learning is installed for comparative experiments.

4.1.3 Evaluation metrics

We adopt the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as evaluation metrics to measure the accuracy of the SE-ARIMA. The formulas for these valuation metrics are as follows:

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

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{17}$$

where $\hat{y}_i$ represents the predicted value, and y_i represents the actual value.

4.1.4 Compared method

To demonstrate the performance of the SE-ARIMA, we compare it with several well-performing prediction methods, including Linear Regression (LR), SVR, LSTM, and ARIMA. The experiments are on conducted on the Beijing air quality dataset.

4.1.5 Parameters setting

We predict the air quality at the moment $T+1$ based on the historical series of last T moments. Therefore, we treat the data of the last time slot as the test sample and the rest of the data as the training samples to compare the proposed SE-ARIMA model with compared methods for the experiment. For the proposed model, we determine the optimal parameters p, q, and d of the three ($p_1 = 0$, $q_1 = 1$, $d_1 = 1$; $p_2 = 1$, $q_2 = 1$, $d_2 = 0$; $p_3 = 1$, $q_3 = 1$, $d_3 = 0$) by the grid search method. During the training of SE-ARIMA, we found that when the number of iterations reached about 20 to 30, as shown in Fig. 4, the algorithm converges and is fully sufficient to reach the best accuracy. Therefore, we set the number of iterations to 30.

4.2 Experimental result

4.2.1 Performance comparison

To demonstrate the performance of our proposed method for multi-step prediction, we conduct air quality prediction for the next 5 hours on the dataset. Table 1 presents the accuracy comparison with other methods over the next 5 hours. The values are mean of the evaluation metrics of the seven air quality indicators. It can be seen that our proposed SE-ARIMA achieves the best result. The RMSE and MAE of our method are reduced by 1.51% to 2.81% and 0.51% to 2.47%

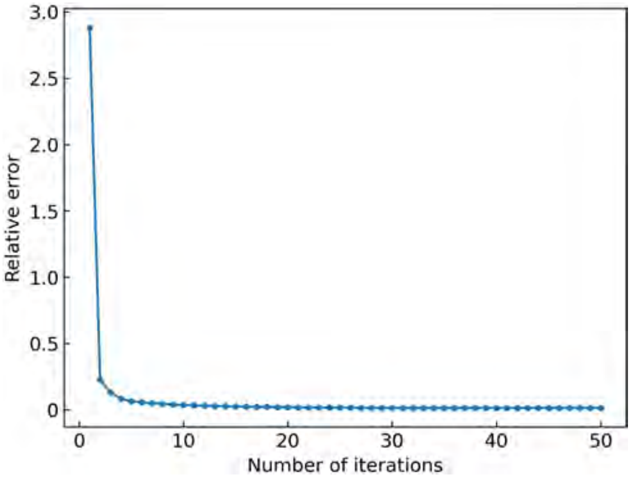


Fig. 4 Convergence curve of the SE-ARIMA.

Table 1 Mean of the evaluation metrics of the seven air quality indicators. The bold fonts indicate the best results.

Hour	Method	RMSE	MAE
1	ARIMA	12.3461	6.4909
	LR	12.395	6.7269
	SVR	13.4959	8.0929
	LSTM	14.8252	7.9143
	Our proposed	**12.1031**	**6.3304**
2	ARIMA	19.1747	10.41
	LR	20.6173	11.2503
	SVR	21.2818	13.2851
	LSTM	22.9538	12.707
	Our proposed	**18.6361**	**10.1965**
3	ARIMA	23.6686	13.4594
	LR	27.4645	14.716
	SVR	26.039	16.879
	LSTM	28.0559	16.295
	Our proposed	**23.0716**	**13.197**
4	ARIMA	27.5181	16.0301
	LR	35.6144	17.578
	SVR	29.6939	19.5749
	LSTM	31.5316	18.9647
	Our proposed	**26.97**	**15.838**
5	ARIMA	30.1616	18.1577
	LR	47.0667	19.8047
	SVR	32.1326	21.5101
	LSTM	33.5604	20.8253
	Our proposed	**29.7071**	**18.0645**

compared to the traditional ARIMA model, respectively. That is because ARIMA can be applied to smoother data after empirical modal decomposition, which is exactly what the ARIMA model needs. In addition, since SVD removes some of the influence of noisy data on the final results and implicitly captures the correlation between air quality at different locations, our model is more accurate compared to the traditional ARIMA. We find that deep learning networks LSTM works poorly instead, probably due to overfitting of the model caused by too many parameters of the deep model. The LR and ARIMA models are superior to the LSTM, which shows that using a linear model to predict is effective if we only focus on the time series itself. The ARIMA model is superior to LR, especially for future multi-step forecasting. This is because the ARIMA model takes into account not only the lag of past time but also the lag of past errors.

To show the prediction performance of our method on different air pollution indicators, we present the evaluation results in Tables 2 and 3. The bold fonts indicate the best results and the underlined fonts indicate the second-best results. The RMSE and MAE of the AQI, $PM_{2.5}$, NO_2, and CO in 1–2 hours predicted by our proposed method in Tables 2 and 3 are the smallest compared to other methods. In the first two hours, the values of the mean RMSE of the predicted AQI, $PM_{2.5}$, NO_2, and CO decrease by 23.91%, 15.30%, 3.22%, and 2.21%, respectively. And the values of the mean MAE of the predicted first two steps decrease by 1.10%, 1.35%, 3.20%, and 2.43%, respectively. The RMSE and MAE of the predicted AQI, $PM_{2.5}$, and CO in the next 3–5 hours are the smallest among all methods. For AQI, $PM_{2.5}$, and CO, the values of the RMSE decrease by 2.55% to 26.92%, 2.77% to 16.52%, and 1.64% to 3.10%, respectively; and the values of the MAE decrease by 0.68% to 1.36%, 1.00% to 1.96%, and 1.82% to 2.56%, respectively. It can be seen that our proposed method can improve the prediction accuracy of AQI, $PM_{2.5}$, and CO. The results of the other four air pollutant predictions can also achieve better results overall. Our method achieves the best results and can be used for long-term and short-term forecasting.

We visualized the results of 300 hours prediction to show intuitively the prediction performance of our method and other methods. The comparison of the

Table 2 RMSE of all methods for multi-step prediction. The bold fonts indicate the best results and the underlined fonts indicate the second best results.

Hour	Method	$PM_{2.5}$(μg/m^3)	PM_{10}(μg/m^3)	AQI	SO_2(μg/m^3)	NO_2(μg/m^3)	O_3(μg/m^3)	CO (mg/m^3)
1	ARIMA	17.9801	28.0261	20.8667	2.0831	8.1348	9.1305	0.2013
	LR	18.5244	28.1723	21.1453	**2.0220**	7.9285	**8.7702**	0.2021
	SVR	21.0023	29.7324	23.9495	2.1445	8.2905	9.1322	0.2196
	LSTM	22.7343	32.1978	25.3295	2.4312	9.9765	10.8652	0.2416
	Our proposed	**17.4203**	**27.8007**	**20.4827**	2.0397	**7.9177**	8.8627	**0.1980**
2	ARIMA	29.2613	40.3637	32.9650	3.1899	12.9545	15.1721	0.3162
	LR	34.2837	**44.0240**	35.7563	3.1226	12.5070	14.2969	0.3309
	SVR	34.9534	43.2421	38.8637	3.3318	13.1506	15.0758	0.3553
	LSTM	37.5642	44.9158	41.0029	3.5588	15.3340	17.9143	0.3868
	Our proposed	**28.3277**	40.1002	**31.9311**	**3.0960**	**12.4656**	**14.2250**	**0.3074**
3	ARIMA	36.8245	**46.9124**	41.2233	4.1229	16.1173	20.0899	0.3901
	LR	49.1285	58.0553	46.9464	**3.9346**	**15.2867**	**18.4847**	0.4156
	SVR	43.6182	50.4531	48.3095	4.1365	16.0074	19.3078	0.4406
	LSTM	47.2025	51.0306	51.2626	4.3516	18.4955	23.5463	0.5022
	Our proposed	**35.8838**	47.0265	**39.9473**	4.0055	15.5212	18.7390	**0.3780**
4	ARIMA	42.6732	**53.0557**	48.3556	4.8044	18.9295	24.3482	0.4600
	LR	69.2486	77.8622	57.9428	**4.4859**	**17.5815**	**21.6890**	0.4911
	SVR	50.1232	56.5775	55.3214	4.6828	18.1967	22.4479	0.5076
	LSTM	52.4083	56.0888	58.0214	4.8602	21.0579	27.6842	0.6006
	Our proposed	**41.4902**	53.9001	**47.0754**	4.6918	18.2782	22.9077	**0.4469**
5	ARIMA	47.3816	**55.5670**	53.7903	5.3511	20.8225	27.7085	0.5099
	LR	103.0363	106.4252	71.8711	**4.8508**	**19.0100**	**23.7227**	0.5511
	SVR	54.8098	60.1122	60.4239	5.0396	19.5028	24.4834	0.5565
	LSTM	56.9858	56.2906	62.8355	5.2086	22.7708	30.1810	0.6506
	Our proposed	**46.0594**	56.9299	**52.4169**	5.2905	20.2388	26.5185	**0.4958**

Table 3 MAE of all methods for multi-step prediction. The bold fonts indicate the best results and the underlined fonts indicate the second best results.

Hour	Method	$PM_{2.5}$ (μg/m^3)	PM_{10} (μg/m^3)	AQI	SO_2 (μg/m^3)	NO_2 (μg/m^3)	O_3 (μg/m^3)	CO (mg/m^3)
1	ARIMA	8.7398	13.3299	10.6256	1.2260	5.3372	6.0625	0.1156
	LR	9.2621	13.8234	11.2432	1.2254	5.3705	6.0414	0.1224
	SVR	12.2349	15.7411	14.3913	1.4039	5.9466	6.7884	0.1441
	LSTM	11.1223	15.3473	13.5206	1.3971	6.5273	7.3383	0.1474
	Our proposed	**8.6523**	**12.7633**	**10.5377**	**1.1874**	**5.1943**	**5.8650**	**0.1127**
2	ARIMA	14.3211	**19.7879**	17.2944	1.9332	8.8220	10.5235	0.1881
	LR	16.5357	20.9746	19.6220	1.9376	8.9387	10.5360	0.2073
	SVR	21.0071	23.5963	24.2764	2.2655	9.8269	11.7838	0.2395
	LSTM	18.7657	22.4316	22.4353	2.1070	10.5166	12.4519	0.2413
	Our proposed	**14.0823**	19.8580	**17.0595**	**1.8569**	**8.4946**	**9.8407**	**0.1837**
3	ARIMA	18.9595	**23.8095**	22.7418	2.5591	11.3678	14.5359	0.2421
	LR	22.7215	25.3926	26.5730	2.4847	11.3435	14.2243	0.2727
	SVR	27.4967	28.3180	31.4548	2.8843	12.2724	15.4199	0.3068
	LSTM	24.9511	26.6958	29.1489	2.6500	13.2557	17.0476	0.3160
	Our proposed	**18.5871**	24.1635	**22.4637**	**2.4590**	**10.9329**	**13.5367**	**0.2359**
4	ARIMA	22.8060	**27.0906**	27.3732	3.0359	13.6636	17.9510	0.2906
	LR	28.0013	29.0276	32.4126	**2.8680**	13.2924	17.1170	0.3273
	SVR	32.3420	31.9373	36.8035	3.3280	14.1319	18.1234	0.3587
	LSTM	29.3591	30.0074	34.0485	2.9882	15.3667	20.6055	0.3777
	Our proposed	**22.4007**	28.0530	**27.1867**	2.9234	**13.1695**	**16.8486**	**0.2842**
5	ARIMA	26.2461	**29.5115**	31.3395	3.4639	15.3489	20.8648	0.3295
	LR	32.8308	31.4995	37.2981	**3.1282**	**14.4666**	**19.0374**	0.3722
	SVR	36.1365	34.3636	40.9648	3.6256	15.2656	19.8140	0.4004
	LSTM	33.0185	31.3311	37.8551	3.2674	16.8357	23.0468	0.4224
	Our proposed	**25.7691**	31.1480	**31.1244**	3.3674	14.8570	19.8619	**0.3235**

predicted and actual values of our proposed method with compared methods is shown in Fig. 5, from which we can see that the predictions of SE-ARIMA are closer to the curve of the actual values. And the results of our proposed SE-ARIMA are better suited to the actual value in predicting peaks and troughs. When there are small fluctuations in the data, the predicted values of ARIMA have large biases with the actual values; LR, SVR, and LSTM have much higher predicted values than the actual values when there are waves in the true values. Our proposed method has a similar error in predicting the crest, but such error is smaller than the compared methods.

Figure 6 shows a box plot of the absolute errors of our model and the compared methods. The error of each method is the sum of the absolute errors of the predicted and actual values of the seven air quality indicators from the 34 stations. In Fig. 6, it is clear to observe that our proposed method has the smallest absolute error among all the methods and its distribution is in a small range. In addition, the box plot shows that the outliers of the absolute errors predicted by the method are smaller compared to the baseline, and the prediction results of our proposed method can reflect the true situation more accurately.

We display the time costs required for our method and the compared methods in Fig. 7. The time of each method is the average cost in conducting 300 predictions of seven air quality indicators in 34 monitoring stations. Because the time of LSTM is too large, we truncate it in Fig. 7. The computational time of the proposed method is the lowest among all the compared methods, only 60.20% of the LR, and 43.66% less than the conventional ARIMA.

In summary, our method can effectively improve the overall accuracy of prediction and reduce the computational time cost.

4.2.2 Ablation study

To demonstrate that each component in our proposed method is effective, we perform ablation experiments by removing a single key part of SE-ARIMA that have the following three parts: (1) Proposal-1: The EMD part for stationary data and the truncated SVD part for denoising data are removed; (2) Proposal-2: The EMD

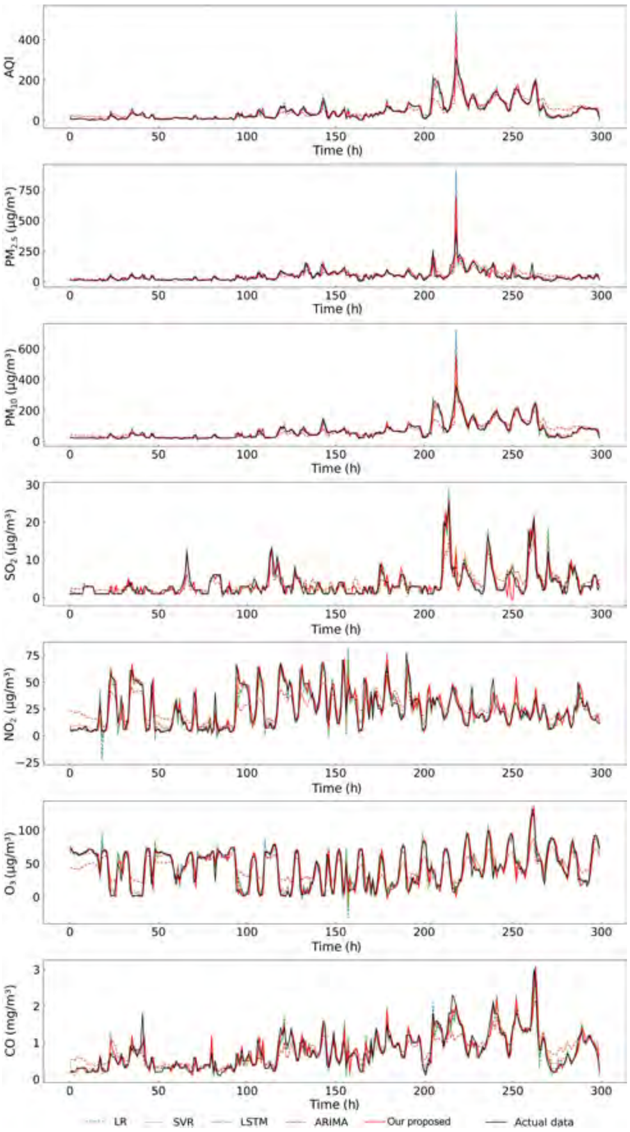


Fig. 5 Predicted and actual values of different methods.

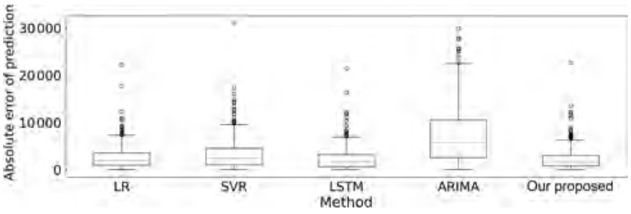


Fig. 6 Box plot of absolute errors of the predicted values of different methods.

part is removed; (3) Proposal-3: The truncated SVD part is removed.

The proposed method is compared with the above three cases with all parameter settings being the same. The RMSE and MAE of the prediction results on the dataset of 34 monitoring stations in Beijing are given in Table 4, where the bold font indicates the best result under these evaluation metrics.

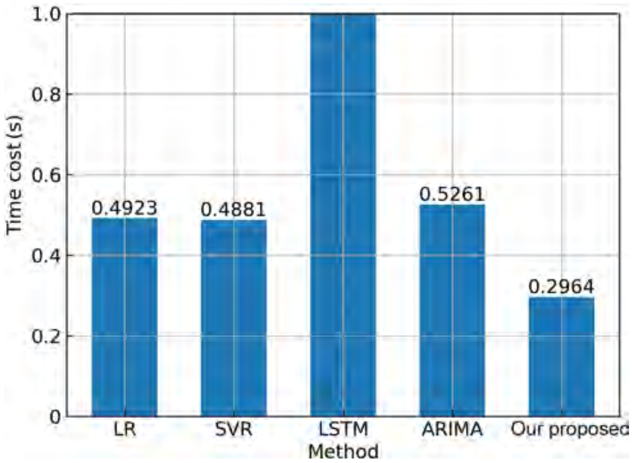


Fig. 7 Time costs of different methods for calculating.

With the EMD part removed, the RMSE of PM$_{2.5}$, PM$_{10}$, and O$_3$ increase by 2.31% , 2.80% and 3.25%, respectively. And the MAE of PM$_{10}$ and O$_3$ increase by 1.55% and 3.05% , respectively. With the SVD part removed, the RMSE of PM$_{2.5}$, PM$_{10}$, and O$_3$ increase by 2.77% , 3.16% and 3.62%, respectively. And the MAE of PM$_{10}$ and O$_3$ increase by 1.80% and 3.40%, respectively. The results of other indicators also show a slight increase. Without the EMD component, the prediction errors increase slightly, because the data are not stable enough. The performance reduces more without the truncated SVD part, which indicates that the truncated SVD can lead to a greater improvement in the performance of the mode. So, it is clear that each part of the SE-ARIMA can lead to improving the performance of overall prediction.

5 Conclusion

In this study, we propose a hybrid model to predict the air quality indicators from multiple monitoring stations in next hours. The proposed model integrated the extended ARIMA model with the EMD method and truncated SVD to improve air quality prediction accuracy. In detail, EMD is used to decompose the original non-stationary air quality indicator series into smooth sub series. Moreover, the traditional ARIMA model is extended to predict all air quality indicators from multiple monitoring stations simultaneously. Experimental results demonstrate our model outperforms state-of-art air quality prediction models in both accuracy and time cost.

However, there are still limitations in this study. The main limitation is that we only use historical air

Table 4 Evaluation results of the SE-ARIMA and the method with the components removed. The bold fonts indicate the best results.

Indicator	Method	$PM_{2.5}$ (μg/m^3)	PM_{10} (μg/m^3)	AQI	SO_2 (μg/m^3)	NO_2 (μg/m^3)	O_3 (μg/m^3)	CO (mg/m^3)
RMSE	Our proposed-1	17.1476	27.4165	19.7113	1.8989	8.0836	9.0955	0.1970
	Our proposed-2	17.0250	27.3826	19.5847	1.8826	8.0288	8.9977	0.1956
	Our proposed-3	17.1018	27.4786	19.6527	1.8952	8.0540	9.0298	0.1965
	Our proposed	**16.6414**	**26.6356**	**19.2348**	**1.8704**	**7.9106**	**8.7142**	**0.1905**
MAE	Our proposed-1	8.6046	13.3080	10.2975	1.1491	5.3972	6.1302	0.1129
	Our proposed-2	8.5339	13.2968	10.2226	1.1436	5.3566	6.0600	0.1123
	Our proposed-3	8.5844	13.3295	10.2698	1.1483	5.3743	6.0804	0.1127
	Our proposed	**8.5047**	**13.0944**	**10.2162**	**1.1397**	**5.3046**	**5.8807**	**0.1114**

quality indicators to make prediction. Air quality is determined by many factors, such as weather conditions and transportation, which are not considered in our study. In the future, we will make full use of those information to achieve higher accuracy[39]. Furthermore, we will also study more models, such as deep models[40] to improve our proposed model, analyze the causal relationships between air quality and various influencing factors through the correlation graph[41], and make further corrections to the predicted values through anomaly detection[42, 43]. In practical applications, monitoring stations are distributed in different locations, we can consider IoT and edge computing technologies to reduce the delay of data transmission for real-time prediction of different stations[44, 45].

References

[1] Y. Zeng, J. Chen, N. Jin, X. Jin, and Y. Du, Air quality forecasting with hybrid LSTM and extended stationary wavelet transform, *Build. Environ.*, vol. 213, p. 108822, 2022.

[2] World Health Organization, WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, World Health Organization, https://www.who.int/publications/i/item/9789240034433, 2021.

[3] P. D. Waggoner, Pandemic policymaking, *J. Soc. Comput.*, vol. 2, no. 1, pp. 14–26, 2021

[4] F. Fourati and M. S. Alouini, Artificial intelligence for satellite communication: A review, *Intell. Conver. Netw.*, vol. 2, no. 3, pp. 213–243, 2021.

[5] J. Evans, Social computing unhinged, *J. Soc. Comput.*, vol. 1, no. 1, pp. 1–13, 2020.

[6] C. Hu, W. Fan, E. Zeng, Z. Hang, F. Wang, L. Qi, and M. Z. A. Bhuiyan, Digital twin-assisted real-time traffic data prediction method for 5G-enabled internet of vehicles, *IEEE Trans. Ind. Inform.*, vol. 18, no. 4, pp. 2811–2819, 2022.

[7] J. Tie, X. Lei, and Y. Pan, Metabolite-disease association prediction algorithm combining DeepWalk and random forest, *Tsinghua Science and Technology*, vol. 27, no. 1, pp. 58–67, 2022.

[8] Y. Liu, Z. Song, X. Xu, W. Rafique, X. Zhang, J. Shen, M. R. Khosravi, and L. Qi, Bidirectional GRU networks-based next POI category prediction for healthcare, *Int. J. Intell. Syst.*, vol. 37, no. 7, pp. 4020–4040, 2022.

[9] S. Zhang, H. Liu, J. He, S. Han, and X. Du, Deep sequential model for anchor recommendation on live streaming platforms, *Big Data Min. Anal.*, vol. 4, no. 3, pp. 173–182, 2021.

[10] X. Xu, Q. Jiang, P. Zhang, X. Cao, M. R. Khosravi, L. T. Alex, L. Qi, and W. Dou, Game theory for distributed IoV task offloading with fuzzy neural network in edge computing, *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 11, pp. 4593–4604, 2022.

[11] A. Agarwal, S. Sharma, V. Kumar, and M. Kaur, Effect of e-learning on public health and environment during COVID-19 lockdown, *Big Data Min. Anal.*, vol. 4 no. 2, pp. 104–115, 2021.

[12] C. Catlett, P. Beckman, N. Ferrier, H. Nusbaum, M. E. Papka, M. G. Berman, and R. Sankaran, Measuring cities with software-defined sensors, *J. Soc. Comput.*, vol. 1, no. 1, pp. 14–27, 2020.

[13] N. Ji, L. Ma, H. Dong, and X. Zhang, EEG signals feature extraction based on DWT and EMD combined with approximate entropy, *Brain Sci.*, vol. 9, no. 8, p. 201, 2019.

[14] C. Yan, Y. Zhang, W. Zhong, C. Zhang, and B. Xin, A truncated SVD-based ARIMA model for multiple QoS prediction in mobile edge computing, *Tsinghua Science and Technology*, vol. 27, no. 2, pp. 315–324, 2022.

[15] A. K. Sandhu, Big data with cloud computing: Discussions and challenges, *Big Data Min. Anal.*, vol. 5, no. 1, pp. 32–40, 2022.

[16] T. Li, C. Li, J. Luo, and L. Song, Wireless recommendations for internet of vehicles: Recent advances, challenges, and opportunities, *Intell. Conver. Netw.*, vol. 1, no. 1, pp. 1–17, 2020.

[17] M. A. Bouras, F. Farha, and H. Ning, Convergence of computing, communication, and caching in internet of things, *Intell. Conver. Netw.*, vol. 1, no. 1, pp. 18–36, 2020.

[18] Y. Zhang, H. Zhang, J. Cosmas, N. Jawad, K. Ali, B. Meunier, A. Kapovits, L. K. Huang, W. Li, L. Shi, et al., Internet of radio and light: 5g building network radio and edge architecture, *Intell. Conver. Netw.*, vol. 1, no. 1, pp. 37–57, 2020.

[19] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, Prediction of COVID-19 confirmed, death, and cured cases

in India using random forest model, *Big Data Mining and Analytics*, vol. 4, no. 2, pp. 116–123, 2021.

[20] H. Zhu and J. Hu, Air quality forecasting using SVR with quasi-linear kernel, in *Proc. Int. Conf. Computer, Information and Telecommunication Systems* (*CITS*), Beijing, China, 2019, pp. 1–5.

[21] Y. T. Tsai, Y. R. Zeng, and Y. S. Chang, Air pollution forecasting using RNN with LSTM, in *Proc. IEEE 16th Int. Conf. Dependable, Autonomic and Secure Computing, 16th Int. Conf. Pervasive Intelligence and Computing, 4th Int. Conf. Big Data Intelligence and Computing and Cyber Science and Technology Congress*, Athens, Greece, 2018, pp. 1074–1079.

[22] X. Zhang, X. Rui, X. Xia, X. Bai, W. Yin, and J. Dong, A hybrid model for short-term air pollutant concentration forecasting, in *Proc. IEEE Int. Conf. Service Operations and Logistics, and Informatics* (*SOLI*), Yasmine Hammamet, Tunisia, 2015, pp. 171–175.

[23] P. Wang, H. Zhang, Z. Qin, and G. Zhang, A novel hybrid-Garch model based on ARIMA and SVM for $PM_{2.5}$ concentrations forecasting, *Atmos. Pollut. Res.*, vol. 8, no. 5, pp. 850–860, 2017.

[24] C. Ding, J. Duan, Y. Zhang, X. Wu, and G. Yu, Using an ARIMA-GARCH modeling approach to improve subway short-term ridership forecasting accounting for dynamic volatility, *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 4, pp. 1054–1064, 2018.

[25] S. Moisan, R. Herrera, and A. Clements, A dynamic multiple equation approach for forecasting $PM_{2.5}$ pollution in Santiago, Chile, *Int. J. Forecast.*, vol. 34, no. 4, pp. 566–581, 2018.

[26] S. Du, T. Li, Y. Yang, and S. J. Horng, Deep air quality forecasting using hybrid deep learning framework, *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2412–2424, 2021.

[27] Y. Qi, Q. Li, H. Karimian, and D. Liu, A hybrid model for spatiotemporal forecasting of $PM_{2.5}$ based on graph convolutional neural network and long short-term memory, *Sci. Total Environ.*, vol. 664, pp. 1–10, 2019.

[28] X. Wang, Y. Zhou, and C. Zhao, Heart-rate analysis of healthy and insomnia groups with detrended fractal dimension feature in edge, *Tsinghua Science and Technology*, vol. 27, no. 2, pp. 325–332, 2022.

[29] S. Fan, D. Hao, Y. Feng, K. W. Xia, and W. Yang, A hybrid model for air quality prediction based on data decomposition, *Information*, vol. 12, no. 5, p. 210, 2021.

[30] X. B. Jin, N. X. Yang, X. Wang, Y. Bai, T. Su, and J. Kong, Integrated predictor based on decomposition mechanism for PM2.5 long-term prediction, *Appl. Sci.*, vol. 9, no. 21, p. 4533, 2019.

[31] A. Altıntaş and L. Davidson, EMD-SVR: A hybrid machine learning method to improve the forecasting accuracy of highway tollgates traveling time to improve the road safety, in *Proc. 4th Int. Conf. Intelligent Transport Systems, from Research and Development to the Market Uptake*, Virtual Event, 2021, pp. 241–251.

[32] H. Zheng, J. Yuan, and L. Chen, Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation, *Energies*, vol. 10, no. 8, p. 1168, 2017.

[33] X. B. Jin, N. X. Yang, X. Wang, Y. Bai, T. Su, and J. Kong, Deep hybrid model based on EMD with classification by frequency characteristics for long-term air quality prediction, *Mathematics*, vol. 8, no. 2, p. 214, 2020.

[34] G. Huang, X. Li, B. Zhang, and J. Ren, PM2.5 concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition, *Sci. Total Environ.*, vol. 768, p. 144516, 2021.

[35] Z. Y. Wang, J. Qiu, and F. Li, Hybrid models combining EMD/EEMD and ARIMA for long-term streamflow forecasting, *Water*, vol. 10, no. 7, p. 853, 2018.

[36] N. Fatema, H. Malik, and M. S. Abd Halim, Hybrid approach combining EMD, ARIMA and monte carlo for multi-step ahead medical tourism forecasting, *J. Intell. Fuzzy Syst.*, vol. 42, no. 2, pp. 1235–1251, 2022.

[37] G. E. P. Box and G. M. Jenkins, Truncated SVD is adopted to capture correlations among air pollutants and neighbor stations. *J. Time Ser. Anal.*, vol. 40, no. 5, pp. 970–971, 1970.

[38] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. Yen, C. C. Tung, and H. H. Liu, The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis, *Proc. Roy. Soc. A: Math. Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903– 995, 1998.

[39] F. Wang, G. Li, Y. Wang, W. Rafique, M. R. Khosravi, G. Liu, Y. Liu, and L. Qi, Privacy-aware traffic flow prediction based on multi-party sensor data with zero trust in smart city, *ACM Trans. Internet Technol.*, doi: 10.1145/3511904.

[40] Y. Ma, H. Sun, Y. Chen, J. Zhang, Y. Xu, X. Wang, and P. Hui, Deep-predict: A zone preference prediction system for online lodging platforms, *J. Soc. Comput.*, vol. 2, no. 1, pp. 52–70, 2021.

[41] L. Qi, W. Lin, X. Zhang, W. Dou, X. Xu, and J. Chen, A correlation graph based approach for personalized and compatible web APIs recommendation in mobile APP development, *IEEE Trans. Knowl. Data Eng.*, doi: 10.1109/TKDE.2022.3168611.

[42] Y. Yang, X. Yang, M. Heidari, M. A. Khan, G. Srivastava, M. Khosravi, and L. Qi, ASTREAM: Data-stream-driven scalable anomaly detection with accuracy guarantee in IIoT environment, *IEEE Trans. Netw. Sci. Eng.*, doi: 10.1109/TNSE.2022.3157730.

[43] L. Qi, Y. Yang, X. Zhou, W. Rafique, and J. Ma, Fast anomaly identification based on multiaspect data streams for intelligent intrusion detection toward secure industry 4.0, *IEEE Trans. Ind. Inform.*, vol. 18, no. 9, pp. 6503–6511, 2022.

[44] X. Xu, H. Tian, X. Zhang, L. Qi, Q. He, and W. Dou, DisCOV: Distributed COVID-19 detection on X-ray images with edge-cloud collaboration, *IEEE Trans. Serv. Comput.*, vol. 15, no. 3, pp. 1206–1219, 2022.

[45] J. Ren, J. Li, H. Liu, and T. Qin, Task offloading strategy with emergency handling and blockchain security in SDN-empowered and fog-assisted healthcare IoT, *Tsinghua Science and Technology*, vol. 27, no. 4, pp. 760–776, 2022.

Yuxuan Cao received the BEng degree from Nantong University, China in 2019. He is currently a master student at Qufu Normal University, China. His research interests include time series prediction and data mining.

Difei Zhang is an undergraduate student at the School of Mathematical Sciences, Qufu Normal University, China. His research interests include data science, statistics, and computer science.

Shaoqi Ding received the BEng degree from Qufu Normal University, China in 2020. He is currently a master student at Qufu Normal University, China. His research interests include recommender system and service computing.

Weiyi Zhong received the MEng degree from Qufu Normal University, China in 2021. She is currently a PhD candidate at Qufu Normal University, China. Her research interests include recommender systems and services computing.

Chao Yan received the MEng degree from Chinese Academy of Sciences, China in 2006. He is currently an associate professor at Qufu Normal University, China. He is also a PhD candidate at Shandong University of Science and Technology, Qingdao, China. His research interests include recommender system and service computing.