Stochastic Environmental Research and Risk Assessment (2024) 38:2095–2117
https://doi.org/10.1007/s00477-024-02693-4

REVIEW PAPER

Data analysis and preprocessing techniques for air quality prediction: a survey

Chengqing Yu[1] · Jing Tan[3] · Yihan Cheng[2] · Xiwei Mi[2]

Accepted: 22 February 2024 / Published online: 18 March 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Air quality prediction technology can provide effective technical means for environmental governance. In recent years, due to the strong nonlinearity of data, there has been extensive research on data analysis and preprocessing techniques. This paper aims to comprehensively summarize and analyze the methods used in air quality forecasting, specifically focusing on four categories: data decomposition, dimensionality reduction, data correction, and spatial interpolation. Each method's purpose, characteristics, improvements, and implementation details are described in detail. The evaluation of data preprocessing methods is based on popularity, accuracy improvements, time consumption, maturity, and implementation difficulty. Among the existing methods, data decomposition and feature selection are commonly used and well-developed. However, outlier detection and spatial interpolation have limited applications and require further research. Furthermore, this paper discusses current challenges in applying these methods and future development trends, providing a valuable reference for future research.

Keywords Air quality forecasting · Data preprocessing · Decomposition · Feature selection

1 Introduction

The rise of industrialization has drawn attention to the impact of increasing emissions and air pollution. Air pollution is characterized by the presence of harmful substances in the atmosphere, including nitric oxide (NO), nitrogen dioxide (NO_2), carbon monoxide (CO), particulate matter (PM), and ozone (O_3). When these substances surpass prescribed normal levels, they can have detrimental effects on both the ecological environment and human health (Dincer and Akkuş 2018). Prolonged exposure to severe air pollution can lead to adverse health effects, resulting in higher morbidity and mortality rates. Additionally, various air pollutants can cause environmental damage, including the greenhouse effect, ozone depletion, haze, acid rain, and other significant environmental disasters (Najjar 2011). Therefore, it is crucial to forecast air quality in advance and provide accurate and effective air quality data for air pollution management and early warning systems.

1.1 Overview of air quality forecasting models

In recent years, there has been increasing attention towards the analysis and prediction of air pollutant concentration, leading to the gradual development of air quality forecasting models. Based on different forecasting approaches, the current mainstream air quality forecasting models can be categorized into three groups: physical models, statistical models, and intelligent models (Zhu et al. 2018b, c).

Physical models are mathematical models used in physics and chemistry to predict the change in pollutant concentrations. These models primarily examine the sources and transport processes of air pollutants (Karimian et al. 2023). Statistical models analyze the statistical relationships between pollutant concentrations and historical meteorological data. Statistical models are easier to calculate and generally offer better forecasting performance compared to physical models (Zhou et al. 2023). However, air pollution

Chengqing Yu and Jing Tan have contributed equally to this work.

✉ Xiwei Mi
mixiwei@bjtu.edu.cn

[1] University of Chinese Academy of Sciences, Beijing, China

[2] School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

[3] School of Traffic and Transportation Engineering, Central South University, Changsha 410075, China

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

data is often disordered and unstable, posing a challenge for traditional statistical models that cannot effectively handle nonlinear data (Fang et al. 2022). Intelligent models are also increasingly used to predict air quality, such as the artificial neural network (ANN) (Jiang et al. 2004), the support vector machine (SVM) (Wang et al. 2017d), and the fuzzy logic model (FLM) (Yildirim and Bayramoglu 2006). These models exhibit strong robustness and high forecasting accuracy, making them superior to physical models and statistical models in terms of development potential. However, it is important to note that single intelligent models still have certain drawbacks, including overfitting, slow convergence speed, and susceptibility to local optimal states.

Hybrid models are widely used in air quality forecasting as they overcome the limitations of individual models. These models combine the advantages of multiple forecasting models, resulting in improved forecasting performance. The key components of hybrid models include data preprocessing methods, forecasting models, optimization algorithms, ensemble methods, and data post-processing methods (Li et al. 2023).

1.2 Motivation of the review

Data analysis and pre-processing methods, as important components of hybrid models, are playing an increasingly important role and are a collective term for a series of data operations that can improve data quality (Liu et al. 2018). Specifically, data pre-processing methods can eliminate redundant information in the raw data, reduce the negative impact of outliers, filter noise in the data, and help predictive models achieve better prediction results (Wang et al. 2022).

Considering the importance of air quality, there have been some review papers on air quality forecasting. (Bai et al. 2018) divided the air pollution forecasting methods into three classical categories and they comprehensively reviewed their theory and application of these forecasting methods. Rybarczyk and Zalakeviciute (2018) summarized the machine learning algorithms for outdoor air pollution modeling and described the principles of these algorithms. Casazza et al. (2019) reviewed the development of three-dimensional air quality monitoring and modeling for applications in the management and planning of urban ports. Byun and Schere (2006) reviewed the governing equations, computational algorithms, and other components of the CMAQ modeling system. Gulia et al. (2015) reviewed the urban air quality management, focusing on the monitoring and modeling methods. Daly and Zannetti (2007) briefly reviewed air pollution modeling techniques, mainly computer methods for simulating air quality processes.

Based on the literature research, it can be found that most reviews focus on the forecasting model itself and there is currently no review on data preprocessing methods in air quality forecasting (Shi et al. 2024). Given the importance of data preprocessing methods, it is necessary to review data preprocessing methods for air quality forecasting (Ojagh et al. 2021).

This paper classifies and summarizes data preprocessing methods used in air quality forecasting, to select the most appropriate data preprocessing method for the corresponding air quality forecasting model in the future. Besides, this paper allows researchers to pay more attention to the less-used data preprocessing methods. By sorting out and summarizing the models and applications, we divided data preprocessing methods in air quality forecasting into four categories: (a) data decomposition, (b) dimensionality reduction, (c) data correction, and (d) spatial interpolation. Each of these models has its own distinct characteristics and is used in air quality forecasting for various purposes. Data decomposition decomposes the original air quality data into several separate sub-sequences (Liang et al. 2020). Feature extraction and feature selection are employed to reduce dimension (Kristiani et al. 2021). Missing data processing is used to deal with the lost data. Outlier detection is used to correct the input dataset (Kang et al. 2023). Spatial interpolation can be used to solve problems related to spatial pollutant concentration, which is spatial reasoning about the same moment (Prihatno et al. 2021). Figure 1 illustrates the overall application of these data preprocessing methods in air quality prediction.

1.3 Organization of the paper

The paper is organized as follows: Sect. 2 classifies and introduces data decomposition methods. Section 3 provides a summary of the dimensionality reduction technology used in data preprocessing, which includes feature extraction and feature selection. Section 4 reviews data preprocessing methods for correcting the input dataset, such as missing data processing and outlier detection. Section 5 presents a general introduction to spatial interpolation methods. A comprehensive discussion of these data preprocessing methods is presented in Sect. 6. Finally, the paper is summarized in Sect. 7.

2 Data decomposition

Because of the combined influence of natural fluctuations and human activities, AQI data usually display a consistent overall pattern with occasional local disturbances over time. Hence, if the model can accurately identify these characteristics in the data, the prediction performance will be significantly enhanced (Jin et al. 2022). In the published literature on air quality forecasting, the first step of data preprocessing is usually data decomposition (Yang et al. 2023b). As illustrated in Fig. 2, most air quality forecasting models that

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

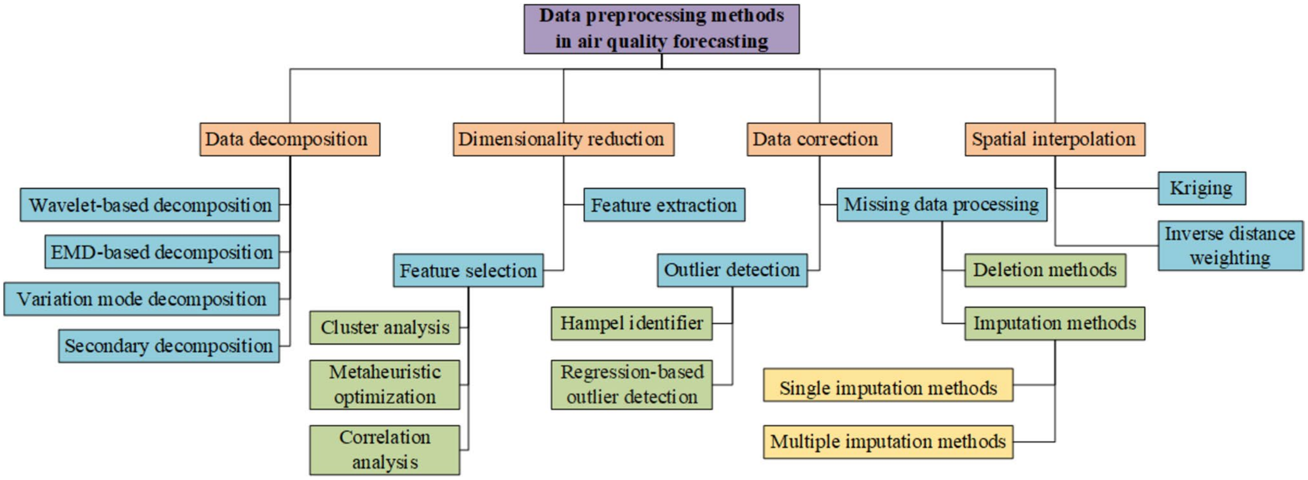


Fig. 1 The overall framework of the application of data preprocessing methods in air quality forecasting

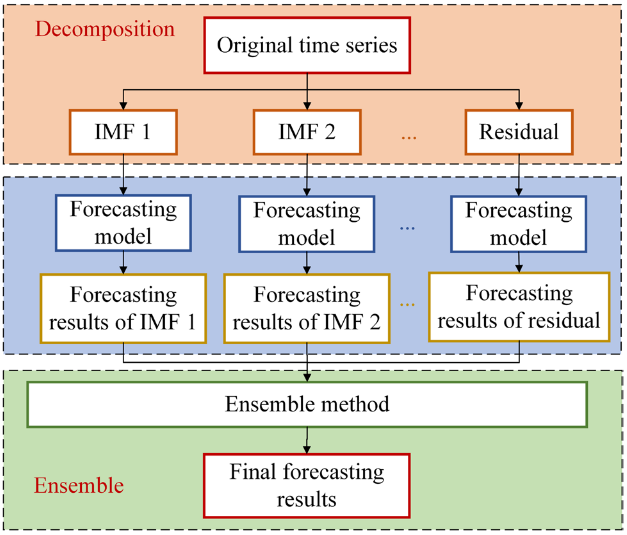


Fig. 2 The “decomposition-ensemble” framework of air quality forecasting

utilize data decomposition follow the 'decomposition-ensemble' framework (Yuan et al. 2023). The data decomposition method essentially involves breaking down the unstable and disordered original series into a specific number of stationary and ordered sub-sequences (Yu et al. 2024). A forecasting model is established for each sub-sequence separately, and relatively independent forecasting results are obtained. The final forecasting result is obtained by combining these individual forecasting results (Mi et al. 2022).

After conducting a thorough review of the existing literature on air quality forecasting, the data decomposition methods employed can be categorized into four main categories: (a) wavelet-based decomposition, (b) EMD-based decomposition, (c) Variational mode decomposition, and (d) secondary decomposition.

2.1 Wavelet-based decomposition

The wavelet-based decomposition can decompose non-stationary air quality time series into two components: the approximate component representing low-frequency features and the detail component representing high-frequency features. This decomposition process helps in obtaining stationary and regular signals (Sun et al. 2013). The choice of decomposition level is crucial. Selecting a higher decomposition level results in smoother and more sensitive components. Nevertheless, errors can arise during the decomposition process, and higher decomposition levels can lead to increased accumulated errors (Chen et al. 2013). The models based on wavelet-based decomposition have good localization characteristics in the time and frequency domain and can dig out the hidden information as much as possible. Cheng et al. (2019) combined ANN, ARIMA, and SVM with the wavelet decomposition (WD) respectively when forecasting $PM_{2.5}$ concentration. The experimental results showed that the hybrid models with wavelet-based decomposition had better forecasting performance than the individual models.

The wavelet transform (WT) is the most basic form of wavelet-based decomposition, which is developed from the Fourier transform (FT). The WT can be divided into the continuous wavelet transform (CWT) and the discrete wavelet transform (DWT). Compared with CWT, DWT takes less time to compute and has fewer implementation requirements. Therefore, DWT is more widely used than CWT (Mittal and Bhardwaj 2011; Osowski and Garanty 2007; Siwek and Osowski 2012). However, it is important to note that DWT also has some limitations. One of the main limitations is that

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

DWT can generate different power spectrums depending on the starting points. This inconsistency in the transformed coefficients estimated by DWT can lead to instability in the analysis (Percival and Walden 2000). To overcome these limitations, (Prakash et al. 2011) applied a method called the maximum overlap wavelet transform (MODWT) to air quality forecasting in the urban atmosphere. The stationary wavelet transform (SWT) is also used in air quality forecasting (Bai et al. 2016; Li and Tao 2017). Compared with WT, SWT has translation invariance, which means the length of the wavelet coefficient is equal to the length of the original data. SWT is more advantageous than DWT. Because the approximation coefficients and detail coefficients obtained by SWT do not need to be down-sampled but are sampled at each level of decomposition without loss of information (Bai et al. 2016).

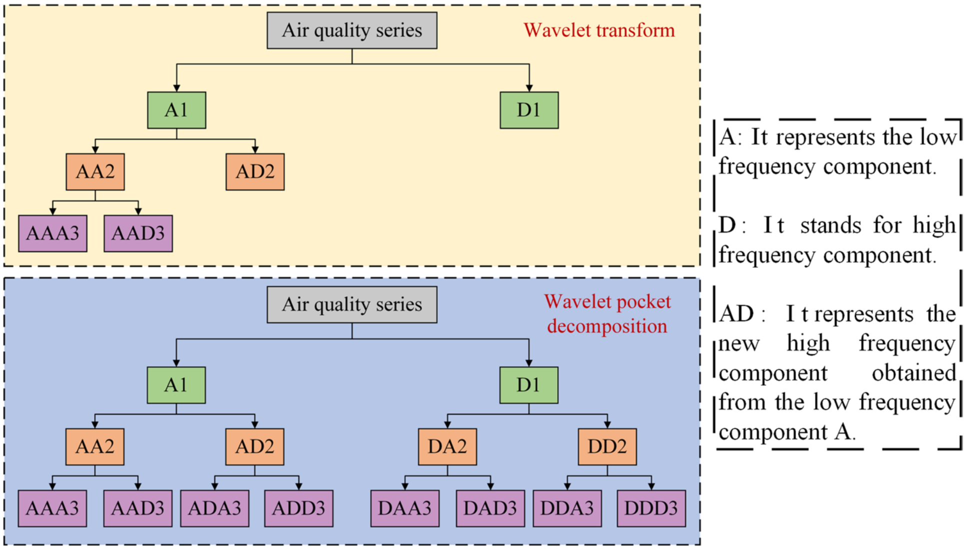
WT can effectively decompose the low-frequency series of each layer. However, in many practical applications, WT becomes necessary to decompose the intermediate and high-frequency series in order to extract more useful information (Chen et al. 2020). To address this issue, wavelet packet decomposition (WPD) has been developed. WPD offers enhanced forecasting performance by disassembling and analyzing the appropriate coefficients and detailed coefficients, thus providing more comprehensive and insightful information as compared to WT (Sun and Li 2020). The comparison of structures between WT and WPD in the decomposition process is illustrated in Fig. 3. as a simplified binary tree with three layers. The selection of the mother wave and the number of layers decomposed is a crucial factor in WPD (Wang et al. 2022).

The wavelet-based decomposition methods require predetermined parameters, such as the number of layers decomposed in WPD (Guo et al. 2023). The empirical wavelet transform (EWT) is an effective time series decomposition method, which can adjust parameters adaptively (Dabin et al. 2022). According to the spectrum of the signal, it can automatically divide the frequency band. Considering the excellent performance of EWT, many researchers have applied EWT to more complex and demanding air quality forecasting (Huang et al. 2023). For example, c used the EWT to further implement data preprocessing and improve the limitation of the classic decomposition method. Compared with the classical wavelet transform method, EWT fully realizes the adaptive data decomposition, and guarantees the performance of the model (Nuhu et al. 2023).

2.2 EMD-based decomposition

Empirical mode decomposition-based decomposition methods are a series of time series preprocessing methods that are based on modal analysis. These methods provide more robust processing and extraction capabilities when dealing with non-linear and non-stationary data compared to other methods, such as wavelet-based decomposition and Fourier decomposition (Yuan et al. 2019). Wavelet-based decomposition methods heavily depend on the selection of wavelet function and decomposition level, which necessitates more experimental data in practical applications. In contrast, EMD-based decomposition methods do not require predefined basic functions and exhibit excellent adaptability, effectively overcoming the limitations of wavelet-based decomposition methods (Zhang et al. 2019).

Empirical mode decomposition was proposed by (Huang et al. 1998), which decomposes the original signal series into a series of inherent mode functions (IMFs) with close frequency components and a residual sequence. These IMFs can show oscillation characteristics at each local. Besides, EMD can reduce the complexity of each IMF, thus improving the forecasting accuracy (Zhu et al.



A: It represents the low frequency component.

D: It stands for high frequency component.

AD: It represents the new high frequency component obtained from the low frequency component A.

Fig. 3 The comparison between three-layer WT and WPD of decomposing structure

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

2017). For example, (Yuan and Yang 2023) used the EMD to effectively improve the prediction effect of LSTM and ensure the stability of the model. However, EMD has the problem of mode mixing, which sometimes results in characteristics that are not representative of the original data (Huang et al. 1998). Moreover, the operation speed of EMD also needs to be improved. To solve the problems, EMD is constantly developed and improved and has many variations.

Based on EMD, (Wu and Huang 2009) added white noise to the original time series to create the ensemble empirical mode decomposition (EEMD). After multiple calculations, the noise will cancel each other out, and the final real decomposition results can be obtained. In the forecasting of air quality, EEMD is widely used in lots of pieces of literature, which plays a good role (Song and Fu 2020; Zhu et al. 2018a; Amanollahi and Ausati 2020). However, it is important to note that the introduction of white noise cannot be entirely eliminated, and the process of decomposing the Intrinsic Mode Functions (IMFs) using Empirical Mode Decomposition (EEMD) may generate additional interference signals during the reconstruction phase.

The complementary ensemble empirical mode decomposition (CEEMD) was proposed by Yeh et al. (2010), which is improved on EEMD. It adds a pair of noises with the same amplitude and opposite phase to the time series to reduce the residual noise. For example, (Fang et al. 2022) used the CEEMD to optimize the nonlinearity of PM2.5 numerical control data and improve the performance of the attention network. Considering the effectiveness of CEEMD, it has been applied in many literatures (Niu et al. 2016; Qi et al. 2019; Wang et al. 2018; Wang et al. 2017c; Xu et al. 2017a; Yang and Wang 2017; Zhu et al. 2018b, 2018c2019a).

In addition to classical EMD methods, some advanced strategies can also improve the performance of EMD. Torres et al. (2011) proposed the complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN), which is an improvement on CEEMD. CEEMDAN introduces adaptive noise. It greatly reduces the number of experiments and avoids the problem of incorrect components. Besides, there are the improved complementary ensemble empirical mode decomposition (ICEEMD) (Li et al. 2019; Xu et al. 2017b), the improved CEEMDAN (ICEEMDAN) (Jiang et al. 2019; Li and Zhu 2018; Sharma et al. 2020), and the fast ensemble empirical mode decomposition (FEEMD) (Luo et al. 2018). To sum up, these advanced data analysis and decomposition methods can provide effective technical support for air quality prediction (Masood and Ahmad 2023; Tian and Gai 2022; Yu et al. 2022). The process of improving EMD and its variations is illustrated in Fig. 4.

2.3 Variational mode decomposition

Although EMD-based methods have made some progress in research, these methods usually have modal aliasing and other phenomena, which affect the effect of the model (Li et al. 2022a). Compared with the EMD method, the variational mode decomposition (VMD) is an adaptive and completely non-recursive signal decomposition preprocessing method. The VMD is proposed by Dragomiretskiy and Zosso (2013), which can transform the signal decomposition problem into a variational problem. It can adaptively decompose the dataset into a certain discrete number of band-limited intrinsic mode functions (BIMFs). Compared with EMD-based methods, VMD avoids errors related to recursion (Wang et al. 2020b; Wu and Lin 2019a; Xu et al. 2020). Therefore, the VMD method can effectively alleviate the modal aliasing and boundary corresponding problems, which ensures the operational efficiency and robustness of the model.

2.4 Secondary decomposition

The air quality index (AQI) is the primary indicator used to assess air quality. Due to the influence of the Earth system, atmospheric diffusion, and human activities, air quality data typically encompass the overall periodicity of natural variability and local disruptions caused by other factors, leading to a significant nonlinearity in the AQI data (Li et al. 2022b). To reduce the non-linear characteristics of AQI and extract useful information from the original data, the secondary decomposition algorithm (SD), also known as the hybrid decomposition method, is sometimes used in air quality forecasting (Liu et al. 2021a). SD is to carry out further decomposition when the first decomposition does not achieve the target forecasting effect.

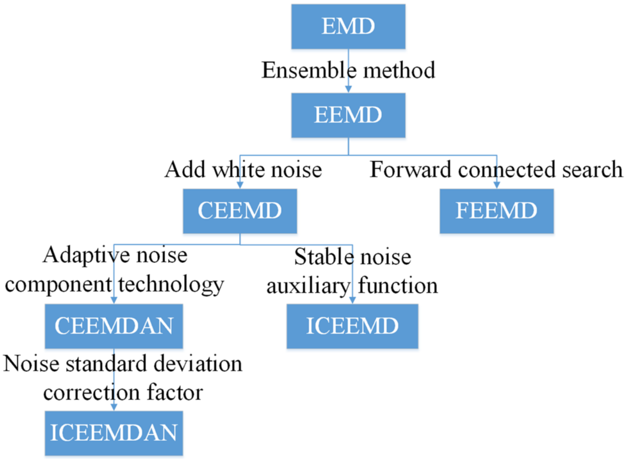

EMD
Ensemble method
EEMD
Add white noise
Forward connected search
CEEMD
FEEMD
Adaptive noise component technology
Stable noise auxiliary function
CEEMDAN
ICEEMD
Noise standard deviation correction factor
ICEEMDAN

Fig. 4 The development process of the EMD and improved variations

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Many scholars have adopted SD and proved that its forecasting ability is better than the single decomposition algorithm in most cases. Wang et al. (2017a) proposed to take WT as the main decomposition method and used VMD to further decompose the highly fluctuating high-frequency detail sequences to generate smoother sub-sequences. In the first layer decomposition, (Wang et al. 2017b) used CEEMD to decompose the AQI data sequence into ten IMFs with inconsistent frequencies. Then they used VMD to further decompose IMF1 with the highest frequency oscillation characteristics. Gan et al. (2018) first decomposed the original time series with WPD, and then used the CEEMD to further decompose the high-frequency components extracted by WPD. (Wu and Lin 2019b) used WD to obtain several IMFs with different frequencies, and then they used VMD to further decompose them into multiple variational modes (VMs) to obtain more satisfying forecasting results. Luo et al. (2018) applied SD by combining FEEMD with VMD for daily PM_{10} forecasting. Hao and Tian (2019) used the modified complete ensemble empirical mode decomposition with adaptive noise (MCCMADAN).

Table 1 provides a summary of the SD algorithms used in the aforementioned air quality forecasting studies.

However, SD often encounters issues such as over-decomposition or lengthy computation times. To address this problem, it is essential to carefully select the appropriate sequence for further decomposition during the initial decomposition. Two commonly employed strategies are used to tackle this issue: (a) utilizing sample entropy (SE) to determine the subseries for further decomposition and (b) directly selecting specific subseries. The SE was proposed by (Richman and Moorman 2000). The smaller the SE value, the higher the autocorrelation between the subseries (Wu and Lin 2019b). In other words, the higher the SE value, the more disordered the subseries. It is sometimes introduced as a subsequent step of SD. First, calculate the SE value of each BIMF. And then regroup them into new subseries based on the approximate SE values. Yu et al. (2023a) used the SE to improve the effect of VMD, which is usually incorporated into air quality forecasting models that employ a single decomposition method. In another reference (Liu et al. 2021a), SE is applied after the VMD, addressing the issue of excessive decomposition and simplifying the calculation process. The second strategy typically involves selecting the sequence with the highest frequency. Overall, the SE approach provides researchers with a means to choose secondary decomposition techniques. Table 2 provides a summary of the aforementioned data decomposition methods.

3 Dimensionality reduction

Compared with traditional univariate time series forecasting tasks, air quality forecasting is usually affected by different locations and different types of air pollutant data (Yu et al. 2023b). Specifically, air quality forecasting requires a large sample of historical data, which belongs to multidimensional data (or high-dimensional data). Multidimensional data analysis is scalable, but it is important to note that the complexity of data mining algorithms also increases with the number of dimensions in the data. Therefore, screening key information from a large number of high-level data can effectively improve the model prediction performance. To obtain a compact and comprehensive dataset and overcome the "curse of dimensionality", it is very necessary to reduce the dimension of multidimensional data (Domańska and Łukasik 2016). Dimensionality reduction algorithms (DR) can reduce redundant variables from the original

Table 1 The summary of SD algorithms in air quality forecasting

SD algorithm	Data source	Forecasting subject	Proposed model	Benchmark model	Improvement percentage*	Using the SE
WT + VMD (Wang et al. 2017a)	Wuhan, China	$PM_{2.5}$	WT-VMD-DE-BP	WT-DE-BP VMD-DE-BP	57.01%(MAPE) 66.72%(MAPE)	No
CEEMD + VMD (Wang et al. 2017b)	Beijing, China	AQI	CEEMD-VMD-DE-ELM	CEEMD-DE-ELM VMD-DE-ELM	74.04%(MAPE) 43.78%(MAPE)	No
WPD + CEEMD (Gan et al. 2018)	Shenyang, China	$PM_{2.5}$	SD-LSSVR-PSOGSA	WPD-LSSVR-PSOGSA CEEMD-LSSVR-PSOGSA	30.42%(MAPE) 42.96%(MAPE)	No
WD + VMD (Wu and Lin 2019b)	Beijing, China	AQI	SD-SE-LSTM-BA-LSSVM	WD-LSTM-BA-LSSVM	35.84%(MAPE)	Yes
CEEMADN + VMD (Hao and Tian 2019)	Jinan, China	AQI	MCEEMDAN-MOSSA-ENN	CEEMDAN-MOSSA-ENN	61.52%(MAPE)	No

*The improvement percentage is related to the error index of the proposed hybrid E_1 and the benchmark model E_2. $Improvement\ percentage = \frac{E_2 - E_1}{E_2} \times 100\%$

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 2 The summary of data decomposition methods in air quality forecasting

Category	Advantages	Limitations	Implementation details	Applied algorithms in references
Wavelet-based decomposition	Good localization characteristics in the time domain and frequency domain WPD can disassemble appropriate coefficients and detailed coefficients EWT can adjust parameters adaptively	The choice of decomposition level is difficult Results in the instability of the transformed coefficients estimated by DWT	It greatly depends on the decomposition structure formed by the wavelet function and decomposition level	WD (Cheng et al. 2019), DWT (Mittal and Bhardwaj 2011; Osowski and Garanty 2007; Siwek and Osowski 2012), MODWT (Prakash et al. 2011), SWT (Bai et al. 2016; Li and Tao 2017),WPD (Sun and Li 2020) (Wang et al. 2022), EWT (Huang et al. 2023) (Nuhu et al. 2023) (Huang et al. 2023)
EMD-based decomposition	Stronger capabilities when dealing with non-linear and non-stationary data It does not need to set up basic functions in advance	Traditional EMD has mode mixing	Improved variations have been proposed to enhance their ability	EMD (Zhu et al. 2017), EEMD (Song and Fu 2020) (Zhu et al. 2018a) (Amanollahi and Ausati 2020), CEEMD (Niu et al. 2016; Qi et al. 2019; Wang et al. 2018; Wang et al. 2017c; Xu et al. 2017a; Yang and Wang 2017; Zhu et al. 2018b, 2019a, 2018c), ICEEMD (Li et al. 2019; Xu et al. 2017b), ICEEMDAN (Jiang et al. 2019; Li and Zhu 2018; Sharma et al. 2020), FEEMD (Luo et al. 2018)
Variational mode decomposition	A multi-resolution and non-recursive method Avoids errors related to recursion	Traditional VMD is non-adaptive	A discrete number of BIMFs	VMD (Li et al. 2022a; Wang et al. 2020b; Wu and Lin 2019a; Xu et al. 2020) (Yu et al. 2023a)
Secondary decomposition	Its forecasting ability is higher than the single decomposition method under a reasonable scheme	Excessive decomposition Increase of computation time The selection of the appropriate sequence for further decomposition in the first decomposition is not easy	To carry out further decomposition when the first decomposition does not achieve the target forecasting effect Sometimes uses SE	WT + VMD (Wang et al. 2017a), CEEMD + VMD (Wang et al. 2017b), WPD + CEEMD (Gan et al. 2018), WD + VMD (Wu and Lin 2019b), CEEMADN + VMD (Hao and Tian 2019)

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

input data. The DR is to change the high dimensional data $Y = [y_1, y_2, \ldots y_m] \in R^{m\times p}$ with p dimensions into the low dimensional data $Z = [z_1, z_2, \ldots z_m] \in R^{m\times k}$ with k dimensions where k is less than p. At present, DR algorithms mainly include feature extraction and feature selection.

3.1 Feature extraction

Feature extraction aims to construct reduced datasets by eliminating redundant and irrelevant features. The newly obtained datasets should retain as much original information as possible.

Principal component analysis (PCA) is a commonly used linear mapping method that is based on the search for eigenvectors. It converts the data into a new set of uncorrelated indices called principal components (PCs) through an orthogonal transformation (Taghavi et al. 2023). PCA effectively preserves the main variance of the original data. Each PC is a linear combination of the original variables and is arranged in descending order based on their variance (Sah et al. 2023). The first k PCs are the ones with the biggest variance and reduce the dimension of data to k. Only they can visualize the data in the low dimensional space (Wen et al. 2023).

Sun and Sun (2016) used PCA to forecast daily $PM_{2.5}$ concentration. First, by calculating the correlation coefficient values, they found that $PM_{2.5}$ concentration had a very significant correlation with other air quality data except for O_3. Then they extracted the major information in the data, except for the $PM_{2.5}$ concentration of the previous day. Through PCA analysis, they determined that the first two components accounted for over 85% of the factors. These two principal components were then incorporated into the inputs to mitigate multicollinearity among predictors. The reference (Kumar and Goyal 2013) proposed a neural network hybrid model based on PCA to predict the daily air quality index. The reference (Kumar and Goyal 2011) combined PCA with the multiple linear regression (MLR) technique to forecast the daily air quality index in Delhi, also known as the principal component regression (PCR) technique. By introducing PCA, they reduced the number of variables to overcome the difficulty of calculating regression coefficients. Similarly, The reference (Azid et al. 2014) employed PCA with varimax rotation to extract five parameters from eight air quality parameters, which effectively represented the air quality variables. Their analysis of the experimental results demonstrated that selecting appropriate forecasting parameters could significantly decrease the number of samples required and reduce computation time.

3.2 Feature selection

Feature selection involves selecting the subset of features that are most relevant to the given problem from the original dataset (Ayesha et al. 2020). This subsection describes the feature selection methods currently used in air quality forecasting.

3.2.1 Cluster analysis

Cluster analysis is a technique used to study the underlying structure of data by selecting relevant features. This method involves dividing the input data into clusters, where objects within a cluster share similar characteristics. Subsequently, certain clusters can be eliminated or selected. In the context of air quality forecasting, commonly used cluster analysis methods include partitional clustering techniques and hierarchical clustering techniques (Govender and Sivakumar 2019). Partitional clustering involves partitioning the data to create all clusters simultaneously. This technique is widely used in the k-means technique due to its simplicity and efficiency. Besides, hierarchical clustering utilizes a distance metric to measure the dissimilarity between data points in clusters, which is an iterative clustering method, with the Ward technique being the most commonly used approach in hierarchical clustering.

Franceschi et al. (2018) utilized the k-means algorithm to cluster air pollutant concentration and meteorological variables. The determination of the optimal number of clusters is crucial in the application of k-means. By using a variation of the k-means technique, the x-means technology, at each station, they found that two clusters of data were the most appropriate. The findings demonstrated that the k-means technique can effectively uncover hidden information in the dataset, and utilizing the clustering results as inputs can enhance forecasting capabilities. Liu et al. (2008) used the Ward hierarchical clustering technique to analyze the similarity of the original data and divided them into several clusters. In hierarchical clustering techniques, Euclidean distance represents the similarity of different objects. Tamas et al. (2016) applied discriminant analysis (DA) based on hierarchical clustering analysis to reduce classification errors proposed a two-stage clustering method combining the Ward hierarchical clustering technique and the k-means technique to subdivide the dataset. In the literature, they used the cluster samples with the most similarities to the target as input. This two-step clustering method combines the advantages of the two clustering methods.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

3.2.2 Metaheuristic optimization

In addition to feature selection methods, some approaches also incorporate metaheuristic optimization algorithms. Firstly, a metaheuristic optimization algorithm is used to obtain a subset of input variables. Then a predictor is trained using this subset to generate the corresponding forecasting results. This process is repeated multiple times to obtain different forecasting results. Based on the forecasting accuracy of each subset, the most accurate subset is considered as the optimal subset. Various metaheuristic optimization algorithms inspired by biological or physical phenomena have been proposed in the literature (Lee and Geem 2005).

The genetic algorithm (GA) is a global optimization method that simulates the evolution process of natural selection and genetic mechanisms, including selection, crossover, and mutation (Niska et al. 2004; Zhai and Chen 2018). Particle swarm optimization (PSO) is an optimization algorithm inspired by the collective behavior of bird and fish populations, aiming to find the optimal solution (Nieto et al. 2018; Wang et al. 2005). The simulated annealing algorithm (SA) is usually used in a large search space. It starts with a relatively high initial temperature, and as the temperature parameter decreases, it randomly searches for the global optimal solution in combination with the probabilistic jump characteristic. Ly et al. (2019) conducted a comparison between two metaheuristic optimization algorithms, simulated annealing (SA) and PSO. The experimental results demonstrated that both algorithms performed well and were statistically significant. They were both statistically significant, while PSO showed slightly better performance compared to SA. Additionally, the differential evolution algorithm (DE), similar to genetic algorithms (GA), was utilized for global optimization in hybrid models (Teng et al. 2018). It guides the optimal search directly through the group intelligence generated by cooperation and competition among individuals in a group. Other metaheuristic optimization algorithms, such as the cuckoo search (CS) (Qin et al. 2014; Sun and Sun 2016), the bat algorithm (BA) (Wu and Lin 2019b), and the grey wolf optimizer (GWO) (Niu et al. 2016; Xu et al. 2017a), have also been commonly used in air quality forecasting to enhance the forecasting performance. In addition to using a single metaheuristic optimization algorithm, combining multiple metaheuristic optimization algorithms can further enhance the optimization performance (Yang and Wang 2017; Zhu et al. 2019b).

3.2.3 Correlation analysis

Some researchers neglect to examine the correlation between features and the relationship between input and output in their model-building process. To address this issue, correlation analysis can be used as a feature selection method, effectively resolving this problem.

The autocorrelation function (ACF) and the partial autocorrelation function (PACF) are common correlation analysis techniques, that identify the variables with the highest correlation to the predictor variable as the input variables (Zhao and Li 2019). The indexes to evaluate the correlation in time series are the correlation coefficient and partial correlation coefficient. Zhao and Li (2019) used ACF and PACF to determine the inherent lag in datasets and select the final predictor inputs. In previous literature, the highest peak value at lag1 in ACF indicated the highest correlation coefficient, leading to the selection of the corresponding variable as an input. On the other hand, when using PACF, they arranged the lag variables in descending order based on the partial correlation coefficient and selected the top-ranked variables.

As an approach for feature selection, grey correlation analysis (GCA) is utilized in air quality forecasting. GCA is a statistical analysis method that considers the geometric proximity between variables using gray correlation. The degree of gray correlation is used to indicate the correlation between meteorological factors and the target variable (Zhu et al. 2018b). The researchers calculated and ranked the gray correlation degrees corresponding to the meteorological factors, and selected the sequences with relatively high degrees of gray correlation. The reference (Qin et al. 2014) used GCA to determine the degree of relationship between PM and other predictors. The results showed that the relationship between atmospheric pollutants and particulate matter was greater than that between meteorological factors.

3.2.4 Other feature selection methods

In addition to classical methods, researchers have proposed some other feature selection methods in recent years. Zhai and Chen (2018) adopted a method called the stability feature selection (SFS) to identify important features and remove noisy features. This method is based on bootstrap and involves sub-sampling the training data to calculate an L1-based regularized estimation. It retains the variables that are consistently selected during the randomization process and disregards those with low probability, thus reducing the dimensionality.

Random forest (RF) is a feature selection method that can handle both linear and nonlinear problems without the need to consider independent and dependent variables separately. RFs consist of a predetermined number of binary decision trees, where each tree uses a subset of samples to make predictions. These predictions are then aggregated and optimized to produce the final result (Kamińska 2019). One advantage of RF is its ability to process both digital and non-digital features (Dotse et al. 2018; Kamińska 2019; Li

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

et al. 2018; Lyu et al. 2017). The reference (Dotse et al. 2018) proposed the GA-RF-BPNN model to predict daily PM_{10} exceedances. By optimizing RF tuning parameters, GA could better control the variable selection process. The reference (Li et al. 2018) used an online method based on RF to forecast $PM_{2.5}$, NO_2, and SO_2 concentrations 24 h in advance. They also utilized a sliding window approach to retrain the model with recent data.

The binary chaotic crow search algorithm (BCCSA) is a method that utilizes chaotic mapping to perform efficient and accurate searches. The reference (Qi et al. 2019) combined CEEMD and BCCSA to select the most appropriate IMFs and reconstruct them into input series. The sequences decomposed by CEEMD contained high-frequency IMFs, and removing the highest-frequency IMFs without discrimination would lead to a reduction in forecasting accuracy. To address this, BCCSA was employed after CEEMD to identify and eliminate the IMFs that negatively impact the forecasting results. BCCSA offers a significant advantage in selecting the optimal feature subset by effectively classifying features and utilizing the minimum number of features possible (Sayed et al. 2019).

Feature selection methods are also employed in air quality forecasting. Chen et al. (2017) utilized information theory based on partial mutual information (MI) to predict the air quality of Beijing. The feature information derived from MI is mutually independent and exhibits strong independence.

The phase space reconstruction (PSR) technology is based on chaos theory, which is related to delay time and the embedding dimension. Niu et al. (2017) applied the C–C method to PSR. This method determines the appropriate input form for each IMF component and the residual component, enabling the reconstruction of the time series into multi-dimensional data. This helps to minimize the impact of individual selective input on the forecasting results. Xu and Ren (2019) also integrated PSR with a forecasting model to forecast $PM_{2.5}$ concentration. They used the PSR to optimize parameters and achieved favorable performance.

In addition to the single feature selection method, some auxiliary methods have been proposed to complement these feature selection methods. One of these methods is the C–C method, as mentioned earlier. Zhang et al. (2016) introduced a feature selection method that combines the multifractal dimension and harmony search algorithm. They used the multifractal dimension as the evaluation criterion for feature subsets and determined the number of selected features based on it. Additionally, they employed an improved harmony search algorithm as the search strategy.

Overall, the feature selection method effectively optimizes the quality of input features for the model and enhances the overall performance. Table 3 provides a summary of the aforementioned feature selection methods.

4 Data correction

The main sources of AQI data are air monitoring stations and sensors located at various locations in space. However, due to device aging, transmission errors, and inclement weather, data collectors often fail, resulting in the unavailability of the collected data (Barkjohn et al. 2022). Therefore, the design of efficient data correction methods can ensure the accuracy of prediction tasks. In the currently published literature on air quality forecasting, the main data correction methods include missing data preprocessing and outlier detection.

Table 3 The summary of feature selection methods in air quality forecasting

Subcategory of feature selection	Characteristics or implementation details	Applied algorithms in references
Cluster analysis	Divides the input data into several clusters Objects in a cluster share similar characteristics	k-means technique (Franceschi et al. 2018), Ward hierarchical clustering technique (Liu et al. 2008), two-stage clustering technique (Tamas et al. 2016)
Metaheuristic optimization	Combines random algorithm with the local search algorithm Some of them are based on biological or physical phenomena in nature	GA (Niska et al. 2004; Zhai and Chen (2018), PSO (Nieto et al. 2018; Wang et al. 2005), SA (Ly et al. 2019), DE (Teng et al. 2018), CS (Qin et al. 2014; Sun and Sun 2016), BA (Wu and Lin 2019b), GBO (Niu et al. 2016; Xu et al. 2017a), hybrid metaheuristic optimization algorithms (Yang and Wang 2017; Zhu et al. 2019b)
Correlation analysis	Focuses on the correlation between features or the correlation between input and output	ACF (Zhao and Li 2019), PACF (Zhao and Li 2019), GCA(Qin et al. 2014; Zhu et al. 2018b)
Others	Some other feature selection methods and the combination with the auxiliary algorithm	SFS (Zhai and Chen 2018), RF (Dotse et al. 2018; Kamińska 2019; Li et al. 2018; Lyu et al. 2017), BCCSA (Qi et al. 2019), MI (Chen et al. 2017), PSR (Gan et al. 2018; Niu et al. 2017; Xu and Ren 2019)

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

4.1 Missing data processing

Missing data, also known as incomplete data matrices, is a common issue in air quality forecasting, particularly in small cities and localities (Quinteros et al. 2019). There are several reasons for data missingness, including insufficient sample data, measurement errors, and data collection failures (Junninen et al. 2004). In air quality forecasting, where data continuity is crucial for time series prediction, it becomes necessary to propose methods for handling missing data. These methods can be broadly categorized into two groups: (a) deletion methods and (b) imputation methods.

4.1.1 Deletion methods

In the field of air quality prediction, the deletion method is the simplest and most direct method to deal with missing data (Abdullah et al. 2016; Tamas et al. 2016). These methods can be categorized into two main types: list-wise deletion and pair-wise deletion. List-wise deletion, also known as the complete case approach, involves completely discarding units with missing data. Deletion methods are the default method used in many statistical packages. However, this approach can result in a loss of data and information. In practical applications, more than half of the samples are often lost, especially when there are a lot of variables. Pair-wise deletion, also known as available-case analysis, removes the units that are associated with any calculation involving variables with missing data (Plaia and Bondi 2006). This method can help mitigate the issue of significant data loss caused by list-wise deletion. However, both these deletion methods, while simple and fast, may result in deviations and unsatisfactory forecasting results. To address these problems, interpolation methods can be used effectively to alleviate them.

4.1.2 Imputation methods

The imputation method replaces the missing data with a new estimated value. Imputation methods are highly advantageous in reconstructing data sets. These methods primarily consist of single imputation, multiple imputation, and other methods for imputing missing data.

A. Single imputation methods

Single imputation methods fill in one value for each missing data. By summarizing and reviewing the literature on air pollution forecasting, single imputation is divided into three categories: (a) mean imputation, (b) random imputation, and (c) conditional mean imputation. The mean imputation method replaces the missing value with the mean value (Niu et al. 2016; Sharma et al. 2020). However, this method will destroy the inherent structure of the dataset easily and greatly reduce the prediction effect. The conditional mean imputation method is to replace the missing value with the conditional mean.

Plaia and Bondi (2006) used conditional mean imputation methods including the hour mean method, row mean method, and last and next method to forecast PM_{10} concentrations. The hourly mean method utilizes information from the same monitoring station on an hourly basis. The missing hourly observation is filled by calculating the average value of all known hourly observations in the same hour at the same monitoring station within the specified time. The row mean method utilizes hourly information from other monitoring stations. The missing hourly observation is filled by calculating the average value of all known hourly observations in the same hour at other monitoring stations within the specified time. Regarding the last and next method, it utilizes the last known and next known observations in the monitoring station. The missing hourly observation is filled by calculating the average of the two known observations. (Freeman et al. 2018) adopted two different single imputation methods according to the number of consecutive gaps in data. When the number of consecutive gaps is less than eight, the missing data are estimated linearly, and the missing data is filled with the values of the first and the last measurement in the gap according to the previous observations. When the number of consecutive gaps is more than eight, the missing data is filled by averaging the corresponding hourly measurement values of the previous 2 days. According to the experimental results, the error of the first method increases as the gap increases. At the same time, the second method has a larger error.

Single imputation methods have many attractive features. They only need to be carried out once for each missing value, thus they can be applied directly when dealing with the missing data. However, a single imputed value cannot completely determine which value to implement, and the accuracy needs to be improved.

B. Multiple imputation methods

Compared with single imputation, multiple imputation considers the errors of interpolation, which puts more attention on the error variance problem and improves the accuracy of forecasting. Multiple imputation generates multiple simulated values instead of one for each missing data (Junninen et al. 2004). In general, the method consists of three steps: Firstly, multiple imputation considers all possible values that can be used to populate missing data and creates multiple copies of the database. Secondly, multiple imputation analyzes each database individually. Finally, multiple imputation performs a comprehensive analysis of the results from each database, taking into

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

account the standard error. In this way, data uncertainty is incorporated into the process. Lei and Wan (2010) used multiple imputation methods in air quality forecasting in Macau. The results proved that the forecasting performance after preprocessing of missing data was better than that of the conventional case without multiple imputation methods.

C. Other missing data imputation methods

Kolehmainen et al. (2001) used the nearest neighbor method. They disregarded the effect of the replacement method on the results, as the percentage of missing data in the entire dataset was only 0.94%. (Qin et al. 2014; Xu et al. 2017b) employed the cubic spline imputation method to address the missing data. Junninen et al. (2004) compared linear imputation, nearest neighbor, linear imputation, and spline imputation methods. The results indicated that the linear and spline imputation methods exhibited similar performance when dealing with gaps of 1–2 values. However, as the gap length increases, the non-systematic physical errors increase sharply. This leads to poor spline imputation performance until non-physical artifacts with gaps greater than 24 values appear randomly. They concluded that the nearest neighbor is the best choice for the univariate method, but the limitation of the gap length should be considered when using it.

Westerlund et al. (2014) proposed a two-step imputation scheme. In the first step, they used the site-dependent effect method (SDEM) which takes into account the correlation between spatial and temporal factors. The performance of SDEM was found to be superior to both single and multiple imputation methods. However, SDEM relies on data from other monitoring stations for imputation. If the available data from other stations is limited, SDEM cannot be applied. To solve this problem, they adopted the cyclisation temporal variability approach as the second step. This method requires less data information compared to SDEM. In this study, the second-step method was used for interpolation with less than 2.1% of the total observations.

In addition to classic methods, deep learning-based frameworks are also commonly used (Freeman et al. 2018). Wu et al. (2022) proposed the inverse mapping generative adversarial network to realize missing data imputation. Compared with classical models, deep learning optimizes the imputation performance. Sayeed et al. (2022) proposed the convolutional neural network-based missing data imputation method. The results prove the effectiveness of deep learning. In summary, the effectiveness of deep learning has been convincingly demonstrated in the field of missing data imputation. As a result, this field is poised to become a promising avenue for future research.

4.2 Outlier detection

In time series analysis, outliers are typically defined as data points that deviate from the expected patterns or significantly differ from other observed data points. These outliers can occur due to external interferences, such as sampling errors or accidental abnormal factors. The presence of outliers can have detrimental effects on time series analysis as they can directly impact the accuracy of forecasting models and potentially lead to incorrect conclusions when analyzing the dataset.

To enhance the accuracy and efficiency of forecasting models and mitigate the impact of outliers, the practice of outlier detection is commonly employed in air quality forecasting. The primary objective of outlier detection is to identify anomalous data and behaviors that do not conform to the expected patterns or deviate from the anticipated outcomes. Various detection techniques are utilized to identify these outliers, which are subsequently rectified or eliminated.

4.2.1 Hampel identifier

Regarding the original time series, the Hampel identifier (HI) is widely recognized as the most effective and commonly used method for outlier detection. The HI is capable of identifying outliers in the original time series and filtering them out. One advantage of using the Hampel Identifier (HI) method to eliminate outliers is that it does not require prior prediction of interference, and the processed series remains undistorted.

HI is also known as a triple standard deviation discriminant algorithm. It sets the sliding window length and evaluation parameters and determines the threshold value according to the rules of the 3σ algorithm. Then determine the data that can be considered as outliers according to the relationship between the threshold value and the sample value. (Wang et al. 2020a) applied the HI algorithm to forecast the daily air quality index. In their study, they used eight common evaluation criteria to demonstrate the comprehensive performance of the models. The HI algorithm was utilized to detect the AQI series and correct the outliers by replacing them with the local median. To assess the effectiveness of the HI algorithm, the experiment included several comparison models, with the only difference being the use of the HI algorithm. The results indicated that the forecasting models based on HI demonstrated superior comprehensive performance and significantly improved forecasting ability. Qi et al. (2019) applied HI to eliminate outliers in the AQI series before data decomposition. They also use the Fast Fourier Transform (FFT) to convert the outliers to be eliminated in the frequency domain.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 4 The summary of data correction methods in air quality forecasting

Subcategory of data correction methods	Further classification	Characteristics or implementation details	Applied algorithms in references
Missing data processing	Deletion methods	The list-wise deletion and pin-wise deletion are commonly used to eliminate missing data	(Abdullah et al. 2016; Tamas et al. 2016) (Plaia and Bondi 2006)
	Imputation methods	Single imputation methods: (a) mean imputation, (b) random imputation, and (c) conditional mean imputation Multiple imputation methods: It puts more attention on the error variance problem and improves the accuracy of forecasting Other missing data imputation methods	(Niu et al. 2016; Sharma et al. 2020) (Plaia and Bondi 2006) (Freeman et al. 2018) (Junninen et al. 2004) (Lei and Wan 2010) (Kolehmainen et al. 2001) (Qin et al. 2014; Xu et al. 2017b) (Junninen et al. 2004) (Wu et al. 2022) (Sayeed et al. 2022)
Outlier detection	Hampel identifier	It is widely recognized as the most effective and commonly used method for outlier detection	(Wang et al. 2020a) (Qi et al. 2019)
	Regression-based outlier detection	It is a method that takes into account the temporal aspect of the data, using contextual outlier detection	(Díaz-Robles et al. 2008; Kumar et al. 2004; Siew et al. 2008)
	Other missing data imputation methods	It includes some more advanced frameworks to improve the performance of traditional models	(Dincer and Akkuş, 2018) (Wu et al. 2018) (Wei et al. 2023) (Lai et al. 2022)

4.2.2 Regression-based outlier detection

Regression-based outlier detection is a method that takes into account the temporal aspect of the data, using contextual outlier detection. In the field of air quality forecasting, the ARIMA linear regression method is commonly employed (Díaz-Robles et al. 2008; Kumar et al. 2004; Siew et al. 2008). To detect outliers using ARIMA, the method fits multiple time series separately, utilizing standard settings. It then calculates the outliers for each data point by taking the absolute value of the residual between the model value and the measured value. However, due to the non-stationary and chaotic nature of the original air pollution data, these regression-based outlier detection and correction methods cannot fully reflect the information of the pollution index sequence.

4.2.3 Other outlier detection methods

The fuzzy time series (FTS) model is advantageous as it requires minimal observation data and does not rely on statistical assumptions. However, FTS models have limitations in handling outliers (Liang et al. 2023a). In order to address this issue, the reference (Dincer and Akkuş 2018) proposed a model based on the fuzzy K-medoid (FKM) clustering algorithm, which can effectively solve these limitations. The selection of the clustering center plays a crucial role in the FKM clustering algorithm and is determined based on actual observations in the dataset. The probability of outliers being selected as clustering centers is very low. Experimental results demonstrate that the proposed forecasting model exhibits robustness against outliers and performs well in forecasting time series models that contain a large number of outliers.

Wu et al. (2018) introduced a fully automatic method based on the probability of residuals for outlier detection. They categorized the abnormal data into four types based on its characteristics. To effectively identify outliers, they employed fitting estimation methods specific to each type of abnormal data. They utilized techniques such as low-pass filtering and spatial regression to fit the monitoring data. The probability of residuals was calculated by examining the distribution characteristics of the residuals, and values with low probability were identified as outliers.

In addition to classic methods, outlier detection methods based on neural networks have also been widely used (Zheng et al. 2022). Wei et al. (2023) proposed the Long short-term memory network (LSTM) based outlier detection method to optimize air quality data. Experiments show that deep learning can significantly improve the effectiveness of the model. Lai et al. (2022) evaluated the performance of four recurrent neural networks (RNN) models in the field of air quality outlier detection. The experimental results demonstrate the superiority of the RNN framework.

Overall, the data correction method effectively optimizes the quality of the raw air quality data and enhances the performance of predictors. Table 4 provides a summary of the data correction methods.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

5 Spatial interpolation

To monitor and forecast air pollutants, many air quality monitoring stations have been established. However, the high cost of establishing these stations has resulted in sparse distribution and irregular spacing. Consequently, the observation data fail to accurately reflect the real-time spatial pollutant concentration. In order to address this issue, several spatial interpolation methods have been proposed (Wei et al. 2022). These methods utilize data from known monitoring sites within the region to estimate observations from non-monitoring sites. These spatial interpolation methods are widely used in air quality forecasting, with Kriging and inverse distance weighting (IDW) being the most commonly used traditional methods (Liang et al. 2022). Moreover, researchers have continuously proposed improved methods based on these traditional techniques.

5.1 Kriging

Kriging is a widely used geostatistical interpolation method for estimating spatial parameters (Andria et al. 2008). It involves estimating data from non-monitoring stations by considering the distance and degree of variation of data points from known monitoring stations. This method utilizes a function called variogram to capture spatial variations. The variogram reflects the spatial autocorrelation property and can be employed to assess the similarity between two observed values at a specific distance. By considering the spatial distribution, kriging helps to minimize the error in predicting estimated values.

Tong et al. (2015) compared the performance and practical feasibility of various models based on the traditional kriging method when studying the air quality index of Wuhan, China. Sahu and Mardia (2005) applied kriging to short-term forecasting of air pollution and established the spatial forecasting surface of the model. Kottur and Mantha (2015) proposed an integrated approach using ANN and kriging to forecast the concentration of air pollution using meteorological data. In addition, the input features such as rainfall, temperature, wind speed, direction, longitude, and dimensional spatial parameters are also considered. As a classic spatial interpolation method, the kriging method can provide estimates of pollutant concentration in non-monitoring stations. Besides, the kriging method considers the location factor when forecasting the parameter value, which is proved to be the best linear unbiased estimation method.

5.2 Inverse distance weighting

As a deterministic spatial interpolation method, Inverse Distance Weighting (IDW) is known for its speed and simplicity (Lu and Wong 2008). The fundamental concept behind IDW is to estimate unsampled values by taking the weighted average of known values. These weights are determined based on the distance between the known sampled locations and the unknown predicted locations, with the weights being inversely proportional to the distance. However, considering that the spatial relation is more complex than a simple spatial distance and position relation, IDW uses power or exponential functions to modify the weights. Empirical studies have shown that, in general, the adaptive IDW method outperforms the constant parameter method. Furthermore, it has been proven that the adaptive IDW method performs better than the ordinary kriging method when the typical variogram is not effective in capturing the spatial structure of the data (Lu and Wong 2008).

Ma et al. (2019b) proposed a method that combines the bi-directional long short-term memory (BLSTM) network with IDW for spatial forecasting of air pollution. The inclusion of Inverse Distance Weighting (IDW) in the analysis allows for the consideration of spatial–temporal correlation and the interpolation of the spatial distribution of air pollution. This approach addresses the limitation of most deep learning methods in current air quality forecasting, which primarily focus on data from monitoring stations and neglect those without. Jumaah et al. (2019) utilized IDW in combination with a geographic information system (GIS) based on the ordinary least squares (OLS) to forecast the air quality index in Kuala Lumpur, Malaysia.

5.3 Other spatial interpolation methods

The kriging and IDW methods have been widely used as traditional spatial interpolation methods in air quality forecasting and have indeed provided more accurate concentration forecasting for those non-monitoring stations. However, these methods are designed for static research objects, which poses significant limitations in actual air quality forecasting. Additionally, the use of predetermined formulas, such as power functions or exponential functions in IDW, can lead to inaccuracies in the forecasting results. In this way, the accuracy of forecasting results is largely dependent on the experimenter (Lu and Wong 2008). To solve these problems, various other spatial interpolation methods have been proposed, utilizing machine learning or neural networks. These methods exhibit strong robustness, non-linear mapping capability, and effective adjustment ability.

Wahid et al. (2013) introduced a spatial interpolation method that utilized the radial basis function neural network (RBFNN) to predict ozone concentration in the Sydney basin, Australia. The authors emphasized the significance of selecting the radial basis centers, basis function radius, and network weights in this method. The findings of their ozone concentration forecasting demonstrated the method's strong computational ability and forecasting performance. To discuss the mean spatial distribution of atmospheric

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

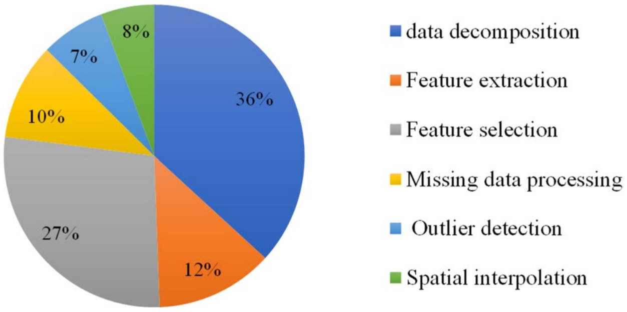


Fig. 5 The usage frequencies of six data preprocessing methods in existing air quality forecasting models: results derived from all application cases

Table 5 The performance evaluation of data preprocessing methods in air quality forecasting (The number of * is proportional to the size of the corresponding index)

Category	Usage frequency	Accuracy improvement	Consuming time	Maturity	Implementation difficulty
Data decomposition	*****	****	*****	****	**
Feature extraction	**	**	**	***	**
Feature selection	****	***	***	***	**
Missing data processing	***	*	**	**	**
Outlier detection	***	***	*	*	***
Spatial interpolation	**	**	*	**	***

pollutants, (Pfeiffer et al. 2009) proposed a spatial interpolation method for examining the average spatial distribution of atmospheric pollutants. Their approach involved diffusion sampling measurements and the evaluation of artificial neural networks (ANN). Diffusion samplers were employed to simultaneously measure the average concentration of pollutants in the specified area. Ma et al. (2019a) utilized a geographic long short-term memory (LSTM) neural network for predicting PM2.5 concentration. The inclusion of the geological layer aided in integrating and analyzing spatial relationships with other known locations, thereby enhancing the accuracy and reliability of interpolating unknown locations. Furthermore, their research demonstrated that the proposed method exhibited a lower root mean square error (RMSE) for spatial distribution compared to traditional spatial interpolation techniques like kriging and IDW.

6 Discussion

This paper examines the application of various data preprocessing methods in air quality forecasting. By analyzing a large number of papers related to air quality prediction, this section discusses the application of different data preprocessing methods. These approaches can be refined into six main technical directions: data decomposition, feature extraction, feature selection, missing data processing, outlier detection, and spatial interpolation. Each method has distinct characteristics and is applied under specific conditions. The frequency of their usage in air quality forecasting varies. The distribution of the four methods is illustrated in Fig. 5, which shows that data decomposition is the most commonly employed data preprocessing method in air quality forecasting. Dimension reduction is commonly used in air quality forecasting, with feature selection being more prevalent than feature extraction. However, methods for missing data processing and outlier detection have not been extensively explored in this field. The application of missing data processing techniques is limited due to the requirement of a significant amount of historical data in air quality forecasting, leading to the occasional neglect of missing data. Outlier detection is rarely employed in air quality forecasting mainly due to the challenge of distinguishing between outlier samples and normal ones. Spatial interpolation is the least utilized method in this context.

To visually compare these data preprocessing methods, Table 5 provides a comprehensive evaluation of their performance. The evaluation takes into account factors such as frequency of use, improvement of accuracy, consuming time, maturity, and difficulty of implementation. The number of stars represents the performance level, with more stars indicating better performance. The results were obtained from our statistical analysis of all kinds of literature. Specifically, this paper considers the frequency of use of different data processing methods and their experimental effects in the paper. However, these methods are not independent, and each has its unique characteristics, making it difficult to evaluate them as inherently good or bad. Additionally, a review of the literature reveals that most studies employ multiple data preprocessing methods rather than relying on a single method. Combining multiple data preprocessing methods often leads to better results in forecasting air quality.

6.1 Challenges in current applications

Although data preprocessing methods have made many achievements in air quality forecasting, they also face some

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

challenges in application, which requires further research and analysis.

6.1.1 Challenges in data preprocessing methods

As can be seen from the previous analysis, AQI data is usually composed of global regularity and local disturbance due to the influence of human activities and natural laws. The data decomposition method can effectively decompose the original data into regular components and disturbance components to improve the modeling effect of the model. However, due to the complexity of natural laws and human activities, it is difficult to fully solve the nonlinear problem of AQI data with the simple decomposition strategy (Guo et al. 2023). In addition, although the secondary decomposition can further improve the decomposition effect, this strategy requires adequate evaluation of the subsequences to prevent overfitting problems caused by over-decomposition (Zhang et al. 2023). Therefore, effectively utilizing expert experience to analyze the intrinsic characteristics of air quality data and improve the efficiency of secondary decomposition remains an effective and challenging research direction.

6.1.2 Challenges in taking advantage of other features

Considering the composition, generation, and diffusion of air pollution, it is important to fully explore the interconnections between different geographical locations and different air pollutants during the modeling process (Udristioiu et al. 2023). Existing methods mainly use feature selection and feature extraction to optimize input features, but they ignore the spatio-temporal correlation between different spatial sites and the heterogeneity between different variables (Chengqing et al. 2023). Therefore, it is a challenging and critical task to design a spatiotemporal prediction model that can make full use of all the characteristic information.

6.1.3 Challenges in data overcorrection

During the process of outlier detection, there is a possibility of data overcorrection. This refers to the correction of non-outliers or outliers beyond the specified requirements. However, overcorrection has several drawbacks, including a decrease in forecasting accuracy. Unfortunately, there is currently no optimal solution to address this issue in air quality forecasting. One potential approach is to introduce certain indicators or methods to effectively tackle this problem. On the one hand, when restoring outliers, the modification of normal values can also be set as an optimized indicator, which can prevent a bad impact on normal values (Yıldız et al. 2022). On the other hand, the purpose of optimizing outliers is to further improve the prediction effect. Therefore, using the prediction results as an additional indicator can further improve the performance of the model (Gilik et al. 2022).

6.2 Possible future development trends

The importance of historical data as the foundation for forecasting is well-documented in existing literature. Moreover, data-driven models are increasingly being employed in air quality forecasting. Hence, it is crucial to give due consideration to data preprocessing methods in future research. The following are some potential areas for future development.

6.2.1 Use the most appropriate data correction method

Among the literature dealing with missing data using imputation methods, it is evident that the majority of studies utilize the mean substitution method. However, this approach proves to be ineffective when dealing with cases that have a limited amount of historical data or a significant amount of missing data (Flores et al. 2023). In order to deal with missing data and abnormal data effectively, it is very important to fully mine and model the hidden association between outliers and valid values (Liang et al. 2023b). At present, graph-based and attention-based strategies have gained wide attention in the field of data correction because they can effectively discover the hidden association between outliers and valid values. For example, (Lin et al. 2022) used the graph neural network to model the temporal and spatial correlation between missing values and normal values, and achieved better results than traditional methods. Feng et al. (2023) proposed the attention-based outlier detection method, which can work better than classic models. In the future, by effectively combing the data characteristics and proposing missing value recovery and outlier detection methods based on deep learning, superior results will be achieved compared to traditional statistical strategies.

6.2.2 Combine with appropriate post-processing methods

Data post-processing and data preprocessing are two methods of data processing, but they serve different purposes in the model. Data preprocessing is performed before feeding the data into prediction models, while data post-processing is focused on improving the initial output. Both these methods are crucial in data-driven models. In future research, it would be beneficial to select suitable data post-processing techniques to complement data preprocessing methods and maximize the effectiveness of data processing. In particular, with the rapid development of machine learning (Mi et al. 2022) and multivariate time series forecasting techniques (Yan et al. 2022), it is important and effective to incorporate more advanced forecasting models.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

6.2.3 Combine with other environmental and natural condition data

In air quality forecasting, historical data is commonly used. However, several other factors also impact the forecasting results, such as terrain, industrial emissions, wind speed, wind direction, humidity, and temperature (Shang et al. 2022). These factors contribute to a comprehensive analysis of the overall environment and natural conditions, which indirectly affect the prediction accuracy. For example, (Zheng et al. 2015) suggested that making full use of big data technology and other data can further optimize the prediction effect of the model. Liu et al. (2021b) combined different air monitoring stations and different air pollutants to achieve the AQI spatiotemporal prediction technology. Compared with univariate modeling, this approach can achieve better results (Liang et al. 2023c). In the future, it would be beneficial to focus on incorporating additional environmental and natural condition data to enhance the overall forecasting capability.

6.2.4 Combine with deep learning techniques

In recent years, with the continuous improvement of data volume and computing resources, deep learning methods have been widely studied in the field of AQI prediction (Lin et al. 2018). In addition to AQI prediction technology, deep learning has also made vigorous development in the fields of feature analysis, outlier detection and so on (Méndez et al. 2023). For example, (Wong et al. 2023) used a spatiotemporal graph neural network to improve the effect of AQI data imputation. Yang et al. (2023a) proposed the self-attention-based AQI data outlier detection method, which can work better than classic models. Therefore, making full use of deep learning technology can further improve the effectiveness of existing methods.

7 Conclusion

With the increasing severity of air pollution, air quality forecasting has become increasingly crucial for early warning systems and human well-being. The utilization of data preprocessing methods in data-driven models greatly impacts the accuracy of these forecasts. This research paper provides a comprehensive analysis of four data preprocessing methods commonly used in current air quality forecasting literature: data decomposition, dimensionality reduction, data correction, and spatial interpolation. This study thoroughly analyzes the purpose, improvements, characteristics, and implementation details of each data preprocessing method from an academic perspective. The evaluation of these methods is based on five perspectives: frequency of use, improvement of accuracy, consuming time, maturity, and difficulty of implementation. Additionally, the study identifies current challenges, potential development trends, and prospects based on the review results. In summary, the previous discussion leads to the following conclusions:

- Among the four data preprocessing methods, spatial interpolation is relatively less used. However, it is an area that can be further explored in future research. On the other hand, data decomposition and feature selection techniques have already been well-established and widely applied in various applications.
- It is difficult to determine the superiority or inferiority of data preprocessing methods. Each method has its advantages and disadvantages, and its performance has improved over time. For instance, preprocessing methods based on data decomposition can effectively enhance forecasting ability. However, these methods may require more computation time due to their specific framework, and there is a possibility of over-decomposition. Therefore, it cannot be said that one data preprocessing method is superior or inferior. Each data preprocessing method has its own advantages and disadvantages, and its performance has improved gradually during the development process.

Appendix

The list of all abbreviations is given in Table 6.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 6 List of abbreviations

ACF	Autocorrelation function	IMFs	Inherent mode functions
ANN	Artificial neural network	LSSVM	Least squares support vector machine
AQI	Air quality index	LSTM	Long short-term memory
ARIMA	The autoregressive integrated moving average model	MAPE	Mean Absolute Percentage Error
ARMA	Autoregressive moving average model	MCCMADAN	Modified complete ensemble empirical mode decomposition with adaptive noise
BA	Bat algorithm	MI	Mutual information
BCCSA	Binary chaotic crow search algorithm	MLR	Multiple linear regression model
BIMFs	Band-limited intrinsic mode functions	MM5	Fifth-generation mesoscale model
BLSTM	Bi-directional long short-term memory	MODWT	Maximum overlap wavelet transform
BPNN	Back propagation neural network	MSE	Mean squared error
RMSE	Root mean square error	NO	Nitric oxide
CEEMD	Complementary ensemble empirical mode decomposition	CEEMDAN	Complete ensemble empirical mode decomposition with adaptive noise
NO_2	Nitrogen dioxide	O_3	Ozone
CMAQ	Community multiscale air quality model	OLS	Ordinary least squares
CO	Carbon monoxide	PACF	Partial autocorrelation function
CS	Cuckoo search	PCA	Principal component analysis
CWT	Continuous wavelet transform	PCR	Principal component regression
DA	Discriminant analysis	PCs	Principal components
DE	Differential evolution algorithm	PM	Particulate matter
DR	Dimensionality reduction	PSO	Particle swarm optimization
DWT	Discrete wavelet transform	PSR	Phase space reconstruction
EEMD	Ensemble empirical mode decomposition	IDW	Inverse distance weighting
EMD	Empirical mode decomposition	RBFNN	Radial basis function neural network
EWT	Empirical wavelet transform	RF	Random forest
FEEMD	Fast ensemble empirical mode decomposition	CAMX	Comprehensive air quality model with extensions
FFT	Fast Fourier Transform	SA	Simulated annealing algorithm
FKM	Fuzzy K-medoid clustering algorithm	SD	Secondary decomposition algorithm
FLM	Fuzzy logic model	SDEM	Site-dependent effect method
FT	Fourier transform	SE	Sample entropy
FTS	Fuzzy time series model	SFS	Stability featureselection
GA	Genetic algorithm	SVD	Singular value decomposition
GCA	Grey correlation analysis	SVM	Support vector machine
GIS	Geographic information system	SWT	Stationary wavelet transform
GRNN	Generalized regression neural network	VMD	Variation mode decomposition
GWO	Grey wolf optimizer	VMs	Variational modes
HI	Hampel identifier	WD	Wavelet decomposition
ICEEMD	Improved complementary ensemble empirical mode decomposition	WPD	Wavelet packet decomposition
ICEEMDAN	Improved CEEMDAN	WT	Wavelet transform

Acknowledgements This study is fully supported by National Natural Science Foundation of China (52102471), and Fundamental Research Funds for the Central Universities (Science and technology leading talent team project 2022JBXT008).

Author contributions Chengqing Yu: Writing, Methodology. Jing Tan: Writing, Modification. Yihan Cheng: Writing. Xiwei Mi: Conceptualization, Methodology.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Declarations

Conflict of interests The authors declare no competing interests.

References

Abdullah S, Ismail M, Fong SY, Ahmed N (2016) Evaluation for long term PM 10 concentration forecasting using multi linear regression (MLR) and principal component regression (PCR) models. EnvironmentAsia, 9.

Amanollahi J, Ausati S (2020) PM 2.5 concentration forecasting using ANFIS, EEMD-GRNN, MLP, and MLR models: a case study of Tehran, Iran. Air Qual Atmos Health 13:161–171.

Andria G, Cavone G, Lanzolla AM (2008) Modelling study for assessment and forecasting variation of urban air pollution. Measurement 41:222–229

Ayesha S, Hanif MK, Talib R (2020) Overview and comparative study of dimensionality reduction techniques for high dimensional data. Inf Fusion 59: 44–58.

Azid A, Juahir H, Toriman ME, Kamarudin MKA, Saudi ASM, Hasnam CNC et al (2014) Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: a case study in Malaysia. Water Air Soil Pollut 225:2063

Bai Y, Li Y, Wang X, Xie J, Li C (2016) Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. Atmos Pollut Res 7:557–566

Bai L, Wang J, Ma X, Lu H (2018) Air pollution forecasts: an overview. Int J Environ Res Public Health 15:780

Barkjohn KK, Holder AL, Frederick SG, Clements AL (2022) Correction and accuracy of PurpleAir PM2. 5 measurements for extreme wildfire smoke. Sensors 22:9669.

Byun D, Schere KL (2006) Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system.

Casazza M, Lega M, Jannelli E, Minutillo M, Jaffe D, Severino V et al (2019) 3D monitoring and modelling of air quality for sustainable urban port planning: Review and perspectives. J Clean Prod 231:1342–1352

Chen Y, Shi R, Shu S, Gao W (2013) Ensemble and enhanced PM10 concentration forecast model based on stepwise regression and wavelet analysis. Atmos Environ 74:346–359

Chen S, Kan G, Liang K, Zhang M, Li J, Hong Y, et al (2017) Air quality analysis and forecast for environment and public health protection: a case study in Beijing, China. Transylvanian Rev 24(12):3575–3591.

Chen X, Yin L, Fan Y, Song L, Ji T, Liu Y, et al (2020) Temporal evolution characteristics of PM2. 5 concentration based on continuous wavelet transform. Sci Total Environ 699:134244.

Cheng Y, Zhang H, Liu Z, Chen L, Wang P (2019) Hybrid algorithm for short-term forecasting of PM2. 5 in China. Atmos Environ 200:264–279.

Chengqing Y, Guangxi Y, Chengming Y, Yu Z, Xiwei M (2023) A multi-factor driven spatiotemporal wind power prediction model based on ensemble deep graph attention reinforcement learning networks. Energy 263:126034

Dabin Z, Boting Z, Liwen L, Liling Z (2022) Carbon price forecasting based on secondary decomposition and aggregation strategy. J Syst Sci Math Sci 42:3094

Daly A, Zannetti P (2007) Air pollution modeling–An overview. Amb Air Pollut, 15–28.

Díaz-Robles LA, Ortega JC, Fu JS, Reed GD, Chow JC, Watson JG et al (2008) A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco. Chile Atmos Environ 42:8331–8340

Dincer NG, Akkuş Ö (2018) A new fuzzy time series model based on robust clustering for forecasting of air pollution. Eco Inform 43:157–164

Domańska D, Łukasik S (2016) Handling high-dimensional data in air pollution forecasting tasks. Eco Inform 34:70–91

Dotse S-Q, Petra MI, Dagar L, De Silva LC (2018) Application of computational intelligence techniques to forecast daily PM10 exceedances in Brunei Darussalam. Atmos Pollut Res 9:358–368

Dragomiretskiy K, Zosso D (2013) Variational mode decomposition. IEEE Trans Signal Process 62:531–544

Fang S, Li Q, Karimian H, Liu H, Mo Y (2022) DESA: a novel hybrid decomposing-ensemble and spatiotemporal attention model for PM2. 5 forecasting. Environ Sci Pollut Res 29:54150–54166.

Feng Y, Kim J-S, Yu J-W, Ri K-C, Yun S-J, Han I-N et al (2023) Spatiotemporal informer: a new approach based on spatiotemporal embedding and attention for air quality forecasting. Environ Pollut 336:122402

Flores A, Tito-Chura H, Centty-Villafuerte D, Ecos-Espino A (2023) Pm2. 5 time series imputation with deep learning and interpolation. Computers 12:165.

Franceschi F, Cobo M, Figueredo M (2018) Discovering relationships and forecasting PM10 and PM2. 5 concentrations in Bogotá, Colombia, using artificial neural networks, principal component analysis, and k-means clustering. Atmos Pollut Res 9:912–922.

Freeman BS, Taylor G, Gharabaghi B, Thé J (2018) Forecasting air quality time series using deep learning. J Air Waste Manag Assoc 68:866–886

Gan K, Sun S, Wang S, Wei Y (2018) A secondary-decomposition-ensemble learning paradigm for forecasting PM2. 5 concentration. Atmos Pollut Res 9:989–999.

Gilik A, Ogrenci AS, Ozmen A (2022) Air quality prediction using CNN+ LSTM-based hybrid deep learning architecture. Environ Sci Pollut Res, pp 1–19.

Govender P, Sivakumar V (2019) Application of k-means and hierarchical clustering techniques for analysis of air pollution: a review (1980–2019). Atmos Pollut Res.

Gulia S, Nagendra SS, Khare M, Khanna I (2015) Urban air quality management-a review. Atmos Pollut Res 6:286–304

Guo Q, He Z, Wang Z (2023) Prediction of hourly PM2. 5 and PM10 Concentrations in Chongqing City in China based on artificial neural network. Aerosol Air Qual Res 23:220448.

Hao Y, Tian C (2019) The study and application of a novel hybrid system for air quality early-warning. Appl Soft Comput 74:729–746

Huang Y, Zhang X, Li Y (2023) A novel hybrid model for PM2. 5 concentration forecasting based on secondary decomposition ensemble and weight combination optimization. IEEE Access.

Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, et al (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc Roy Soc Lond Ser A Math Phys Eng Sci 454:903–995.

Jiang D, Zhang Y, Hu X, Zeng Y, Tan J, Shao D (2004) Progress in developing an ANN model for air pollution index forecast. Atmos Environ 38:7055–7064

Jiang P, Li C, Li R, Yang H (2019) An innovative hybrid air pollution early-warning system based on pollutants forecasting and Extenics evaluation. Knowl-Based Syst 164:174–192

Jin W, Dong S, Yu C, Luo Q (2022) A data-driven hybrid ensemble AI model for COVID-19 infection forecast using multiple neural networks and reinforced learning. Comput Biol Med 146:105560

Jumaah HJ, Ameen MH, Kalantar B, Rizeei HM, Jumaah SJ (2019) Air quality index prediction using IDW geostatistical technique and

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

OLS-based GIS technique in Kuala Lumpur, Malaysia. Geomat Nat Haz Risk 10:2185–2199

Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. Atmos Environ 38:2895–2907

Kamińska JA (2019) A random forest partition model for predicting NO2 concentrations from traffic flow and meteorological conditions. Sci Total Environ 651:475–483

Kang J, Zou X, Tan J, Li J, Karimian H (2023) Short-Term PM2. 5 concentration changes prediction: a comparison of meteorological and historical data. Sustainability 15:11408.

Karimian H, Li Y, Chen Y, Wang Z (2023) Evaluation of different machine learning approaches and aerosol optical depth in PM2. 5 prediction. Environ Res 216:114465.

Kolehmainen M, Martikainen H, Ruuskanen J (2001) Neural networks and periodic components used in air quality forecasting. Atmos Environ 35:815–825

Kottur SV, Mantha S (2015) An integrated model using Artificial Neural Network (ANN) and Kriging for forecasting air pollutants using meteorological data. Int J Adv Res Comput Commun Eng 4:146–152.

Kristiani E, Kuo T-Y, Yang C-T, Pai K-C, Huang C-Y, Nguyen KLP (2021) PM2. 5 forecasting model using a combination of deep learning and statistical feature selection. IEEE Access 9:68573–68582.

Kumar A, Goyal P (2011) Forecasting of air quality in Delhi using principal component regression technique. Atmos Pollut Res 2:436–444

Kumar A, Goyal P (2013) Forecasting of air quality index in Delhi using neural network based on principal component analysis. Pure Appl Geophys 170:711–722

Kumar K, Yadav A, Singh M, Hassan H, Jain V (2004) Forecasting daily maximum surface ozone concentrations in Brunei Darussalam—an ARIMA modeling approach. J Air Waste Manag Assoc 54:809–814

Lai W-I, Chen Y-Y, Sun J-H (2022) Ensemble machine learning model for accurate air pollution detection using commercial gas sensors. Sensors 22:4393

Lee KS, Geem ZW (2005) A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice. Comput Methods Appl Mech Eng 194:3902–3933

Lei KS, Wan F (2010) Pre-processing for missing data: a hybrid approach to air pollution prediction in Macau. In: 2010 IEEE international conference on automation and logistics. IEEE, New York, pp. 418–422.

Li Y, Tao Y (2017) PM10 concentration forecast based on wavelet support vector machine. In: 2017 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC). IEEE, New York, pp 383–386.

Li C, Zhu Z (2018) Research and application of a novel hybrid air quality early-warning system: a case study in China. Sci Total Environ 626:1421–1438

Li J, Shao X, Zhao H (2018) An online method based on random forest for air pollutant concentration forecasting. In: 37th Chinese Control Conference (CCC). IEEE 2018:9641–9648

Li R, Dong Y, Zhu Z, Li C, Yang H (2019) A dynamic evaluation framework for ambient air pollution monitoring. Appl Math Model 65:52–71

Li H, Jiang Z, Shi Z, Han Y, Yu C, Mi X (2022a) Wind-speed prediction model based on variational mode decomposition, temporal convolutional network, and sequential triplet loss. Sustain Energy Technol Assess 52:101980

Li Y, Guo J-e, Sun S, Li J, Wang S, Zhang C (2022b) Air quality forecasting with artificial intelligence techniques: A scientometric and content analysis. Environ Model Softw 149:105329.

Li Y, Xue L, Tao Y, Li Y, Wu Y, Liao Q, et al. (2023) Exploring the contributions of major emission sources to PM2. 5 and attributable health burdens in China. Environ Pollut 322:121177.

Liang T, Xie G, Mi D, Jiang W, Xu G (2020) PM2. 5 concentration forecasting based on data preprocessing strategy and LSTM neural network. Int J Mach Learn Comput 10:729–734.

Liang K, Meng L, Liu M, Liu Y, Tu W, Wang S, et al (2022) Reasoning over different types of knowledge graphs: static, temporal and multi-modal. arXiv preprint arXiv:2212.05767.

Liang K, Liu Y, Zhou S, Tu W, Wen Y, Yang X, et al. (2023a) Knowledge graph contrastive learning based on relation-symmetrical structure. IEEE Trans Knowl Data Eng, pp 1–12.

Liang K, Meng L, Liu M, Liu Y, Tu W, Wang S, et al (2023b) Learn from relational correlations and periodic events for temporal knowledge graph reasoning. In: Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval, pp. 1559–1568.

Liang K, Zhou S, Liu Y, Meng L, Liu M, Liu X (2023c) Structure guided multi-modal pre-trained transformer for knowledge graph reasoning. arXiv preprint arXiv:2307.03591

Lin X, Wang H, Guo J, Mei G (2022) A deep learning approach using graph neural networks for anomaly detection in air quality data considering spatiotemporal correlations. IEEE Access 10:94074–94088

Lin Y, Mago N, Gao Y, Li Y, Chiang Y-Y, Shahabi C, et al (2018) Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning. In: Proceedings of the 26th ACM SIGSPATIAL international conference on advances in geographic information systems, pp 359–368.

Liu Y, Guo H, Mao G, Yang P (2008) A Bayesian hierarchical model for urban air quality prediction under uncertainty. Atmos Environ 42:8464–8469

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

Liu H, Yu C, Yu C (2021a) A new hybrid model based on secondary decomposition, reinforcement learning and SRU network for wind turbine gearbox oil temperature forecasting. Measurement 178:109347

Liu X, Qin M, He Y, Mi X, Yu C (2021b) A new multi-data-driven spatiotemporal PM2. 5 forecasting model based on an ensemble graph reinforcement learning convolutional network. Atmos Pollut Res 12:101197.

Lu GY, Wong DW (2008) An adaptive inverse-distance weighting spatial interpolation technique. Comput Geosci 34:1044–1055

Luo H, Wang D, Yue C, Liu Y, Guo H (2018) Research and application of a novel hybrid decomposition-ensemble learning paradigm with error correction for daily PM10 forecasting. Atmos Res 201:34–45

Ly H-B, Le LM, Phi LV, Phan V-H, Tran VQ, Pham BT et al (2019) Development of an AI model to measure traffic air pollution from multisensor and weather data. Sensors 19:4941

Lyu B, Zhang Y, Hu Y (2017) Improving PM2. 5 air quality model forecasts in China using a bias-correction framework. Atmosphere 8:147.

Ma J, Ding Y, Cheng JC, Jiang F, Wan Z (2019a) A temporal-spatial interpolation and extrapolation method based on geographic Long Short-Term Memory neural network for PM2. 5. J Clean Prod 237:117729.

Ma J, Ding Y, Gan VJ, Lin C, Wan Z (2019b) Spatiotemporal prediction of PM2. 5 concentrations at different time granularities using IDW-BLSTM. IEEE Access 7:107897–107907.

Masood A, Ahmad K (2023) Prediction of PM2. 5 concentrations using soft computing techniques for the megacity Delhi, India. Stochastic Environ Res Risk Assess 37:625–638.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Méndez M, Merayo MG, Núñez M (2023) Machine learning algorithms to forecast air quality: a survey. Artif Intell Rev, pp 1–36.

Mi X, Yu C, Liu X, Yan G, Yu F, Shang P (2022) A dynamic ensemble deep deterministic policy gradient recursive network for spatiotemporal traffic speed forecasting in an urban road network. Digital Signal Process 129:103643

Mittal A, Bhardwaj R (2011) Prediction of daily air pollution using wavelet decomposition and Adaptive Network-based fuzzy inference system. Int J Environ Sci 2:174–184

Najjar YS (2011) Gaseous pollutants formation and their harmful effects on health and environment. Innovative Energy Policies 1:1–9

Nieto PG, García-Gonzalo E, Sánchez AB, Miranda AR (2018) Air quality modeling using the PSO-SVM-based approach, MLP neural network, and M5 model tree in the metropolitan area of Oviedo (Northern Spain). Environ Model Assess 23:229–247

Niska H, Hiltunen T, Karppinen A, Ruuskanen J, Kolehmainen M (2004) Evolving the neural network model for forecasting air pollution time series. Eng Appl Artif Intell 17:159–167

Niu M, Wang Y, Sun S, Li Y (2016) A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2. 5 concentration forecasting. Atmos Environ 134:168–180.

Niu M, Gan K, Sun S, Li F (2017) Application of decomposition-ensemble learning paradigm with phase space reconstruction for day-ahead PM2. 5 concentration forecasting. J Environ Manage 196:110–118.

Nuhu SN, Duan Z, Li Y (2023) PM2. 5 prediction method using back propagation neural network. In: International conference on internet of things and machine learning (IoTML 2022). 12640. SPIE, pp 434–439.

Ojagh S, Cauteruccio F, Terracina G, Liang SH (2021) Enhanced air quality prediction by edge-based spatiotemporal data preprocessing. Comput Electr Eng 96:107572

Osowski S, Garanty K (2007) Forecasting of the daily meteorological pollution using wavelets and support vector machine. Eng Appl Artif Intell 20:745–755

Percival DB, Walden AT (2000) Wavelet methods for time series analysis. Vol 4: Cambridge University Press, Cambridge

Pfeiffer H, Baumbach G, Sarachaga-Ruiz L, Kleanthous S, Poulida O, Beyaz E (2009) Neural modelling of the spatial distribution of air pollutants. Atmos Environ 43:3289–3297

Plaia A, Bondi A (2006) Single imputation method of missing values in environmental pollution data sets. Atmos Environ 40:7316–7330

Prakash A, Kumar U, Kumar K, Jain V (2011) A wavelet-based neural network model to predict ambient air pollutants' concentration. Environ Model Assess 16:503–517

Prihatno AT, Nurcahyanto H, Ahmed MF, Rahman MH, Alam MM, Jang YM (2021) Forecasting PM2. 5 concentration using a single-dense layer BiLSTM method. Electronics 10:1808.

Qi X, Chen G, Li Y, Cheng X, Li C (2019) Applying neural-network-based machine learning to additive manufacturing: current applications, challenges, and future perspectives. Engineering 5(4): 721–729.

Qin S, Liu F, Wang J, Sun B (2014) Analysis and forecasting of the particulate matter (PM) concentration levels over four major cities of China using hybrid models. Atmos Environ 98:665–675

Quinteros ME, Lu S, Blazquez C, Cárdenas-R JP, Ossa X, Delgado-Saborit J-M et al (2019) Use of data imputation tools to reconstruct incomplete air quality datasets: a case-study in Temuco. Chile Atmos Environ 200:40–49

Richman JS, Moorman JR (2000) Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol-Heart Circulatory Physiol 278:H2039–H2049

Rybarczyk Y, Zalakeviciute R (2018) Machine learning approaches for outdoor air quality modelling: a systematic review. Appl Sci 8:2570

Sah D, Verma PK, Kumari KM, Lakhani A (2023) Characterisation, sources and health risk of heavy metals in PM2. 5 in Agra, India. Exposure Health 15:585–596.

Sahu SK, Mardia KV (2005) A Bayesian kriged Kalman model for short-term forecasting of air pollution levels. J Roy Stat Soc: Ser C (appl Stat) 54:223–244

Sayed GI, Hassanien AE, Azar AT (2019) Feature selection via a novel chaotic crow search algorithm. Neural Comput Appl 31:171–188

Sayeed A, Choi Y, Pouyaei A, Lops Y, Jung J, Salman AK (2022) CNN-based model for the spatial imputation (CMSI version 1.0) of in-situ ozone and PM2. 5 measurements. Atmos Environ 289:119348.

Shang P, Liu X, Yu C, Yan G, Xiang Q, Mi X (2022) A new ensemble deep graph reinforcement learning network for spatio-temporal traffic volume forecasting in a freeway network. Digital Signal Process 123:103419

Sharma E, Deo RC, Prasad R, Parisi AV (2020) A hybrid air quality early-warning framework: an hourly forecasting model with online sequential extreme learning machines and empirical mode decomposition algorithms. Sci Total Environ 709:135934

Shi Z, Li J, Jiang Z, Li H, Yu C, Mi X (2024) WGformer: a Weibull-Gaussian Informer based model for wind speed prediction. Eng Appl Artif Intell 131:107891

Siew LY, Chin LY, Wee PMJ (2008) ARIMA and integrated ARFIMA models for forecasting air pollution index in Shah Alam. Selangor Malays J Anal Sci 12:257–263

Siwek K, Osowski S (2012) Improving the accuracy of prediction of PM10 pollution by the wavelet transformation and an ensemble of neural predictors. Eng Appl Artif Intell 25:1246–1258

Song C, Fu X (2020) Research on different weight combination in air quality forecasting models. J Clean Prod 261:121169

Sun W, Li Z (2020) Hourly PM2. 5 concentration forecasting based on mode decomposition-recombination technique and ensemble learning approach in severe haze episodes of China. J Clean Prod 263:121442.

Sun W, Sun J (2016) Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. J Environ Manage 188:144–152.

Sun W, Zhang H, Palazoglu A, Singh A, Zhang W, Liu S (2013) Prediction of 24-hour-average PM2. 5 concentrations using a hidden Markov model with different emission distributions in Northern California. Sci Total Environ 443:93–103.

Taghavi M, Ghanizadeh G, Ghasemi M, Fassò A, Hoek G, Hushmandi K, et al (2023) Application of functional principal component analysis in the spatiotemporal land-use regression modeling of PM2. 5. Atmosphere 14: 926.

Tamas W, Notton G, Paoli C, Nivet M-L, Voyant C (2016) Hybridization of air quality forecasting models using machine learning and clustering: an original approach to detect pollutant peaks. Aerosol Air Qual Res 16:405–416

Teng Y, Huang X, Ye S, Li Y (2018) Prediction of particulate matter concentration in Chengdu based on improved differential evolution algorithm and BP neural network model. In: 2018 IEEE 3rd international conference on cloud computing and big data analysis (ICCCBDA).

Tian Z, Gai M (2022) New PM2. 5 forecasting system based on combined neural network and an improved multi-objective optimization algorithm: taking the economic belt surrounding the Bohai Sea as an example. J Clean Prod 375:134048.

Tong Y, Yu Y, Hu X, He L (2015) Performance analysis of different kriging interpolation methods based on air quality index in wuhan. In: 2015 sixth international conference on intelligent

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

control and information processing (ICICIP). IEEE, New York, pp 331–335.

Torres ME, Colominas MA, Schlotthauer G, Flandrin P (2011) A complete ensemble empirical mode decomposition with adaptive noise. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE , New York, 4144–4147

Udristioiu MT, Mghouchi YE, Yildizhan H (2023) Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning. J Clean Prod 421:138496

Wahid H, Ha QP, Duc H, Azzi M (2013) Neural network-based metamodelling approach for estimating spatial distribution of air pollutant levels. Appl Soft Comput 13:4087–4096

Wang D, Wei S, Luo H, Yue C, Grunder O (2017a) A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Sci Total Environ 580:719–733

Wang J, Niu T, Wang R (2017b) Research and application of an air quality early warning system based on a modified least squares support vector machine and a cloud model. Int J Environ Res Public Health 14:249

Wang D, Liu Y, Luo H, Yue C, Cheng S (2017c) Day-ahead PM2. 5 concentration forecasting using WT-VMD based decomposition method and back propagation neural network improved by differential evolution. Int J Environ Res Public Health 14:764.

Wang P, Zhang H, Qin Z, Zhang G (2017d) A novel hybrid-Garch model based on ARIMA and SVM for PM2. 5 concentrations forecasting. Atmos Pollut Res 8:850–860.

Wang J, Li H, Lu H (2018) Application of a novel early warning system based on fuzzy time series in urban air quality forecasting in China. Appl Soft Comput 71:783–799

Wang L, Shi X, Li M, Chen G, Ge H, Lee H, et al (2005) Applications of PSO algorithm and OIF Elman neural network to assessment and forecasting for atmospheric quality. adaptive and natural computing algorithms. Springer, Cham, pp 251–254.

Wang J, Du P, Hao Y, Ma X, Niu T, Yang W (2020a) An innovative hybrid model based on outlier detection and correction algorithm and heuristic intelligent optimization algorithm for daily air quality index forecasting. J Environ Manage 255:109855

Wang Y, Wang J, Li Z (2020b) A novel hybrid air quality early-warning system based on phase-space reconstruction and multi-objective optimization: a case study in China. J Clean Prod, p 121027.

Wang J, Wang R, Li Z (2022) A combined forecasting system based on multi-objective optimization and feature extraction strategy for hourly PM2. 5 concentration. Appl Soft Comput 114:108034.

Wei Y, Jang-Jaccard J, Xu W, Sabrina F, Camtepe S, Boulic M (2023) LSTM-autoencoder-based anomaly detection for indoor air quality time-series data. IEEE Sens J 23:3787–3800

Wei P, Xie S, Huang L, Liu L, Tang Y, Zhang Y, et al (2022) Spatial interpolation of PM2. 5 concentrations during holidays in south-central China considering multiple factors. Atmos Pollut Res 13:101480.

Wen W, Hua T, Liu L, Liu X, Ma X, Shen S, et al (2023) Oxidative potential characterization of different PM2. 5 sources and components in Beijing and the surrounding region. Int J Environ Res Public Health 20:5109.

Westerlund J, Urbain J-P, Bonilla J (2014) Application of air quality combination forecasting to Bogota. Atmos Environ 89:22–28

Wong P-Y, Su H-J, Lung S-CC, Wu C-D (2023) An ensemble mixed spatial model in estimating long-term and diurnal variations of PM2. 5 in Taiwan. Sci Total Environ 866:161336.

Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv Adapt Data Anal 1:1–41

Wu Q, Lin H (2019a) Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network. Sustain Cities Soc 50:101657

Wu Q, Lin H (2019b) A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. Sci Total Environ 683:808–821

Wu H, Tang X, Wang Z, Wu L, Lu M, Wei L et al (2018) Probabilistic automatic outlier detection for surface air quality measurements from the China national environmental monitoring network. Adv Atmos Sci 35:1522–1532

Wu Z, Ma C, Shi X, Wu L, Dong Y, Stojmenovic M (2022) Imputing missing indoor air quality data with inverse mapping generative adversarial network. Build Environ 215:108896

Xu Y, Du P, Wang J (2017a) Research and application of a hybrid model based on dynamic fuzzy synthetic evaluation for establishing air quality forecasting and early warning system: a case study in China. Environ Pollut 223:435–448

Xu Y, Yang W, Wang J (2017b) Air quality early-warning system for cities in China. Atmos Environ 148:239–257

Xu Y, Liu H, Duan Z (2020) A novel hybrid model for multi-step daily AQI forecasting driven by air pollution big data. Air Qual Atmos Health 13:197–207

Xu X, Ren W (2019) Application of a hybrid model based on echo state network and improved particle swarm optimization in PM2. 5 concentration forecasting: a case study of Beijing, China. Sustainability 11:3096.

Yan G, Bai Y, Yu C, Yu C (2022) A Multi-factor driven model for locomotive axle temperature prediction based on multi-stage feature engineering and deep learning framework. Machines 10:759

Yang H-C, Yang M-C, Wong G-W, Chen MC (2023a) Extreme event discovery with self-attention for PM2. 5 anomaly prediction. IEEE Intell Syst 38:36–45

Yang H, Wang W, Li G (2023b) Prediction method of PM2. 5 concentration based on decomposition and integration. Measurement 216:112954.

Yang Z, Wang J (2017) A new air quality monitoring and early warning system: air quality assessment and air pollutant concentration prediction. Environ Res 158:105–117

Yeh J-R, Shieh J-S, Huang NE (2010) Complementary ensemble empirical mode decomposition: a novel noise enhanced data analysis method. Adv Adapt Data Anal 2:135–156

Yildirim Y, Bayramoglu M (2006) Adaptive neuro-fuzzy based modelling for prediction of air pollution daily levels in city of Zonguldak. Chemosphere 63:1575–1582

Yıldız AY, Koç E, Koç A (2022) Multivariate time series imputation with transformers. IEEE Signal Process Lett 29:2517–2521

Yu C, Yan G, Yu C, Liu X, Mi X (2024) MRIformer: a multi-resolution interactive transformer for wind speed multi-step prediction. Inf Sci 661:120150

Yu Y, Li H, Sun S, Li Y (2022) PM2. 5 concentration forecasting through a novel multi-scale ensemble learning approach considering intercity synergy. Sustain Cities Soc 85:104049.

Yu C, Yan G, Ruan K, Liu X, Yu C, Mi X (2023a) An ensemble convolutional reinforcement learning gate network for metro station PM2. 5 forecasting. Stochastic Environ Res Risk Assess, pp 1–16.

Yu C, Yan G, Yu C, Mi X (2023b) Attention mechanism is useful in spatio-temporal wind speed prediction: evidence from China. Appl Soft Comput 148:110864

Yuan E, Yang G (2023) SA–EMD–LSTM: A novel hybrid method for long-term prediction of classroom PM2. 5 concentration. Expert Syst Appl, 120670.

Yuan Z, Gao S, Wang Y, Li J, Hou C, Guo L (2023) Prediction of PM2. 5 time series by seasonal trend decomposition-based dendritic neuron model. Neural Comput Appl, pp 1–17.

Yuan W, Wang K, Bo X, Tang L, Wu J (2019) A novel multi-factor & multi-scale method for PM2. 5 concentration forecasting. Environ Pollut 255:113187.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Zhai B, Chen J (2018) Development of a stacked ensemble model for forecasting and analyzing daily average PM2. 5 concentrations in Beijing, China. Sci Total Environ 635:644–658.

Zhang C, Ni Z, Ni L, Tang N (2016) Feature selection method based on multi-fractal dimension and harmony search algorithm and its application. Int J Syst Sci 47:3476–3486

Zhang L, Xu L, Jiang M, He P (2023) A novel hybrid ensemble model for hourly PM2. 5 concentration forecasting. Int J Environ Sci Technol 20:219–230.

Zhang C, Wang X, Chen S, Zou L, Zhang X, Tang C (2019) A study on daily PM2. 5 concentrations in Hong Kong using the EMD-based MFDFA method. Physica A Stat Mech Appl 530:121182.

Zhao F, Li W (2019) A combined model based on feature selection and woa for pm2. 5 concentration forecasting. Atmosphere 10:223.

Zheng J, Wang C, Liang Y, Liao Q, Li Z, Wang B (2022) Deeppipe: a deep-learning method for anomaly detection of multi-product pipelines. Energy 259:125025

Zheng Y, Yi X, Li M, Li R, Shan Z, Chang E, et al (2015) Forecasting fine-grained air quality based on big data. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 2267–2276.

Zhou Y, Yang Z, Sun Q, Yu C, Yu C (2023) An artificial intelligence model based on multi-step feature engineering and deep attention network for optical network performance monitoring. Optik 273:170443

Zhu S, Lian X, Liu H, Hu J, Wang Y, Che J (2017) Daily air quality index forecasting with hybrid models: A case in China. Environ Pollut 231:1232–1244

Zhu J, Wu P, Chen H, Zhou L, Tao Z (2018a) A hybrid forecasting approach to air quality time series based on endpoint condition and combined forecasting model. Int J Environ Res Public Health 15:1941

Zhu S, Lian X, Wei L, Che J, Shen X, Yang L, et al (2018b) PM2. 5 forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors. Atmos Environ 183:20–32.

Zhu S, Yang L, Wang W, Liu X, Lu M, Shen X (2018c) Optimal-combined model for air quality index forecasting: 5 cities in North China. Environ Pollut 243:842–850

Zhu S, Qiu X, Yin Y, Fang M, Liu X, Zhao X et al (2019a) Two-step-hybrid model based on data preprocessing and intelligent optimization algorithms (CS and GWO) for NO2 and SO2 forecasting. Atmos Pollut Res 10:1326–1335

Zhu W, Wang M, Zhang B (2019b) The effects of urbanization on PM2. 5 concentrations in China's Yangtze River Economic Belt: New evidence from spatial econometric analysis. J Clean Prod 239:118065

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com