Pol. J. Environ. Stud. Vol. 34, No. 3 (2025), 3063-3076
DOI: *10.15244/pjoes/188850* ONLINE PUBLICATION DATE: 2024-09-09

Original Research

A Novel Hybrid Forecasting Model for $PM_{2.5}$ Concentration Based on Optimized VMD Decomposition, Multi-Objective Feature Selection, and Error Correction

Chenhao Cai[1]*, Leyao Zhang[2], Jianguo Zhou[1], Luming Zhou[1]

[1]Department of Economics and Management, North China Electric Power University, 689 Huadian Road, Baoding 071000, China
[2]National Engineering Research Center for E-Learning, Central China Normal University, Luoyu Road, Wuhan , 430079 , Hubei, China

Received: 24 April 2024
Accepted: 17 May 2024

Abstract

Accurate prediction of $PM_{2.5}$ concentration is crucial for public health and environmental protection. This paper develops a novel forecasting model that combines optimized signal decomposition with multi-objective feature selection techniques and error correction to enhance the accuracy of $PM_{2.5}$ concentration predictions. Initially, the RIME algorithm is employed to precisely set the parameters of Variational Mode Decomposition (VMD), which decomposes the raw $PM_{2.5}$ data into high, medium, and low-frequency components based on sample entropy values. Subsequently, a multi-objective feature selection approach is utilized to identify key feature subsets that significantly influence each frequency domain component. Finally, an optimized Informer model is deployed for comprehensive forecasting, complemented by an error correction mechanism to obtain the final $PM_{2.5}$ concentration predictions. Experimental results indicate that the optimized decomposition effectively extracts key information from the data, reducing prediction complexity. The multi-objective feature selection approach provides superior identification of feature subsets compared to traditional single-objective methods. The enhanced Informer model, coupled with error correction, significantly improves the model's accuracy and robustness.

Keywords: $PM_{2.5}$ forecasting, VMD, multi-objective feature selection, RIME algorithm, informer, error correction

*e- mail: caichenhao@126.com;
Tel.: +86-147-7491-7117

Introduction

As industrialization accelerates, human society's demand for energy continues to rise. However, this increase in energy demand is accompanied by worsening air pollution issues, particularly the rising concentration of $PM_{2.5}$ (fine particulate matter), posing a significant threat to public health. Due to its small particle size, $PM_{2.5}$ can penetrate deep into the lungs and even enter the bloodstream, closely associated with various respiratory and cardiovascular diseases [1]. Consequently, monitoring and predicting $PM_{2.5}$ concentrations has become a critical task in urban environmental management. Accurate prediction of $PM_{2.5}$ concentrations can greatly assist in environmental quality assessment, the establishment and improvement of public health warning systems, and the formulation of strategies for air pollution control [2, 3].

Recent studies have highlighted the significant constraints imposed by climate change and smog pollution on high-quality economic development in China. The urgency of addressing these environmental challenges is underscored by the adverse effects on public health and economic stability, particularly in urban areas where $PM_{2.5}$ levels are critically high. According to recent findings, low-carbon city pilot (LCCP) policies have shown potential for synergistically governing carbon and smog emissions, thus aiding cities in achieving dual environmental and economic benefits [4]. These policies have been instrumental in reducing energy consumption intensities and optimizing industrial structures, which are crucial steps toward sustainable urban development.

Furthermore, the development of public transportation systems has been identified as a crucial strategy for improving urban air quality. The "Transit Metropolis" construction demonstration project, a significant initiative in China, has been effective in reducing private car ownership and upgrading industrial structures, thereby contributing positively to air quality [5]. This policy's success provides a compelling context for our model's application. The significance of this research is further magnified by the pressing need to implement effective air quality management practices in light of the ongoing challenges posed by rapid urbanization and industrialization in major Chinese cities. By providing a robust tool for predicting $PM_{2.5}$ concentrations, this paper assists policymakers and urban planners in crafting strategies that align with sustainability goals.

Current research indicates that $PM_{2.5}$ concentration prediction techniques primarily fall into two categories: models based on physical laws and data-driven intelligent algorithms. Physical models often require researchers to have an extensive background in atmospheric science, understanding the chemical and physical evolution processes and the trans-regional migration of the atmosphere [6]. These models use meteorological data and environmental information to make predictions through complex numerical calculations. Although suitable for long-term forecasting, these models involve cumbersome computational processes and require high-quality data [7]. In contrast, data-driven intelligent algorithms, by mining statistical information and features from historical data, can establish a mapping relationship between historical data and forecasting targets. These methods can effectively predict $PM_{2.5}$ concentrations even with incomplete historical data and are more cost-effective. Traditional statistical methods like Markov chains [8] and Autoregressive Integrated Moving Average (ARIMA) models [9], as well as machine learning approaches such as Support Vector Machines [10], have been widely adopted in the field of $PM_{2.5}$ prediction.

The Transformer, known for its capability to learn complex patterns and features from time series and its ability to capture long-distance dependencies, has been applied in time series forecasting [11]. The Informer model introduces the ProbSparse self-attention mechanism, which reduces computational complexity by calculating attention only for a subset of key elements. Autoregressive sparsification further reduces the attention scores that need to be calculated, focusing only on those most likely to be important. This reduces the computational burden and improves the model's efficiency and performance in handling long sequences. The Informer model optimizes the encoding of time features, ensuring the model better understands and utilizes temporal information, which is particularly important for time series forecasting. This encoding approach helps the model capture seasonal and trend dynamics in the time series [12]. With increasing demands for prediction accuracy, more research is exploring hybrid forecasting models to more accurately capture the temporal characteristics of $PM_{2.5}$ concentration variations and enhance prediction accuracy.

In the data preprocessing stage, to mitigate the effects of the nonlinearity, volatility, and instability of $PM_{2.5}$ concentration data on the forecasting models, researchers often employ decomposition algorithms. These algorithms can break down complex raw data into several more stable subsequences, thereby reducing the non-stationary characteristics of the data. Widely used decomposition algorithms include Empirical Mode Decomposition (EMD), Empirical Wavelet Transform (EWT), and Variational Mode Decomposition (VMD). For instance, some studies have used EMD to split $PM_{2.5}$ data into multiple intrinsic mode components and residue components, which are then sequentially fed into a GRU neural network for training [13]. The Empirical Wavelet Transform (EWT) algorithm can adaptively partition the Fourier spectrum and select appropriate wavelet filter banks. Research has been conducted using EWT to obtain several $PM_{2.5}$ time series components and then construct predictors based on the Echo State Network (ESN) for each decomposed sublayer within each cluster group to perform multi-step forecasting calculations

and form the final predictions [14]. Although EMD and EWT have demonstrated certain effects in enhancing prediction accuracy, they also have issues with endpoint effects and mode mixing. To overcome these issues, some studies have used VMD for data decomposition, inputting each decomposed subsequence (including the residual sequence) into a GRU, and then calculating the prediction loss of the subsequences [15], achieving relatively good results. However, the selection of VMD parameters remains a challenge. In other forecasting fields, some studies have adjusted VMD parameters through optimization algorithms and combined them with Convolutional Neural Networks and Bidirectional Long Short-Term Memory networks (CNN-BILSTM) to predict decomposed components, proving the effectiveness of this method in parameter selection [16].

In the feature selection stage of $PM_{2.5}$ concentration forecasting, selecting highly correlated and significant features is crucial for enhancing the model's prediction speed and accuracy. There are numerous methods for feature selection, including Mutual Information (MI) and Random Forest (RF) [13], which have proven effective in identifying features strongly correlated with $PM_{2.5}$ concentrations. For example, some studies have used a method combining Variance Inflation Factor and Mutual Information (VIF-MI) to select features, demonstrating that this method can effectively reduce model complexity and enhance prediction efficiency [17]. While feature selection can improve prediction accuracy [18], although single-objective feature selection methods are effective in reducing redundancy and error, they often overlook the interrelationships among multiple objectives, which may not achieve the optimal overall effect. To this end, in the field of wind speed forecasting, some research has proposed a method based on K-means clustering and the non-dominated sorting differential evolution algorithm (FWNSDEC), a multi-objective feature selection method that integrates the advantages of different algorithms and shows better prediction performance than single feature selection methods [19].

After a systematic review of the literature, several research gaps seem to await further supplementation and exploration. Firstly, while previous studies have successfully applied various decomposition techniques to predict $PM_{2.5}$ concentrations, common algorithms like EMD and EWT still have issues with endpoint effects and mode mixing, and parameter selection for various decomposition algorithms remains a challenge; single-objective feature selection has been applied in the field of $PM2_{2.5}$ concentration prediction, but integrating the interrelationships among multiple objectives to further reduce redundancy and error could potentially yield better prediction results; moreover, the error series generated in predictions also contains some potential features, and in some prediction fields, error correction has been proven to improve prediction accuracy to some extent [20, 21], but existing $PM_{2.5}$ concentration predictions have paid little attention to post-model processing, especially error correction model with decomposition (ECD). Based on the above analysis, this paper proposes an integrated forecasting framework. The first step of the framework uses the RIME algorithm to select parameters for Variational Mode Decomposition (VMD) to stabilize the fluctuations in $PM_{2.5}$ data. Then, by integrating similar components using sample entropy, high, medium, and low-frequency data components are formed. Next, an optimized multi-objective feature selection algorithm (MOFS) is used, combining the functions of filters and wrappers, to precisely select the best input feature set. Finally, an optimized Informer model is used to predict different frequency data components and error correction is combined to obtain the final prediction results. The specific research contributions include:

1) In the data preprocessing stage, the RIME algorithm is innovatively used to select the parameters for Variational Mode Decomposition (VMD), enhancing the effectiveness of VMD decomposition and reducing the volatility of the raw data. This is crucial as the RIME algorithm simulates the movement and interaction of frostbite particles, exhibiting high exploratory capabilities, especially important when dealing with long-duration data containing complex patterns and trends, such as $PM_{2.5}$ concentrations. The forward greedy selection mechanism ensures that the algorithm always chooses solutions with higher fitness during updates, which helps continuously optimize the quality of results in long-term series analysis. Additionally, this selection mechanism aids in rapid convergence, particularly when dealing with large datasets, effectively reducing computational time and resource consumption. Unlike traditional methods that often struggle with parameter selection [13, 16], the RIME-enhanced VMD stabilizes data fluctuations and reduces mode mixing issues inherent in techniques like EMD and EWT. This approach has not been observed in previous $PM_{2.5}$ concentration forecasting literature.

2) For the first time, Multi-Objective Feature Selection (MOFS) is introduced into the field of $PM_{2.5}$ concentration prediction. MOFS optimizes the objective functions of both the designed filter and wrapper, thus integrating the advantages of these two different feature selection criteria. The Minimum Redundancy Maximum Relevance (mRMR) and Time Series Cross-Validation (TSCV) are used as the filter and wrapper objective functions, respectively. This approach selects the best feature inputs from multiple meteorological factors and historical data, enhancing the model's prediction speed and accuracy. By employing MOFS, our model effectively addresses the limitations seen in traditional methods like Mutual Information and Random Forest, which do not consider the multivariate interrelationships crucial for understanding complex environmental data [13, 18].

3) Innovatively, an optimized Informer model is introduced into the field of $PM_{2.5}$ concentration prediction. This paper compares it with common

models such as LSTM, CNN, and Bi-LSTM. Studies [22, 23] have utilized LSTM and CNN for short-term $PM_{2.5}$ forecasting, demonstrating their effectiveness in capturing linear relationships but often struggling with complex nonlinear patterns and long-term dependencies that are critical in environmental datasets. Similarly, Bi-LSTM models have improved upon LSTM by processing data in both forward and backward states but still face limitations in handling very long sequences and high computational loads. The ProbSparse self-attention mechanism and time feature encoding of the Informer model ensure that the model can understand and utilize time information effectively while reducing computational complexity, helping to capture the dynamic changes in the $PM_{2.5}$ concentration time series. Experimental results show that the model used in this study achieved the best results across all three evaluation metrics.

4) Error correction techniques are seldom applied in the field of $PM_{2.5}$ concentration prediction, particularly the error correction model with decomposition (ECD), which is introduced into the hybrid model proposed in this paper to further enhance prediction accuracy.

5) Aligning with global and national initiatives to combat air pollution, our model's capabilities support the objectives of policies like China's "Transit Metropolis" policy and LCCP, which aim to achieve dual environmental and economic benefits [4, 5]. Our research not only provides technological insights but also offers practical tools for policymakers to assess and refine strategies aimed at reducing $PM_{2.5}$ levels and their associated health risks. The hybrid model proposed in this paper considers potential issues to various stages of prediction, combining the advantages of each module to accurately predict $PM_{2.5}$ concentrations. This enriches the technological tools in the field, providing valuable quantitative references for environmental quality assessment, the establishment and improvement of public health warning systems, and the formulation of strategies for air pollution control.

Materials and Methods

Rime Optimization Algorithm

The Rime Optimization Algorithm was proposed by Hang Su in February 2023. It is inspired by the growth mechanism of rime frost. The algorithm simulates the movement of soft rime frost particles and introduces a frost impulse search strategy for algorithmic search. The components of the algorithm are as follows:

1) Initialization of rime frost cluster. Where i represents the index of a RIME agent and j the sequence number of rime frost particles, R denotes the population, and $F(S_i)$ is used to represent the growth state of each individual, i.e., the fitness value of an agent in metaheuristic algorithms.

$$R = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_i \end{bmatrix}; \quad S_i = [x_{i1} x_{i2} \cdots x_{ij}] \tag{1}$$

$$R = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} \\ x_{21} & x_{22} & \cdots & x_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} \end{bmatrix} \tag{2}$$

2) Soft rime frost search strategy. Reflecting the five movement characteristics of rime frost particles, this simplifies the simulation of each rime frost particle's condensation process and calculates the position of the rime frost particles as shown in Equation (3).

$$R_{ij}^{new} = R_{best,j} + r_1 \cdot \cos\theta \cdot \beta \cdot (h \cdot (Ub_{ij} - Lb_{ij}) + Lb_{ij}), r_2 < E \tag{3}$$

Where R_{ij}^{new} is the new position of the updated particle. $R_{best,j}$ is the particle of the best rime frost agent R within the rime frost population j. The parameter r is a random number within the range (1, 1) and r_1 controls the direction of particle movement. cosθ changes with the iteration number as shown in Equation (4). β is the environmental factor, which follows the iteration number to simulate the impact of the external environment and ensures the convergence of the algorithm as shown in Equation (5). h is the adhesion factor, a random number within the range (0, 1), controlling the distance between the centers of two rime frost particles. Ub_{ij} and Lb_{ij} define the upper and lower limits of the escape space, restricting the particle's effective movement area. E is the adhesion coefficient, influencing the probability of agent cohesion and increasing with the iteration number, as shown in Equation (6).

$$\theta = \pi \cdot \frac{t}{10 \cdot T} \tag{4}$$

Where t,T represent the current iteration number and the maximum iteration number, respectively.

$$\beta = 1 - [\frac{w \cdot t}{T}] / w \tag{5}$$

Where the default value of w is 5, it is used to control the number of segments of β.

$$E = \sqrt{(t / T)} \tag{6}$$

3) Hard-rime puncture mechanism. Inspired by the piercing phenomenon, a hard-rime puncture mechanism is proposed to update the algorithm among agents, enabling the exchange of algorithmic particles to enhance the convergence of the algorithm and the ability to escape local optima. The particle interchange formula is shown in Equation (7).

$$R_{ij}^{new} = R_{best,j}, r_3 < F^{normr}\left(S_i\right) \tag{7}$$

Where $F^{normr}(S_i)$ denotes the normalized value of the current agent's fitness.

4) Positive Greedy Selection Mechanism. The idea is to compare the fitness value of an agent after the update with the value before the update. If the updated fitness is superior, a replacement occurs, and the solutions of both agents are swapped.

Variational Mode Decomposition (VMD)

Variational Mode Decomposition (VMD) decomposes a complex signal into a predetermined number of band-limited Intrinsic Mode Functions (IMFs) through a variational approach. The core advantage of VMD is its ability to adaptively determine the central frequency and bandwidth of each component, effectively isolating different frequency components of the signal.

In practical applications, the performance of VMD depends on the appropriate selection of the modal number K and penalty factor α. Suitable K and α can decompose the signal more effectively and enhance the forecasting performance. To effectively select these two parameters, this paper utilizes RIME with the objective of minimizing the sample entropy value for optimization.

Sample Entropy

Sample Entropy (SampEn) is a tool used to measure the complexity of a time series, quantifying the rate of new information production within the series. Compared to other measures of complexity, Sample Entropy has the advantages of being computationally straightforward, independent of data length, and insensitive to noise. Its primary purpose is to determine the frequency of similar patterns within a time series. A lower Sample Entropy indicates a higher frequency of repeating patterns and, consequently, a lower complexity. Conversely, a higher Sample Entropy suggests that the time series is more complex and unpredictable. In this paper, Sample Entropy is used to measure the complexity of the components, and on this basis, components of high, medium, and low frequencies are constructed.

Multi-Objective Feature Selection (MOFS)

Filter objective function: mRMR (minimum Redundancy Maximum Relevance) obtains an optimal feature set by considering both relevance and redundancy, improving machine learning efficiency without sacrificing a significant amount of accuracy. In mRMR, mutual information coefficients are used to calculate the relationship between variables. The specific calculation method is shown in Equation (8), where $p(x)$ and $p(y)$ represent the marginal probability density of variables x and y, respectively; $p(x, y)$ represents the joint probability density of x and y; and $I(x, y)$ represents the mutual information coefficient value between x and y, with higher values indicating stronger relevance.

$$I(x,y) = \iint p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy \tag{8}$$

Therefore, for a given feature set S, the redundancy and relevance within the set can be measured by Equations (9) and (10), respectively.

$$R(S) = \frac{1}{|S|^2}\sum_{x_i,x_j\in S} I(x_i,x_j) \tag{9}$$

$$D(S) = \frac{1}{|S|}\sum_{x_i\in S} I(x_i,y) \tag{10}$$

Based on the mRMR criterion, the filter objective function can be constructed to obtain both minimal redundancy and maximal relevance, as computed by Equation (11).

$$f_1(S) = R(S) - D(S) \tag{11}$$

Wrapper objective function: Mean Squared Error (MSE) is commonly used to measure the magnitude of error between predicted and actual values. To achieve a more scientifically valid error metric, MSE is combined with time series cross-validation techniques to derive the wrapper objective function, as shown in Equation (12). In the equation, n represents the number of samples in each subset, and v represents the number of validation sets. $y_{i,t}$ and $\hat{y}_{i,t}(S)$ respectively represents the actual value and the predicted value with the given feature set S for the i^{th} sample in the subset. The paper sets the number of cross-validations to 10. As Kernel Extreme Learning Machine (KELM) is efficient and provides good fitting results, it is used as the predictor for the wrapper.

$$f_2(S) = \mathrm{TSCV} = \frac{1}{nv}\sum_{i=1}^{v}\sum_{t=1}^{n}\left(y_{i,t} - \hat{y}_{i,t}(S)\right)^2 \tag{12}$$

Thus, the combined filter-wrapper objective function can be formulated as shown in Equation (13):

$$\min\begin{cases} f_1(S) = R(S) - D(S) \\ f_2(S) = \mathrm{TSCV} = \dfrac{1}{nv}\sum_{i=1}^{v}\sum_{t=1}^{n}\left(y_{i,t} - \hat{y}_{i,t}(S)\right)^2 \end{cases} \tag{13}$$

Informer

Informer, introduced by Zhou et al. (2021), represents a significant advancement in the field of time-series forecasting, particularly for long sequences. Building on the well-established Transformer architecture, the Informer introduces several key innovations that address the limitations of its predecessor, especially in handling sequences that are orders of magnitude longer than those typically processed by standard Transformers.

Enhanced Attention Mechanism

One of the most critical challenges addressed by the Informer is the inefficiency of the Transformer's self-attention mechanism when dealing with long sequences. The standard self-attention has a quadratic computational complexity with respect to the sequence length, which makes it impractical for long sequence forecasting due to excessive memory and computational resource requirements.

ProbSparse Self-Attention Mechanism

The Informer mitigates this issue through its ProbSparse self-attention mechanism, which strategically reduces the number of attention calculations required. This is achieved by identifying and attending only to a sparse subset of the most informative key-query pairs, rather than exhaustively computing attention weights across all pairs.

Complexity Reduction

$$Attention(Q,K,V) = Soft\max\left(\frac{\bar{Q}K^T}{\sqrt{d_k}}\right)V \tag{14}$$

Here, in Equation (14), Q is a matrix of queries, K is a matrix of keys, V is a matrix of values, and d_k is the scaling factor determined by the dimensionality of the keys. The Informer's attention mechanism significantly reduces computational complexity by sparsifying the matrix Q.

In the conventional Transformer, the computation of QK^T has a complexity of $O(n^2)$, where n is the sequence length. The Informer, by contrast, reduces this to $O(nlogn)$, thereby enabling the processing of significantly longer sequences.

Sparsity Measure

The significance of queries is ascertained using a sparsity measure, which is essential for the ProbSparse technique:

$$M(q_i \mid K) = \ln\sum_{l=1}^{L_K} e^{\frac{q_i k_i}{\sqrt{d}}} - \frac{1}{L_K}\sum_{l=1}^{L_K}\frac{q_i k_j^T}{\sqrt{d}} \tag{15}$$

Where q_i is a specific query and L_K is the length of the key sequence.

Self-Attention Distilling Encoder

Traditional Transformer encoders are limited by the memory constraints imposed by long input sequences. The Informer's encoder circumvents this limitation through a process called Self-Attention Distilling.

Distillation Mechanism

$$X_{j+1}^t = MaxPool\left(ELU\left(Conv1D\left(\left[X_j^t\right]_{AB}\right)\right)\right) \tag{16}$$

$[X_j^t]_{AB}$ is the attention block. By applying a convolutional layer followed by a downsampling operation, the Informer encoder is able to reduce the temporal dimension of the data while preserving essential information, thus facilitating the processing of long sequences without a proportional increase in memory usage.

Generative Decoding for Forecasting

The Informer's decoder extends the Transformer's capabilities by incorporating a Generative Inference approach designed to efficiently generate predictions for long sequences.

Generative Inference Process

$$X_{t_{\text{de}}} = Concat\left(X_{t_{\text{token}}}, X_0^t\right) \in \mathbf{R}^{(L_{token}+L_y)\times d_{\text{model}}} \tag{17}$$

where $X_{t_{\text{de}}}$ feed de represents the input of the decoder, $X_{t_{\text{token}}}$ token is the beginning token of the sequence, and X_0^t is the placeholder of the target sequence. This process allows the Informer to generate forecasts without the need to recompute attention weights for the entire sequence, thereby avoiding the performance bottleneck encountered by standard Transformer decoders during long sequence generation.

Proposed Model

The hybrid model proposed in this paper consists of four modules: optimized decomposition, feature selection, deep learning prediction, and error correction. Fig. 1 illustrates the flowchart of this hybrid model.

Step 1: Optimized Decomposition.

The raw $PM_{2.5}$ data sequence is decomposed using Variational Mode Decomposition (VMD) optimized

by the RIME algorithm. This step extracts useful information by calculating Sample Entropy (SE) for each component derived from the decomposition. Based on these calculations, the main frequency components – high, medium, and low – are reconstructed to capture the different dynamic characteristics of $PM_{2.5}$ concentration changes.

Step 2: Multi-objective Feature Selection.

This step involves using a multi-objective feature selection strategy to filter meteorological elements and endogenous variables from the decomposed frequency components. This process aims to select the most influential feature subsets from complex data, creating an optimal feature combination for each frequency component. This approach simplifies the data for the prediction model, reduces the computational burden, and enhances prediction accuracy.

Step 3: Deep Learning Prediction.

The Informer model's hyperparameters are optimized using the RIME algorithm. Predictions are then made for all sub-sequences, and the results from each component are linearly combined to produce preliminary prediction results and an error sequence.

Step 4: Error Correction.

The error sequence is decomposed using VMD, and each error sub-sequence is predicted using the Informer model. The results are then summed to obtain an error prediction result. The error sequence is added to the initial prediction results to obtain the final $PM_{2.5}$ concentration prediction results.

Results and Discussion

Data

As the capital of China, Beijing is not only the center for politics, culture, education, and international exchange, but it is also a highly populated and industrialized metropolis. In the past few years, Beijing has frequently faced severe air pollution issues, particularly concerning high concentrations of $PM_{2.5}$, which have significantly impacted the quality of life and health of urban residents. Therefore, predicting $PM_{2.5}$ concentrations in Beijing holds particular importance.

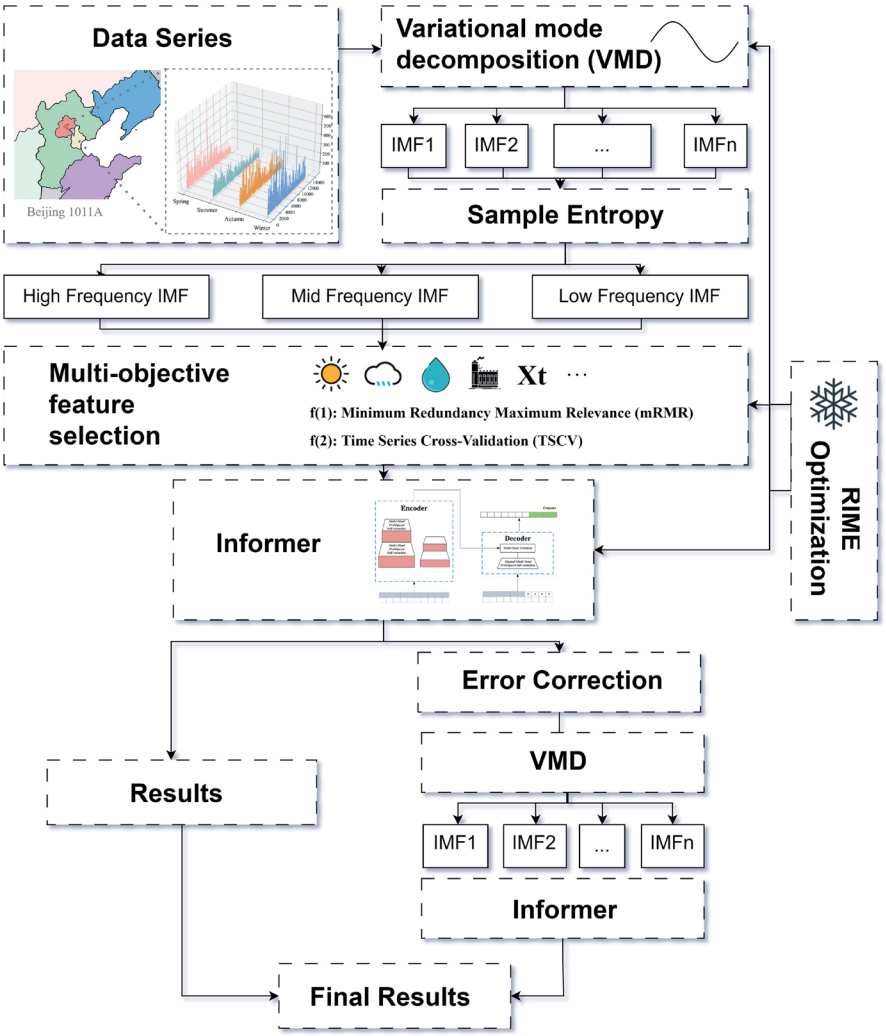

Fig. 1. The flow chart of the proposed model.

The data used in this study are sourced from the public datasets provided by the China National Meteorological Information Center, specifically from the monitoring station numbered 1011A located near the Beijing Olympic Sports Center. The data encompass meteorological features and common pollutants, including $PM_{2.5}$ measurements, collected from May 13, 2014, at 8:00 AM to October 10, 2020, at 11:00 PM, with data sampled every hour. Considering the significant seasonal variations in $PM_{2.5}$ concentrations in the Beijing area, to improve prediction accuracy, the original dataset is divided into four subsets corresponding to the seasons: spring, summer, autumn, and winter.

Each dataset includes meteorological features such as temperature, solar zenith angle, cloud opacity, dew point temperature, wind speed at 10 m height, wind direction at 10 m height, relative humidity, and atmospheric precipitable water. The selected common pollutants for the study are SO_2, NO_2, CO, O_3, and PM_{10}. The data splitting strategy for model training involves using the first 70% of the data as the training set, the following 20% as the validation set, and the last 10% as the test set.

Evaluation Metrics

This paper selects three evaluation metrics, including Root Mean Square Error *RMSE*, Mean Absolute Error *MAE*, and the coefficient of determination R^2. Their calculation formulas are as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{X}_i\right)^2} \tag{18}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left| X_i - \hat{X}_i \right| \tag{19}$$

$$R^2 = 1 - \sum_{i=1}^{n}\frac{\left(X_i - \bar{X}_i\right)^2}{\left(X_i - \hat{X}_i\right)^2} \tag{20}$$

Optimized Decomposition Results

In Variational Mode Decomposition (VMD), the number of decomposition layers, K, and the penalty factor, α, are critical factors affecting the performance of the algorithm and significantly influencing the outcomes of the decomposition. The decomposition layer K refers to the number of modes preset during the VMD process. If K is set too low, multiple modes may be merged; if K is too high, it may treat noise or irrelevant information as independent modes, increasing computational complexity and potentially introducing errors.

The penalty factor α serves as a balance parameter for the constraints and is used to adjust the decomposition precision and smoothness in VMD. The value of α determines the bandwidth during modal extraction, and the optimal value of α can vary significantly depending on the type of signal and the level of noise. A larger α results in smoother modal components but might overlook some subtle signal features; conversely, a smaller α could lead to overly wide bandwidths, causing unclear boundaries between modes. In this case study, the RIME optimization algorithm is effectively used for parameter tuning. Arranged according to the seasons – spring, summer, autumn, and winter – the obtained values of K are 6, 6, 6, and 7, respectively, while the values of α are 100, 100, 200, and 100, respectively.

Feature Selection Results

The results of feature selection for the four datasets corresponding to the seasons of spring, summer, autumn, and winter are shown in Table 1, where a checkmark (✓) indicates that the feature was selected. The results in Table 1 illustrate that the features required for each seasonal dataset vary. Since PM_{10} includes $PM_{2.5}$, the correlation between them might be high, making this feature frequently used. Additionally, wind speed can influence the dispersion and dilution of pollutants, and typically, higher wind speeds help reduce $PM_{2.5}$ concentrations. When humidity is high, atmospheric particulates can grow through vapor absorption and chemical reactions, potentially increasing $PM_{2.5}$ concentrations, which might explain why meteorological features such as Height 10 m, wind speed, and Relative humidity are often selected.

Comparative Experiments

We initially conducted forecasts using un-decomposed data for the four seasons – spring, summer, autumn, and winter – employing predictive models such as Bi-LSTM, LSTM, CNN, and Informer, with their parameters optimized using the RIME algorithm. According to the results shown in Table 2, the Informer model consistently demonstrated the best predictive performance across all original datasets.

Further, the data for the four seasons were processed through modal decomposition, and the same four benchmark predictive models – Bi-LSTM, LSTM, CNN, and Informer – were applied again. After modal decomposition, the predictive results of each mode were summed to derive the final forecast for each season. As indicated by the results in Table 2, the Informer model continued to exhibit the best forecasting performance on the seasonally decomposed datasets. Therefore, in this case study, employing the Informer as the forecasting model is an appropriate choice.

Comparison of Feature Selection Methods

Building on the decomposition of the original time series, this study combines Multi-objective Feature Selection (MOFS) for predicting $PM_{2.5}$ concentrations

Table 1. Multi-objective feature selection results and feature abbreviations.

Features	Spring			Summer			Autumn			Winter		
	H	M	L	H	M	L	H	M	L	H	M	L
TEM	✓	✓		✓	✓	✓	✓	✓		✓	✓	
AZM			✓				✓					
CO(a)		✓					✓			✓	✓	
DPT	✓											
APW	✓		✓			✓		✓		✓	✓	
RH	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
GP	✓		✓	✓		✓	✓		✓	✓	✓	✓
HWD			✓			✓	✓		✓			✓
HWS	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
ZA												
SO2	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓
CO	✓		✓	✓	✓			✓	✓	✓		✓
NO2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
O3	✓	✓		✓			✓	✓		✓	✓	✓
PM10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Xt-1	✓			✓			✓			✓		
Xt-2		✓		✓								
Xt-3			✓			✓			✓			✓
Xt-4			✓		✓						✓	
Xt-5			✓			✓			✓			✓
Xt-6		✓			✓			✓			✓	
Xt-7		✓			✓			✓			✓	
Xt-8												
Xt-9		✓	✓		✓	✓		✓	✓		✓	✓
Xt-10		✓		✓	✓	✓	✓		✓	✓	✓	

Features	Abbreviations	Features	Abbreviations
Temperature	TEM	Relative humidity	RH
Azimuth	AZM	Ground pressure	GP
Clouds opacity	CO(a)	Height 10m wind direction	HWD
Dew point temperature	DPT	Height 10m wind speed	HWS
Atmospheric precipitable water	APW	Zenith angle	ZA

Notes: Features starting with Xt- represent lag variables.

and compares the results with those obtained using two common single-objective optimization algorithms. The results demonstrate that the MOFS method outperforms the others when predicting summer data using the Informer model. Specifically, the Root Mean Square Error (RMSE) for MOFS is 42.1987, which is significantly lower than that of the Random Forest (RF) method at 65.3129 and the Mutual Information (MI) method at 68.4523, indicating superior performance. Similarly, in the prediction of winter data, the MOFS method excels, with a Coefficient of Determination (R^2) reaching 0.9712, compared to 0.9258 for RF and 0.9207 for MI.

Therefore, compared to traditional single-objective feature selection methods, the multi-objective feature selection approach exhibits better performance.

It effectively identifies the appropriate feature subsets, thereby enhancing the accuracy of predictions. Table 3 displays the error metrics of the predictions using the MOFS method post-modal decomposition, alongside those using MI and RF methods. Fig. 2 illustrates the prediction results for each model after RIME optimization using the MOFS approach.

Table 2. Benchmark model predictions before and after decomposition.

Model (undecomposed)	Season	RMSE	MAE	R^2	Model (decomposed)	RMSE	MAE	R^2
Informer	Spring	72.4729	40.5198	0.9132	Informer	64.3245	37.2569	0.9351
	Summer	77.8127	35.3712	0.9471		69.9871	33.4892	0.9589
	Autumn	60.1376	26.6723	0.9015		54.7293	21.3891	0.9196
	Winter	47.4692	25.5471	0.8943		42.3985	19.5128	0.9156
LSTM	Spring	101.5612	62.9837	0.8601	LSTM	93.4862	50.7823	0.8805
	Summer	112.8794	60.6598	0.8107		103.5471	57.3219	0.8413
	Autumn	85.2486	51.8713	0.7952		78.1129	49.5678	0.8287
	Winter	81.6428	45.2156	0.8214		74.5832	40.3891	0.8496
CNN	Spring	91.7946	56.3427	0.8471	CNN	84.3562	53.4789	0.8712
	Summer	92.4269	51.5698	0.8104		88.2173	48.7891	0.8329
	Autumn	75.2073	35.6841	0.8268		69.3945	33.2567	0.8503
	Winter	59.8715	28.5403	0.8689		55.6784	26.9812	0.8916
Bi-LSTM	Spring	76.6591	43.3286	0.8947	Bi-LSTM	71.2984	38.6789	0.9134
	Summer	81.7834	40.4532	0.9097		72.4562	38.2945	0.9261
	Autumn	67.3129	29.8798	0.8625		62.6793	28.1679	0.8853
	Winter	48.5732	25.2314	0.8867		35.4891	24.3896	0.9148

Table 3. Results of prediction error metrics based on MI, RF, and MOFS.

Model	Season	RMSE	MAE	R^2	Model	RMSE	MAE	R^2
MI-Informer	Spring	61.2896	35.6783	0.9447	RF-Informer	58.4673	34.1897	0.9512
	Summer	68.4523	32.1297	0.9623		65.3129	30.7894	0.9671
	Autumn	51.8732	22.9874	0.9258		49.6894	21.6579	0.9304
	Winter	40.1597	22.3789	0.9207		38.2145	21.5893	0.9258
MI-LSTM	Spring	90.2148	57.4691	0.8914	RF-LSTM	87.1293	55.7896	0.8997
	Summer	100.2347	56.1982	0.8532		97.3185	54.3981	0.8639
	Autumn	75.8973	48.2145	0.8369		73.5768	46.7892	0.8442
	Winter	71.4321	40.7894	0.8573		68.9892	39.2146	0.8649
MI-CNN	Spring	81.2345	51.3579	0.8798	RF-CNN	78.4876	49.7893	0.8879
	Summer	82.9784	47.2147	0.8427		80.9321	45.6312	0.8521
	Autumn	65.7892	32.1543	0.8597		63.5821	31.4729	0.8674
	Winter	53.4987	26.1298	0.8985		51.1298	25.3982	0.9052
MI-Bi-LSTM	Spring	69.1583	39.4671	0.9205	RF-Bi-LSTM	67.4329	38.7891	0.9273
	Summer	74.2398	37.4892	0.9317		72.1894	36.6712	0.9392
	Autumn	60.4571	27.5893	0.8924		58.6712	26.9374	0.8991
	Winter	44.3126	23.8795	0.9107		43.2985	23.1247	0.9184
Model		Season	RMSE		MAE		R^2	

Table 3. Continued.

MOFS-Informer	Spring	32.8749	21.2678	0.9791
	Summer	42.1987	30.8756	0.9853
	Autumn	20.8567	16.2348	0.9815
	Winter	17.3468	9.1562	0.9712
MOFS-LSTM	Spring	38.7654	28.9345	0.9573
	Summer	66.4123	48.3579	0.8942
	Autumn	50.2341	35.9786	0.9001
	Winter	26.7861	16.4672	0.9531
MOFS-CNN	Spring	34.1286	24.5691	0.9641
	Summer	36.4823	19.8732	0.9759
	Autumn	33.8562	24.1123	0.9589
	Winter	39.2471	28.6839	0.9617
MOFS- Bi-LSTM	Spring	41.3571	26.7342	0.9602
	Summer	38.6745	27.5938	0.9564
	Autumn	28.9147	21.3572	0.9742
	Winter	23.5821	18.4673	0.9758

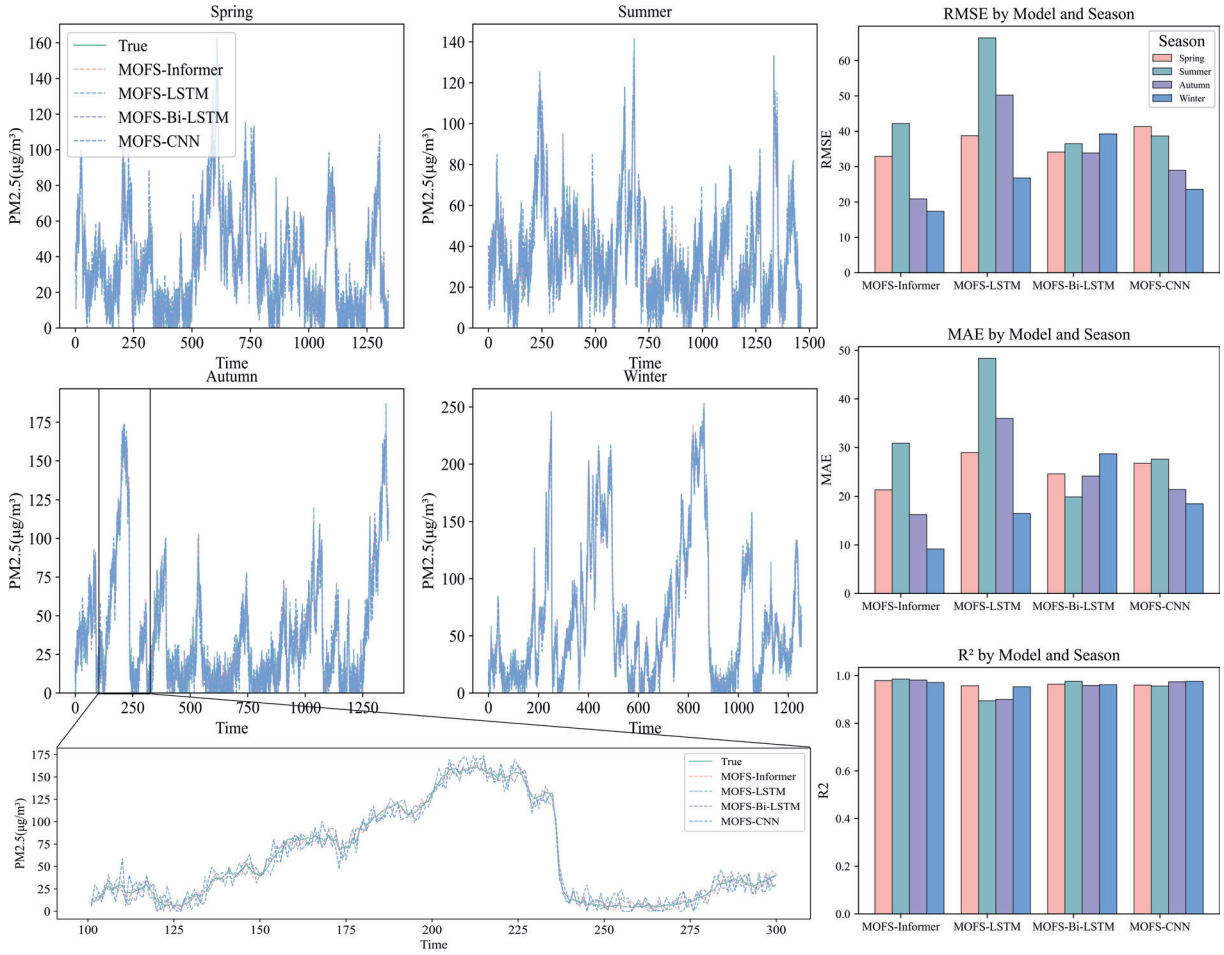

Fig. 2. Predictions for each model after RIME optimization using the MOFS approach.

Error Correction with Decomposition

The error sequence is decomposed using Variational Mode Decomposition (VMD), and each decomposed error subsequence is then predicted using the Informer model. The predictions of these subsequences are summed to derive the error prediction results. The error sequence is then added to the initial prediction results to yield the final $PM_{2.5}$ concentration predictions.

To validate the adaptability of the proposed model, forecasts were also carried out using data collected from monitoring stations coded 1052A in Baoding and 1352A in Guangzhou. Baoding, located to the south of Beijing, shares a similar geographical position and is likewise subjected to severe smog pollution, characteristic of this region. Both cities are influenced by surrounding industrial zones and agricultural activities; however, Baoding also possesses its own unique sources of pollution, such as significant industrial emissions. In contrast, Guangzhou is situated in the southern part of China, within the Pearl River Delta, and experiences a subtropical monsoon climate that is markedly different from Beijing's temperate monsoon climate. The high temperatures and humidity typical of Guangzhou, combined with the complex interactions between land and sea, pose distinct challenges for air quality models. These conditions are invaluable for testing the adaptability of the model under varying climatic conditions and diverse pollution sources. Table 4 displays the final prediction results of the model proposed in this paper, and Fig. 3 illustrates a comparison of the prediction results before and after error correction. It is evident that the model proposed in this paper exhibits excellent accuracy and performance, effectively accomplishing the task of accurately predicting $PM_{2.5}$ concentrations.

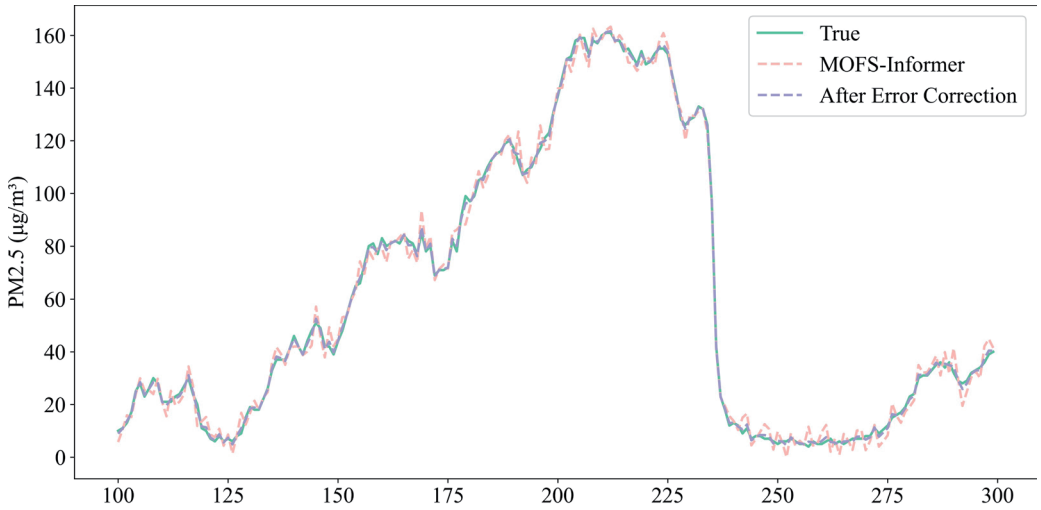

Fig. 3. Comparison of partial prediction results before and after error correction (Beijing 1011A, Autumn).

Table 4. Final predictions of the proposed model.

Site	Season	RMSE	MAE	R^2
Beijing 1011A	Spring	27.3265	16.3870	0.9901
	Summer	35.8367	21.8365	0.9953
	Autumn	17.3290	10.2351	0.9908
	Winter	13.3218	7.3261	0.9836
Baoding 1052A	Spring	25.4261	13.0428	0.9892
	Summer	37.1246	22.5873	0.9848
	Autumn	18.1456	10.7481	0.9904
	Winter	22.8971	15.6394	0.9875
Guangzhou 1352A	Spring	36.7893	28.9821	0.9797
	Summer	36.5812	22.3186	0.9951
	Autumn	22.8569	13.5509	0.9910
	Winter	18.7640	11.0108	0.9842

Conclusions

Accurate prediction of $PM_{2.5}$ concentrations is of paramount importance for public health and environmental protection, as the level of these concentrations directly relates to air quality and has become a global focal point due to its impact on human health. Long-term exposure to high concentrations of $PM_{2.5}$ increases the risk of respiratory and cardiovascular diseases, especially in urban areas. Thus, predicting $PM_{2.5}$ is crucial for environmental monitoring, pollution alerts, and health risk assessments.

To address the nonlinearity and complexity of predicting $PM_{2.5}$ concentrations, this paper presents a novel hybrid model for $PM_{2.5}$ concentration prediction. This model initially decomposes the historical $PM_{2.5}$ data using VMD, optimized by the RIME algorithm, to extract several intrinsic modal components and a residual component. Subsequently, based on sample entropy values, these components are restructured into high, medium, and low-frequency components. Then, the MOFS method is employed to select the most influential feature subsets from meteorological conditions and historical data. Finally, the optimized Informer model is used for concentration prediction and error correction to obtain the final prediction results.

Experimental results demonstrate that this model surpasses other single and hybrid models in all evaluation metrics. These achievements indicate that RIME effectively selects the key parameters for VMD, enhancing the applicability and decomposition efficiency of VMD in processing $PM_{2.5}$ data while also reducing the volatility and chaos of the original data. The MOFS method integrates the advantages of wrappers and filters, providing a more stable and reasonable feature combination compared to standalone feature selection algorithms, thereby reducing model complexity and enhancing prediction efficiency and accuracy. Additionally, RIME effectively explores the hyperparameter space of the Informer model, aiding the model in achieving higher prediction accuracy and stability. The optimized Informer, combined with an error correction model with decomposition (ECD), achieves accurate $PM_{2.5}$ concentration predictions. The hybrid model proposed in this paper considers potential issues in the prediction process and combines the strengths of each module to accurately predict $PM_{2.5}$ concentrations, enriching the technological tools in this field and providing valuable quantitative references for public health and environmental protection.

Conflict of Interest

The authors declare no conflict of interest.

References

1. HÄHNEL P., MAREČEK J., MONTEIL J., O'DONNCHA F. Using deep learning to extend the range of air pollution monitoring and forecasting. Journal of Computational Physics. **408**, 109278, **2020**.
2. JIA W., LI L., LEI Y., WU S. Synergistic effect of CO2 and $PM_{2.5}$ emissions from coal consumption and the impacts on health effects. Journal of Environmental Management. **325**, 116535, **2023**.
3. HAN X., LIU Y., GAO H., MA J., MAO X., WANG Y., MA X. Forecasting $PM_{2.5}$ induced male lung cancer morbidity in China using satellite retrieved $PM_{2.5}$ and spatial analysis. Science of The Total Environment. **607-608**, 1009, **2017**.
4. LI Z., BAI T., TANG C. How does the low-carbon city pilot policy affect the synergistic governance efficiency of carbon and smog? Quasi-experimental evidence from China. Journal of Cleaner Production. **373**, 133809, **2022**.
5. BI S., HU J., SHAO L., FENG T., APPOLLONI A. Can public transportation development improve urban air quality? Evidence from China. Urban Climate. **54**, 101825, **2024**.
6. LV B., COBOURN W.G., BAI Y. Development of nonlinear empirical models to forecast daily $PM_{2.5}$ and ozone levels in three large Chinese cities. Atmospheric Environment. **147**, 209, **2016**.
7. KUMAR A., PATIL R.S., DIKSHIT A.K., ISLAM S., KUMAR R. Evaluation of control strategies for industrial air pollution sources using American Meteorological Society/Environmental Protection Agency Regulatory Model with simulated meteorology by Weather Research and Forecasting Model. Journal of Cleaner Production. **116**, 110, **2016**.
8. SUN W., ZHANG H., PALAZOGLU A., SINGH A., ZHANG W., LIU S. Prediction of 24-hour-average $PM_{2.5}$ concentrations using a hidden Markov model with different emission distributions in Northern California. Science of The Total Environment. **443**, 93, **2013**.
9. WANG P., ZHANG H., QIN Z., ZHANG G. A novel hybrid-Garch model based on ARIMA and SVM for $PM_{2.5}$ concentrations forecasting. Atmospheric Pollution Research. **8**, (5), 850, **2017**.
10. SUN W., SUN J. Daily $PM_{2.5}$ concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. Journal of Environmental Management. **188**, 144, **2017**.
11. WANG L., HE Y., LI L., LIU X., ZHAO Y. A novel approach to ultra-short-term multi-step wind power predictions based on encoder–decoder architecture in natural language processing. Journal of Cleaner Production. **354**, 131723, **2022**.
12. ZHOU X., LIU C., LUO Y., WU B., DONG N., XIAO T., ZHU H. Wind power forecast based on variational mode decomposition and long short term memory attention network. Energy Reports. **8**, 922, **2022**.
13. SUN W., XU Z. A novel hourly $PM_{2.5}$ concentration prediction model based on feature selection, training set screening, and mode decomposition-reorganization. Sustainable Cities and Society. **75**, 103348, **2021**.
14. LIU H., LONG Z., DUAN Z., SHI H. A New Model Using Multiple Feature Clustering and Neural Networks for Forecasting Hourly $PM_{2.5}$ Concentrations, and Its Applications in China. Engineering. **6**, (8), 944, **2020**.

15. HUANG H., QIAN C. Modeling $PM_{2.5}$ forecast using a self-weighted ensemble GRU network: Method optimization and evaluation. Ecological Indicators. **156**, 111138, **2023**.
16. LIU L., GUO K., CHEN J., GUO L., KE C., LIANG J., HE D. A Photovoltaic Power Prediction Approach Based on Data Decomposition and Stacked Deep Learning Model. Electronics. **12**, (13), 2764, **2023**.
17. YIN T., CHEN H., YUAN Z., SANG B., HORNG S.-J., LI T., LUO C. LEFMIFS: Label enhancement and fuzzy mutual information for robust multilabel feature selection. Engineering Applications of Artificial Intelligence. **133**, 108108, **2024**.
18. NEMATIRAD R., PAHWA A. Solar Radiation Forecasting Using Artificial Neural Networks Considering Feature Selection. IEEE Kansas Power and Energy Conference. **2022**.
19. LV S.-X., WANG L. Multivariate wind speed forecasting based on multi-objective feature selection approach and hybrid deep learning model. Energy. **263**, 126100, **2023**.
20. ZHOU J., XU Z., WANG S. A novel dual-scale ensemble learning paradigm with error correction for predicting daily ozone concentration based on multi-decomposition process and intelligent algorithm optimization, and its application in heavily polluted regions of China. Atmospheric Pollution Research. **13**, (2), 101306, **2022**.
21. DUAN J., ZUO H., BAI Y., DUAN J., CHANG M., CHEN B. Short-term wind speed forecasting using recurrent neural networks with error correction. Energy. **217**, 119397, **2021**.
22. ZHOU H., ZHANG F., DU Z., LIU R. Forecasting $PM_{2.5}$ using hybrid graph convolution-based model considering dynamic wind-field to offer the benefit of spatial interpretability. Environmental Pollution. **273**, 116473, **2021**.
23. KOW P.-Y., CHANG L.-C., LIN C.-Y., CHOU C.C.K., CHANG F.-J. Deep neural networks for spatiotemporal $PM_{2.5}$ forecasts based on atmospheric chemical transport model output and monitoring data. Environmental Pollution. **306**, 119348, **2022**.