RESEARCH ARTICLE

A Novel Coupled Optimization Prediction Model for Air Quality

QICHEN SHAO [1,2]**, JIAHAO CHEN** [1,2]**, AND TAO JIANG** [1]

[1]College of Mathematical Science, Yangzhou University, Yangzhou 225000, China
[2]College of Innovation and Entrepreneurship, Yangzhou University, Yangzhou 225000, China

Corresponding author: Tao Jiang (jtrjl_2007@126.com)

ABSTRACT PM2.5 is a significant pollutant that negatively affects atmospheric environmental sustainability, and accurate prediction of its concentration is crucial. Most existing prediction models face challenges such as inadequate data feature capture, dismissal of influential factors, and subjective model parameter tuning. To address these issues, this paper introduces a novel coupled air quality optimization prediction model based on Variational Mode Decomposition (VMD), the Informer time series algorithm, Extreme Gradient Boosting (XGBoost), and the Dung Beetle Optimization Algorithm (DBO). The coupling approach screens influential features using the Spearman coefficient method, optimizes VMD with DBO, decomposes time series data, and classifies various feature data according to approximate entropy. The Informer algorithm and DBO-optimized XGBoost process different feature data separately, then superimpose and reconstruct the predicted values to obtain results. Using air quality prediction in Nanjing as an example, the new model achieves superior performance (R-squared=0.961, RMSE=1.988, MAE=1.624). Compared to the WANNs model with the highest accuracy in recent relevant studies, our model demonstrates a 2.96% increase in R-squared, a 21.89% decrease in RMSE, and a 20.05% decrease in MAE. This comparison illustrates that the proposed DBO-VMD-Informer-XGBoost prediction model effectively addresses the limitations of existing air quality prediction models and offers increased prediction accuracy. By employing the advanced DBO algorithm for prediction and innovatively combining VMD, Informer, and XGBoost, this model presents high potential in air quality prediction and is anticipated to have broader applications.

INDEX TERMS Air quality prediction, dung beetle algorithm, time series informer, variational mode decomposition, XGBoost.

I. INTRODUCTION
A. BACKGROUND
Since the industrial revolution, global urbanization and industrialization have been steadily accelerating, resulting in persistent air quality problems affecting various countries worldwide. The 2022 air pollution prevention and control funding budget, issued by China's National Ministry of Finance, reveals that China's budget for air pollution prevention and control amounts to 20.7 billion, a 65.6% increase compared to 12.5 billion in 2021. The primary pollutants impacting the atmospheric environment include PM2.5, PM10, nitrogen dioxide, sulfur dioxide, carbon monoxide, and ozone. PM2.5 is the most critical contaminant, and its presence in the environment escalates the risk of humans developing atherosclerosis, hypertension, cardiac arrhythmia, and other diseases [1]. As lung function in infants and children is still developing, air pollution exacerbates respiratory diseases, impairs lung function, and increases asthma rates [2]. Consequently, an accurate and effective air quality prediction model is essential in fostering significant improvements in the atmospheric environment and

The associate editor coordinating the review of this manuscript and approving it for publication was Deepak Mishra.

protecting humans from the harmful effects of atmospheric pollution.

B. LITERATURE REVIEW

Air quality is influenced by various factors, including time series, atmospheric conditions, pollutant emissions, and human activities [3]. Abirami proposed a DL-Air deep learning model to analyze the correlation between spatial and temporal dependencies in air quality data and meteorological parameters using a Long Short-Term Memory network variant (STAA-LSTM) for air quality forecasting [4]. Chen applied the Spearman coefficient (SC) method and Random Search algorithm (RS) in conjunction with the XGBoost for air quality prediction [5]. Spatiotemporal coordinated predict models outperform traditional models due to their incorporation of temporal and spatial dependencies. Future forecasting models should focus more on factors beyond pollutant time series.

Air quality prediction methods can be broadly categorized into three groups: numerical prediction methods, statistical prediction methods, and artificial intelligence methods. Among these, statistical and artificial intelligence methods are more widely adopted. Czernecki employed a single XGBoost prediction model to predict PM2.5 concentrations in a Polish city, demonstrating that XGBoost outperformed other machine learning algorithms in this task [6]. Van utilized Decision Tree, Random Forest (RF), and XGBoost to predict AQI values, with the comparison showing that XGBoost excelled in predicting AQI values [7]. Kothandaraman employed KNN, Linear Regression, RF, Lasso Regression, Ridge Regression, XGBoost, and AdaBoost models to forecast PM2.5 levels. The XGBoost, AdaBoost, Random Forest, and KNN models demonstrated improved predictions and reduced error rates [8]. Dong initially applied a single Informer model to establish a nonlinear relationship between impact factors and air quality for air quality prediction [9]. Subsequently, numerous coupled models have been proposed to enhance the accuracy of single-model predictions. Li utilized a Convolutional Neural Network (CNN) to extract air quality features and employed Long Short-Term Memory (LSTM) for air pollution data forecasting [10]. Pranolo used PSO-LSTM to analyze multivariate data for air quality forecasting [11]. Gao's combined prediction model employed Empirical Mode Decomposition (EMD) to classify data and subsequently applied Extreme Learning Machine (ELM) and XGBoost for air quality forecasting [12]. Coupled models can adapt to various systems and situations more efficiently, flexibly, and accurately than single models, which exhibit superior predictive performance [13]. Nonetheless, there remains potential for further improvement in prediction performance. Future research should focus on developing more robust coupled models to minimize the negative impact of human tuning on prediction outcomes.

With the continuous advancement of intelligent algorithms, various sophisticated modern algorithms have emerged. In addition to the basic coupled models discussed above, numerous new coupled prediction models have been proposed recently. Guo employed Artificial Neural Networks (ANNs) and Wavelet Neural Networks (WANNs) to predict daily PM2.5 concentrations in Shanghai [14]. Lee utilized a recently developed novel spatial multiple receptor model (BSMRM) to estimate the potential contribution of selected unmonitored sites to ambient PM2.5 [15]. Gao applied a Generalized Additive Model (GAM) to assess historical trends and future projections of PM2.5 concentrations in Southern California, considering emissions and climate impacts [16]. Zhang developed a spatiotemporal model (LDSPM) for predicting PM2.5 based on the K-Core concept and label distribution [17]. Sun used a Distributed Lag Nonlinear Model (DLNM) to establish warning thresholds for air quality data for PM2.5 prediction [18]. N. Kapoor employed Artificial Neural Networks (ANN) to predict the event-specific spread of SARS-CoV-2 in a mixed-mode ventilation office [19]. Kapoor also compared multiple machine learning models for office CO2 prediction and found that the optimized Gaussian Process Regression (GPR) model performed best [20]. In recent years, to further reduce prediction errors, a class of improved coupled models has been increasingly utilized in forecasting. These enhanced models employ various forecasting algorithms to leverage the advantages of different approaches and better capture the intrinsic characteristics of historical data while avoiding overfitting. Hybrid prediction models have been applied to photovoltaic power prediction [21], [22], wind power prediction [23], [24], and power load prediction models [25], [26]. However, their use in air quality forecasting is less prevalent. Future models could explore combining multiple prediction algorithms to better capture data features and improve accuracy and generalization.

After analyzing the matter in question, it has been identified that the present research on air quality projections suffers from a few critical issues. Firstly, whilst predicting air quality, other factors alongside pollutant time series demand thorough attention. Secondly, experimenters subjectively regulate the parameters in air quality prediction models, thereby substantially compromising their predictive accuracy. Therefore, it is essential to rationally choose such parameters via intelligent optimization algorithms. Last but not least, time series data are volatile and erratic. Without data processing, individual prediction models will fail to fully grasp the underlying characteristics of air pollution, resulting in limited prediction accuracy. Hence, it is crucial to decompose the data and reduce its complexity.

C. CONTRIBUTIONS OF THE PAPER

Based on the above analysis, a new coupled optimization prediction model based on DBO optimization VMD and Informer and XGBoost is proposed and applied to the air quality prediction problem. We first select influencing features by Spearman coefficient method. This paper uses the

advanced Dung beetle algorithm (DBO) to optimize the VMD and determine the number of Intrinsic mode fraction (IMF) decomposition layers with the penalty factor. Compared to the popular PSO, GWO, SSA, SMA and AVOA optimisation algorithms, the DBO optimisation parameters have a higher stability and accuracy. A detailed and scientific demonstration of the superiority of the DBO algorithm will be presented later. VMD achieves the exact decomposition of complex data and then classify the IMF based on approximate entropy. Then use Informer prediction for the low frequency decomposition group and XGBoost prediction for the high frequency decomposition group and optimize the required parameters: Maximum depth of the tree, Learning rate, Number of trees, and Subsample using the Dung beetle algorithm (DBO). Informer is good at predicting low-frequency data at low approximate entropy and has better generalization ability for such long time series. XGBoost performs well in solving high-frequency data predictions with high approximate entropy. Finally, reconstruct and superimpose the decomposition predictied results to obtain the air quality prediction values. Compared to most of the latest and most advanced air quality prediction models [14], [15], [16], [17], [18], [19], [20], this paper is the first innovative use of coupled VMD-Informer-XGBoost for air quality prediction, which enables the model to capture the characteristics of the dataset better and improve the accuracy and generalization of the prediction model. The model proposed in this paper is more innovative, advanced, and accurate regarding parameter optimization, model structure, and prediction performance when compared. We conduct several scientifically sound numerical experiments later to prove the above points.

There are four innovations and enhancements in this paper: (1) This paper innovatively proposes DBO-VMD-Informer-XGBoost model and applies it to air quality prediction, which greatly improves the prediction accuracy of the model and has strong stability and generalization ability; (2) The advanced DBO algorithm with excellent performance is firstly adopted in the field of data prediction, which relatively reduces the running time and greatly reduces the errors caused by human empirical tuning. The DBO algorithm has a faster convergence rate and better convergence results; (3) The innovative coupled VMD-Informer-XGBoost is first used in air quality prediction, which enables the model to capture the characteristics of the data set to a great extent and improve its accuracy and generalisation. After using DBO-VMD decomposition, the data complexity is greatly reduced. Informer and DBO-XGBoost algorithms are adopted to deal with the decomposed data with different features. The prediction error due to the complexity of the dataset is substantially reduced; (4) The influencing factors are filtered by Spearman coefficient method to avoid over-fitting or feature loss and reduce computational effort.

All of the above contributions are combined, as shown in Table 1.

TABLE 1. The main contributions list.

	Contributions
1st	The Spearman coefficient was used to screen the influencing factors to avoid overfitting or loss of features and to reduce the computational effort.
2nd	The DBO algorithm is used for the first time to reduce the errors caused by human adjustment and to reduce the running time.
3rd	The model innovatively utilized coupled VMD-Informer-XGBoost to minimize dataset complexity, accurately capture dataset features, and improve pattern prediction accuracy and generalization.
4th	The innovative DBO-VMD-Informer-XGBoost model was proposed and applied to PM2.5 prediction, exhibiting improved accuracy along with robust stability and generalization ability.

D. PAPER ORGANIZATION

The rest of this study is organized as follows. Section II explains the principle of the algorithm, the proposed process, and the implementation steps of the DBO-VMD-Informer-XGBoost hybrid model. Section III presents the numerical experiments we have conducted and analyses the results in detail. These include data collection and pre-processing, correlation analysis, DBO performance testing, parameter optimization and data decomposition, experimental setup, experimental results, and analysis, and the trade-off between accuracy and workload. Section IV summarises the model's strengths, limitations, and future perspectives in this study.

II. MODELS AND ALGORITHMS

A. DUNG BEETLE OPTIMIZATION ALGORITHM (DBO)

The Dung Beetle Optimizer (DBO) is a novel optimization algorithm based on the behaviors of dung beetles, developed in November 2022 [27]. Unlike existing optimization algorithms such as the Bat, Ant, and Whale algorithms, DBO boasts rapid convergence, high accuracy, and stability [27]. Despite its significant advantages, DBO has not been widely applied yet. The DBO algorithm primarily obtains the global best position and fitness value via the position change of dung beetles, using the following five behaviors: dancing, rolling, stealing, foraging, and reproduction. Suppose n represents the population size of dung beetles.

(1) The position is updated and defined when the dung beetle rolls the ball (forward unobstructed) as follows, where x_i^t represents the ith dung beetle's position at the tth iteration, t represents the number of iterations, a represents a random number of -1 or 1, k represents a random number in $(0, 0.2]$, x_{worst}^t denotes the global worst position, and b denotes a random number in $(0, 1]$:

$$x_i^{t+1} = x_i^t + akx_i^{t-1} + b\left|x_i^t - x_{worst}^t\right| \qquad (1)$$

(2) The position is updated and defined when the dung beetle is dancing (forward with obstacles) as follows:

$$x_i^{t+1} = x_i^t + \tan\theta \left| x_i^t - x_i^{t-1} \right| \tag{2}$$

where θ is a random number in $[0, \pi]$. When $\theta = 0, \pi/2, \pi$, the dung beetle position does not change.

(3) Update the existing position of the post-breeding Dung beetle brood and define it as follows b_1, b_2 is an independent and random $1 \times d$ matrix, Lb^* and Ub^* are the upper and lower bounds of the reproduction region, and x_{gbest}^t is the current global best position.

$$x_i^{t+1} = x_{gbest}^t + b_1(x_i^t - Lb^*) + b_2 \left| x_i^t - Ub^* \right| \tag{3}$$

(4) Update the existing position of a small Dung beetle while foraging and define it as follows, where Lb^l and Ub^l are the upper and lower bounds of the foraging area, C_1 obeys a normal distribution, and C_2 is a random $1 \times d$ matrix.

$$x_i^{t+1} = x_i^t + C_1(x_i^t - Lb^l) + C_2 \left| x_i^t - Ub^l \right| \tag{4}$$

(5) Update the existing position of the dung beetle and define it when stealing as follows:

$$x_i^{t+1} = x_{lbest}^t + Sg(\left| x_i^t - x_{gbest}^t \right| + \left| x_i^t - x_{lbest}^t \right|) \tag{5}$$

where g is the $1 \times d$ matrix obeying a normal distribution, x_{lbest}^t is the local best position, and S is a constant value.

Finally, the dung beetles that rolled the ball and danced were combined into one group, and the remaining dung beetles for each behavior were grouped alone, dividing the Dung beetle population according to the ratio of $6:6:7:11$. The positions of each group are updated according to the behavior of the group to which they belong, and several iterations are performed to derive the global best position and its fitness value.

In the new model, we use DBO to optimize the parameters of VMD and XGBoost.

B. DBO-OPTIMIZED VMD
Dragomiretskiy et al. introduced Variational Mode Decomposition (VMD) in 2014. VMD is an adaptive signal decomposition method that draws on the classic Wiener filter, Hilbert transforms, and frequency mixing concepts [28]. As opposed to other mode decomposition techniques like Empirical Mode Decomposition (EMD) and Signal Mode Decomposition, VMD boasts higher accuracy when processing data center frequency components, faster computation, and more excellent noise immunity [29]. VMD employs a fully non-recursive mode decomposition for the calculation of the Intrinsic Mode Function (IMF) and selectively chooses the number of modes available. This has several advantages, such as autonomous mode selection and non-recursive solutions.

VMD processes the signal components to minimize the sum of the bandwidths of the center frequencies of each mode component and establishes a constrained fractional variable optimization model. The formula is as follows:

$$\min_{\{u_k\},\{w_k\}} \left\{ \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}$$

$$s.t. \sum_{k=1}^K u_k = f(t) \tag{6}$$

where $\{u_k\}$ is the kth mode component of the decomposition, $\{w_k\}$ is the kth center frequency, $*$ is the convolution operation, $f(t)$ is the original signal and ∂_t is the gradient operation. And then VMD uses Lagrange multipliers to convert the constrained form into an unconstrained form to solve the mode component u_k and the center frequency w_k. There are the equations:

$$u_k^{n+1}(w) = \frac{f(w) - \sum_{i\neq k} u_i(w) + \frac{\lambda(w)}{2}}{1 + 2\alpha(w - w_k)^2} \tag{7}$$

$$w_k^{n+1} = \frac{\int_0^\infty w \left| u_k^{n+1}(w) \right|^2 dw}{\int_0^\infty \left| u_k^{n+1}(w) \right|^2 dw} \tag{8}$$

where n represents the number of iterations, $u_k^n(w)$ represents the kth mode component of the nth iteration, w denotes the component frequency, α denotes the penalty factor, w_k^n denotes the kth center frequency of the nth iteration, and λ denotes the Lagrange multiplier. A mode component k that is too small will cause modal mixing and too large will over-decompose. Too large a penalty will narrow the modal function and lose helpful information, and too small will carry too much interference information.

From the above VMD decomposition process and the decomposition principle, it can be seen that the number of decompositions k and the penalty factor α have a significant impact on the results. In this paper, DBO is used to optimise the VMD parameters, with the envelope entropy E_p as the fitness function.

$$E_p = -\sum_{j=1}^m p_j \lg p_j \tag{9}$$

$$p_j = a(j) \Big/ \sum_{j=1}^m a(j) \tag{10}$$

where m is the number of samples and $a(j)$ is the Hilbert demodulated envelope sequence of the modal components decomposed by the VMD. The principle and procedure of the DBO optimized-VMD algorithm are as follows.

In the new model, we use the DBO-optimized VMD to achieve the exact decomposition of complex data.

C. DBO-OPTIMIZED XGBoost
XGBoost is an integrated algorithm based on Boosting that combines basis functions and weights. Also known as
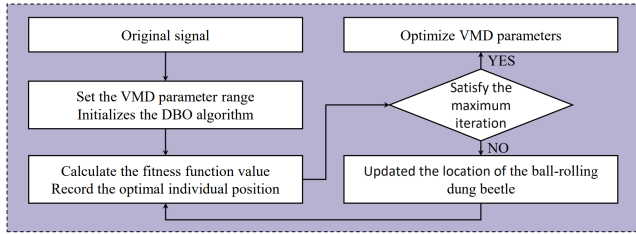
FIGURE 1. Flowchart of DBO-optimized VMD.

extreme gradient boosting trees, XGBoost offers high prediction accuracy, fast running speed, and negligible feature requirements, making it a popular big data processing method across several industries [30]. Relative to Gradient Boosting Decision Tree (GBDT), XGBoost enhances model performance and flexibility by regulating overfitting, an approach that proves particularly effective when handling datasets with high frequency [31]. The following subsection outlines XGBoost's principal steps in brief.

Define $D = \{(x_i, y_i)\}$, $i = 1, 2, \cdots, n$, and n is the sample quantity. The XGBoost model is constructed by iteratively adding trees with the output function shown below:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i) \tag{11}$$

where T denotes the leaf nodes quantity in the tree and $f_t(x_i)$ denotes the newly added function for the tth round of prediction.

According to the core idea of XGBoost, the residual fitting formula is as follows:

$$\hat{y}_i^{(0)} = 0$$
$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{12}$$

where $\hat{y}_i^{(0)}$ represents the initial value, $\hat{y}_i^{(t)}$ represents the predicted result of the tth round, and $\hat{y}_i^{(t-1)}$ represents the predicted result of the $(t-1)$th round.

Construct optimization function based on XGBoost models:

$$obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + C \tag{13}$$

where $\Omega(f) = \gamma T + \lambda/2 \sum_{j=1}^{T} w_j^2$, γ and λ represent the parameters indicating the complexity of the model, w_j represents the jth weight parameter and C means a constant term.

XGBoost requires selecting parameters: maximum tree depth, learning rate, number of trees, and subsample. These four parameter values play a key role in the predictive performance and stability of XGBoost. So we need to adjust them more accurately. This paper uses DBO to optimize the XGBoost parameters and the prediction error RMSE is used as the fitness value.

In the new model, we use the DBO-optimized XGBoost with excellent performance to predict the high frequency components of the DBO-optimized VMD decomposition.

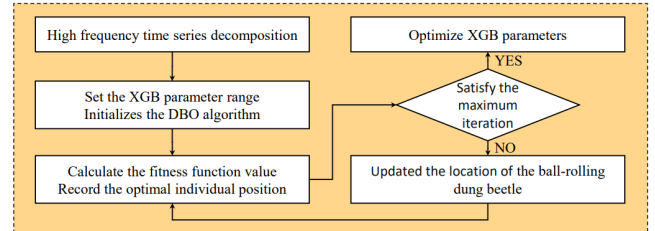The principle and procedure of the DBO optimized-XGBoost algorithm are as follows.

FIGURE 2. Flowchart of DBO-optimized XGBoost.

D. INFORMER

Informer is a supervised learning model with an attention mechanism that comprises two key components: an encoder, which establishes long-term dependencies regarding the stability of the original input sequence, and a decoder, which undertakes sequence prediction [9]. The classical Transformer algorithm uses an encoder to establish a hidden representation of the input sequence, while the decoder decodes the output sequence to derive the predicted value, as explained by [32]. Compared to Transformer, Informer reduces computational complexity and memory usage while improving long-time sequence prediction efficiency and accuracy. This is achieved in large part by the reduction in memory requirements [33].

Use the standardized influencing factor data $X = (X_1, X_2, \ldots, X_t)^T \in R^{L \times d}$ from the previous t moments as input data, where L denotes the number of types of influencing factors, d denotes the input dimension.

The sequence data's position information is encoded, and global time information is incorporated to enhance model stability. Subsequently, the influential factor data imbued with position and time information is fed into the decoder and encoder of the model. The encoder, which is composed internally of a multi-headed probabilistic sparse self-attentive model and a stack of "distillation" mechanism modules, achieves the desired outcome. The equation for the probabilistic sparse self-attentive mechanism is as follows:

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

where $Q \in R^{LQ \times d}$, $K \in R^{LK \times d}$, $V \in R^{LV \times d}$, d denotes the input dimensions, $\bar{Q}$ denotes obtained from Q based on probabilistic sparse, and $Soft$ max denotes the activation function. The "distillation" mechanism reduces memory consumption and computation time by compressing the feature dimensions and runs as the following equation:

$$X_{j+1}^t = MaxPool(ELU(Convld([X_j^t]_{AB}))) \tag{15}$$

where $[X_j^t]_{AB}$ contains the multi-head probabilistic sparse self-attention operation, $Convld$ denotes the convolution
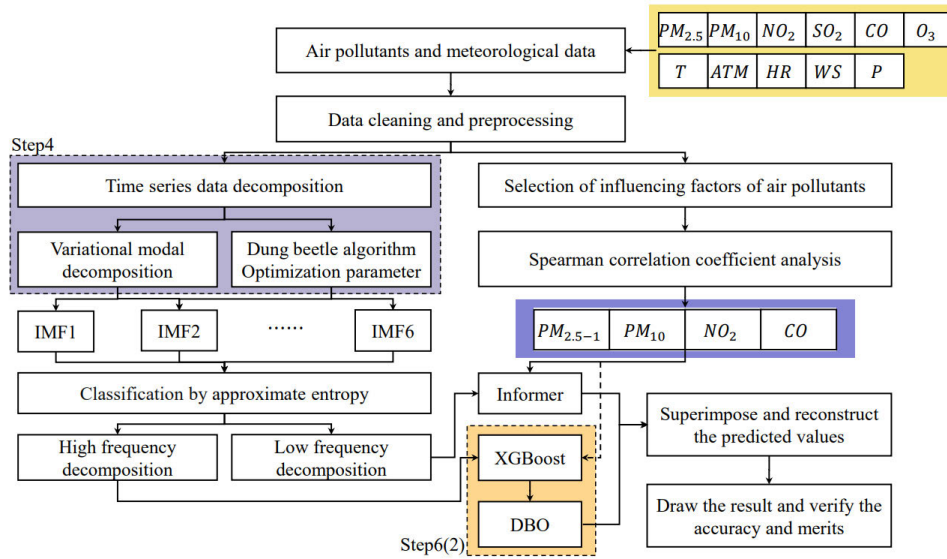
FIGURE 3. Flowchart of air quality prediction system.

operation, *ELU* denotes the activation function, and *MaxPool* denotes the maximum pooling operation. After the two modular operations, the encoder outputs one intermediate result.

Inside the decoder, the input data after probabilistic sparsification and the intermediate results of the encoder are subjected to self-attentive operations and dimensional adjustment to output the prediction result $X = (X_{t+1}, X_{t+2}, \ldots, X_{t+n})^T$. Finally, the model is gradually optimized based on the loss function of the prediction result.

In the new model, we use the Informer to predict the low frequency components of the DBO-optimized VMD.

E. HYBRID DBO-VMD-INFORMER-XGBoost MODEL
Combining the DBO algorithm, DBO-optimized VMD, DBO-optimized XGBoost, and Informer algorithm, this paper proposes a new coupled optimized air quality prediction model DBO-VMD-Informer-XGBoost. The new model addresses some of the shortcomings of the existing air quality prediction field, such as prediction bias resulting from complex data and inadequate feature capture, neglecting the influence of factors other than time series, and the imprecision that comes with human intervention. The specific steps to implement this new model are as follows:

Step 1: Collect data on the six principal air pollutants and five meteorological variables.

Step 2: Perform data cleaning and preprocessing, which entails standardizing the data, filling in any missing values using linear interpolation, and removing outliers, if any.

Step 3: Employ the Spearman coefficient method to select the *n* impact features of the meteorological and pollutant data.

Step 4: Validate the DBO algorithm's efficiency and performance. Use the DBO-optimized VMD algorithm to decompose historical PM2.5 data to obtain IMF components.

Step 5: The approximate entropy values of each of the *m*IMF components are derived separately. The *m* IMF

components are divided into *f* low frequency decomposition parts and $m-f$ high frequency decomposition parts according to the approximate entropy values so as to select the optimal model for predicting each component.

Step 6: Use *n* impact features selected in step 3 as the feature inputs of the model and use the historical PM2.5 data as the model prediction outputs.

(1) The Informer time series prediction model is built for the *f* low frequency decomposition parts and predicts the result $p_1, p_2, \ldots, p_f$.

(2) The DBO-optimized XGboost prediction model is built for $m-f$ high frequency decomposition parts and predict the result $p_{f+1}, p_{f+2}, \ldots, p_m$.

Step 7: Reconstruct and superimpose the forecast results of the *m* components in Step 6 to obtain the final forecast results. The model is evaluated based on the relative error index calculation and comparison with other models.

A diagrammatic representation of the airflow of the novel coupled optimization prediction model based on DBO-VMD-Informer-XGboost can be found in Fig.3.

Affected by time series, atmospheric conditions, emergencies, and other factors, air quality shows instability and randomness, but it also maintains a certain periodicity. In this case, we use Spearman correlation coefficient method to select characteristics of influencing factors of air pollutants to avoid overfitting, improve accuracy and reduce the amount of calculation. We have tuned the parameters of DBO and tested the performance of the tuned DBO in the benchmark problem and in the real-world problem. The Dung beetle algorithm (DBO) is used to optimize the Variational modal decomposition (VMD) parameters: decomposition layers *k* quantity and the penalty factor α, to avoid errors caused by artificial tuning of the parameters, and to decompose the PM2.5 historical data, thus reducing the complexity of the data and improving the prediction accuracy. Calculating each

component's approximate entropy, low frequency and high frequency decomposition series are divided. To fit data sets with different frequency and play a better performance than a single prediction model, we use the Informer prediction model, which is good at predicting low frequency data for low approximate entropy and has better generalisation for such long time series. Meanwhile, we use the XGBoost model, which performs well in solving high-frequency data prediction for high approximate entropy. The DBO algorithm is adopted to automatically optimize the parameters of XGBoost: the maximum depth of the tree, the learning rate, the number of trees, and the ratio of subsamples, to achieve accurate prediction of air quality and avoid errors caused by human adjustment of parameters. Finally, the air quality forecast results are obtained by superimposing and reconstructing the prediction data of each decomposition.

The novel coupled DBO-VMD-Informer-XGBoost forecast model proposed here greatly reduces the complexity of time-series data and accurately captures data features. Fully exploit the patterns in historical air quality data. Meanwhile, it integrates the influence of various affecting factors and avoids the error caused by artificial adjustment of model parameters, which significantly improve predictive performance and reduces the error rates. It greatly improves the prediction accuracy of the model, and has strong model stability and generalization ability. In order to reflect the superior performance of the novel coupled model, we use the excellent models in the reference literature as the control group in this paper for comparison and validation, and the specific evaluation results are shown in Table 9.

III. NUMERICAL EXPERIMENTS AND ANALYSIS OF RESULTS

A. DATA COLLECTION AND PREPROCESSING
The experimental data were obtained from https://rp5.ru/, including six air pollutant measurements and temperature, air pressure, relative humidity, wind speed, and precipitation. In this paper, we take the data of Nanjing city from 2020.1.1 to 2022.12.6 days as experimental sample. Nanjing is located in the middle and lower reaches of the Yangtze River in China and has a subtropical monsoon climate. Its economic level, cultural level, and productivity level are moderate. Among urban agglomerations with similar characteristics, Nanjing is highly representative. The period of the data set is three years. Daily data are measured from multiple meteorological base stations in different parts of Nanjing. The dataset is sufficiently representative. The dataset also contains data on a wide range of air quality and climate conditions and is diverse enough. Using the Nanjing data for the model experiments provides more substantial evidence of the model's novelty and applicability and the experimental results' efficiency and credibility. The forecast target is the concentration value of PM2.5 ($\mu g/\text{m}^3$). Next, we pre-process the collected data.

(1) Deal with missing data. Due to the small span of missing data, this paper uses linear interpolation to reasonably supplement the original data set, which is closer to the changing trend of the original data set.

$$y = \frac{x_1 - x}{x_1 - x_0}y_0 + \frac{x - x_0}{x_1 - x_0}y_1 \qquad (16)$$

where, y is the data to be added and x is the time point to which the data belongs.

(2) Hand of abnormal data. There are also a few nodes with abrupt data changes in the dataset, which the authors consider to be sharp changes in air quality caused by unexpected events, such as dust explosions, extensive human gathering activities, etc. This paper performs univariate anomaly correction for such nodes, assuming that the data are based on a normal distribution. The threshold is set to 99%, and all those falling outside the threshold are judged as outliers and removed.

(3) Normalize the processed dataset. Following data preprocessing, normalization was performed on the dataset.

$$y^* = \frac{y - y_{\min}}{y_{\max} - y_{\min}} \qquad (17)$$

where y^* is the normalized value.

(4) The collected data was partitioned and 15% was taken as the test set, 10% as the cross-validation set and 75% as the training set.

B. CORRELATION ANALYSIS
The literature [34], [35], [36] confirms that the concentration levels of the six air pollutants and meteorological variables, such as temperature, barometric pressure, relative humidity, wind speed, and precipitation, can impact PM2.5 predictions.

Consequently, it is necessary to consider the time factor and the influence derived from these variables while making PM2.5 predictions. However, since each variable has a different impact level, we filtered out certain variables with little or no impact. Feature selection is crucial in pre-processing meteorological and pollutant data, as it helps avoid overfitting, improves accuracy, and reduces computational effort. The Pearson and Spearman methods are typically used to analyze the correlation between two variables. The Spearman coefficient method is preferred when one of the variable sets follows a fixed order, while the Pearson coefficient applies when analyzing customarily distributed variables. Given that we do not observe normal distribution in our data set, we utilize the Spearman coefficient method to calculate the correlation between each feature and the predictor variables.

The SC uses monotonic equations to evaluate the correlation between two statistical variables in statistics. The correlation coefficient between two variables ranges from -1 to 1, with a positive value indicating that the changing trend between the two variables is the same and a negative value suggesting that the changing direction is the opposite. A correlation coefficient of 0 reveals no movement in the change between the variables. The formula for the correlation

coefficient between two variables is as follows:

$$\rho(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}} \quad (18)$$

where ρ is the correlation coefficient between x and y, n is the size of the data set, and $\bar{x}$ and $\bar{y}$ are the means. The heat map of feature correlation coefficients is shown in Fig.4.
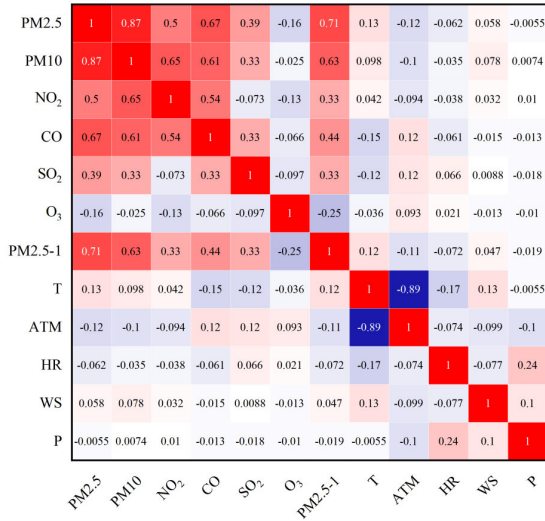

FIGURE 4. Feature correlation coefficient heat map.

The description of each symbol in the above figure is shown in Table 2.

TABLE 2. Symbol description.

Symbol	Meaning	Unit
PM2.5-1	PM2.5 concentration on the previous day	$\mu g/m^3$
T	Temperature	°C
ATM	Pneumatic Pressure	atm
HR	Relative Humidity	RH
WS	Wind Speed	m/s
P	Precipitation	mm

The graph presented herein focuses on the correlation coefficients between PM2.5 and other variables. The magnitude of the coefficients determines which variables significantly influence PM2.5. A coefficient value greater than 0.5 in absolute terms indicates a higher correlation with the prediction target. The correlation coefficient matrix indicates that all meteorological features possess weak correlations. Therefore, if these variables are imposed as influential components of PM2.5, overfitting, reduced accuracy, and operational complications may arise. However, only PM2.5-1, PM10, CO, and NO_2 are highly correlated with PM2.5 and hence have been selected as the influencing features.

C. DBO PERFORMANCE TEST

The article tests the performance of the DBO to illustrate the advantages of the DBO and reflect the rationale for use. Firstly, we select suitable parameters for the DBO. The literature [27] gives recommended ranges for each parameter of the DBO, i.e. $k \in (0, 0.2]$, $b \in (0, 1]$, $S \in (0, 2]$ in equation (1-5). In this paper, we choose $k = \{0.05, 0.1, 0.15, 0.02\}$, $b = \{0.05, 0.25, 0.45, 0.65, 0.85\}$, $S = \{0.1, 0.5, 1, 1.5, 2\}$. We test the 125 sets of parameter values above. After comparison, we set $n = 50$, $t = 500$, $k = 0.14$, $b = 0.25$, $S = 0.5$ in equations (1-5). This set of parameters performs best. After that, the performance tests in the benchmark problem as well as the performance tests in the real problems are conducted for the DBO after the parameterization.

In addition to testing the performance of DBO, the same tests were conducted on Particle Swarm Algorithm (PSO), Sparrow Search Algorithm (SSA), Gray Wolf Optimization Algorithm (GWO), Slime Mould Algorithm (SMA), and African Vultures Optimization Algorithm (AVOA). The specific test procedure, test results, and analysis of the results are placed in Appendix A of this paper.

Finally, after comparison, we conclude that the tuned DBO has the best accuracy and stability in finding the optimal global solution. Moreover, the DBO can effectively handle real-world optimization problems. In other words, the tuned DBO applies to problems of various types and dimensions. Therefore, the VMD and XGBoost after DBO optimization have higher stability and accuracy, avoiding the significant errors caused by the human choice of parameters.

D. PARAMETER OPTIMIZATION AND DATA DECOMPOSITION

For the VMD decomposition and XGBoost prediction procedures, this paper utilizes DBO optimization with appropriate parameterization ($n = 50$, $t = 500$, $k = 0.14$, $b = 0.25$, $S = 0.5$) to determine the parameters required, avoiding the significant errors that artificially selected parameters can cause.

(1) The optimal number of decomposition layers k and the optimal penalty factor α are derived by DBO optimization VMD using the envelope entropy as the fitness value, which is shown in Table 3.

TABLE 3. Parameters of VMD model.

Parameter	k	α
Range	3-10	500-2000
Optimum	6	1682

If k is too small, the mode is aliased. Otherwise, the decomposition is excessive. If α is too large, the frequency band of the mode function is too narrow, and the effective information is ignored. Otherwise, the frequency band of the mode function is too broad, and the interference information is too much. After DBO optimization, we obtain VMD algorithm parameters $k = 6$ and $\alpha = 1682$. In this case, the value

of k is appropriate, and the modal decomposition effectively decomposes the data complexity. At the same time, α is scientific and retains most of the original modal adequate information, while the interference information is reduced as much as possible.

The parameter combination (6,1682) is chosen as the optimal parameter setting for the VMD decomposition of PM2.5 historical data. After the DBO optimization, we used the VMD to decompose the PM2.5 historical data and obtained the time domain and frequency diagrams of the six IMF components, as shown in Fig.5.
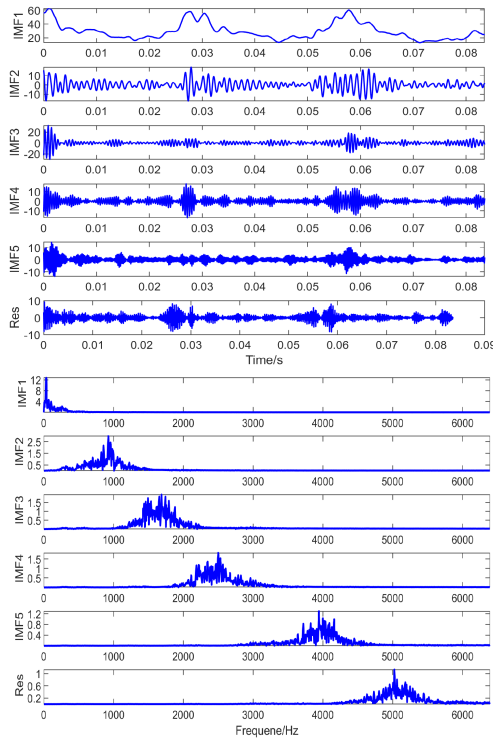

FIGURE 5. The time and frequency domain of six IMF components.

In turn, each component's approximate entropy [37] was calculated and shown in Table 4.

TABLE 4. Approximate entropy of each component.

Component	Approximate entropy
IMF1	0.3674220
IMF2	0.3593639
IMF3	0.4038877
IMF4	0.3745767
IMF5	0.5986284
RES（IMF6）	0.6717569

According to the approximate entropy values obtained from the above table, we designate IMF1-IMF4 as low-frequency decomposition and IMF5-IMF6 as high-frequency decomposition components. The low-frequency component group (IMF1-IMF4) preserves the trend of the original air quality data and removes the influence of random noise effectively, displaying clear time-series characteristics. As Informer exhibits better generalization performance for long-time series, we utilized it for handling the decomposition group of IMF1-IMF4. The high-frequency component group (IMF5-IMF6) presents a data type resembling Gaussian white noise. As XGBoost has a remarkable performance in predicting high-frequency sequences, we employed DBO-XGBoost for the prediction of IMF5-IMF6.

(2) Four parameters of XGBoost were optimized using DBO: maximum tree depth, learning rate, number of trees, and subsample, and the results are shown in Table 5.

TABLE 5. Parameters of the XGBoost model.

Parameter Name	Parameter Value
Learning Rate	0.7
Maximum tree depth	7
Number of trees	226
Subsample	0.96

E. NUMERICAL EXPERIMENT SETUP

Three sets of comparison experiments were designed in this paper to verify the prediction effect of the DBO-VMD-Informer-XGBoost model.

Comparison experiment 1

This experimental group includes some base models and improved models related to the model proposed in this paper, as shown in Table 6-I.

Comparison experiment 2

To reflect the competitiveness of the model, this experimental group contains a part of the newly proposed machine learning model and a part of the coupled optimization model, as shown in Table 6-II.

Comparison experiment 3

In order to determine the strong performance of DBO in this model, we change the optimization algorithm of the model and compare them with DBO to highlight the superior performance of the DBO used in the new model, as shown in Table 6-III.

The evaluation metrics selected for the comparison experiments are R-squared (R^2), root mean square error (RMSE), and mean absolute error (MAE) [7]. The formulas for calculating the three indicators are as follows.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\frac{1}{n}\sum_{i=1}^{n}y_i - y_i)^2} \quad (19)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (20)$$

$$MAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|} \quad (21)$$

TABLE 6. Prediction models of experiment group 1.

Predictive Model	Model Introduction
VMD-Informer-XGBoost	The novel coupled prediction model proposed in this paper without DBO optimization.
EMD-ELM-XGBoost [12]	X. Gao used EMD to classify the data and then used ELM and XGBoost to forecast the air quality of the data.
SC-RS-XGBoost [5]	H. Chen first used Spearman coefficients for correlation analysis, followed by a Random search algorithm with XGBoost coupled model without data decomposition.
Informer [9]	H. Dong used the Informer single prediction model.
XGBoost [6]	B. Czernecki used the XGBoost single prediction model.

TABLE VI- II
PREDICTION MODELS OF EXPERIMENT GROUP 2

Predictive Model	Model Introduction
WANNs [14]	Q. Guo used ANNs and WANNs to predict daily PM2.5 concentrations in Shanghai.
BSMRM [15]	Y. Lee used the recently developed BSMRM to predict the potential contribution of selected unmonitored sites to ambient PM2.5.
GAM [16]	Z. Gao used GAM to estimate historical trends and future projections of emissions and climate impacts of PM2.5 concentrations in California.
LDSPM [17]	Y. Zhang developed an LDSPM spatiotemporal model to predict PM2.5 concentrations based on the K-Core idea and label distribution.
DLNM [18]	H. Sun used DLNM to set warning thresholds for air quality prediction.
CNN-LSTM [10]	T. Li used CNN to extract air quality features and LSTM for short-term time series prediction.

TABLE VI-III
PREDICTION MODELS OF EXPERIMENT GROUP 3

Predictive Model	Model Introduction
GWO-VMD-Informer-XGBoost	Optimize parameters of VMD and XGBoost by GWO
SSA-VMD-Informer-XGBoost	Optimize parameters of VMD and XGBoost by SSA
PSO-VMD-Informer-XGBoost	Optimize parameters of VMD and XGBoost by PSO
SMA-VMD-Informer-XGBoost	Optimize parameters of VMD and XGBoost by SMA
AVOA-VMD-Informer-XGBoost	Optimize parameters of VMD and XGBoost by AVOA

F. NUMERICAL EXPERIMENT RESULTS

The experimental comparison of the three groups of models established by the above numerical experiments is conducted to obtain the predicted values of each of the three groups of comparison experiments. We analyze the experimental results visually and obtain the comparison graphs of the predicted results of the three groups of experiments, as shown in Fig. 8.

Comparing the predicted data graphs with the original data, we observe that each model's expected results follow similar trends, and each group's results show similar trends with the observed values. It indicates that each group's experimental setup yielded relatively valid prediction data. However, the models differ in their prediction accuracy, highlighting differences in air quality prediction performance between the models. In Experiment 1, the differences in the models' prediction performances are evident in intervals [46,56] and [90,100]. Fig. 6(a) demonstrates that the new model is closest to the observed values, while the more traditional inform model yields inferior results. In comparison to the new hybrid model proposed in this paper, the prediction accuracy of the five models in the comparison group is lower, and the variation nodes differ more from the original observations.
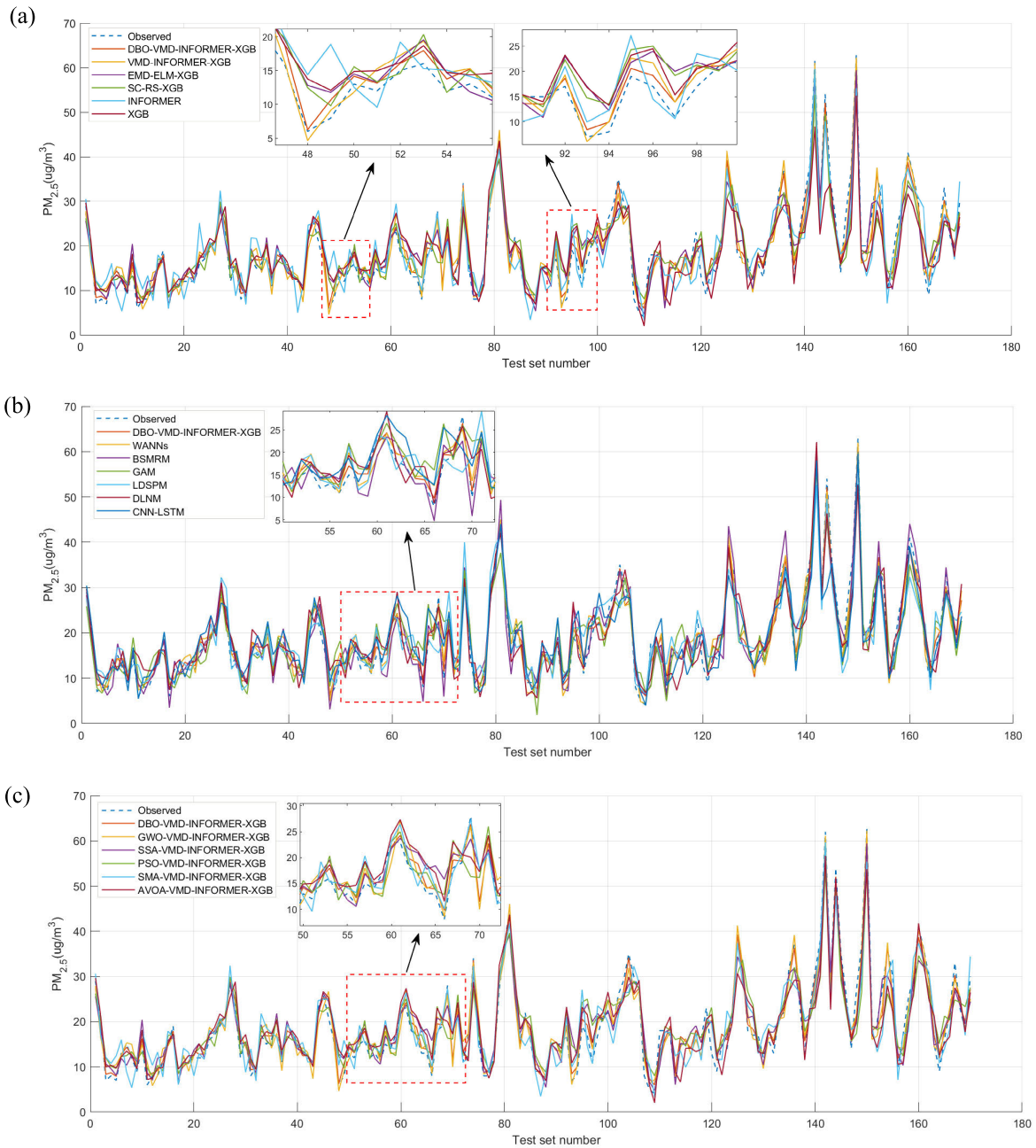
FIGURE 6. Multiple models comparison diagram.

In Experiment 2, the difference in the prediction performance of the models occurs in the interval [57,67]. Fig. 6(b) shows that the new model is closest to the observed values, while the GAM and BSMRM models are poorly fitted in this interval. The prediction accuracy of the six models in the comparison group is lower than the new hybrid model proposed in this paper. The prediction results differ from the original observations. In Experiment 3, the difference in prediction performance occurs in the interval [58,67]. Fig. 6(c) reveals that the DBO optimization parameters model is closest to

the observed values, whereas the PSO optimization model produces the worst prediction results. Comparatively, the prediction accuracy of the other three models in the comparison group is lower than the new hybrid model's accuracy. The predicted effects of each model group show a roughly similar trend. Notably, the forecast results of the new hybrid model are closest to the original observations, reflecting the superior performance of the new model in the field of air quality prediction. We conducted a detailed comparison analysis of the predicted values of DBO-VMD-Informer-XGBoost with
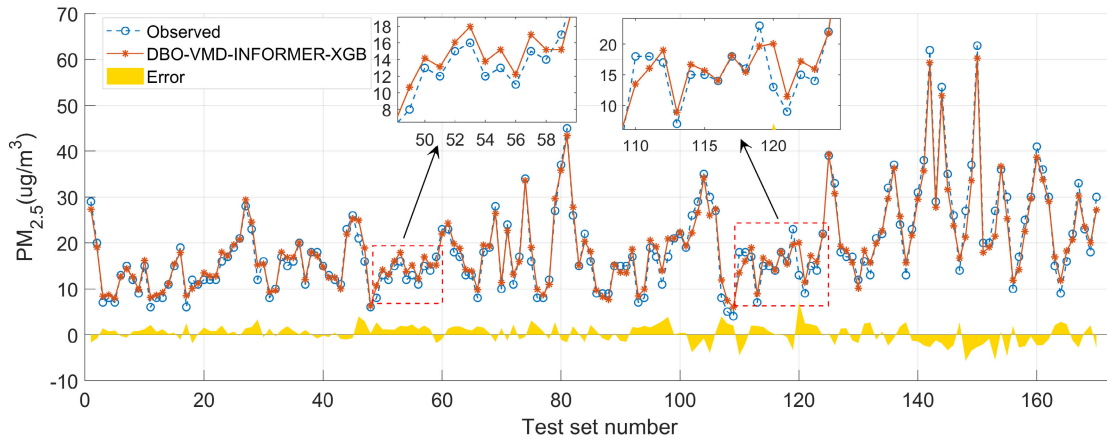
FIGURE 7. Comparison of new model prediction and observation.

the observed values, as shown in Fig. 7. A visual inspection of the elaborate comparison plots shows the differences between the new model and the original data, which visually reflects the model's superior performance.

The yellow area in Fig.7 shows the residuals after model prediction, the solid blue line indicates the actual PM2.5 data, and the red dashed line indicates the prediction result of the new model. The trends of the two curves are close, and the degree of proximity is high. The residuals do not fluctuate beyond [-7,8], and 97% of the errors are within the range of [-4,4]. The comparison results show that the forcast results of the new model are more accurate, and the residual difference between the prediction results and the original observations is smaller, which reflects the superior effect of the new hybrid predict model.

To demonstrate the superior performance of the novel coupled model in air quality prediction, we used R-Square, MAE, and RMSE error indicators shown in (19), (20), and (21) to analyze the results quantitatively. The evaluation results for different models are presented in Table 7.

In Experiment 1, we compared the new model with the basic and improved models proposed in this study to demonstrate the new model's rationality. The new model outperformed the other five models, exhibiting the best prediction performance. Compared with the best EMD-ELM-XGBoost model, R^2 improved by 9.3% to 0.961, RMSE decreased by 42.9% to 1.988, and MAE decreased by 40.1% to 1.624. Moreover, the new model outperformed the VMD-Informer-XGBoost model, which was not optimized by DBO, with improved R^2 by 3.7%, RMSE decreased by 27.0%, and MAE decreased by 26.5%. These results confirm the new model's superior prediction accuracy.

In Experiment 2, we compared the new model to a selection of newly proposed machine learning models and coupled optimization models to reflect the model's competitiveness. Among the other six new prediction models, WANNs performed the best, with the highest R^2 and the lowest RMSE and MAE ($R^2 = 0.935$, RMSE = 2.456, MAE = 2.032).

TABLE 7. Evaluation indicators for each method.

Predictive Model	R^2	RMSE	MAE
DBO-VMD-Informer-XGBoost	0.961	1.988	1.624
VMD-Informer-XGBoost	0.926	2.723	2.209
EMD-ELM-XGBoost [12]	0.879	3.481	2.710
SC-RS-XGBoost [5]	0.822	4.220	3.269
Informer [9]	0.813	4.327	3.411
XGBoost [6]	0.776	4.734	3.530
WANNs [14]	0.935	2.546	2.032
BSMRM [15]	0.889	3.331	2.597
GAM [16]	0.850	3.880	3.038
LDSPM [17]	0.829	4.146	3.278
DLNM [18]	0.855	3.820	3.118
CNN-LSTM [10]	0.835	4.069	3.102
GWO-VMD-Informer-XGBoost	0.905	3.094	2.405
SSA-VMD-Informer-XGBoost	0.876	3.534	2.862
PSO-VMD-Informer-XGBoost	0.842	3.981	3.206
SMA-VMD-Informer-XGBoost	0.915	2.889	2.219
AVOA-VMD-Informer-XGBoost	0.906	2.986	2.370

Nonetheless, the new hybrid model outperformed WANNs, with R^2 improved by 2.69%, RMSE decreased by 21.89%, and MAE decreased by 20.05%. After evaluating and comparing the proposed coupled model to a series of novel models, our findings demonstrate that the proposed model operates very satisfactorily.

In Experiment 3, we built the experiment using five different optimization strategies to demonstrate the strong performance of the DBO algorithm. The model combining the SMA optimization algorithm performed the best among the other five compared algorithms, with a score of $R^2 = 0.915$, RMSE $= 2.889$, and MAE $= 2.219$. However, the new model improved R^2 by 5.02%, reduced RMSE by 31.19%, and reduced MAE by 26.81% compared to the model with GWO optimization parameters, thereby reflecting its suitability and accuracy for PM2.5 concentration prediction. The use of DBO significantly improved the performance of the new model.

Additionally, Fig. 8 indicates that DBO converges the fastest and best when optimizing the same set of parameters, reaching the best fitness value in less than 20 iterations. DBO's convergence rate is much faster than the other three algorithms, and it converges to the smallest fitness value, demonstrating optimal convergence. The new model demonstrates accurate and stable prediction performance.
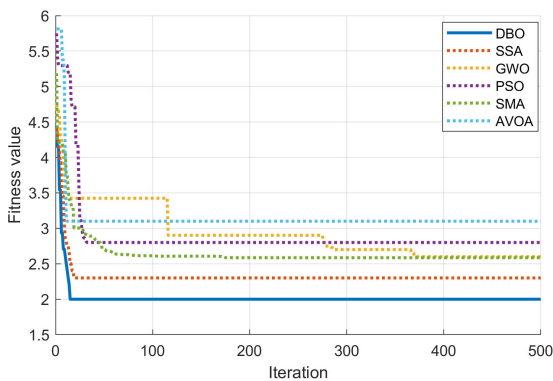

FIGURE 8. Convergence curves of the four algorithms.

Meanwhile, to ensure that the data predicted by the new model is statistically significant, the Spearman coefficient method is used to calculate the correlation between the results and the original data, and the correlation coefficient is shown below.

TABLE 8. The correlation coefficient.

	Original observed	Predict results
Original observed	1	0.968
Predict results	0.968	1

And the significance p value was 0.000, indicating that there is a strong correlation between the two groups of data, and the predict result is statistically significant.

Considered together, the DBO-VMD-Informer-XGBoost model performs the best compared with other optimized models in the three sets of experiments with $R^2 = 0.961$, RMSE $= 1.988$, and MAE $= 1.624$. The most advanced WANNs model predicted results with evaluation metrics of

$R^2 = 0.935$, RMSE $= 2.456$, and MAE $= 2.032$. Compared to this model, the new proposed model effectively improves the R^2 of the prediction results and reduces the RMES and MAE of the forecast results. The new model fits the original data better, significantly improves prediction accuracy, and reflects higher stability. The Spearman coefficient method filters the influencing factors to avoid overfitting or feature loss. The DBO algorithm significantly reduces the error caused by human empirical tuning. The coupled VMD-Informer-XGBoost is used to reduce the dataset's complexity, capture the dataset's features well and improve the accuracy and generalization of the pattern prediction. However, we are concerned about the limitation that the prediction time consumption will increase relatively due to the increase of model steps and the high computational complexity.

G. THE TRADEOFF BETWEEN ACCURACY AND WORKLOAD

To address the previously proposed limitations of the new model, we must balance model accuracy and workload. Following an in-depth study and review of numerous literature sources, we have found that algorithm complexity analysis can be assessed using two aspects: time complexity and space complexity. Space complexity primarily measures the extra space required for an algorithm to execute. Given that modern computers have significantly expanded storage capacity and the proposed model in this paper does not require an exceptionally large amount of data, we consider time complexity to be the primary measure of a new algorithm's complexity. The time complexity of swarm intelligence optimization algorithms mainly depends on the population size, the number of iterative evolutions, and the complexity of iterative operations. Generally, the iterative operation complexity of swarm optimization algorithms is lower than that of the other two influencing factors. As a result, we believe its complexity primarily relies on the population size and the number of iterations. Assuming a population size of 'N' and 'T' evolutionary iterations, we conduct a complexity analysis of the DBO, GWO, SSA, PSO, SMA, and AVOA algorithms in this paper and calculate their time complexity to be O(NT). The VMD-Informer-XGBoost algorithm's time complexity is less than O(NT). The time complexity of DBO-VMD-Informer-XGBoost, GWO-VMD-Informer-XGBoost, SSA-VMD-Informer-XGBoost and, other algorithms in the control group 3 is O(NT). However, the time complexities of experiments 1 and 2 are less than O(NT), making it challenging to directly compare the specific workload differences between the new model and other models by analyzing time complexity alone. Therefore, we visually compare workloads by measuring each model's running time. We measured the computational time of all models in the aforementioned experiments, as presented in Table 9.

The DBO-VMD-Informer-XGBoost model proposed in this paper took 25.302 seconds to process the same dataset,

TABLE 9. Multiple model runtime.

Predictive Model	Running Time (s)
DBO-VMD-INFORMER-XGBOOST	25.302
VMD-INFORMER-XGBOOST	10.018
EMD-ELM-XGBOOST	18.528
SC-RS-XGBOOST	85.302
INFORMER	4.192
XGBOOST	3.239
WANNs	21.916
BSMRM	23.022
GAM	17.389
Predictive Model	Running Time (s)
LDSPM	26.904
DLNM	17.578
CNN-LSTM	53.392
GWO-VMD-INFORMER-XGBOOST	27.294
SSA-VMD-INFORMER-XGBOOST	33.662
PSO-VMD-INFORMER-XGBOOST	117.366
SMA-VMD-INFORMER-XGBOOST	41.029
AVOA-VMD-INFORMER-XGBOOST	37.982

which was faster than the average operation time of all models (32.595 seconds), despite the model's complexity. In Experimental Group 1, the new model's runtime was prolonged by approximately 6.8 seconds compared to the best model EMD-ELM-XGBOOST, which had a runtime of 18.528 seconds. However, the proposed algorithm significantly improved prediction accuracy, reducing the RMSE value from 3.481 to 1.988, a reduction of 42.9%. Meanwhile, although the basic model Informer and XGBoost had the shortest running time, their prediction accuracy was far inferior to the new model. Similarly, in Experimental Group 2, the new model's runtime was prolonged by approximately 3.4 seconds compared to the best model WANNs, which had a runtime of 21.916 seconds. However, the new model's prediction performance was competitive, with the RMSE value decreasing from 2.546 to 1.988, a reduction of 21.89%. In Experimental Group 3, the runtime of the model optimized by GWO was 27.294 seconds, while the new proposed model took 25.302 seconds. The new model's adoption of DBO significantly reduced workload and computation time while improving prediction accuracy. We can thus weigh the relationship between accuracy and workload and conclude that despite the workload increase, the new model's prediction performance is greatly improved, and the workload and time requirements are acceptable and moderate. The proposed model is reasonable and contributes to the development of forecasting models.

IV. CONCLUSION

In order to improve the accuracy of air quality prediction model, this paper absorbs the advantages of several existing optimization prediction models and proposes the nem coupled DBO-VMD-Informer-XGBoost optimization prediction model. We first select the influencing factors by Spearman coefficient method. To test the optimization performance of DBO after tuning the parameters, we select and use six benchmark functions and 18 problems from 57 real-world optimization problems. Moreover, we adopt PSO, GWO, SSA, SMA and AVOA optimization algorithms as comparisons to verity the outstanding performance of DBO. In order to reduce the impact of data complexity on the prediction, the new model adopts DBO-VMD, accurately decomposes the complex data, and uses the appropriate prediction algorithm Informer and DBO-XGBoost for data decomposition groups with different characteristics. R^2, RMSE, MAE, and the running time of each model are used as comparison indexes to verify the prediction performance and rationality of the proposed model. The comparison of various metrics shows that the proposed new model has the highest prediction accuracy. At the same time, the running time of the model is acceptable and moderate. In conclusion, the proposed model is reasonable and greatly contributes.

In conclusion, this paper contributes to the reasonable and improved the coupled DBO-VMD-Informer-XGBoost optimization prediction model, which enhances air quality prediction accuracy and contributes to relevant fields. The previously mentioned analysis indicates that the proposed model accurately characterizes the original data set, mitigates data complexity, and avoids errors from the artificial adjustment of model parameters. The model significantly enhances prediction accuracy, possessing robust stability and generalization abilities since it incorporates relevant influencing factors and is not solely a time series prediction model. The model captures the data set's inherent characteristics simultaneously with the influencing factors, ensuring prediction accuracy remains unaffected despite changes in the prediction object. The new model is believed to exhibit similar excellent performance in predicting pollutants beyond PM2.5.

The innovations and advantages of the model are: (a) For the first time in the field of data prediction, the advanced DBO algorithm with excellent performance is adopted, which relatively reduces the running time and, at the same time, greatly reduces the errors caused by human empirical tuning. The prediction accuracy is greatly improved and the performance far exceeds the similar prediction models in the recent literature. (b) For the first time in the field of air quality prediction, the innovative use of coupled VMD-Informer-XGBoost enables the model to capture the characteristics of the data set to a great extent, improving the accuracy and generalisation of the model prediction. After using DBO-VMD decomposition, the data complexity is greatly reduced. And the Informer and DBO-XGBoost algorithms are adopted to deal with the decomposed data with different features.

TABLE 10. Selected benchmark functions.

Functions	Dimension	Search space
$F_1(x) = \sum_{i=1}^{n} x_i^2$	30	$-100 \le x_i \le 100$
$F_7(x) = \sum_{i=1}^{n} i x_i^4 + \text{random}[0,1)$	30	$-1.28 \le x_i \le 1.28$
$F_9(x) = \sum_{i=1}^{n} [x_i^2 - 10\cos(2\pi x_i) + 10]$	30	$-5.12 \le x_i \le 5.12$
$F_{10}(x) = -20\exp(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}) - \exp(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)) + 20 + e$	30	$-32 \le x_i \le 32$
$F_{16}(x) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1 x_2 - 4x_2^2 + 4x_2^4$	2	$-5 \le x_i \le 5$
$F_{18}(x) = [1 + (x_1 + x_2 + 1)^2 (19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1 x_2 + 3x_2^2)]$ $\times [30 + (2x_1 - 3x_2)^2 (18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1 x_2 + 27x_2^2)]$	2	$-2 \le x_i \le 2$

The prediction error due to the complexity of the dataset is substantially reduced. (c) The influencing factors are filtered by Spearman coefficient method to avoid over-fitting or feature loss and reduce computational effort.

The model's limitations and future perspectives include: (a) The new model mixes multiple algorithms to improve prediction accuracy; however, this increases workload. The runtime of 25.302s compared to other models in the experimental group is acceptable. Nevertheless, its complex workload needs simplification since it requires higher technical expertise from the operator, which impedes the model's future forecasting application. Future research can develop new techniques to save prediction time. (b) The present experiments are limited to Nanjing City and require further extension across different regions for air quality or other prediction applications in the future.

APPENDIX A
DBO PERFORMANCE TEST
The paper sets $n = 50$, $t = 500$, $k = 0.14$, $b = 0.25$, $S = 0.5$ in equations (1-5). This set of parameters performs best. After that, the performance tests in the benchmark problem as well as the performance tests in the real problems are conducted for the DBO after the parameterization.

(1) Testing of DBO on six sets of benchmark problems

There are a total of three classes of classical test problems [38], namely unimodal benchmarks, multimodal benchmarks, and complex modal benchmarks. In this paper, we select two benchmark functions from each type of problem and evaluate the search performance of DBO. F_1 and F_7 are unimodal benchmark functions, F_9 and F_{10} are multimodal benchmark functions, and F_{16} and F_{18} are composite modal benchmark functions, as shown in Table 10.

In Fig. 9, the 3D shape of test functions used to evaluate the efficiency of DBO has been illustrated, with function values appearing in blue, yellow, and red from low to high. The DBO algorithm's performance is tested using the unimodal

benchmark F_1 versus F_7, resulting in the derivation of the global optimal solution with ease. However, the algorithm may fall into local optimal solutions when dealing with multimodal and composite modal problems compared to the unimodal benchmark problem. Nevertheless, as depicted in Fig. 5, DBO can obtain optimal solutions effectively within the search space. Therefore, the six benchmark function problems above can successfully identify DBO's superior search performance. Fig. 10 portrays the convergence function of the best fitness value when DBO searches for the globally optimal solutions for the aforementioned six benchmark problems. The horizontal axis indicates the number of iterations, while the vertical axis represents the best fitness value.

Fig. 10 reveals that DBO can efficiently converge towards the globally optimal solution when dealing with the last five problems. Therefore, DBO exhibits the ability to search for the optimal global solutions to the six benchmark problems mentioned above after parameter tuning. Besides the six benchmark functions tested on DBO, similar tests are conducted on Particle Swarm Algorithm (PSO), Sparrow Search Algorithm (SSA), Gray Wolf Optimization Algorithm (GWO), Slime Mould Algorithm (SMA), and African Vultures Optimization Algorithm (AVOA), resulting in a total of 36 benchmark experiments. In this study, each group of benchmark experiments is performed independently 30 times. The average value of the 30 independent experiments is then determined as the best value for each group of the benchmark experiments using the formula shown below. Where p is the number of optimization experiments, $F_{i,j}$ is the best value of the jth independent repetition experiment, and M_i is the best value of the ith benchmark experiment for each algorithm.

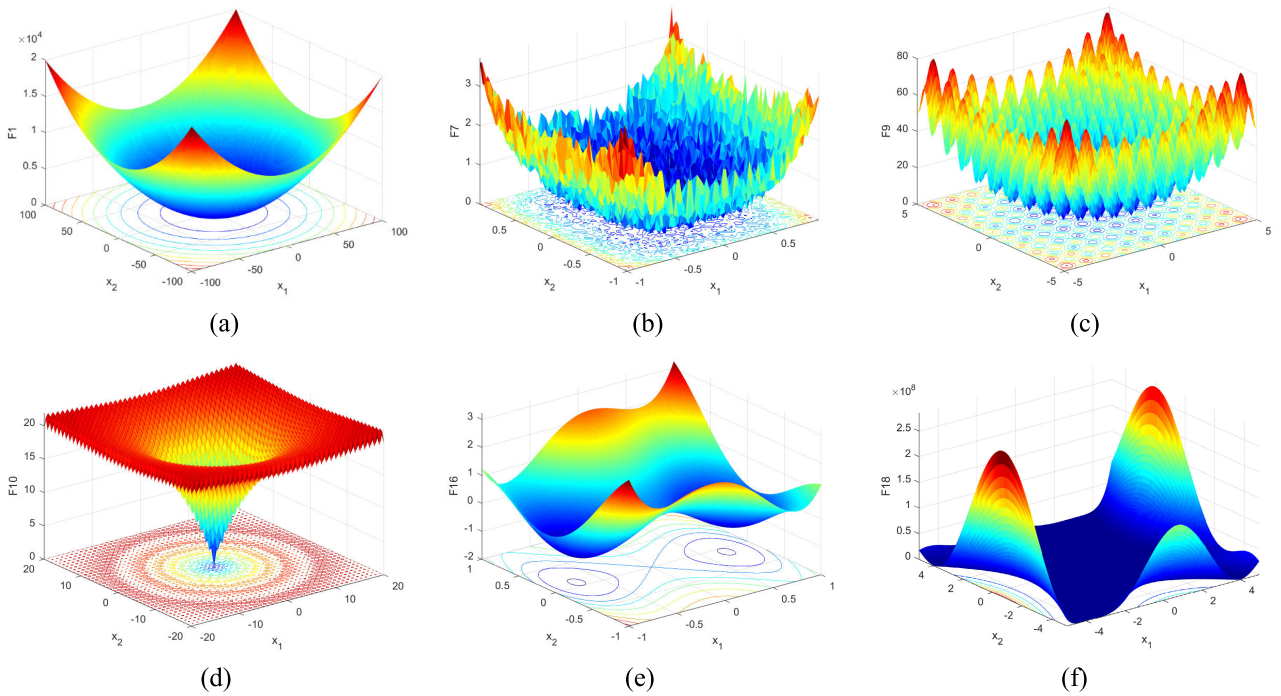$$M_i = \frac{1}{p}\sum_{j=1}^{p} F_{i,j} \qquad (22)$$

FIGURE 9. Three-dimensional surface plots of the six benchmark functions.
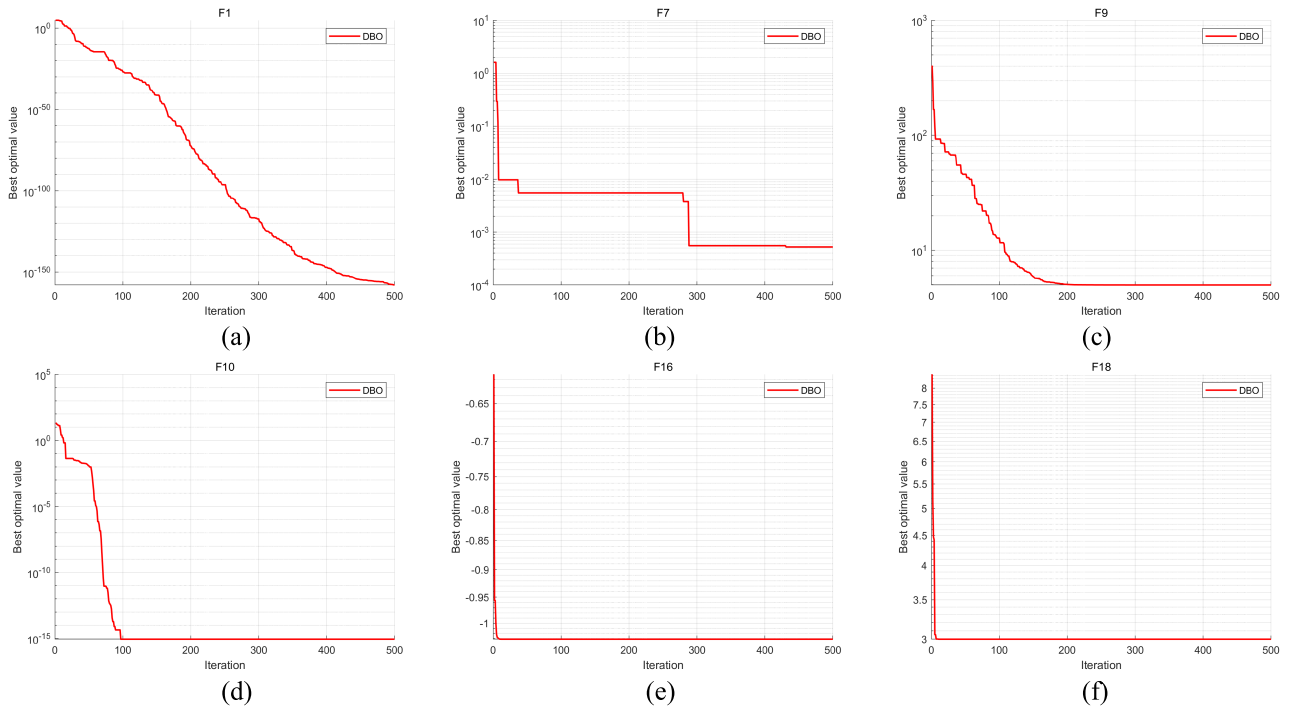

FIGURE 10. Convergence curve of the best fitness value.

In this paper, the standard deviation STD is derived from the theoretical best value M of the six benchmark problem experiments and the average solution best value M_i. When optimizing different dimensional problems, the more

stable the performance of the algorithm in searching for the best solution is when the STD value is smaller. The difference between the solution and the theoretical best value is minor and more accurate. Therefore, the performance

TABLE 11. Optimal values for different benchmark questions (mean values of 30 independent repetitions).

Functions	PSO	GWO	SSA	SMA	AVOA	DBO	Theoretical best value
F_1	4.09E-05	1.82E-13	2.49E-11	2.09E-21	2.37E-11	1.75E-153	0
F_7	1.3892	0.008248	0.02751	0.006529	0.01783	0.001514	0
F_9	35.2838	5.2371	19.2934	3.0213	10.9382	0	0
F_{10}	2.4679	7.24E-04	0.05943	6.28E-06	9.43E-04	8.88E-16	0
F_{16}	-1.03162	-1.03162	-1.03163	-1.03162	-1.03162	-1.0316	-1
F_{18}	3	3	3.002	3	3	3	3
STD	15.83014	2.34214	8.62833	1.35124	4.89173	0.01414	

TABLE 12. Optimal values for real-word problems (mean values of 30 independent repetitions).

Category	Problem	PSO	GWO	SSA	SMA	AVOA	DBO	Best result
ICP	RC1	193.0012	191.0245	188.9281	187.3825	190.0247	189.34423	189.3116
	RC2	7059.2343	7065.9701	7032.8203	7044.9271	7036.3829	7049.0216	7049.0369
	RC3	-4491.276	-4539.298	-4512.723	-4541.638	-4519.827	-4527.803	-4529.119
PSDP	RC8	2.8937	2.9722	1.9221	2.3283	2.2903	2.0011	2
	RC9	2.9601	2.0214	2.4527	2.6375	2.4781	2.5591	2.5576
	RC10	1.3229	1.1009	0.9312	1.2088	1.1379	1.0771	1.0765
MEP	RC15	3005.0187	2998.7908	2985.2395	2998.6719	2987.5346	2998.4613	2994.4244
	RC16	0.0298	0.0423	0.0412	0.0412	0.0399	0.0322	0.0322
	RC17	0.0126	0.0129	0.0128	0.0129	0.0126	0.0126	0.0126
PSP	RC34	7.1209	4.2192	2.4219	3.2032	2.1628	1.0231	0
	RC35	15.5913	4.2102	1.3986	1.7826	2.6702	0.8772	0.0890
	RC36	11.6938	3.2497	4.9674	1.9927	3.6983	1.8292	0.0720
PEP	RC45	3.0021	0.0399	0.0402	0.0526	0.0312	0.0381	0.0380
	RC46	0.0402	0.0207	0.0296	0.0286	0.0335	0.0211	0.0212
	RC47	0.0255	0.0377	0.0174	0.0189	0.0328	0.0127	0.0151
Category	Problem	PSO	GWO	SSA	SMA	AVOA	DBO	Best result
LFRO	RC51	4596.9784	4549.4982	4551.3233	4554.2885	4550.2917	4559.0135	4550.8511
	RC52	3326.7593	3389.2345	3392.8791	3370.8923	3385.6429	3346.3711	3348.9821
	RC53	4997.9003	5003.2648	4988.0017	5001.7872	4991.3804	5000.2381	4997.6069
STD		17.173379	10.898629	12.443288	6.4808385	9.8977726	3.2241672	

of each algorithm in six benchmark experiments can be evaluated comprehensively using STD with the following equation where q is the type of benchmark experiment for each algorithm.

$$STD = \sqrt{\frac{1}{q-1} \sum_{i=1}^{q} (M_i - M)^2} \qquad (23)$$

The test results are shown in Table 11.

In the six benchmark problems, there are multiple locally optimal solutions for the multimodal F_9, F_{10}, which requires a high performance for the global search of the algorithm.

The performance differences between the algorithms can be well reflected. From Table 9, it can be seen that PSO has the worst search result, and DBO optimization produces the best value closest to the theoretical optimum, which can successfully avoid the dilemma of falling into the local optimum solution. Using the STD values of each algorithm, the performance of the algorithms is considered comprehensively. DBO has the smallest STD value. In other words, the optimal global solution searched by DBO is always the closest to the theoretical optimal value under different benchmark problems. None of the algorithms in PSO, SSA, GWO, SMA or AVOA can outperform DBO. the tuned DBO has high

accuracy and stability in searching for the optimal global solution.

(2) Testing of DBO on real-world optimization problems

Due to the complexity of real-world optimization problems, the requirements for the algorithms are high. Therefore, it is essential to evaluate the effectiveness of DBO in solving real-world problems. It follows from (1) that DBO can successfully pass a portion of the benchmark problems. Kumar A presented 57 real-world optimization problems [39]. A total of 6 broad categories were classified, including seven Industrial chemical processes (ICP), seven Process synthesis and design problems (PSDP), nineteen Mechanical engineering problems (MEP), eleven Power system problems (PSP), six Power electronic problems (PEP), seven Livestock feed ration optimization (LFRO). This paper selects three problems from each broad category for quizzing, i.e., 18 problems. Furthermore, the four algorithms in (1) are adopted to deal with the 18 problems to obtain 72 sets of quizzes. We experiment in the same way as the benchmark problem quiz. Each group of quizzes is repeated 30 times independently, and the mean value is taken as the optimized realistic best value. The theoretical best result of each real-world problem is obtained based on [40], and calculate the STD value of the experimental best value and the theoretical best result.

This paper uses SSA, GWO, PSO, SMA and AVOA as control groups for the DBO algorithm. Based on Table 12, DBO can effectively handle real-world optimization problems. In addition, DBO has the smallest STD value. In other words, the tuned DBO applies to various types and dimensions of problems. It also shows the high stability and accuracy of using DBO to optimize the required parameters of VMD and XGBoost, avoiding the significant errors caused by artificially selected parameters. Therefore, the article chooses DBO as the algorithm for optimising the parameters.

REFERENCES

[1] L. Yang, Y. Zhang, W. Qi, T. Zhao, L. Zhang, L. Zhou, and L. Ye, "Adverse effects of PM$_{2.5}$ on cardiovascular diseases," *Rev. Environ. Health*, vol. 37, no. 1, pp. 71–80, Mar. 2022.

[2] H. Brumberg, C. J. Karr, A. Bole, S. Ahdoot, S. J. Balk, A. S. Bernstein, and L. G. Byron, "Ambient air pollution: Health hazards to children," *Pediatrics*, vol. 147, no. 6, pp. 1–12, Jun. 2021.

[3] C.-C. Liu, T.-C. Lin, K.-Y. Yuan, and P.-T. Chiueh, "Spatio-temporal prediction and factor identification of urban air quality using support vector machine," *Urban Climate*, vol. 41, Jan. 2022, Art. no. 101055.

[4] S. Abirami and P. Chitra, "Regional air quality forecasting using spatiotemporal deep learning," *J. Cleaner Prod.*, vol. 283, Feb. 2021, Art. no. 125341.

[5] H. Chen, G. Deng, and Y. Liu, "Monitoring the influence of industrialization and urbanization on spatiotemporal variations of AQI and PM$_{2.5}$ in three provinces, China," *Atmosphere*, vol. 13, no. 9, p. 1377, Aug. 2022.

[6] B. Czernecki, M. Marosz, and J. Jędruszkiewicz, "Assessment of machine learning algorithms in short-term forecasting of PM10 and PM2.5 concentrations in selected Polish agglomerations," *Aerosol Air Quality Res.*, vol. 21, no. 7, 2021, Art. no. 200586.

[7] N. H. Van, P. Van Thanh, D. N. Tran, and D.-T. Tran, "A new model of air quality prediction using lightweight machine learning," *Int. J. Environ. Sci. Technol.*, vol. 20, no. 3, pp. 2983–2994, May 2022.

[8] D. Kothandaraman, N. Praveena, K. Varadarajkumar, B. M. Rao, D. Dhabliya, S. Satla, and W. Abera, "Intelligent forecasting of air quality and pollution prediction using machine learning," *Adsorption Sci. Technol.*, vol. 2022, pp. 1–15, Jun. 2022.

[9] H. Dong and L. Sun, "Prediction of PM$_{2.5}$ concentration based on informer," *Environ. Eng.*, vol. 6, no. 40, pp. 48–54, Mar. 2022.

[10] T. Li, M. Hua, and X. Wu, "A hybrid CNN-LSTM model for forecasting particulate matter (PM$_{2.5}$)," *IEEE Access*, vol. 8, pp. 26933–26940, 2020.

[11] A. Pranolo, Y. Mao, A. P. Wibawa, A. B. P. Utama, and F. A. Dwiyanto, "Robust LSTM with tuned-PSO and bifold-attention mechanism for analyzing multivariate time-series," *IEEE Access*, vol. 10, pp. 78423–78434, 2022.

[12] X. Gao, K. Zhou, and L. Li, "Mixed air quality warning system based on XGBOOST and ELM: A case study of Nanjing," *Chin. J. Manag. Sci.*, vol. 2022, pp. 1–12, Jan. 2022.

[13] Y. Chen, R. Shi, S. Shu, and W. Gao, "Ensemble and enhanced PM$_{10}$ concentration forecast model based on stepwise regression and wavelet analysis," *Atmos. Environ.*, vol. 74, pp. 346–359, Aug. 2013.

[14] Q. Guo, Z. He, and Z. Wang, "Predicting of daily PM$_{2.5}$ concentration employing wavelet artificial neural networks based on meteorological elements in Shanghai, China," *Toxics*, vol. 11, no. 1, p. 51, Jan. 2023.

[15] Y. S. Lee, J. Y. Kim, S.-M. Yi, H. Kim, and E. S. Park, "Predicting latent source-specific PM$_{2.5}$ pollution from regional sources at unmonitored sites by Bayesian spatial multivariate receptor modeling," *Environ. Pollut.*, vol. 324, May 2023, Art. no. 121389.

[16] Z. Gao, C. E. Ivey, C. L. Blanchard, K. Do, S.-M. Lee, and A. G. Russell, "Emissions, meteorological and climate impacts on PM$_{2.5}$ levels in Southern California using a generalized additive model: Historic trends and future estimates," *Chemosphere*, vol. 325, Jun. 2023, Art. no. 138385.

[17] Y. Zhang and Q. Yan, "A spatiotemporal model for PM$_{2.5}$ prediction based on the K-core idea and label distribution," *Meteorolog. Appl.*, vol. 30, no. 1, pp. 1–12, Feb. 2023.

[18] H. Sun, S. Chen, X. Li, L. Cheng, Y. Luo, and L. Xie, "Prediction and early warning model of mixed exposure to air pollution and meteorological factors on death of respiratory diseases based on machine learning," *Environ. Sci. Pollut. Res.*, vol. 30, no. 18, pp. 53754–53766, Mar. 2023.

[19] N. R. Kapoor, A. Kumar, A. Kumar, D. A. Zebari, K. Kumar, M. A. Mohammed, A. S. Al-Waisy, and M. A. Albahar, "Event-specific transmission forecasting of SARS-CoV-2 in a mixed-mode ventilated office room using an ANN," *Int. J. Environ. Res. Public Health*, vol. 19, no. 24, p. 16862, Dec. 2022.

[20] N. R. Kapoor, A. Kumar, A. Kumar, A. Kumar, M. A. Mohammed, K. Kumar, S. Kadry, and S. Lim, "Machine learning-based CO$_2$ prediction for office room: A pilot study," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–16, Mar. 2022.

[21] X. Yao, X. Fu, and C. Zong, "Short-term load forecasting method based on feature preference strategy and LightGBM-XGboost," *IEEE Access*, vol. 10, pp. 75257–75268, 2022.

[22] J. Li and Q. Liu, "Forecasting of short-term photovoltaic power generation using combined interval type-2 Takagi–Sugeno-Kang fuzzy systems," *Int. J. Electr. Power Energy Syst.*, vol. 140, Sep. 2022, Art. no. 108002.

[23] O. Abedinia, M. Bagheri, M. S. Naderi, and N. Ghadimi, "A new combinatory approach for wind power forecasting," *IEEE Syst. J.*, vol. 14, no. 3, pp. 4614–4625, Sep. 2020.

[24] L. Ye, B. Dai, M. Pei, P. Lu, J. Zhao, M. Chen, and B. Wang, "Combined approach for short-term wind power forecasting based on wave division and Seq2Seq model using deep learning," *IEEE Trans. Ind. Appl.*, vol. 58, no. 2, pp. 2586–2596, Mar. 2022.

[25] X. Wang, X. Gao, Z. Wang, C. Ma, and Z. Song, "A combined model based on EOBL-CSSA-LSSVM for power load forecasting," *Symmetry*, vol. 13, no. 9, p. 1579, Aug. 2021.

[26] Z. Shang, Z. He, Y. Song, Y. Yang, L. Li, and Y. Chen, "A novel combined model for short-term electric load forecasting based on whale optimization algorithm," *Neural Process. Lett.*, vol. 52, no. 2, pp. 1207–1232, Jul. 2020.

[27] J. Xue and B. Shen, "Dung beetle optimizer: A new meta-heuristic algorithm for global optimization," *J. Supercomput.*, vol. 79, no. 7, pp. 7305–7336, Nov. 2022.

[28] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 531–544, Feb. 2014.

[29] F. Li, R. Li, L. Tian, L. Chen, and J. Liu, "Data-driven time-frequency analysis method based on variational mode decomposition and its application to gear fault diagnosis in variable working conditions," *Mech. Syst. Signal Process.*, vol. 116, pp. 462–479, Feb. 2019.

[30] O. Sagi and L. Rokach, "Approximating XGBoost with an interpretable decision tree," *Inf. Sci.*, vol. 572, pp. 522–542, Sep. 2021.

[31] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[32] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI*, May 2021, vol. 35, no. 12, pp. 11106–11115.

[33] H. Wei, W.-S. Wang, and X.-X. Kao, "A novel approach to ultra-short-term wind power prediction based on feature engineering and informer," *Energy Rep.*, vol. 9, pp. 1236–1250, Dec. 2023.

[34] X. Li, L. C. Abdullah, S. Sobri, M. S. Md Said, S. A. Hussain, T. P. Aun, and J. Hu, "Long-term air pollution characteristics and multi-scale meteorological factor variability analysis of mega-mountain cities in the Chengdu-Chongqing economic circle," *Water, Air, Soil Pollut.*, vol. 234, no. 5, May 2023.

[35] D. Han, T. Zhang, X. Zhang, and Y. Tan, "Study on spatiotemporal characteristics and influencing factors of pedestrian-level $PM_{2.5}$ concentrations in outdoor open spaces of Harbin in winter, using a generalized additive model (GAM)," *Urban Climate*, vol. 46, Dec. 2022, Art. no. 101313.

[36] A. Pires, "Analysis and modelling of $PM_{2.5}$ temporal and spatial behaviors in European cities," *Sustainability*, vol. 11, no. 21, p. 6019, Oct. 2019.

[37] A. Delgado-Bonal and A. Marshak, "Approximate entropy and sample entropy: A comprehensive tutorial," *Entropy*, vol. 21, no. 6, p. 541, May 2019.

[38] E. Naderi, L. Mirzaei, M. Pourakbari-Kasmaei, F. V. Cerna, and M. Lehtonen, "Optimization of active power dispatch considering unified power flow controller: Application of evolutionary algorithms in a fuzzy framework," *Evol. Intell.*, vol. 2023, pp. 1–12, Jan. 2023.

[39] A. Kumar, G. Wu, M. Z. Ali, R. Mallipeddi, P. N. Suganthan, and S. Das, "A test-suite of non-convex constrained optimization problems from the real-world and some baseline results," *Swarm Evol. Comput.*, vol. 56, Aug. 2020, Art. no. 100693.

[40] P. Suganthan. (2020). *Guidelines for Real-World Single-Objective Constrained Optimisation Competition*. [Online]. Available: https://github.com/P-N-Suganthan/2020-RW-Constrained-Optimisation

QICHEN SHAO was born in Wuxi, Jiangsu, China, in May 2002. He is currently pursuing the B.Sc. degree in mathematics and applied mathematics with the College of Mathematical Sciences, Yangzhou University, China. His primary research interests include machine learning and deep learning, focusing on interdisciplinary mathematical modeling problems. He has won several awards in national mathematical modeling competitions and was awarded the talent qualification by the Ministry of Industry and Information Technology of China, in 2022.

JIAHAO CHEN was born in Changzhou, Jiangsu, China, in February 2002. He is currently pursuing the B.Sc. degree in mathematics and applied mathematics with the College of Mathematical Sciences, Yangzhou University, China. His research interests include operations research, cybernetics, and marginal solutions of partial differential equations with a focus on interdisciplinary mathematical modeling research. He has received several awards from national mathematical modeling competitions and national mathematical competitions.

TAO JIANG was born in Weifang, Shandong, China, in 1978. He received the M.Sc. degree in computational mathematics from the Harbin Institute of Technology, in 2007, and the Ph.D. degree in applied mathematics from Northwestern Polytechnic University, in 2012.

He conducted postdoctoral research with the Federal University of Paraná, Brazil, from 2017 to 2018 and visited Brown University, USA, in 2019. He is currently a Professor with the College of Mathematical Sciences, Yangzhou University, China. He is mainly engaged in the research of numerical algorithms and numerical simulation in the field of fluid mechanics in engineering mechanics. He focuses on the research of interdisciplinary problems, mainly involving the knowledge of the numerical solution of partial differential equations, nonlinear dynamics, fluid mechanics, multiphase flow, and material form. In the past ten years, he has published nearly 40 SCI papers on the study of pure mesh-free particle methods for viscoelastic flow problems and high-dimensional simulations and has presided over or participated in completing several national and provincial research projects.

• • •