Applied Mathematical Modelling 106 (2022) 177–198

Contents lists available at ScienceDirect

Applied Mathematical Modelling

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

# Novel hybrid extreme learning machine and multi-objective optimization algorithm for air pollution prediction

Lu Bai [a], Zhi Liu [a,*], Jianzhou Wang [b]

[a] Department of Mathematics, University of Macau, Taipa, Macao, China
[b] Institute of Systems Engineering, Macau University of Science and Technology, Taipa, Macao, China

ARTICLE INFO

*Article history:*
Received 5 July 2021
Revised 19 January 2022
Accepted 23 January 2022
Available online 14 February 2022

*Keywords:*
Mathematical modelling
Hybrid prediction model
Improved extreme learning machine
Data decomposition
Multi-objective optimization approach
Deterministic and interval predictions

ABSTRACT

A novel system regarding deterministic and interval predictions of pollutant concentration is constructed in this study, which can not only obtain higher prediction accuracy in deterministic prediction and also provide effective interval prediction of air pollutant concentration. In the deterministic prediction stage, the improved extreme learning machine combines outlier detection and correction algorithm, data decomposition strategy, and a multi-objective optimization algorithm to form a hybrid model for predicting pollutant concentration. Moreover, the applicability of the optimization algorithm was verified from theoretical and experimental analysis. In the interval prediction stage, three distributions are compared to mine, the traits of deterministic prediction errors are analyzed, and interval prediction is designed to quantify the uncertainties associated with pollutant concentration. To investigate the prediction performance of the proposed system, comparison experiments have been executed using the $PM_{2.5}$ concentration series from three cities. The results indicate that the system proposed in this paper outperforms comparison models in forecasting accuracy and has advantages for pollution prediction.



## 1. Introduction

The currently prevailing air pollution prediction methods can be divided into physical-driven models and data-driven models [1]. The physical-driven models predict the concentrations of air pollutants by simulating the transport and diffusion of pollutants in the atmosphere. The most commonly used physical-based models are based on WRF (Weather Research and Forecasting) systems. For example, based on the WRF coupled with chemisty (WRF-Chem) model, a prototype air quality modeling system was developed for Macedonia to forecasting high-resolution weather and air pollution [2]. By using the WRF system to get the input of the prediction model, a generalized additive model was developed to capture the relationships between $PM_{2.5}$ and aerosol optical depth, relative humidity, and other features to predict the $PM_{2.5}$ concentration [3]. Considering that $PM_{2.5}$ concentration in Taiwan influenced by variation of synoptic weather patterns significantly, a bias-correction method was proposed to improve the prediction performance of the WRF-CMAQ-based real-time air quality forecasting system [4].

Compared to the physical-driven models, the data-driven models are easy to implement, which has received much attention. Classification and regression tree was used to predict $PM_{10}$ concentration by capturing the non-linear relationship

* Corresponding author.
*E-mail addresses:* yb97466@umac.mo (L. Bai), liuzhi@umac.mo, liuzhi@um.edu.mo (Z. Liu), wangjz@dufe.edu.cn (J. Wang).

https://doi.org/10.1016/j.apm.2022.01.023
0307-904X/

**Nomenclature**

*Abbreviates*

| | |
|---|---|
| $A_i$ | $i_{th}$ actual value |
| $\bar{A}$ | Mean value of actual series |
| ADMM | Alternate direction method of multipliers |
| AERMOD | American Meteorological Society/Environmental Policy Agency Regulatory Model |
| MAE | Mean Absolute Error |
| AIC | Akaike information criterion |
| ANFIS | Adaptive Neuro-Fuzzy Inference System |
| ApEn | Approximate entropy |
| AQI | Air Quality Index |
| AWD | Accumulated Width Deviation |
| BLIMFs | Band-limited intrinsic mode functions |
| CDFs | Cumulative distribution functions |
| Chem | Chemistry |
| CMAQ | Community Multi-scale Air Quality |
| DM | Diebold-Mariano test |
| DP | Determinstic prediction |
| EEMd | Ensemble Empirical Mode Decomposition |
| ELM | Extreme learning machine |
| EMD | Empirical Mode Decomposition |
| $F_i$ | $i_{th}$ predicted value |
| $\bar{F}$ | Mean value of predicted series |
| $\overline{\overline{F}}$ | The maximum value of forecasting series |
| $\underline{F}$ | The minimum value of forecasting series |
| $H_1$ | moALo-$\ell_{2,1}$RFELM |
| $H_2$ | mSSa-$\ell_{2,1}$RFELM |
| $H_3$ | VMd-moALo-$\ell_{2,1}$RFELM |
| $H_4$ | VMd-mSSa-$\ell_{2,1}$RFELM |
| $H_5$ | HI-EEMd-mSSa-$\ell_{2,1}$RFELM |
| HI | Hampel Identifier |
| IMFs | Intrinsic mode functions |
| IP | Interval prediction |
| IPCP | Interval Prediction Coverage Probability |
| IPNAW | Interval Prediction Normalized Average Width |
| $L_i$ | The predicted value of lower bound of interval |
| $\ell_{2,1}$RFELM | Random Fourier ELM with $\ell_{2,1}$-norm regularization |
| LSTM | Long Short-Term Memory |
| MAD | Median Absolute Deviation |
| $MAE_{GZ}$ | The value of MAE in Guangzhou |
| $MAE_{SZ}$ | The value of MAE in Shenzhen |
| $MAE_{ZH}$ | The value of MAE in Zhuhai |
| MAPE | Mean Absolute Percentage Error |
| $MAPE_{GZ}$ | The value of MAPE in Guangzhou |
| $MAPE_{SZ}$ | The value of MAPE in Shenzhen |
| $MAPE_{ZH}$ | The value of MAPE in Zhuhai |
| MLE | Maximum likelihood estimation |
| mSSa | Multi-objective Salp Swarm Algorithm |
| P | HI-VMd-mSSa-$\ell_{2,1}$RFELM |
| PDFs | Probability density functions |
| PRD | Pearl River Delta |
| $R^2$ | R square |
| RMSE | Root Mean Square Error |
| $RMSE_{GZ}$ | The value of RMSE in Guangzhou |
| $RMSE_{SZ}$ | The value of RMSE in Shenzhen |
| $RMSE_{ZH}$ | The value of RMSE in Zhuhai |
| $RMSE_{IP}$ | Root Mean Square Error of Interval Prediction |

| | |
|---|---|
| $S_1$ | Arima |
| $S_2$ | $\ell_{2,1}$RFELM |
| SampEn | Sample Entropy |
| TLS | T Location-Scale |
| $U_i$ | The predicted value of upper bound of interval |
| VMd | Variational Mode decomposition |
| WRF | Weather Research and Forecasting |

*Functions*

| | |
|---|---|
| $erfc(\cdot)$ | Complementary error function |
| $obf_i(\cdot)$ | $i$th objective function |
| $\Gamma(\cdot)$ | Gamma function |

*Variables*

| | |
|---|---|
| $k$ | Window size (HI) |
| $r$ | Threshold of the tolerance for acceptable matches (SampEn) |
| $m$ | Length of reconstructed series (SampEn) |
| $\alpha$ | Quadratic penalty term (VMd) |
| $\lambda$ | Lagrange multipliers (VMd) |
| $\nu$ | Initialize center frequency (VMd) |
| $\tau$ | Noise tolerance (VMd) |
| $\varepsilon$ | The tolerance of the convergence criterion (VMd) |
| $G$ | Number of modes (VMd) |
| AMS | Size of archive (mSSa) |
| $\mathbf{Z}$ | The number of salp chain (mSSa) |
| $D$ | The dimension of search (mSSa) |
| $T_{mSSa}$ | Number of maximum iteration (mSSa) |
| $\mathbf{Fit}(\cdot)$ | Fitness of salp chain |
| $p_1^i$ | The position of lead in $i$th salp chain (mSSa) |
| $\overline{\overline{p_j^1}}$ | The upper bound of $p_j^1$ (mSSa) |
| $\underline{\underline{p_j^1}}$ | The lower bound of $p_j^1$ (mSSa) |
| $\overline{\overline{[P_{lb}, P_{ub}]}}$ | The dynamics interval of the parameters to be optimized (mSSa) |
| $\mathbf{P}_t$ | The position matrix of salps at $t$th iteration (mSSa) |
| $t$ | The current iteration number (mSSa) |
| $D_{hidden}$ | Number of neutons in hidden layer ($\ell_{2,1}$RFELM) |
| $p$ | Number of lags of the time series (Arima) |
| $d$ | Differential order of the time series (Arima) |
| $q$ | Number of lags of the prediction error items (Arima) |
| $\tilde{C}$ | Penalty coefficient ($\ell_{2,1}$RFELM) |
| $T_{ELM}$ | Number of maximum iteration ($\ell_{2,1}$RFELM) |

betweent $PM_{10}$ concentration series and a group of predictors [5]. The prediction performance was compared between the Adaptive Neuro-Fuzzy Inference System (ANFIS) and the semi-experimental nonlinear regression model, and pointed out that ANFIS can predict pollutant concentration more accurately [6]. Based on the data from Kuala Lumpur, Malaysia, researchers found that Singh fuzzy time series model was the most accurate and efficient forecasting model by comparing the prediction accuracy of other seven models [7]. In recent years, with the continuous development of data-driven models, hybrid models have emerged, which exhibit good robustness and adaptability by combining the advantages of different models, and are widely used in various fields. For example, a novel combined prediction system based on extrem learning machine and multi-objective grey wolf optimization was proposed to predict wind speed [8]; an improved least squares twin support vector regression model based on the robust $\ell_1$-norm distance was bilut to predict short-term traffic flow, in which the fruit fly optimization algorithm is used to optimiz the parameters of the prediction model [9]. An intelligent air quality forecasting system was proposed to predict AQI, which constructed an second-stage feature selection module and combined the multi-objective Bonobo optimization algorithm [10]. By combining the Hampel identifier, empirical wavelet transform, Elman neural network and Outlier-robust extreme learning machine, a novel hybrid algorithm was proposed, which improved the forecasting accuracy of fine particle concentrations [11]. To accurately predict the non-methane volatile organic compounds in the air caused by traffic, a multivariate nonlinear grey model based on the Gaussian kernel was proposed, which improves prediction accuracy by measuring the relationship between traffic emission and relevant factors [12]. Some scholars have also proposed air pollution early warning systems. For air quality assessment, a dynamic evaluation system was proposed in [13]. This system can not only provide accurate prediction results of various pollutants, but also evaluate the air

quality based on the prediction results by using fuzzy mathematical synthetic evaluation. In another research, a hybrid prediction model consisting of a hybrid data preprocessing module, extreme learning machine, and multi-objective grasshopper optimization algorithm was proposed to provide accurate pollutants concentration prediction for evaluating the air quality of the upcoming day by using the fuzzy evaluation method [14]. Moreover, an online air quality evaluation framework was also developed in [15] for assessing air quality in Zhengzhou, China by using cloud model based on dynamic steaming data. More recently, according to the idea of ensemble learning, an air pollution early-warning system that based on incremental extreme learning machine is proposed to forecasting the air quality in China [16]. Considering that an important purpose of pollution prediction is to control air pollution, there are also some studies on the synergistic control of pollution. Based on the situation in China, a new dynamic evolutionary game model of haze cooperative control between the regional governance was developed [17]. The analysis results pointed out that the penalties on the uncooperative governments should be imposed to promote cooperation control of air pollution. However, just like the literature mentioned above, the majority of studies on air pollution forecasting are deterministic forecasts, while few studies focus on interval prediction.

No matter what kind of deterministic prediction models, each prediction result has inherent and irreducible uncertainties. If the uncertainties are ignored, it may lead to evaluation errors, resulting in unreasonable supervision or preventive measures [18]. To fill this gap, a novel hybrid system incorporating deterministic prediction and interval prediction is proposed in this study, which contains three modules: data preprocessing module, prediction module, and interval prediction module. In the data preprocessing module, Hampel Identifier **(HI)** method, Sample Entropy **(SampEn)** and Variational Mode decomposition **(VMd)** technique are introduced in this study. In detail, **HI** is first used to detect and correct the outliers in the original series, and **SampEn** is utilized to measure the complexity of series before and after outlier correction. Then, the corrected series is decomposed into a set of modes by using the **VMd**. In the prediction module, each mode is predicted by using **mSSa-**$\ell_{2,1}$**RFELM** (Random Fourier Extreme Learning Machine with $\ell_{2,1}$-norm Regularization based on multi-objective Salp Swarm algorithm). After then, the prediction results of the modes are summed to get the prediction value of the pollutant concentrations. Further, based on the deterministic prediction values and the best fitting distribution, the upper and lower bounds of pollutant concentrations can be achieved in the interval prediction module.

**The main innovations of this study are summarized as follows:**

***(1) A system contains deterministic prediction and interval prediction is proposed, which provides accurate concentration predictions and effective concentration intervals. Moreover, both theoretical and experimental analysis verify the applicability of the mSSa.*** In the prediction module, an outlier detection and correction method, a data decomposition strategy, a multi-objective optimization algorithm, and an improved ELM method are utilized to improve prediction accuracy. In the improved ELM method, $\ell_{2,1}$-norm is used to make the hidden layer more compact and discriminative, and Random Fourier Mapping is used to approximate the kernel to improve the extendibility of ELM. And the interval prediction module provides effective intervals of pollutant concentrations, which quantify the range of concentration change of air pollutants caused by uncertainties, providing decision-makers with more valuable uncertainty information.

***(2) An effective data preprocessing module is introduced to reduce the complexity, the noisy, and chaotic characteristics of original series.*** The outliers and noise in the original series will affect the performance of the prediction models. However, many previous studies have neglected the impact of outliers on modeling and failed to further improve the prediction accuracy of their models. In this paper, an outlier detection and correction method is used to preprocess the original series and apply the Sample Entropy to measure the complexity of the series before and after processing to prove the effectiveness of the outlier processing. Moreover, most previous studies used EMD-series decomposition methods for denoising, and such methods have disadvantages, for instance, endpoint effects, mode aliasing, etc. The VMd decomposition technique used in this paper can overcome these drawbacks and can effectively extract the features from the original series.

***(3) Used interval prediction technique further mined and analyzed the uncertainty associated with deterministic prediction.*** Considering that different series have different statistical differences, the Maximum Likelihood Estimation is used to estimate the parameters of the distributions, and then based on the best fitting distribution, the uncertainty of pollutant concentration prediction is analyzed.

***(4) Comprehensive evaluations of the proposed prediction model are presented.*** In the discussion part, the computational complexity of the proposed model is analyzed, and the sensitivity of the model to input perturbation is studied. Moreover, in addition to the use of predictive evaluation indicators, the Diebold-Mariano test is also introduced to study the significance of the model. The results indicated that the proposed model has low sensitivity to perturbed input, and is significantly better than the comparison models.

The rest sections of this study can be organized as follows: the methods and techniques used in this paper, and the framework of the research are introduced in Sections 2; Section 3 shows the study data, and evaluation indices used in each section; the experimental results are listed in Section 4; Section 5 gives several discussions of the developed model in this study, and the conclusions are listed in Section 6.

## 2. Method and framework of proposed system

This section briefly introduce the related approaches used in deterministic prediction section and the proposed system.

### 2.1. Data preprocessing

There are three parts in deterministic prediction section, data preprocessing part, predictive model optimization part, and optimal model prediction part. Assuming that the analyzed sequence is $\mathbf{X} = \{x_1, x_2, \ldots, x_n, \ldots, x_N\}$, the steps of deterministic prediction are as follows.

#### 2.1.1. Outlier processing

An outlier is defined as an observation that is greatly different from other observations. So it is suspected to have been generated by a different mechanism than the other observations [19]. The presence of outliers in the modeling will affect the accuracy of the prediction. The Hampel Identifier **(HI)** method is adopted to identify and correct outliers among series. **HI** detects outliers under a sliding window. If the detected data $\boldsymbol{x_s}$ is greater than the threshold value, then $\boldsymbol{x_s}$ is considered an outlier. Details of **HI** are illustrated as follows [20].

Compute the local median and median absolute deviation **(MAD)** in the window by using the following formulae:

$$m_s = \begin{cases} median(x_1, \ldots, x_{s+(k-1)/2}), \ s = 1, \ldots, \frac{k-1}{2}, \\ median(x_{s-(k-1)/2}, \ldots, x_s, \ldots, x_{s+(k-1)/2}), \ s = \frac{k-1}{2}+1, \frac{k-1}{2}+2, \ldots, N-\frac{k-1}{2}, \\ median(x_{s-(k-1)/2}, \ldots, x_n), \ s = N-\frac{k-1}{2}+1, \ldots, N, \end{cases} \tag{1}$$

$$\mathbf{MAD}_s = median(|x_{s-k} - m_s|, \ldots, |\boldsymbol{x_s} - m_s|, \ldots, |x_{s+k} - m_s|), \tag{2}$$

where $x_s$ is the detected data, $k$, $0 < k \le (N-1)/2$ is window size, which is determined at the beginning, and must be odds. There are $(k-1)/2$ adjacent samples to the left and right of the $x_s$, respectively. The filter algorithm appends $(k-1)/2$ zeros before $x_1$ when $s = 1$, and $(k-1)/2$ zeros after $x_N$ when $s = N$.

Determine whether $x_s$ is an outlier by comparing $x_s$ with threshold value. If $x_s$ is an outlier, replace it by local median $m_s$.

$$x_s = \begin{cases} x_s, & |x_s - m_s| \le sd_s, \\ m_s, & |x_s - m_s| > sd_s, \end{cases} \tag{3}$$

where $sd_s = \sigma \mathbf{MAD}_s$, $\sigma = \frac{1}{\sqrt{2} erfc^{-1}(1/2)} \approx 1.4286$ and $erfc(\cdot)$ is complementary error function. The series after outlier correction denoted as $\ddot{\mathbf{X}} = \{\ddot{x}_1, \ddot{x}_2, \ldots, \ddot{x}_n, \ldots, \ddot{x}_N\}$.

After outlier processing, the complexity of the series before and after processing is tested using the Sample Entropy to verify the effectiveness of the outlier processing. Sample Entropy **(SampEn)** is an improved method based on Approximate Entropy (ApEn) and is a measure tool of the complexity of the time series. The lower the value of the SampEn, the more self-similarity the series and the lower the complexity of the series. Given a series, such as **X**, the details of calculating **SampEn** are as follows [21].

First, reconstruct the series. Give the parameter $m$ as the length of the compared series, the reconstructed matrix can represented as:

$$\begin{aligned} &\mathbf{Y}^m = \left\{ \overrightarrow{\boldsymbol{y_1^m}}, \overrightarrow{\boldsymbol{y_2^m}}, \ldots, \overrightarrow{\boldsymbol{y_j^m}}, \ldots, \overrightarrow{\boldsymbol{y^m}}_{\boldsymbol{N-m+1}} \right\}, \\ &\overrightarrow{\boldsymbol{y_j^m}} = \left\{ x_j, x_{j+1}, \ldots, x_{j+m-1} \right\}^{\mathrm{T}}, \ j = 1, 2, \ldots, N-m+1. \end{aligned} \tag{4}$$

Then, calculate the maximum distance between two sub-sequences. This distance is equal to the maximum absolute value of the difference between the corresponding elements of the two sub-sequences, and the mathematical process is represented as follows:

$$\begin{aligned} &d_{ij}^m = \max\left\{ |x_{i+r} - x_{j+r}| : \ 0 \le r \le m-1 \right\}, \\ &i = 1, 2, \ldots, N-m+1; \ j = 1, 2, \ldots, N-m+1; \ i \ne j. \end{aligned} \tag{5}$$

Define $\mathbf{B}_{ij}^m(r)$:

$$\mathbf{B}_{ij}^m(r) = \begin{cases} 1, \ d_{ij}^m \le r, \\ 0, \ \text{otherwise}, \end{cases} \tag{6}$$

where $r$ is a threshold given in advance representing tolerance for acceptable matches. Then, the number of sub-sequences whose distance from $\overrightarrow{\boldsymbol{y_j^m}}$ is less than $r$, denoted as $\mathbf{B}_j^m(r)$:

$$\mathbf{B}_j^m(r) = \sum_{i=1, \ j \ne i}^{N-m+1} \mathbf{B}_{ij}^m(r). \tag{7}$$

For $1 \le j \le n-m+1$ define $\check{\mathbf{B}}_j^m(r)$ as

$$\check{\mathbf{B}}_j^m(r) = \frac{1}{N-m}\mathbf{B}_j^m(r). \tag{8}$$

Calculated the average of $\check{\mathbf{B}}_j^m(r)$ over all $j$

$$\overline{\mathbf{B}^m}(r) = \frac{\sum_{j=1}^{N-m+1}\check{\mathbf{B}}_j^m(r)}{N-m+1}. \tag{9}$$

Adjust the dimension of the sub-sequences to $m+1$ and repeat the above calculation process to obtain $\mathbf{B}^{m+1}(r)$. Finally, **SampEn** is obtained as:

$$\mathbf{SampEn}(m,r,n) = -\ln\left[\frac{\mathbf{B}^{m+1}(r)}{\mathbf{B}^m(r)}\right]. \tag{10}$$

Following these computational steps, the complexity of the series $\mathbf{X}$ and $\ddot{\mathbf{X}}$ is calculated and compared.

*2.1.2. Data decomposition*

Since the original series are fluctuating, it is difficult to analyze the useful signal of the series. Therefore, a data decomposition technique is adopted. In order to reduce the noise in series, improve the signal-to-noise ratio, and improve the prediction accuracy, a new time-frequency analysis method, Variational Mode decomposition **(VMd)**, is used in this paper.

VMd was proposed by Dragomiretskiy [22]. It is entirely non-recursive, so it avoids the endpoint effect and spurious component problems encountered during iteration. The VMd decomposition technique determines the component center frequency and bandwidth of each sub-series by searching for the optimal solution of the variational model. This technique has been well received in many field [23–26]. The core of **VMd** is to construct and solve variational problems. The theory will be briefly introduced.

**Assumption 1.** All components are band-limited signals concentrated near their respective center frequencies.

Suppose the processed series $\ddot{\mathbf{X}}$ is eventually decomposed into $G$ sub-series, denoted as $\ddot{\mathbf{X}}_{\boldsymbol{g}}, g = 1, \ldots, G$. These sub-series are band-limited intrinsic mode functions **(BLIMFs)** with center frequency $\boldsymbol{\omega_g}$. The variational problem of this method is to minimize the sum of the estimated bandwidth of each BLIMF, provided that the sum of all BLIMFs is equal to the original signal $\ddot{\mathbf{X}}$. Based on the Assumption 1, the constrained variational problem can be expressed as follows:

$$\begin{aligned} &\mathbf{Min}_{\ddot{\mathbf{X}}_{\boldsymbol{g}},\boldsymbol{\omega_g}}\left\{\sum_g \left\|\partial_n\left[\left(\delta(n)+\tfrac{i}{\pi n}\right)*\ddot{\mathbf{X}}_{\boldsymbol{g}}(n)\right]e^{-i\boldsymbol{\omega_g}n}\right\|_2^2\right\},\\ &\mathbf{s.t.}\ \sum_g \ddot{\mathbf{X}}_{\boldsymbol{g}} = \ddot{\mathbf{X}}. \end{aligned} \tag{11}$$

where $n$ is the time script, $\partial_n$ denotes the variance of the white noise, $\delta(n)$ stands for the Dirac distribution, $*$ represents the convolution operator and $i^2 = -1$.

To solve the variational problem, a quadratic penalty term $\alpha$ is introduced in addition to the Lagrange multipliers $\boldsymbol{\lambda}$. Then, (11) can be transformed into an unconstrained variational problem with the following augmented Lagrangian expressions:

$$\begin{aligned} \mathcal{L}\left(\ddot{\mathbf{X}}_{\boldsymbol{g}},\boldsymbol{\omega_g},\boldsymbol{\lambda}\right) :=& \alpha\sum_g \left\|\partial_n\left[\left(\delta(n)+\tfrac{i}{\pi n}\right)*\ddot{\mathbf{X}}_{\boldsymbol{g}}(n)\right]e^{-i\omega_g n}\right\|_2^2\\ &+\left\|\ddot{\mathbf{X}}(n)-\sum_g\ddot{\mathbf{X}}_{\boldsymbol{g}}(n)\right\|_2^2+\left\langle\boldsymbol{\lambda}(n),\ddot{\mathbf{X}}(n)-\sum_g\ddot{\mathbf{X}}_{\boldsymbol{g}}(n)\right\rangle. \end{aligned} \tag{12}$$

To solve (12), the alternate direction method of multipliers (ADMM) algorithm is used. Finally, all the solutions in the Fourier domain are written as follows:

$$\begin{aligned} &\hat{\ddot{\mathbf{X}}}_{\boldsymbol{g}}^{T_{vmd}+1}(\omega) = \frac{\hat{\ddot{\mathbf{X}}}(\omega)-\sum_{j\neq g}\hat{\ddot{\mathbf{X}}}_j(\omega)+0.5*\hat{\boldsymbol{\lambda}}(\omega)}{1+2\alpha(\omega-\boldsymbol{\omega_g})^2},\\ &\hat{\boldsymbol{\omega}}_{\boldsymbol{g}}^{T_{vmd}+1} = \frac{\int_0^\infty \omega\left|\hat{\ddot{\mathbf{X}}}_{\boldsymbol{g}}(\omega)\right|^2 d\omega}{\int_0^\infty\left|\hat{\ddot{\mathbf{X}}}_{\boldsymbol{g}}(\omega)\right|^2 d\omega},\\ &\hat{\boldsymbol{\lambda}}^{T_{vmd}+1}(\omega) = \hat{\boldsymbol{\lambda}}^{T_{vmd}}(\omega)+\tau\left(\hat{\ddot{\mathbf{X}}}(\omega)-\sum_g\hat{\ddot{\mathbf{X}}}_g^{T_{vmd}+1}(\omega)\right). \end{aligned} \tag{13}$$

In this step, $\alpha$ is a balancing parameter used to reduce the interference of Gaussian noise. Moreover, the tolerance of the convergence criterion, denoted as $\varepsilon$, is set as the related stop condition:

$$\sum_{g} \frac{\left\| \hat{\ddot{\mathbf{X}}}_{\mathbf{g}}^{T_{vmd}+1}(\omega) - \hat{\ddot{\mathbf{X}}}_{\mathbf{g}}^{T_{vmd}}(\omega) \right\|_2^2}{\left\| \hat{\ddot{\mathbf{X}}}_{\mathbf{g}}^{T_{vmd}}(\omega) \right\|_2^2} < \varepsilon. \tag{14}$$

here, $T_{vmd}+1$ is the number of iterations when the stopping condition is satisfied.

### 2.2. Optimization of predictive model

In order to make the prediction results more accurate, this paper uses an optimization algorithm to optimize the parameters of the prediction model. Multi-objective salp swarm algorithm **(mSSa)** is a novel swarm intelligence-based optimization algorithm that simulates the salp swarms' behavior in seeking food [27]. As with other multi-objective optimization algorithms, **mSSa** introduces related concepts such as Pareto optimality to compare solutions to find the optimal solution. The relevant concepts are defined as follows.

**Definition 1** ***Pareto domination.*** Given two vectors $\overrightarrow{\mathbf{X}} = (x_1, x_2, \ldots, x_n)$ and $\overrightarrow{\mathbf{Y}} = (y_1, y_2, \ldots, y_n)$, vector $\overrightarrow{\mathbf{Y}}$ dominates $\overrightarrow{\mathbf{X}}$ or called vector $\overrightarrow{\mathbf{X}}$ is dominated by vector $\overrightarrow{\mathbf{Y}}$ denoted as $\overrightarrow{\mathbf{Y}} \prec \overrightarrow{\mathbf{X}}$ if and only if

$$\forall\, i \in [1, o], \left[ f_i(\overrightarrow{\mathbf{Y}}) \le f_i(\overrightarrow{\mathbf{X}}) \right] \wedge \exists\, i \in [1, o], \left[ f_i(\overrightarrow{\mathbf{Y}}) < f_i(\overrightarrow{\mathbf{X}}) \right], \tag{15}$$

where $f_i(\cdot)$ represents $i$th objective function.

The definition means that a solution is better than some other solutions if it has equal or at least one better value in the objectives compared to some other solutions.

**Definition 2** ***Pareto optimality.*** For $x \in \overrightarrow{\mathbf{X}}$, if $\{ \nexists\, z \in \overrightarrow{\mathbf{X}} \mid z \prec x \}$, then $x$ is a Pareto-optimal solution.

In multi-objective optimization problems, the archive mechanism is often introduced to store the non-dominated solutions at each iteration. These solutions are a subset of the Pareto optimal set (Definition 3).

**Definition 3** ***Pareto optimal set.*** A set including all the non-dominated solutions is called Pareto optimal set. The mathematical description is as follows:

$$\mathbf{P}_s := \left\{ x, z \in \overrightarrow{\mathbf{X}} \mid \nexists\, z \prec x \right\}. \tag{16}$$

After the introduction of the basic concepts, the process of the algorithm will introduced. In this study, the mSSa is used to find the optimal initial parameters in the forecasting model, including the original weights **W** between the input layer and the hidden layer, and biases **B** of the hidden layer. Assuming the number of parameters to be optimized is **D**, the number of search agent is **Z**, then the specific process of finding the optimal solution is as follows.

The location of all the salps at $t$th iteration can be defined as a matrix [28]:

$$\mathbf{P}(t) = \begin{bmatrix} p_1^1(t) & p_2^1(t) & \ldots & p_{\mathbf{D}}^1(t) \\ p_1^2(t) & p_2^2(t) & \ldots & p_{\mathbf{D}}^2(t) \\ \vdots & \vdots & \ldots & \vdots \\ p_1^{\mathbf{Z}}(t) & p_2^{\mathbf{Z}}(t) & \ldots & p_{\mathbf{D}}^{\mathbf{Z}}(t) \end{bmatrix}, \tag{17}$$

here, each row is a group of candidate solution to the optimization problem, called a search agent, and each column is a group of candidate solution to a certain parameter.

Then, calculate fitness of each salp chain by follows:

$$\mathbf{Fit}(\bar{p}^z) = \{ obf_1(\bar{p}^z), obf_2(\bar{p}^z), \ldots, obf_o(\bar{p}^z) \}, z = 1, 2, \ldots, \mathbf{Z}, \tag{18}$$

where $\mathbf{Fit}(\bar{p}^z)$ represents the fitness of $z$th search agent, $obf_o(\bar{p}^z)$ is the value of $o$th objective function of $z$th search agent. Based on the concept of multi-objective optimization problem mentioned in [27], we set two conflict objective functions in this paper, they are

$$\mathbf{Min:} \begin{cases} obf_1 = \frac{1}{N}(F - A)^2, \\ obf_2 = std(F - A), \end{cases} \tag{19}$$

where $N$ is the number of samples, $F$ is predict value and $A$ is actual value.

After then, determine the non-dominated salp chains according to Definition 1, and update the repository. Select a salp chain as food source from the repository, denoted as **S**.

Since salp swarm often exists in chains, the individuals in the chain are divided into two groups: leader and followers. The leader is at the front of the chain, and others are followers. And at each iteration, leader $p_j^1$ guides the swarm toward the food source in a D-dimensional search space. The positions of the leaders are updated at each iteration as follows:

$$p_j^1 = \begin{cases} \mathbf{S}_j + c_1\left[\left(\overline{\overline{p_j^1}} - \underline{\underline{p_j^1}}\right)c_2 + \underline{\underline{p_j^1}}\right], & c_3 \geq 0, \\ \mathbf{S}_j - c_1\left[\left(\overline{\overline{p_j^1}} - \underline{\underline{p_j^1}}\right)c_2 + \underline{\underline{p_j^1}}\right], & c_3 < 0, \end{cases} \tag{20}$$

where $p_j^1$ is the position of leader in the $j$th dimension. $\mathbf{S}_j$ represents the food source position in the $j$th dimension. $\underline{\underline{p_j^1}}$ and $\overline{\overline{p_j^1}}$ are the lower bound and the upper bound of $p_j^1$. And $c_1 = 2e^{-(4t/\mathrm{T})^2}$ is a parameter that controls the balance of exploration and exploitation, where T is the number of maximum iteration, and $t$ is the number of the current iteration. $c_2$ and $c_3$ are random numbers generated in [0, 1], among them $c_2$ determines the distance to move and $c_3$ determines the direction of movement.

The positions of the followers are mathematically updated as

$$p_j^z = \frac{1}{2}\left(p_j^z + p_j^{z-1}\right),\ 2 \leq z,\ j = 1, 2, \ldots, \mathrm{D}. \tag{21}$$

Whereafter, the processes of calculating fitness, updating the repository, selecting food source and updating the salps location are repeated until satisfied with the end condition.

**Proof.** Assume the Pareto optimal solution is $\mathcal{V}^* = \{v_1^*, v_2^*, \ldots, v_o^*\}$, and it satisfies $\nexists\, \phi \in \vec{\mathbf{X}}\ s.t.\ \phi \prec \mathcal{V}^*$ according to Definition 1 and Definition 2. Let $\check{A}$ is an archive to store the non-dominated solutions, and the maximum size of $\check{A}$ is $\tilde{N}$. If a new non-dominated solution $\mathcal{V}'$ appears, compare $\mathcal{V}'$ with all the solutions in the $\check{A}$. If $\mathcal{V}' \prec \mathcal{V}_n, n \in [1, \tilde{N}]$, $\mathcal{V}'$ will be included in the $\check{A}$.

When the number of non-dominated solutions in the $\check{A}$ is equal to $\tilde{N}$, the Roulette wheel mechanism is applied to remove the most crowded solution. The probability of ecah solution can be selected is $P_n = k/C_n,\ 1 < k < C_n$, here $C_n$ is the number of $n$th solution in the $\check{A}$. Therefore, the larger $P_n$ indicates the more crowded, and the greater the possibility of being deleted, the deleting probability is $Prop_n = 1/P_n$. After update the $\check{A}$, using sorting mechanism to determine the Pareto optimal solution $\mathcal{V}^*$. So the optimal solution can always be found. □

### 2.3. Prediction using optimized model

The prediction model used in this paper, $\ell_{2,1}$ **RFELM**, was proposed by Zhou et al. [29]. It combines the **Random Fourier Mapping** as an activation function to improve the extendibility of ELM. And uses $\ell_{2,1}$**-norm** to cut irrelevant neurons that make the hidden layer more compact and discriminative.

Before introducing $\ell_{2,1}$RFELM, some basic definitions are introduced.

**Definition 4** $\ell_{2,1}$ ***-norm*** [30]**.** Let $\mathbf{M} = (m_{ij})_{d\times n},\ i = 1, 2, \ldots, d;\ j = 1, 2, \ldots, n$, the $\ell_{2,1}$-norm of $\mathbf{M}$ is defined as follows:

$$\| \mathbf{M} \|_{2,1} = \sum_{j=1}^{n}\left(\sum_{i=1}^{d} m_{ij}^2\right)^{\frac{1}{2}}. \tag{22}$$

The main idea of the **Random Fourier Mapping** is to explicitly map the data to a low-dimensional Euclidean inner product space by using a randomized feature map $z$, so that the inner product between a pair of transformed points approximates their kernel evaluation. According to this idea, the **Random Fourier Mapping** used in this paper is defined as follow.

**Definition 5** ***Random Fourier Mapping*** [31]**.** Give a positive definite shift-invariant kernel $\boldsymbol{k(\delta)}$, to obtain a real-valued random feature for kernel function, replace $e^{-iw^{\mathrm{T}}(x-y)}$ with its real part $cos(w^{\mathrm{T}}(x-y))$. And in order to lower the variance of $z_w(x)^{\mathrm{T}} z_w(y)$, draw **L** i.i.d samples, denoted as $w_1, \ldots, w_{\mathbf{L}} \in \mathcal{R}^{\boldsymbol{d}}$ from $p(w)$, then the **Random Fourier Mapping** can be defined as follow:

$$z(x) = \frac{1}{\sqrt{\mathbf{L}}}[\cos(w_1^{\mathrm{T}}x), \ldots, \cos(w_{\mathbf{L}}^{\mathrm{T}}x), \sin(w_1^{\mathrm{T}}x), \ldots, \sin(w_{\mathbf{L}}^{\mathrm{T}}x)]^{\mathrm{T}}. \tag{23}$$

In this study, the previous seven days' pollutant concentrations are used as input to predict the next day's pollutant concentration, so the series is reconstructed into the input matrix $\mathcal{X}$ and the output vector $\mathcal{Y}$ as follows:

$$\mathcal{X} = [\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{I}] = \begin{bmatrix} \ddot{x}_1 & \ddot{x}_2 & \cdots & \ddot{x}_{N-7} \\ \ddot{x}_2 & \ddot{x}_3 & \cdots & \ddot{x}_{N-6} \\ \vdots & \vdots & \ddots & \vdots \\ \ddot{x}_7 & \ddot{x}_8 & \cdots & \ddot{x}_{N-1} \end{bmatrix},$$
$$\mathcal{Y} = [\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_\mathrm{I}] = [\ddot{x}_8, \ddot{x}_9, \ldots, \ddot{x}_{N-1}, \ddot{x}_N]. \tag{24}$$

Firstly, initialization of parameters. The connection weights $w_{hi}$ between the input and hidden layers, and the bias $b_h$ of the hidden layer are equal to the optimal solution obtained by mSSa. Assume the hidden layer has $\mathrm{D}_{hidden}$ neurons, the weight matrix **W** and bias vector **B** are represented as follows:

$$\begin{aligned} &\mathbf{W} = [w_1, w_2, \ldots, w_h, \ldots, w_{\mathrm{D}_{hidden}}]^\mathrm{T}, \\ &\mathbf{B} = [b_1, b_2, \ldots, b_h, \ldots, b_{\mathrm{D}_{hidden}}], \quad h = 1, 2, \ldots, \mathrm{D}_{hidden}. \end{aligned} \tag{25}$$

Based on the input set $\mathcal{X}$ and the initial parameters **W** and **B**, calculate output matrix **H** of the hidden layer. Moreover, the **Random Fourier Mapping** is used to approximate the kernel in this study, so this step maps the input data of the hidden layer into the Random Fourier feature space according to (23), then the output of Hidden layer **H** can be written as follows:

$$\mathbf{H} = \begin{bmatrix} g(w_1\mathrm{X}_1 + b_1) & g(w_2\mathrm{X}_1 + b_2) & \cdots & g(w_{\mathrm{D}_{hidden}}\mathrm{X}_1 + b_{\mathrm{D}_{hidden}}) \\ g(w_1\mathrm{X}_2 + b_1) & g(w_2\mathrm{X}_2 + b_2) & \cdots & g(w_{\mathrm{D}_{hidden}}\mathrm{X}_2 + b_{\mathrm{D}_{hidden}}) \\ \vdots & \vdots & \ddots & \vdots \\ g(w_1\mathrm{X}_\mathrm{I} + b_1) & g(w_2\mathrm{X}_\mathrm{I} + b_2) & \cdots & g(w_{\mathrm{D}_{hidden}}\mathrm{X}_\mathrm{I} + b_{\mathrm{D}_{hidden}}) \end{bmatrix}, \tag{26}$$

where $g(\cdot)$ represents Random Fourier Mapping that is defined in the Definition 5.

Denote the connection weight vector between the hidden layer and the output layer as $\boldsymbol{\beta}$, then the output function of the network is modeled as below:

$$\sum_{h=1}^{\mathrm{D}_{hidden}} \beta_h g(w_h \mathrm{X}_i + b_h) = O_i, \ i = 1, 2, \ldots, \mathrm{I}. \tag{27}$$

For a given dataset $(\mathcal{X}, \mathcal{Y})$, the output function of this network can be represented in a matrix form as $\mathbf{H}\boldsymbol{\beta} = \mathcal{Y}^\mathrm{T}$, where $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_{\mathrm{D}_{hidden}}]^\mathrm{T}$.

The only parameter need to solve is $\boldsymbol{\beta}$ in $\ell_{2,1}$RFELM, so based on the given dataset and parameters, $\boldsymbol{\beta}$ can be obtained by solving the following optimization problem:

$$\begin{aligned} &\underset{\boldsymbol{\beta}, \xi}{\mathbf{Min}} \ \frac{1}{2}\|\boldsymbol{\beta}^\mathrm{T}\|_{2,1} + \frac{1}{2}\tilde{\mathrm{C}}\sum_{i=1}^{N}\|\xi_i\|^2, \\ &\mathbf{s.t.}\ g(\mathbf{W}\mathrm{X}_i + \mathbf{B})\,\boldsymbol{\beta} = y_i - \xi_i, i = 1, 2, \ldots, \mathrm{I}, \end{aligned} \tag{28}$$

where $\xi$ represents the training error, and $\tilde{\mathrm{C}}$ is the penalty coefficient assigned manually. For more details about solving this optimization problem, please refer to [29]. Finally, $\beta$ can be deduced as follows:

$$\hat{\boldsymbol{\beta}} = \left(\frac{\mathbf{D}}{\tilde{\mathrm{C}}} + \mathbf{H}^\mathrm{T}\mathbf{H}\right)^{-1}\mathbf{H}^\mathrm{T}\mathcal{Y}^\mathrm{T}, \tag{29}$$

where **D** is a diagonal matrix with $\mathbf{D}_{ii} = (1/2)\|\boldsymbol{\beta}\|_2$. It can easily found that **D** depends on $\boldsymbol{\beta}$, so an iterative algorithm is used in this step to obtain the exact solution for (29). And at the beginning of the iterative, **D** is an identity matrix.

### 2.4. Framework of the proposed prediction system

To improve the prediction accuracy of pollutant concentration, this study proposes a novel hybrid prediction model based on the outlier detection and correction method and "decomposition and ensemble" techniques, named HI-VMd-mSSa-$\ell_{2,1}$RFELM. After that, the interval prediction of pollutant concentration is implemented. Additionally, the flowchart of the proposed system is shown in Fig. 1.

- **Outlier processing**. Outliers in the data series usually affect the robustness of the model, so we applied the Hampel Identifier method to detect and correct the outliers in the series. The implementation of HI is shown in Fig. 1 (Step 1).

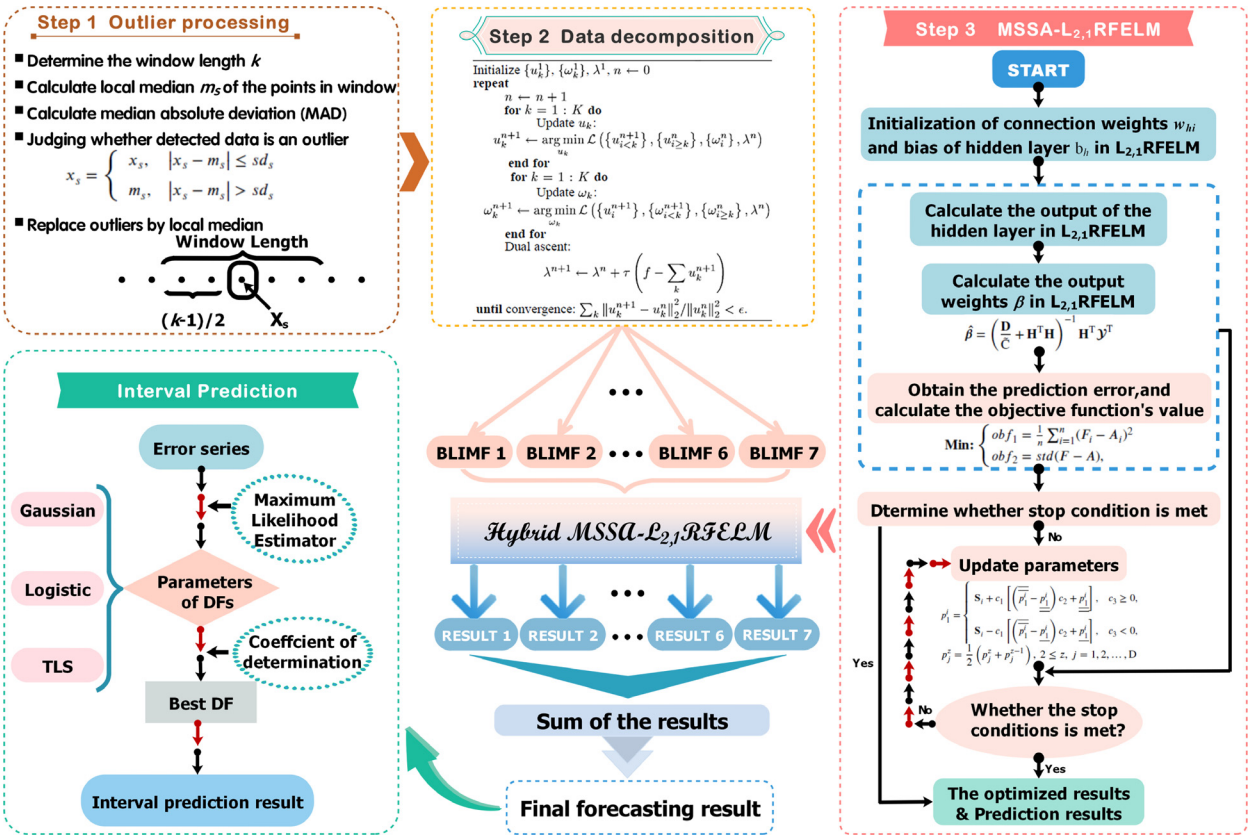


**Fig. 1.** The framework of the developed hybrid architecture. Step 1 in the graph lists the specific flow of the Hampel Identifier method. And Step 2 is given the pseudo-code of the Variational Mode decomposition technique [22] and the decomposition results. Each BLIMF will be predicted in Step 3. After the deterministic prediction, the prediction error will be analyzed in the interval prediction module.

- **Data decomposition**. In order to improve the prediction accuracy, an effective denoising technique, VMd, was adopted to remove the noise in time series. In this part, the original series were decomposed into a finite set of IMFs, and each IMF contains different features hidden in the pollutant concentration series. The pseudo-code of VMd is shown in Fig. 1 (Step 2).
- **Optimization**. In this study, a novel multi-objective optimization algorithm was used to optimize the parameters of the $\ell_{2,1}$RFELM model. The flowchart of mSSa is shown in Fig. 1 (Step 3).
- **Prediction**. The IMFs obtained from VMd were forecasted by $\ell_{2,1}$RFELM model. Subsequently, summed all the forecasting results as the final prediction values, and added up all the prediction errors for the analysis of interval prediction. The network of $\ell_{2,1}$RFELM is illustrated in the Fig. 1 (Step 3).
- **Interval Prediction**. We compared three distributions, and found a better distribution that can fit prediction errors well, so based on the selected distribution and deterministic prediction results, the interval of sample point concentrations could be found. The details of this part are given in Fig. 1 (Interval Prediction).

## 3. Data description and model evaluation criteria

To test the prediction performance of the proposed model, three $PM_{2.5}$ concentration series of study cities from the Pearl River Delta (PRD) region in China and eight error measures are selected in this study.

### *3.1. Data description and analysis*

Located in the south of China, the PRD region is the pioneer region of Chinese economic reform and an important economic center of China. Therefore, the analysis of $PM_{2.5}$ in the PRD region is not just beneficial to the health of residents, but also to the sustainable development of the economy.

In this study, daily $PM_{2.5}$ concentration series of three major cities from the PRD region, Guangzhou, Shenzhen, and Zhuhai, are taken as case studies. These daily $PM_{2.5}$ time series are collected from 2019.01.01 to 2020.11.30. Specifically, each dataset consists of two subsets: training dataset ranging from 2019.01.01 to 2020.09.22 for establishing the forecasting

**Table 1**
Descriptive statistics of $PM_{2.5}$ (Unit:μg/$m^3$).

| Study areas | Central Tendency | | Degree of Dispersion | | | Distribution Shape | |
|---|---|---|---|---|---|---|---|
| | Mean | Median | Std. | Min | Max | Skew. | Kurt. |
| Guangzhou | 26.4213 | 23 | 14.4150 | 4 | 92 | 1.0167 | 4.0337 |
| Shenzhen | 21.7565 | 20 | 12.4202 | 4 | 74 | 0.8845 | 3.6590 |
| Zhuhai | 21.5691 | 18 | 14.1101 | 3 | 76 | 1.0900 | 3.9976 |

**Table 2**
Description and equations of the performance metric rules.

| | Metrics | Definitions | Equations |
|---|---|---|---|
| **DP** | **MAE** | Mean absolute error of the prediction results | $\mathbf{MAE} = \frac{1}{N}\sum_{i=1}^{N}\lvert F_i - A_i\rvert$ |
| | **RMSE** | Root of mean of prediction error squares | $\mathbf{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - A_i)^2}$ |
| | **MAPE** | Average of absolute percentage error | $\mathbf{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\lvert\frac{F_i - A_i}{A_i}\rvert \times 100\%$ |
| **IP** | **IPCP** | Converge probability of interval prediction | $\mathbf{IPCP} = \frac{1}{N}\sum_{i=1}^{N} c_i$, $c_i = \begin{cases}1, \text{ if } F_i \in [L_i, U_i]\\ 0, \text{ otherwise}\end{cases}$ |
| | **AWD** | Accumulated width deciation | $\mathbf{AWD} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{AWD}_i^{\alpha}$, $\mathrm{AWD}_i^{\alpha} = \begin{cases}\frac{L_i^{\alpha} - A_i}{U_i^{\alpha} - L_i^{\alpha}}, & A_i < L_i^{\alpha}\\ 0, & A_i \in I_i^{\alpha}\\ \frac{A_i - U_i^{\alpha}}{U_i^{\alpha} - L_i^{\alpha}}, & A_i > L_i^{\alpha}\end{cases}$ |
| | **IPNAW** | Normalized average width of interval prediction | $\mathbf{IPNAW} = \frac{1}{N}\sum_{i=1}^{N}\frac{U_i - L_i}{\overline{\overline{F}} - \underline{\underline{F}}}$ |
| **Fitting** | $\mathbf{R}^2$ | Coefficient of determination | $\mathbf{R}^2 = 1 - \frac{\sum_{k=1}^{K}(\hat{P}_k - P_k)^2}{\sum_{k=1}^{K}(P_k - \bar{P})^2}$ |
| | $\mathbf{RMSE_{IP}}$ | Root of mean of fitting error squares | $\mathbf{RMSE_{IP}} = \sqrt{\frac{\sum_{k=1}^{K}(\hat{P}_k - P_k)^2}{K}}$ |

1.DP represents Determinstic Prediction; IP represents Interval Prediction.
2. $F_i$ represents i-th forecasting value, $A_i$ represents i-th actual value. $\bar{F}$ denotes the mean value of forecasting series, $\bar{A}$ denotes the mean value of actual series.
3. $L_i$ is the predicted value of the upper bound of the interval, and $U_i$ is the predicted value of the lower bound of the interval. $L_i^{\alpha}$ and $U_i^{\alpha}$ are the lower and upper bounds of the interval under the significant level $\alpha$. $\overline{\overline{F}}$ and $\underline{\underline{F}}$ are the maximum and minimum values of forecasting series.
4. $\hat{P}_k$ is the estimated probability density at k, and $P_k$ is the actual probability density at k. $\bar{P}$ represents the mean of actual probability density.

models, and testing dataset ranging from 2020.09.23 to 2020.11.30 for verifying the performance of the designed model. The descriptive statistics for the three series are shown in Table 1.

### 3.2. Performance evaluation

To validate the prediction performance of the models, many evaluation criteria have been widely used in several studies. In this research, seven error criteria are adopted to evaluate the effectiveness of the models, including Mean Absolute Error (**MAE**), Root Mean Square Error (**RMSE**), Mean Absolute Percentage Error (**MAPE**). Among these criteria, **the lower the value of MAE, RMSE, and MAPE, the better the model**.

The interval prediction analyzes the uncertainty information about the $PM_{2.5}$ concentration forecasting results by finding a prediction interval with high coverage and small interval. This study compares the performance of the three distributions in interval forecasting, and the primary indicators compared are Interval Prediction Coverage Probability (**IPCP**), Interval Prediction Normalized Average Width (**IPNAW**), and Accumulated Width Deviation (**AWD**). According to the analysis demand, **the larger IPCP and the smaller IPNAW and AWD, the better the interval prediction.** In addition, we evaluate the fitting performance of different distributions by using the following two metrics. R square ($\mathbf{R}^2$) is utilized to evaluate the correlation between the recorded and the estimated cumulative probability. Moreover, the deviation between the estimated cumulative probability and the recorded cumulative probability is measured by $\mathbf{RMSE}_{IP}$. The details of these indicators are listed in Table 2.

## 4. Empirical study

In this section, the parameter settings of the compared models and the proposed model are provided. The experimental results are presented to verify the efficiency of the proposed hybrid model **(HI-VMd-mSSa-$\ell_{2,1}$RFELM, P)**. Moreover, the effectiveness of each component of the proposed model is verified. In this study, seven models are selected as the baseline

**Table 3**
The hyper-parameters of proposed model.

| Method | Meaning | Symbol | Value |
|---|---|---|---|
| Hampel Identifier (HI) | Sliding window length | $k$ | 3 |
| Sample Entropy (SampEn) | Length of reconstructed series | $m$ | 3 |
| | Matching Tolerance | $r$ | $0.2\times Std.$ |
| Variational Mode | Balance parameter | $\alpha$ | 2000 |
| Decomposition (VMd) | Noise tolerance | $\tau$ | 0 |
| | Number of modes | $G$ | 8 |
| | Tolerance of convergence criterion | $\varepsilon$ | $10^{-9}$ |
| Multi-objective Salp Swarm | Size of archive | AMS | 100 |
| Algorithm (mSSa) | Search agents' number | $\mathbf{Z}$ | 30 |
| | Maximum iterations | $T_{mSSa}$ | 50 |
| | Individual value range | $[P_{lb}, P_{ub}]$ | $[-5, 5]$ |
| $\ell_{2,1}$-norm Random Fourier | Penalty coefficient | $\tilde{C}$ | 5 |
| ELM ($\ell_{2,1}$RFELM) | Maximum iterations | $T_{ELM}$ | 50 |
| | Number of neurons in hidden layer | $D_{hidden}$ | 15 |

models, including Auto-regressive integrated Moving Average **(Arima, $S_1$)**, $\ell_{2,1}$-norm and Random Fourier Mapping-based Extreme Learning Machine (**$\ell_{2,1}$**RFELM, $S_2$)**,** $\ell_{2,1}$RFELM based on the multi-objective Ant Lion Optimization **(moALo-$\ell_{2,1}$**RFELM, $H_1$)**,** $\ell_{2,1}$RFELM based on the multi-objective Salp Swarm Optimization **(mSSa-$\ell_{2,1}$**RFELM, $H_2$)**,** moALo-$\ell_{2,1}$RFELM based on Variational Mode decomposing technique **(VMd-moALo-$\ell_{2,1}$**RFELM, $H_3$)**, VMd-mSSa-$\ell_{2,1}$**RFELM ($H_4$)**,** mSSa-$\ell_{2,1}$RFELM combined with Hampel Identifier method and Ensemble Empirical Mode decomposition technique **(HI-EEMd-HmSSa-$\ell_{2,1}$**RFELM, $H_5$)**.**

### 4.1. Hyper-Parameter settings

Since the values of hyper-parameters are mostly empirical values in some fields, even taking the same values in other fields may lead to different results. Therefore, the settings of hyper-parameters in this paper refer to studies with the same application background. According to [11], the length of the reconstructed series of SampEn is set to be 3, and the matching tolerance of SampEn is set to be 0.2 × Standard deviation. Then the major hyper-parameter of HI, sliding window length is selected in a range of [3,50] according to the results of SampEn. The sliding window length with the smallest SampEn will be selected. The selections of the hyper-parameters in VMD lack adaptive processing procedure [32]. Therefore, after referring to several models on time series forecasting, the balance hyper-parameter, noise tolerance, initialize center frequency, and convergence tolerance are set as shown in Table 3 [32,33]. The center frequencies of all the modes are initialized by using the uniform distribution [34]. In addition, the key hyper-parameters of mSSa and $\ell_{2,1}$RFELM are search agents' number, individual value range, and penalty coefficient. To make the predicted value of each point close to the true value, RMSE is used to select these main hyper-parameters. Since RMSE gives more weight to the points with large prediction errors, which means that large errors will make the RMSE values poor, so the optimal hyper-parameters of the model are selected by comparing the values of RMSE. Where the search agents' number is picked in [10, 20, 30, 40, 50], the individual value range is selected from [−1,1], [−2,2], [−3,3], [−4,4], [−5,5], and the penalty coefficient is chosen in [0.01, 0.1, 0.5, 1, 2, 3, 4, 5] [29,35,36]. Furthermore, after several experiments, it is found that the better performance of prediction for each study site can be obtained when the number of BLIMF in the data decomposition processing is 8. The results of the experiments based on the data from Guangzhou of China show that the value of RMSE is optimal when the hyper-parameters are set as shown in Table 3. The prediction results for Shenzhen and Zhuhai of China confirm that the model constructed in this paper predicts the best results based on the selected hyper-parameters.

For Arima, the optimal values of $\boldsymbol{p}$, $\boldsymbol{d}$, $\boldsymbol{q}$ are determined based on the Akaike information criterion (AIC). To guarantee the fair experiments between the proposed model and the benchmark models, the hyper-parameters of the benchmark models are set the same as that of the proposed model. The experiments performed in our study are implemented on MATLAB R2020a, run on a Windows 10 Professional operating system.

### 4.2. Experiment I: Comparison of deterministic forecasts

Based on the above settings, testing experiments are conducted on the three study series. This subsection describes four types of model comparisons: comparisons between the single models and the optimized models (**Comparison I**), comparisons of models before and after combining data decomposition strategy (**Comparison II**), comparisons of models before and after combining the outlier detection and correction method (**Comparison III**), comparisons between the proposed model and all the benchmark models (**Comparison IV**). The detailed values of the performance metrics of the proposed and comparison models are presented in Table 5, where the values marked in bold are the best values of each evaluation metric, indicate that the corresponding model is the best one among all the eight models.

**Table 4**
Sample Entropy of the original series and the processed series.

| Series | Sample Entropy | | |
|---|---|---|---|
| | Guangzhou | Shenzhen | Zhuhai |
| Original series | 0.9627 | 0.6382 | 0.7116 |
| Processed series | 0.9268 | 0.6157 | 0.6617 |

**Table 5**
Results of deterministic prediction.

| Metrics | Guangzhou | | | Shenzhen | | | Zhuhai | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAPE | MAE | RMSE | MAPE | MAE | RMSE | MAPE |
| $S_1$ | 6.2300 | 7.7987 | 26.2359% | 4.1595 | 5.3211 | 18.1223% | 5.4782 | 7.6782 | 24.6947% |
| $S_2$ | 6.2054 | 8.0392 | 26.9886% | 4.0358 | 5.1551 | 17.8757% | 5.4997 | 7.1832 | 24.8322% |
| $H_1$ | 5.9139 | 7.8553 | 25.5637% | 4.1029 | 5.3926 | 17.7560% | 5.5079 | 7.1955 | 24.8277% |
| $H_2$ | 5.9655 | 7.7322 | 25.4945% | 3.9124 | 5.0631 | 17.2753% | 5.3688 | 7.2335 | 24.2870% |
| $H_3$ | 2.2102 | 2.7282 | 9.0225% | 2.0464 | 2.4241 | 9.4584% | 2.5164 | 3.2272 | 11.2619% |
| $H_4$ | 2.0073 | 2.4875 | 8.1709% | 2.0156 | 2.4065 | 8.7864% | 2.2681 | 2.9346 | 10.4727% |
| $H_5$ | 2.8009 | 3.5438 | 12.2059% | 2.0667 | 2.5437 | 9.0461% | 2.4876 | 3.2007 | 10.8669% |
| P | **1.7878** | **2.2362** | **7.4723%** | **1.6437** | **1.9536** | **7.2571%** | **2.1629** | **2.6873** | **10.1336%** |

1. The bold numbers indicate the optimal value of the indicators.
2. $S_1$: Arima; $S_2$: $\ell_{2,1}$RFELM; $H_1$: moALo-$\ell_{2,1}$RFELM; $H_2$: mSSa-$\ell_{2,1}$RFELM; $H_3$: VMd-moALo-$\ell_{2,1}$RFELM; $H_4$: VMd-mSSa-$\ell_{2,1}$RFELM; $H_5$: HI-EEMd-mSSa-$\ell_{2,1}$RFELM; P: HI-VMd-mSSa-$\ell_{2,1}$RFELM.

*4.2.1. Comparison I: Comparisons of models before and after combining optimization algorithms*

To measure the effectiveness of the optimization algorithms in improving the prediction, this paper compares the values of prediction performance metrics between the single models and the optimized models. The single models including Arima ($S_1$) and $\ell_{2,1}$RFELM ($S_2$), the optimized models are moALo-$\ell_{2,1}$RFELM ($H_1$) and mSSa-$\ell_{2,1}$RFELM ($H_2$), and the prediction results are shown in the Table 5. The $\mathbf{MAPE}_{GZ}$ of $S_1$ ($\boldsymbol{p=3, d=1, q=4}$) is 26.2359%, and that of $S_2$ is 26.9886%. After combining optimization algorithms, the $\mathbf{MAPE}_{GZ}$ of the model is improved. For $H_1$ and $H_2$, $\mathbf{MAPE}_{GZ}$=[25.5637%, 25.4945%], the values are lower than $\mathbf{MAPE}_{GZ}$ of $S_1$ and $S_2$. In addition, results of $\mathbf{MAE}_{GZ}$ indicate that the predictive performance of the models that combine with optimization algorithm is better than that of the single models. However, values of remaining four metrics show that the predictive capability of the models combined with the moALo is not better than that of the single models.

Based on the data from Shenzhen, the best values of parameters of the $S_1$ are $\boldsymbol{p=7}$, $\boldsymbol{d=1}$ and $\boldsymbol{q=7}$, and $\mathbf{MAPE}_{SZ}$ of $S_1$ is 18.1223%. The $\mathbf{MAPE}_{SZ}$ of $S_2$, $H_1$, and $H_2$ are 17.8757%, 17.7560%, and **17.2753**%, respectively. Moreover, both the $\mathbf{MAE}_{SZ}$ and $\mathbf{RMSE}_{SZ}$ values of $H_2$ are lower than those values of $S_1$ and $S_2$, but those values of $H_1$ are greater than those values of $S_2$. That is, compared with the traditional statistical model, the optimized model has better predictive ability, but compared with an individual neural network model, the result is not necessarily obtained.

For Zhuhai, the best values of parameters of Arima are $\boldsymbol{p=6}$, $\boldsymbol{d=1}$ and $\boldsymbol{q=10}$. The $\mathbf{MAPE}_{ZH}$ of $S_1$, $S_2$, $H_1$, and $H_2$ are 24.6947%, 24.8322%, 24.8277%, and **24.2870**%, respectively. This result indicates that the predictive ability of $S_1$ is better than that of $S_2$ and $H_1$, and $H_2$ performs the best. In addition, the comparison results of other metrics also indicate that $H_1$ performs similarly to the single models, and $H_2$ outperforms the single models.

In summary, the comparison results based on the data from Guangzhou show that $S_1$ and $S_2$ have similar prediction ability. Among the three indicators, the values of RMSE and MAPE show that $S_1$ outperforms $S_2$, and MAE show that $S_2$ is better than $S_1$. Therefore, it can be assumed that the **predictive capabilities of the $S_1$** and **$S_2$are similar in the case of Guangzhou**. However, **for Shenzhen and Zhuhai, the comparison results show that $S_2$has better prediction performance**, and almost all indicators' results support this result. In addition, the performance of the two optimization algorithms is also compared. Based on the analysis of the comparison results, it is clear that **the optimization effect of mSSa is stronger than that of moALo**. Because the indicators' values of the mSSa-based model are all better than those of the moALo-based model.

**Remark.** The analysis of the comparison results shows that the predictive ability of $S_1$ and $S_2$ are similar, and the mSSa optimized model performs better than the moALo optimized model. However, the improvement in predictive ability of the single models by combining only optimization algorithms is limited.

*4.2.2. Comparison II: Comparisons of models before and after combining data decompose strategy*

The original air pollution series are characterized by instability and noise, and lead to imprecise prediction. So, decomposition and ensemble technique is used in this paper to extract useful information from the original series. In order to verify the effectiveness of the data decomposition strategy, and to compare the effects of different decomposition methods, the prediction results of $H_1$, $H_2$, $H_3$, $H_4$, $H_5$ and the proposed model are compared. Among them, $H_1$ and $H_2$ are hybrid models

without data decomposition strategy, $H_3$, $H_4$ and the proposed model are combined with VMd decomposition technique, and $H_5$ is combined with EEMd decomposition technique. And **the parameters of EEMd are: (1) the standard deviation of white noise is 0.2; (2) the number of white noise is 100; (3) the maximum number of sifting iterations allowed is 100.**

The results of comparing indicators' values of $H_3$ and $H_1$, $H_4$ and $H_2$ show that the prediction performance of the model combined with the data decomposition strategy is substantially improved. The **MAPE** values of $H_1$ for the three study areas are 25.5637%, 17.7560%, and 24.8277%, respectively, while the **MAPE** values of $H_3$ are **9.0225**%, **9.4584**% and **11.2619**%, respectively, with an average improvement in **MAPE** of 55% compared to $H_1$. And the MAE values of $H_2$ for the study areas are 5.9655, 3.9124, and 5.3688, that of $H_4$ are **2.0073, 2.0156** and **2.2681**. It can be easily found that in all the study areas, the values of **MAE** of $H_4$ are significantly improved compared with that of $H_2$. The results of RMSE also show that **the predictive ability of the model is significantly improved after combining the data decomposition strategy**.

In order to compare different data decomposition method, this paper compares the prediction results of $H_5$ and the proposed model. The values of three indicators in these three cities show that the proposed model has better prediction ability than $H_5$. Therefore, it can be concluded that **VMd is more effective than EEMd in improving the prediction ability of the model**.

**Remark.** The data decomposition strategy reduces the interference between different features in the original data, making the linear and non-linear features within the series more readily available and reducing the difficulty of modeling. Therefore, the prediction performance of the hybrid model combined with data decomposition strategy can be greatly improved. Furthermore, the VMd-based models outperform the EEMd-based model. The possible reasons are: (1) compared with EEMd, VMd overcomes the weakness of mode aliasing, noise sensitivity, boundary effects, etc.; (2) VMd is less sensitive to noise; (3) VMd can distinguish sub-sequences with similar frequencies; (4) VMd decomposes are more thorough, and the residual noise of each component is very small.

*4.2.3. Comparison III: Comparisons of models before and after outlier processing*

In this study, the forecasting results of the proposed model are compared with that of $H_4$ to verify the effectiveness of the outlier process on the improvement of prediction accuracy. It can be concluded that **the hybrid models combined with outlier processing have better prediction performance**. More details, for case in Guangzhou, the $\mathbf{MAPE}_{GZ}$ before using HI is 8.1709%, and after HI processing, $\mathbf{MAPE}_{GZ}$ is reduced to 7.4723%, the $\mathbf{MAE}_{GZ}$ value of the proposed model (after HI processing) is 1.7878, and $\mathbf{MAE}_{GZ}$ before HI processing is 2.0073. As for the case in Shenzhen, the $\mathbf{MAPE}_{SZ}$ of the proposed model is 7.2571%, that of the model without HI is 8.7864%. And the $\mathbf{MAE}_{SZ}$ of the proposed model is 1.6437, that of $H_4$ is 2.0156. Combining the comparison results of the another metric, it can be concluded that the HI method can improve the predictive ability of models. Based on the data from Zhuhai, the $\mathbf{MAPE}_{ZH}$ values of the proposed model and $H_4$ are **[10.1336**%, **10.4727**%**]**, and $\mathbf{RMSE}_{ZH}$ values of the two models are **[2.6873, 2.9346]**. Same as the results for Shenzhen, the results for all indicators in Zhuhai indicate that the proposed model outperforms $H_4$ in terms of prediction .

In addition, to quantify and analyze the effectiveness of HI, the Sample Entropy of the original series and processed series is calculated. The results in Table 4 show that after HI, the complexity of the original series is reduced, and the learnability of the data is improved.

**Remark.** The presence of outliers in a series will increase the burden and difficulty of modeling, so outlier processing is very important. HI is a variant of the 3-sigma rule in statistics, which is robust to outlier and is an effective way to handle them.

*4.2.4. Comparison IV: Comparisons of the proposed model and the benchmark models*

The prediction results of the proposed model and the benchmark models are shown in Fig. 2. , and the results of the three evaluation metrics are shown in Table 5. The results indicate that the prediction accuracy of the proposed model is superior to the other models. Taking MAPE and MAE as examples, it is clearly seen from Fig. 3 that the proposed model has the lowest MAPE and MAE among all models. Therefore, **the model proposed in this paper can be considered superior to the comparison models.**

Using (30), improvement percentages of each indicator can be calculated for the proposed model relative to the comparison model. Where $V_m^{Model1}$ represents the value of the metric $m$ for Model 1, and $V_m^{Model2}$ is the value of metric $m$ for Model 2. The calculation results are shown in Table 6. Taking Guangzhou as an example, $\mathbf{P_{MAPE}=71.52\%}$ is the improvement percentage of the proposed model vs. $S_1$, which means that the $MAPE_{GZ}$ value of the proposed model is reduced by 71.52% for $S_1$, and it can also be said **that the predictive ability of the proposed model is improved by 71.52%with respect to** $\mathbf{S}_1$. The results of other indicators are interpreted in the same way. According to the results in the Table 6, it can be seen that compared with other comparison models, the prediction ability of the proposed model has been improved, especially compared with the individual models, the prediction ability of the proposed model has been improved the most.

$$\mathbf{P}_{metric} = \frac{V_m^{Model1} - V_m^{Model2}}{V_m^{Model1}}. \tag{30}$$

**Remark.** The prediction accuracy of almost all the hybrid models is higher than that of the single models. And the optimization algorithms have limited effect on improving the prediction accuracy, while the decomposition techniques have

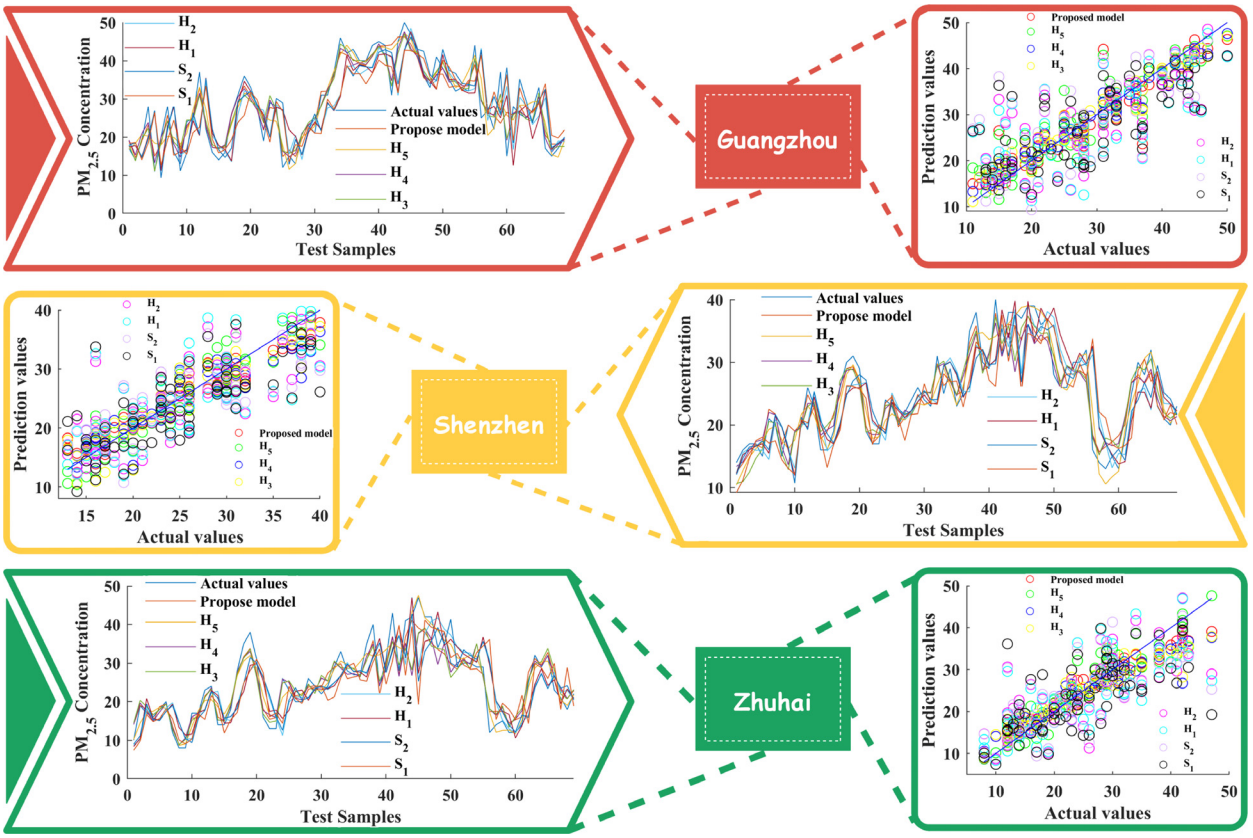


**Fig. 2.** The lines in the figure represent actual values and the prediction results of different prediction models. The closer the curve of the prediction result is to the actual value curve, the better the prediction of the method. The circles in the figure represent the relationship between the predicted value and the actual value. When the predicted value is equal to the actual value, the circle is on the diagonal, that is, the closer the circle is to the diagonal, the better the prediction. If the circle is above the diagonal line, it indicates that the predicted value is greater than the true value, otherwise, the predicted value is less than the true value.

**Table 6**
Comparison between the developed model and other models.

| Compared Models | Guangzhou | | | | Shenzhen | | | | Zhuhai | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $P_{MAE}$ | $P_{RMSE}$ | $P_{MAPE}$ | DM | $P_{MAE}$ | $P_{RMSE}$ | $P_{MAPE}$ | DM | $P_{MAE}$ | $P_{RMSE}$ | $P_{MAPE}$ | DM |
| P vs. $S_1$ | 71.30 | 71.33 | 71.52 | 5.6219* | 60.48 | 63.29 | 59.95 | 4.1718* | 60.52 | 65.00 | 58.96 | 3.6816* |
| P vs. $S_2$ | 71.19 | 72.18 | 72.31 | 5.2066* | 59.27 | 62.10 | 59.40 | 4.4394* | 60.67 | 62.59 | 59.19 | 4.4014* |
| P vs. $H_1$ | 69.77 | 71.53 | 70.77 | 5.8037* | 59.94 | 63.77 | 59.13 | 4.2233* | 60.73 | 62.65 | 59.18 | 4.0025* |
| P vs. $H_2$ | 70.03 | 71.08 | 70.69 | 5.6902* | 57.99 | 61.41 | 57.99 | 4.3491* | 59.71 | 62.85 | 58.28 | 3.9476* |
| P vs. $H_3$ | 19.11 | 18.04 | 17.18 | 2.6222* | 19.68 | 19.41 | 23.27 | 2.4556** | 14.05 | 16.73 | 10.02 | 2.8658* |
| P vs. $H_4$ | 10.94 | 10.10 | 8.55 | 1.5410*** | 18.45 | 18.82 | 17.41 | 3.0771* | 4.64 | 8.43 | 3.24 | 1.3567*** |
| P vs. $H_5$ | 36.17 | 36.90 | 38.78 | 3.8598* | 20.47 | 23.20 | 19.78 | 2.2020** | 13.05 | 16.04 | 6.75 | 1.3321*** |

1. $S_1$: Arima; $S_2$: $\ell_{2,1}$RFELM; $H_1$: moALo-$\ell_{2,1}$RFELM; $H_2$: mSSa-$\ell_{2,1}$RFELM; $H_3$: VMd-moALo-$\ell_{2,1}$RFELM; $H_4$: VMd-mSSa-$\ell_{2,1}$RFELM; $H_5$: HI-EEMd-mSSa-$\ell_{2,1}$RFELM; P: The proposed model.
2. $P_{MAE}$, $P_{RMSE}$, $P_{MAPE}$ repersent improvement percentages of the developed model compared with other models (%).
3. $*$ indicates the 1% significance level $Z_{0.01/2} = 2.58$; $**$ indicates the 5% significance level $Z_{0.05/2} = 1.96$; $***$ indicates the 10% significance level $Z_{0.10/2} = 1.64$.

significant effect on improving the prediction accuracy of the model. The reason for the better performance of hybrid models can be summarized as follows: (1) the original series contain complex non-linearity and non-stationarity, which are often difficult to be captures by a single model; (2) the hybrid model combines the advantages of different methods, thus improving the performance of the model.

In summary, the prediction model proposed in this paper has the best prediction effect. Besides, the results of the comparative experiments in this paper show that **the optimization algorithm has limited effect on the improvement of prediction accuracy**, and different optimization algorithms differ in terms of improving the prediction accuracy. In this study mSSa is better than moALo. Moreover, **data decomposition** is very important, particularly more efficient decompo-

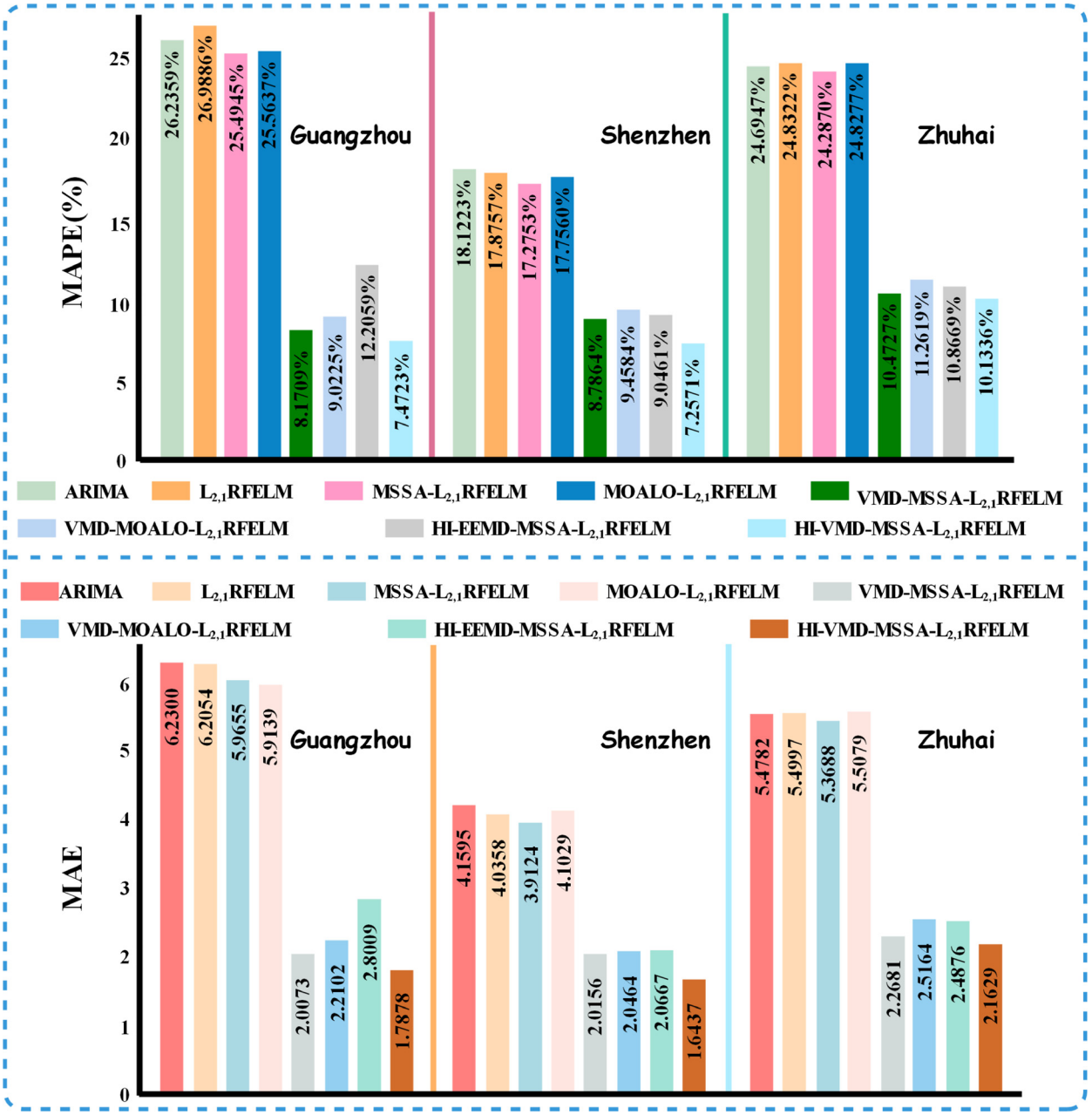


**Fig. 3.** Graph showing: the upper part are the MAPE values of different models' predictions based on the data from three study cities, and the lower part are the MAE values of different models predicting based on the data from three study cites. The height of the bars in the graph represents the magnitude of the indicator value. It is obvious from the figure that the values of MAE and MAPE for the prediction method incorporating the data decomposition algorithm decrease significantly, and the values of MAE and MAPE for the prediction model proposed in this paper are the smallest compared to other methods.

sition techniques, such as VMd, **can significantly improve the predictive ability of the models**. In addition, **the outlier processing** is also **beneficial to improve the prediction accuracy**.

### *4.3. Experiment II: Interval prediction*

In order to quantify the uncertainties associated with $PM_{2.5}$ concentration prediction, a novel scenario of interval prediction is proposed based on the deterministic prediction errors. Obtaining the finding that TLS distribution is appropriate to model the distribution of the prediction error series. Then, based on the distribution, the concentration intervals can be obtained. More details are given in the following.

**Table 7**
Estimated values of the parameters of the distribution.

| Study areas | Gaussian | | Logistic | | TLS | | |
|---|---|---|---|---|---|---|---|
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\nu$ |
| Guangzhou | −0.0360 | 2.6816 | −0.0907 | 1.4210 | −0.1163 | 1.9842 | 4.1136 |
| Shenzhen | −0.0400 | 2.3199 | −0.1149 | 1.2375 | −0.1445 | 1.7238 | 4.0537 |
| Zhuhai | −0.0023 | 2.8017 | −0.1220 | 1.5046 | −0.1564 | 2.1624 | 4.6178 |

1.Using the maximum likelihood estimation method to estimate.

**Table 8**
Evaluation of fitting effect of three distributions.

| Study cities | $\mathbf{R^2}$ | | | $\mathbf{RMSE_{IP}}$ | | |
|---|---|---|---|---|---|---|
| | Gaussian | Logistic | TLS | Gaussian | Logistic | TLS |
| Guangzhou | 0.9488 | 0.9871 | **0.9901** | 0.0124 | 0.0062 | **0.0055** |
| Shenzhen | 0.9324 | 0.9791 | **0.9879** | 0.0170 | 0.0095 | **0.0072** |
| Zhuhai | 0.9521 | 0.9897 | **0.9928** | 0.0116 | 0.0054 | **0.0045** |

1.The bold numbers indicate the optimal value of the indicators.

*4.3.1. Distribution fitting of the prediction error*

Multifarious distributions have been introduced to describe time series in different fields. Such as, applied Burr, Logistic, Gamma, Weibull, and Rician distributions to simulate the $PM_{2.5}$ series, than based on the simulated results calculated the health-related economic loss under different pollution levels [37]. Compared the fitting effect of the Gaussian distribution and TLS distribution to the prediction error of wind power [38]. In this study, three distributions are utilized to represent the prediction error series, Gaussian distribution, Logistic distribution and T Location-scale distribution.

Based on the given series, the parameters of three distributions are estimated by using maximum likelihood estimation, the results shown in the Table 7. Both the Gaussian and Logistic distributions have two parameters, and the meaning of each parameter is the same for both distributions. $\mu$ is the location parameter, which determines the location of distribution, and $\sigma$ is the scale parameter that determines the magnitude of the distribution. The shape of the logistic distribution is similar to the shape of the normal distribution, but the logistic distribution has heavier tails. The TLS distribution has three parameters, $\mu$ is location parameter, $\sigma$ is scale parameter and $\nu$ is shape parameter.

In the fitting part, two indicators shown in the Table 2 (Distr Fitting) are used to evaluate the fitting, and the evaluation results are given in Table 8. As we can see, for Guangzhou, the $R^2$ values of different distributions are **[0.9488, 0.9871, 0.9901]**. According to the rule that the greater the value of $R^2$, the better the fit, it can be tentatively determined that the TLS distribution fits the data from Guangzhou the best. Further, observing the results of $\mathbf{RMSE_{IP}}$ in the Table 8, we can conclude that the TLS distribution fits better based on the data from Guangzhou. The results for the other two study cities are also show that **the TLS distribution fits the series the best**.

*4.3.2. Interval prediction of concentration series*

Different from the deterministic prediction, interval prediction offers the upper and lower bounds of the observed values. In this study, based on the error series and the best-fit distribution, interval predictions are performed at a given significant level $\alpha$. Additionally, three evaluation indicators, **IPCP, IPNW, and AWD**, are utilized to evaluate the performance of the interval prediction.

The interval prediction results are listed in Table 9. Based on the data from Guangzhou, the coverage probability is 100% of three distributions under 95% confidence level, thus the out-of-interval deviation (AWD) is 0.0000 for these three distributions. The three distributions perform differently in interval prediction under the other confidence levels. When the confidence level is 85%, the interval derived by the Gaussian distribution covers 89.86% of the sample points, so there is 10.14% points out of the interval, the normalized deviation is 0.2780. Taking Zhuhai as an example, when $\boldsymbol{\alpha = 0.05}$ the coverage of the interval obtained by all the three distributions is 95.65% **(IPCP = 95.65%)**. And average interval widths obtained by the three distributions are different. The interval obtained by the Gaussian distribution is narrowest, while the widest average interval width is obtained by the TLS distribution **($\mathbf{IPNAW}_t$= 0.3574)**. But, the out-of-interval deviation (AWD) is the largest for the Gaussian distribution. For $\boldsymbol{\alpha = 0.1}$, the Gaussian distribution holds the largest interval coverage **($\mathbf{IPCP}_g$** = 89.86%**)** and the smallest out-of-interval deviation **($\mathbf{AWD}_g$= 0.0099)**, but it has the widest average interval width **($\mathbf{IPNAW}_g$= 0.2890)**. In contrast, the AWD of the TLS distribution is **only 0.0012** larger than that of the Gaussian distribution, however its average interval width is much smaller than that of the Gaussian distribution. The interval prediction results in other significant levels are similar to these two cases.

Figure 4 visualizes the interval prediction results of three distributions with two significance levels: **0.05 and 0.2** based on the data from **Shenzhen**. In the figure, the points are the actual values of the series and the shadow areas are the predicted intervals. For example, the **sub-graph (a)** illustrates the predicted intervals for $PM_{2.5}$ concentration based on Gaussian

**Table 9**
Interval prediction results under different significant levels.

| Confidence level (%) | Distributions | Guangzhou | | | Shenzhen | | | Zhuhai | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IPCP(%) | AWD | IPNAW | IPCP(%) | AWD | IPNAW | IPCP(%) | AWD | IPNAW |
| 95 | G | 100.00 | 0.0000 | 0.3313 | 98.55 | 0.0012 | 0.3579 | 95.65 | 0.0034 | **0.3444** |
| | L | 100.00 | 0.0000 | **0.3281** | 98.55 | 0.0014 | **0.3569** | 95.65 | 0.0028 | 0.3457 |
| | T | 100.00 | 0.0000 | 0.3435 | 98.55 | **0.0010** | 0.3748 | 95.65 | **0.0019** | 0.3574 |
| 90 | G | 94.20 | **0.0024** | 0.2780 | **95.65** | **0.0034** | 0.3004 | **89.86** | **0.0099** | 0.2890 |
| | L | 94.20 | 0.0039 | **0.2637** | 94.20 | 0.0045 | **0.2868** | 88.41 | 0.0114 | **0.2778** |
| | T | 94.20 | 0.0037 | 0.2645 | 94.20 | 0.0043 | 0.2882 | 88.41 | 0.0111 | 0.2783 |
| 85 | G | **89.86** | **0.0072** | 0.2433 | **91.30** | **0.0082** | 0.2629 | 85.51 | **0.0201** | 0.2529 |
| | L | 84.06 | 0.0120 | 0.2250 | 89.86 | 0.0118 | 0.2447 | 85.51 | 0.0249 | 0.2370 |
| | T | 84.06 | 0.0135 | **0.2210** | 88.41 | 0.0129 | **0.2405** | 85.51 | 0.0258 | **0.2339** |
| 80 | G | **84.06** | **0.0159** | 0.2433 | **88.41** | **0.0149** | 0.2340 | **84.06** | **0.0319** | 0.2252 |
| | L | 79.71 | 0.0263 | 0.2250 | 85.51 | 0.0223 | 0.2140 | 75.36 | 0.0412 | 0.2073 |
| | T | 79.71 | 0.0303 | **0.2210** | 85.51 | 0.0252 | **0.2075** | 71.01 | 0.0449 | **0.2027** |
| 75 | G | **79.71** | **0.0278** | 0.1944 | **85.51** | **0.0240** | 0.2101 | **76.81** | **0.0468** | 0.2021 |
| | L | 76.81 | 0.0436 | 0.1743 | 82.61 | 0.0350 | 0.1895 | 68.12 | 0.0654 | 0.1836 |
| | T | 76.81 | 0.0503 | **0.1674** | 81.16 | 0.0404 | **0.1820** | 68.12 | 0.0721 | **0.1783** |
| 70 | G | **76.81** | **0.0425** | 0.1752 | **82.61** | **0.0352** | 0.1893 | **69.57** | **0.0674** | 0.1821 |
| | L | 72.46 | 0.0638 | 0.1554 | 78.26 | 0.0511 | 0.1690 | 66.67 | 0.0929 | 0.1637 |
| | T | 71.01 | 0.0737 | **0.1482** | 75.36 | 0.0596 | **0.1611** | 66.67 | 0.1020 | **0.1582** |

1. G represents Gaussian distribution; L represents Logistic distribution; T represent T Location-Scale distribution.
2. The bold numbers indicate the optimal value of the indicators.
3. Degree of confidence = 1 - $\alpha$.
4. **IPCP** $= (1/N)\sum_{i=1}^{N} c_i$, where $c_i = \begin{cases} 1, \text{ if } F_i \in [L_i, U_i] \\ 0, \text{ otherwise} \end{cases}$ ; **AWD** $= (1/N)\sum_{i=1}^{N} \text{AWD}_i^{\alpha}$, where $\text{AWD}_i^{\alpha} = \begin{cases} (L_i^{\alpha} - A_i)/(U_i^{\alpha} - L_i^{\alpha}),\ A_i < L_i^{\alpha} \\ 0,\ A_i \in I_i^{\alpha} \\ (A_i - U_i^{\alpha})/(U_i^{\alpha} - L_i^{\alpha}),\ A_i > U_i^{\alpha} \end{cases}$ ; **IPNAW** $= (1/N)\sum_{i=1}^{N}(U_i - L_i)/(\overline{\overline{F}} - \underline{\underline{F}})$.

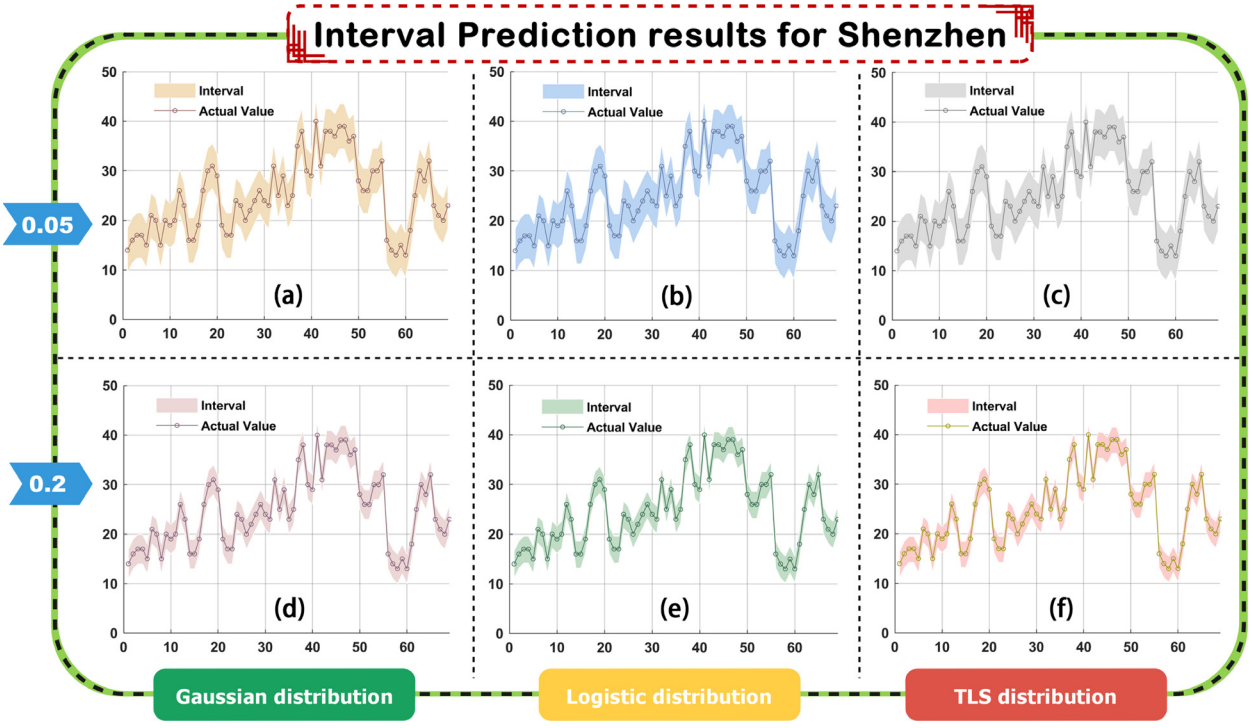


**Fig. 4.** The interval prediction results in Shenzhen at two significance levels. The prediction intervals at the significance level of 0.05 are depicted in (a)–(c), and the intervals at the significance level of 0.2 are shown in (d)–(f). Here, (a) and (d) are results of the Gaussian distribution, it can be seen that the predicted interval at the significance level of 0.05 is wider than that when the significance level is 0.2, and all the points are in the shadow area. The same is true for the other two distributions, such as (b) and (e) of Logistic distribution, (c) and (f) of TLS distribution.

distribution at the significance level 0.05, the maximum value of predicted interval no more than 50. From Fig. 4, the interval prediction performance is excellent as most of the sample points lie in the shaded areas.

**Remark.** In practice, intervals with high IPCP are not always the best intervals. When the IPCP is primarily considered, a smaller $\alpha$ is preferred, while a larger $\alpha$ will be chosen when a smaller IPNAW is mainly considered.

## 5. Discussions

To verify the prediction performance of the developed combined model, further discussions are also performed in this section, including computational complexity analysis, sensitivity analysis and statistical test.

### 5.1. Computational complexity analysis

To investigate the efficiency of the prediction system, the complexity analysis of the data decomposition module and the prediction module is revealed theoretically in this section.

$\mathcal{A}$ *Complexity measuring of data preprocessing module.* Given a series of length $n$, used the outlier processing method HI to correct the abnormal points in this series, utilized SampEn to calculate the complexity of the series before and after processing, and applied VMd to reduce the noise in the series. In the HI, the first step is to calculate the local median in (2.1.1), the computational of this operation is $\mathcal{O}(n)$. And the following operations are to calculate the difference and median, the complexity of these two operations is $\mathcal{O}(n)$. Therefore, the computational complexity of HI is $\mathcal{O}(n)$. For SampEn, the computation is based on the measure of the correlation integral, and it requires a computational complexity of $\mathcal{O}(n^2)$ to direct implementation of the correlation integral [39].

The computational complexity of each step of VMd can be specified as follows. According to the introduce in Section 2, assuming that the original series is decomposed into $G$ sub-series, the final decomposition results are obtained after $T_{vmd}$ iterations. In the first step, initialize BLIMFs $\ddot{\mathbf{X}}_g$ and center frequencies $\omega_g$, $g = 1, \ldots, G$, and Lagrange variables $\lambda$, needs $\mathcal{O}(2nlog_2 2n)$ computational complexity. The computational complexity of the process of updating $\ddot{\mathbf{X}}_g$ and $\omega_g$ is $(6A + 2M + 2D) \cdot G \cdot T_{vmd} \cdot 2n$[1] and $(2C + 3M + 2A) \cdot G \cdot T_{vmd} \cdot n^2$ And the computational complexity of find Lagrange variable is $(4A + M) \cdot T_{vmd} \cdot 2n$. As a result, the computational complexity of VMD in the worest case is $\mathcal{O}(2nlog_2 2n)$ [33].

$\mathcal{B}$ *Complexity consideration of prediction module.* In the proposed prediction system, mSSa-$\ell_{2,1}$RFELM was developed as a prediction model, in which mSSa plays an important role in parameter optimization of $\ell_{2,1}$RFELM. The computation complexity of mSSa has an important impace on that of the prediction system. Thus, the complexity of mSSa is conducted. As introduced in Section 2, the number of objectives in this study is 2, the number of variables is $\mathbf{D}$, the number of the search agent is $\mathbf{Z}$, the computational complexity of one iteration of mSSa is $\mathcal{O}(\mathrm{D} \times \mathrm{Z} + cof \times \mathrm{Z} + 2\mathrm{Z}^2)$, here $cof$ denoted the cost of objective functions [27]. In this study, two objective functions are selected, as shown in the (19), the computational complexity of the $obf_1$ is $\mathcal{O}(n^2)$, and that of the $obf_2$ is also $\mathcal{O}(n^2)$. Therefore, the computational complexity of one iteration of mSSa in this study is $\mathcal{O}(\mathrm{D} \times \mathrm{Z} + n^2 \times \mathrm{Z} + 2\mathrm{Z}^2)$.

In the $\ell_{2,1}$RFELM, the computational complexity is mainly contributed by the following operations: $\mathbf{H}^{\mathrm{T}}\mathbf{H}$, $\mathbf{H}^{\mathrm{T}}\boldsymbol{\mathcal{Y}}^{\mathrm{T}}$, $\mathbf{D}$, and $\left(\frac{\mathbf{D}}{\bar{\mathrm{C}}} + \mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}$. The computation complexity of $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ is $\mathcal{O}(\mathrm{D}^2_{hidden})$, where $\mathrm{D}_{hidden}$ is the number of neurons in hidden layer. For an output matrix of size $\mathrm{I} \times 1$, the computational complexity of $\mathbf{H}^{\mathrm{T}}\boldsymbol{\mathcal{Y}}^{\mathrm{T}}$ and $\mathbf{D}$ are $\mathcal{O}(\mathrm{D}_{hidden}\mathrm{I})$ and $\mathcal{O}(\mathrm{D}^2_{hidden})$ [40]. And in one iteration, it takes $\mathcal{O}(\mathrm{D}^3_{hidden})$ computational efforts to calculate $\left(\frac{\mathbf{D}}{\bar{\mathrm{C}}} + \mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}$. Therefore, the final computational complexity of one iteration is $\mathcal{O}(\mathrm{D}^3_{hidden})$.

### 5.2. Sensitivity to input perturbation

To test the sensitivity to input perturbation of the proposed prediction system, the data sets from three study areas and one random series were used assess the effect of small changes in input data on the prediction output. By adding white noise with a variance of 0.1 to 0.5 to the input data, the input data is small variations, and then the MAE, RMSE and MAPE of the predicted results after adding different white noises are compared.

Table 10 shows that the MAPE value for the Guangzhou data without added noise is 7.4723%, and the maximum MAPE (8.2263%) obtained when the variance of white noise is 0.4, the number of MAPE only changes by 0.754. This means that when white noises with a variance range of [0.1,0.5] is added, the change in MAPE value is less than 0.754, and the minimum change value is 0.0569 compared with the prediction result without adding perturbed data. In addition, the percentage of MAPE change is in the range of [0.76%, 7.79%]. The results are similar for other study areas. The percentage of MAPE change for Shenzhen and Zhuhai is in the range of [0.33%, 5.34%] and [1.79%, 7.40%], respectively. Besides, from the perspective of MAE and RMSE, the variations between the prediction results with and without perturbations are slight. Figure 5 shows the predicted performance of different study areas with different input disturbances. It can be found that the prediction

[1] $A$ represents the addition operation; $M$ represents the multiplication operation; $D$ represents the division operation.

[2] $C$ represents the comparison operation.

**Table 10**
Prediction results of sensitivity to input perturbations.

Prediction results based on data with white noise and data without white noise

| Variance of white noise | Guangzhou | | | Shenzhen | | | Zhuhai | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAPE (%) | MAE | RMSE | MAPE (%) | MAE | RMSE | MAPE (%) |
| Without white noise | 1.7878 | 2.2362 | 7.4723 | 1.6437 | 1.9536 | 7.2571 | 2.1629 | 2.6873 | 10.1336 |
| 0.1 | 1.8506 | 2.2737 | 7.6009 | 1.6280 | 1.9198 | 7.3071 | 2.2663 | 2.8598 | 10.3146 |
| 0.2 | 1.8763 | 2.3581 | 7.9606 | 1.6769 | 1.9950 | 7.4995 | 2.3312 | 2.8869 | 10.8837 |
| 0.3 | 1.8311 | 2.3024 | 7.7194 | 1.6817 | 2.0021 | 7.4744 | 2.2409 | 2.8030 | 10.4262 |
| 0.4 | 1.9099 | 2.4401 | 8.0540 | 1.6458 | 1.9031 | 7.2810 | 2.2169 | 2.7470 | 10.3281 |
| 0.5 | 1.8310 | 2.2389 | 7.5292 | 1.7088 | 2.0085 | 7.6444 | 2.2549 | 2.8018 | 10.4905 |

Improvement percentages of the prediction without white noise compared with the prediction with white noise (%)

| | $P_{mae}$ | $P_{rmse}$ | $P_{mape}$ | $P_{mae}$ | $P_{rmse}$ | $P_{mape}$ | $P_{mae}$ | $P_{rmse}$ | $P_{mape}$ |
|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 3.51 | 1.68 | 1.72 | 0.96 | 1.73 | 0.69 | 4.78 | 6.42 | 1.79 |
| 0.2 | 4.95 | 5.45 | 6.54 | 2.02 | 2.12 | 3.34 | 7.78 | 7.43 | 7.40 |
| 0.3 | 2.42 | 2.96 | 3.31 | 2.31 | 2.48 | 2.99 | 3.60 | 4.30 | 2.89 |
| 0.4 | 6.83 | 9.12 | 7.79 | 0.13 | 2.59 | 0.33 | 2.50 | 2.22 | 1.92 |
| 0.5 | 2.42 | 0.12 | 0.76 | 3.96 | 2.81 | 5.34 | 4.26 | 4.26 | 3.52 |

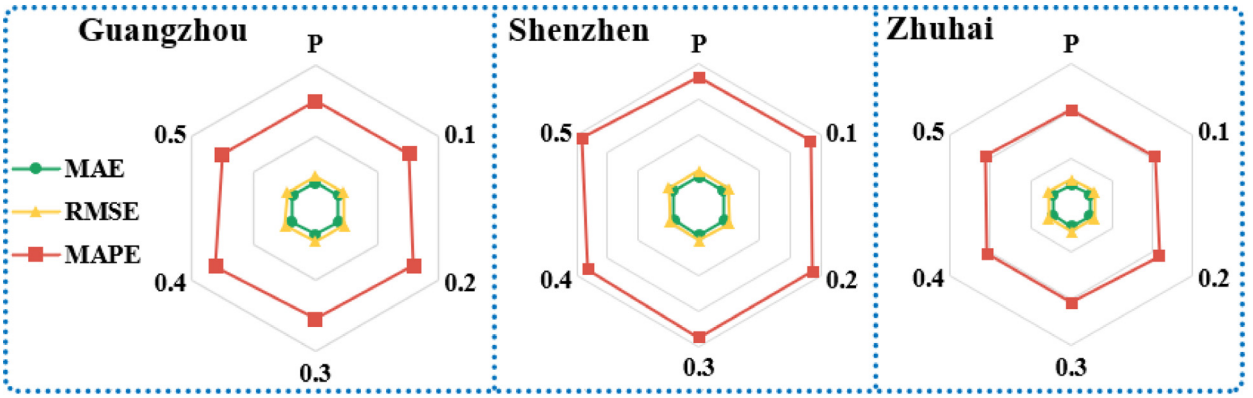


**Fig. 5.** Prediction results based on different input data. In the figure, 0.1–0.5 represent the variance of the white noise added to the input series, and P represents that no white noise is added to the input series. The results from three study cities show that the prediction results are similar after adding different perturbations to the original input series, since the values of the three evaluation indicators are similar based on the six input series.

results after adding random perturbation to the input data are slightly different from the result without adding random perturbation. Thus, the proposed model has low sensitivity to perturbed input.

### 5.3. Statistical analysis and comparisons

In general, prediction performance evaluation indicators, such as MAE, RMSE, and MAPE, can effectively evaluate the prediction model. However, relying only on these indicators is one-sided, because the prediction results are affected by the sampling results, and the test data is just a small part of all the data. Therefore, the Diebold-Mariano (DM) test [41] is introduced to determine whether the two prediction models are significantly different at a given significance level, and to discuss the superiority of the proposed model compared to other comparison models from the statistical significance.

Let $e_1$ and $e_2$ are the prediction error from method 1 and method 2, respectively. Given the loss function $L(e_i^n) = (e_i^n)^2, n = 1, 2, \ldots, N$, the loss differential between these two methods can be calculated as $d_{12}^n = L(e_1^n) - L(e_2^n)$. Then, the DM statistic is calculated as follows:

$$\mathbf{DM} = \frac{\sum_{n=1}^{N} d_{12}^n / N}{\sqrt{S^2/N}} \tag{31}$$

here, $S^2$ is the variance estimator of the $d_{12}$.

The null hypothesis of the DM test is $\mathbf{H}_0$: $d_{12} = 0$, which means that the two prediction results have no significant difference. On the contrary, the alternative hypothesis is $\mathbf{H}_1$: $d_{12} \neq 0$, which means there are significant differences in methods. DM statistical follows standard normal distribution, if the value of DM statistic falls outside the confidence intercal, the null hypothesis will be rejected. Specific results of DM test are shown in the DM columns in Table 6. For Guangzhou and Shenzhen, the values of the DM statistic are greater than the limit corresponding to the 5% significance level, which means that

the proposed method is significantly different from the other compared model under 5% significance level. And for Zhuhai, except $H_4$ and $H_5$, all the values of DM test are greater than the limit corresponding to the 1% significance level. These test results reveal the superiority of the proposed prediction model.

## 6. Conclusions

As pointed out in this paper, the single model can not capture complex non-linearity and non-stationarity characteristics of the original series. In addition, the uncertainty information of the pollutant concentration are rarely studied. To fill this gap, in this study, a novel hybrid system containing deterministic and interval predictions is designed to overcome the above problems. Theoretical and simulation experiments demonstrate the effectiveness of the proposed system. The prediction system comprises the following steps.

(1) Using HI method detect and correct the outliers in the original concentration series. This data processing step can reduce the complexity of the original series.
(2) The series after corrected is decomposed into several sub-series. These sub-series contain different characteristics hidden in the original concentration series.
(3) Applied the optimized model, mSSa-$\ell_{2,1}$RFELM, to predict each sub-series. The prediction results of the sub-series are summed to obtain the final prediction result.
(4) Based on the deterministic prediction results, interval prediction is subsequently performed to analyze the uncertainties of pollutant concentration.

The proposed system is tested for $PM_{2.5}$ prediction in three study cities. The results of the experiments show that the proposed deterministic prediction model, HI-VMd-mSSa-$\ell_{2,1}$RFELM, achieves optimal prediction performance for these three cities. Moreover, seven benchmark models are selected to validate the prediction ability of the proposed model. Relative to the Arima model, the **improvement percentages of MAPE for the proposed model are 71.52**%, 59.95%**and 58.96**% for the three study areas, respectively. For Guangzhou, relative to the optimized models $H_1$ and $H_2$, **the proposed model is improved by 70.77**% and 70.69%**in terms of the MAPE**. As for the other comparison experiments of the three series, the results also show that the proposed model in this paper has the best prediction performance. Furthermore, the results also demonstrated the robust ability of the decomposition technique in improving forecasting accuracy. On the basis of deterministic forecasting, three distributions are introduced to fit the prediction error series, and the parameters of distributions are determined by Maximum Likelihood Estimation (MLE). The results show that TLS distribution is the best-fitted distribution with the highest $R^2$ and the lowest RMSE. Then the confidence interval under different significant levels and different parameters can be obtained. The interval prediction results indicate excellent performance owing to the most of the sample points fell into the shaded areas, and the predicted intervals with higher values of IPCP and smaller values of IPNAW at different significant levels. Overall, this research developed a novel hybrid model with high prediction accuracy, and found a better distribution to predict the intervals of $PM_{2.5}$ concentrations. Additionally, it was found that data decomposition can significantly improve the prediction ability of models, and TLS distribution can better fit the prediction error series, making the results of uncertainty prediction better.

The advantages of this study are the use of data preprocessing strategy to correct the abnormal data and filters the noise in the original series to make the prediction results more accurate, which is of great significance in the prediction, and combining interval prediction in the prediction model to measure the uncertainty associated with the pollution concentration. However, some limitations are also existing in this work needing to be further researched, such as, only $PM_{2.5}$ data is utilized to verify the proposed hybrid model, and other factors that affect the concentration of pollutants are not considered. In addition, there are many new prediction technologies, such as deep learning, which are not studied in this paper. In future research, some possible factors should be studied in detail to develop a prediction model that can better predict the pollutant concentration. In addition, an attempt is made to investigate prediction methods related to deep learning, and to consider the interpretability of the models.

## CRediT authorship contribution statement

**Lu Bai:** Conceptualization, Methodology, Software, Investigation, Visualization, Writing – original draft, Data curation. **Zhi Liu:** Supervision, Methodology, Writing – review & editing. **Jianzhou Wang:** Supervision, Methodology, Writing – review & editing.

## Acknowledgment

This research is partially supported by National Natural Science Foundation of China (No. 11971507).

## References

[1] Q. Zhang, J. C. K. Lam, V. O. K. Li, Y. Han, Deep-air: a hybrid CNN-LSTM framework forfine-grained air pollution forecast, (2020) arXiv preprint arXiv:2001.11957

[2] V. Spiridonov, B. Jakimovski, I. Spiridonova, G. Pereira, Development of air quality forecasting system in macedonia, based on WRF-chem model, Air Qual. Atmos. Health. 12 (7) (2019) 825–836, doi:10.1007/s11869-019-00698-5.
[3] S.K. Sahu, S. Sharma, H. Zhang, V. Chejarla, H. Guo, J. Hu, Q. Ying, J. Xing, S.H. Kota, Estimating ground level PM2.5 concentrations and associated health risk in India using satellite based AOD and WRF predicted meteorological parameters, Chemosphere 255 (2020) 126969, doi:10.1016/j.chemosphere.2020.126969.
[4] F.-Y. Cheng, C.-Y. Feng, Z.-M. Yang, C.-H. Hsu, K.-W. Chan, C.-Y. Lee, S.-C. Chang, Evaluation of real-time PM2.5 forecasts with the WRF-CMAQ modeling system and weather-pattern-dependent bias-adjusted PM2.5 forecasts in taiwan, Atmos. Environ. 244 (2021) 117909, doi:10.1016/j.atmosenv.2020.117909.
[5] S.G. Gocheva-Ilieva, D.S. Voynikova, M.P. Stoimenova, A.V. Ivanov, I.P. Iliev, Regression trees modeling of time series for air pollution analysis and forecasting, Neural Comput. Appl. 31 (12) (2019) 9023–9039, doi:10.1007/s00521-019-04432-1.
[6] M. Zeinalnezhad, A.G. Chofreh, F.A. Goni, J.J. Klemeš, Air pollution prediction using semi-experimental regression model and adaptive neuro-fuzzy inference system, J. Clean. Prod. 261 (2020) 121218, doi:10.1016/j.jclepro.2020.121218.
[7] J.W. Koo, S.W. Wong, G. Selvachandran, H.V. Long, L.H. Son, Prediction of air pollution index in Kuala Lumpur using fuzzy time series and statistical models, Air Qual. Atmos. Health 13 (1) (2020) 77–88, doi:10.1007/s11869-019-00772-y.
[8] Y. Shao, J. Wang, H. Zhang, W. Zhao, An advanced weighted system based on swarm intelligence optimization for wind speed prediction, Appl. Math. Model. 100 (2021) 780–804, doi:10.1016/j.apm.2021.07.024.
[9] H. Yan, T. Zhang, Y. Qi, D.-J. Yu, Short-term traffic flow prediction based on a hybrid optimization algorithm, Appl. Math. Model. 102 (2022) 385–404, doi:10.1016/j.apm.2021.09.040.
[10] J. Wang, H. Li, H. Yang, Y. Wang, Intelligent multivariable air-quality forecasting system based on feature selection and modified evolving interval type-2 quantum fuzzy neural network, Environ. Pollut. 274 (2021) 116429, doi:10.1016/j.envpol.2021.116429.
[11] H. Liu, Y. Xu, C. Chen, Improved pollution forecasting hybrid algorithms based on the ensemble method, Appl. Math. Model. 73 (2019) 473–486, doi:10.1016/j.apm.2019.04.032.
[12] M. Xie, L. Wu, B. Li, Z. Li, A novel hybrid multivariate nonlinear grey model for forecasting the traffic-related emissions, Appl. Math. Model. 77 (2020) 1242–1254, doi:10.1016/j.apm.2019.09.013.
[13] R. Li, Y. Dong, Z. Zhu, C. Li, H. Yang, A dynamic evaluation framework for ambient air pollution monitoring, Appl. Math. Model. 65 (2019) 52–71, doi:10.1016/j.apm.2018.07.052.
[14] Y. Wang, J. Wang, Z. Li, A novel hybrid air quality early-warning system based on phase-space reconstruction and multi-objective optimization: a case study in China, J. Cleaner Prod. 260 (2020) 121027, doi:10.1016/j.jclepro.2020.121027.
[15] J. Li, G. Lu, T. Niu, J. Zhang, Developing an online air quality warning system based on streaming data for dynamic environmental management, J. Cleaner Prod. 273 (2020) 122953, doi:10.1016/j.jclepro.2020.122953.
[16] Z. Du, J. Heng, M. Niu, S. Sun, An innovative ensemble learning air pollution early-warning system for China based on incremental extreme learning machine, Atmos. Pollut. Res. 12 (9) (2021) 101153, doi:10.1016/j.apr.2021.101153.
[17] M. Zhang, H. Li, New evolutionary game model of the regional governance of haze pollution in China, Appl. Math. Model. 63 (2018) 577–590, doi:10.1016/j.apm.2018.07.008.
[18] Y. Song, S. Qin, J. Qu, F. Liu, The forecasting research of early warning systems for atmospheric pollutants: a case in Yangtze River Delta region, Atmos. Environ. 118 (2015) 58–69, doi:10.1016/j.atmosenv.2015.06.032.
[19] J. Stevens, Intermediate Statistics: A Modern Approach, Third Edition, 1999, doi:10.4324/9781410601643.
[20] Z. Yao, R. Wang, J. Zhi, Q. Shi, Robust locally weighted regression for profile measurement of magnesium alloy tube in hot bending process, Math. Probl. Eng. 2020 (2020) 7952649, doi:10.1155/2020/7952649.
[21] D. Cuesta-Frau, D. Novák, V. Burda, A. Molina-Picó, B. Vargas, M. Mraz, P. Kavalkova, M. Benes, M. Haluzik, Characterization of artifact influence on the classification of glucose time series using sample entropy statistics, Entropy. 20 (11) (2018), doi:10.3390/e20110871.
[22] K. Dragomiretskiy, D. Zosso, Variational mode decomposition, IEEE Trans. Signal Process. 62 (3) (2014) 531–544, doi:10.1109/TSP.2013.2288675.
[23] P. Du, J. Wang, W. Yang, T. Niu, Point and interval forecasting for metal prices based on variational mode decomposition and an optimized outlier-robust extreme learning machine, Resour. Policy 69 (2020) 101881, doi:10.1016/j.resourpol.2020.101881.
[24] S. Sun, F. Jin, H. Li, Y. Li, A new hybrid optimization ensemble learning approach for carbon price forecasting, Appl. Math. Model. 97 (2021) 182–205, doi:10.1016/j.apm.2021.03.020.
[25] H. Li, H. Lv, T. Zhang, Q. Han, J. Liu, J. Xiong, Z. Guan, Modeling and evaluation of dynamic degradation behaviours of carbon fibre-reinforced epoxy composite shells, Appl. Math. Model. 104 (2022) 21–33, doi:10.1016/j.apm.2021.11.015.
[26] W. Song, H. Liu, E. Zio, Long-range dependence and heavy tail characteristics for remaining useful life prediction in rolling bearing degradation, Appl. Math. Model. 102 (2022) 268–284, doi:10.1016/j.apm.2021.09.041.
[27] S. Mirjalili, A.H. Gandomi, S.Z. Mirjalili, S. Saremi, H. Faris, S.M. Mirjalili, Salp swarm algorithm: a bio-inspired optimizer for engineering design problems, Adv. Eng. Softw. 114 (2017) 163–191, doi:10.1016/j.advengsoft.2017.07.002.
[28] Z. Cheng, J. Wang, A new combined model based on multi-objective salp swarm optimization for wind speed forecasting, Appl. Soft Comput. J. 92 (2020) 106294, doi:10.1016/j.asoc.2020.106294.
[29] S. Zhou, X. Liu, Q. Liu, S. Wang, C. Zhu, J. Yin, Random Fourier extreme learning machine with $\ell_{2,1}$-norm regularization, Neurocomputing 174 (2016) 143–153, doi:10.1016/j.neucom.2015.03.113.
[30] J. Yuan, High dimensional data reconstruction based on l2,1 norm, Appl. Math. Model. 89 (2021) 1764–1774, doi:10.1016/j.apm.2020.08.055.
[31] A. Rahimi, B. Recht, Random features for large-scale kernel machines, in: Proceedings of the 20th International Conference on Neural Information Processing Systems, Curran Associates Inc., 2007, pp. 1177–1184.
[32] C. Tian, Y. Hao, Point and interval forecasting for carbon price based on an improved analysis-forecast system, Appl. Math. Model. 79 (2020) 126–144, doi:10.1016/j.apm.2019.10.022.
[33] Y. Liu, G. Yang, M. Li, H. Yin, Variational mode decomposition denoising combined the detrended fluctuation analysis, Sig. Process. 125 (2016) 349–364, doi:10.1016/j.sigpro.2016.02.011.
[34] Y. Wang, R. Markert, J. Xiang, W. Zheng, Research on variational mode decomposition and its application in detecting rub-impact fault of the rotor system, Mech. Syst. Signal Process. 60-61 (2015) 243–251, doi:10.1016/j.ymssp.2015.02.020.
[35] W. Yang, J. Wang, H. Lu, T. Niu, P. Du, Hybrid wind energy forecasting and analysis system based on divide and conquer scheme: a case study in China, J. Cleaner Prod. 222 (2019) 942–959, doi:10.1016/j.jclepro.2019.03.036.
[36] Y. Hao, C. Tian, The study and application of a novel hybrid system for air quality early-warning, Appl. Soft Comput. J. 74 (2019) 729–746, doi:10.1016/j.asoc.2018.09.005.
[37] J. Wang, L. Zhang, X. Niu, Z. Liu, Effects of PM2.5 on health and economic loss: evidence from Beijing-Tianjin-Hebei region of China, J. Cleaner Prod. 257 (2020) 120605, doi:10.1016/j.jclepro.2020.120605.
[38] J. Wang, T. Niu, H. Lu, W. Yang, P. Du, A novel framework of reservoir computing for deterministic and probabilistic wind power forecasting, IEEE Trans. Sustain. Energy 11 (1) (2020) 337–349, doi:10.1109/TSTE.2019.2890875.
[39] Y.-H. Wang, I.-Y. Chen, H. Chiueh, S.-F. Liang, A low-cost implementation of sample entropy in wearable embedded systems: an example of online analysis for sleep EEG, IEEE Trans. Instrum. Meas. 70 (2021) 1–12, doi:10.1109/TIM.2020.3047488.
[40] L.-R. Ren, Y.-L. Gao, J.-X. Liu, R. Zhu, X.-Z. Kong, L2, 1-extreme learning machine: an efficient robust classifier for tumor classification, Comput. Biol. Chem. 89 (2020) 107368, doi:10.1016/j.compbiolchem.2020.107368.
[41] X. Francis, S. Roberto, Comparing predictive accuracy, J. Bus. Econ. Stat. 13 (3) (1995) 253–263, doi:10.2307/1392185.