


Received 17 August 2022, accepted 28 August 2022, date of publication 14 September 2022, date of current version 3 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3206542

## RESEARCH ARTICLE

# Two Stage Prediction Model of Sunspots Monthly Value Based on CEEMDAN and Particle Swarm Optimization ELM

BEIJIA ZHANG<sup>1</sup>, LIN SUN<sup>2</sup>, AND WENBO WANG<sup>1</sup> 

<sup>1</sup>School of Automobile and Traffic Engineering, Wuhan University of Science and Technology, Wuhan 430081, China

<sup>2</sup>All-purpose Quality Education College, Wuchang University of Technology, Wuhan 430223, China

<sup>3</sup>School of Science, Wuhan University of Science and Technology, Wuhan 430081, China

Corresponding author: Wenbo Wang (wangwenbo@wust.edu.cn)


This work was supported in part by the National Natural Science Foundation of China under Grant 61671338 and Grant 51877161.

**ABSTRACT** Sunspot number is the most basic parameter to describe the level of solar activity. The accurate prediction of sunspot number can reflect the electromagnetic disturbance level of the electromagnetic layer, ionosphere and the middle and high layers in the future in advance, so as to provide important reference information for navigation, positioning, communication and the prediction of orbital attenuation of LEO satellites. Aiming at the characteristics of sunspot time series such as non-stationary, chaotic and difficult to predict, this paper proposed a two-stage combined prediction model based on complete Ensemble Empirical mode decomposition with adaptive noise (CEEMDAN), particle swarm optimization (PSO) and extreme learning machine(ELM) network. In the first stage of prediction: firstly, the original sunspot monthly mean series is decomposed by CEEMDAN to reduce the non-linearity and non-stationary of the original series. Then, an ELM prediction model is established for the sub-sequences decomposed by CEEMDAN, and the input layer dimension, hidden layer dimension, input layer weight and hidden layer bias of ELM are optimized by PSO algorithm. Finally, the prediction results of the first stage are obtained by superimposing the prediction results of each sub-sequence. The second stage is the error self-correction stage. Firstly, the prediction error sequence of the first stage is obtained. Then, the CEEMDAN-PSO-ELM prediction model is used to self-correct the prediction error of the first stage. Finally, the prediction results of the first stage and the self-correction results of the second stage are superimposed to obtain the final prediction value of the monthly sunspot number. In this paper, CEEMDAN is used to reduce the non-linearity and non-stationary of the sunspot series, and PSO is used to determine the best parameters of ELM network, and the useful information in the prediction error is fully considered, which effectively improves the prediction accuracy of sunspot monthly mean series. The prediction experiment is carried out by using the measured sunspot monthly mean series, and the proposed model is compared with wavelet neural network (WNN), back propagation neural network (BPNN), ELM, CEEMDAN-ELM and CEEMDAN-PSO-ELM. The experimental results show that the prediction accuracy of the proposed two-stage prediction model has been significantly improved, and has better prediction stability.

**INDEX TERMS** Sunspots, CCEMDAN, particle swarm optimization, extreme learning machine.

## I. INTRODUCTION

The number of sunspots is the most basic parameter describing the level of solar activity. Because the number of sunspots

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed M. A. Moustafa .

is closely related to solar flares and coronal mass ejections and other eruptive phenomena, these eruptive phenomena cause geomagnetic disturbances. Accurate prediction of sunspot numbers can reflect the state of disturbance levels in the magnetosphere, ionosphere, and middle and upper atmospheres, which can provide important reference information

for navigation, positioning, communications, and low-orbit satellite orbit attenuation predictions. Therefore, the accurate prediction of sunspot number has very important research significance.

At present, many scholars have conducted in-depth research on the prediction of the monthly average of sunspots. Among them, two physical prediction models based on similar week and engine models have achieved good prediction results [1], [2], [3]. However, it is another important research direction to regard the smoothed monthly mean value of sunspot number as a time series, and construct the time series using historical data accumulated in long-term records, and forecast by a certain prediction algorithm [4]. In recent years, nonlinear prediction models represented by neural networks have been widely used in the prediction of nonlinear systems due to their extensive adaptability and learning capabilities. The most widely used in the prediction of the monthly mean value of sunspots are the nonlinear prediction models represented by radial basis function neural networks [5], [6], echo state networks [7], and least square support vector machines [8].

However, different prediction models have their own advantages and disadvantages, so, the single prediction model is difficult to achieve the optimal prediction results [9]. In order to further improve the prediction performance, researchers have proposed the hybrid prediction model, which can integrate the advantages of different models. The hybrid prediction model mainly includes the following aspects: data preprocessing technology, such as empirical mode decomposition (CEEMDAN), variational mode decomposition (VMD) and singular spectrum analysis (SSA), intelligent optimization algorithms, such as genetic algorithm (GA) [10], particle swarm optimization (PSO) [11] and grey wolf optimizer (GWO) [12], as well as prediction models, such as extreme learning machine (ELM), back propagation neural network (BPNN) and support vector machine (SVM), etc [13].

In recent years, in order to improve the prediction performance of data, different types of hybrid prediction models have been proposed. Reference [14] used empirical mode decomposition and autoregressive models to study the long-term prediction of sunspot numbers. Reference [15] used complementary ensemble empirical mode decomposition (CEEMD) to decompose the time series of sunspot numbers, and then used wavelet neural network (WNN) to predict the decomposed components. Reference [16] combines empirical mode decomposition with long- and short-term network models in deep learning, and uses a cyclic model to predict sunspot sequences. Reference [17] proposed a hybrid wind power prediction model by using the gravity search algorithm (GSA) to optimize the hyper-parameters of least squares support vector machine (LSSVM). Reference [18] proposed a hybrid prediction model based on wavelet transform and convolutional neural network. Reference [19] first uses wavelet packet transform to decompose the original data, then uses LSSVM to predict

the decomposed data, and uses the combined optimization method based on simulated annealing and particle swarm optimization (SA-PSO) to optimize the super-parameters of LSSVM. The experimental results of the above hybrid models show that the hybrid prediction model has higher prediction accuracy than the single prediction model.

In addition, most of the research on prediction models only carried out simple error correction, ignoring the importance of prediction error, resulting in the inability to make full use of the useful information in the prediction error. However, the current research shows that considering the error factor can significantly improve the prediction performance. For example, [20] shows that the accuracy of the prediction model based on error correction is significantly better than that before error correction. Reference [21] proposed a hybrid prediction model with error correction. The prediction results show that the prediction model based on error correction has better prediction ability than other prediction models without error correction. The prediction results of [20], [22], [23], and [24] also show that error correction can improve the prediction accuracy. Based on the results of the above literatures, this paper considers the error factor in constructing the prediction model of the monthly mean series of sunspots, and further improves the prediction accuracy through multi-scale error correction.

Therefore, this paper proposes a two-stage prediction model based on multi-scale decomposition, swarm intelligence optimization algorithm and multi-scale correction of error sequence, which successfully solves the above important problems. Specifically, in the first stage of prediction, aiming at the nonlinear problem of sunspot monthly mean time series, CEEMDAN method is used to decompose the original data into a series of intrinsic mode functions (IMF), and then particle swarm optimization (PSO) ELM model (PSO-ELM) is used to predict all the intrinsic mode components. Then, in the second stage of prediction, an error correction prediction method based on multi-scale PSO-ELM is constructed to correct the prediction error in the first stage. Finally, the prediction results of all IMF in the first stage are integrated with the error prediction results in the second stage to obtain the final prediction value. Experimental results show that the proposed two-stage prediction model performs well in predicting the monthly mean of sunspots.

The main innovation and contributions of this paper are summarized as follows:

- (1) In this paper, a new two-stage prediction framework is proposed, which can better improve the prediction accuracy of sunspot monthly mean. Moreover, the effectiveness of the proposed two-stage prediction model is verified by several datasets, and the prediction results of the proposed method are compared and analyzed with other five classical prediction results..

- (2) The proposed sunspot prediction model takes into account the error factors, and successfully solves a major problem of the previous models, that is, it only improves the prediction ability of sunspot monthly mean series, without

**TABLE 1. Properties descriptive statistics of smoothed monthly mean value of sunspot number.**

Data	Numbers	Mean	Max	Min	Std	Ske w	Kurt
All samples	3236	82.24	285.0	0.0	63.16	0.77	2.81
Training	2589	80.71	285.0	0.0	62.76	0.84	3.05
Testing	647	88.38	232.9	2.2	64.43	0.48	2.02

considering the influence of error factors on the effectiveness of the prediction model. Therefore, the multi-scale PSO-ELM hybrid model is used to correct the prediction error, and then the correction results of the prediction error are integrated into the proposed two-stage prediction model.

(3) The accuracy and stability of the prediction are considered in the proposed prediction model. Specifically, PSO algorithm is used to optimize super-parameters of ELM to achieve high prediction accuracy and stability. The prediction results show that the prediction model based on PSO-ELM has high prediction accuracy and stability, which reflects the effectiveness of PSO-ELM in predicting the monthly mean value of sunspots.

## II. MATERIALS AND METHOD

### A. MONTHLY MEAN DATA OF SUNSPOTS

The monthly average data of sunspots comes from the official website of the Solar Influence Data Analysis Center (SIDC) of the Royal Observatory of Belgium (<http://sidc.oma.be/silso/datafiles>) [25]. The sunspot data on this website has been recorded since 1749, and the data was revised significantly in July 2015. This article uses the corrected monthly average of sunspots. The smoothed monthly average of the number of sunspots is the statistics of the monthly average of the number of sunspots in the current 6 months, the next 6 months and the current month. The weight of the first month of the first 6 months and the last month of the next 6 months is 0.5, and the weight of the other 11 months is 1. Divide the sum of the monthly averages of the 13-month sunspots by 12. The time series data is the smoothed monthly average of the number of sunspots from July 1949 to February 2019. Table 1 is a descriptive statistics of the smoothed monthly average of the number of sunspots.

### B. CEEMDAN

Huang *et al.* [26] proposed an EMD method that can decompose any signal into intrinsic mode functions (IMF). M. A. Colorminas *et al.* [27] proposed the Complete ensemble empirical mode decomposition with adaptive noise(CEEMDAN) method based on the research of Huang *et al.* CEEMDAN uses the zero-mean characteristic of Gaussian white noise to make the decomposition effect of signal data more complete. The specific processing process is as follows:

Step 1: Add the standard normal distribution white noise  $w^i(n)$  of different amplitudes to the given target signal  $x(n)$

and construct the signal sequence of the  $i$ -th experiment as

$$x^i(n) = x(n) + \gamma_0 w^i(n), \quad (i = 1, 2, \dots, I) \quad (1)$$

where:  $\gamma_0$  is the standard deviation of noise.

Step 2: In the first stage, the EMD method is used to decompose the target signal, and the first intrinsic modal component(IMF) is obtained and the average value is

$$C_1(n) = \frac{1}{I} \sum_{i=1}^I IMF_1^i(n) \quad (2)$$

(IMF refers to a function that satisfies the following two conditions: 1) the number of extreme points is equal to or differs from the number of zero crossings by at most one; 2) the mean of the envelope defined by the local maximum point and the envelope defined by the local minimum point is zero)

The first stage margin signal is expressed as

$$r_1(n) = x(n) - C_1(n) \quad (3)$$

Step 3: Define  $E_k(n)$  as the  $K$ -th IMF component after EMD decomposition of the signal data. Decomposing the sequence  $r_1(n) + \gamma_1 E_1(w^i(n))$ , the IMF component of the second stage can be obtained as

$$C_2(n) = \frac{1}{I} \sum_{i=1}^I E_1\{r_1(n) + \gamma_1 E_1[w^i(n)]\} \quad (4)$$

The second remaining component is

$$r_2(n) = r_1(n) - C_2(n) \quad (5)$$

Step 4: By analogy, the  $K$ -th remaining component is

$$r_k(n) = r_{k-1}(n) - C_k(n) \quad (6)$$

The  $(k + 1)$ -th IMF component is

$$C_{k+1}(n) = \frac{1}{I} \sum_{i=1}^I E_1\{r_k(n) + \gamma_k E_k[w^i(n)]\} \quad (7)$$

Step 5: Repeat Step 4 until the remaining components cannot meet the EMD decomposition conditions or the iteration ends. Finally, all  $K$  intrinsic mode functions ( $\{imf_k\}_{1 \leq k \leq K}$ ) of CEEMDAN are obtained, and the residual  $R(n)$  is

$$R(n) = x(n) - \sum_{k=1}^K C_k(n) \quad (8)$$

the target data sequence is decomposed into

$$x(n) = \sum_{k=1}^K C_k(n) + R(n) \quad (9)$$

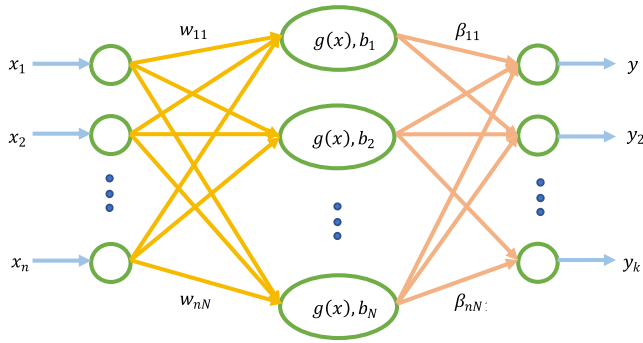


FIGURE 1. Network structure of ELM.

C. PARTICLE SWARM OPTIMIZATION EXTREM LEARNING MACHINE

1) EXTREM LEARNING MACHINE

ELM is a new algorithm for Single-hidden Layer Feed Forward Networks (SLFNs) proposed by Huang *et al.* [28]. Compared with traditional neural networks based on gradient learning, ELM has a unique advantage, that is, the algorithm only needs to set the number of hidden layer units, and then randomly generate input connection weight vectors and hidden layer biases, and calculate the optimal solution. This process avoids the traditional multiple iterations. In particular, ELM has a fast learning rate, which greatly reduces training time, thereby improving the generalization of neural networks and improving the accuracy of network operation results. The ELM network structure is shown in Figure 1.

For any  $N$  different samples  $(x_i, y_i)$ , where

$$x_i = [x_{i1}, x_{i2}, \dots, x_{iN}]^T \in R^N$$

$$y_i = [y_{i1}, y_{i2}, \dots, y_{iK}]^T \in R^K$$

the output of a feed-forward neural network with hidden nodes and an activation function  $G(x)$  is

$$y_i = f(x_i)$$

$$= \sum_{j=1}^Q \beta_j G(w_j \cdot x_i + b_j), \quad w_j \in R^N, \quad \beta_j \in R^K \quad (10)$$

where  $i = 1, 2, \dots, N, j = 1, 2, \dots, Q, w_j$  is the input weight connecting the input layer to the  $j$ -th hidden layer node,  $\beta_j = [\beta_{j1}, \beta_{j2}, \dots, \beta_{jK}]^T$  is the output weight connecting the  $j$ -th hidden layer node to the output node,  $b_j$  is the deviation value of the  $j$ -th hidden layer node,  $w_j \cdot x_i$  represents the inner product of vectors  $w_j$  and  $x_i$ , the excitation function  $G(x)$  can be selected as ‘‘Sigmoid’’, ‘‘Tansig’’, ‘‘Sine’’ or ‘‘RBF’’ and so on.

Converting Equation (10) into matrix form, we can get:

$$Y = H\beta \quad (11)$$

where  $H$  is the hidden layer output matrix of the network.

In the ELM algorithm, the input weight and hidden layer can be given randomly without adjustment during the training process, and the hidden layer matrix  $H$  is a definite matrix

before training. The training of the feed-forward neural network is actually transformed into a problem of solving the least square solution  $\hat{\beta}$  of the output weight matrix. The output weight matrix  $\hat{\beta}$  can be expressed as

$$\hat{\beta} = H^+ Y \quad (12)$$

where  $H^+$  is the generalized inverse of matrix  $H$ .

According to Equations (10) to (12), the output weight matrix is determined by the deviation between the input weight matrix and the hidden layer. Since ELM randomly assigns the initial input weight matrix and hidden layer deviation, there may be some input weight matrix and hidden layer deviation of 0, resulting in some hidden layer nodes are invalid. Therefore, in some practical applications, the accuracy and time of ELM training will be affected by randomness.

2) PSO ALGORITHM

The Particle swarm optimization (PSO) algorithm was first proposed by Kennedy and Eberhart [29] in 1995. The PSO algorithm originated from the study of the predation behavior of birds. When birds prey, the easiest and most effective way for each bird to find food is to search the area around the bird closest to the food.

Suppose that in a  $D$ -dimensional search space, there is a population of  $n$  particles  $X = (X_1, X_2, \dots, X_n)$ . The  $i$ -th particle represents a  $D$ -dimensional vector  $X_i = [X_{i1}, X_{i2}, \dots, X_{iD}]^T$ , which represents the position of the  $i$ -th particle in the  $D$ -dimensional search space, and also represents a potential solution of the problem. According to the objective function, the fitness value corresponding to each particle position  $X_i$  can be calculated. The velocity of the  $i$ -th particle is  $V_i = [V_{i1}, V_{i2}, \dots, V_{iD}]^T$  and its individual extreme value is  $P_i = [P_{i1}, P_{i2}, \dots, P_{iD}]^T$ . The global extreme minimum of the population is  $P_g = [P_{g1}, P_{g2}, \dots, P_{gD}]^T$ .

In each iteration, the particle updates its speed and position through individual extreme values and global extreme values. The update formula is as follows:

$$V_{id}^{k+1} = \omega V_{id}^{k+1} + c_1 r_1 (P_{id}^k - X_{id}^k) + c_2 r_2 (P_{gd}^k - X_{id}^k) \quad (13)$$

$$X_{id}^{k+1} = X_{id}^k + V_{id}^{k+1} \quad (14)$$

In Equations (13) and (14),  $\omega$  is the inertia weight,  $X_{id} \in X_i$  is the  $d$ -th element of the  $i$ -th particle,  $d = 1, 2, \dots, D, i = 1, 2, \dots, n, k$  is the current iteration number,  $V_{id}$  is the velocity of the particle,  $c_1$  and  $c_2$  are non-negative constants, called acceleration factors,  $r_1$  and  $r_2$  are random numbers distributed between  $[0, 1]$ . In order to prevent blind search of particles, it is generally recommended to limit its position and speed to a certain interval  $[-X_{max}, X_{max}], [-V_{max}, V_{max}]$ .

3) PSO-ELM MODEL

When using the ELM model to predict the time series, the input layer dimension and the hidden layer dimension are

artificially given, and the optimal network dimension cannot be guaranteed. The input weights and implicit deviation vectors of the ELM network are randomly assigned, which leads to the weak generalization and stability of the network, which affects the prediction accuracy of the ELM network. The PSO algorithm is an effective algorithm for parameter optimization of the ELM model. This article uses PSO to optimize the ELM parameters, the specific steps are as follows:

Step 1: Load the data and divide the data into 80% training set and 20% test set.

Step 2: Initialize the particle population and set the relevant parameters of the PSO algorithm. Individuals (particles) in the population are composed of input layer dimensions, hidden layer dimensions, input weights and hidden deviations. The particle length is  $L = 2K + Q + KQ$ , where  $Q$  is the number of hidden layer nodes and  $K$  is the number of neurons in the input layer.

Step 3: The input layer dimensions, hidden layer dimensions, input weights and implicit deviations corresponding to each particle are brought into the ELM training algorithm, namely Equations (10)~(12) to obtain the output weights and the predicted values of the matrix. The mean square error (MSE) of the training set output of the ELM network is used as the fitness of the particle swarm optimization algorithm, the fitness value of each particle is calculated, and the individual extreme value and the global extreme value are updated.

Step 4: In the iterative process, the velocity and position of the particles are updated according to Equations (13) and (14). When the maximum number of iterations or the best fitness is reached, the optimization iteration process is stopped.

Step 5: The optimal input layer dimension, hidden layer dimension, input weight and implicit deviation obtained by performing the above steps are substituted into Equation (12) to calculate the output weight matrix, and the prediction result is obtained.

### III. TWO-STAGE PREDICTION OF SUNPOT MONTHLY VALUE BASED ON CEEMDAN-PSO-ELM

The monthly mean value of sunspots is nonlinear, non-stationary and time-varying.

This paper proposes a two-stage sunspot prediction model based on the combination of CEEMDAN and ELM, and uses the PSO algorithm to optimize the input layer dimension, hidden layer dimension, input weight and implicit deviation of the ELM model. The forecast flow chart is shown in Figure 2.

The two stages prediction process based on CEEMDAN-PSO-ELM is described as follows:

Step 1: The original sequence is decomposed into  $C_1, C_2, \dots, C_K$  and  $R$  by CEEMDAN.

Step 2: Normalize  $C_1, C_2, \dots, C_K$  and  $R$  by the following formula

$$y = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (15)$$

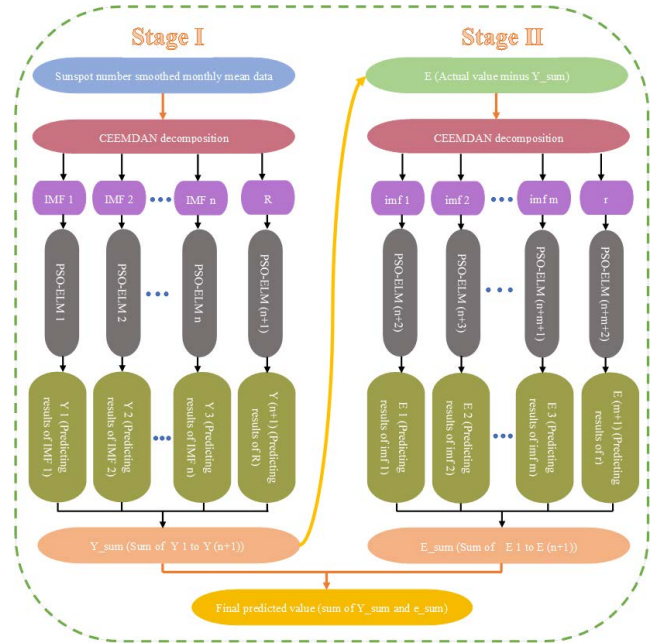


FIGURE 2. Flow of two-stage based on CEEMDAN-PSO-ELM.

where  $x$  is the input data, and  $x_{max}$  and  $x_{min}$  are the maximum and minimum of  $x$ , respectively, and  $y$  is the normalized result of the input data.

Step 3: Bring each normalized sub-sequence into the PSO-ELM model for prediction, and de-normalize the output result to get the prediction result  $Y_1, Y_2, \dots, Y_{K+1}$ .

Step 4: Sum  $Y_1, Y_2, \dots, Y_{K+1}$  to get the prediction result of the first stage  $Y_{sum}$ . Subtract  $Y_{sum}$  from the actual value to get the error sequence  $E$ .

Step 5: Use the CEEMDAN method to decompose the error sequence  $E$  to obtain  $C_1^E, C_2^E, \dots, C_M^E$  and  $R^E$ .

Step 6: Normalize  $C_1^E, C_2^E, \dots, C_M^E$  and  $R^E$ .

Step 7: Bring each normalized sub-sequence into the PSO-ELM model for prediction, and get the prediction result  $Y_1^E, Y_2^E, \dots, Y_{M+1}^E$ .

Step 8: Sum  $Y_1^E, Y_2^E, \dots, Y_{M+1}^E$  to get the second stage prediction result  $Y_{sum}^E$ .

Step 9: Sum the prediction result of the first stage  $Y_{sum}$  and the prediction result of the second stage  $Y_{sum}^E$  to get the final prediction value, that is prediction value  $\hat{Y} = Y_{sum} + Y_{sum}^E$ .

### IV. MONTHLY SUNSPOT PREDICTION EXPERIMENT

The monthly mean number of sunspots is from the official website of the solar action data analysis center of the Royal Observatory of Belgium (source: silica data, Royal Observatory of Belgium, Brussels). We selected the sunspot smoothing monthly observations from August 1949 to March 2021 as the experimental data, and the total length of the data set is 3260. In the first experiment, we selected all 3260 data as experimental data (recorded as dataset1), in which the length of the training set is 2600 and the length of the test set is 660. In the second group of experiments, we selected the

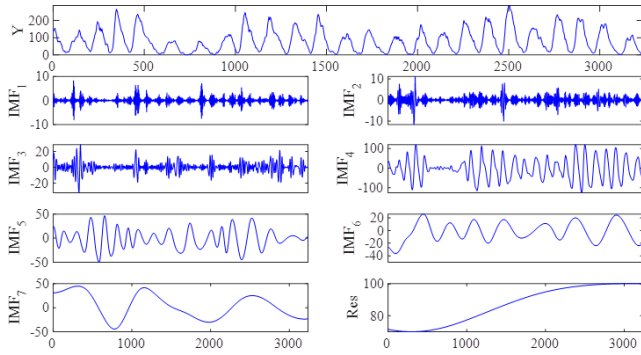


FIGURE 3. CEEMDAN decomposition result of the original sequence.

data interval of [1621, 2435], a total of 815 data (denoted as dataset2). The first 695 data of this data set were used as the training set, and the last 120 data were used as the test set.

**A. MODEL EVALUATION CRITERIA**

This paper uses the following four error indicators to measure the prediction effect of the proposed prediction model: Mean Absolute Error(MAE), Root Mean Squared Error(RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination ( $R^2$ ). The formulas are:

$$MAE = \frac{1}{n} \sum_{t=1}^n |\hat{x}(t) - x(t)| \tag{16}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n |\hat{x}(t) - x(t)|^2} \tag{17}$$

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{\hat{x}(t) - x(t)}{x(t)} \right| \tag{18}$$

$$R^2 = \frac{\sum_{t=1}^n (\hat{x}(t) - \bar{x}(t))^2}{\sum_{t=1}^n (x(t) - \bar{x}(t))^2} \tag{19}$$

In Equations (16)~(19),  $\hat{x}(t)$  represents the prediction value,  $x(t)$  represents the original data,  $\bar{x}(t)$  represents the mean of original data, and  $n$  represents the test set data length. MAE represents the error between  $\hat{x}(t)$  and  $x(t)$ . RMSE reflects the error distribution, that is, the deviation between  $\hat{x}(t)$  and  $x(t)$ . MAPE is used to measure the quality of a model’s prediction results. The smaller the MAE, RMSE and MAPE, the better the model. The larger the  $R^2$ , the better the model.

**B. PREDICTION EXPERIMENT OF DATASET1**

**1) FIRST STAGE PREDICTION OF DATASET1**

Use CEEMDAN to decompose the monthly mean time series to obtain 7 IMF ( $C_1, C_2, \dots, C_7$ ) and a residual component  $R$ . The decomposition result is shown in Figure 3. The PSO-ELM model is established for each normalized IMF and  $R$ . Set the number of PSO iterations to 300, the population

TABLE 2. Elm parameters optimization results by PSO in first stage.

Subsequence of $Y$	ELM model parameters	
	Input layer dimension	Hidden layer dimension
$C_1$	9	30
$C_2$	6	30
$C_3$	6	33
$C_4$	5	23
$C_5$	5	27
$C_6$	5	32
$C_7$	4	26
$R$	13	23

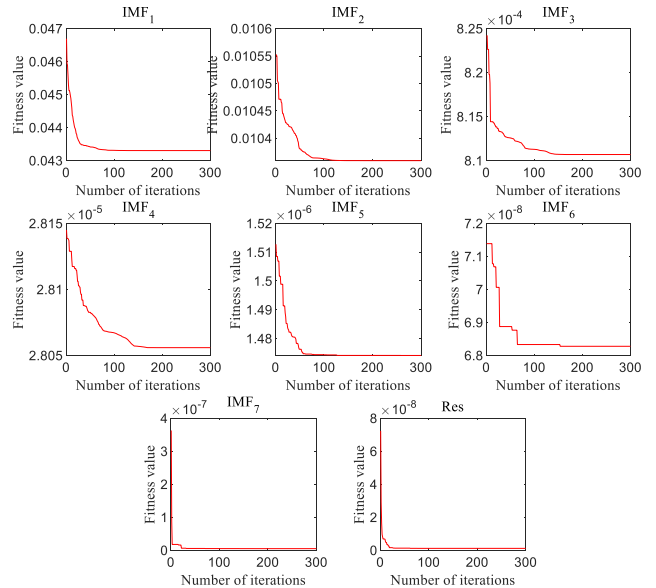


FIGURE 4. Fitness curve of first stage PSO optimization.

size to 30, the inertia factor to 0.8, and the learning rate to 2. The optimization results of the PSO algorithm for the input dimension and hidden layer dimension of the ELM model are shown in Table 2. The fitness curve of the PSO algorithm for optimizing the ELM model of each IMF and  $R$  is shown in Figure 4.

**2) SECOND STAGE PREDICTION OF DATASET1**

After the predicted value of the first stage is obtained, the error sequence  $E$  is decomposed by CEEMDAN, and the parameter settings in CEEMDAN are the same as those in the first stage. The CEEMDAN method decomposes  $E$  to obtain 11 IMFs ( $C_1^E, C_2^E, \dots, C_{11}^E$ ) and a residual component  $R^E$ . The decomposition result is shown in Figure 5. The initial parameter settings of the second stage PSO algorithm are the same as those of the first stage. The optimization results of the PSO algorithm for the input dimension and hidden layer dimension of the ELM model are shown in Table 3.

**3) PREDICTIVE EFFECTS OF MULTIPLE MODELS**

The prediction result of the first stage and the prediction result of the second stage are summed to obtain the final

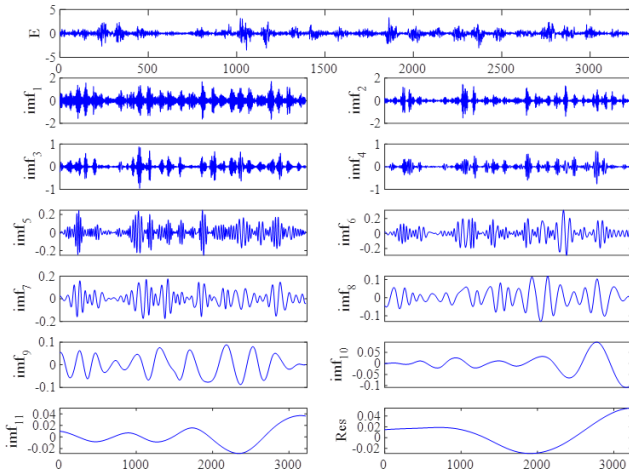


FIGURE 5. CEEMDAN decomposition result of residual sequence.

TABLE 3. Elm parameters optimization results by PSO in second stage.

Subsequence of $E$	ELM model parameters	
	Input layer dimension	Hidden layer dimension
$C_1^E$	12	22
$C_2^E$	4	24
$C_3^E$	4	16
$C_4^E$	6	31
$C_5^E$	5	21
$C_6^E$	5	30
$C_7^E$	4	17
$C_8^E$	14	17
$C_9^E$	4	27
$C_{10}^E$	4	27
$C_{11}^E$	4	25
$R^E$	4	19

predicted value. In order to verify the superiority of the proposed method, the proposed method is compared with wavelet neural network (WNN), back propagation neural network (BPNN), ELM, CEEMDAN-ELM and CEEMDAN-PSO-ELM. The prediction results of the proposed method and other methods are shown in Figure 6. It can be seen from Figure 6 that there are 9 Parts ( $Part_1, Part_2, \dots, Part_9$ ) in total in the peaks and troughs of the test set sequence. When the number of sunspots is continuously rising or falling, the fitting effects of each model are better. At the peaks and troughs, the single model's fit to the mutation point deviated significantly, while the combined model's fit to the mutation point more closely matched the original data. The prediction effects of the combined models are all good, and it is difficult to see the quality of the combined models from the fitting

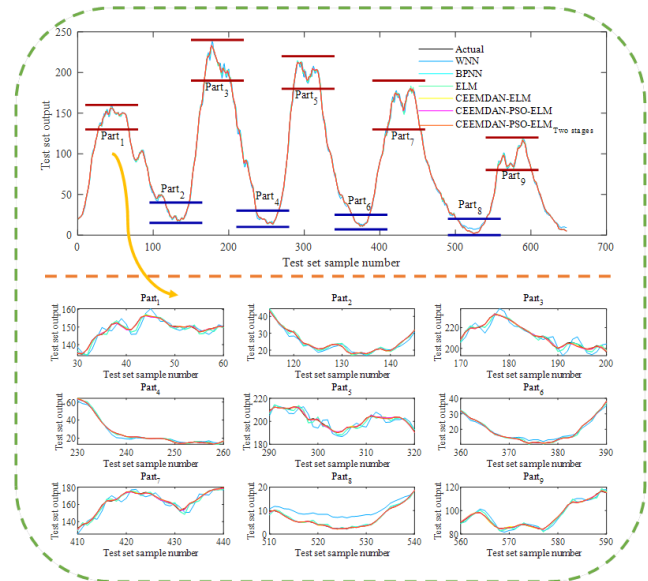


FIGURE 6. Comparison of prediction results of various models.

TABLE 4. Evaluation of prediction results of dataset 1.

Models	MAE	RMSE	MAPE	$R^2$
WNN	1.9888	2.5007	0.0595	0.9987
BPNN	1.0646	1.4357	0.0204	0.9995
ELM	1.0599	1.4254	0.0203	0.9995
CEEMDAN-ELM	0.3800	0.5247	0.0072	0.9999
CEEMDAN-PSO-ELM	0.3402	0.4962	0.0058	0.9999
Proposed method	<b>0.2141</b>	<b>0.3260</b>	<b>0.0039</b>	<b>1.0000</b>

graph. It is necessary to further calculate, analyze and compare the model evaluation indicators.

#### 4) PREDICTIVE EFFECT EVALUATION OF DATASET1

Using the aforementioned model evaluation criteria to evaluate the model. The evaluation results are shown in Table 4 and Figure 7. In terms of a single model, the ELM model is better than the WNN model and the BPNN model in predicting the monthly mean sequence of sunspots. CEEMDAN decomposition effectively improves the prediction accuracy of the ELM model. The PSO algorithm optimizes the parameters of the CEEMDAN-ELM model to further improve the prediction accuracy. The introduction of the two-stage method makes the accuracy of the model higher. The three error indicators of the model used in this article are all smaller than other forecasting models, and the coefficient of determination is higher than other models. Among them, MAE is 0.2141, RMSE is 0.3260, MAPE is 0.0039, and  $R^2$  is 1. Compared with ELM, the proposed two-stage model has decreased MAE by 79.80%, RMSE by 77.13%, and MAPE by 80.79%. Compared with CEEMDAN-ELM, the proposed two-stage model has decreased MAE by 43.66%, RMSE by 37.87%, and MAPE by 45.83%. Compared with CEEMDAN-PSO-ELM, the proposed two-stage model has decreased MAE by 37.07%, RMSE by 34.30%, and MAPE by 32.76%.

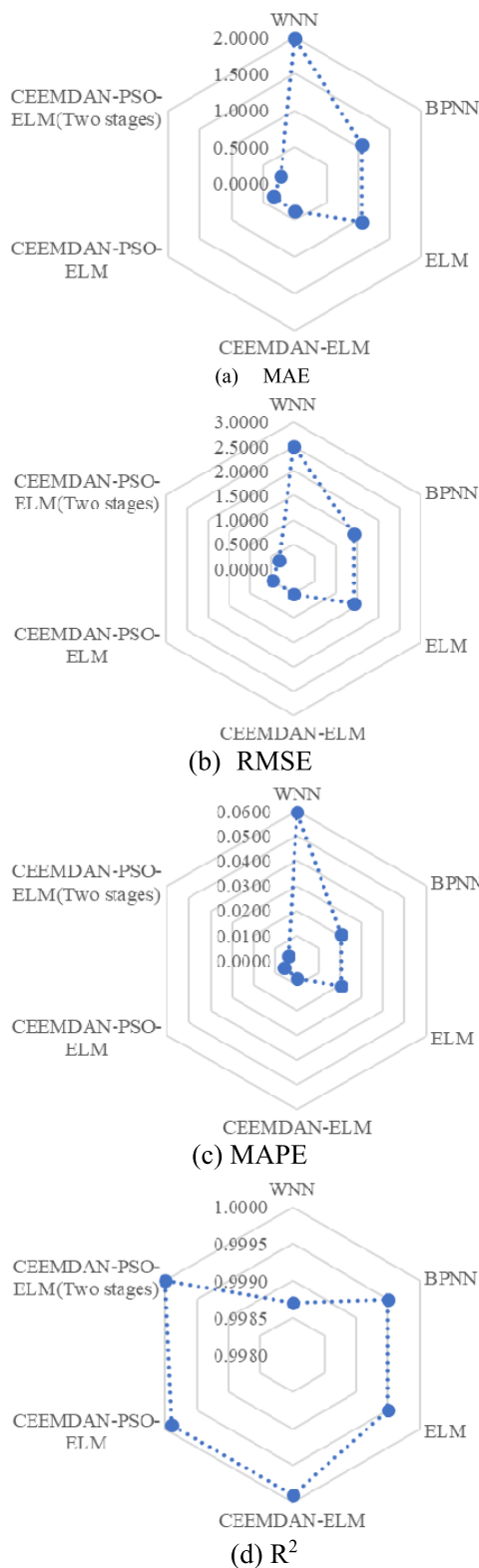


FIGURE 7. Comparison of MAE, RMSE, MAPE,  $R^2$  of each prediction model.

5) COMPARISON OF PREDICTION ERROR OF DATASET1

The error sequence of each model is obtained by subtracting the true value of the monthly mean value of sunspots from the predicted value of each model. The error sequence of

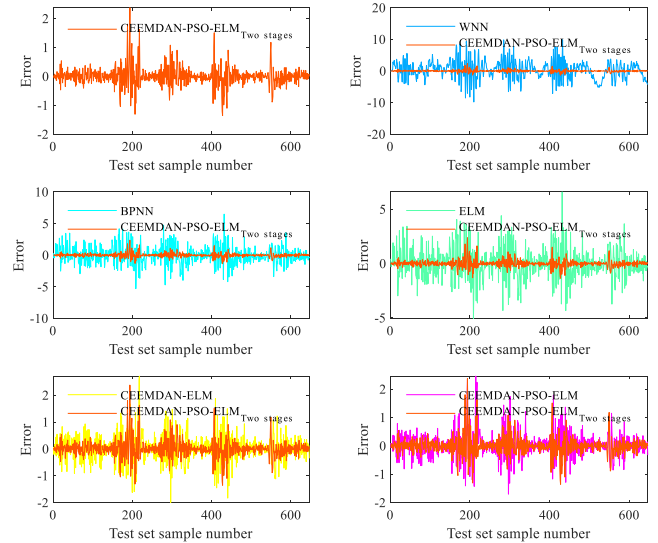


FIGURE 8. Comparison of prediction errors of various models.

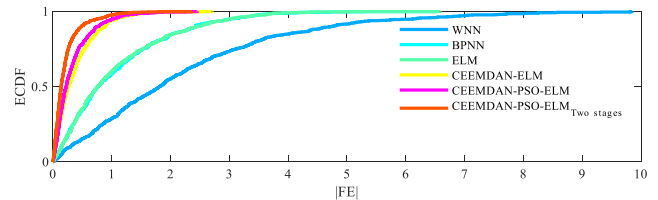


FIGURE 9. Empirical cumulative distribution diagram of the absolute prediction errors of each model.

the model proposed in this paper is compared with the error sequence of other models as shown in Figure 8. It can be seen that the model proposed in this paper has a smaller error sequence amplitude and closer to zero than other models. Taking the absolute value of the prediction error sequence to obtain the absolute prediction error sequence  $|FE|$ , the empirical cumulative distribution diagram of the absolute prediction error sequence of each model is shown in Figure 9. Combining with the descriptive statistics of the absolute prediction error sequence  $|FE|$  in Table 5, the mean value of  $|FE|$  of the model proposed in this paper is 0.2198, and the variance is 0.2594, which is smaller and more stable. The combined forecasting model adopted in this paper is more suitable for forecasting the changing trend of the monthly mean time series of sunspots. It is a feasible forecasting method and has practical significance.

C. PREDICTION EXPERIMENT OF DATASET 2

1) OPTIMIZATION OF PREDICTION OF DATASET 2

The prediction process of two-stage CEEMDA-PSO-ELM for dataset 2 is as follows:

In the first stage: firstly, the monthly mean sunspot time series of dataset 2 is decomposed by CCEEMDAN. The dataset 2 is decomposed into seven components, which are six IMFs ( $C_1, C_2, \dots, C_6$ ) and one residual component. Then,



TABLE 5. Evaluation of prediction results.

Models	FE				
	Mean	Max	Min	Median	Std
WNN	2.2276	9.8523	0.0106	1.8103	1.8473
BPNN	1.0732	6.4898	5.4476e-04	0.7675	0.9582
ELM	1.0603	6.5908	4.8841e-04	0.7515	0.9539
CEEMDAN-ELM	0.3787	2.7093	9.0642e-04	0.2650	0.3654
CEEMDAN-PSO-ELM	0.3347	2.4584	2.8828e-04	0.2238	0.3437
Proposed method	<b>0.2198</b>	<b>2.3754</b>	<b>5.7222e-04</b>	<b>0.1380</b>	<b>0.2594</b>

TABLE 6. Optimized parameters of ELM in first stage for dataset2.

Component of CCEEMDAN	Input layer dimension	Hidden layer dimension
$C_1$	17	24
$C_2$	1	47
$C_3$	17	29
$C_4$	17	33
$C_5$	16	21
$C_6$	19	26
$R$	42	29

IMFs and  $R$  are normalized, and the normalized IMFs and  $R$  are predicted by PSO-ELM model respectively. The optimization results of PSO on the input dimension and hidden layer dimension of ELM network are shown in Table 6. After optimization, an optimized ELM model is obtained for each component. Thirdly, the optimized model is used to predict each component, and the prediction results of each component are superimposed to obtain the prediction results of the first stage.

In the second stage, the error sequence is decomposed by CEEMDAN into 8 subsequences, which are 7 IMFs ( $C_1^E, C_2^E, \dots, C_7^E$ ) and 1 residual component ( $R^E$ ) respectively. IMFs and  $R$  are normalized, and the normalized IMFs and  $R$  are predicted by PSO-ELM model respectively. The optimization results of the input dimension and hidden layer dimension of ELM network are shown in Table 7.

The optimized model is used to predict each component, and the prediction results of each component are superimposed to obtain the prediction results of the second stage, as shown in Figure 10. It can be seen that the fitting effect of the error is good, especially the place with large error. In this way, the place with large error in the first stage prediction can be effectively corrected, so that the predicted value is closer to the real value.

TABLE 7. Optimized parameters of ELM in second stage for dataset2.

component of error	Input layer dimension	Hidden layer dimension
$C_1^E$	25	35
$C_2^E$	4	26
$C_3^E$	22	20
$C_4^E$	8	29
$C_5^E$	9	34
$C_6^E$	5	31
$C_7^E$	26	26
$R^E$	45	29

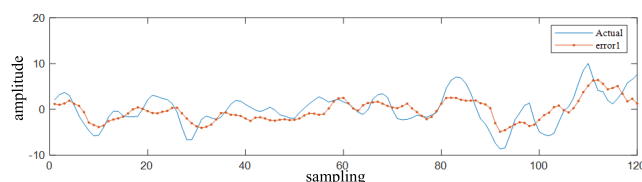


FIGURE 10. Error correction results of dataset2 in second stage.

TABLE 8. Evaluation of prediction results of dataset 2.

model	MAE	RMSE	MAPE	$R^2$
WNN	2.0590	2.6599	0.0813	0.8187
BPNN	1.2042	1.6169	0.0439	0.7370
ELM	1.1077	1.4720	0.0401	0.6700
CEEMDAN-ELM	0.4373	0.6049	0.01391	0.9698
CEEMDAN-PSO-ELM	0.4149	0.5266	0.0116	0.9839
Proposed method	<b>0.2701</b>	<b>0.3630</b>	<b>0.0065</b>	<b>0.9906</b>

## 2) PREDICTIVE EFFECT EVALUATION OF DATASET 2

In order to verify the prediction effect of the proposed model, the prediction results of the proposed model are compared with those of LSTM, LSSVM, BP, ELM, CEEMDAN-ELM and CEEMDAN-PSO-ELM models. The index comparisons of the prediction results of the six models are shown in Table 8.

It can be seen from Table 8 that CEEMDAN decomposition effectively improves the prediction accuracy of ELM, PSO further improves the prediction accuracy on the basis of CEEMDAN-ELM, and the self-correction of errors in the two-stage prediction makes the model achieve higher accuracy. The three error indexes of the proposed model are all smaller than other prediction models, and the coefficient of determination is higher than other models, MAE is 0.2701, RMSE is 0.3603, MAPE is 0.0065 and  $R^2$  is 0.9906. Compared with CEEMDAN-ELM, MAE decreased by 38.23%,

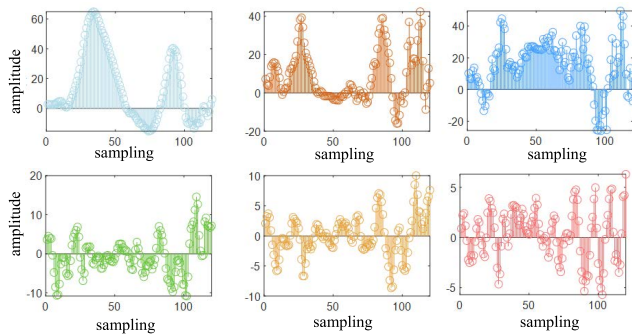


FIGURE 11. Prediction error of dataset 2 by different models.

TABLE 9. Evaluation of prediction results of dataset 2.

Models	FE				
	Mean	Max	Min	Median	Std
WNN	2.2276	9.8523	0.0106	1.8103	1.8473
BPNN	1.0732	6.4898	5.4476e-04	0.7675	0.9582
ELM	1.0603	6.5908	4.8841e-04	0.7515	0.9539
CEEMDAN-ELM	0.3787	2.7093	9.0642e-04	0.2650	0.3654
CEEMDAN-PSO-ELM	0.3347	2.4584	2.8828e-04	0.2238	0.3437
Proposed method	<b>0.2198</b>	<b>2.3754</b>	<b>5.7222e-04</b>	<b>0.1380</b>	<b>0.2594</b>

RMSE decreased by 39.99%, MAPE decreased by 53.27% and  $R^2$  increased by 2.14%. Compared with CEEMDAN-PSO-ELM, MAE decreased by 34.89%, RMSE decreased by 31.07%, MAPE decreased by 43.96% and  $R^2$  increased by 0.68%.

The error sequence of each model was obtained by subtracting the real value of monthly sunspots from the predicted value. The error of the proposed model and the error order of other models are shown in Figure 11. It can be seen that the error amplitude of the proposed model is the smallest, which is closer to 0 and closer to the positive distribution. This also indicates that it is necessary to correct the error of the first stage in the prediction of the second stage.

Comparing the absolute value of the prediction result of different models, that is, the absolute prediction error sequence  $|FE|$ , and calculating the descriptive statistical value of  $|FE|$  as shown in Table 9. It can be seen from value is closer to the real value.

Table 9 that the mean value of  $|FE|$  of the proposed model is 2.1701 and the variance is 1.4926. Compared with other models, the statistical value of the proposed method is smaller and more stable, indicating that the proposed two-stage prediction model is more suitable for predicting the change trend of the sunspots monthly mean.

## V. CONCLUSION

This article uses a two-stage modeling method. The first stage: use CEEMDAN for the smoothing of the monthly

mean sequence of sunspots, establish an ELM model for each sub-sequence after decomposition, and use the PSO algorithm to optimize the ELM parameters of each sub-model, and superimpose the prediction results of the sub-sequences. The second stage: CEEMDAN-PSO-ELM modeling is performed on the residuals obtained in the first stage, and the prediction results of the second stage are obtained. The final prediction result is obtained by summing the prediction results of the first stage and the second stage. Through simulation experiment analysis, the following conclusions are drawn:

(1) The monthly mean value model of sunspots predicted after the sequence is decomposed by CEEMDAN has higher overall prediction accuracy than the direct prediction model. Sequence decomposition can effectively reduce the impact of non-stationary features of the sequence on the prediction results.

(2) The PSO algorithm has good parameter optimization capabilities, which can effectively solve the influence of the randomness of the ELM model parameters on the model, and improve the prediction accuracy of the prediction model.

(3) The two-stage prediction method can further improve the prediction accuracy on the basis of the above model.

The experimental results show that the CEEMD-PSO-ELM sunspot forecasting model has achieved good forecast accuracy.

## REFERENCES

- [1] Z. L. Du and H. N. Wang, "The prediction method of similar cycles," *Res. Astron. Astrophys.*, vol. 11, no. 12, pp. 1482–1492, 2011.
- [2] Z. L. Du and H. N. Wang, "Predicting the solar maximum with the rising rate," *Sci. China-Phys. Mech. Astron.*, vol. 55, pp. 365–370, Feb. 2012.
- [3] J. Jiang, P. Chatterjee, and A. R. Choudhuri, "Solar activity forecast with a dynamo model," *Monthly Notices Roy. Astron. Soc.*, vol. 381, no. 4, pp. 1527–1542, Nov. 2007.
- [4] Z. Pala and R. Atici, "Forecasting sunspot time series using deep learning methods," *Sol. Phys.*, vol. 294, no. 5, pp. 1–14, May 2019.
- [5] L. Ding, Y. Jiang, and R. Lan, "Prediction of the smoothed monthly mean sunspot area using artificial neural network," in *Proc. 5th Int. Conf. Inf. Comput. Sci.*, Washington, DC, USA, Jul. 2012, pp. 33–36.
- [6] D. Zhang and Y. Han, "Time series prediction with RBF neural networks," *Inf. Technol. J.*, vol. 12, no. 14, pp. 2815–2819, Jul. 2013.
- [7] M.-H. Yusoff and Y. Jin, "Modeling neural plasticity in echo state networks for time series prediction," in *Proc. 14th UK Workshop Comput. Intell. (UKCI)*, Piscataway, NJ, USA, Sep. 2014, Art. no. 6930163.
- [8] S. Ismail, A. Shabri, and R. Samsudin, "A hybrid model of self-organizing maps (SOM) and least square support vector machine (LSSVM) for time-series forecasting," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 10574–10578, Aug. 2011.
- [9] L. Xiao, W. Shao, T. Liang, and C. Wang, "A combined model based on multiple seasonal patterns and modified firefly algorithm for electrical load forecasting," *Appl. Energy*, vol. 167, pp. 135–153, Apr. 2016.
- [10] Y. Li, B. Tang, X. Jiang, and Y. Yi, "Bearing fault feature extraction method based on GA-VMD and center frequency," *Math. Problems Eng.*, vol. 2022, Jan. 2022, Art. no. 2058258.
- [11] Y. Li, L. Mu, and P. Gao, "Particle swarm optimization fractional slope entropy: A new time series complexity indicator for bearing fault diagnosis," *Fractal Fractional*, vol. 6, no. 7, p. 345, Jun. 2022.
- [12] S. Mirjalili, S. Saremi, S. M. Mirjalili, and L. D. S. Coelho, "Multi-objective grey wolf optimizer: A novel algorithm for multi-criterion optimization," *Expert Syst. Appl.*, vol. 47, pp. 106–119, Apr. 2016.
- [13] J. Naik, S. Dash, P. K. Dash, and R. Bisoi, "Short term wind power forecasting using hybrid variational mode decomposition and multi-kernel regularized pseudo inverse neural network," *Renew. Energy*, vol. 118, pp. 180–212, Apr. 2018.

- [14] T. Xu, J. Wu, Z.-S. Wu, and Q. Li, "Long-term sunspot number prediction based on EMD analysis and AR model," *Chin. J. Astron. Astrophys.*, vol. 8, no. 3, pp. 337–342, Jun. 2008.
- [15] G. Li and S. Wang, "Sunspots time-series prediction based on complementary ensemble empirical mode decomposition and wavelet neural network," *Math. Problems Eng.*, vol. 2017, pp. 1–7, Jan. 2017.
- [16] T. Lee, "EMD and LSTM hybrid deep learning model for predicting sunspot number time series with a cyclic pattern," *Sol. Phys.*, vol. 295, no. 6, p. 82, Jun. 2020.
- [17] X. Yuan, C. Chen, Y. Yuan, Y. Huang, and Q. Tan, "Short-term wind power prediction based on LSSVM-GSA model," *Energy Convers. Manag.*, vol. 101, pp. 393–401, Sep. 2015.
- [18] H.-Z. Wang, G.-Q. Li, G.-B. Wang, J.-C. Peng, H. Jiang, and Y.-T. Liu, "Deep learning based ensemble approach for probabilistic wind power forecasting," *Appl. Energy*, vol. 188, pp. 56–70, Feb. 2017.
- [19] J.-Z. Wang, Y. Wang, and P. Jiang, "The study and application of a novel hybrid forecasting model—A case study of wind speed forecasting in China," *Appl. Energy*, vol. 143, pp. 472–488, Apr. 2015.
- [20] Z. Yu, C. Yang, Z. Zhang, and J. Jiao, "Error correction method based on data transformational GM(1,1) and application on tax forecasting," *Appl. Soft Comput.*, vol. 37, pp. 554–560, Dec. 2015.
- [21] H. Luo, D. Wang, C. Yue, Y. Liu, and H. Guo, "Research and application of a novel hybrid decomposition-ensemble learning paradigm with error correction for daily PM<sub>10</sub> forecasting," *Atmos. Res.*, vol. 201, pp. 34–45, Mar. 2018.
- [22] F. Pianosi, A. Castelletti, L. Mancusi, and E. Garofalo, "Improving flow forecasting by error correction modelling in altered catchment conditions," *Hydrol. Processes*, vol. 28, no. 4, pp. 2524–2534, Feb. 2014.
- [23] Z. Liang, J. Liang, C. Wang, X. Dong, and X. Miao, "Short-term wind power combined forecasting based on error forecast correction," *Energy Convers. Manag.*, vol. 119, pp. 215–226, Jul. 2016.
- [24] Y. Jiang and G. Huang, "Short-term wind speed prediction: Hybrid of ensemble empirical mode decomposition, feature selection and error correction," *Energy Convers. Manag.*, vol. 144, pp. 340–350, Jul. 2017.
- [25] N. R. Rigozo, M. P. Souza Echer, H. Evangelista, D. J. R. Nordemann, and E. Echer, "Prediction of sunspot number amplitude and solar cycle length for cycles 24 and 25," *J. Atmos. Solar-Terr. Phys.*, vol. 73, nos. 11–12, pp. 1294–1299, Jul. 2011.
- [26] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London A, Math. Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, 1998.
- [27] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 4144–4147.

- [28] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.
- [29] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Proc. 6th Int. Symp. Micro Mach. Hum. Sci.* Piscataway, NJ, USA: IEEE Service Center, 1995, pp. 39–43.



**BEIJIA ZHANG** was born in Wuhan, Hubei, China, in 2002. She received the bachelor's degree in traffic engineering from the Wuhan University of Science and Technology, in 2020, where she is currently pursuing the master's degree in traffic engineering.

Her research interests include signal processing technology, neural networks, and signal prediction. She received the Scholarship from the School of Science, Wuhan University of Science and Technology.



**LIN SUN** received the M.S. degree from Central China Normal University, Wuhan, China, in 2005, and the Ph.D. degree from the College of Information and Communication, National University of Defense Technology, Wuhan, in 2012.

Currently, she is an Associate Professor with the Wuchang University of Technology. She has published more than 40 papers in various academic journals and conference proceedings. Her current research interests include mathematical algorithms, modeling and simulation, and big data analysis.



**WENBO WANG** received the M.S. degree in applying mathematics from Wuhan University, Wuhan, in 2003, where he received the Ph.D. degree from the School of Mathematics and Statistics, in 2006. From 2007 to 2014, he was a Research Assistant at the College of Science, Wuhan University of Science and Technology. Since 2015, he has been a Professor with the College of Science, Wuhan University of Science and Technology. His research interests include signal processing, wavelet analysis theory, and artificial neural networks.

• • •