Short-Term Load Forecasting and Temperature Load Extraction Based on CEEMDAN and TDIC

Min Wang, Peng Zhang*, Chao Wu, Zongyin Fan, Zixuan Yu, Yuan Chen
College of Energy and Electrical Engineering, Hohai University, Nanjing, China
**Corresponding Author.*

Abstract

With the intensification of urbanization in various countries worldwide, the temperature load which is greatly affected by ambient temperature, such as summer cooling loads and winter heating loads, accounts for a rising proportion of the total urban load. It causes an increasing peak-to-valley load difference. However, due to the complex composition and strong randomness of the load, it is necessary to study the multi-scale and multi-period correlation between temperature. Based on this, the complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) is used to decompose the temperature and load into multi-scale components. The time-dependent intrinsic correlation (TDIC) is proposed to analyze the local correlation between temperature and load in multiple periods under a multi-scale framework, and obtain the dynamic change characteristics of the correlation between temperature and load. Based on the TDIC analysis results, a suitable sample period for short-term load forecasting (STLF) and input temperature data can be selected. Finally, extreme learning machine optimized by particle swarm optimization (PSO-ELM) is used to forecast each component of the load. The proposed STLF method is validated on real-time data from the Pennsylvania-New Jersey-Maryland (PJM) Company in the United States. The proposed method has greatly reduced in both mean absolute percentage error (MAPE) and root mean square error (RMSE) compared with other traditional methods, and the temperature load that fluctuates with temperature in the day to be forecasted is extracted.

Keywords: *Complete ensemble empirical mode decomposition with adaptive noise, temperature load extraction, particle swarm optimization, extreme learning machine, short-term load forecasting.*

I. Introduction

Accurate load forecasting is the key to promote power system planning and dispatching, reduce the power system operation costs, and ensure the orderly and efficient development of competitive power markets **Error! Reference source not found.**, **Error! Reference source not found.**. Load forecasting can be divided into long-term load forecasting (LTLF), medium-term load forecasting (MTLF), short-term load forecasting (STLF), and very short-term load forecasting (VSTLF) according to the time scale and different power generation and operation plans of the power system **Error! Reference source not found.**.

At present, the commonly used load forecasting models can be divided into two types. One of which is the traditional forecasting models based on statistics, such as multiple linear regression (MLR), time series models [4], grey model (GM) [5], kalman filter (KF) [6], etc. However, traditional forecasting models are only suitable for processing linear data, and large errors are often generated when forecasting non-linear data. The traditional models largely depend on the stability of historical data. In STLF, the load has strong non-linear and volatility. Therefore, in recent years, the second type of load forecasting models based on artificial intelligence (AI) has become a research hotspot, such as artificial neural network (ANN) [7], fuzzy model [8], and support vector regression (SVR) [9]. ANN has been widely used in many fields with its good learning ability. However, some inherent shortcomings of traditional ANN, such as back propagation neural network (BPNN), has become the main bottleneck restricting their development. Huang et al. [10]proposed a new type of single-hidden layer feed-forward neural network (SLFNN) in 2006 and named it extreme learning machine (ELM). ELM randomly generates the weights between the input layer and the hidden layer and the threshold of the hidden layer neurons. Compared with the traditional ANN and SVR, it has the advantages of strong generalization ability and fast calculation speed [11]. However, randomly assigned input weights and hidden layer

ISSN: 0010-8189

thresholds may affect the generalization performance of ELM. Therefore, Cai *et al.* [12] and Zeng *et al*. [13] utilized particle swarm optimization (PSO) to optimize the weights and biases of ELM network nodes, and compared PSO optimized ELM with other models like GM, SVR, ANN, and unoptimized ELM, and they concluded that PSO-ELM could effectively improve the accuracy of load forecasting.

As the load variation is affected by external factors, we should pay more attention to the correlation characteristics between external factors and load characteristic statistical indicators [14]. With the development of the economy and the improvement of living standards, the temperature load that is greatly affected by ambient temperature, such as summer cooling loads and winter heating loads, account for a rising proportion of the total urban load, which leads to more obvious characteristics of the load variation with the temperature. In the analysis of traditional correlation relations, correlation coefficient method, mutual information method, and Copula function method are often used [15]. These methods can only analyze the static correlation between temperature and load, and can only get the total correlation within a certain period, such as a year, a month, and a week. Dynamic correlation like time-variation Copula proved to be an effective method to study the dynamic correlation that changes with time. However, the current dynamic correlation research methods only analyze the correlation between the original time series, ignoring that the difference in correlation may come from the components of different time scales within a time series. In recent years, global temperature changes have become increasingly complex, resulting in an increasing proportion of temperature loads in the load. At the same time, the load composition of an area is often very complicated, and not every type of load is affected by temperature, which makes it difficult to obtain satisfactory results only by improving a single load forecasting algorithm. Moreover, accurately extracting the temperature load has also become the key to load forecasting.

Therefore, many scholars propose to preprocess the load data series before load forecasting [16]. Recent studies have shown that a framework called “decomposition and ensemble” can improve the forecasting performance [17]. For example, in [18], they propose to decompose and denoise original load data by using Wavelet Transform (WT), and establish different forecasting models for different load components after the decomposition. However, the disadvantage of WT is that the basic function and time scale need to be given in advance according to experience, and the decomposition effect of non-linear time series is not good, so it is difficult to explore the potential characteristics and components of load fully.

Empirical Mode Decomposition (EMD) is a new adaptive signal time-frequency processing method proposed by N. E. Huang in 1998 [19]. Compared with traditional signal processing methods, EMD has strong self-adaptability, completeness, and time-invariance. It is mainly used to process non-linear and non-stationary signal data, especially suitable for signals with strong randomness, high volatility, and low stability. In 2009, Huang proposed Ensemble Empirical Mode Decomposition (EEMD) [20] to solve the mode mixing phenomenon existing in EMD. At present, EMD and EEMD have been widely used in forecasting electricity prices [21], gas consumption [22], and new energy output [23]. In the field of STLF, M. R. Haq et al. [24] used EMD to decompose the load series into multiple low-frequency components and used T-Copula to analyze the upper tail correlation between the load and four types of meteorological variables to improve the load forecasting accuracy of peak load period. Q. Liu et al. [25] proposed a fuzzy weight combination theory from the perspective of data preprocessing to improve the accuracy of similar day selection, so as to reduce the original data to be processed by EMD. Then, KF-BA-SVM was used to forecast each load component after EMD decomposition. X. GAO et al. [26] used the Pearson coefficient to find IMF with high correlation with the original load series after the EMD decomposition of the load, and used it together with the original load series as the input of the forecasting model. D. Deng et al. [27] used EEMD to decompose the load, combined each mode into high-frequency and low-frequency parts according to IMFs frequency, and used different forecasting algorithms to forecast the two parts of the load. Most studies have used EMD and EEMD to decompose the load data series before load forecasting, and they have obtained better forecasting results.

In recent years, some researchers have found that the EEMD method has problems with low decomposition

efficiency, and it is difficult to completely eliminate noise. Therefore, Torres et al. [28] proposed Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) to solve these problems. At present, CEEMDAN has been gradually applied in the field of short-term load forecasting. Zhang et al. [29] proposed a combined load forecasting model based on CEEMDAN, quantum dragonfly algorithm (QDA), and SVR. Mohanad et al. [30], adopted an improved CEEMDAN based load forecasting model to reduce the non-linear and volatility of the load. By comparing with the deep learning (DL) algorithm, Li et al. [17] proved that CEEMDAN could improve the forecasting accuracy of the Multiple Kernel Extreme Learning Machine (MKELM) and reduce the training time of intelligent forecasting algorithm. Li et al. [31] and combined EMD, EEMD, and CEEMDAN with artificial intelligence forecasting algorithms. Through STLF and wind speed forecasting experiments, they both proved that CEEMDAN is better than EMD and EEMD in improving forecasting accuracy. The above studies prove that CEEMDAN has advantages in load decomposition and forecasting compared with EMD and EEMD. And compared to using the DL algorithms, such as deep belief networks (DBN) and long short-term memory (LSTM), SVR or ELM combined with CEEMDAN can achieve better performance.

In conclusion, most studies have shown that decomposing the load before load forecasting could improve the accuracy of load forecasting and found that components with different frequency characteristics will appear after the load is decomposed. However, in the current STLF research, only the load data to be forecasted is decomposed, and the load components are only combined according to the frequency. The current research ignores the randomness and volatility of the influencing factors and ignores the local feature information of different time scales contained in influencing factors. Meanwhile, they only pursue the improvement of load forecasting accuracy and fail to accurately classify and extract components with different characteristics such as temperature load. Therefore, it is impossible to extract more implicit information from the data, and it is difficult to obtain more meaningful conclusions. In summer, the load is greatly affected by temperature, and the temperature fluctuates greatly, and the frequency is uncertain. Therefore, the decomposition of temperature is also essential.

Because the load is composed of multiple scales, direct correlation analysis can produce unstable results. Simultaneously, the load tends to have dynamic changes in different periods. Therefore, it is inappropriate to use static correlation analysis to analyze the correlation between temperature and load. Suppose a traditional sliding window-based dynamic correlation analysis method is used. In that case, because the sliding window size cannot be accurately determined, the local relevant information is difficult to obtain [15]. Chen et al. proposed a time-dependent intrinsic correlation (TDIC) analysis method. The application of TDIC is based on EMD, which can adaptively adjust the size of the sliding window according to the instantaneous frequency of the time series [32]. TDIC method has been applied in many fields. Wang et al. [15] use CEEMDAN to decompose wind power and photovoltaics, then use TDIC to analyze the correlation between wind power and photovoltaics in different scales, and obtains the dynamic characteristics that vary with time. Through TDIC analysis, Huang et al. [33] obtained the inherent correlation between temperature and dissolved oxygen in the ocean at different frequency components, and concluded that there is a strong negative correlation in the components of the mean year in 3 years and 1 year. Peng et al. [34] used CEEMDAN and TDIC to analyze the multi-scale and time-varying correlations between crude oil and the US dollar in the short, medium, and long term.

This paper proposes a research method of temperature load extraction and STLF based on CEEMDAN, TDIC, and PSO-ELM based on the complex change characteristics of load and temperature. The main contributions of this paper are as follows:

Firstly, we propose to use CEEMDAN to decompose the temperature data and load data throughout the year, to reduce the nonlinearity of the data and separate the data with different time scale characteristics.

Secondly, we propose to use the TDIC-based correlation analysis method to analyze the dynamic and local correlation between temperature and load under different time scales. At the same time, the TDIC method can also determine the required sample data period in the STLF model, because too much data used for forecasting model

training will result in the addition of useless factors, and too little will result in insufficient model training. After that, CEEMDAN is used again to decompose the temperature and load data in the sample period and analyze the local characteristics of temperature and load. The components in the temperature load that fluctuates with temperature in the day be forecasted can be determined.

Finally, we use the PSO-ELM model to forecast each load component. And according to the TDIC analysis results, for some load IMFs, the temperature IMFs with the highest correlation are used as the model input temperature. The case study proves that the forecasting method based on CEEMDAN decomposition, TDIC correlation analysis, and PSO-ELM proposed in this paper can effectively improve load forecasting accuracy and extract the temperature load.

This paper is organized as follows. Section II presents the model of CEEMDAN and TDIC. Section III presents the model of forecasting algorithms based on PSO and ELM. In section IV, the load and temperature are decomposed by CEEMDAN, and the multi-scale dynamic correlation between temperature and load is analyzed by TDIC. Section V gives the temperature load extraction results, load forecasting results and corresponding analysis. The conclusion is in section VI.

II. Load Decomposition and Characteristic Analysis Model

2.1 Basic principles of EMD

EMD can decompose the signal into multiple intrinsic mode functions (IMFs) and a residual function. The criteria to generate IMFs are as follows [35]: (i) in the whole data time series, the maximum difference between the number of extremum points and the number of zero points is 1; (ii) At any point, the mean of the maximum envelope and the minimum envelope is equal to 0. The specific steps of decomposing the signal time series with EMD are as follows:

Step 1: Determine the local maximum point and the local minimum point of the original signal series $x(t)$.

Step 2: Connect all local maxima and local minima to form an upper envelope and a lower envelope, and then calculate the average envelope m_1.

Step 3: Determine whether the $h_1(t)=x(t)-m_1$ meets the IMF criteria. If the criteria are met, then $c_1(t)=h_1(t)$ is the first IMF; If the criteria are not met, then h_1 is regarded as the new original series, and the above steps are repeated k times to make $h_{1k}(t)$ meet the criteria of IMF. $c_1(t)=h_{1k}(t)$ is denoted as the first IMF of the original signal series.

Step 4: The first IMF is separated from the original series, and the remainder $r_1(t)$ is treated as the original series.

Step 5: Repeat steps 1-4 to get n IMFs. When the residual term $r_n(t)$ satisfies the termination condition and cannot extract any IMF, the decomposition process ends.

Finally, the signal series can be expressed as the sum of the IMFs and the final residual function, i.e.,

$$x(t)=\sum_{j=1}^{N} c_j(t)+r_n(t) \tag{1}$$

2.2 CEEMDAN

EEMD solves the mode mixing problem of EMD by adding Gaussian white noise and averaging it. However, due to the large white noise residual, the number of sieving increases, and the decomposition may fail, so the calculation efficiency is not high. Given the above problems, Torres et al. [28] proposed the CEEMDAN method, which adds adaptive white noise to the original signal based on EEMD, to overcome the shortcomings of EMD and

EEMD. The detailed steps are as follows:

Step 1: Suppose that the reconstructed signal after adding white noise to the original signal is represented as $x(i) = x + \beta_0 n(i)$, where β_0 is the signal-to-noise ratio of the noise relative to the original signal, $n(i)$ ($i = 1, 2, ..., L$) is the Gaussian white noise, i is the number of times that carries out EMD decomposition. $E_j(\square)$ ($j = 1, 2, ..., N$) is defined as the process of obtaining the j IMF through EMD decomposition, then the first IMF expression is:

$$\overline{c_1} = \frac{1}{L}\sum_{i=1}^{L} E_1(x(i)) = \frac{1}{L}\sum_{i=1}^{L} c_1(i) \tag{2}$$

The residual function is expressed as:

$$r_1 = x - \overline{c_1} \tag{3}$$

Step 2: When calculating the second IMF, the signal to be decomposed is $r_1 + \beta_1 E_1(n(i))$, then the second IMF is:

$$\overline{c_2} = \frac{1}{L}\sum_{i=1}^{L} E_1(r_1 + \beta_1 E_1(n(i))) \tag{4}$$

The residual function is expressed as:

$$r_2 = r_1 - \overline{c_2} \tag{5}$$

Step 3: Repeat step 2 to find out the number $j+1$ IMF is:

$$\overline{\overline{c_{j+1}}} = \frac{1}{L}\sum_{i=1}^{L} E_1(r_j + \beta_j E_j(n(i))) \tag{6}$$

Then the original signal series can be expressed as the sum of the IMFs and the final residual function, i.e.:

$$x = \sum_{j=1}^{N} \overline{c_j} + r \tag{7}$$

2.3 TDIC

The TDIC method is used to analyze the correlation of non-linear data in different time scales, and adjust the time window through adaptive criteria. Based on the decomposition of the CEEMDAN method, the TDIC calculation steps are as follows:

Step 1: Use CEEMDAN to decompose two time series $x_1(t)$ and $x_2(t)$ into the same number of IMF:

$$x_p(t) = \sum_{j=1}^{N} c_{pj}(t) + r_p(t),\ p = 1, 2 \tag{8}$$

where the $c_{pj}(t)$ is j th IMF of $x_p(t)$ and $r_p(t)$ is the residual function.

Step 2: Use the HHT method to find the instantaneous period $T_{pj}(t)$ of $c_{pj}(t)$. The sliding window centered at any time t_k is expressed as:

$$t_w^n = [t_k - nT_d / 2, t_k + nT_d / 2] \tag{9}$$

where n is any positive real number [32] , and normally n is selected as 1. The smallest sliding window suitable for calculation is $T_d = \max(T_{1j}(t_k), T_{2j}(t_k))$. The sliding window is chosen in this way to ensure that at least one complete cycle is included in the correlation calculation of the two time series.

Step 3: The TDIC correlation of each IMF pair is defined as follows:

$$R_j(t_k^n) = Corr(c_{1j}(t_w^n), c_{2j}(t_w^n)) \tag{10}$$

where *Corr* represents the overall correlation coefficient of two time series.

III. Load Forecasting Model

3.1 ELM

The ELM model structure includes an input layer, a hidden layer, and an output layer. The algorithm randomly generates the connection weight between the input layer and the hidden layer and the threshold value of the hidden layer neurons, and there is no need to adjust during the training process. By solving the generalized inverse matrix, the output weight matrix is obtained, and we can complete the training. The principle is as follows:

Suppose there are N training samples, which is (x_i, t_i) (among them, the input sample is $x_i = [x_{i1}, x_{i2}, \cdots, x_{in}]^T$ and the output sample is $t_i = [t_{i1}, t_{i2}, \cdots, t_{im}]^T$), there are L neurons in the hidden layer, and the excitation function is $g(x)$. The output of the model is:

$$o_j = \begin{bmatrix} o_{1j} \\ o_{2j} \\ \vdots \\ o_{mj} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{L} \beta_{i1} g(w_i x_j + b_i) \\ \sum_{i=1}^{L} \beta_{i2} g(w_i x_j + b_i) \\ \vdots \\ \sum_{i=1}^{L} \beta_{im} g(w_i x_j + b_i) \end{bmatrix} (j = 1, 2, \cdots, N) \tag{11}$$

Among them, $w_i = [w_{i1}, w_{i2}, \cdots, w_{in}]$ represents the connection weight between the i th neuron in the hidden layer and each input layer neuron, $\beta_i = [\beta_{i1}, \beta_{i2}, \cdots, \beta_{im}]$ represents the connection weight between the i th neuron in the hidden layer and each output layer neuron, and b_i represents the threshold of i th neuron.

To minimize the output error of the model, it needs to meet:

$$\sum_{j=1}^{N} \| o_j - t_j \| = 0 \tag{12}$$

Then, the formula (11) can transfer to the formula (13).

$$H\beta = T \tag{13}$$

$$H(w_1, \cdots, w_L, b_1, \cdots, b_L, x_1, \cdots, x_N) = \begin{bmatrix} g(w_1 \times x_1 + b_1) & \cdots & g(w_L \times x_1 + b_L) \\ g(w_1 \times x_2 + b_1) & \cdots & g(w_L \times x_2 + b_L) \\ \vdots & & \\ g(w_1 \times x_N + b_1) & \cdots & g(w_L \times x_N + b_L) \end{bmatrix}_{N' L} \tag{14}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_2^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L' M} \quad and \; T = \begin{bmatrix} t_1^T \\ t_2^T \\ \vdots \\ t_N^T \end{bmatrix}_{N' M} \tag{15}$$

Therefore, the training process of the entire neural network is equivalent to finding the least squares solution of the linear system $H\beta = T$.

$$\left\| H\hat{\beta} - T' \right\| = \min_{\beta} \left\| H\beta - T' \right\| \tag{16}$$

$$\hat{\beta} = H^{+}T' \tag{17}$$

where, H^{+} is the Moore–Penrose generalized inverse of the hidden layer output matrix H [12].

3.2 PSO

To improve ELM forecasting accuracy, PSO is proposed to optimize the input weights and hidden layer thresholds in ELM. The optimal parameters are selected by iterative calculation of optimal fitness value in multi-dimensional space. The optimal fitness value here is the minimum value of two error indicators shown in the next subsection.

The particle position update formula of PSO is:

$$v_i^d = \zeta v_i^d + c_1 r_1 (p_i^d - x_i^d) + c_2 r_2 (p_g^d - x_i^d) \tag{18}$$

$$x_i^d = x_i^d + \delta v_i^d \tag{19}$$

where, ξ is the inertia factor, c_1 and c_2 are acceleration constants, r_1 and r_2 are random numbers changing within [0,1]. v_i^d x_i^d , p_i^d , and p_g^d are respectively the velocity, position, optimal individual position, and optimal swarm position of the i th particle in the d dimension. δ is the constraint factor used to control the weight of velocity.

3.3 Forecasting Error Evaluation Indicator

To evaluate the accuracy of the forecasting models, two error evaluation indicators, namely mean absolute percentage error (MAPE) and root mean square error (RMSE) [24] is selected, with the following formula:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \tag{20}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{21}$$

where, n is the number of simulated points, y_i is the true value of the i th simulation point, $\hat{y}_i$ is the simulation value of the i th simulation point.

IV. Load Decomposition and Correction Analysis Case Based on CEEMDAN and TDIC

All the calculation and simulation in this paper are implemented by MATLAB R2020b. The computer used in this work has an Intel Core i5-8265U processor, CPU @ maximum 4.9 GHz. The hourly temperature and hourly load data used in this paper are from the PEPCO area of Pennsylvania-New Jersey-Maryland (PJM) Company in the United States [36, 37].

4.1 Decomposition case of CEEMDAN

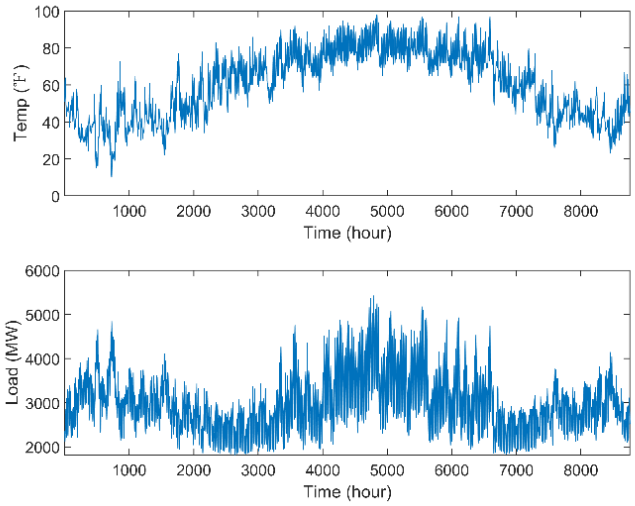
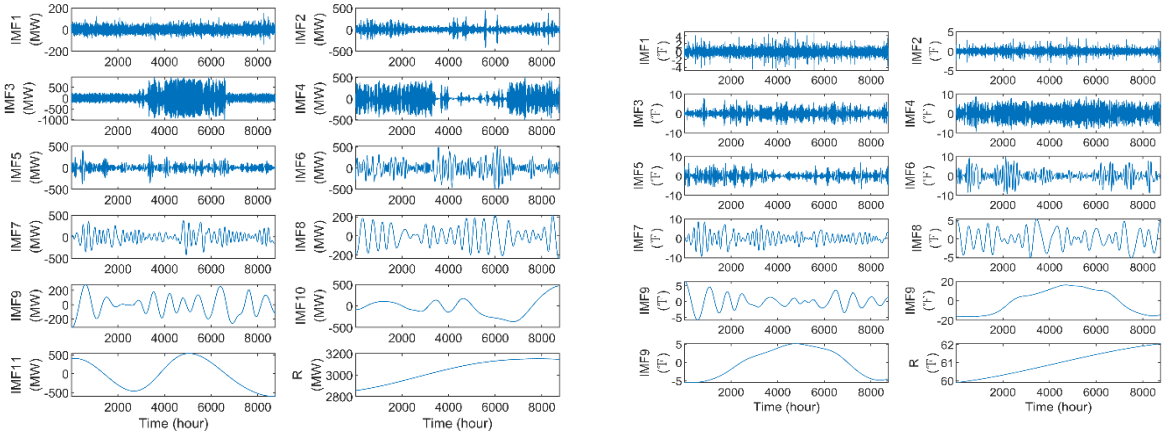


Fig 1: Annual temperature and load data.



(a) Temperature. **(b) Load.**

Fig 2. Annual temperature (a) and load (b) decomposition results using CEEMDAN.

4.2 TDIC analysis

Here, we first calculate the average cycles of each component after decomposition of temperature and load, and the results are shown in Table 1.

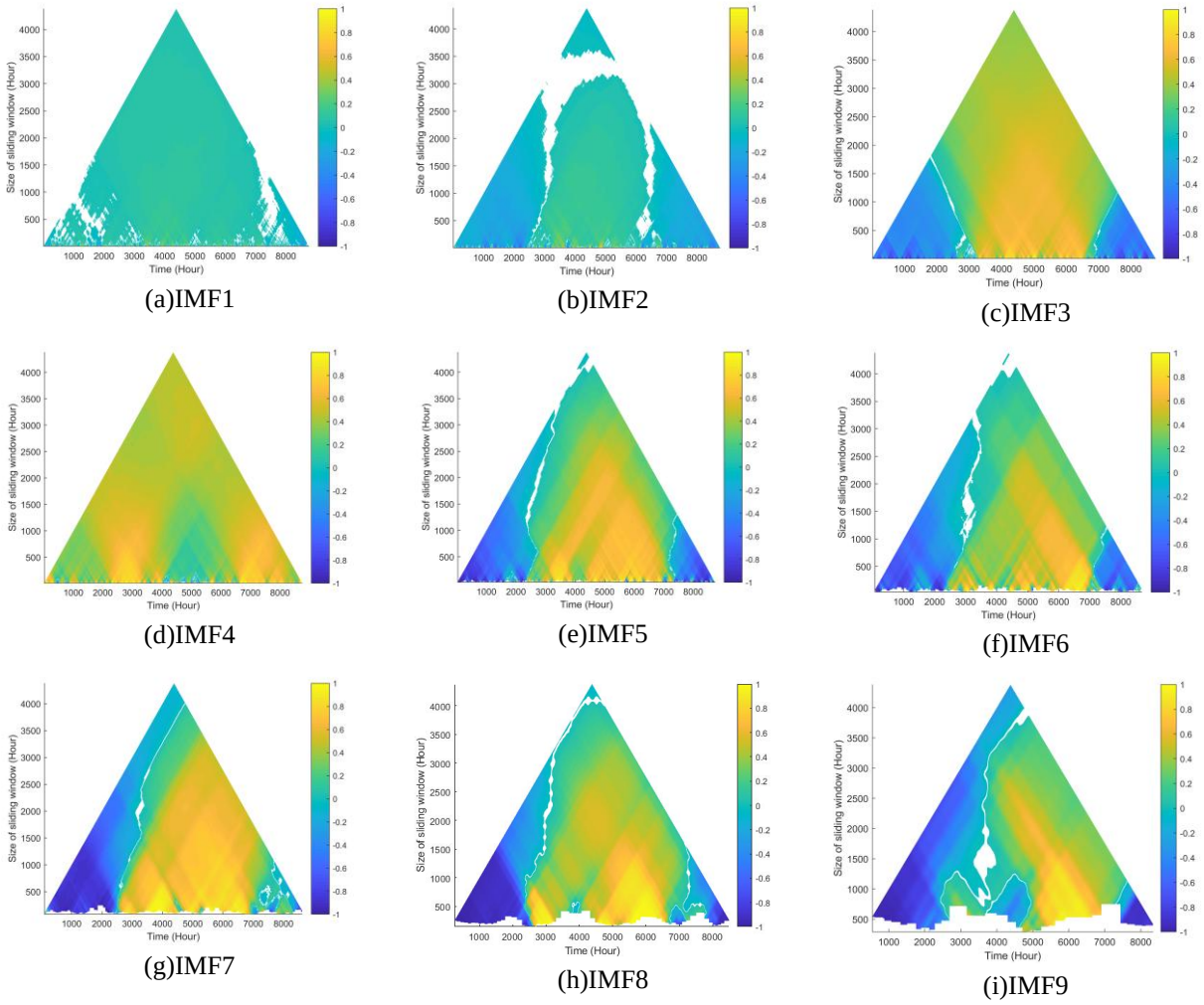
Table 1 Average cycles of temperature and load components.

	Average cycles (hour)	
	Temperature	Load
IMF1	3.37	3.17
IMF2	6.47	5.31
IMF3	14.61	12.42
IMF4	26.22	23.95
IMF5	45.69	46.39
IMF6	105.96	93.52
IMF7	197.95	201.44
IMF8	377.34	414.85
IMF9	792.90	792.90
IMF10	2488.57	5607.33
IMF11	5756	8077
Residual		

ISSN: 0010-8189

It can be seen from Table 1 that in the temperature and load data pairings of different time scales, except for IMF10, IMF11, and the residual, the average cycles of the rest of the IMFs are relatively close. This shows that temperature and load may have some correlation characteristics in these time scales. The cycles of IMF1-3 are all within one day, the cycle of IMF4 is about one day, the cycles of IMF5 and IMF6 are within one week, the cycles of IMF7 and IMF8 are about one week and two weeks, and the cycle of IMF9 is about one month.

Based on this, we use TDIC to analyze the dynamic correlation between temperature and load of IMF1 to IMF9, and the results are shown in Figure 3. Because the average cycles of temperature and load of IMF10, IMF11, and the residual are too large, TDIC analysis has no statistical significance.



(a)IMF1 (b)IMF2 (c)IMF3

(d)IMF4 (e)IMF5 (f)IMF6

(g)IMF7 (h)IMF8 (i)IMF9

Fig 3: TDIC plots between temperature and load of each IMFs.

In every plot of Figure 3, the horizontal axis represents the time of the signal and corresponds to the center position of the sliding window. The vertical axis represents the size of the sliding window. When the boundary of the sliding window exceeds the left or right end of the time domain, the TDIC correlation will not be calculated **Error! Reference source not found.**. Therefore, the TDIC plot is triangular. The sliding window size corresponding to the vertice of a plot is the entire time domain, and its value is the overall correlation coefficient of the two time series. The yellow part in the plot represents a positive correlation, and the blue part represents a negative correlation. The strength of the correlation can be seen by the color depth.

The TDIC plots in Figure 3 clearly shows the dynamic correlation between temperature and load time series on different time scales and different time windows. In terms of IMF1 and IMF2, there is no strong positive or negative correlation in any period within a year. This is because the frequency of these two components is

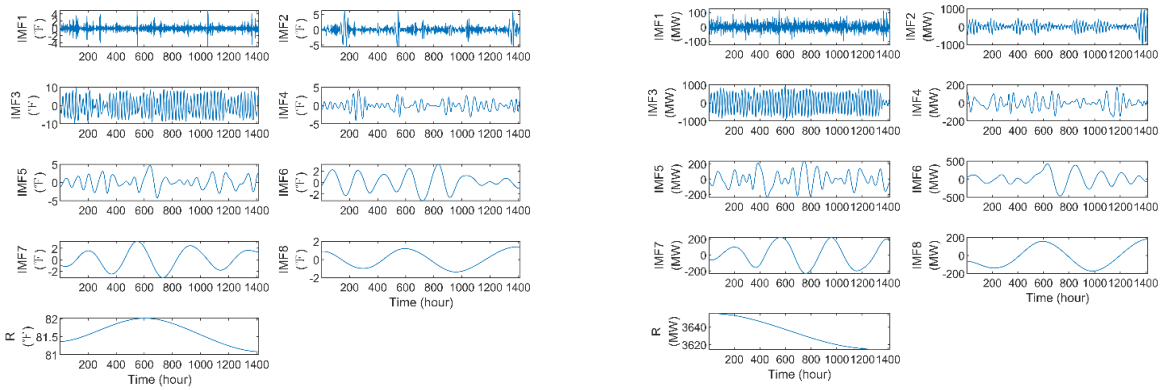
relatively high, the fluctuations are relatively severe, and they are greatly affected by random factors. The correlation of IMF4 is not obvious in the summer period, but shows a positive correlation in the spring and autumn, and shows a trend of positive correlation as the time window enlarging. IMF9 only shows a positive correlation in autumn and a negative correlation in winter. The rest of the IMFs are positively correlated in the summer-centered period, and have a negative correlation in the winter-centered period, and the correlation is strong, which shows strong seasonal characteristics. Besides, most intermediate frequency IMFs exhibit a dynamic alternate variation between positive and negative correlation in a short time window, showing strong dynamic characteristics.

It can be seen that in a period of one month or less, the correlation between temperature and load will show a trend of dynamic variation, and the correlation between temperature and load components of different scales is quite different.

4.3 Summer sample data selection and analysis

According to the analysis in section IV-B, temperature and load have a strong correlation in multiple time scales in summer. At the same time, summer is also the season with the highest load of the year. Since this paper will perform STLF during the summer high-temperature period in the next section, we will select the training sample data used by the forecasting model. The temperature and load data period used for the forecasting model should be selected to ensure that they have a strong correlation in more time scales, and there is not much correlation fluctuation in the period. Therefore, we choose the hourly temperature and hourly load data in the PEPCO area of Pennsylvania-New Jersey-Maryland (PJM) Company in the United States from 01:00 25 June 2019, to 24:00 22 August 2019, as the forecasting sample. This period is just in the green triangle area of the TDIC plot of IMF4 in Figure 3 and outside the yellow area of the TDIC plot of IMF9. It is also in the yellow triangle area of the more components. So this period can be used for better forecasting research. At the same time, this period is also the period when the highest load peak occurs in the year, and the load is greatly affected by temperature.

Because the reduction of the data scale used for decomposition will inevitably affect the number of extreme points of the data, and then affect the result of CEEMDAN decomposition, we re-decompose the temperature and load data in the forecasting sample period by CEEMDAN. The result is shown in Figure 4. Figure 4 shows that temperature and load are divided into 8 IMFs and a residual function, respectively.

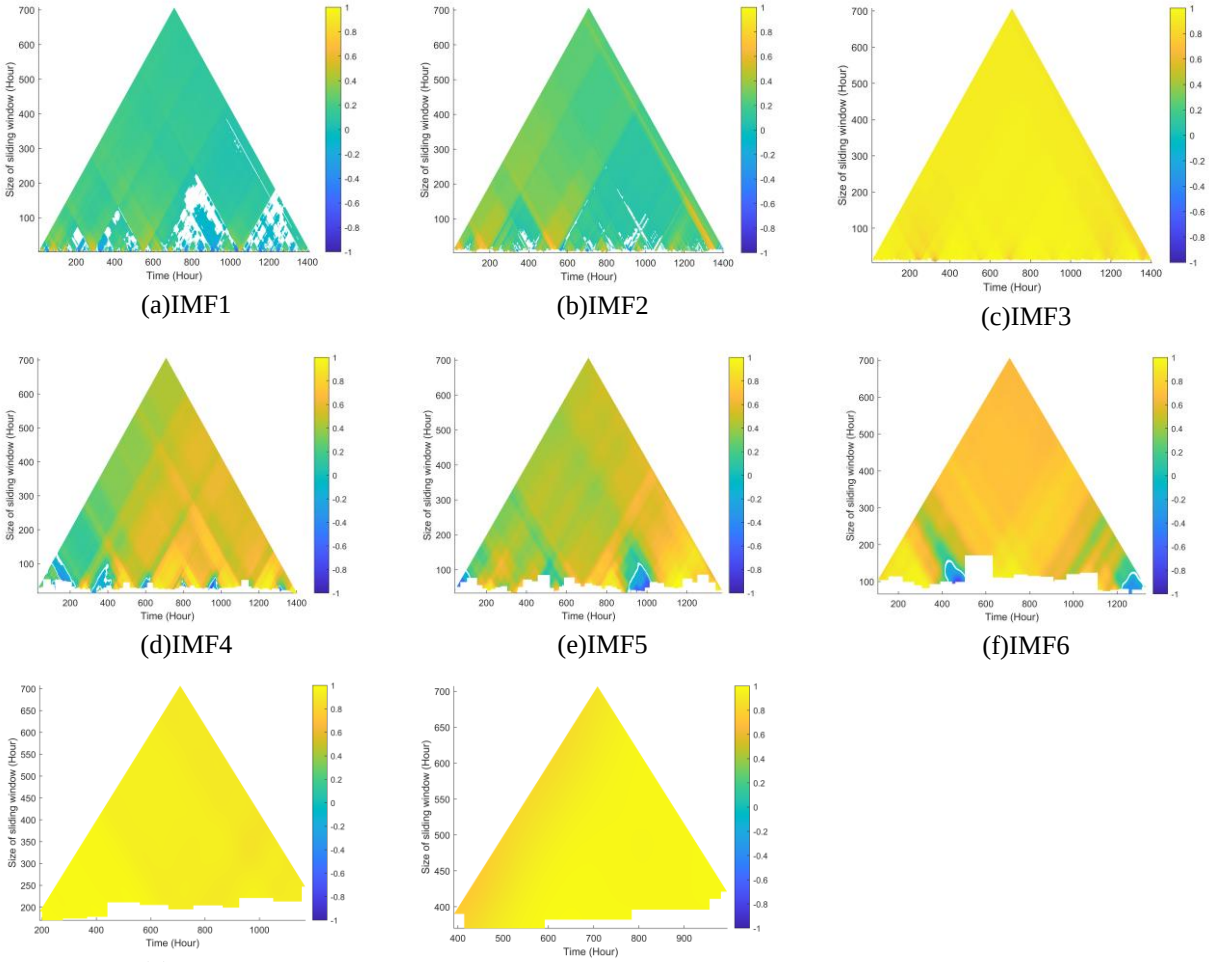


(a) Temperature **(b) Load**

Fig 4: TDIC plots between temperature and load of each IMFs.

Figure 5 shows the TDIC plots of eight IMFs corresponding to temperature and load. It can be clearly seen that the TDIC plots of IMF3, IMF7, and IMF8 are all yellow, and the correlation is higher than 0.8 throughout the triangle, indicating that the three load components are highly correlated with temperature throughout the sample period. The TDIC plots of IMF1 and IMF2 are basically green, indicating that these two load components have basically no

relationship with temperature during the summer period, and their changing laws should be affected by other internal and external factors. When the sliding window of the TDIC plot of IMF4 is about 100 hours, the correlation presents a phenomenon of alternating positive and negative. The overall correlation of IMF5 is not high, but it shows a trend of increasing correlation in the later period of the sample period. On the contrary, the correlation of IMF6 showed a negative correlation trend in the later period. Therefore, the correlation between the eight IMFs of load and the IMFs of temperature is not the same. Among them, IMF3, IMF7, and IMF8 belong to the temperature load component during the entire sample period, while IMF4, IMF5, and IMF6 belong to the temperature load in certain specific periods.



(a)IMF1 (b)IMF2 (c)IMF3

(d)IMF4 (e)IMF5 (f)IMF6

(g)IMF7 (h)IMF8

Fig 5: TDIC plots between temperature and load of each IMFs.

The significance of the analysis in Section 4 is that by decomposing the temperature and load based on CEEMDAN, and performing a correlation analysis based on TDIC for each corresponding IMF, it can be found that IMFs of different scales of load are affected by temperature differently. At the same time, the same IMF is affected differently by the temperature at different times. Therefore, the temperature load can be accurately extracted in a specific period.

It should be noted here that the temperature load extracted by the method proposed in this paper is only a fluctuation component, not the actual summer cooling load or winter heating load. This is because the residual of the load is constant or changes slowly in a certain period, and cannot be analyzed by TDIC, but it also includes constant cooling or heating loads. However, the constant load will not affect the accuracy of STLF in the next section.

V. Load Forecasting Case Study

5.1 Forecasting flow

Based on the analysis and calculation results in Section IV, this paper uses PSO-ELM to forecast each load component after CEEMDAN decomposition. The specific flow chart is shown in Figure 6.

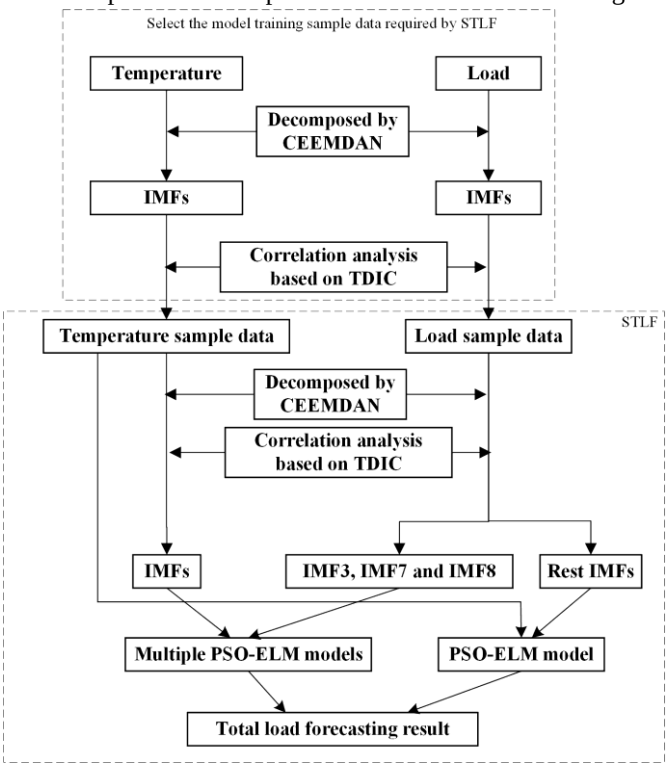


Fig 6: Flow chart of load forecasting method proposed in this paper.

5.2 Case study

The case study is to verify the advantages of the CEEMDAN-based decomposition and TDIC-based correlation analysis of each IMF of temperature and load proposed in this paper in STLF and temperature load extraction. We use the data from 01:00, 25 June 2019, to 24:00, 22 August 2019 in section IV-C as the training data in the forecasting model, to forecast the load from 01:00, 23 August 2019, to 24:00, 23 August 2019.

This paper is aimed at future 24-hours STLF. Assuming that the day to be forecasted is t, the output of the forecasting model is the load data of the day to be forecasted for 24 hours, namely $L(t,1)$ to $L(t,24)$. The input data of the model includes the load and temperature data of the previous day, that is $L(t\text{-}1,1)$ to $L(t\text{-}1,24)$ and $T(t\text{-}1,1)$ to $T(t\text{-}1,24)$, and the temperature data of the day to be forecasted, that is $T(t,1)$ to $T(t,24)$. In the selection of load forecasting factors, in addition to considering the impact of temperature and historical load on the current load, the impact of day types and holidays are also added [16]. So the input data also includes the input day types $DT(t)$ (Monday = 1, Tuesday = 2, and so on) and the input type holiday $HD(t)$ (Holidays = 1, otherwise 0). Therefore, there are 74 input nodes and 24 outputs nodes in the PSO-ELM model. In the selection of model parameters, for PSO, we set c_1=2.8 and c_2=1.3, and the population number and the maximum iteration number are 50. For ELM, the number of hidden neurons is 50, and we use the sigmoid function as the activation function.

5.2.1 Temperature load forecasting

Since the load IMF3, IMF7, and IMF8 are highly correlated with the corresponding temperature IMFs in the forecasting sample period, it can be considered that these three IMFs belong to the temperature load of the day to be forecasted. At the same time, starting from the 1070th hour, IMF5 has a positive correlation with IMF5 of temperature higher than 0.8. It can be considered that IMF5 also a temperature load component of the day to be forecasted. Therefore, the temperature load of the day to be forecasted is a combination of IMF3, IMF5, IMF7, and IMF8.

Taking load IMF3 as an example, when forecasting it, we choose IMF3 of temperature as the input temperature data of the model. The forecasting results and error analysis table are shown in Figure 7 and Table 2, respectively. As shown in the legend of Figure 7, the meaning of "(1)" means that the input temperature data uses the original temperature, and "(2)" means that the input temperature data uses the IMF3 of temperature decomposed by CEEMDAN.

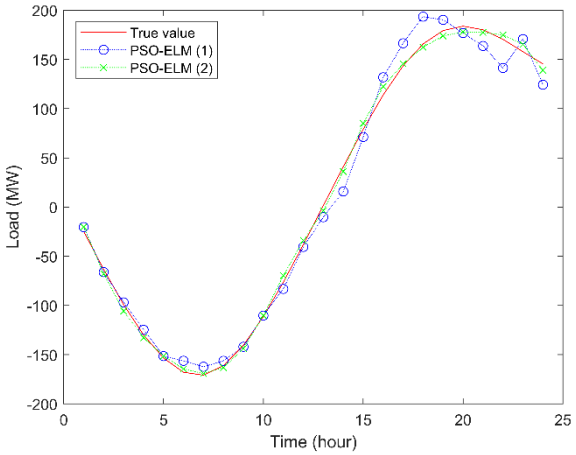


Fig 7: Forecasting results of load IMF3.

Table 2 Error analysis table of load IMF3.

Method	MAPE (%)	RMSE (MW)
PSO-ELM (1)	19.63	4.82
PSO-ELM (2)	43.3	13.98

As seen in Figure 7 and Table 2, PSO-ELM have better forecasting results when the input temperature data is the IMF3 of temperature, and the errors are reduced. Therefore, we use IMF3, IMF7, and IMF8 of temperature as the PSO-ELM input temperature data to forecast the IMF3, IMF7, and IMF8 of the load. Since the overall correlation between IMF5 of load and IMF5 of temperature during the sample period is less than 0.6, we still choose to use the original temperature as the input temperature data for its forecasting. Based on this, we can forecast the temperature load on the day to be forecasted, and the result is shown in Figure 8. It can be concluded that in a day, due to the influence of temperature, the load fluctuation reaches close to 400MW.

ISSN: 0010-8189

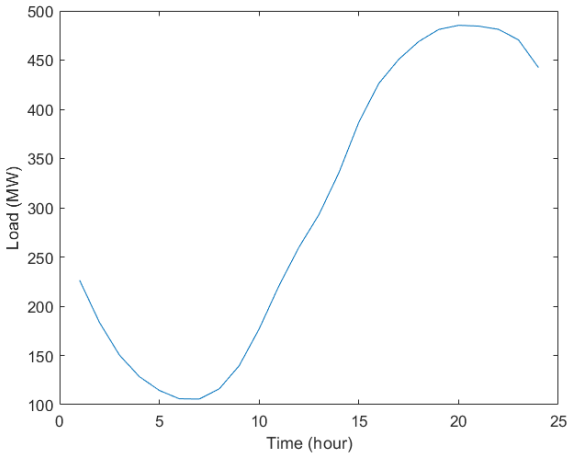


Fig 8: Forecasting results of temperature load.

The significance of the temperature load forecasting is that, based on the temperature forecasting data of the next day, the temperature load that fluctuates with temperature can be forecasted. At the same time, the amount of fluctuation in load affected by temperature can be quantified. This is conducive to grid managers to formulate power dispatch strategies and determine the reserve generation capacity based on temperature changes.

5.2.2 Total load forecasting

In summary, the forecasting method of this paper is divided into two steps. First, the temperature and load are decomposed by CEEMDAN, and TDIC is used to analyze the correlation between temperature and load at different scales. Secondly, for the selected IMF3, IMF7, and IMF8 of load, the corresponding IMFs of temperature are used as the forecasting input, and the original temperature is used as the forecasting input for the other IMFs and the residual, then the PSO-ELM is used to forecast each component of the load.

To verify the superiority of the load forecasting method proposed in this paper, five methods, including ARIMA, ELM, and LSTM, is used to do the future 24-hours STLF of the same sample data to make a comparison with the method proposed in this paper. The five methods are as follows:

- ARIMA: Use ARIMA to forecast the original load.
- ELM: Use ELM to forecast the original load.
- PSO-ELM: Use PSO-ELM to forecast the original load.
- LSTM: Use LSTM to forecast the original load (The LSTM model contains an input layer, two LSTM layers, a fully connected layer, and an output layer. The number of neurons of each LSTM layer is 50. The learning rate is 0.01, and the number of iterations is 100. We use the sigmoid function as the activation function).
- CEEMDAN-PSO-ELM: Decompose the original load using CEEMDAN and combine each component into three parts according to the frequency, namely high-frequency part, intermediate-frequency part, and low-frequency part. Then use PSO-ELM to forecast each part of the load.

The forecasting results are shown in Figure 9:

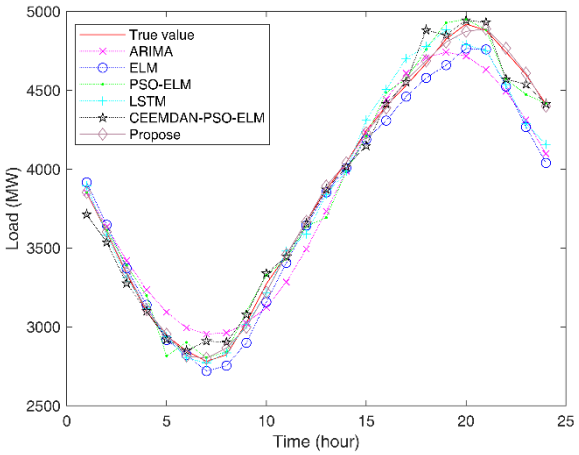


Fig 9: Forecasting results of the total load.

As seen in Figure 9, when ARIMA, ELM, and LSTM are used to forecast the original load, the forecasting accuracy drops sharply after the 17th and 18th hours, while the forecasting results of the other three methods are close to the actual value. For further comparison, Table 3 presents two specific error indicators of the six forecasting methods.

Table 3 Error analysis table of total load forecasting results.

Method	MAPE (%)	RMSE (MW)
ARIMA	3.43	156.47
ELM	2.48	135.68
PSO-ELM	1.65	79.77
LSTM	1.94	115.24
CEEMDAN-PSO-ELM	1.47	78.76
Propose	0.51	23.28

In Table 3, by comparing the forecasting results of ELM and PSO-ELM, it can be concluded that the forecasting performance of PSO-ELM is better, which benefits from the adjustment and optimization of ELM model parameters by PSO. By comparing the first four methods, PSO-ELM has the smallest forecasting error than other methods. It shows that without decomposing load and temperature, PSO-ELM has the highest forecasting accuracy. Because the influence of external factors on the load is not considered, the ARIMA model will produce large errors, and it is not suitable for processing more non-linear data. The reason for the large LSTM forecasting error is that the forecasting accuracy of the LSTM model depends on the selection of internal threshold, weight, the number of neurons, and the number of LSTM layers. Compared with PSO-ELM, LSTM needs to determine and optimize more unknown parameters. Therefore, an improper selection will lead to an error increase. For the case of this paper, LSTM does not perform well. The MAPE and RMSE of CEEMDAN-PSO-ELM are slightly reduced compared to PSO-ELM. This is because the load is decomposed and then recombined, which is beneficial to separate the linear and non-linear parts of the load, but the forecasting performance is not apparent.

The method proposed in this paper uses TDIC to carry out correlation analysis after CEEMDAN decomposition of temperature and load, mining the multi-scale and multi-period correlation of temperature and load. And in load forecasting, for different load components, we use the original temperature data or the corresponding temperature component data as the forecasting input temperature data. Therefore, it performs best in STLF. Compared with CEEMDAN-PSO-ELM method which has no TDIC-based temperature and load correlation analysis and input temperature data selection, the MAPE of the proposed method is reduced by 65%, and the RMSE is reduced by 70%.

VI. Conclusion

This paper proposes a method of temperature load extraction and short-term load forecasting based on CEEMDAN and TDIC. First, temperature and load data are decomposed by CEEMDAN. Second, TDIC is proposed to analyze the multi-scale and multi-period dynamic correlation of temperature and load within a year and summer. Finally, PSO-ELM model is selected to forecast the different components of the load. Through the analysis and calculation of temperature and load data in the PEPCO area of PJM Company in the United States, the results show four aspects: (1) Through CEEMDAN-based decomposition, temperature and load data can be decomposed into multiple time scale components. (2) Through the correlation analysis based on TDIC, it can be found that the correlation between temperature and load at different scales in a year is different. Specifically, in the high-frequency components, temperature and load have a poor correlation. In the intermediate-frequency components, multiple IMFs exhibit a strong positive correlation in summer and a strong negative correlation in winter, and it shows dynamic change characteristics of positive and negative correlations in short time window. In the low-frequency components, the correlation between temperature and load is relatively stable. (3) After redecomposing the sample data used for forecasting model training before the forecast day and analyzing the correlation, multiple load IMFs exhibit a strong correlation with the corresponding temperature IMFs higher than 0.8 during the full sample period. Based on this result, selecting the corresponding temperature IMFs as the input of the forecasting model can improve the accuracy of load forecasting. At the same time, some load IMFs have a high correlation with the temperature only in local periods, so the composition of the temperature load on the day to be forecasted can be accurately extracted through TDIC analysis. (4) The STLF method based on CEEMDAN, TDIC, PSO-ELM proposed in this paper can greatly reduce the error of load forecasting and has good forecasting performance.

Because a part of the temperature load remains constant throughout the day under high or low-temperature conditions, the temperature load extracted in this paper is only a fluctuating part of the actual grid temperature load. Our work on temperature load extraction helps grid managers obtain the amount of fluctuation in the load affected by temperature during a day, and the quantitative relationship between temperature fluctuation and load fluctuation. It provides more accurate multi-time-scale load decomposition data for the formulation of power grid multi-scale operation plan and the optimization of real-time dispatch under random weather conditions. So the study in this paper has important engineering application value.

References

[1] Fan Shu, Chen Luonan. Short-term load forecasting based on an adaptive hybrid method. IEEE transactions on power systems. 2006, 21(1): 392-401.
[2] Wang Yi, Zhang Ning, Tan Yushi, Hong Tao, Kirschen Daniel S, Kang Chongqing. Combining Probabilistic Load Forecasts. IEEE transactions on smart grid. 2019, 10(4): 3664-3674.
[3] Singh Priyanka, Dwivedi Pragya. Integration of new evolutionary approach with artificial neural network for solving short term load forecast problem. Applied energy. 2018, 217: 537-549.
[4] Taylor JamesW. An evaluation of methods for very short-term load forecasting using minute-by-minute British data. International journal of forecasting. 2008, 24(4): 645-658.
[5] Xue Yang, Zhang Ning, Wu Haidong, Yu Zhicheng, Li Rui. Short-term load forecasting method for user side microgrid based on UTCI-MIC and amplitude compression grey model. Power System Technology. 2020, 44: 556-563.
[6] Guan Che, Luh Peter B, Michel Laurent D, Chi Zhiyi. Hybrid Kalman Filters for Very Short-Term Load Forecasting and Prediction Interval Estimation. IEEE transactions on power systems. 2013, 28(4): 3806-3817.
[7] Kuo Pinghuang, Huang Chioujye. A High Precision Artificial Neural Networks Model for Short-Term Energy Load Forecasting. Energies (Basel). 2018, 11: 213.
[8] Malekizadeh M, Karami H, Karimi M, Moshari A, Sanjari MJ. Short-term load forecast using ensemble neuro-fuzzy model. Energy (Oxford). 2020, 196: 117-127.
[9] Duan Min, Darvishan Ayda, Mohammaditab Rasoul, Wakil Karzan, Abedinia Oveis. A novel hybrid

prediction model for aggregated loads of buildings by considering the electric vehicles. Sustainable cities and society. 2018, 41: 205-219.

[10] Huang Guangbin, Zhu Qinyu, Siew Cheekheong. Extreme learning machine: Theory and applications. Neurocomputing (Amsterdam). 2006, 70(1): 489-501.

[11] Huang Guangbin, Zhou Hongming, Ding Xiaojian, Zhang Rui. Extreme Learning Machine for Regression and Multiclass Classification. IEEE transactions on systems, man and cybernetics. Part B, Cybernetics. 2012, 42(2): 513-529.

[12] Cai Weihong, Yang Junjie, Yu Yidan, Song Youyi, Zhou Teng, Qin Jing. PSO-ELM: A Hybrid Learning Model for Short-Term Traffic Flow Forecasting. IEEE access. 2020, 8: 6505-6514.

[13] Zeng Nianyin, Zhang Hong, Liu Weibo, Liang Jinling, Alsaadi Fuad E. A switching delayed PSO optimized extreme learning machine for short-term load forecasting. Neurocomputing (Amsterdam). 2017, 240: 175-182.

[14] Ma Rui, Zhou Xie, Peng Zhou, Liu Daoxin, Xu Huiming, Wang Jun, Wang Xiliang. Data Mining on Correlation Feature of Load Characteristics Statistical Indexes Considering Temperature. Proceedings of the CSEE. 2015, 35(16): 43-51.

[15] Wang Min, Wu Chao, Zhang Peng, Fan Zongyin, Yu Zixuan. Multiscale Dynamic Correlation Analysis of Wind-PV Power Station Output Based on TDIC. IEEE access. 2020, 8: 200695-200704.

[16] Liang Yi, Niu Dongxiao, Hong Weichiang. Short term load forecasting based on feature extraction and improved general regression neural network model. Energy (Oxford). 2019, 166: 653-663.

[17] Li Taiyong, Qian Zijie, He Ting. Short-Term Load Forecasting with Improved CEEMDAN and GWO-Based Multiple Kernel ELM. Complexity. 2020, 2020: 1-20.

[18] Sudheer G, Suseelatha A. Short term load forecasting using wavelet transform combined with Holt–Winters and weighted nearest neighbor models. International journal of electrical power & energy systems. 2015, 64: 340-346.

[19] Huang Norden E., Shen Z., Long S.R. The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis. Proceedings of the Royal Society. A, Mathematical, physical, and engineering sciences. 1998, 454: 95-903.

[20] Wu Z., Huang Norden E. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Advances in Adaptive Data Analysis. 2009, 1: 1-14.

[21] Zhang Jinliang, Zhang Yuejun, Li Dezhi, Tan Zhongfu, Ji Jianfei. Forecasting day-ahead electricity prices using a new integrated model. International journal of electrical power & energy systems. 2019, 105: 541-548.

[22] Qiao Weibiao, Huang Kun, Azimi Mohammadamin, Han Shuai. A Novel Hybrid Prediction Model for Hourly Gas Consumption in Supply Side Based on Improved Whale Optimization Algorithm and Relevance Vector Machine. IEEE access. 2019, 7: 88218-88230.

[23] Wang Sen, Sun Yonghui, Zhou Yan, Mahfoud Rabea Jamil, Hou Dongchen. A New Hybrid Short-Term Interval Forecasting of PV Output Power Based on EEMD-SE-RVM. Energies (Basel). 2019, 13(1): 87.

[24] Haq Md Rashedul, Ni Zhen. A New Hybrid Model for Short-Term Electricity Load Forecasting. IEEE access. 2019, 7: 125413-125423.

[25] Liu Qingzhen, Shen Yuanbin, Wu Lei, Li Jie, Zhuang Lirong, Wang Shaofang. A Hybrid FCW-EMD and KF-BA-SVM Based Model for Short-term Load Forecasting. CSEE JOURNAL OF POWER AND ENERGY SYSTEMS. 2018, 4(2): 226-237.

[26] Gao Xin, Li Xiaobing, Zhao Bing, Ji Weijia, Jing Xiao, He Yang. Short-Term Electricity Load Forecasting Model Based on EMD-GRU with Feature Selection. Energies (Basel). 2019, 12(6): 1140.

[27] Deng Daiyu, Li Jian, Zhang Zhenyuan, Teng Yufei, Huang Qi. Short-term electric load forecasting based on EEMD-GRU-MLR. Power System Technology. 2020, 44(2): 593-602.

[28] Torres M.E., Colominas M.A., Schlotthauer G., Flandrin P., A complete ensemble empirical mode decomposition with adaptive noise. Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, 4144–4147.

[29] Zhang Zichen, Hong Weichiang. Electric load forecasting by complete ensemble empirical mode decomposition adaptive noise and support vector regression with quantum-based dragonfly algorithm. Nonlinear dynamics. 2019, 98(2): 1107-1136.

[30] AL-Musaylh Mohanad S, Deo Ravinesh C, Li Yan, Adamowski JF. Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting. Applied energy. 2018, 217: 422-439.

[31] Li Wenwu, Shi Qiang, Sibtain Muhammad, Li Dan, Mbanze Daniel Eliote. A Hybrid Forecasting Model for Short-Term Power Load Based on Sample Entropy, Two-Phase Decomposition and Whale Algorithm Optimized Support Vector Regression. IEEE access. 2020, 8: 166907-166921.

[32] Chen X., Wu Z., Huang Norden E. The Time-Dependent Intrinsic Correlation Based on The Empirical Mode Decomposition. Advances in Adaptive Data Analysis, 2010, 2(2): 233-265.

[33] Huang Yongxiang, Schmitt FrancoisG. Time Dependent Intrinsic Correlation Analysis of Temperature and Dissolved Oxygen Time Series Using Empirical Mode Decomposition. Journal of marine systems. 2014, 130: 90-100.

[34] Peng Qing, Wen Fenghua, Gong Xu. Time-Dependent Intrinsic Correlation Analysis of Crude Oil and The US Dollar Based on CEEMDAN. International journal of finance and economics. 2021, 26(1): 834-48.

[35] Huang Norden E., Shen Samuel S.P., Hilbert-Huang Transform and its Applications: 2nd Edition. Singapore: World Scientific Publishing, 2014.

[36] Data Miner 2-Hourly Load: Metered, PJM, Accessed: 4 May, 2020. [Online]. Available: https://dataminer2.pjm.com/feed/hrl_load_ metered /definition/.

[37] Hourly/Sub-Hourly Observational Data Map, Accessed: 3 May, 2020. [Online]. Available: https://gis.ncdc.noaa.gov/maps/ncei/cdo/ hourly?layers=001/.