Zhang, G., Tian, C., Li, C., Zhang, J., and Zuo, W., 2020
Accurate Forecasting of Building Energy Consumption Via A Novel Ensembled Deep Learning Method Considering the Cyclic Feature. Energy.
This paper has been accepted by Energy on 03/31/2020.

Accurate Forecasting of Building Energy Consumption Via A Novel Ensembled Deep Learning Method Considering the Cyclic Feature

Guiqing Zhang[a], Chenlu Tian[a], Chengdong Li[a,d,*], Jun Jason Zhang[b], Wangda Zuo[c]

[a]*Shandong Key Laboratory of Intelligent Buildings Technology, School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan 250101, China*
[b]*School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China*
[c]*Department of Civil, Environmental and Architectural Engineering, University of Colorado Boulder, Boulder, CO 80309, U.S.A*
[d]*Shandong Co-Innovation Center of Green Building, Jinan 250101, China*

Abstract

Short-term forecasting of building energy consumption (BEC) is significant for building energy reduction and real-time demand response. In this study, we propose a new method to realize half-hourly BEC prediction. In this new method, to fully utilize the existing data features and to further promote the forecasting performance, we divide the BEC data into the stable (cyclic) and stochastic components, and propose a novel hybrid model to model the stable and stochastic components respectively. The cyclic feature (CF) is extracted via the spectrum analysis, while the stochastic component is approximated by a novel Deep Belief Network (DBN) and Extreme Learning Machine (ELM) based ensembled model (DEEM). This novel hybrid model is named DEEM+CF. Furthermore, two real-world BEC experiments are performed to verify the proposed method. Also, to display the superiorities of the proposed DEEM+CF, this model is compared with the DBN, DBN+CF, ELM, ELM+CF, Support Vector Regression (SVR) and SVR+CF. Experimental results indicate that the CF has a great influence on the promotion of forecasting accuracy for approximately 20%, and DEEM+CF performance is the best among the comparative models, with at least 3%, 6%, 10% better accuracy than the DBN+CF, ELM+CF and SVR+CF respectively under the criteria of MAE.

Keywords: Building energy consumption, Cyclic feature, Deep belief network, Extreme learning machine, Spectrum analysis

1. Introduction

The building energy consumption (BEC) accounts for about 30% of the whole energy usage in the world, and it is still increasing in a fast speed [1]. The growing BEC has attracted

*Corresponding author
Email addresses: qqzhang@sdjzu.edu.cn (Guiqing Zhang), chenlutian2017@sdjzu.edu.cn (Chenlu Tian), lichengdong@sdjzu.edu.cn (Chengdong Li), jun.zhang.ee@whu.edu.cn (Jun Jason Zhang), Wangda.Zuo@Colorado.edu (Wangda Zuo)

much attention worldwide due to the environmental degradation [2]. On the other aspect, recently, lots of advanced information technologies applied in buildings and grid make it possible to realize end-to-end connection, pushing the building and grid to a new area where the building's role is transformed from the pure customer to multiple identical prosumer [3]. In such conditions, the hourly or half hourly short-term prediction of BEC has become a foundation task in the real-time demand response, building energy optimization, etc., which play a great role in both building energy reduction and grid operation and management [4].

To achieve short-term forecasting of BEC, lots of researches are conducted using various methods. The methods applied in this domain mainly include physical models [5, 6], statistic models [7], and machine learning methods [8]. Among these methods, machine learning has become one of the most promising methods recently because of its good capacity in nonlinear approximation without the need for some detailed or unavailable building and environmental knowledge. Machine learning can be divided into the traditional machine learning and deep learning. Each machine learning method has specific advantages and application circumstances [9, 10, 11]. Aiming at improving the prediction performance of BEC, some traditional machine learning methods are always integrated together according to the application requirements. For example, in [12], the random forest is combined with the back propagation neural network to generate a hybrid model for performance forecasting of the ground source heat pump system. Jung [13] utilized the improved least-squared Support Vector Regression (SVR) to realize more accurate BEC forecasting. Yuan et al. [14] adopted the particle swarm optimization in an improved ELM for the robust forecasting of the BEC. Huang et al. [15] constructed an ensemble forecasting model which combined the extreme gradient boosting, SVR, ELM, and the multiple linear regression for energy demand forecasting. These ensemble methods have achieved good results, and ELM is one of the most popular method for its fast computing and good capability of approximation.

Another popular idea to achieve and improve the forecasting of BEC is to combine the deep learning model with traditional model, because the deep learning has deeper computing layers and allow higher levels of feature and relation abstractions [16], while the traditional machine learning has lower computational complexity. Inspired by the idea of parallel system and parallel learning [17, 18, 19], Tian et al. [20] utilized GAN to achieve data enhancement which was applied in some traditional machine learning methods to improve the forecasting results. Fu [21] presented a hybrid model adopting the empirical mode decomposition and DBN to realize the forecasting of the building cooling load. Li et al. [22] proposed a modified DBN utilizing ELM to boost the forecasting accuracy of BEC. However, in existing studies, the abstracted features from various layers of the deep learning models are not fully utilized.

Even though, people have made lots of contributions to the improvement of machine learning method in BEC forecasting, one unavoidable problem is that the performance of machine learning method relies greatly on the input data, thus, it is significant to extract and utilize the valuable features which are inherent in the original data. Recently, deep learning began to be applied in feature engineering of BEC forecasting. In [23], Autoencoders and GAN are used in feature extraction to improve the prediction accuracy of BEC. Some other deep learning models such as DBN are also professional in feature abstraction via layer-by-layer processing, but such features from each layer of deep learning model are not fully

utilized. Besides the inherent features extracted by some deep learning methods, BEC has its obvious cyclic feature – the daily periodic feature. People always leave home at about 8-9 am, and go back at 5-7 pm. Even in their work place, they work for a while then have a rest. Also, the temperature are also periodic in one day [24]. All of these periodic components combine to lead to the daily cyclic feature of BEC. In [25], the cyclic feature of electricity demand is analyzed deeply and is utilized to generate the synthetic sequences. However, to the authors' knowledge, such cyclic feature has barely been taken into account in the present algorithms for the BEC forecasting.

In this paper, a novel DBN and ELM based ensembled method considering the cyclic feature of the observed data, named DEEM+CF, is proposed to achieve half hourly short-term prediction of BEC. In this new method, the main steps are listed below:

- Firstly, the cyclic feature of daily BEC is extracted by spectrum analysis, and the original data is divided into stable (cyclic) and the stochastic ones.
- Secondly, the DEEM is utilized to predict the stochastic ones. In the DEEM, different layers of the DBN are used to abstract different levels of stochastic data features, and the new constructed feature sets from each layer of DBN are then used to train the corresponding ELMs. Such ELMs output the preliminary forecasting results, and further being integrated by another ELM to generate the final predicted results for the stochastic components. The DEEM takes full use of all abstracted features from each layer of DBN.
- Thirdly, the predicted results from DEEM are combined with the cyclic feature to give the final forecasting outputs of BEC.

What's more, to prove the effectiveness and the superiorities of the proposed DEEM+CF model, two experiments utilizing two real-world datasets are conducted in this paper, and comparisons with the pure DBN, the DBN+CF, the ELM, the ELM+CF, the SVR and the SVR+CF are made. Experimental results and comparisons demonstrate that the utilization of the cyclic feature can greatly promote the BEC prediction accuracy approximately 20%, and the DEEM+CF performs at least 3%, 6%, 10% better than the DBN+CF, ELM+CF and SVR+CF.

The remainder of this paper is organized as follows. Section 2 gives a basic introduction of the DBN and ELM. In section 3, the DEEM+CF model is proposed and illustrated in details. In Section 4, two experiments utilizing two real-world datasets are performed to prove the superiorities and effectiveness of the DEEM+CF. Finally, we draw the conclusions of this research in Section 5.

2. Methodologies

The DBN and ELM models are the basic components of the proposed DEEM. In this section, DBN and ELM will be introduced briefly.

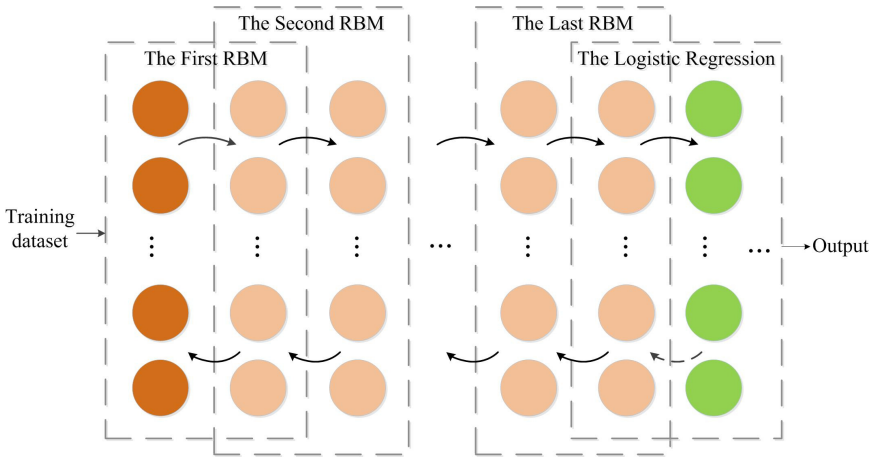


Figure 1: The architecture of DBN [26]

2.1. Deep Belief Networks (DBN)

DBN is stacked by several Restricted Bolzmann Machines (RBMs) one by one [26] as depicted in Figure 1. It is expected to extract high levels of features out of the input data space via layer-by-layer processing.

A single RBM is typically constituted by a hidden layer and a visible layer, and the nodes of various layers are fully connected. The visible layer nodes are regarded as the inputs, while the hidden layer nodes are seen as the outputs. The node values in each layer constitute the binary vector as follows

$$\boldsymbol{v} = \{v_1, v_2, \cdots, v_i, \cdots, v_m\}^{\mathsf{T}} \in \{0,1\}^m, \tag{1}$$

$$\boldsymbol{h} = \{h_1, h_2, \cdots, h_j, \cdots, h_n\}^{\mathsf{T}} \in \{0,1\}^n, \tag{2}$$

where v_i is the visible variable in the visible layer, h_i is the hidden variable in the hidden layer, m is the number of the visible layer nodes, and n is the total number of the hidden layer nodes. The RBM is a model based on energy which is always expected to be lowest. The energy function could be described as [26]

$$E(\boldsymbol{v}, \boldsymbol{h}|\Theta) = -\boldsymbol{a}^{\mathsf{T}}\boldsymbol{v} - \boldsymbol{b}^{\mathsf{T}}\boldsymbol{h} - \boldsymbol{v}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{h} = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i w_{ij} h_j, \tag{3}$$

in which $\Theta = \{\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\}$ represents the set of the model parameters, $\boldsymbol{W} \in R^{I \times J}$ is the weighting matrix, $w_{ij} \in \boldsymbol{W}$ is the weighting variable between v_i and h_j, $\boldsymbol{a} \in R^I$ and $\boldsymbol{b} \in R^J$ are the bias vectors, $a_i \in \boldsymbol{a}$ is the bias of each v_i, and $b_j \in \boldsymbol{b}$ is the bias of each h_j.

To obtain well trained RBMs, the partial derivative of Θ needs to be computed via Gibbs sampling, however it is time-consuming to run the Gibbs sampling for many times. To solve this problem, Hinton [27] proposed the contrastive divergence method to train RBMs, and this method just needs to run Gibbs sampling for K times. Usually, when $K = 1$, the RBM is trained well.

In DBN, the hidden layer of the former RBM is the input layer of the next RBM, and the output of the ultimate RBM is fed into logistic regression. The training process of the initial DBN is constituted by two stages which are the pre-training and fine-tuning process. To begin, suppose that there is a training dataset $(\boldsymbol{X}, \boldsymbol{y})$ which has N samples $\{(\boldsymbol{x}_k, y_k)\}_{k=1}^N$ where $\boldsymbol{x}_k = [x_k^1, x_k^2, \cdots, x_k^m]$. The detailed training processes for the DBN are listed below [26]:

- **Step 1:** Initialize the parameters of DBN including the number of input nodes m , the number of hidden and output nodes n, and the number of the hidden layers L.
- **Step 2:** Input $\boldsymbol{X}$ to the visible layer to train the weighting matrix Θ_1^2 that connects the input layer and the second layer. Θ_1^2 is computed. From this training process, the node values $\boldsymbol{h}(2)$ in the second layer of DBN will be obtained.
- **Step 3:** The node values $\boldsymbol{h}(2)$ in the second layer are then used to determine Θ_2^3. Then, the node values $\boldsymbol{h}(3)$ in the third layer will be gained.
- **Step 4:** Let $l = 3$, the node values $\boldsymbol{h}(l)$ in the lth layer are used to train Θ_l^{l+1}, and the output results $\boldsymbol{h}(l+1)$ of the $(l+1)$th layer in the DBN will be got.
- **Step 5:** Set $l = l + 1$, and then the step 4 is iterated until $l > L + 1$.
- **Step 6:** The outputs from the last hidden layer are fed into a logistic regression part to generate the final output of the DBN. Then, utilize the training data set $(\boldsymbol{X}, \boldsymbol{y})$ again to realize the fine tune of all the parameters of the DBN by the backward propagation algorithm.

2.2. Extreme Learning Machine (ELM)

Suppose that the ELM has n hidden nodes and one output node. The architecture of the ELM is depicted in Figure 2. For the input $\boldsymbol{x} = [x^1, x^2, \cdots, x^m]$, the output of the ELM can be presented as

$$f(\boldsymbol{x}) = \sum_{j=1}^{n} \beta_j g(\boldsymbol{x}, \boldsymbol{a}_j, b_j) \tag{4}$$

where $\boldsymbol{w}_j = (\boldsymbol{a_j}, b_j)^{\mathsf{T}}$ is the weighting vector that connects the input and hidden nodes, and it is randomly given, $\boldsymbol{\beta}$ is the output weighting vector that connects the hidden and output layers, and g represents the activation function.

In the training process of the ELM, no iteration is needed. For the given training dataset $(\boldsymbol{X}, \boldsymbol{y})$ which has N samples $\{(\boldsymbol{x}_k, y_k)\}_{k=1}^N$ where $\boldsymbol{x}_k = [x_k^1, x_k^2, \cdots, x_k^m]$, we firstly compute the training matrix as

$$\boldsymbol{H} = \begin{bmatrix} g(\boldsymbol{a}_1\boldsymbol{x}_1 + b_1) & ... & g(\boldsymbol{a}_n\boldsymbol{x}_1 + b_n) \\ g(\boldsymbol{a}_1\boldsymbol{x}_2 + b_1) & ... & g(\boldsymbol{a}_n\boldsymbol{x}_2 + b_n) \\ ... & ... & ... \\ g(\boldsymbol{a}_1\boldsymbol{x}_N + b_1) & ... & g(\boldsymbol{a}_n\boldsymbol{x}_N + b_n) \end{bmatrix}, \tag{5}$$

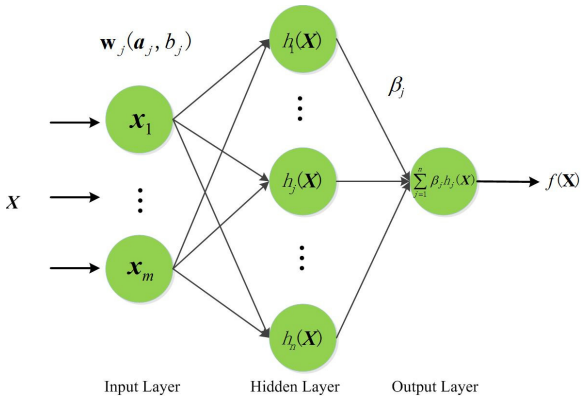


Figure 2: The architecture overview of ELM [28]

in which the parameters $\boldsymbol{a}_i, b_i$ $(i = 1, \cdots, n)$ are randomly given. Then, the weights $\boldsymbol{\beta}$ connecting the hidden layer and the output layer are directly computed via the least square estimation method as

$$\boldsymbol{\beta} = \boldsymbol{H}^{+}\boldsymbol{y} \tag{6}$$

where " $+$ " means the Moore-Penrose generalized inverse, and $\boldsymbol{y} = [y_1, ..., y_N]^{\mathrm{T}}$.

3. The Proposed Forecasting Model Considering the Cyclic Feature

This section presents the proposed DBN and ELM based ensemble method considering the cyclic feature, named DEEM+CF. For clear elaborations, the scheme of the proposed forecasting model will be introduced firstly, and then the cyclic feature extraction will be given, and finally, how to construct the DEEM will be illustrated.

3.1. The Scheme of the Proposed DEEM+CF

The scheme for constructing the proposed DEEM+CF model is shown in Figure 3 and is briefly illustrated as follows:

- **Step 1:** Extract the cyclic feature which is the stable component of the original BEC time series data.
- **Step 2:** Generate the stochastic time series data which is the residual part of the original BEC data after removing the stable component – the cyclic feature. Then, transform the stochastic time series data to the stochastic training dataset.
- **Step 3:** Utilize the stochastic training dataset to optimize the DEEM to achieve the optimal forecasting performance for the stochastic data.
- **Step 4:** Integrate the predicted stochastic results with cyclic features to achieve the final prediction of BEC.

Below, we will give the design details of the proposed DEEM+CF.

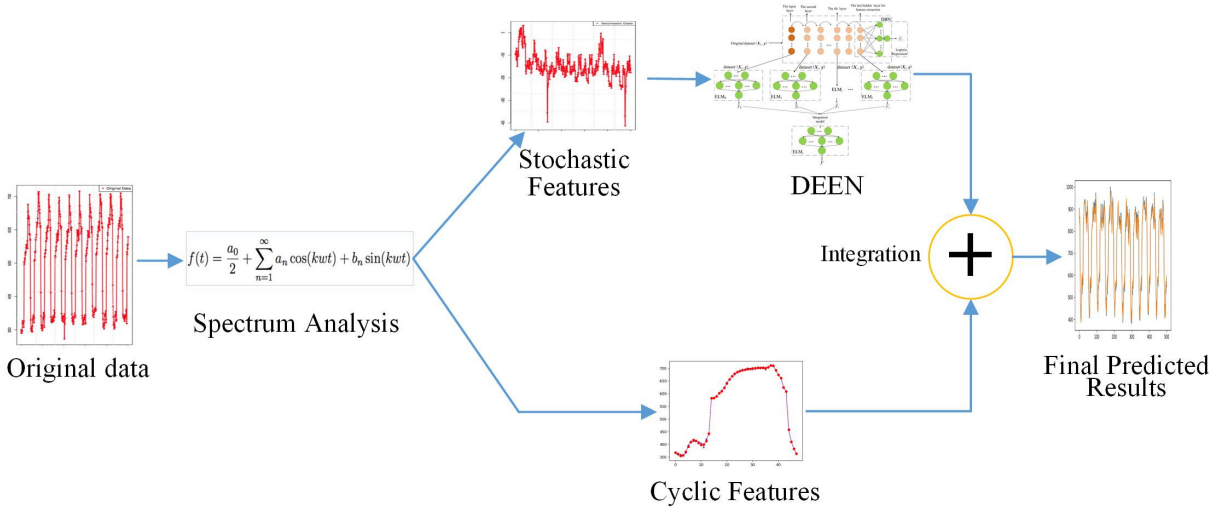


Figure 3: The proposed forecasting scheme

3.2. The Cyclic Feature Extraction via Spectrum Analysis

3.2.1. Spectrum Analysis

One complicated signal can be transformed to simple waves which have specific cyclic periods [29, 30]. Spectrum Analysis is able to achieve such decomposition in format of Fourier series [31, 32] . Recently, this method is always adopted to analyze the inherent information in many domains such as transportation [33, 34], electricity forecasting [35], fault detection [36] and solar radiation analysis [37, 38]. Here, the spectrum analysis is selected to extract the daily cyclic features of BEC series for its good capacity in finding cyclic components. The following of this part is the definition of cyclic spectrum function .

Assume that $f(t)$ is a periodic series with sampling period T. Then, $f(t)$ could be expressed to be Fourier series as

$$f(t) = \sum_{j=1}^{n_i} c_j \cdot e^{jkwt} \tag{7}$$

where a_j is the coefficient, and $w = \frac{2\pi}{T}$.

The Fourier series can also be expanded to be trigonometric polynomials series as

$$f(t) = \frac{c_0}{2} + \sum_{\gamma=1}^{\infty} c_\gamma \cos(kwt) + \sum_{\gamma=1}^{\infty} d_\gamma \sin(kwt) \tag{8}$$

where c_0, c_γ, d_γ can be presented as

$$c_0 = \frac{T}{2}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)dt, \quad c_\gamma = \frac{T}{2}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)cos(\gamma wt)dt, \quad d_\gamma = \frac{T}{2}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)sin(\gamma wt)dt. \tag{9}$$

3.2.2. The Extraction of the Cyclic Feature

To get the cyclic features, firstly, the average of daily BEC value p^t is calculated and obtained. The average of the daily BEC is expressed as

$$p^t = \frac{1}{D}\sum_{i=1}^{D} v_i^t \tag{10}$$

where p^t is the average of the daily values at time t, v_i^t is the original value at time t on the ith day, D is the number of total days.

Then the spectrum function is utilized to extract the cyclic features of daily BEC. To obtain more reasonable cyclic features, BIC is adopted to evaluate the performance of spectrum function. The number of cyclic components (trigonometric waves) will be increased, and BIC of each different spectrum function is calculated. We select the spectrum function which has the lowest BIC as the cyclic model of BEC, and the daily stable components p^t are obtained as

$$\hat{p^t} = c_0 + c_1 \sin\left(\frac{2\pi t}{N}\right) + d_1 \cos\left(\frac{2\pi t}{N}\right) + \cdots + c_n \sin\left(\frac{2n\pi t}{N}\right) + d_n \cos\left(\frac{2n\pi t}{N}\right), \tag{11}$$

where N is the number of daily collected data, $c_0, c_1, \cdots, c_n, d_1, \cdots, d_n$ are computed via the least square estimation method as follows

$$\begin{bmatrix} c_0 \\ c_1 \\ d_1 \\ \cdots \\ c_n \\ d_n \end{bmatrix} = \begin{bmatrix} 1 & \sin\left(\frac{2\pi}{N}\right) & \cos\left(\frac{2\pi}{N}\right) & \cdots & \sin\left(\frac{2n\pi}{N}\right) & \cos\left(\frac{2n\pi}{N}\right) \\ 1 & \sin\left(\frac{4\pi}{N}\right) & \cos\left(\frac{4\pi}{N}\right) & \cdots & \sin\left(\frac{4n\pi}{N}\right) & \cos\left(\frac{4n\pi}{N}\right) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & \sin\left(\frac{2N\pi}{N}\right) & \cos\left(\frac{2N\pi}{N}\right) & \cdots & \sin\left(\frac{2N*n\pi}{N}\right) & \cos\left(\frac{2N*n\pi}{N}\right) \end{bmatrix}^{+} \begin{bmatrix} p^1 \\ p^2 \\ \cdots \\ p^N \end{bmatrix} \tag{12}$$

where " $+$ " means the Moore-Penrose generalized inverse.

After the cyclic features are obtained, the stochastic components are computed via getting rid of the stable ones (cyclic features) from the original data. Each original data can be divided into the stable component and stochastic component as

$$v_i^t = \hat{p}^t + x_i^t \tag{13}$$

where x_i^t is the stochastic component at time t on the ith day.

The stable components reflect the trend of the BEC, while the stochastic components present the specific and random features of the BEC. The stochastic components are combined to 1-D time series data $\{x_1, x_2, \cdots\}$. This remaining stochastic BEC data series is then transformed to the stochastic training dataset $(\boldsymbol{X}_0, \boldsymbol{y})$ which has N samples $\{(\boldsymbol{x}_{0,k}, y_k)\}_{k=1}^{N}$ where $\boldsymbol{x}_{0k} = [x_{0,k}^1, x_{0,k}^2, \cdots, x_{0,k}^m]$.

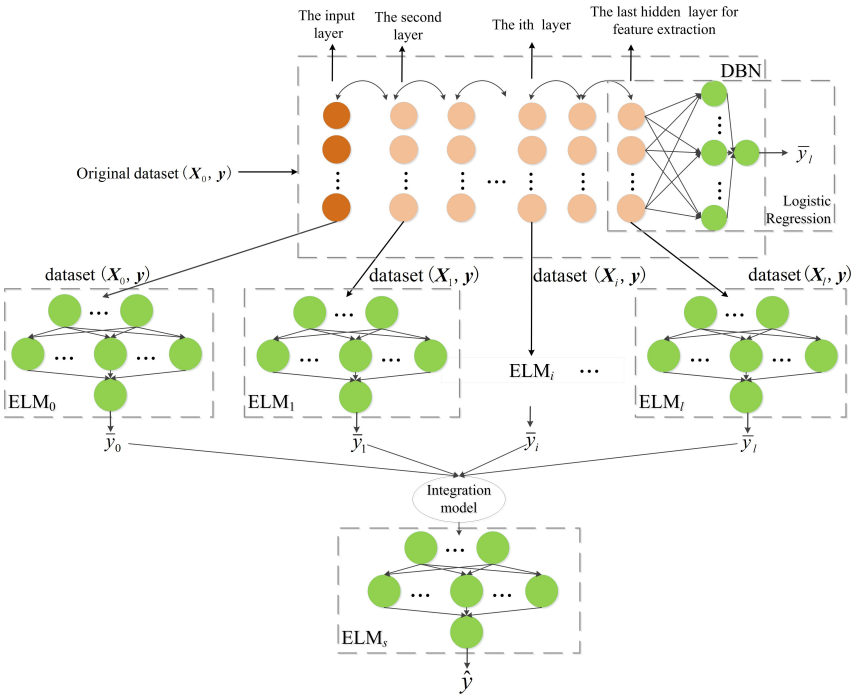


Figure 4: The architecture overview of the proposed DEEM.

3.3. Remaining Stochastic Data-Driven Design of the DEEM

3.3.1. The Framework of the DEEM

In this section, the DBN and ELMs are integrated in an ensemble forecasting method for the forecasting of the stochastic BEC data. The architecture of the DEEM is shown in Figure 4, in which the DBN is utilized to generate the new representative feature datasets, and the ELMs are selected to be the premier and ensemble forecasting models. Firstly, the original training dataset is input to DBN model, and DBN extracts the new data features from the stochastic training dataset via layer-by-layer processing. Each layer of DBN outputs one new feature dataset which combines the target values to be the new training dataset. Then the new training datasets are utilized to train separate ELMs to get the premier predicted results of target values. Finally, all of the premier results are integrated and then combined with the target values again to train another ELM and get the final predicted results.

The construction steps of the DEEM are listed below.

* **Input:** The stochastic training data sets $(\boldsymbol{X}_0, \boldsymbol{y})$, the number of the hidden layers of the DBN.

* **Output:** The final predicted result $\hat{\boldsymbol{y}}$ for the stochastic component.

- **Step 1:** Input the stochastic training dataset $(\boldsymbol{X}_0, \boldsymbol{y})$ to the DBN model, and train the DBN model. Suppose that the outputs from the *ith* hidden layer are $\boldsymbol{X}_i$ $(i = 1, 2, \cdots, l)$, then from the *ith* hidden layer, one new dataset $(\boldsymbol{X}_i, \boldsymbol{y})$ will be generated.

- **Step 2:** Input the generated dataset $(\boldsymbol{X}_i, \boldsymbol{y})$ to one corresponding ELM model to train it and get individual predicted results $\overline{\boldsymbol{y}}_i$ $(i = 1, 2, \cdots, l)$. Besides, the initial training dataset $(\boldsymbol{X}_0, \boldsymbol{y})$ is also used to train a single ELM to obtain the predicted results $\overline{\boldsymbol{y}}_0$.

- **Step 3:** Integrate all of the individual predicted results $\overline{\boldsymbol{y}}_{\boldsymbol{i}}s$ $(i = 0, 1, \cdots, l)$ by another ELM model to generate the final predicted result $\hat{\boldsymbol{y}}$.

In the DEEM model, from each hidden layer of DBN, we will generate one new dataset. The input stochastic training dataset and the newly constructed datasets will all be used to participate the forecasting of the BEC. Compared with the conventional deep learning models, in the DEEM, the initial training dataset and all of the abstracted features from the hidden layers are fully utilized.

Below, we will explain the details of such steps.

3.3.2. Training Data Generation and Learning of ELMs

The newly generated training datasets are obtained from each layer of the DBN. The newly constructed training dataset for the ith hidden layer is $(\boldsymbol{X}_i, \boldsymbol{y})$, and $\boldsymbol{X}_i$ is obtained as

$$\boldsymbol{X}_i = \hat{g}_i(\boldsymbol{X}_{i-1}, \boldsymbol{W}_i, \boldsymbol{a}_i, \boldsymbol{b}_i) \tag{14}$$

where $\hat{g}_i(\cdot)$ represents the activation function in the ith hidden layer of the DBN, and $(\boldsymbol{W}_i, \boldsymbol{a}_i, \boldsymbol{b}_i)$ is the weighting matrix connecting the $(i-1)$th and ith hidden layers of the DBN.

The input part $\boldsymbol{X}_i$ of the newly generated dataset can be finally presented as

$$\boldsymbol{X_i} = \begin{bmatrix} \boldsymbol{x}_{i,1} \\ \boldsymbol{x}_{i,2} \\ \cdots \\ \boldsymbol{x}_{i,N} \end{bmatrix} = \begin{bmatrix} x_{i,1}^1 & \cdots & x_{i,1}^m \\ x_{i,2}^1 & \cdots & x_{i,2}^m \\ \cdots & \cdots & \cdots \\ x_{i,N}^1 & \cdots & x_{i,N}^m \end{bmatrix} \tag{15}$$

where m is the number of the ith hidden layer nodes in the DBN.

If the DBN has l hidden layers for feature abstraction, we will obtain $l+1$ training datasets which include l newly generated training datasets and one original stochastic training dataset. The $l+1$ training datasets will be utilized to construct $l+1$ corresponding ELMs and to obtain $l+1$ premier individual forecasting results $\overline{y}_i$s, which could be computed as

$$\overline{y}_{i,k} = \sum_{j=1}^{n_i} \beta_{i,j} g_i(\boldsymbol{a}_{i,j} \boldsymbol{x}_{i,k} + b_{i,j}) \tag{16}$$

where $i = 0, 1, \cdots, l$, $k = 1, \cdots, N$, $g_i(\cdot)$ is the activation function in the ith ELM, $(\boldsymbol{a_{i,j}}, b_{i,j})$ is the weighting vector which connects the input layer and the hidden layer of the ith ELM, and there are n_i hidden nodes in the ith ELM. $\boldsymbol{\beta}_i = [\beta_{i,1}, \cdots, \beta_{i,n_i}]^{\mathsf{T}}$ is the weighting vector

that connects the hidden and output layers of the ith ELM, and can be determined as

$$\begin{aligned}\boldsymbol{\beta}_i &= \left[\beta_{i,1}, \beta_{i,2}, \cdots, \beta_{i,n_i}\right]^{\mathsf{T}} \\ &= \begin{bmatrix} g_i(\boldsymbol{a}_{i,1}\boldsymbol{x}_{i,1} + b_{i,1}) & \ldots & g_i(\boldsymbol{a}_{i,n_i}\boldsymbol{x}_{i,1} + b_{i,n_i}) \\ g_i(\boldsymbol{a}_{i,1}\boldsymbol{x}_{i,2} + b_{i,1}) & \ldots & g_i(\boldsymbol{a}_{i,n_i}\boldsymbol{x}_{i,2} + b_{i,n_i}) \\ \vdots & \vdots & \vdots \\ g_i(\boldsymbol{a}_{i,1}\boldsymbol{x}_{i,N} + b_{i,1}) & \ldots & g_i(\boldsymbol{a}_{i,n_i}\boldsymbol{x}_{i,N} + b_{i,n_i}) \end{bmatrix}^{+} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}\end{aligned} \tag{17}$$

3.3.3. Design of the Ensemble Part

In the ultimate ensemble part, the $l+1$ premier predicted results will be firstly combined to be a new training dataset, and then the newly generated dataset will be utilized to construct the ensemble model which is chosen to be the ELM again due to its low computation complexity and good capability in nonlinear approximation.

Assume that the integrated training dataset for the final training is $(\overline{\boldsymbol{Y}}, \boldsymbol{y})$, where $\overline{\boldsymbol{Y}}$ can be expressed as

$$\overline{\boldsymbol{Y}} = \left[\overline{\boldsymbol{y}}_0, \overline{\boldsymbol{y}}_1, \cdots, \overline{\boldsymbol{y}}_l\right] = \begin{bmatrix} \overline{y}_{0,1} & \cdots & \overline{y}_{l,1} \\ \overline{y}_{0,2} & \cdots & \overline{y}_{l,2} \\ \vdots & \vdots & \vdots \\ \overline{y}_{0,N} & \cdots & \overline{y}_{l,N} \end{bmatrix} \tag{18}$$

in which $\overline{y}_{i,k}$ is the predicted result for the input data $\boldsymbol{x}_{i,k}$, and can be obtained by (16).

$\overline{\boldsymbol{Y}}$ can also be expressed as

$$\overline{\boldsymbol{Y}} = [\overline{\boldsymbol{y}}^{(0)}, \overline{\boldsymbol{y}}^{(1)}, \cdots, \overline{\boldsymbol{y}}^{(N)}]^{\mathsf{T}} \tag{19}$$

where $\overline{\boldsymbol{y}}^{(i)} = [\overline{y}_{0,i}, \overline{y}_{1,i}, \cdots, \overline{y}_{l,i}]^{\mathsf{T}}$ for $i = 1, 2, \cdots, N$.

Then, the integrated training data set will be employed to construct another ELM. To begin, suppose that the ELM in the ensemble part has q hidden nodes and its input-output mappings can be given as

$$\hat{y}_i = \sum_{p=1}^{q} \hat{\boldsymbol{\beta}}_p \hat{g}(\hat{\boldsymbol{a}}_p \overline{\boldsymbol{y}}^{(i)} + \hat{b}_p) \tag{20}$$

where $i = 1, 2, \cdots, N$, $(\hat{\boldsymbol{a}}_p, \hat{b}_p)$ is the weighting matrix that connects the input and hidden layers of the integration ELM model, $\hat{g}(\cdot)$ represents the activation function in the integration ELM, and $\hat{\boldsymbol{\beta}}$ is the weighting vector connecting the hidden and output layers.

To assure the performance of the ensemble ELM, its weighting vector is also obtained

by the least square estimation as

$$\hat{\boldsymbol{\beta}} = \left[\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_q\right]^{\mathsf{T}} = \begin{bmatrix} \hat{g}(\hat{\boldsymbol{a}}_1\overline{\boldsymbol{y}}^{(1)} + \hat{b}_1) & \cdots & \hat{g}(\hat{\boldsymbol{a}}_q\overline{\boldsymbol{y}}^{(1)} + \hat{b}_q) \\ \hat{g}(\hat{\boldsymbol{a}}_1\overline{\boldsymbol{y}}^{(2)} + \hat{b}_1) & \cdots & \hat{g}(\hat{\boldsymbol{a}}_q\overline{\boldsymbol{y}}^{(2)} + \hat{b}_q) \\ \cdots & \cdots & \cdots \\ \hat{g}(\hat{\boldsymbol{a}}_1\overline{\boldsymbol{y}}^{(N)} + \hat{b}_1) & \cdots & \hat{g}(\hat{\boldsymbol{a}}_q\overline{\boldsymbol{y}}^{(N)} + \hat{b}_q) \end{bmatrix}^{+} \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} \tag{21}$$

4. Experiments and Comparisons

To verify the advantages of the proposed DEEM+CF model, two comparative experimental studies will be conducted in this section.

4.1. Experimental Setting and Applied Datasets

4.1.1. Comparative Methods

For the purpose of showing the advantages of the proposed DEEM+CF method, firstly, several popular regression models including lasso regression [39], ridge regression [40] and multi-polynomial [41] are adopted to be the comparative methods of spectrum analysis in cyclic feature extraction. ELM is utilized to be the prediction model. Secondly, several popular machine learning models, including the DBN, ELM, and the SVR, are selected to be the comparative models of DEEM+CP. Besides, to verify the effectiveness of the cyclic feature furthermore, the hybrid models that combine the cyclic feature with the DBN, ELM and SVR are respectively constructed to be the comparative models too, i.e. the DBN+CF, ELM+CF, and SVR+CF are also designed to be the comparative models.

4.1.2. Evaluation Indices

To evaluate the forecasting performance, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Pearson Correlation Coefficient (r) are selected as the evaluation indices. The four comparative indices have been widely used for forecasting accuracy evaluation and are computed as

$$MAE = \frac{1}{M}\sum_{m=1}^{M}|\hat{y}_m - y_m| \tag{22}$$

$$MAPE = \frac{1}{M}\sum_{m=1}^{M}\frac{|\hat{y}_m - y_m|}{y_m} \times 100\% \tag{23}$$

$$RMSE = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{y}_m - y_m)^2} \tag{24}$$

$$r = \frac{\sum_{m=1}^{M}(\hat{y}_m - E(\hat{y}_m))(y_m - E(y_m))}{\sqrt{(\hat{y}_m - E(\hat{y}_m)^2)}\sqrt{(y_m - E(y_m)^2)}} \tag{25}$$

where y_m and $\hat{y}_m$ are respectively the observed values and predicted values, $E(\cdot)$ represents the average of the samples.

Besides, to determine the number of cyclic features, Bayesian Information Criterion (BIC) is adopted for model construction of the spectrum function. BIC can balance parameter adding and overfitting, and lower BIC means better model. BIC is calculated as

$$BIC = \ln(M)k - 2ln(\hat{L}) \tag{26}$$

where M is the number of data, k is the number of parameters adopted by model, $\hat{L}$ is the maximum likelihood function of the model.

When the errors of model are independent and distributed according to normal distribution, BIC can be presented as

$$BIC = M\ln(\hat{\sigma^2}) + k\ln(M) \tag{27}$$

where $\hat{\sigma^2}$ is the error variance which is computed as

$$\hat{\sigma^2} = \frac{1}{M}\sum_{m=1}^{M}(\hat{y}_m - y_m)^2 \tag{28}$$

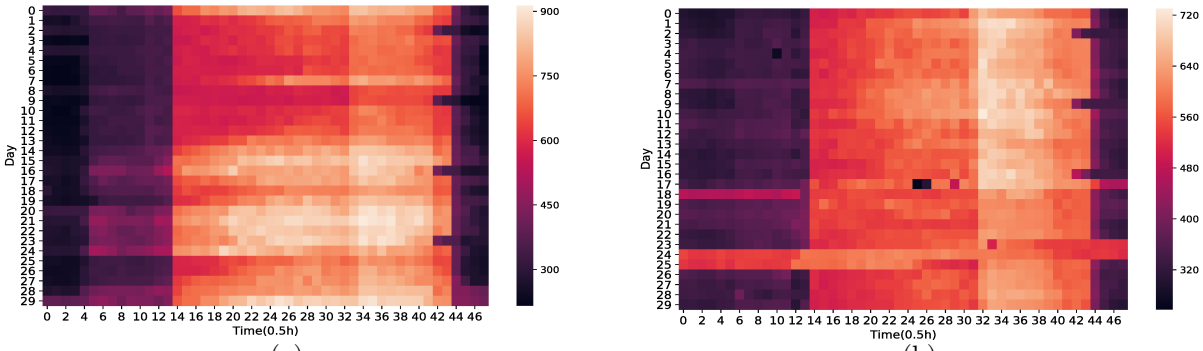
4.1.3. Applied Dataset

Two buildings are chosen as the testing buildings to prove the effectiveness and superiorities of the DEEM+CF method. The BEC datasets are retrieved from https://trynthink.github.io/buildingsdatasets/.

The first building is located in Hialeah which is one of the warmest place in America. Its energy consumption status was collected every 15 minutes from January 1, 2010 to December 31, 2010. There are 34940 samples in the dataset. Comparatively, the energy consumption in summer is higher than the other seasons in this building. The original data was processed and aggregated into the 30 minutes interval, and 17470 samples are obtained finally. In the newly dataset, the value scale of energy consumption in half an hour is between 219 to 1032 kW.

The second building is from Pico Rivera, CA where the climate is comfortable. There are four collected data points in one hour, and the data from January 1, 2010 to October 31, 2010 were selected. The data collected in summer is more fluctuant than in winter. The data from this building was integrated into the 30 minutes interval too, and there are 14592 samples at last. The highest BEC in half an hour is 997 kW and the lowest value is 191 kW.

The original daily BEC data of the two buildings in one month is displayed in Figure 5(a) and Figure 5(b). In each experiment, the stochastic data and the original data will be divided into two parts, we use the first 70% data as training dataset and the left 30% for testing, and the size of the input sequence is set to be 10.



(a) (b)

Figure 5: (a) Daily BEC data (kW) in the first experiment, (b) Daily BEC data (kW) in the second experiment.

4.2. The First Experiment

4.2.1. Configuration of the Forecasting Models

The proper configuration of parameters is important for the forecasting accuracy of machine learning models. In this experiment, the optimal parameters of the models, including the spectrum function, DEEM, and the comparative models are detailed below.

(a) Configuration of the spectrum function

To obtain proper number of the cyclic waves in spectrum functions, the average of the BEC time series in the first building is computed firstly via (10) and then input to the spectrum function model. The number of the cyclic waves in the form of trigonometric functions changes from 1 to 30. The performance of each spectrum function with specific number of cyclic waves is evaluated via BIC. The lower BIC means more reasonable cyclic features are extracted without significant overfitting.

Figure 6(a) illustrates the performances of the spectrum functions with different number of cyclic waves. From this figure, we can see that when the spectrum function has 25 cyclic waves, it obtains the best performance in this experiment. To show the cyclic features more clearly, the spectrum map which reflect the amplitude of cyclic waves which have different frequencies is show in Figure 6(b). It is clear that there are two significant cycles in period of 3-4 hours and 24 hours, and these two significant cycles are combined with other 23 cycles to present the daily cyclic feature of BEC. The stable and the stochastic time series data are then obtained. Figure 6(c) demonstrates the first 500 original BEC data of the first building, and Figure 6(d) presents all of the remaining stochastic BEC data of the first building.

On the other aspect, to evaluate the performance of spectrum function in cyclic feature extraction, lasso regression, ridge regression and multi polynomial regression are also used to model the cyclic features, and ELM is selected to be the prediction model. Here, the penalty coefficient of lasso regression is set to be 1, and the number of dimensions is set to be 26. The penalty coefficient and the number of dimensions in ridge regression is set to be 0.01 and 26 separately. The highest degree of independent variable in multi-polynomial regression is set to be 10. All of the experiments are conducted for ten times, and the

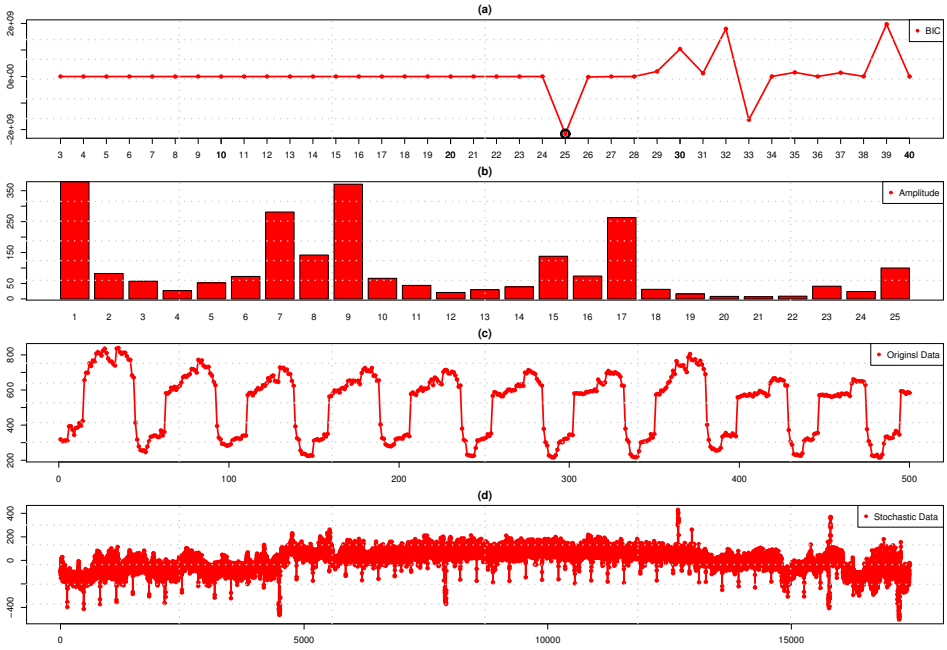


Figure 6: (a) The performance of the spectrum functions with different number of trigonometric waves in the first experiment,(b) The spectrum map of cycle features in the first experiment (c) The former 500 original BEC data in the first experiment, (d) The stochastic BEC data in the first experiment.

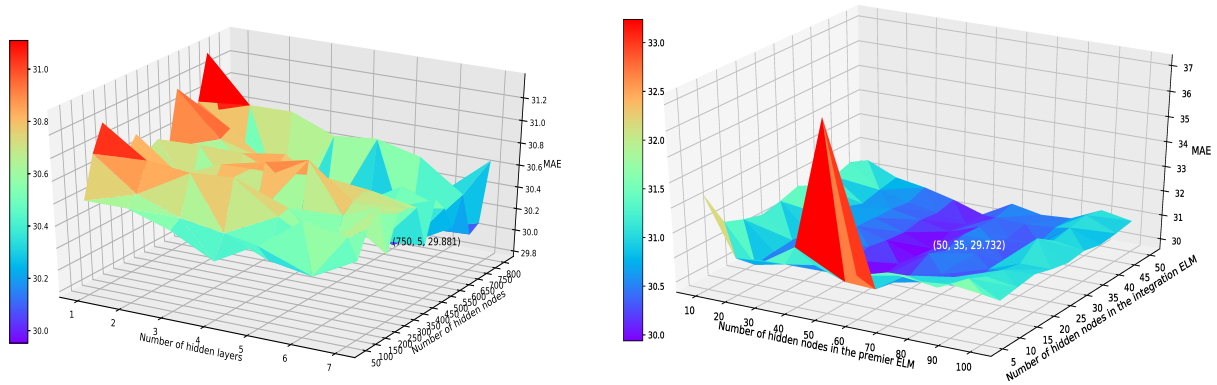
predicted results using four cyclic feature models are compared under the criteria of MAE, MAPE, RMSE and r.

(b) Configuration of the DEEM

The DEEM is composed of the DBN and ELMs, thus, for the purpose of achieving accurate forecasting, it is significant to determine the proper numbers of the hidden layers and the nodes in each hidden layer of the DBN and the ELMs. In order to seek the best structure of the DEEM, the parameter searching experiment of the DEEM is conducted in two stages. In the first stage, we fix the numbers of the hidden nodes in the ELMs, while changing the numbers of the hidden layers and the nodes in each hidden layer of the DBN. In the second stage, the selected best structure for the DBN in the first stage is fixed, while the numbers of the nodes in the hidden layer of the premier and integration ELMs are changed. Here, the original data are utilized to determine the best structure of DEEM for evaluation of the proposed method.

In the first stage, we set the number of hidden layers from 1 to 7, and change the number of the nodes in each hidden layer of the DBN from 50 to 800 at interval of 50. The predicted performance of each DEEM with different number of hidden layers and hidden nodes is evaluated under the criteria of MAE. Figure 7(a) shows the MAEs of different DEEMs. The lower value of MAE means better forecasting performance of DEEM. It is clear that when we choose 5 hidden layers and 750 hidden nodes in each layer of the DBN, MAE reaches the minimal value throughout all of the results.

In the second stage, we fix the structure of the DBN with 5 hidden layers and 750 hidden



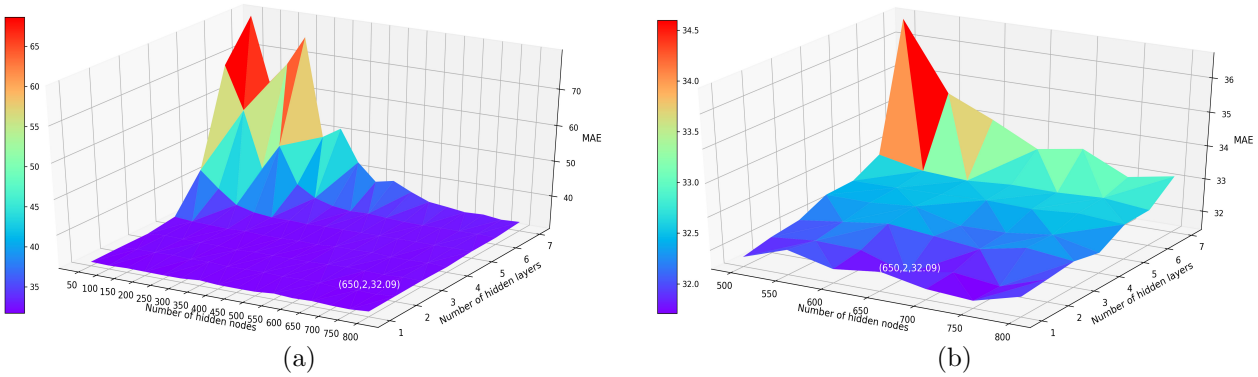
(a) (b)

Figure 7: (a) The MAEs of the DEEMs with different numbers of hidden layer and hidden nodes in DBN when the premier and integration ELMs are fixed in the first experiment, (b) The MAEs of the DEEMs with different numbers of hidden nodes in premier and integration ELMs when the DBN is fixed in the first experiment.

nodes in each layer, while the number of hidden nodes in the premier ELMs are set from 5 to 50 at the interval of 5, and the number of hidden nodes in the integration ELM is changed from 10 to 100 at the interval of 10. As the first stage, the performances of all of the DEEMs with different number of hidden nodes in premier and integration ELMs are compared under the criteria of MAE. Figure 7(b) shows the MAE comparison of such DEEMs. It can be seen from this figure that the best premier and integration ELMs have 50 and 35 hidden nodes respectively.

(c) Configuration of the other comparative models

To achieve the rationality of performance comparison, the optimal parameter searching processes of the DBN, the ELM and the SVR using the original data are also carried to accomplish the best performance of the comparative models.



(a) (b)

Figure 8: (a) The MAEs of the DBNs with different numbers of hidden layer and nodes when the regression part is fixed in the first experiment, (b) Fine details of the forecasting performance in Figure 8(a).

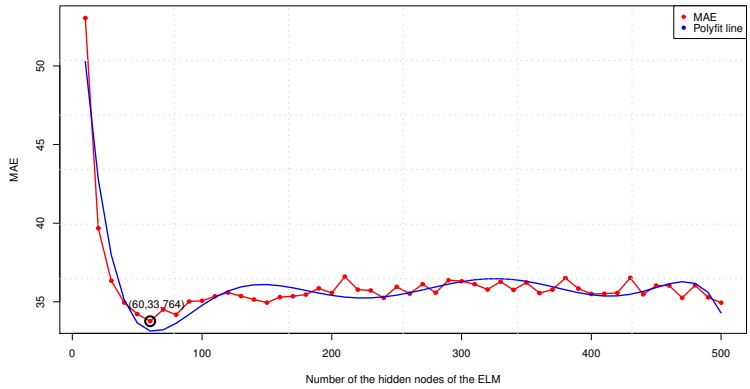


Figure 9: The MAEs of the ELMs in the first experiment.

In this paper, the adopted DBN is composed of several RBMs and one fully connected layer for logistic regression. For the DBN, the numbers of the hidden layers and the nodes in each hidden layer are also key factors affecting the forecasting accuracy. To obtain the best parameters, the number of the nodes in each hidden layer is also changed from 50 to 800 at interval of 50, the number of the hidden layers is set from 1 to 7, and the number of hidden nodes in the regression part changes from 5 to 50 at interval of 5. MAE is selected again to evaluate the performance of DBNs when the number of hidden layer and the numbers of nodes in hidden layer and regression part are changed separately. The MAE achieves the lowest result when the DBN has 2 hidden layers, 650 nodes in each hidden layer, and 35 hidden nodes in the regression part. Figure 8(a) shows the forecasting performance of the DBNs with different numbers of hidden nodes and layers but fixed regression part. To trace the important details of Figure 8(a), the key part of Figure 8(a) is zoomed in Figure 8(b).

The best structure of the ELM for comparison is also explored. The number of hidden nodes in the ELM is set from 10 to 500 at the interval of 10. Figure 9 shows the MAEs of such ELMs. We can observe that the best ELM has 60 hidden nodes.

For the SVR, we choose the RBF function to be its kernel function again. And, through testing, we set the penalty and kernel coefficients to be 0.5 and 0.6 respectively.

4.2.2. Experimental Results

Table 1 shows the average values and standard derivations of MAE, RMSE, MAPE, and r of the forecasting performance using different cyclic feature models when the ELM is selected to be the prediction model.

Table 2 demonstrates the average values and standard derivations of the MAE, RMSE, MAPE, and r of the forecasting models considering or without considering the cyclic feature which is obtained by spectrum analysis. The predicted residential errors of the proposed DEEM+CF and the other comparative forecasting models are recorded and their kernel density histograms are shown in Figure 10.

Table 1: Forecasting performance using different cyclic feature models when ELM is selected to be the prediction model in the first experiment.

Model	MAE	RMSE	MAPE(%)	r
ELM+Lasso	31.627 ±1.274	44.692 ±1.184	5.044 ±0.189	$0.971 \pm 1.674 \times 10^{-3}$
ELM+Ridge	32.698 ±0.489	44.235±1.141	5.032 ±0.175	$0.971 \pm 8.751 \times 10^{-4}$
ELM+Multi-polynomial	30.468 ±1.230	41.787 ±0.571	4.920 ±0.133	$0.975 \pm 5.253 \times 10^{-4}$
ELM+Spectrum	25.450 ±0.338	39.995 ±0.389	5.121 ±0.137	$0.977 \pm 3.695 \times 10^{-4}$
ELM	34.009 ±1.054	48.165 ±1.159	5.826 ±0.223	$0.964 \pm 1.701 \times 10^{-3}$

Table 2: Performances of the forecasting models in the first experiment ("model+CF" is the model considering the cyclic feature which is extracted by spectrum analysis).

Model	MAE	RMSE	MAPE(%)	r
SVR	36.751 ±0.000	48.065 ±0.000	6.479 ±0.000	$0.965 \pm 0.000 \times 10^{-4}$
SVR+CF	26.544 ±0.000	36.852 ±0.000	4.869 ±0.000	$0.982 \pm 0.000 \times 10^{-4}$
ELM	34.009 ±1.054	48.165 ±1.159	5.826 ±0.223	$0.964 \pm 1.701 \times 10^{-3}$
ELM+CF	25.450 ±0.338	39.995 ±0.389	5.121 ±0.137	$0.977 \pm 3.695 \times 10^{-4}$
DBN	32.197 ±0.593	46.565 ±0.444	5.504 ±0.108	$0.966 \pm 6.311 \times 10^{-4}$
DBN+CF	24.690 ±0.213	34.036 ±0.241	4.497 ±0.058	$0.983 \pm 2.275 \times 10^{-4}$
DEEM	30.462 ±0.450	43.892 ±0.385	5.159 ±0.084	$0.970 \pm 5.116 \times 10^{-4}$
DEEM+CF	**23.832** ±0.069	**33.259** ±0.109	**4.200** ±0.046	$\mathbf{0.984} \pm 1.071 \times 10^{-4}$

4.3. The Second Experiment

4.3.1. Configuration of the Forecasting Models

Similar configuration schemes are utilized in this experiment. Details will be given below.

(a) Configuration of the spectrum function in the second experiment

Figure 11(a) shows the performances of the spectrum functions which have different number of cyclic waves. According to this figure, the best spectrum function model has 26 trigonometric functions. Figure 11(b) shows the spectrum map of cyclic feature model. We can see that there are two specific cycles in the period of 2 hours and 6 hours in 26 cycles. The 26 cycles are combined to illustrate the daily BEC cyclic features in second building. Figure 11(c) shows the first 500 original BEC data from the second building. Figure 11(d) presents the remaining stochastic BEC data after removing the cyclic feature.

Besides, the ridge regression, lasso regression and multi-polynomial regression are also selected as the comparative methods of spectrum function. The configurations of these three comparative models are as the first experiment.

(b) Configuration of the DEEM

Again, the parameter searching process of the DEEM is constituted by two stages.

In the first stage, the structures of the premier and integration ELMs are fixed, while the numbers of the hidden nodes and layers of the DBN are changed. Figure 12(a) shows the MAEs of the DEEMs which have different numbers of hidden nodes and layers in the DBN. According to Figure 12(a), the MAE of the DEEM obtains the lowest value when the DBN has 3 hidden layers and 700 nodes in each hidden layer.

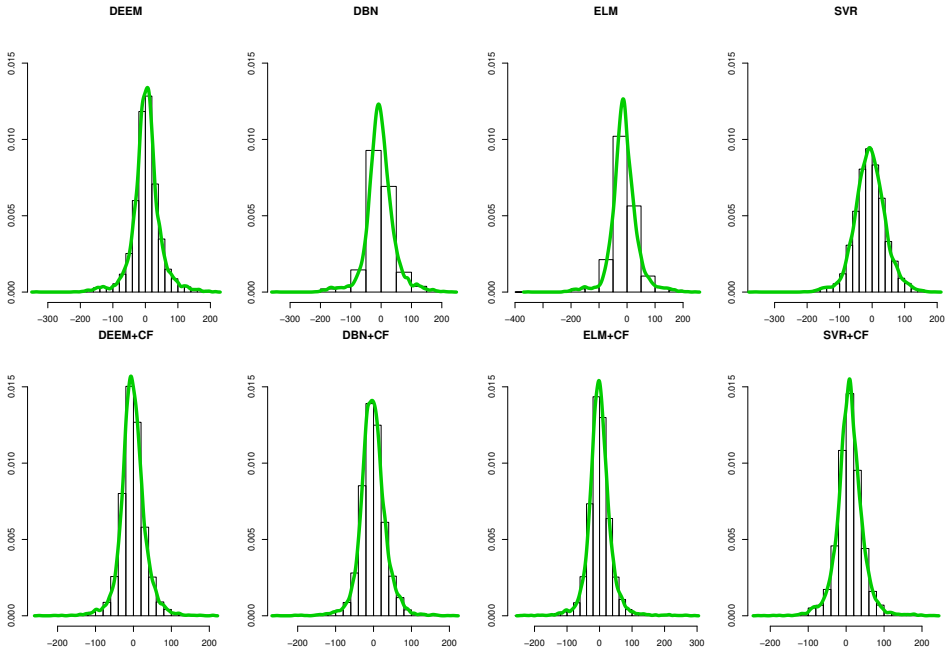


Figure 10: The error histograms of the eight forecasting models in the first experiment.

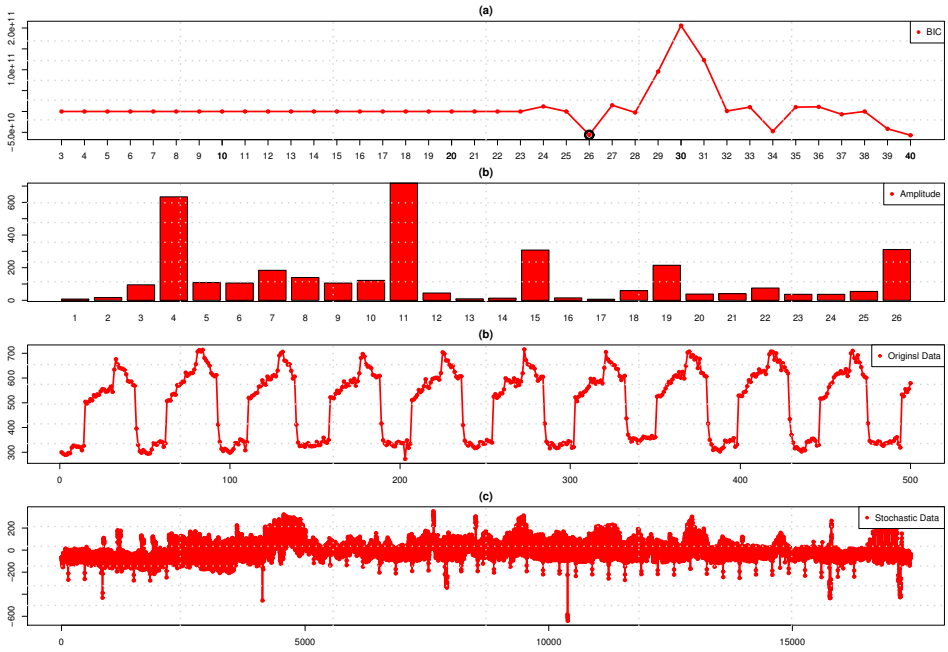
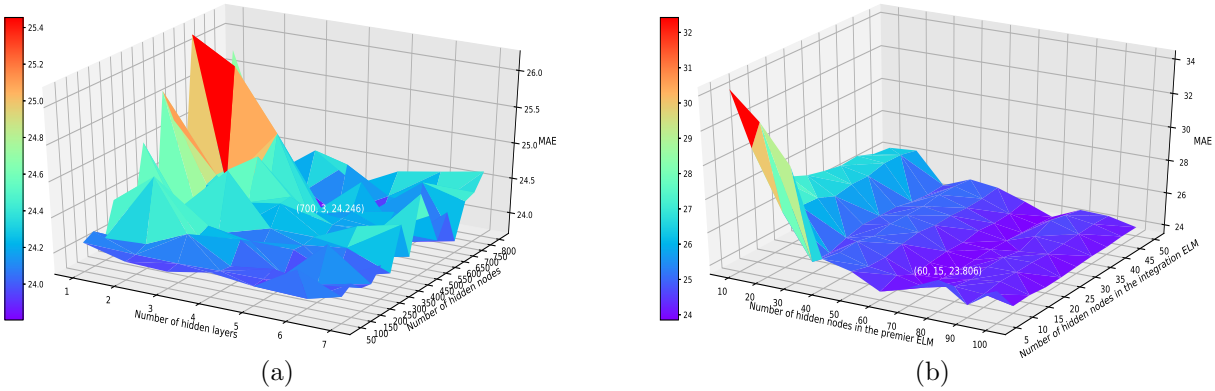
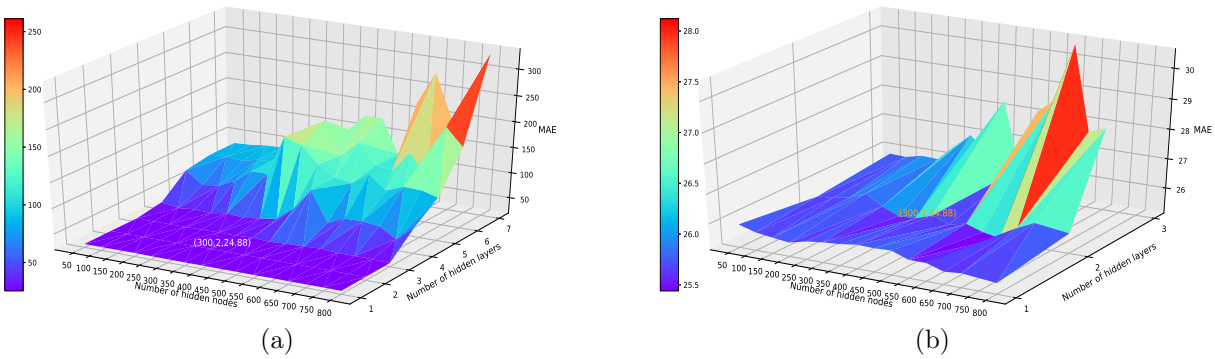


Figure 11: (a) The performance of the spectrum functions with different number of trigonometric waves in the second experiment, (b) The spectrum map of cycle features in the second experiment , (c) The first 500 original BEC data from the second building, (d) The remaining stochastic BEC data after removing the cyclic feature in the second building.



(a) (b)

Figure 12: (a) The MAEs of the DEEMs with different numbers of hidden layers and hidden nodes in DBN when the ELMs are fixed in the second experiment, (b) The MAEs of the DEEMs with different premier and integration ELMs when the DBN is fixed in the second experiment.



(a) (b)

Figure 13: (a) The MAEs of the DBNs in the second experiment, (b) Fine details of the forecasting performance in Figure 13(a).

In the second stage, the DBN in the DEEM is fixed as determined in the first stage, and, we change the numbers of the hidden nodes in the premier and integration ELMs. Figure 12(b) illustrates the forecasting performance of such DEEMs with different ELMs. From this figure, the best premier ELMs have 60 hidden nodes, and the best integration ELM have 15 hidden nodes.

(c) Configuration of the other comparative models

In the second experiment, aiming at exploring the best structure of the DBN, we evaluate the performance of different DBNs whose hidden layers and the hidden nodes in each hidden layer are respectively set from 1 to 7 and from 50 to 800 at the interval of 50. Figure 13(a) illustrates the MAEs of such DBNs. The best comparative DBN model has two hidden layers, 300 nodes in each hidden layer, and 30 hidden nodes for regression.

In order to acquire the best structure of the ELM in the second experiment, we change the number of the hidden nodes from 10 to 500 at the interval of 10. Figure 14 shows the

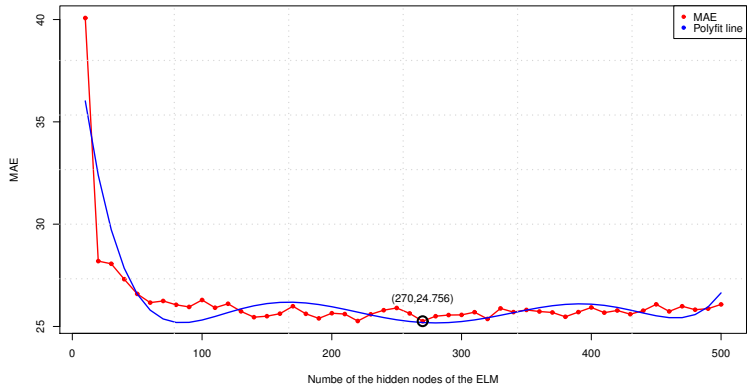


Figure 14: The MAEs of the ELMs in the second experiment.

Table 3: Performances of the forecasting performance using different cyclic feature models when ELM is selected to be the prediction model in the second experiment.

Model	MAE	RMSE	MAPE(%)	r
Lasso	27.380 ±1.143	42.007 ±0.887	5.840 ±0.271	$0.965 \pm 1.400 \times 10^{-3}$
Ridge	27.313 ±1.082	41.678 ±2.143	6.492 ±0.205	$0.973 \pm 2.670 \times 10^{-3}$
Multi-polynomial	25.146 ±1.082	38.917±1.665	5.330 ±0.342	$0.969 \pm 2.576 \times 10^{-3}$
Spectrum function	**23.343** ±0.637	**34.343** ±1.502	**4.935** ±0.346	$\mathbf{0.977} \pm 3.111 \times 10^{-3}$

MAEs of such ELMs. According to this figure, the best ELM has 270 hidden nodes.

Furthermore, the optimal structure exploration procedures for the SVR are as the first experiment. For the SVR, we also utilzie the RBF activation function, and set the penalty and kernel coefficients to be 0.7 and 0.5 respectively.

4.3.2. Experimental results

In this experiment, the prediction of ELM using different cyclic features extracted by spectrum analysis, lasso regression, ridge regression and multi-polynomial were performed for 10 times separately again. Table 3 presents the average of MAE, RMSE, MAPE, and r of the forecasting performance of the ELM using different cyclic feature models in the second experiment.

Besides, the proposed DEEM+CF, the proposed DEEM, the DBN+CF, the DBN, the ELM+CF, the ELM+CF, the ELM, the SVR+CF and the SVR were also conducted for 10 times again. The MAE, RMSE, MAPE, and r are chosen to be the comparative indices too. Table 4 lists the comparison results of these models.

The forecasting errors of these eight models are also recorded. The kernel density histograms of the forecasting errors of the eight forecasting models are displayed in Figure 15.

4.4. Comparison and Discussion

From the figures and tables above, we have the following observations and conclusions.

Table 4: Performances of the forecasting models in the second experiment.

Model	MAE	RMSE	MAPE(%)	r
SVR	38.700 ±0.000	50.790 ±0.000	8.260 ±0.000	$0.954 \pm 0.000 \times 10^{-4}$
SVR+CF	22.124 ±0.000	31.620 ±0.000	5.237 ±0.000	$0.980 \pm 0.000 \times 10^{-4}$
ELM	26.937 ±2.466	44.522 ±2.760	5.566 ±0.541	$0.962 \pm 4.930 \times 10^{-3}$
ELM+CF	22.449 ±0.975	32.413 ±1.356	4.755 ±0.243	$0.978 \pm 2.414 \times 10^{-3}$
DBN	25.219 ±0.855	42.235 ±0.360	5.241 ±0.064	$0.966 \pm 6.069 \times 10^{-4}$
DBN+CF	20.252 ±0.212	29.683 ±0.170	4.584 ±0.052	$0.982 \pm 2.275 \times 10^{-4}$
DEEM	23.793 ±0.141	41.032 ±0.228	4.875 ±0.045	$0.968 \pm 3.702 \times 10^{-4}$
DEEM+CF	**19.063** ±0.132	**28.685** ±0.122	**4.247** ±0.041	**0.983** $\pm 1.071 \times 10^{-4}$

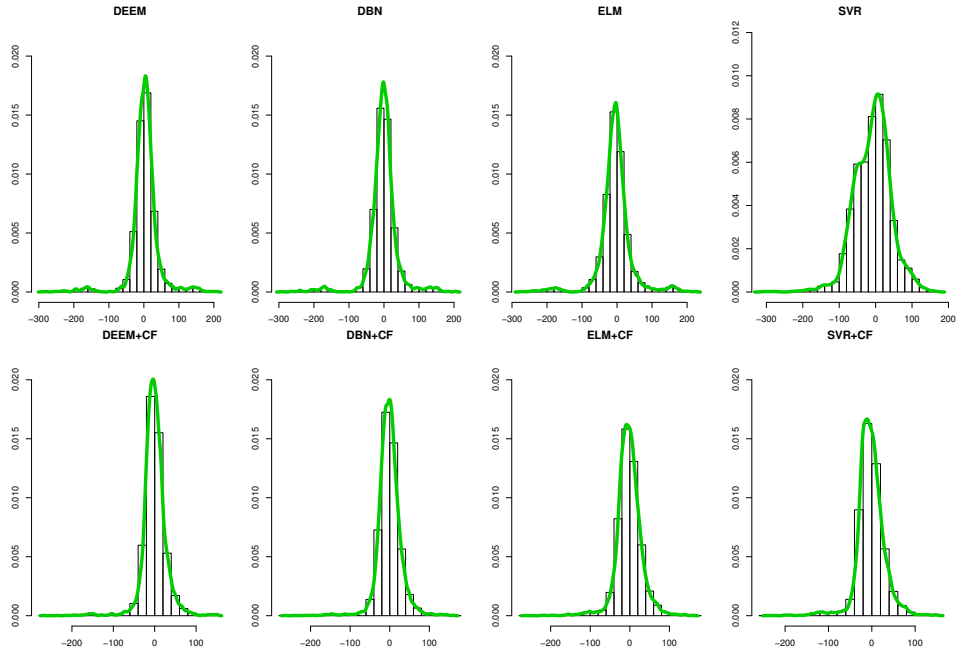


Figure 15: The error histograms of the eight forecasting models in the second experiment.

- From Figures 6 (c) (d) and 11 (c) (d), we can see that the original data has clear stable and periodic feature, but the remaining data is much more stochastic than the original data.

- From Figures 7(a), 12(a), 8 and 13, it can be clearly seen that, with the increase of the hidden layers of the DBN, the MAE value of the DEEMs has a downtrend, but the MAE of the pure DBN model increases rapidly when the number of the hidden layer is higher than a threshold. Consequently, we can conclude that the full utilization of the extracted features from different layers of the DBN can help to achieve more accurate forecasting performance.

- According to Figures 7(b) and 12(b), the number of hidden nodes in the premier and integration ELMs can influence the performance of the DEEM, and the hidden nodes

in the premier ELMs has greater influence compared with those in the integration ELM.

- Table 1 and Table 3 present that the prediction utilizing cyclic feature extracted by spectrum analysis obtain the best performance. Utilizing ridge regression, lasso regression, and multi polynomial models to extract the cyclic features can also improve the forecasting performance of BEC. From our results the multi-polynomial also performs better than the other two models.

- Table 2 and Table 4 demonstrate that the accuracy of the forecasting models considering the cyclic feature extracted by spectrum analysis is much better than those that do not combine the cyclic feature. According to the criteria of MAE, in the first experiment, the DEEM+CF, DBN+CF, ELM+CF and SVR+CF are respectively 21.765%, 23.316%, 25.167%, 27.773% better than the DEEM, DBN, ELM and SVR which don't consider the cyclic feature. In the second experiment, the DEEM+CF, DBN+CF, ELM+CF and SVR+CF are respectively 19.880%, 19.695%, 16.661%, 42.832% better. In addition, the DEEM+CF has the best accuracy, and it is 3.475%, 6.538%, 10.217% better than the DBN+CF, ELM+CF and SVR+CF respectively in the first experiment, and in the second experiment 5.871%, 15.083%, 13.836% better than these three comparative models. Furthermore, according to the standard derivations of the evaluation indices, the forecasting performance of the DEEM+CF is relatively more stable in contrast to the other comparative models except the SVR.

- From Figures 10 and 15, we can clearly observe that there exist more errors around zero in the histograms of the models that consider the cyclic feature. This also implies that the cyclic feature can improve the forecasting accuracy. Again, the error histograms of the DEEM+CF are the narrowest and highest ones which also imply the most accurate forecasting performance of this proposed model.

Overall, the prediction models, that consider the cyclic feature, have much higher forecasting accuracy than the models that are directly trained by the original data, and the proposed DEEM+CF performs more stably and accurately compared with the other models. This proves that the cyclic feature has great influence on the accuracy promotion of the BEC forecasting, and the full utilization of the abstracted features from all the layers of the DBN is also useful and helpful to improve the forecasting performance.

5. Conclusion

Short-term forecasting of BEC is helpful to the real-time building energy-demand response, the energy planning and the building management. In this paper, a novel deep belief network and extreme learning machine based ensemble method considering the cyclic feature is proposed to promote the accuracy of half hourly BEC forecasting. In the proposed ensemble model, the stable component – the cyclic feature of the BEC is extracted via the spectrum analysis, while the remaining stochastic component after removing the stable component from the original BEC data is used to construct the DEEM. Two experiments are

performed to prove the effectiveness and superiorities of the proposed DEEM+CF model. As demonstrated by the experimental results and comparisons, the cyclic feature can improve the prediction performance for about 20% better than those without the utilization of cyclic feature, and what's more, the proposed DEEM+CF model has much higher accuracy than the other comparative models, separately 3%, 6%, 10% better at least than DBN+CF, ELM+CF and SVR+CF under the criteria of MAE in our experiments.

In this study, we achieve the parameter optimization of the DEEM via fixing and changing the structures of the ELMs and the DBN alternately. However, the parameter optimization method is not the best, and it still needs further exploration. Besides, the cyclic features are greatly related to occupancy which is one of the key factors in BEC forecasting. It is valuable to study and apply the relationships between them. In the future, the relationships and their applications will be one of our key researches.

Acknowledgments

This study is partly supported by the National Natural Science Foundation of China (61573225), the Taishan Scholar Project of Shandong Province (TSQN201812092), the Key Research and Development Program of Shandong Province (2019GGX101072), the State Scholarship Fund and the Youth Innovation Technology Project of Higher School in Shandong Province (2019KJN005).

References

[1] Y. Ye, W. Zuo, G. Wang, A comprehensive review of energy-related data for U.S. commercial buildings, Energy and Buildings 186 (2019) 126–137. doi:10.1016/j.enbuild.2019.01.020.

[2] Y. Lou, W. M. Jayantha, L. Shen, Z. Liu, T. Shu, The application of low-carbon city (lcc) indicatorsa comparison between academia and practice, Sustainable Cities and Society 51 (2019) 101677. doi:https://doi.org/10.1016/j.scs.2019.101677.

[3] C. Fan, G. Huang, Y. Sun, A collaborative control optimization of grid-connected net zero energy buildings for performance improvements at building group level, Energy 164 (2018) 536 – 549. doi:https://doi.org/10.1016/j.energy.2018.09.018.

[4] P. Huang, Y. Sun, A collaborative demand control of nearly zero energy buildings in response to dynamic pricing for performance improvements at cluster level, Energy 174 (2019) 911 – 921. doi:https://doi.org/10.1016/j.energy.2019.02.192.

[5] Y. Ye, K. Hinkelman, J. Zhang, W. Zuo, G. Wang, A methodology to create prototypical building energy models for existing buildings: A case study on us religious worship buildings, Energy and Buildings 194 (2019) 351 – 365. doi:https://doi.org/10.1016/j.enbuild.2019.04.037.

[6] Y. Fu, W. Zuo, M. Wetter, J. W. VanGilder, X. Han, D. Plamondon, Equation-based object-oriented modeling and simulation for data center cooling: A case study, Energy and Buildings 186 (2019) 108–125. doi:10.1016/j.enbuild.2019.01.018.

[7] J. Chambers, P. Hollmuller, O. Bouvard, A. Schueler, J.-L. Scartezzini, E. Azar, M. K. Patel, Evaluating the electricity saving potential of electrochromic glazing for cooling and lighting at the scale of the swiss non-residential national building stock using a monte carlo model, Energy 185 (2019) 136 – 147. doi:https://doi.org/10.1016/j.energy.2019.07.037.

[8] K. Amasyali, N. M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renewable and Sustainable Energy Reviews 81 (2018) 1192–1205. doi:10.1016/j.rser.2017.04.095.

[9] Y. Zhou, S. Zheng, G. Zhang, Machine learning-based optimal design of a phase change material integrated renewable system with on-site pv, radiative cooling and hybrid ventilationsstudy of modelling and application in five climatic regions, Energy 192 (2020) 116608. doi:https://doi.org/10.1016/j.energy.2019.116608.

[10] M. Alizamir, S. Kim, O. Kisi, M. Zounemat-Kermani, A comparative study of several machine learning based non-linear regression methods in estimating solar radiation: Case studies of the usa and turkey regions, Energy (2020) 117239doi:https://doi.org/10.1016/j.energy.2020.117239.

[11] H. Liu, B. Xu, D. Lu, G. Zhang, A path planning approach for crowd evacuation in buildings based on improved artificial bee colony algorithm, Applied Soft Computing 68 (2018) 360 – 376. doi:https://doi.org/10.1016/j.asoc.2018.04.015.

[12] S. Lu, Q. Li, L. Bai, R. Wang, Performance predictions of ground source heat pump system based on random forest and back propagation neural network models, Energy Conversion and Management 197 (2019) 111864. doi:https://doi.org/10.1016/j.enconman.2019.111864.

[13] H. C. Jung, J. S. Kim, H. Heo, Prediction of building energy consumption using an improved real coded genetic algorithm based least squares support vector machine approach, Energy and Buildings 90 (2015) 76–84. doi:10.1016/j.enbuild.2014.12.029.

[14] Y. Xu, M. Zhang, L. Ye, Q. Zhu, Z. Geng, Y. He, Y. Han, A novel prediction intervals method integrating an error & self-feedback extreme learning machine with particle swarm optimization for energy consumption robust prediction, Energy 164 (2018) 137–146. doi:10.1016/j.energy.2018.08.180.

[15] Y. Huang, Y. Yuan, H. Chen, J. Wang, Y. Guo, T. Ahmad, A novel energy demand prediction strategy for residential buildings based on ensemble learning, Energy Procedia 158 (2019) 3411–3416. doi:10.1016/j.egypro.2019.01.935.

[16] Y. Lv, Y. Duan, W. Kang, Z. Li, F. Wang, Traffic flow prediction with big data: A deep learning approach, IEEE Transactions on Intelligent Transportation Systems 16 (2) (2015) 865–873. doi:10.1109/TITS.2014.2345663.

[17] X. Xue, Y. Guo, S. Chen, S. Wang, Analysis and controlling of manufacturing service ecosystem: A research framework based on the parallel system theory, IEEE Transactions on Services Computing (Early Access)doi:10.1109/TSC.2019.2917445.

[18] X. Xue, S. Wang, L. Zhang, Z. Feng, Y. Guo, Social learning evolution (sle): Computational experiment-based modeling framework of social manufacturing, IEEE Transactions on Industrial Informatics 15 (6) (2019) 3343–3355. doi:10.1109/TII.2018.2871167.

[19] X. Xue, H. Han, S. Wang, C. Qin, Computational experiment-based evaluation on context-aware o2o service recommendation, IEEE Transactions on Services Computingdoi:10.1109/TSC.2016.2638083.

[20] C. Tian, C. Li, G. Zhang, Y. Lv, Data driven parallel prediction of building energy consumption using generative adversarial nets, Energy and Buildings 186 (2019) 230–243. doi:10.1016/j.enbuild.2019.01.034.

[21] G. Fu, Deep belief network based ensemble approach for cooling load forecasting of air-conditioning system, Energy 148 (2018) 269–282. doi:10.1016/j.energy.2018.01.180.

[22] C. Li, Z. Ding, J. Yi, Y. Lv, G. Zhang, Deep belief network based hybrid model for building energy consumption prediction, Energies 11 (1) (2018) 242. doi:10.3390/en11010242.

[23] C. Fan, Y. Sun, Y. Zhao, M. Song, J. Wang, Deep learning-based feature engineering methods for improved building energy prediction, Applied Energy 240 (2019) 35 – 45. doi:https://doi.org/10.1016/j.apenergy.2019.02.052.

[24] M. Ashouri, B. C. Fung, F. Haghighat, H. Yoshino, Systematic approach to provide building occupants with feedback to reduce energy consumption, Energy 194 (2020) 116813. doi:https://doi.org/10.1016/j.energy.2019.116813.

[25] J. W. Boland, Generation of synthetic sequences of electricity demand with applications, in: Filar J., Haurie A. (eds) Uncertainty and Environmental Decision Making. International Series in Operations Research & Management Science, Springer, Boston, MA, 2010, pp. 275–314. doi:https://doi.org/10.1007/978-1-4419-1129-2_10.

[26] G. E. Hinton, A practical guide to training restricted boltzmann machines, Neural Networks: Tricks of the Trade 7700 (2012) 599–619. doi:10.1007/978-3-642-35289-8_32.

[27] T. Tieleman, G. E. Hinton, Using fast weights to improve persistent contrastive divergence, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 1033–1040. doi:http://doi.acm.org/10.1145/1553374.1553506.

[28] G. Huang, Q. Zhu, C. K. Siew, Extreme learning machine: Theory and applications, Neurocomputing 70 (1) (2006) 489–501. doi:10.1016/j.neucom.2005.12.126.

[29] A. D. Lima, L. F. Silveira, S. X. de Souza, Spectrum sensing with a parallel algorithm for cyclostationary feature extraction, Computers & Electrical Engineering 71 (2018) 151 – 161. doi:https://doi.org/10.1016/j.compeleceng.2018.07.016.

[30] J. Kostrzewa, Time series forecasting using clustering with periodic pattern, in: 2015 7th International Joint Conference on Computational Intelligence (IJCCI), Vol. 3, 2015, pp. 85–92. doi:10.5220/0005586900850092.

[31] A. I. Taiwo, T. O. Olatayo, A. F. Adedotun, Modeling and forecasting periodic time series data with fourier autoregressive model, Iraqi Journal of Science (2019) 1367–1373.

[32] Y. Zou, X. Hua, Y. Zhang, Y. Wang, Hybrid short-term freeway speed prediction methods based on periodic analysis, Canadian Journal of Civil Engineering 42 (8) (2015) 570–582. doi:10.1139/cjce-2014-0447.

[33] J. Tang, F. Liu, Y. Zou, W. Zhang, Y. Wang, An improved fuzzy neural network for traffic speed prediction considering periodic characteristic, IEEE Transactions on Intelligent Transportation Systems 18 (9) (2017) 2340–2350.

[34] W. Zhang, Y. Zou, J. J. Tang, Y. Wang, Short-term prediction of vehicle waiting queue at ferry terminal based on machine learning method, Journal of Marine Science & Technology 21 (4) (2016) 1–13. doi:10.1007/s00773-016-0385-y.

[35] R. Li, P. Jiang, H. Yang, C. Li, A novel hybrid forecasting scheme for electricity demand time series, Sustainable Cities and Society 55 (2020) 102036. doi:https://doi.org/10.1016/j.scs.2020.102036.

[36] C. Liu, K. Gryllias, A semi-supervised support vector data description-based fault detection method for rolling element bearings based on cyclic spectral analysis, Mechanical Systems and Signal Processing 140 (2020) 106682. doi:https://doi.org/10.1016/j.ymssp.2020.106682.

[37] J. Boland, Characterising seasonality of solar radiation and solar farm output, Energies 13 (2) (2020) 471. doi:https://doi.org/10.3390/en13020471.

[38] J. Boland, A. Grantham, Nonparametric conditional heteroscedastic hourly probabilistic forecasting of solar radiation, J-Multidisciplinary Scientific Journal 1 (1) (2018) 174–191. doi:https://doi.org/10.3390/j1010016.

[39] F. Al-Obeidat, B. Spencer, O. Alfandi, Consistently accurate forecasts of temperature within buildings from sensor data using ridge and lasso regression, Future Generation Computer Systemsdoi:https://doi.org/10.1016/j.future.2018.02.035.

[40] A. Satre-Meloy, Investigating structural and occupant drivers of annual residential electricity consumption using regularization in regression models, Energy 174 (2019) 148 – 168. doi:https://doi.org/10.1016/j.energy.2019.01.157.

[41] C. Fan, Y. Ding, Cooling load prediction and optimal operation of HVAC systems using a multiple nonlinear regression model, Energy and Buildings 197 (2019) 7 – 17. doi:https://doi.org/10.1016/j.enbuild.2019.05.043.