Short-Term Load Forecasting Based on VMD and Deep TCN-Based Hybrid Model with Self-Attention Mechanism

Qingliang Xiong 1, Mingping Liu 1,∗, Yuqin Li 1, Chaodan Zheng 1 and Suhui Deng 1,2

1 School of Information Engineering, Nanchang University, Nanchang 330031, China; 13723120151@163.com (Q.X.); yvqinli@163.com (Y.L.); dan629@ncu.edu.cn (C.Z.); shdeng@ncu.edu.cn (S.D.)
2 Jiangxi Provincial Key Laboratory of Interdisciplinary Science, Nanchang University, Nanchang 330031, China
∗ Correspondence: mpliu@ncu.edu.cn

Abstract: Due to difficulties with electric energy storage, balancing the supply and demand of the power grid is crucial for the stable operation of power systems. Short-term load forecasting can provide an early warning of excessive power consumption for utilities by formulating the generation, transmission and distribution of electric energy in advance. However, the nonlinear patterns and dynamics of load data still make accurate load forecasting a challenging task. To address this issue, a deep temporal convolutional network (TCN)-based hybrid model combined with variational mode decomposition (VMD) and self-attention mechanism (SAM) is proposed in this study. Firstly, VMD is used to decompose the original load data into a series of intrinsic mode components that are used to reconstruct a feature matrix combined with other external factors. Secondly, a three-layer convolutional neural network is used as a deep network to extract in-depth features between adjacent time points from the feature matrix, and then the output matrix captures the long-term temporal dependencies using the TCN. Thirdly, long short-term memory (LSTM) is utilized to enhance the extraction of temporal features, and the correlation weights of spatiotemporal features are future-adjusted dynamically using SAM to retain important features during the model training. Finally, the load forecasting results can be obtained from the fully connected layer. The effectiveness and generalization of the proposed model were validated on two real-world public datasets, ISO-NE and GEFCom2012. Experimental results indicate that the proposed model significantly improves the prediction accuracy in terms of evaluation metrics, compared with other contrast models.

Keywords: short-term load forecasting; variational modal decomposition; temporal convolutional network; long short-term memory; self-attention mechanism

1. Introduction

Power load forecasting is the foundation of the operation and planning of a power system. The generation, transmission, distribution and consumption of electricity occur almost simultaneously. It is well known that massive storage of electric power is still difficult for current technologies. Thus, power supply sides and demand sides should be balanced dynamically to ensure the safety and reliability of the power grid [1]. However, the high penetration of renewable energy sources and electric devices in modern power systems result in significant uncertainties in the load demand. Motivated by this standpoint, power load forecasting has become increasingly important in modern power systems.

Power load forecasting involves forecasting the load demand for a time span in the future, using historical load data and other external factors such as meteorological conditions, seasonal effects, economic factors and social activities. Accurate load forecasting can help power companies to make efficient decisions, avoid resource waste and improve grid stability and reliability [2]. Therefore, load forecasting has attracted increasing attention from researchers all over the world. Short-term load forecasting (STLF) is an approach that predicts power load consumption within an interval of an hour to a week, based
on historical load data and other external factors [3]. STLF provides a basis for generating units to schedule their start-up and shut-down times, prepares the rotating reserve and carries out in-depth analyses of the limitations in the transmission system [4]. It has been demonstrated that a 1% increase in STLF error increases the operating cost by GBP 17.7 million [5] or EUR 4.55 to 9.1 million [6]. In addition, the use of distributed generation and intelligent devices generates load data with more nonlinear patterns and dynamics [7]. Furthermore, the scale of the load data also sharply increases with the widespread use of smart meters [8]. Therefore, an accurate STLF model is a premise for the safe, stable and economic operation of an electric grid.

In the early years, a variety of prediction models were developed based on statistical methods and artificial intelligence methods. The statistical methods usually include autoregressive moving average [9], autoregressive integrated moving average [10], exponential smoothing [11], etc. Although these methods are easy to implement without additional feature inputs, they are not suitable for handling load data with nonlinear patterns and dynamics [12,13]. In recent years, artificial intelligence (AI) methods have been widely used in power load forecasting, along with progressive computer technologies. As a typical AI algorithm, artificial neural networks (ANNs) have strong nonlinear modeling capabilities, and can obtain any nonlinear function without having information of the relationships between the training model and data in advance [14]. However, ANNs set hyperparameters manually, and are thus unable to effectively tackle large-scale load data. Thus, it is difficult to obtain greater prediction accuracies of load data in modern power systems [15]. Moreover, the output of the training model depends not only on the current input, but also on the previous input [16]. Therefore, traditional artificial intelligence methods cannot fully extract the features of time series to achieve greater prediction accuracies.

Recently, deep learning methods have become increasingly popular in the field of time series. Long short-term memory (LSTM) developed from a recurrent neural network (RNN) can deal with long-term dependencies of time series, avoiding gradient vanishing and exploding of the RNN [17]. A convolutional neural network (CNN) can effectively capture the local relationships among adjacent time points [18] and is also widely used in many fields, such as image recognition [19], renewable energy forecasting [20], as well as load forecasting [21]. However, a large number of hyperparameters of CNN can easily result in overfitting of the training model. It should be pointed out that one-dimensional (1D) CNNs can overcome the overfitting problem due to there being fewer learnable parameters. Furthermore, the temporal convolutional network (TCN) developed from CNN has a unique dilated convolutional module to obtain a large receptive field with fewer layers [22], which is conducive to extracting nonlinear features of time series. Although the TCN has better predictive performance in time series compared with LSTM and CNN [23,24], it cannot learn the dependency of long-range positions inside the sequence and extract internal correlations of input. In addition, self-attention mechanism (SAM) focuses on capturing internal correlations of input features and can effectively deal with long time series, which helps the training model to identify key features. For instance, a novel model based on bidirectional LSTM, XGBoost and SAM was proposed for power load forecasting in [25]. An improved LSTM optimized with SAM was used to predict the concentration of air pollutants in [26]. A new model based on SAM and multi-task learning was developed to predict ultra-short-term photovoltaic power generation in [27]. These experimental results mentioned above indicate that SAM has the capability to select key information from complex features and avoid irrelevant information during model training.

Although these deep learning methods improve the prediction accuracy, single models cannot fully extract the input features, especially in-depth features. Therefore, scholars are increasingly focusing on hybrid models, which combine the advantages of each single model [28]. For example, a hybrid approach combining singular spectrum analysis-based decomposition and ANN for day-ahead hourly load forecasting was proposed in [29]. A TCN–LSTM hybrid model was used to forecast weather on meteorological data in [30].
superior performance in [31]. A hybrid model based on TCN combined with attention mechanism was proposed to fully exploit the nonlinear relationships between load data and external factors in [32]. Although the above hybrid models improved the prediction performances in various aspects, the raw load data were directly fed into the training models without being smoothed and denoised in advance, which cannot be conducive to further enhancing the accuracy of STLF.

In order to avoid the effects of nonlinear patterns and dynamics of load data on the forecasting accuracy, decomposition methods have been used to decompose the load data into multiple smoothing and stable sub-series at different frequency bands. For instance, Wang et al. [33] decomposed the load data into a series of sub-signals using wavelet decomposition (WD). The sub-signals were predicted by different models according to their frequencies. Liang et al. [34] used empirical mode decomposition (EMD) to decompose load signals into multiple intrinsic mode functions (IMFs) to weaken the nonlinearity and dynamics of the load data. Zhu et al. [35] proposed a hybrid model for carbon price forecasting based on EMD and evolutionary least squares to support vector regression. An ensemble empirical mode decomposition (EEMD) was used to decompose load data into different frequency components in [36]. The low- and high-frequency components were predicted via multivariable linear regression LSTM. However, WD cannot meet the requirements of complex and dynamic load series in time-frequency analysis [33]. In addition, EMD and EEMD have the disadvantages of mode aliasing and end effect, and lack the support of mathematical theory [37]. Then, a variational mode decomposition (VMD) algorithm with a solid mathematical model was used as an alternative to overcome their drawbacks mentioned above. VMD has good data decomposition accuracy, and obtains a group of stable sub-signals without noise interference [38]. A hybrid model based on VMD and LSTM for short-term load forecasting was proposed in [39]. A CNN and TCN hybrid model combined with VMD-based data processing was developed to predict power load forecasting in [40]. A hybrid prediction model based on VMD, TCN and an error correction strategy for electricity load forecasting was presented in [41]. A hybrid model based on TCN and VMD for short-term wind power forecasting was proposed in [42]. A hybrid model based on GRU and TCN combined with VMD decomposition was developed to predict load forecasting in [43].

Although these studies mentioned above verified that deep learning methods based on VMD techniques can improve the prediction accuracy of load demand, some fields still require further study and improvement. For example, many studies often adopted deep learning models with shallow networks to improve the training efficiency, and they rarely considered using the models with deep networks to learn full features from the decomposed sub-series. Moreover, most studies scarcely focused on the correlations of sub-series and external factors, such as temperature, seasons, holidays, etc. Deep learning models with multiple hidden layers can indeed improve their feature extraction capabilities, but it is difficult to train these models with some sub-series and external factors. Therefore, it is necessary to construct a deep learning model with shallow networks to extract in-depth features from several sub-series and external factors. As a result, a novel STLF model based on VMD and a deep TCN-based hybrid method with SAM is proposed to fully capture the in-depth features of multiple sub-series and external factors. The main advantages of this research are as follows:

(1) The raw load data are decomposed into multiple stable sub-series using VMD. Furthermore, the external factors are also considered as input variables and reconstructed as a feature matrix, along with the sub-series for the training model.

(2) A three-layer 1D-CNN network is constructed as a deep network to eliminate overfitting of the training model due to fewer hyperparameters, and reshapes the extracted features into time series for the TCN module. The TCN extracts the long-term temporal dependencies of the input matrix. Moreover, LSTM is used to further enhance the extraction of temporal features.
SAM can amplify important information, and then weaken irrelevant information of the feature matrix. Thus, SAM dynamically adjusts the correlation weight of complex features to obtain important information from the input data.

Compared with traditional models or other benchmarking models, the novel hybrid model performs more effectively and achieves greater prediction accuracies.

The rest of this study is organized as follows. Section 2 introduces the basic methodologies of VMD, 1D-CNN, TCN, LSTM, SAM, as well as the framework of the proposed hybrid model. Section 3 describes the data preprocessing, evaluation criteria and experimental analysis. Finally, conclusion and future research are drawn in Section 4.

2. Methodologies

In this section, an STLF integration framework will be introduced in detail, including VMD, TCN, 1D-CNN, LSTM and SAM. The following sections briefly describe each of the modules used in this study. Finally, the architecture of the proposed hybrid model will be presented, and its main hyperparameters are also listed.

2.1. VMD

VMD is an adaptive and fully non-recursive method with modal change and signal processing [44], which decomposes time series data into multiple IMFs. It can not only retain the inherent information of original signals, but also avoids the overlap of variable information. Moreover, VMD has superior performance in sampling and denoising compared to WD, EMD and EEMD [45]. The process of decomposition using VMD can be described as follows:

Step 1: The analytic signal \( f(t) \) can be transformed by the Hilbert theory into a unilateral frequency spectrum for a given mode, as follows:

\[
\text{Hilbert}[f(t)] = \frac{1}{\pi} \int_{-\infty}^{+\infty} f(v) \frac{t}{t-v} dv
\]  

(1)

Step 2: Each mode is added by an exponential term, and its frequency spectrum is shifted to the baseband as follows:

\[
\left[ \left( \delta(t) + \frac{j}{\pi t} \right) u_k(t) \right] e^{-j\omega_k t}
\]  

(2)

where \( \delta(t) \) is the Dirac distribution, \( u_k \) is the \( k \)th mode, \( \omega \) is the angular frequency and \( j \) is an imaginary unit.

Step 3: The Gaussian smoothness, i.e., the square norm of the gradient, is used to calculate the bandwidth of the signal. Then, the constrained variational expression can be written as follows:

\[
\min \{ u_k \}, \{ \omega_k \} \left\{ \sum_k \left\| \frac{\partial}{\partial t} \left[ \left( \delta(t) + \frac{j}{\pi t} \right) u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \text{s.t.} \sum_k u_k = f \}
\]  

(3)

where \( \omega_k \) is the center frequency of the \( k \)th mode.

Step 4: Lagrange multiplier theory is an effective method to enforce constraints. For the convenience of analysis, the constrained variational problem needs to be transformed into an unconstrained one, as shown below:

\[
\mathcal{L}(\{ u_k \}, \{ \omega_k \}, \lambda) := a \sum_k \left\| \frac{\partial}{\partial t} \left[ \left( \delta(t) + \frac{j}{\pi t} \right) u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \| f(t) - \sum_k u_k(t) \|_2^2 + \langle \lambda(t), f(t) - \sum_k u_k(t) \rangle
\]  

(4)

where \( a \) is a quadratic penalty factor and \( \lambda \) is the Lagrange multiplier.
Step 5: It is easy to achieve the saddle point of Equation (4) using the alternate direction method of multipliers to further achieve the frequency domain update of each mode. Thus, the expressions of modal component and center frequency are given as follows:

$$\hat{u}_{n+1}^k(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k} \hat{u}_i(\omega) + \lambda(\omega)}{1 + 2\alpha(\omega - \omega_k)^2}$$

(5)

$$\omega_{n+1}^k = \frac{\int_0^\infty \omega |\hat{u}_k(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k(\omega)|^2 d\omega}$$

(6)

where $n$ represents the number of iterations. Additionally, Lagrange multiplier $\lambda$ should be satisfied with the following expression:

$$\hat{\lambda}_{n+1}(\omega) \leftarrow \hat{\lambda}_n(\omega) + \tau \left( \hat{f}(\omega) - \sum_k \hat{u}_{n+1}^k(\omega) \right)$$

(7)

Step 6: Solve the saddle point of Equation (2) repeatedly until the modal component satisfies the following convergence criterion:

$$\sum_k \left\| \hat{u}_{n+1}^k - \hat{u}_k^n \right\|_2^2 / \left\| \hat{u}_k^n \right\|_2^2 < \varepsilon$$

(8)

where $\varepsilon$ is the convergence tolerance that determines the accuracy and iteration times.

2.2. 1D-CNN

The 1D-CNN is often used in signal processing and time series models [46]. It has the ability of translation and scaling due to the local links and weight sharing. Compared with multi-dimensional CNNs, the 1D-CNN can change the number of channels with retaining the feature size to realize dimensional reduction. Furthermore, it deepens the network structure and introduces more nonlinear calculations without increasing the receptive field. The 1D-CNN is calculated as follows:

$$y_t = \sum_{k=1}^K Z_k X_{t-k+1}$$

(9)

where $Z_k$ is the convolution kernel and $X_{t-k+1}$ is the time series.

2.3. TCN

The TCN has been developed on the basis of the 1D-CNN, and has shown excellent performance in many fields of time series processing [47]. It includes causal convolution, dilated convolution and residual block, features which are briefly introduced in the following.

2.3.1. Causal Convolution

Causal convolution has the capability to overcome the disclosure of future information. It means that an output $y_t$ at moment $t$ only depends on the input at previous moments, i.e., $(x_t, x_{t-1}, x_{t-2}, x_{t-3}, \ldots)$. This just reflects the strong causal relationship of time series [48]. The calculation process of causal convolution is presented in Figure 1.
2.3.2. Dilated Convolution

The expansion of the receptive field leads to an increase in the hyperparameters, and then cause gradient vanishing and exploding. Dilated convolution can address this problem [49]. According to Figure 2, the formula for dilated convolution is listed in the following.

For an input series $X = [x_0, \ldots x_N]$ and a filter $f$ with size $k$, the dilated convolution operation $F$ of the sequence element $N$ can be written as follows:

$$F(N) = (X_{\text{dil}}f)(N) = \sum_{i=0}^{k-1} f(i) \times X_{N-d \cdot i}$$  \hspace{1cm} (10)

where $k$ is the filter size, $d$ is the dilation factor and $d \cdot i$ indicates the orientation of the convolution. An increase in the dilation factor $d$ enables the expansion of the receptive field without increasing the computational cost, which significantly improves the training efficiency.

2.3.3. Residual Block

Figure 3 illustrates the structure of residual block [50]. It is obvious that the residual block includes two layers of dilated causal convolution and nonlinear units. Thus, the rectified linear unit (ReLU) is used as an activation function, and regularizes after each dilated convolution. A $1 \times 1$ convolution connects back to the input to make the input and output compatible. This suggests that the stacked residual blocks not only improve the training efficiency of the network, but also avoid gradient vanishing. Moreover, it can continuously update the residual characteristics to improve the transmission efficiency of the feature information.
The unique gates system is introduced in the LSTM model to control the information flow. Firstly, the input gate $i_t$ is used to decide how much the memory cell state will be updated by the block input. Secondly, what information from the previous cell state should be forgotten is decided by the forget gate $f_t$. Finally, the output gate $o_t$ controls which part of the current memory cell state should be output.

The formulas of the control parameters can be written as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$  \hspace{1cm} (11)

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f)$$  \hspace{1cm} (12)

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i)$$  \hspace{1cm} (13)

$$\tilde{c} = \tanh(W_c \times [h_{t-1}, x_t] + b_c)$$  \hspace{1cm} (14)
\[
\text{tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (15)
\]

\[
c_t = f_t c_{t-1} + i_t \tilde{c} \quad (16)
\]

\[
o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \quad (17)
\]

\[
h_t = o_t \times \text{tanh}(c_t) \quad (18)
\]

where \(x\) is the input variable, \(x_t\) and \(h_t\) are the input and output at time step \(t\), respectively, \(W\) and \(b\) are weight matrices and bias matrices, respectively, \(\sigma(\cdot)\) and \(\text{tanh}(\cdot)\) are two activation functions for the update of cell state and the selective output, respectively, and \(c_t\) and \(\tilde{c}\) denote the current values and new candidate values for the cell state, respectively.

2.5. SAM

SAM pays attention to the internal correlations of input features and assigns different weights according to their importance [52]. Firstly, the input sequence \(X\) is linearly transformed into the query (\(Q\)), key (\(K\)) and value (\(V\)) using three different weight matrices \(W^Q\), \(W^K\) and \(W^V\), respectively. Secondly, the similarity between \(Q\) and \(K\) can be calculated, and is further normalized by a softmax function to achieve the self-attention matrix (\(W\)). Finally, the output matrix \(H\) can be obtained by multiplying the matrix \(W\) with \(V\). SAM can dynamically adjust the weights of different features and obtain the long-range dependencies of the sequences. Figure 5 depicts the structure of SAM, and the process of SAM is formulated as follows:

\[
\begin{align*}
Q &= W^Q X \\
K &= W^K X \\
V &= W^V X
\end{align*} \quad (19)
\]

\[
W = \text{Softmax} \left( \frac{QK^T}{\sqrt{D_K}} \right) \quad (20)
\]

\[
H = \text{Attention}(Q, K, V) = WV = V \text{Softmax} \left( \frac{K^T Q}{\sqrt{D_K}} \right) \quad (21)
\]

where \(D_K\) is the dimension of key \(K\) and Softmax (\(\cdot\)) is the function of normalization by column.

Figure 5. The structure of SAM.
2.6. Proposed Model

The framework of the proposed model is illustrated in Figure 6, and mainly includes three parts, i.e., feature engineering, feature extraction and load forecasting. In the feature engineering stage, the load data are decomposed into multiple sub-series by VMD. A new feature matrix is constructed with these sub-series and external factors. In the feature extraction stage, the 1D-CNN network is used as a deep network to extract in-depth spatial features, and the TCN extracts the long-term temporal dependencies of the input matrix. Furthermore, the temporal features hidden in load series are further extracted using LSTM. Then, SAM can dynamically adjust the weight of spatiotemporal features to obtain an important feature matrix. Finally, load forecasting can be obtained through the fully connected layer. Table 1 lists the hyperparameters of each algorithm used in the proposed model. Moreover, the steps of the proposed model are described in the following:

![Flow chart of the proposed model](image)

**Figure 6.** Flow chart of the proposed model.

Step 1: VMD is employed to decompose the raw load data into eight IMFs. These IMFs are normalized along with temperature, while seasons, holidays and weekends are processed by using a one-hot encoder. The feature matrix can be reconstructed by combining IMFs and external factors in parallel.

Step 2: The three-layer 1D-CNN is used as a deep network to extract deep spatial features between adjacent time points, and then the TCN is utilized to globally capture temporal features from the load data.

Step 3: LSTM can further enhance the extraction of long-term dependencies. Furthermore, the SAM dynamically adjusts the correlations of different features, strengthens long-range dependencies and obtains the important spatiotemporal features.

Step 4: The feature matrix processed from SAM is reshaped into time series that are used to predict load data through the fully connected layer.
Table 1. The hyperparameters of each algorithm used in the proposed model.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Dropout</th>
<th>Convolutional Kernel</th>
<th>Dilation</th>
<th>Output Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st 1D-CNN</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>[128, 72, 17]</td>
</tr>
<tr>
<td>2nd 1D-CNN</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>[128, 84, 14]</td>
</tr>
<tr>
<td>3rd 1D-CNN</td>
<td>-</td>
<td>5</td>
<td>-</td>
<td>[128, 42, 10]</td>
</tr>
<tr>
<td>TCN</td>
<td>0.2</td>
<td>2</td>
<td>[1, 2, 4, 8, 16]</td>
<td>[128, 20, 10]</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.2</td>
<td>-</td>
<td>-</td>
<td>[128, 20, 16]</td>
</tr>
<tr>
<td>SAM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>[128, 20, 16]</td>
</tr>
<tr>
<td>Reshape</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>[128, 256, -]</td>
</tr>
<tr>
<td>Linear</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>[128, 160, -]</td>
</tr>
<tr>
<td>Linear1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>[128, 1, - , - ]</td>
</tr>
</tbody>
</table>

3. Experiments and Results Analysis

In this section, two real-world datasets collected from different regions were adopted in this study to verify the effectiveness of the proposed model. The two datasets are described in the following.

The first dataset is the ISO-NE (New England) public dataset [53], which includes load data, temperatures, as well as day types from 1 March 2003 to 31 December 2014. The dataset was collected every hour, and the data from 1 March 2003 to 8 October 2007 were selected for this study. The second selected dataset is the GEFCom2012 public dataset [54], including power load and temperature. The dataset was collected every hour, and a total of 38,065 sets of data from 1 January 2004 to 29 June 2008 were selected as the study sample. The two datasets were divided into a training set, a validation set and a test set at a ratio of 8:1:1.

Experiments on the two datasets were performed on a PC with AMD-4800H CPU and NVIDIA GeForce RTX GPU with 16 GB of memory. The software platform used was Python 3.7 based on Pytorch 1.7.

3.1. Data Preprocessing

In this subsection, the ISO-NE dataset was chosen as a typical case to elaborate the characteristics of load data in detail. In addition, the GEFCom2012 dataset, which is similar to the ISO-NE dataset, will no longer be discussed.

The first step was to determine the number of decomposition models \( k \), the penalty factor \( \alpha \) and the updating step size \( \tau \) using the central frequency observation and the residual index minimization method [55]. In this research, it was convenient to obtain these parameters, i.e., \( k = 8, \alpha = 419 \) and \( \tau = 0.19 \). Figure 7 shows the decomposition results of load data using VMD. Firstly, the changing trend of the low-frequency component is smoother than that of the original signal, and is roughly consistent with the latter. It indicates that the noise interference would be effectively removed by VMD. Secondly, the influence of seasonal factors on load forecasting needed to be considered in the data preprocessing due to the smoothing and stability of the low-frequency component. Finally, all IMF components showed good periodicity and stability without irregular information, which facilitated the extraction of features in the subsequent stage.

Changing trends in electric demand usually present continuous and periodic characteristics without breaking. However, the power consumption is easily influenced by external factors such as seasons, temperature, day types, etc. This results in complex patterns and dynamics in the load data [56]. The relationship between power load and temperature is illustrated in Figure 8, where it shows the distributions of power load and temperature from 1 March 2003 to 8 October 2007. One can clearly see that the power load demand will reach a peak as the temperature becomes very high or low. Thus, temperature has a strong correlation with power load data. Moreover, Figure 9 presents the evolutions of electric consumption in different time horizons. The time horizons of Figure 9a–c are one year from 1 January 2004 to 1 January 2005, one week from 7 to 13 April 2003, and the Christmas period from 22 to 28 December 2003, respectively. One can see from Figure 9a that the power
consumption is higher in January, June, July, August and December due to the summer and winter climates. Additionally, the power consumption on weekdays slightly fluctuates and is higher than that on weekends, as shown in Figure 9b. As an important holiday, one can see from Figure 9c that the power consumption during Christmas is significantly lower than at other times. Therefore, the day types such as weekdays and holidays also have important influences on prediction accuracy of the power load.

![Figure 7](image.png)

**Figure 7.** Decomposition results of the load data using VMD for the ISO-NE dataset.

![Figure 8](image.png)

**Figure 8.** Evolutions in power load and temperature from 1 March 2003 to 8 October 2007 for ISO-NE dataset.

According to the discussions mentioned above, this study selected power demand, holidays, quarters, temperature and weekends as the input variables for load forecasting. One-hot encoding was performed on quarters, holidays, weekends and weekdays. However, a min–max normalization was used to normalize the power load and temperature over a range of (0, 1). The formula of min–max normalization is shown as follows:

$$
\hat{x} = \frac{x - \min(X)}{\max(X) - \min(X)}
$$

(22)

where \(X\) is the input series, \(x\) indicates the data to be normalized, \(\hat{x}\) is the normalized data and \(\max(\cdot)\) and \(\min(\cdot)\) represent the maximum and minimum of the input series, respectively.
3.2. Evaluation Criteria

To evaluate the performance of the proposed model, the mean absolute percentage error (MAPE), root mean square error (RMSE) and R-squared ($R^2$) were selected as evaluation indices. These statistical metrics are defined as follows:

\[
\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right|
\]

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}}
\]

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - y)^2}
\]

where $N$ is the number of samples, and $y_i$ and $\hat{y}_i$ represent the actual value and the predicted value, respectively.

3.3. Comparative Analysis of Experimental Results

Generally, the performances of single models, such as TCN, LSTM and CNN, are inferior to those of hybrid models. Moreover, the TCN-based hybrid models in STLF are comparatively inferior to other hybrid models. Thus, the experimental results of the proposed model were mainly compared with other TCN-based hybrid models with VMD decomposition that include VMD–CNN–TCN–SAM (VCTA), VMD–TCN–LSTM–SAM (VTLA), VMD–CNN–LSTM–SAM (VCLA), VMD–CNN–TCN–LSTM (VCTL) and VMD–CNN–TCN–GRU–SAM (VCTGA). It must be stressed that the CNN in these hybrid models means a three-layer 1D-CNN network. Furthermore, in order to overcome the randomness of the AI algorithms, all of the models mentioned above were trained and tested enough times until the results became stable and reliable.

3.3.1. ISO-NE Dataset

In order to further elaborate its performance for load forecasting, the proposed model was mainly compared to the hybrid models with VMD decomposition. Tables 2–4 show
the maximum, minimum and average values of the MAPE, RMSE and $R^2$, respectively. Compared with the VCLA, one can find that the VTLA reduces the MAPE by 13.6% and the RMSE by 10.4%, due to the large receptive field of TCN. The VCTA model is conducive to the extraction of long-term in-depth features, and improves the performance of load forecasting. For example, the VCTA greatly decreases the MAPE by 48.6% and the RMSE by 48% compared to the VTLA. When the VCTA is further stacked with GRU, i.e., VCTGA, one finds that its MAPE and RMSE can further be reduced by 20.8% and 30.8%, respectively. It means that the GRU can enhance the extraction of long-term dependencies. However, the proposed model even outperforms the VCTGA by 17.5% and 7% in terms of the MAPE and RMSE, respectively. This again demonstrates that LSTM has a stronger capability to capture temporal dependencies compared with GRU. Compared with VCTL, one can see that SAM is very important to globally enhance the key features for the proposed model. For example, compared with the VCTL, the proposed model significantly decreases the MAPE by 39% and the RMSE by 49%. Therefore, SAM is extremely important for globally adjusting and retaining key features to improve the superior performance of the proposed model. Additionally, the $R^2$ of the proposed model increases from 0.04% to 0.02% compared with that of the contrast models and reaches up to 99.9%, which infers that VDM decomposition is indeed effective in improving prediction accuracy, and that the proposed model results can be trusted.

### Table 2. Maximum, minimum and average values of MAPE for ISO-NE dataset.

<table>
<thead>
<tr>
<th>Models</th>
<th>VCLA</th>
<th>VTLA</th>
<th>VCTA</th>
<th>VCTL</th>
<th>VCTGA</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max MAPE (%)</td>
<td>2.08</td>
<td>1.97</td>
<td>1.11</td>
<td>1.28</td>
<td>0.71</td>
<td>0.56</td>
</tr>
<tr>
<td>Min MAPE (%)</td>
<td>1.23</td>
<td>1.07</td>
<td>0.53</td>
<td>0.56</td>
<td>0.45</td>
<td>0.41</td>
</tr>
<tr>
<td>Average MAPE (%)</td>
<td>1.62</td>
<td>1.40</td>
<td>0.72</td>
<td>0.77</td>
<td>0.57</td>
<td>0.47</td>
</tr>
</tbody>
</table>

### Table 3. Maximum, minimum and average values of RMSE for ISO-NE dataset.

<table>
<thead>
<tr>
<th>Models</th>
<th>VCLA</th>
<th>VTLA</th>
<th>VCTA</th>
<th>VCTL</th>
<th>VCTGA</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max RMSE(MW)</td>
<td>387.17</td>
<td>398.52</td>
<td>210.26</td>
<td>303.41</td>
<td>136.25</td>
<td>116.51</td>
</tr>
<tr>
<td>Min RMSE(MW)</td>
<td>305.54</td>
<td>213.04</td>
<td>118.88</td>
<td>122.25</td>
<td>88.22</td>
<td>80.46</td>
</tr>
<tr>
<td>Average RMSE(MW)</td>
<td>344.87</td>
<td>308.32</td>
<td>159.65</td>
<td>201.29</td>
<td>110.17</td>
<td>102.23</td>
</tr>
</tbody>
</table>

### Table 4. Maximum, minimum and average values of $R^2$ for ISO-NE dataset.

<table>
<thead>
<tr>
<th>Models</th>
<th>VCLA</th>
<th>VTLA</th>
<th>VCTA</th>
<th>VCTL</th>
<th>VCTGA</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max $R^2$ (%)</td>
<td>99.31</td>
<td>99.53</td>
<td>99.88</td>
<td>99.87</td>
<td>99.91</td>
<td>99.93</td>
</tr>
<tr>
<td>Min $R^2$ (%)</td>
<td>97.90</td>
<td>98.27</td>
<td>99.51</td>
<td>99.32</td>
<td>99.80</td>
<td>99.87</td>
</tr>
<tr>
<td>Average $R^2$ (%)</td>
<td>98.88</td>
<td>99.20</td>
<td>99.75</td>
<td>99.73</td>
<td>99.86</td>
<td>99.90</td>
</tr>
</tbody>
</table>

To further see the deviations between each model and the actual load data, Figure 10 shows the predicted load over 24 h and the actual load data when the MAPE is a maximum and a minimum, respectively. One can find that the VMD–CNN–TCN (VCT)-based hybrid models approximately capture the changing trends in the actual load. However, the VCLA and VTLA models obviously deviate from the actual load. This means that the stacked model based on CNN and TCN plays a critical role in feature extraction for the training model. Similarly, load forecasting during the peak or valley areas is still difficult for the models to accurately predict. However, it can be seen from the subplots that the hybrid model proposed in this article achieves the smallest deviation compared with other VCT-based hybrid models, especially during the turning area. Furthermore, Figure 11 shows bar charts of the RMSE for each model. One can clearly see that the VCTA-based hybrid models are more stable compared with other models, and the proposed model has the best stability among all of these models. At the same time, SAM is also important to improve the stability of the VCT-based models.
14.5% in terms of the MAPE and RMSE, respectively, which further proves that LSTM can
demonstrate the effectiveness of VDM decomposition; moreover, the R
compared with the VCTL. At the same time, the R
effectively extract the in-depth features from the load series. Furthermore, the proposed

3.3.2. GEFCom2012 Dataset

To further validate the generalizations of the proposed model, all of the contrast
experiments mentioned above were performed using the GEFCom2012 dataset. Tables 5–7
show the maximum, minimum and average values of the MAPE, RMSE and R², respectively. One can see that compared with VCLA, the VCTA model significantly reduces the MAPE
by 45.7% and the RMSE by 30%. This again demonstrates that the VCT-based models
effectively extract the in-depth features from the load series. Furthermore, the proposed
model that is stacked with LSTM based on the VCTA outperforms the latter by 12.7% and
14.5% in terms of the MAPE and RMSE, respectively, which further proves that LSTM can
enhance the capture of temporal dependencies. Additionally, the influence of SAM on the
extraction of important features from the load data can also demonstrated. One can find
that the proposed model further decreases the MAPE by 15.2% and the RMSE by 12.2%
compared with the VCTL. At the same time, the R² results of all of the models are over 99%,
which demonstrates the effectiveness of VDM decomposition; moreover, the R² value of the
proposed model is the highest among these models. Therefore, these above analyses show
that the proposed model has good generalization capabilities in power load forecasting.
Table 5. Maximum, minimum and average values of MAPE for GEFCom2012 dataset.

<table>
<thead>
<tr>
<th>Models</th>
<th>VCLA</th>
<th>VTLA</th>
<th>VCTA</th>
<th>VCTL</th>
<th>VCTGA</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max MAPE (%)</td>
<td>1.93</td>
<td>1.33</td>
<td>1.11</td>
<td>1.13</td>
<td>1.03</td>
<td>0.91</td>
</tr>
<tr>
<td>Min MAPE (%)</td>
<td>1.12</td>
<td>1.02</td>
<td>0.93</td>
<td>0.98</td>
<td>0.91</td>
<td>0.85</td>
</tr>
<tr>
<td>Average MAPE (%)</td>
<td>1.50</td>
<td>1.18</td>
<td>1.02</td>
<td>1.05</td>
<td>0.96</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Table 6. Maximum, minimum and average values of RMSE for GEFCom2012 dataset.

<table>
<thead>
<tr>
<th>Models</th>
<th>VCLA</th>
<th>VTLA</th>
<th>VCTA</th>
<th>VCTL</th>
<th>VCTGA</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max RMSE (MW)</td>
<td>412.50</td>
<td>437.96</td>
<td>278.36</td>
<td>273.77</td>
<td>257.04</td>
<td>236.49</td>
</tr>
<tr>
<td>Min RMSE (MW)</td>
<td>348.79</td>
<td>253.50</td>
<td>261.19</td>
<td>243.01</td>
<td>229.45</td>
<td>217.92</td>
</tr>
<tr>
<td>Average RMSE (MW)</td>
<td>378.57</td>
<td>325.29</td>
<td>264.35</td>
<td>256.76</td>
<td>241.00</td>
<td>225.32</td>
</tr>
</tbody>
</table>

Table 7. Maximum, minimum and average values of $R^2$ for GEFCom2012 dataset.

<table>
<thead>
<tr>
<th>Models</th>
<th>VCLA</th>
<th>VTLA</th>
<th>VCTA</th>
<th>VCTL</th>
<th>VCTGA</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max $R^2$ (%)</td>
<td>99.73</td>
<td>99.76</td>
<td>99.83</td>
<td>99.77</td>
<td>99.85</td>
<td>99.85</td>
</tr>
<tr>
<td>Min $R^2$ (%)</td>
<td>99.14</td>
<td>99.59</td>
<td>99.75</td>
<td>99.51</td>
<td>99.78</td>
<td>99.82</td>
</tr>
<tr>
<td>Average $R^2$ (%)</td>
<td>99.42</td>
<td>99.69</td>
<td>99.79</td>
<td>99.71</td>
<td>99.81</td>
<td>99.83</td>
</tr>
</tbody>
</table>

Figure 12 shows the predicted load over 24 h and the actual load data when the MAPE is a maximum and a minimum, respectively. As shown in Figure 12, all of the models are in accord with the trends in the actual load during the rising or falling stages. However, the proposed model is better able to capture trends in the actual load compared with the deviations of the other models, especially during the turning area. Furthermore, in comparing Figures 10 and 12, one can see that the changing trends of the GEFCom2012 dataset have more peaks and valleys, which proves that the GEFCom2012 dataset has more complexity and volatility compared to the ISO-NE dataset. Thus, the evaluation indices of the load forecasting for all of the models in the GEFCom2012 dataset are higher than those of the corresponding models with the ISO-NE dataset. In addition, Figure 13 shows the maximum, minimum and average values of the RMSE for all of the models with the GEFCom2012 dataset. It can be seen that although the GEFCom2012 dataset has obviously nonlinear and volatile characteristics, the proposed model still achieves the best reliability and the highest prediction accuracy and stability.
Figure 13. Maximum, minimum and average values of RMSE for each model based on GEFCom2012 dataset.

3.3.3. Discussion

After VMD decomposition of the two datasets, the predictive results of all of the VMD-based models in this research became stable, which infers the importance of smoothing and stabilizing the load data for accurate prediction. Moreover, the combination of TCN and LSTM is helpful to fully extract the long-term temporal dependencies and local dependencies of the load data. Due to stacking multiple single modules, there is a lot of irrelevant feature information that is extracted from the load data. Then, SAM is used to dynamically adjust the correlation weights of different important features to obtain the key formation. Therefore, the novel architecture of the proposed model is very beneficial for feature extraction from large-scale load data with nonlinear patterns and dynamics. It should be pointed out that the computational cost of the proposed model increases greatly compared to other single models, due to its complicated structure. However, the computational cost problem can easily be overcome with the rapid development of computing power. In addition, all of the models were trained and tested several times in this study to obtain a series of stable results. Figure 14 shows a boxplot of the MAPE for all of the models based on the ISO-NE and GEFCom2012 datasets. One can see that the proposed model not only achieves the lowest MAPE value among all of the models for the two datasets, but it also achieved the smallest variation range in its MAPE value. This indicates that the novel model proposed in this study has excellent generalization and reliable prediction accuracy.

Figure 14. Boxplots of MAPE for each model based on (a) ISO-NE dataset and (b) GEFCom2012 dataset.

3.3.4. Comparison with State-of-the-Art Models

It is necessary to compare the proposed model with other state-of-the-art hybrid models, all of which are TCN-based or data decomposition models published in recent years. Table 8 lists the comparative results of the proposed model and other baseline models. It is obvious that the proposed model significantly outperforms state-of-the-art models.
in terms of the MAPE. Even if the same datasets are used in [14,54] for one-hour-ahead forecasting, the MAPE values of the proposed model are significantly reduced by 37% and 36%, respectively. It should be emphasized that [14] used an EMD decomposition method, and achieved high prediction accuracy. However, compared with the results of [14], the proposed model in this research reduces the MAPE by 37% and the RMSE by 39%. This also demonstrates that VMD decomposition outperforms EMD decomposition in terms of data smoothing and linearization. Furthermore, the hybrid models with data decomposition all achieved high accuracies in load forecasting, which proves that the smoothing and stabilizing of load data is a very important process for load forecasting. It should be noted that the RMSE values of some comparative models are lower than that of the proposed model due to their simpler structures and fewer hyperparameters. Overall, the proposed model can extract in-depth spatiotemporal features, and then globally enhance key features to achieve reliable load forecasting with greater accuracy.

<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>Method</th>
<th>Dataset</th>
<th>Forecast Horizon</th>
<th>MAPE (%)</th>
<th>RMSE (MW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>He et al. [31]</td>
<td>2021</td>
<td>CNN–TCN</td>
<td>Guangzhou, China</td>
<td>24 h</td>
<td>2.39</td>
<td>378.29</td>
</tr>
<tr>
<td>Tang et al. [32]</td>
<td>2022</td>
<td>SAM–TCN</td>
<td>Chongqing, China</td>
<td>1 h</td>
<td>5.35</td>
<td>52.14</td>
</tr>
<tr>
<td>Liu et al. [16]</td>
<td>2022</td>
<td>EMD–CNN–TCN–SAM–LSTM</td>
<td>ISO-NE</td>
<td>1 h</td>
<td>0.75</td>
<td>168.19</td>
</tr>
<tr>
<td>He et al. [57]</td>
<td>2017</td>
<td>Parallel CNN–RNN</td>
<td>GEFCom2012</td>
<td>1 h</td>
<td>1.405</td>
<td>-</td>
</tr>
<tr>
<td>Xie et al. [58]</td>
<td>2018</td>
<td>RH–ANN</td>
<td>GEFCom2012</td>
<td>24 h</td>
<td>4.90</td>
<td>-</td>
</tr>
<tr>
<td>Nowotarski et al. [59]</td>
<td>2016</td>
<td>Sister model</td>
<td>ISO-NE GEFCom2012</td>
<td>1 h</td>
<td>(2.10–5.44)</td>
<td>-</td>
</tr>
<tr>
<td>Proposed model</td>
<td>-</td>
<td>VCTLA</td>
<td>ISO-NE GEFCom2012</td>
<td>1 h</td>
<td>0.47</td>
<td>102.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.89</td>
<td>225.32</td>
</tr>
</tbody>
</table>

### 4. Conclusions

In order to globally extract in-depth features from load data with nonlinear patterns and dynamics, a novel model needs to be developed to achieve greater accuracy in load forecasting to ensure the reliable and economic operations of an electric grid. This study proposed a new hybrid model based on VMD and deep TCN-based networks with SAM for STLF. The load data were decomposed into multiple sub-series that were reshaped into a feature matrix, along with external factors. A three-layer 1D-CNN network was used as a deep network to extract in-depth features, thus avoiding the over-fitting and gradient vanishing of the training model. The TCN extracted the long-term temporal dependencies, and LSTM further enhanced the extraction of temporal features. These features were globally adjusted via SAM to obtain important features. The load forecast was eventually obtained from the fully connected layer. The effectiveness and generalization of the proposed model were validated on two real-world datasets, ISO-NE and GEFCom2012. Compared with other benchmarking models, the experimental results demonstrated that the proposed model improved the prediction accuracy by 17% to 71% in terms of the MAPE and 7% to 70% in terms of the RMSE for the ISO-NE dataset, or by 15% to 53% in terms of the MAPE and 12% to 43% in terms of the RMSE for the GEFCom2012 dataset. Moreover, the $R^2$ of the proposed model achieved up to 99.9% for the ISO-NE dataset and 99.83% for the GEFCom2012 dataset. Therefore, the proposed model can be effective in globally extracting the in-depth features from load data and external factors to obtain greater accuracy for load forecasting.

In future research, we would like to further optimize the deep learning techniques to reduce their computational costs and achieve greater prediction accuracies. Moreover, future research could improve the TCN-based models to better adapt to complex time
series, and then apply these models to other areas such as wind speed, photovoltaic power and integrated energy systems.

Author Contributions: Conceptualization, Q.X. and M.L.; methodology, Q.X. and M.L.; software, Q.X. and Y.L.; validation, C.Z. and S.D.; formula analysis, M.L. and Q.X.; writing—original draft preparation, M.L. and Q.X.; writing—review and editing, M.L. and Y.L.; data curation, Q.X. and Y.L.; supervision, C.Z. and S.D.; funding acquisition, M.L. and S.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China under grant 62065012, and in part by the Natural Science Foundation of Jiangxi Province of China under grant 20212BAB202031, and in part by the Interdisciplinary Innovation Fund of Natural Science, Nanchang University under grant 9167-28220007-YB2111, and in part by the Innovation Fund Designated for Graduate Students of Jiangxi Province of China under grant YC2022-S132.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The datasets can be found here: https://www.iso-ne.com/isoexpress/web/reports/load-and-demand (accessed on 10 November 2020), and https://www.sciencedirect.com/science/article/pii/S0169207013000745#s000065 (accessed on 8 June 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References
2. Hao, J.; Sun, X.; Feng, Q. A novel ensemble approach for the forecasting of energy demand based on the artificial bee colony algorithm. Energy 2020, 13, 550. [CrossRef]
11. Liu, M.; Qin, H.; Cao, R.; Deng, S. Short-term load forecasting based on improved TCN and DenseNet. IEEE Access 2022, 10, 115945–115957. [CrossRef]
16. Liu, M.; Sun, X.; Wang, Q.; Deng, S. Short-term load forecasting using EMD with feature selection and TCN-based deep learning model. Energy 2022, 15, 7170. [CrossRef]
17. Ciechulski, T.; Osowski, S. High precision LSTM model for short-time load forecasting in power systems. Energy 2021, 14, 2983. [CrossRef]


34. Liang, Y.; Niu, D.; Hong, W. Short term load forecasting based on feature extraction and improved general regression neural network model. *Energy* 2018, 166, 653–663. [CrossRef]


45. Xiao, D.; Ding, J.; Li, X.; Huang, L. Gear fault diagnosis based on kurtosis criterion VMD and SOM neural network. *Appl. Sci.* 2019, 9, 5424. [CrossRef]


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.