EAI Endorsed Transactions
on Energy Web
Research Article EAI.EU

Dynamic Spatial-Temporal Load Forecasting for Multi-Regional Power Systems via Graph Attention Networks

Junxiong Ge[1*], Zhaojun Wang[3], Zhen Jiao[2], Lijun Liu[3], Xiao Li[3], Haimin Hong[1]

[1]Gridcom Technology Communication Co., Ltd Research and development department, Shenzhen, China
[2]State Grid Intelligence Technology Co., Ltd, Jinan, China
[3]State Grid Shandong Electric Power Company Marketing Service Center. Jinan, China

Abstract

Accurate multi-region load forecasting (MRLF) can effectively ensure the stable operation of the power system and is essential for the economic dispatch of the power system. However, with the gradual deepening of the dynamic spatial dependence among multi-region loads and the unmeasured nature of loads, the challenge is posed to MRLF. To address this challenge, a graph attention learning-based load forecasting model for multi-region considering dynamic spatial correlations aggregation is presented in this paper. Firstly, the regional load series is decomposed into a trend component and a fluctuation component through the discrete wavelet transform. Secondly, a generative adversarial network with the concept of zero-sum game is proposed for adversarial training of spatio-temporal prediction models. Furthermore, reasonable future multi-regional load forecasting values are obtained through a aggregation module that aggregate the trend component and the fluctuation component. Finally, the multi-regional load data of the New York Independent System Operator (NYISO) is used as a case. Compared with the evaluation metrics of mainstream models, the model presented in this paper is effective and superior in MRLF.

Keywords: Decoupling, Generative Adversarial Networks, Graph Attention Networks, Multi-Region Load Forecasting

Received on 15 October 2025 accepted on 14 November 2025, published on 09 February 2026

doi: 10.4108/ew.11818

1. Introduction

With the rapid development of the global economy and sustained population growth, electricity demand has exhibited a rapid upward trend. Coupled with the rapid advancement of technology and shifts in lifestyle, electricity consumption patterns have become increasingly complex and diverse. Meanwhile, load forms have grown more diversified, with enhanced volatility.

However, accurate load forecasting is beneficial for maintaining the stable operation of the power system. Reasonably adjust the power supply by forecasting changes in load. Thus, it achieves the purpose of maintaining the stable operation of the power system and realizing the balance between the supply and demand of the power system. Accurate load forecasting can help power companies formulate scheduling programs in advance to ensure the stable and reliable operation of the power system. It can also help to rationally regulate generator power, enhance the economic effectiveness of the power system, and decrease the influence on environmental pollution. Therefore, accurate load forecasting is an indispensable part of ensuring the stable operation of the power system.

Short-term load forecasting (STLF) methods include traditional methods and artificial intelligence methods. Traditional STLF models include regression models[1], Autoregressive Integrated Moving Average (ARIMA) models[2][3], grey relational degree models[4], and exponential smoothing models[5]. However, most traditional

*Corresponding author. Email: 2125786732@qq.com

models cannot effectively forecast nonlinear load series due to their simple structure[6]. With the continuous development of computing power and artificial intelligence algorithms, artificial intelligence models are gradually being applied in the field of STLF[7].

As a subset of machine learning, deep learning utilizes multi-layer neural networks to capture intricate mappings between inputs and outputs, while also mitigating the gradient vanishing problem commonly encountered in shallower models[8]. Currently, many scholars have applied deep learning models to load forecasting. Özüpak et al. [9] proposed a hybrid deep learning model for predicting short-term load demand by combining recurrent neural networks and long short term memory network (LSTM). Sun et al. [10] proposed a short-term load forecasting method based on gated recurrent units (GRU) and stochastic configuration network (SCN). This method achieves multi-scale power load forecasting by adaptively capturing high-frequency intrinsic mode functions through GRU and processing low-frequency IMF through SCN. Han et al. [11] proposed a short-term load forecasting model that combines an improved sand cat swarm algorithm (MSCSO) with a self attention time convolutional network. MSCSO effectively optimized the key parameters of the model. This significantly improves the accuracy and robustness of the model.

The methods described above primarily model the temporal dependencies in load series to improve forecasting accuracy. However, power load also exhibits significant spatial dependencies[12]. Therefore, to enhance load prediction accuracy, the spatial dependencies among loads should not be neglected. Scholars are increasingly focusing on this aspect to improve the precision of load forecasts. Vieira et al.[13] proposed to apply spatial convolution to the load forecasting problem to capture the spatial dependencies between neighboring load nodes through spatial convolution. This method significantly improves forecasting performance compared to autoregressive methods. Lin et al. [14] used graph neural networks to capture hidden spatial dependencies between different houses. Compared with the baseline, the accuracy of residential load forecasting has been significantly improved. Liu et al. [15] proposed a short-term load forecasting model based on dynamic adaptive adversarial graph convolutional network. It generates a dynamic adaptive graph through node embedding method for the spatio-temporal relationship between loads.

Although the above methods have improved prediction accuracy to some extent, previous work has overlooked some problems. Firstly, the spatial dependencies among loads are dynamically changing, and most existing methods are unable to capture these dynamic spatial dependencies[16][17]. As shown in Figure 1, edges represent the spatial dependencies among nodes. The Figure 1 shows that the spatial dependencies of nodes exhibit dynamic variations over time. Secondly, as a type of time series data, load data can be decomposed into various components, including trend, seasonality, and residuals, each exhibiting distinct characteristics[18][19]. However, most existing methods overlook this characteristic. There is no specific network designed for different components to capture features, which seriously affects the prediction accuracy. In addition, most previous load forecasting models simply aimed to minimize L1 loss or L2 loss. However, overly simplified loss functions may not be sufficient to improve some of the model parameters, which affecting the improvement of model prediction performance[20][21].

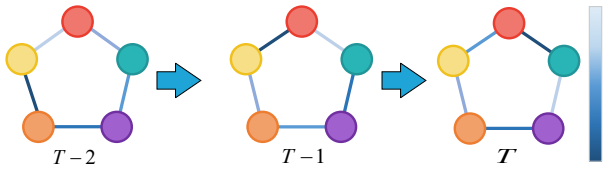

Figure 1. Dynamic spatial dependencies

To address the above problems, a hybrid model with temporal mechanism and LSTM graph attention network (TLGAT) is presented in this paper. This model not only effectively decouples the multi-regional load series into trend component and fluctuation component, but also captures the dynamic temporal and spatial dependencies among regional loads through specific spatio-temporal networks, improving forecast accuracy. The main work of this paper is summarized as follows:

(i) A MRLF model based on graph attention learning is proposed, which is able to learn the interactions among nodes at different moments and effectively extracts the dynamic spatial correlations among multi-region loads, thus improving the load forecasting accuracy;
(ii) Discrete Wavelet Transform (DWT) is employed as a method to decouple multi-region load series. The original region load series are decoupled into low-frequency and high-frequency series by the DWT, which clearly represents the changing features of the region load series. Subsequently, temporal convolution and LSTM are leveraged to capture more accurate and effective temporal dependencies in each series;
(iii) Generative adversarial networks(GAN) are harnessed for adversarial training of models. The model obtains adversarial losses through adversarial training, and combines adversarial loss with L1 loss as the training objective to optimize model parameters and improve model prediction performance;
(iv) The multi-regional load data of the NYISO is used to validate the performance of the TLGAT. Compared with several advanced models, the case study results demonstrate that the TLGAT has superiority in MRLF.

2. Problem Description

Accurate regional load forecasting can balance the relationship between power supply-demand and improve the efficiency of power dispatch. The problem of MRLF aims to predict the regional load in the future by studying the potential relationships among historical multi-regional load

data and relevant influencing factors such as geographical location, meteorological conditions, and economy. Specifically, $x_t^i \in \square$ denotes the multi-regional load of the i th regional load node at time step t. denotes the regional load of all regional load nodes at time step t. $X_{:T_1} = \{X_1, \ldots X_{T_1}\} \in \square^{T_1 \times N}$ signifies the regional load of all regional load nodes in time slice T_1. The aim of this paper is to learn a mapping function F_θ based on historical regional load data from the past T_1 steps and use this mapping function F_θ to map the regional load for the future T_2 steps. The problem of MRLF can be summarized as:

$$\{X_{t+1}, X_{t+2}, \ldots, X_{t+T_2}\} = \mathsf{F}_\theta \{X_{t-T_1+1}, X_{t-T_1+2}, \ldots, X_t\} \quad (1)$$

where θ represents the learnable parameters.

However, in addition to temporal dependencies, spatial dependencies are also present in multi-regional load series. Therefore, it is necessary to construct a power system topology graph $G = (V, E, A)$ to summarize the spatial dependencies among regional loads. V represents the set of regional load nodes, E denotes the set of busbars connecting two regional load nodes, and the graph adjacency matrix A is utilized to describe the spatial dependencies between any two regional load nodes. Therefore, the problem of MRLF will be redefined as:

$$\{X_{t+1}, X_{t+2}, \ldots, X_{t+T_2}\} = \mathsf{F}_\theta (\{X_{t-T_1+1}, X_{t-T_1+2}, \ldots, X_t\}, G) \quad (2)$$

Model Overview: The Figure 2 shows the structure of the TLGAT, which mainly includes three main parts:

1) Decoupling module: Firstly, the decoupling module consists of a multi-level DWT, an inverse discrete wavelet transform (IDWT), and fully connected layers (FC). The multi-historical regional load series is decoupled into a low-frequency series and several high-frequency series by the multi-level DWT. Subsequently, the trend and fluctuation are derived through the IDWT and FC.

2) Dual channel Spatio-temporal Module: Due to the different characteristics between trend and fluctuation, temporal attention and causal convolution are employed to capture the temporal features of trend and fluctuation, respectively. The graph attention networks (GAT) is leveraged to capture the dynamic spatial dependencies among multi-regional loads.

3) Fusion and Adversarial Training Module: Based on the spatio-temporal features learned by the dual channel spatio-temporal module, future trends and fluctuations can be predicted by two predictors. Then, the adaptive fusion module is employed to fuse future trend and fluctuation to obtain future regional loads. Subsequently, the GAN is introduced for adversarial training.

3. Methodology

3.1. Decoupling Module

The complex data from the real world typically consists of multiple components[22][23]. To clearly represent the data, it can be decomposed into various components. Therefore, in order to clearly represent the load data for capturing the feature of load series, the regional load series is decomposed into trends and fluctuations. Furthermore, there is no mutual influence between trends and fluctuations. In other words, when a component of the regional load series changes, other components will not be affected [24].

To decompose the regional load series, the DWT is applied in this paper to decompose the load series in the framework of MRLF. The reason for choosing the DWT is that the load series is similar to the wind power series, both of which are discrete series[25]. The DWT is designed specifically for processing discrete series [26]. Figure 3 shows an example of a simple DWT. In the Figure 3, the first level DWT decouples the input series $X_{:T_1} \in \square^{T_1}$ into a low-frequency series $X_{L_1} \in \square^{\frac{T_1}{2}}$ and a high-frequency series $X_{H_1} \in \square^{\frac{T_1}{2}}$ through a low-pass filter f_{LP} and a high-pass filter f_{HP}. Similarly, the second level DWT decomposes $X_{L_1} \in \square^{\frac{T_1}{2}}$ into $X_{L_2} \in \square^{\frac{T_1}{4}}$ and $X_{H_2} \in \square^{\frac{T_1}{4}}$ .The process can be depicted as:

$$\begin{cases} X_{H_1} = (f_{HP} \otimes X_{:T_1})_{(\downarrow 2)}, \\ X_{L_2} = (f_{LP} \otimes (f_{LP} \otimes X_{:T_1})_{(\downarrow 2)})_{(\downarrow 2)}, \\ X_{H_2} = (f_{HP} \otimes (f_{LP} \otimes X_{:T_1})_{(\downarrow 2)})_{(\downarrow 2)}, \end{cases} \quad (3)$$

where, $\otimes$ represents convolution calculation, $\downarrow 2$ represents reducing the sampling rate by 2.

According to (4), in the process of the DWT, the time series length of low-frequency and high-frequency series will become shorter. To ensure consistent series length, it is necessary to perform recovery sampling operations on $\bar{X}_{L_2}$, $\bar{X}_{H_2}$ and $\bar{X}_{H_1}$. Then, an inverse low-pass filter and an inverse high-pass filter are used to implement the IDWT. Finally, a FC is employed to obtain trend $X_{trend} \in \square^{T_1 \times N \times d}$ and fluctuation $X_{fluc} \in \square^{T_1 \times N \times d}$. The IDWT can be described as:

$$X_{trend} = W^{LP} f_{LP}^T \otimes (f_{LP}^T \otimes (\bar{X}_{L_2})_{\uparrow 2})_{\uparrow 2} + b^{LP} \quad (4)$$

$$X_{flu} = \begin{cases} W^{HP} (f_{LP}^T \otimes (f_{HP}^T \otimes (\bar{X}_{H_2})_{\uparrow 2})_{\uparrow 2} \\ + f_{HP}^T \otimes (\bar{X}_{1,h})_{\uparrow 2}) + b^{HP} \end{cases} \quad (5)$$

where, f_{LP}^{T} and f_{HP}^{T} respectively denotes an inverse low-pass filter and an inverse high-pass filter, $\uparrow 2$ represents the recovery sampling operation, $W^{LP}, W^{HP} \in \mathbb{R}^{d}$ and $b^{LP}, b^{HP} \in \mathbb{R}^{d}$ represent learnable parameters.

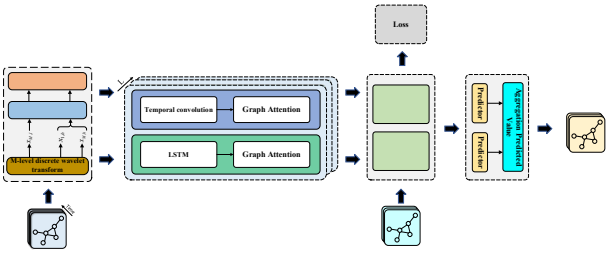

Figure 2. The structure of the TLGAT

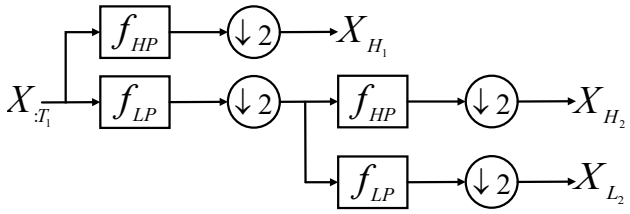

Figure 3. Two level discrete wavelet transform

3.2. LSTM and Temporal Convolution

The trend and fluctuation obtained through decoupling module decomposition have completely opposite temporal features. The change in trend lasts for a long time and has a relatively small amplitude, while the change in fluctuation is usually sudden and has a short duration and a large amplitude.

Due to the rapid and significant fluctuations trend, temporal convolution is used to capture temporal characteristics. Temporal convolution is a method specifically designed for processing time series data within CNNs, characterized by the fact that the output at the current time step depends only on the current and a limited number of previous time steps of input. This structure enables it to effectively capture the short-term dependencies that change over time within the load series. The temporal convolution structure is shown in Figure 4.

Temporal convolution can be described as:

$$(X * f_C)(t) = \sum_{\tau=0}^{K} f_C(\tau) X(t-\tau), \tag{6}$$

where, $X \in \mathbb{R}^{T_1}$ represents time series, $f_C \in \mathbb{R}^{K}$ denotes the filter, and K indicates the size of the convolution kernel for temporal convolution. Therefore, the temporal representation of fluctuation X^{conv} as:

$$X^{conv} = ReLU(\theta * X_{fluc}), \tag{7}$$

where, θ signifies the learnable parameters of the model, and $ReLU(\cdot)$ denotes the linear rectification function.

After LSTM and Temporal Convolution to extract temporal dependencies, $X_{trend}^{att}, X_{fluc}^{conv} \in \mathbb{R}^{T_1 \times N \times d}$ are obtained.

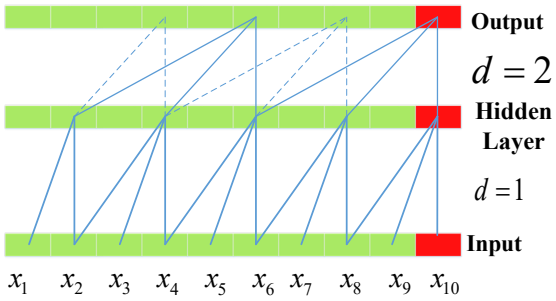

Figure 4. Temporal convolution network

Therefore, for trend, there is a strong correlation between any two time slices of the trend. For fluctuation, time slices that are strongly correlated are often adjacent time slices, and in comparison, those with time intervals further apart have relatively weaker correlations. Based on trend characteristics, LSTM is selected to capture the temporal features of the trend. The LSTM is shown in Figure 5.

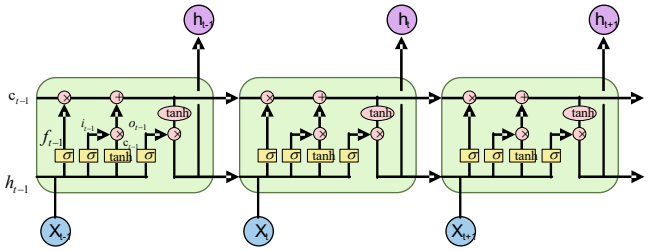

Figure 5. The sstructure of LSTM

As a powerful sequence modeling unit, LSTM 's gating mechanism can dynamically adjust the information flow and determine when to update and retain what information. This enables LSTM to effectively capture the importance differences of different time steps in the input sequence, thereby focusing on key information and improving the performance and efficiency of the model. Therefore, the characteristics of LSTM enable it to reasonably solve the problem of capturing the temporal features of trends. It can not only realize the interaction of all historical information through its circular connection, but also selectively retain

more important historical information and integrate new information through the forgetting gate and the input gate. The process of obtaining the trend output on the time slice can be described as:

$$\begin{cases} i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{(t-1)} + b_{hi}), \\ f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf}), \\ g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{(t-1)} + b_{hg}), \\ o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{(t-1)} + b_{ho}), \\ \mathrm{c}_t = f_t * \mathrm{c}_t + i_t * g_t \\ h_t = o_t * \tanh(c_t) \end{cases} \tag{8}$$

where, h_t signifies the output of trend of node at time slice t, W and b denotes learnable parameters, σ represents the exponential function, f_t, i_t and o_t indicates the temporal correlation between the time slices t and i of node n on the trend.

3.3. Graph Attention Networks

In fact, the future regional load not only depends on its own historical load data, but also closely related to the historical load status of other regions. Therefore, the spatial correlation characteristics should be regarded as the intrinsic attributes in the regional load sequence. However, the load nodes are often non-uniform and irregularly distributed in space, which constitutes non-Euclidean structural data. Obviously, it is difficult for traditional neural networks to effectively extract spatial features from such data with complex topological relationships. Graph neural network is designed to deal with the data of irregular topology, which can mine the implicit spatial correlation between regional loads. In addition, the interaction between load nodes in different regions is not static and fixed, but dynamically evolves with time and external conditions. These dynamic interactions often contain key information. Therefore, the GAT can effectively capture the potential dynamic spatial dependencies between regional loads.

Unlike the traditional GCN, the GAT employs an attention mechanism to adaptively determine the weights between neighboring nodes, enabling weighted aggregation of neighborhood information. Moreover, this mechanism can capture time-varying weights of neighbors across different time steps, thereby dynamically modeling the spatial dependencies among nodes. In contrast, the GCN relies on fixed weights for neighborhood aggregation, which limits its ability to represent dynamic spatial interactions. For this reason, the GAT is adopted in this study to effectively capture the spatial dependencies across regional load data. The spatial correlation $\beta_t^{n,i}$ between node n and node i in time slice t can be described as:

$$\beta_t^{n,i} = \frac{exp(W_S^K x_t^n, W_S^Q x_t^i)}{\sum_{j=1}^{N} exp(W_S^K x_t^n, W_S^Q x_t^i)}, \tag{9}$$

where, $W_S^Q, W_S^K \in \mathbb{R}^{d\times d}$ indicates learnable parameters. The output of node n in time slice t can be described as:

$$x_t^n = \sum_{i=1}^{N} \beta_t^{n,i}(W_S^V x_t^n) \tag{10}$$

where, $W_S^V \in \mathbb{R}^{d\times d}$ denotes learnable parameters. X_{trend}^{att} and X_{fluc}^{conv} can be represented as $X_{trend}^{gat}, X_{fluc}^{gat} \in \mathbb{R}^{T_1\times N\times d}$ after passing through the GAT.

By overlaying carefully designed dual channel spatio-temporal modules L times, different spatio-temporal features in trend and fluctuation can be obtained. Subsequently, future regional load values can be inferred through Aggregation module.

3.4. Aggregation and Adversarial Training

The predictor composed of two FC converts $X_{trend}^{gat}, X_{fluc}^{gat}$ into the predicted representation $\hat{y}_{trend}^{f}, \hat{y}_{fluc}^{f} \in \mathbb{R}^{T_2\times N\times d}$ of trend and fluctuation. To obtain the representation $\hat{y}^f \in \mathbb{R}^{T_2\times N\times d}$ of future regional load values, it is necessary to Aggregation $\hat{y}_{trend}^{f}$ and $\hat{y}_{fluc}^{f}$. The changes in regional load are always frequent, leading to the problem of distribution shift in the fluctuating components of regional load. To address distribution bias and obtain more accurate prediction values. It is necessary to sum the fluctuations within time slice t with weights, which are calculated by LSTM. It effectively solves the problem of regional distribution offset and obtains more accurate future regional load values. The entire process of Aggregation can be described as:

$$\begin{cases} \hat{y}_t^{f^n} = \hat{y}_{trend_t}^{f^n} + \sum_{i=T_1+1}^{T_1+t} \eta_{t,i}^n (W_F^V \hat{y}_{fluc_i}^{f^n}) \\ \eta_{t,i}^n = \frac{exp((W_F^K \hat{y}_{trend_t}^{f^n})^T (W_F^Q \hat{y}_{fluc_i}^{f^n}))}{\sum_{j=T_1+1}^{T_1+t} exp((W_F^K \hat{y}_{trend_t}^{f^n})^T (W_F^Q \hat{y}_{fluc_j}^{f^n}))}, \end{cases} \tag{11}$$

where, W_F^Q, W_F^K, W_F^V signifies learnable parameters. After obtaining the future regional load representation $\hat{y}^f$, the future regional load values $\hat{y} \in \mathbb{R}^{T_2\times N}$ can be obtained through a FC. Subsequently, the GAN is introduced for adversarial training to further optimize the parameters of the model.

The basic construction of the GAN is illustrated in Figure 6. The GAN consist of a generator and a discriminator, which continuously compete to optimize their parameters. To obtain more appropriate parameters for the model, adversarial training is performed based on future regional load values. Firstly, the discriminator judges the future regional load values and returns a discrimination result. Then, based on the discrimination results, the model parameters of the generator are further optimized to enable the generator to forecast more accurate future regional load values. However, more accurate future regional load values will also make the discriminator more rigorous. Therefore, in the process of the generator and discriminator playing against each other. Appropriate parameters can be obtained by the model to accurately predict future regional loads.

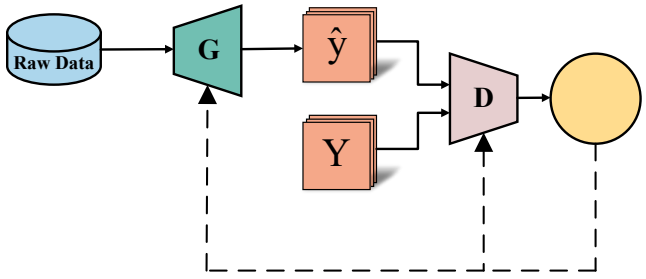

Figure 6. The structure of GAN

For the GAN, the judgment result of the discriminator can be expressed through a binary cross entropy, which can be expressed as:

$$V(\mathrm{G},\mathrm{D}) = \frac{1}{n}\sum_{i=1}^{m}[\log \mathrm{D}(y) + \log(1-\mathrm{D}(\mathrm{G}(X)))] \quad (12)$$

Equation (12) denotes the average expression of binary cross entropy for n samples, where $\mathrm{D}(\mathrm{G}(X))$ is denoted as the probability of determining the generated data to be true, X represents random data, and y denotes real data, $\mathrm{G}(y)$ signifies the data generated by the generator, $\mathrm{D}(y)$ indicates the probability that the discriminator judges the real data y as true.

For the discriminator, it aims to maximize the loss. For the generator, it seeks to minimize the loss. The discriminator and generator always play against each other during the training process, alternating to become stronger. Finally, both will only tend towards a balance point. Therefore, the adversarial training process can be described as:

$$\min_{\mathrm{G}} \max_{\mathrm{D}} V(\mathrm{D},\mathrm{G}) = \begin{cases} \mathrm{E}_{y\sim P_{data}(y)}[\log \mathrm{D}(y)] \\ +\mathrm{E}_{X\sim P_X(X)}[\log(1-\mathrm{D}(\mathrm{G}(X)))] \end{cases} \quad (13)$$

where, $\mathrm{E}_{y\sim P_{data}(y)}[\log \mathrm{D}(y)]$ denotes the expectation of $\log \mathrm{D}(y)$ when y is the real data, $\mathrm{E}_{X\sim P_X(X)}[\log(1-\mathrm{D}(\mathrm{G}(X)))]$ denotes the expectation of $\log(1-\mathrm{D}(\mathrm{G}(X)))$ when X is the random data.

3.5. Objective Function

The process of obtaining an accurate short-term regional load forecasting model is the process of minimizing the objective function. The entire objective function is mainly composed of two parts, one is the L1 loss:

$$L_{Load} = \sum_{t=T_1+1}^{T_1+T_2}\sum_{n=1}^{N}|x_t^n - \hat{y}_t^n| \quad (14)$$

The other part is the adversarial loss of the GAN. The model task is to generate regional load values that are close to the actual values. This is equivalent to the discriminator in the GAN being unable to distinguish the authenticity of the regional load predicted by the model. Adversarial losses can be described as:

$$L_{\mathrm{D}} = -\log(1-\mathrm{D}(\mathrm{G}(X))) \quad (15)$$

Therefore, the goal of the regional load forecasting model is to minimize the objective functions:

$$L = L_{Load} + \lambda L_{\mathrm{D}} \quad (16)$$

where λ is the weight.

4. Experiment Analysis

4.1. Dataset Preprocessing

The regional load data of 11 load areas in New York for the whole year of 2022 are used to validate the performance of the TLGAT proposed in this paper. The load profile of area N.Y.C is shown in Figure 7.

The load data sampling interval is 1 hour with 8760 data points per regional load node data. The detailed descriptive statistical data of the dataset is shown in Table 1.

Table 1. The statistical data of the dataset

Nodes	Time Range	Granularity	Time steps	Unit
11	January 1, 2022-December 31, 2022	1hour	8760	MW

The hyperparameters (including GAN) of each module are shown in Table 2.

The Z-score method is employed to normalize the dataset in this paper. The dataset is split across a training, validation, and test set with a proportion of 6:2:2. The normalization process can be described as:

$$x_{norm} = \frac{x - x_{mean}}{x_{std}} \tag{17}$$

where, x_{mean} represents the average value of all regional loads, x_{std} denotes the standard deviation of all regional loads, x signifies the true value of regional loads, and x_{norm} indicates the normalized values. In this paper, the regional load values of historical $T_1 = 12$ hours are used to forecast the regional load values of future $T_2 = 12$ hours.

Table 2. The hyperparameters of each module

Name	Value	Name	Value
Batch-size	64	DWT level	1
Input-dim	1	Output-dim	1
Features	128	Kernel-size	2
Num-layer	2	Enpoch	100
Wavelet	Coiflets	Learning rate	0.001
Weight(λ)	1	Dropout	0.2

4.2. Model parameter setting and evaluation metrics

The Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) are utilized to evaluate the predictive performance of the TLGAT. The three evaluation metrics can be described as:

$$MAE = \frac{1}{n}\sum_{i}^{n}|\hat{y}_i - y_i| \tag{18}$$

$$RMSE = \sqrt{(\frac{1}{n}\sum_{i}^{n}(\hat{y}_i - y_i)^2)} \tag{19}$$

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

where, n denotes the prediction step size, $\hat{y}_i$ represents the future regional load values, and y_i signifies the regional load real values.

4.3. Case Results and Discussion

To validate the performance of the TLGAT, various mainstream models such as Informer, LSTM, STGCN, AGCRN, GWNet, etc. are selected in this paper to compare their predictive performance. All comparison models and the TLGAT evaluation metrics are shown in Table 3 and Figure 8. The evaluation metrics obtained are the average of three experiments.

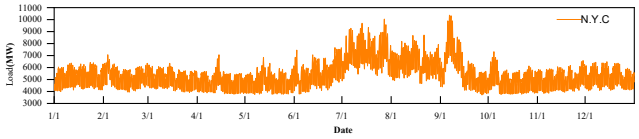

Figure 7. The load profile of region N.Y.C

It can be inferred from Table 2 and Figure 8. Firstly, the SVR model has the worst predictive performance because it can only predict based on linear features in regional load series. Secondly, TLGAT is compared to models such as Dlinear, LSTM, and Informer which only capture temporal features. TLGAT can more accurately predict future regional load values. Compared to SVR, the MAE of TLGAT has reduced by 46.87, RMSE by 76.22 and MAPE by 3.68%. Compared to LSTM, the evaluation metrics for TLGAT were reduced by 13.98, 17.07, and 0.84%, respectively.

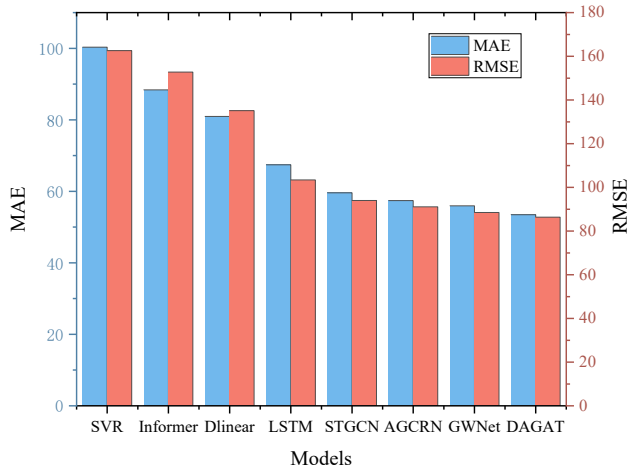

Figure 8. The performance comparison of models

Furthermore, models capable of capturing spatial dependencies between regional loads—such as GWNet outperform those like LSTM which only capture temporal dependencies within individual regional load series. This highlights that regional loads exhibit not only strong temporal autocorrelation but also significant spatial interdependencies. In this context, the proposed TLGAT employs GAT to explicitly model the spatial relationships between regional loads, thereby further enhancing forecasting accuracy. Compared to STGCN the TLGAT model achieves reductions in MAE RMSE and MAPE of 0.249, 0.216, and 0.87%

respectively. This improvement can be attributed to the adaptive weighting mechanism of GAT which aggregates information from neighboring nodes in a learned manner, unlike STGCN which applies fixed and uniform weights during feature aggregation.

In the TLGAT model, the GAT module assigns distinct aggregation weights to neighboring nodes via an attention mechanism, thereby highlighting the relative importance of information from different neighbors. This weighting strategy not only enhances the representation of significant neighbor features but also facilitates the effective capture of spatial dependencies across regional loads. Moreover, these aggregation weights are computed per time slice and adapt dynamically over time. As a result, the model is able to capture dynamic spatial dependencies that evolve across time steps

The regional load sequence is decomposed into trend and fluctuation components via DWT. This decomposition makes the underlying patterns within the load series more distinguishable, thereby facilitating the tailored design of a subsequent dual-channel spatio-temporal module. Each channel is specifically adapted to the distinct characteristics of the trend and fluctuation components, leading to more accurate predictions. Furthermore, adversarial training contributes to parameter optimization, which enhances the overall predictive performance of the model.

Table 3. The performance comparison of models

Models	MAE	RMSE	MAPE
SVR	100.32	162.57	7.88%
Informer	88.40	152.77	6.10%
Dlinear	80.95	135.11	5.75%
LSTM	67.43	103.42	5.04%
STGCN	59.60	94.00	4.76%
AGCRN	57.42	91.10	4.64%
GWNet	55.94	88.51	4.60%
TLGAT	53.45	86.35	4.20%

To further evaluate the contributions of adversarial training and the GAT, ablation studies were conducted on both components. The baseline model, referred to as DCA, employs DWT to decompose the regional load series but incorporates neither GAT nor adversarial training. Subsequently, the temporal dependencies of each component are captured through temporal convolution and LSTM. To further assess the impact of each component, we denote the model without adversarial training as TCN, which is trained to minimize the L1 loss. First, as shown in Table 3, the TCN achieves reductions of 7.1774, 3.3358, and 0.24% across the evaluation metrics compared to LSTM, indicating that the GAT effectively captures dynamic spatial dependencies among regional loads. Second, by integrating GAT and Discrete DWT with adversarial training, we obtain the complete TLGAT model. The evaluation metrics of TLGAT further decrease relative to those of TCN—by 1.7482, 1.0021, and 0.26%, respectively. These results demonstrate that adversarial training facilitates a competitive process between the prediction network and the adversarial network, continuously optimizing the parameters of the prediction model. As a result, the prediction network learns underlying features of regional loads more effectively, leading to more accurate forecasts.

Table 4. The performance comparison of each module

Methods	MAE	RMSE	MAPE
w/o LSTM	62.9756	90.6879	4.70%
w/o TCN	55.1982	87.3521	4.46%
w/o Temporal	60.725	88.472	4.80%
w/o GAN	61.439	87.542	4.83%
TLGAT	53.45	86.35	4.20%

The LSTM denotes the model that does not introduce the GAT and adversarial training and only uses the LSTM to extract the regional load series. The temporal dependencies within each decomposed component are captured using temporal convolution This model is trained with the objective of minimizing the L1 loss function. As summarized in Table 4, the TCN model shows notable improvements over the LSTM baseline, with evaluation metrics decreasing by 7.1774, 3.3358, and 1.04%, respectively. These results indicate that the integration of GAT effectively captures dynamic spatial dependencies among regional loads, contributing significantly to forecasting accuracy. the complete TLGAT model is constructed by further incorporating adversarial training alongside GAT and DWT. The evaluation metrics of TLGAT experience additional reductions of 1.7482, 1.0021, and 0.26% compared to TCN, highlighting the benefit of adversarial training. This outcome demonstrates that the adversarial mechanism facilitates a competitive learning process between the prediction network and the adversarial network. leading to enhanced model robustness and refined parameter optimization.

This process continuously optimizes the parameters of the predictive network model. Ultimately, the prediction network can accurately learn the potential feature of regional loads to obtain accurate future regional load values.

The predictive performance of the model can be intuitively seen from the error distribution diagram. The Figure 9 clearly shows the distribution of prediction errors between the TLGAT and multiple comparison models.

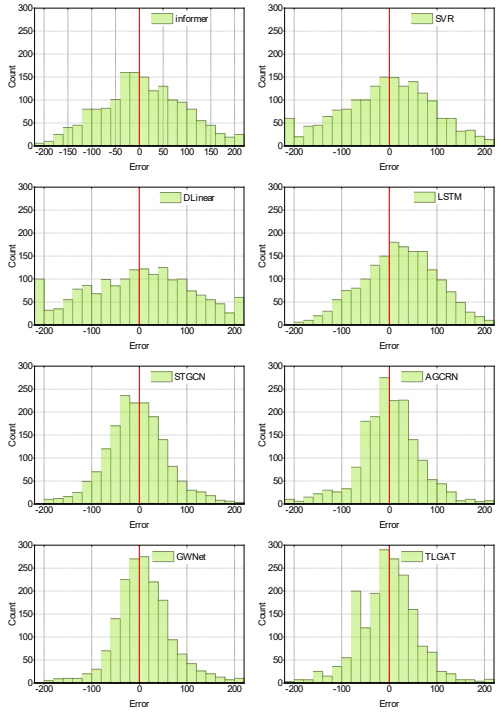

Figure 9. The error distribution comparison of models

From the graph, it can be seen that models with poor predictive performance, such as ARIMA, have dispersed error distributions. Or like Autoformer, the error is mainly distributed far away from 0. It is not difficult to see that the error distribution of the TLGAT is relatively concentrated, mainly around 0. This situation demonstrates that compared to other forecasting models, the TLGAT can achieve smaller prediction errors in most cases and demonstrate its enormous potential in regional load forecasting.

The Figure 10 shows the comparative analysis among the predicted and real values of the TLGAT. As can be seen in the Figure 10, in most cases the predicted curves always overlap with the actual curves and change accordingly as the actual values change. To further analyze the predictive performance of the TLGAT, four prediction windows are randomly selected, as shown in the Figure 11. The left part of the reference line in the figure presents the historical values of regional load, while the right part shows the predicted values and corresponding real values. Compared to other models, the TLGAT model has a smaller deviation between its predicted values and the real values. Several results indicate the superiority of the TLGAT in solving the regional load forecasting problem been effectively improved.

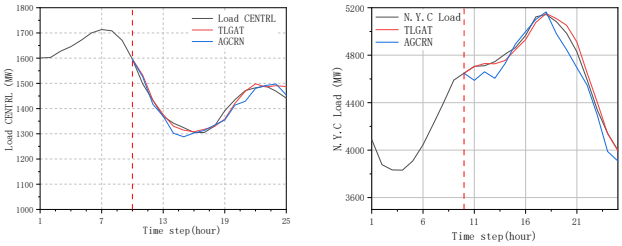

Figure 10. The comparison of real and predicted values in different sliding windows

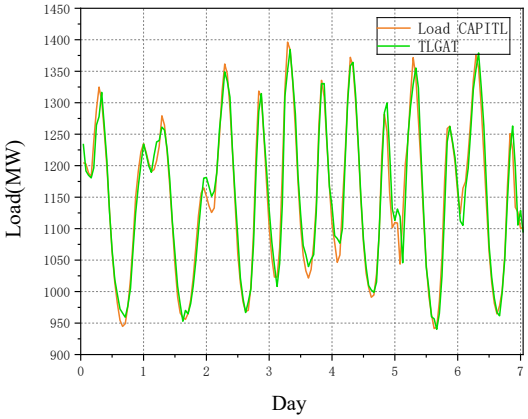

Figure 11. The comparison of true and predicted values for a week

5. Conclusion

Considering the dynamic spatial dependencies among regional loads and the difficulty in characterizing load time series features. A short-term regional load forecasting model based on decoupling mechanism and the GAT is presented in this paper. It is intended to solve the problem of difficulty in capturing the dynamically changing spatial dependencies among regional loads. The DWT is introduced by models to decompose regional load series into the trend and the fluctuation. The decoupling mechanism based on the DWT makes the temporal and spatial dependencies captured by spatio-temporal networks more accurate. Secondly, the time-varying aggregation weights between nodes can be learned by the GAT in the model. Dynamic aggregation weights solve the problem of difficulty in capturing dynamic spatial dependencies among regional loads. The prediction accuracy of the model has In addition, adversarial training based on GAN optimized model parameters and improved model prediction performance. The calculation results on the dataset of the NYISO show that compared with other mainstream models. The model proposed in this paper has effectiveness and superiority in short-term regional load forecasting, and the predicted values are closer to the real values. It effectively helps to clearly represent the feature of regional load series and solves the problem of previous methods being unable to capture dynamic spatial dependencies among loads. In summary, this model not only improves the accuracy of

regional load forecasting but also provides a basis for capturing the dynamically varying spatial dependencies among regional loads, thereby enhancing its effectiveness and credibility in engineering applications.

References

[1] V. Dordonnat, A. Pichavant, A. Pierrot, “GEFCom2014 probabilistic electric load forecasting using time series and semi-parametric regression models,” International Journal of Forecasting, vol. 32, no. 3, pp. 1005-1011, Jul. 2016.

[2] J. W. Taylor and P. E. McSharry, “Short-Term Load Forecasting Methods: An Evaluation Based on European Data,” IEEE Transactions on Power Systems, vol. 22, no. 4, pp. 2213-2219, Nov. 2007.

[3] Rehan Jamil, “Hydroelectricity consumption forecast for Pakistan using ARIMA modeling and supply-demand analysis for the year 2030,” Renewable Energy, vol. 154, pp. 1-10, Jul. 2020.

[4] Saadat Bahrami, Rahmat-Allah Hooshmand, Moein Parastegari, “Short term electric load forecasting by wavelet transform and grey model improved by PSO (particle swarm optimization) algorithm,” Energy, vol. 72, pp. 434-442, Aug. 2014.

[5] Juan F. Rendon-Sanchez, Lilian M. de Menezes, “Structural combination of seasonal exponential smoothing forecasts applied to load forecasting,” European Journal of Operational Research, vol. 275, no. 3, pp. 916-924, Jun. 2019.

[6] Zhining Cao, Jianzhou Wang, Yurui Xia, “Combined electricity load-forecasting system based on weighted fuzzy time series and deep neural networks,” Engineering Applications of Artificial Intelligence, vol. 132, pp. 108375, Jun. 2024.

[7] Hui Hou et al., “Review of load forecasting based on artificial intelligence methodologies, models, and challenges,” Electric Power Systems Research, vol. 210, pp. 108067, Sep. 2022.

[8] J. Wang et al., “An annual load forecasting model based on support vector regression with differential evolution algorithm,” Applied Energy, vol. 94, pp. 65-70, Jun. 2012.

[9] Y. Özüpak and S. Mansurov, "Optimizing electricity demand forecasting with a novel RNN-LSTM hybrid model," Energy Sources, Part B: Economics, Planning, and Policy, vol. 20, no. 1, Art no. 2531448, 2025.

[10] J. Sun, K. Ma and H. Zhao, "Short-term power load hybrid forecasting using GRU and SCN," Energy and Buildings, vol. 347, no. Part B, pp. 116372, 2025.

[11] H. Han, J. Peng, J. Ma, H. Liu and S. Liu, "Research on Load Forecasting Prediction Model Based on Modified Sand Cat Swarm Optimization and SelfAttention TCN," Symmetry, vol. 17, pp. 1270, 2025.

[12] Y. Wang, T. Han, L. Rui, J. Ma and Q. Jin, "Ego-centric multiple-correlation and temporal graph neural networks based residential load forecasting," Engineering Applications of Artificial Intelligence, vol. 160, no. Part B, pp. 111933, 2025.

[13] D.A.G. Vieira et al., “Large scale spatial electric load forecasting framework based on spatial convolution,” International Journal of Electrical Power & Energy Systems, vol. 117, pp.105582, May. 2020.

[14] W. Lin, D. Wu and B. Boulet, “Spatial-Temporal Residential Short-Term Load Forecasting via Graph Neural Networks,” IEEE Transactions on Smart Grid, vol. 12, no. 6, pp. 5373-5384, Nov. 2021.

[15] J. Liu et al., “Load forecasting based on dynamic adaptive and adversarial graph convolutional networks,” Energy and Buildings, vol. 312, pp. 114206, Jun. 2024.

[16] Y. Zhou and M. Wang, "Empower Pre-Trained Large Language Models for Building-Level Load Forecasting," in IEEE Transactions on Power Systems, vol. 40, no. 5, pp. 4220-4232, Sept. 2025

[17] Y. Liu, Y. Wang, P. Xu, Y. Xue, Y. Chen and D. Zhang, "BuildSTG: A multi-building energy load forecasting method using spatio-temporal graph neural network," Energy and Buildings, vol. 347, no. Part B, pp. 116190, 2025.

[18] G. Fan, H. Wei, H. Huang and W. Hong, "Application of ensemble empirical mode decomposition with support vector regression and wavelet neural network in electric load forecasting," Energy Sources, Part B: Economics, Planning, and Policy, vol. 20, no. 1, 2025.

[19] B. Sun, X. Chen, T. Shen and L. Ma, "Enhancing long-term load forecasting with convolutional informer-based hybrid model," Engineering Applications of Artificial Intelligence, vol. 161, no. Part A, pp. 112051, 2025.

[20] H. Dong, J. Zhu, S. Li, Y. Miao, C. Y. Chung and Z. Chen, "Probabilistic Residential Load Forecasting with Sequence-to-Sequence Adversarial Domain Adaptation Networks," in Journal of Modern Power Systems and Clean Energy, vol. 12, no. 5, pp. 1559-1571, September 2024.

[21] X. Ouyang et al., “CityTrans: Domain-Adversarial Training With Knowledge Transfer for Spatio-Temporal Prediction Across Cities,” IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 1, pp. 62-76, Jan. 2024.

[22] H. Zhang, M. Zhou, Y. Chen and W. Kong, "Short-term power load forecasting for industrial buildings based on decomposition reconstruction and TCN-Informer-BiGRU," Energy and Buildings, vol. 347, no. Part B, pp. 116317, 2025.

[23] Z. Tian, Z. Dong and S. Lv, "WDLformer: A novel mid-term load forecasting network considering weather score and distribution shift based on multi-encoders frequency fusion approach," Information Fusion, vol. 126, no. Part B, pp. 103674, 2026.

[24] Y. Fang, Y. Qin, H. Luo, F. Zhao and K. Zheng, "STWave++: A Multi-Scale Efficient Spectral Graph Attention Network With Long-Term Trends for Disentangled Traffic Flow Forecasting," in IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 6, pp. 2671-2685, June 2024.

[25] D. Hangawatta, A. Gargoom and A. Kouzani, "A novel method for electrical vehicle charging load extraction from low sampling rate data," Sustainable Energy, Grids and Networks, vol. 43, pp. 101903, 2025.

[26] R. Gong, A. Jiang, H. Hu, D. Liu, X. Wu and S. Zhang, "Short-term multi-featured power load forecasting model based on GBKA-VMD-Mambaformer with noise separation error correction under scarce load data," Electric Power Systems Research, vol. 247, pp. 111846, 2025.