International Journal of Power and Energy Systems, Vol. 46, No. 10, 2026
An Open Access Paper
Special Issue
Next Generation Energy Systems

MULTI-SCALE POWER LOAD FORECASTING BASED ON ADAPTIVE GCN AND MULTI-HEAD ATTENTION MECHANISM

Xufeng Wu,* Min Chen,* Nan Dong,** Yuwen Wu,** Zhanzhi Liu,** and Buyun Su**

Abstract

With the high proportion of renewable energy connected to the grid, the non-stationarity and multi-scale characteristics of power load are becoming increasingly prominent, which puts higher demands on prediction accuracy. Therefore, a multi-scale power load forecasting model combining an adaptive graph convolutional network and a multi-head attention mechanism was proposed. The core innovation lies in the fact that adaptive graph convolutional networks dynamically evolve the spatial dependency relationships between nodes through learnable adjacency matrices, breaking through the limitations of traditional graph convolutional networks that rely on fixed topologies. The multi-head attention mechanism extracts multi-scale temporal features in parallel from the load sequence. The two work together to achieve deep integration of spatiotemporal features. Experiments on the ISO-NE and BuildingsBench datasets show that the model maintains the highest accuracy in both short-term and medium to long-term predictions, with root mean square error, mean absolute error, and mean absolute percentage error of 46.88MW, 34.29MW, and 5.13%, respectively. Its anti-interference ability and inference speed are also superior to mainstream comparison models. The results indicate that the MS-AGCN-MHA model can effectively improve the robustness and accuracy of load forecasting in complex power grid environments, providing reliable technical support for real-time scheduling of smart grids.

Key Words

Power load forecasting; Adaptive graph convolutional network; Multi-head attention mechanism; Multi-scale features; Spatiotemporal modeling

* Shenzhen Power Supply Bureau Co., Ltd, Shenzhen, 518000, China; e-mail: lizien@mailpo.uu.me
** Southern Power Grid Energy Development Research Institute Co., Ltd, Guangzhou, 511466, China; e-mail: WuXufeng11@126.com
Corresponding author: Xufeng Wu

Recommended by: Zhenling Liu.
(DOI:10.2316/J.2025.203-0657)

1. Overview

In recent years, the large-scale integration of new energy represented by wind power and photovoltaics, as well as the diversification of electricity load structures, have jointly promoted the "dual high" characteristics of a high proportion of renewable energy and a high proportion of power electronic equipment in the power system. As of the end of 2024, the proportion of wind and solar installed capacity in China has reached 42.0%, and the proportion of power generation has exceeded 18%, resulting in a significant increase in the non stationarity and multi-scale volatility of load sequences, which has led to a shift in load forecasting (LF) from single time series modeling to comprehensive modeling that integrates multidimensional features. The high proportion of distributed energy and extreme weather events further increases the uncertainty of load fluctuations, bringing greater pressure to the safe operation and economic dispatch of the power grid [1]-[3].

Deep learning (DL) technology is widely used in LF tasks due to its powerful feature extraction and pattern recognition capabilities, and has demonstrated superior performance in a variety of scenarios. To increase the precision of power LF, Pentsos et al. suggested a novel optimized hybrid model integrating Long Short-Term Memory (LSTM) networks and Transformers. By leveraging the strengths of both architectures and incorporating geographical and user behavior factors, the model achieves reliable electricity load predictions [4]. To increase the precision of short-term power demand forecasting, Duan et al. suggested a deep neural network model that combines deep LSTM networks, threshold periodic units, and boosting techniques. This method enhanced model fusion through the Boosting algorithm, significantly improving the operational efficiency and reliability of power systems [5]. To increase the precision of power load data prediction, Yang et al. suggested a prediction model built on an enhanced extreme learning machine. This method simplified the model structure and reduced training errors by introducing the Pinball Huber robust loss function and genetic algorithm optimization [6]. To increase the grid integration efficiency of renewable energy power generation, Banik et al. sug-

gested a stacked ensemble model that combined extreme gradient boosting and random forests. This method improved the long-term and short-term prediction accuracy of power load by predicting power load and integrating environment-dependent data [7].

In terms of spatial information utilization, the graph convolutional network (GCN) has attracted attention for its ability to model spatial relationships between nodes using power grid topology. Wu et al. proposed a prediction model based on GCN and gated recurrent units to improve the accuracy of multi-area power LF. By constructing an adjacency matrix to analyze spatiotemporal correlations, they achieved superior prediction performance [8]. To increase the precision of short-term LF, Zhang et al. suggested a DL model that combines GCN and tree-like neural models. They significantly improved predicting performance by extracting geographical information from sample load data using GCN [9]. In addition, the introduction of attention mechanisms (AMs) improves long-range dependency capture capabilities and shows potential in terms of feature extraction diversity [10]. To enhance short-term power LF performance, Feng et al. suggested a hybrid model that combined bidirectional LSTM networks, temporal convolutional networks, and AMs. This method identified and weighted key information in multi-dimensional time series through AMs, achieving more accurate short-term LF [11]. Jiang et al. proposed a new dynamic time-dependent model to improve the accuracy of short-term LF. This method achieved more effective multi-step time-dependent learning by capturing similarities between different timestamps through the multi-head AM (MHA) [12].

The research paradigm of power load forecasting is shifting from a single model structure innovation to collaborative optimization of feature engineering, loss function, and model architecture. This transformation is due to the higher requirements placed on prediction systems by the high proportion of renewable energy grid integration and frequent extreme weather events. Traditional feature engineering often directly uses meteorological observations, while the latest research in 2025 emphasizes the exploration of deep dynamic correlations between meteorological factors and loads. For example, Pu et al. used the maximum information coefficient to assign weights to meteorological variables in different weather scenarios, and then combined them with a gate mechanism and a multi-layer self-attention network to effectively capture their global dependence on load [13]. Meanwhile, in order to enhance the robustness of the model under extreme fluctuations, the design of the loss function is no longer limited to mean square error. Quantile regression and uncertainty-based loss functions have become cutting-edge directions. Research and practice have shown that quantile regression models can effectively construct load prediction intervals, providing a basis for risk assessment [14].

Although the aforementioned methods have made progress in prediction accuracy, feature extraction, and spatiotemporal modeling, they still exhibit limitations in handling dynamic spatial topology changes, integrating spatiotemporal multiscale features, and achieving model generalization capabilities. This is particularly evident when addressing scenarios with high renewable energy penetration and significant load fluctuations. To address these challenges, this study proposes a multi-scale power load forecasting model integrating Adaptive Graph Convolutional Networks (AGCN) and Multi-Heterogeneous Analysis (MHA), termed MS-AGCN-MHA. By adaptively constructing the spatial correlation matrix of power system nodes, it integrates multi-order topological information to enhance spatial feature representation capabilities. Additionally, it utilizes the MHA mechanism to capture key information across different time scales in the temporal dimension, achieving deep fusion modeling of load spatiotemporal characteristics. The research aims to improve the accuracy and stability of multiscale load forecasting, providing more reliable decision-making support for intelligent dispatch in complex power systems. The innovation lies in simultaneously achieving dynamic spatial topology updates, multi-scale temporal dependency modeling, and frequency-domain feature integration. This approach enhances prediction performance in complex power system environments, providing efficient and reliable technical support for secure dispatch and scientific planning in smart grids.

2. Methods and Materials

2.1 Prediction Algorithm Based on AGCN and MHA

To address the multi-scale characteristics of power load data in terms of time and space, this study proposes an AGCN-MHA algorithm. Fig. 1 depicts the AGCN-MHA algorithm's fundamental structure.

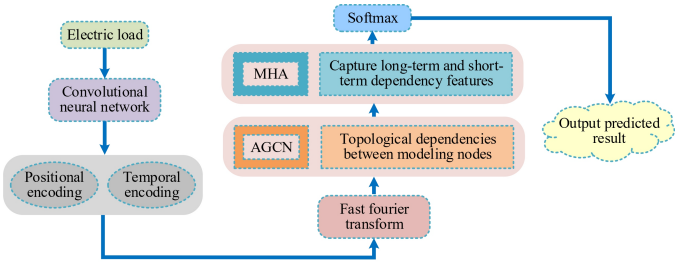

Figure 1. AGCN-MHA Prediction Algorithm Diagram

In Fig. 1, the input end first receives power load sequence data from multiple regions. The original load sequences undergo local feature encoding through the convolutional neural network feature extraction unit. Position encoding and time encoding are introduced at this stage to enhance the model's ability to represent spatiotemporal sequences. Next, the encoded features are sent to the Fast Fourier Transform (FFT) module for frequency domain transformation to extract multi-scale information of the load signal in the frequency domain. To effectively integrate frequency domain and time domain information, the significant frequency domain components extracted by FFT are studied

as independent feature vectors, and they are concatenated with the spatial enhanced features output by AGCN in the feature dimension to form the input of the MHA module. On this basis, each attention head in the MHA mechanism generates its key and value matrices from a composite feature that combines spatiotemporal and frequency-domain information when calculating temporal dependencies. This enables the model to simultaneously balance spatiotemporal context and frequency domain characteristics when capturing temporal dependencies, thereby achieving deep adaptive fusion of frequency domain spatiotemporal features. Subsequently, the core component AGCN dynamically characterizes the correlations between power load nodes through a learnable adjacency matrix, overcoming the limitations of traditional GCNs that rely on static adjacency matrices. This enables adaptive adjustment of node relationship weights across different prediction stages. For temporal modeling, MHA captures long-range dependencies in time series, enabling the model to focus on distinct temporal patterns across different subspaces. To mitigate gradient vanishing, the research introduces residual unit structures between the GCN and attention network, while applying SoftMax normalization after each layer to stabilize feature distributions.

In addition, AGCN and MHA mechanisms achieve deep fusion through a cascading approach of "spatial priority, temporal successor". The model first uses AGCN to dynamically aggregate the spatial features of each power node on each time slice, generating a spatially enhanced node feature sequence. Subsequently, the MHA mechanism acts on the temporal dimension of the sequence, capturing long-term dependencies by calculating attention weights across time steps. This "space-time" alternating processing mode can stack multiple layers in the network, allowing deep AGCN to further optimize spatial relationships based on the temporal context provided by MHA, while deep MHA can mine complex temporal patterns based on more accurate spatial features, thereby achieving collaborative evolution and joint modeling of spatiotemporal features. Finally, a multi-layer perceptron is used to perform regression output for the predicted values. Meanwhile, the learnable dynamic adjacency matrix does not directly participate as a parameter in the calculation of queries, keys, and values in MHA. The spatially enhanced node feature sequence output by AGCN is the direct input for MHA's time attention calculation. An optimized adjacency matrix can aggregate more relevant spatial neighbor information, resulting in feature sequences with less noise and clearer spatial semantics. When the MHA mechanism performs operations on the high-quality sequence, its calculated attention weights can more accurately focus on the truly critical temporal patterns, rather than being misled by local spatial noise. Fig. 2 depicts the AGCN module's basic structure.

In Fig. 2, first, the input power load data goes through a feature extraction module, which converts the time series data into node feature representations suitable for graph convolution operations. Next, AGCN calculates the similarity between node embedding vectors to generate a learnable adjacency matrix that reflects the spatial relationships between nodes [15]. The node features are aggregated using graph convolution operations, and the neighborhood range of feature aggregation is dynamically determined by a learnable adjacency matrix rather than relying on a predefined topological structure [16]-[17]. The model performs deep mining of local spatiotemporal information through feature-weighted aggregation of neighboring nodes, while also taking into account global features at different scales. Finally, after undergoing multi-layer graph convolution and nonlinear activation function (AF) processing, the network outputs a high-dimensional representation that integrates multi-scale features, which serves as input for subsequent prediction modules. The adaptive adjacency matrix is dynamically generated by calculating similarities among node features, with its core being a learnable transformation matrix, as shown in Equation (1).

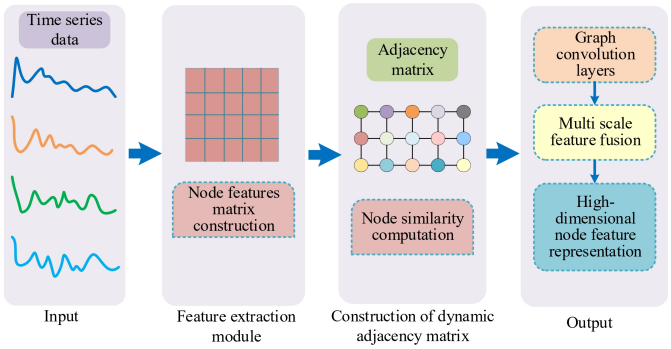

Figure 2. AGCN Basic Structure Diagram

$$A_{(l)} = softmax(ReLU(X_{(l)}W_{a(l)}X(l)T)) \quad (1)$$

In Equation (1), $A_{(l)}$ denotes the adaptive adjacency matrix of layer l, reflecting the dynamic spatial dependency strength between nodes. $softmax$ represents the normalization operation, ensuring the sum of weights in each row of the adjacency matrix equals 1. $ReLU$ denotes the rectified linear unit activation function. $X_{(l)}$ represents the feature matrix of input nodes in layer l. $W_{a(l)}$ denotes the learnable weight matrix constructed from the adjacency matrix of layer l, whose function is to map node features onto a latent space where similarity can be efficiently computed. The study employs the Xavier uniform distribution strategy to initialize the learnable weight matrix. This strategy automatically adjusts the initialization range based on the number of input and output neurons in the layer, helping maintain stable gradient flow during early training and accelerating model convergence. During the model training process, the learnable weight matrix is optimized along with all other model parameters using the backpropagation algorithm. Its gradient is calculated based on the total loss function and updated through the Adam optimizer. The update frequency of this weight matrix is consistent with the main network of the model, that is, it is updated once per training batch. The expression for the convolution operation is shown in Equation (2).

$$H_{(l+1)} = \sigma(A_{(l)}H_{(l)}W_{(l)}) \quad (2)$$

In Equation (2), $H_{(l+1)}$ denotes the FM of the $l+1$-th layer nodes. σ denotes the nonlinear AF. $H_{(l)}$ denotes the FM of the $l-th$ layer nodes. $W_{(l)}$ denotes the learnable WM of the $l-th$ layer graph convolution, used to perform a linear transformation on aggregated neighbor features. Fig. 3 depicts the MHA module's fundamental structure.

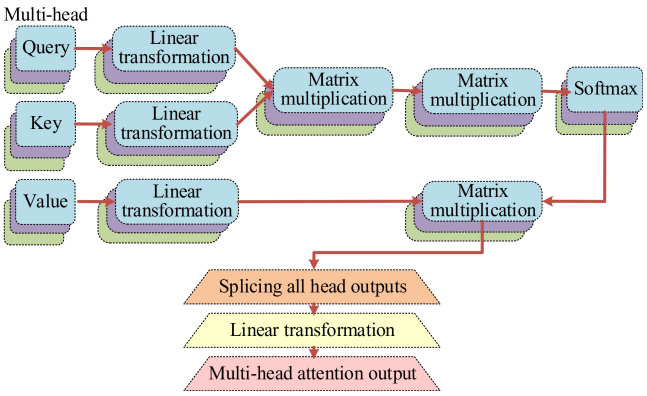

Figure 3. Schematic Diagram of the MHA Mechanism

Fig. 3 illustrates the core workflow of the MHA mechanism in temporal feature modeling. First, the input feature vector is parallelly mapped to multiple attention heads. Each head generates three matrices—query, key, and value—through independent linear transformations. These matrices compute attention weights via dot products and undergo SoftMax normalization, reflecting the correlations between different time points in the sequence. Subsequently, each head performs a weighted summation of the value vector based on these weights, extracting feature information from distinct subspaces. The concatenated outputs from all heads are then fused through a linear transformation to form the final feature representation. This enables the model to simultaneously capture multiple temporal dependency patterns, enhancing its ability to model long-range dependencies and complex temporal relationships [18]. Furthermore, the multi-head attention mechanism leverages parallel computation advantages, improving computational efficiency and model expressiveness. The MHA mechanism captures diverse temporal dependency patterns by computing multiple attention heads in parallel. Its final output is obtained by concatenating and linearly mapping the outputs from each attention head, as shown in Equation (3).

$$MHA(X) = Concat(head_1, head_2, \ldots, head_h)W_O \quad (3)$$

In Equation (3), $MHA(X)$ represents the final MHA output. $Concat$ represents concatenating the outputs of all heads in the feature dimension. $head_h$ is the output of the $h-th$ attention head. W_O represents the linear mapping WM. The calculation of $head_h$ is shown in Equation (4).

$$head_h = Attention(Q \cdot W_h^Q, K \cdot W_h^K, V \cdot W_h^V) \quad (4)$$

In Equation (4), Q, K, and V represent the query, key, and value matrices, respectively. W_h^Q, W_h^K and W_h^V represent the learnable projection matrices for the $h-th$ attention head, respectively.

2.2 Construction of a Multi-scale Power LF Model Based on the AGCN-MHA Algorithm

After completing the construction based on the AGCN-MHA algorithm, it is necessary to embed it into a complete power LF framework to achieve end-to-end processing from raw data to prediction results. Based on this, the study designs a multi-scale power LF model, MS-AGCN-MHA, based on the AGCN-MHA algorithm. Its overall structure is shown in Fig. 4.

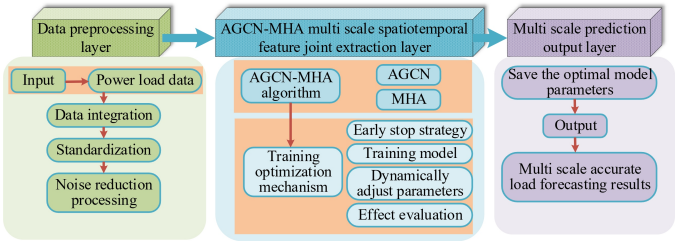

Figure 4. MS-AGCN-MHA Model Architecture

As shown in Figure 4, the MS-AGCN-MHA model first integrates historical load data with multi-domain external factors such as temperature and date type through data acquisition. It then constructs a standardized spatiotemporal dataset via data preprocessing and denoising. Subsequently, it enters the AGCN-MHA multi-scale spatiotemporal feature joint extraction layer, where it learns real-time spatial dependencies between grid nodes through a dynamic adjacency matrix while concurrently extracting features at different temporal scales via the MHA mechanism. During training, an optimizer with early stopping is employed for parameter iteration. The graph convolution order and attention head weights are dynamically adjusted based on the validation set. A multi-scale composite loss function was developed to synchronously optimize the accuracy of all prediction steps for multi-scale prediction tasks. The loss function is the weighted sum of losses over multiple prediction time scales, defined as equation (5).

$$\zeta_{\text{total}} = \sum_{t \in S} \lambda_t \cdot \mathcal{L}(\hat{Y}_t, Y_t) + \eta \parallel \Theta \parallel_2^2 \quad (5)$$

In equation (5), ζ_{total} represents the total loss function value; λ_t represents the adjustable weight coefficient corresponding to the time scale t; $\hat{Y}_t$ and Y_t respectively represent the predicted values of the model and the actual load values on the time scale t; $\zeta(\hat{Y}_t, Y_t)$ represents the basic loss function at time scale t; η represents the regularization coefficient; Θ represents the set of all trainable parameters in the model. Through this composite loss function, the model is forced to learn and optimize the predictive ability of all target time scales simultaneously during the training process, rather than focusing solely on a single scale, ensuring the balance and robustness of its multi-scale predictive performance. It guarantees that the model performs at its best on the validation set by continuously modifying parameters to enhance prediction performance and performing real-time evaluation of model efficacy during training.

After training is complete, the model weights with the best performance are saved, and the trained model is ultimately applied to the task of predicting power load, producing multi-scale, accurate load prediction results. Fig. 5 illustrates the power data preparation procedure in detail.

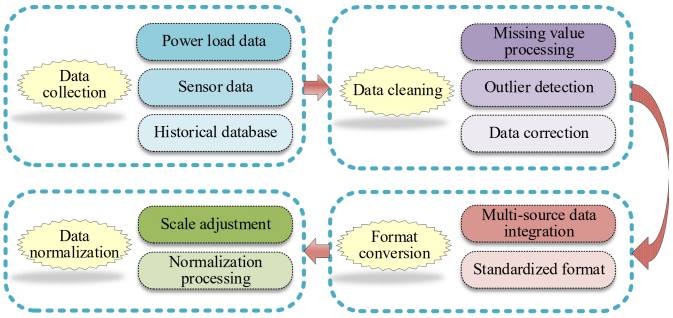

Figure 5. Power Load Dataset Preprocessing Process

Fig. 5 shows that the process primarily consists of four steps: data collection, data cleaning, data format conversion, and data normalization. First, the data collection module is responsible for collecting raw time-series data on power load and related auxiliary information from the power monitoring system, sensors, and historical databases. Next, the data cleaning module detects and corrects missing values and outliers to ensure data integrity and accuracy. In this study, the K-nearest neighbor imputation method is used to fill in long-term missing data, as shown in Equation (6) [19].

$$x_t = \frac{\sum_{i=1}^{k} w_i x_{t_i}}{\sum_{i=1}^{k} w_i} \tag{6}$$

In Equation (6), x_t is the load value at time t to be filled. i represents the sample index. w_i is the weight of the i th nearest neighbor sample. k is the number of nearest neighbors. x_{t_i} represents the load value at a similar time t_i. Next, the data format conversion module unifies multi-source heterogeneous data into a standardized format for subsequent modeling use. Finally, the data normalization module performs scale adjustment and normalization processing on the data to eliminate dimensional effects and improve the stability and convergence speed of model training. Furthermore, the dataset must undergo noise reduction processing before entering the modeling stage. The detailed steps are shown in Fig. 6.

In Fig. 6, data denoising primarily involves three stages: feature extraction, noise identification, and noise filtering. The purpose of performing feature extraction prior to anomaly detection and filtering is as follows: raw power load sequences are a mixture of signals and noise, making it difficult to effectively distinguish legitimate load fluctuations from genuine noise interference through direct manipulation. By first extracting multidimensional features from the raw data—including time-domain statistics, frequency-domain components, and external factors like temperature and date—a richer informational context is established. This makes statistical deviations of outliers and noise patterns more pronounced relative to normal load behavior.

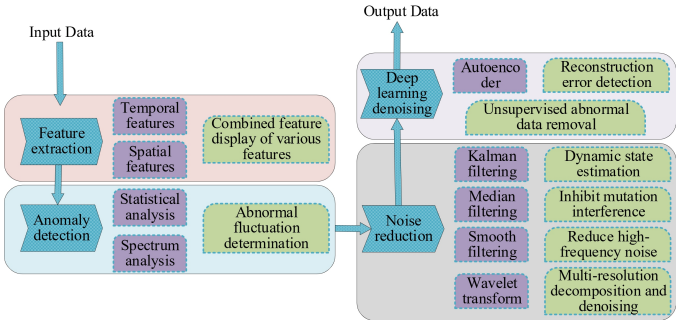

Figure 6. Basic Steps for Dataset Denoising (Feature Representation, Feature Selection, and Data Denoising)

Following noise identification, multiple complementary filtering techniques are applied to address different noise types in power load data. Kalman filtering is suitable for online processing of linear systems with Gaussian noise, but it heavily relies on model accuracy and lacks the capability for handling nonlinear, non-stationary noise. Median filtering effectively removes spike-like anomalies with simple computation, but may obscure genuine rapid load variation details when suppressing high-frequency fluctuations. Smoothing filters effectively mitigate high-frequency random fluctuations but introduce phase lag and obscure the true dynamic trends of the load. Wavelet transforms achieve separation of noise and detail in non-stationary signals through multi-resolution analysis, with their effectiveness significantly influenced by the selection of basis functions and thresholds. In addition, in the deep learning stage, the research uses the reconstruction error detection mechanism of autoencoders to remove abnormal data. For an input feature vector, its reconstruction error is defined as equation (7).

$$\mathcal{L} = ||x - fdec(fenc(x))||_2 \tag{7}$$

In Equation (7), $\mathcal{L}$ displays the reconstruction error. $\mathbf{x}$ displays the input feature vector. f_{dec} represents the decoder function. f_{enc} represents the encoder function. In research, autoencoders are not a complete replacement for traditional filters, but rather serve as a supplementary and enhancing step. Specifically, traditional filters excel at handling noise with clear statistical patterns or specific frequency bands, while deep models such as autoencoders can detect and reconstruct complex nonlinear anomalies that do not conform to normal load patterns by learning the intrinsic manifold distribution of data. These anomalies often exceed the effective processing range of traditional filters. By combining traditional methods for initial denoising with deep learning for fine cleaning, a hybrid strategy is implemented to maximize the retention of key temporal information and comprehensively suppress noise interference. The final result is a high-quality dataset that retains key temporal information and has low noise interference, providing reliable input for subsequent feature selection and prediction models.

Table 1
Experimental Environment Configuration Table

Category	Device name	Specifications	Category	Device name	Specifications
Hardware	GPU	NVIDIA Tesla V100 32GB	Software	Operating system	Ubuntu 18.04 LTS
	CPU	Intel Xeon Gold 5218 @ 2.3 GHz		Programming language	Python 3.8.10
	Memory	128 GB DDR4		DL framework	PyTorch 1.10.0
	Storage	2TB SSD		Visualization tools	Matplotlib 3.4.3, Seaborn 0.11.2
	-	-		Compute unified device architecture	CUDA 11.3

3. Results

3.1 Experimental Environment and Parameter Sensitivity Verification

Simulation tests are carried out to verify the MS-AGCN-MHA model's efficacy. Table 1 displays the experimental environment's precise configuration. The main training parameters include a learning rate of 0.001, a batch size of 64, an optimizer selection of Adam, and 100 training rounds. The study used the Daubechies 4 wavelet basis for 5-layer decomposition, and adaptive thresholding was applied to the detail coefficients of each layer. The threshold was determined according to the Donoho-Johnstone criterion. The process noise covariance of Kalman filtering is set to 0.01, and the observation noise covariance is set to 0.1. This parameter combination has been verified through grid search to have the best smoothing effect and tracking ability on load data. All denoising hyperparameters are determined through cross-validation and maintain consistency across different datasets to ensure comparability. The datasets used are the publicly available ISO-NE power load dataset and the BuildingsBench dataset.

The study initially examines how the number of graph convolution layers (GCLs) and attention heads affects model performance using the experimental setup displayed in Table 1. The evaluation metrics of mean absolute error (MAE) and root mean squared error (RMSE) have been chosen. The results are shown in Fig. 7.

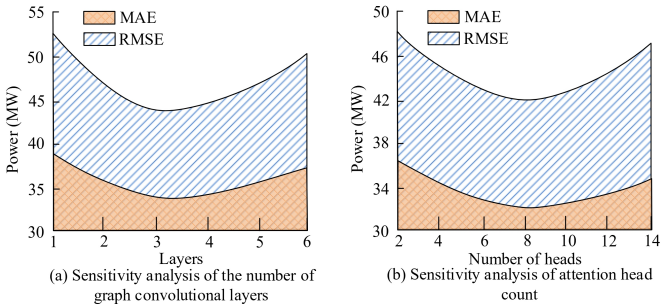

Figure 7. Sensitivity Analysis of the Number of Convolutional Layers and the Number of Attention Heads

As shown in Fig. 7(a), as the number of layers increases from 1 to 3, the Mean Absolute Error (MAE) decreases from approximately 38.79 MW to 34.23 MW, while the Root Mean Square Error (RMSE) decreases from 52.48 MW to 44.76 MW. This phenomenon stems from the difficulty of adequately modeling cross-regional power grid topology using shallow networks. When the number of layers exceeds 3, over-smoothing caused by deep convolutions leads to the loss of spatial details, causing the MAE to rebound to 35.32 MW. This is because in deep GCN, node features are iteratively aggregated through adjacency matrices, resulting in similar feature representations of different nodes in the graph, thereby blurring the key spatial details that originally helped distinguish different node load patterns. As shown in Fig.7(b), when the attention head rises from 2 to 8, the MAE reduces from 36.37 MW to 32.82 MW, and the RMSE decreases from 48.03 MW to 42.79 MW. But when the number of heads exceeds 8, performance drops due to multi-head redundancy. Redundancy is reflected in the fact that some attention heads may learn highly similar or unimportant temporal patterns, leading to inefficient utilization of model capacity and increasing the risk of overfitting due to an increase in parameters. The study uses independent linear projection and final concatenation mapping operations in the MHA mechanism to naturally differentiate and integrate the focus points of different heads. During the training process, the model spontaneously drives different heads to focus on different feature subspaces through gradient descent. When the number of heads is 8, this mechanism achieves the optimal balance between pattern diversity and parameter efficiency. Based on Figure 7, it can be seen that the number of graph convolutional layers dominates the depth and receptive field of spatial topology modeling, while the number of attention heads determines the granularity and diversity of temporal multi-scale feature decomposition. Both have clear optimal values, and excessive depth or quantity can impair generalization performance due to an unreasonable increase in model capacity. For this purpose, the study selected 3 layers of graph convolutional layers and 8 attention heads as the core parameter configurations of the model.

3.2 MS-AGCN-MHA Model Prediction Performance Verification

To verify the contribution of each core module in the MS-AGCN-MHA model to power LF performance, ablation experiments are designed. The experimental variants are set as follows: The adaptive adjacency matrix is removed and replaced with a distance-based static adjacency matrix (V1). Single-head attention is used to replace MHA (V2). The fast Fourier transform module is removed (V3). The residual connection between GCN and attention is removed (V4). The complete model configuration is used (V5). Mean absolute percentage error (MAPE) and RMSE are used in the study as detection indicators. Fig. 8 dis-

plays the findings.

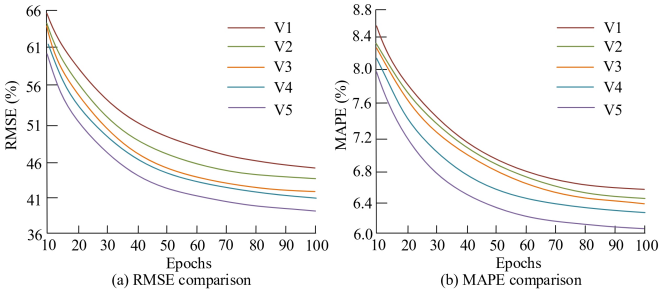

Figure 8. Comparison of RMSE and MAPE for Different Variants in Ablation Experiments Under Different Training Iterations

In Fig. 8(a), the complete model (V5) exhibits the lowest RMSE across all training iterations and converges steadily, reaching 38.98 MW at 100 iterations, demonstrating excellent convergence and prediction accuracy. In contrast, the V1 model, which removes the adaptive adjacency matrix, has the highest error, remaining at 45.31 MW even at 100 iterations. This indicates that the static adjacency matrix struggles to dynamically capture spatial dependencies between nodes. The V4 model, which removes residual connections, converges more slowly and has larger errors during training. This implies that the residual structure improves training stability and successfully reduces gradient vanishing. In Fig. 8(b), the MAPE of the complete model V5 decreases from 7.96% to 6.06%, demonstrating the best error control capability. V1 exhibits a high MAPE in all rounds, with a maximum of 8.54%. This validates the importance of the adaptive adjacency matrix for spatial dependency modeling. In summary, each core module is indispensable for improving model performance, promoting the accuracy and robustness of LF through synergistic effects. On this basis, the study introduces three mainstream LF models for comparison, namely spatio-temporal GCN with attention (ST-GCN-Attention), multi-scale convolutional neural network with MHA (MSCNN-MHA), and GCN with gated recurrent unit fusion (GCN-GRU). Fig. 9 displays the findings.

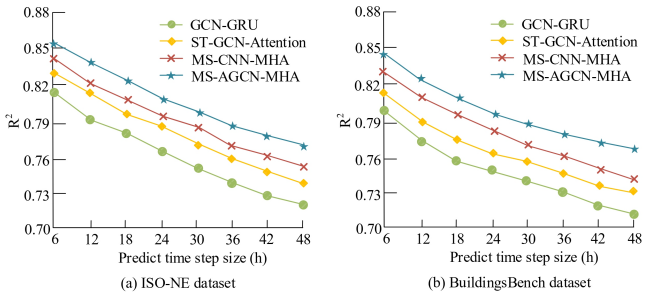

Figure 9. Comparison of R2 for Four Models Under Different Datasets

In Fig. 9(a), the MS-AGCN-MHA model outperforms the comparison models at all time steps on the ISO-NE power load dataset. The advantage is particularly evident in short-term predictions, achieving 0.854 and 0.837 at the 6-hour and 12-hour prediction steps, respectively. MS-AGCN-MHA maintains the highest value of 0.769 at 48 hours, indicating its strong spatial-temporal feature fusion capability. In Fig. 9(b), the R2 variation curve on the BuildingsBench dataset shows that the R2 values of all models are generally lower than those on the ISO-NE dataset. When making short-term forecasts for 6h and 12h, MS-AGCN-MHA still performs best, achieving 0.843 and 0.826, respectively, but slightly lower than the ISO-NE data. The gap becomes more pronounced in medium- and long-term forecasts. This difference indicates that in high-noise, heterogeneous building energy consumption scenarios, the MS-AGCN-MHA model can effectively mitigate accuracy degradation. Meanwhile, the study tests the anti-interference capabilities of the four models under four interference scenarios: Gaussian noise interference (SNR=10dB), pulse interference (5% data points), load sudden changes (± 30% step), and extreme cold wave events (-15°C continuous). The results are shown in Fig. 10.

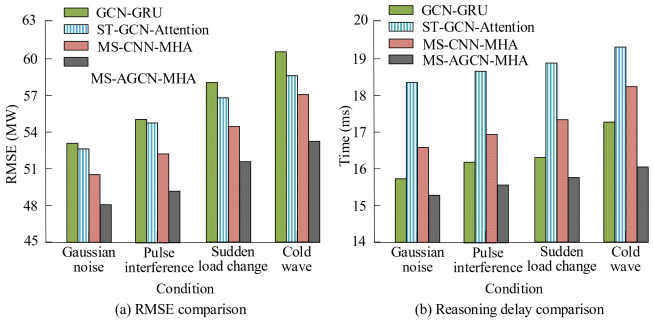

Figure 10

As shown in Fig. 10(a), the MS-AGCN-MHA model exhibits an RMSE of 47.96 MW under Gaussian noise interference, representing an average reduction of 8.70% compared to other models and demonstrating its outstanding robustness in noisy environments. When confronted with non-stationary anomalies such as pulse disturbances and load surges, the MS-AGCN-MHA model achieved RMSEs of 49.10 MW and 51.31 MW, respectively. This indicates that its adaptive adjacency matrix dynamically adjusts spatial dependencies, effectively mitigating the impact of anomalous data on predictions. Combined with Fig. 10(b), MS-AGCN-MHA shows stable performance in delay control, with inference delays of 15.17ms, 15.52ms, 15.76ms, and 15.95ms under four interference scenarios: Gaussian noise interference, pulse interference, load mutation, and extreme cold wave events. This delay level has significant engineering application value: compared to the 2-4s data refresh cycle of modern SCADA systems, the model can complete more than 100 full network load predictions in a single cycle, providing ample margin for scheduling systems to achieve multi-scenario simulation and rolling optimization. Finally, the study further introduced two advanced models, GraphWaveNet and MTGNN, as comparative benchmarks to briefly investigate the effectiveness of the methods. The results are shown in Table 2.

Table 2
A Comprehensive Performance Comparison of the Four Models

Metric	ST-GCN-Attention	MSCNN-MHA	GCN-GRU	Graph Wave Net	MTGNN	MS-AGCN-MHA
RMSE (MW)	51.23	49.17	52.41	48.95	48.15	46.88
MAE (MW)	38.47	36.72	39.05	36.9	36.25	34.29
MAPE (%)	5.87	5.43	6.12	5.38	5.22	5.13
Peak Load Error (%)	7.18	6.83	7.52	6.71	6.45	6.13
Number of Parameters (M)	6.23	5.84	4.52	7.15	6.92	6.83
Training Time per Epoch (s)	134.76	120.34	109.88	145.83	142.17	139.62
GPU Memory Usage (GB)	7.48	6.81	5.36	7.86	7.64	7.23
Pearson Correlation Coefficient	0.91	0.92	0.90	0.93	0.93	0.94

Table 3
Performance Comparison of Various Models in High Renewable Energy Penetration Scenarios

Model	RMSE (MW)	MAE (MW)	MAPE (%)	Pearson correlation coefficient	CRPS	Pinball Loss ($\times 10^{-2}$)
ST-GCN-Attention	68.54	51.89	7.82	0.87	5.42	3.89
MSCNN-MHA	65.91	49.73	7.45	0.88	5.18	3.71
GCN-GRU	71.02	53.45	8.16	0.86	5.65	4.02
MS-AGCN-MHA	61.23	45.16	6.73	0.91	4.75	3.38

According to Table 2, compared with the two newly added advanced benchmark models, the MS-AGCN-MHA model still maintains a comprehensive advantage. Specifically, GraphWaveNet and MTGNN demonstrate superior performance compared to traditional ST-GCN Attention and GCN-GRU models by introducing mechanisms such as diffusion convolution and graph learning layers. This validates the importance of dynamic graph structure learning in modern load forecasting. However, MS-AGCN-MHA achieved finer spatiotemporal feature extraction through closer collaborative design between AGCN and MHA. It achieved the lowest values in RMSE and MAE, with 46.88MW and 34.29MW, respectively, indicating its significant advantages in capturing the complex spatiotemporal dependencies of power loads. Although the parameter count of MS-AGCN-MHA is comparable to that of MTGNN, its training time and video memory usage are 139.62s and 7.23GB, respectively, which are lower than the comparison model, confirming the computational efficiency of the model structure design. In addition, the peak load error of the model is 6.13%, and the Pearson coefficient is 0.94, further verifying its adaptability to extreme power grid conditions and trend capture accuracy, providing reliable support for safety warning in actual power grid dispatch.

3.3 Generalization Ability Verification in High Renewable Energy Penetration Scenarios

To validate the universality and robustness of the MS-AGCN-MHA model in renewable energy-intensive systems, the study supplemented the dataset obtained from the European Grid Open Platform (ENTSO-E) for testing, with a 48-hour prediction horizon. The results are shown in Table 3.

According to Table 3, the MS-AGCN-MHA model still maintains the best overall performance in complex scenarios with high penetration of renewable energy. Its RMSE and MAPE are 61.23 MW and 6.73%, respectively, both significantly lower than the comparison model. The Pearson correlation coefficient reached 0.91, indicating that the predicted curve is highly consistent with the actual trend of net load changes. This is mainly due to the core mechanism of the model: the AGCN module adaptively captures the spatial correlations between grid nodes that undergo drastic changes due to renewable energy injection through a dynamic adjacency matrix. Meanwhile, the MHA mechanism can effectively decouple and learn non-stationary temporal features caused by wind and solar fluctuations, which are superimposed on traditional load patterns. In contrast, the GCN-GRU model performs relatively poorly due to its insufficient ability to capture spatiotemporal dynamic changes; Although the ST-GCN Attention and MSCNN-MHA models can partially cope, their fixed spatial assumptions or single-scale temporal modeling methods limit their performance limits in strong fluctuation scenarios. MS-AGCN-MHA also performs the best in probability prediction tasks, with the lowest Continuous Ranked Probability Score (CRPS) and Pinball Loss of 4.75 and 3.38 × 10-2, respectively. The predicted distribution of the research model is closer to the true data distribution and has good calibration throughout the entire prediction interval. According to Table 3, the MS-AGCN-MHA model is not only applicable to traditional load datasets but also demonstrates excellent prediction accuracy and generalization ability in modern power system environments with higher penetration and greater uncertainty of renewable energy, providing a more comprehensive basis for its practical deployment in future smart grids.

3.4 Model's Generalization Ability Verification on Chinese Power Grid Data

To test the performance of the MS-AGCN-MHA model in actual power grid scenarios in China, a generalization experiment was conducted on the 2023 load dataset of the Guangdong power grid. The load in this region is significantly affected by temperature, with a high proportion of

Table 4
Performance Comparison of Various Models in the Guangdong Power Grid Scenario in China

Model	RMSE (MW)	MAE (MW)	MAPE (%)	Peak load error (%)	CRPS	Pinball Loss ($\times 10^{-2}$)
ST-GCN-Attention	86.45	63.27	8.91	9.23	6.88	4.85
MSCNN-MHA	83.92	61.84	8.65	8.97	6.59	4.66
GCN-GRU	89.13	65.42	9.27	9.58	7.12	5.03
MS-AGCN-MHA	78.36	56.19	7.82	8.14	6.05	4.27

Table 5
Interface Design Specifications

Interface layer	Specification content	Standards/Protocol recommendations
Data input	1. Real-time load data (all nodes)	IEC 61970 CIM/CIS
	2. Breaker status and topology	IEC 61850 SCL
	3. Renewable energy ultra-short-term forecast	IEC 61400-25
	4. Meteorological monitoring data	Custom JSON/AVRO
Service interface	1. Synchronous prediction request/response	RESTful API/gRPC
	2. Asynchronous prediction task management	AMQP/MQTT
	3. Model metadata query	Swagger/OpenAPI 3.0
Data output	1. Node load prediction values (multi-scale)	Standard JSON Schema
	2. Prediction uncertainty intervals	(Includes timestamp, node ID,
	3. Model confidence and health status	value, quality code)

cooling load in summer and large changes in daily load rate, which puts higher demands on the prediction model. The experimental results are shown in Table 4.

According to Table 4, the MS-AGCN-MHA model still maintains optimal performance on the data of the Chinese power grid. Compared to international datasets, the absolute error of various models on Guangdong power grid data has increased, which is due to the higher load base and more complex operating characteristics of China's power grid. The RMSE and MAPE of the MS-AGCN-MHA model are 78.36 MW and 7.82%, respectively, which are significantly lower than the comparison model, while maintaining the lowest peak load error of 8.14%. In addition, the CRPS of the MS-AGCN-MHA model is 6.05, and the Pinball Loss is 4.27 $\times 10^{-2}$, which is still significantly lower than the comparison model. MS-AGCN-MHA effectively captures the power interaction patterns between regions within Guangdong Province through an adaptive adjacency matrix, and the MHA mechanism exhibits better adaptability to the unique holiday load changes in China.

3.5 Interface Design Specification

To achieve seamless integration between the model and the production control system, an interface architecture and core data interaction specification for the model and real-time control system have been proposed. It mainly includes five modules: historical database, data interface, MS-AGCN-MHA model, predictive service bus, and advanced application software. The specific interface design specifications are shown in Table 5.

According to Table 5, this design follows the microservice architecture concept, encapsulating the prediction function as independent and reusable services, and decoupling them from various advanced applications such as power flow calculation and security-constrained economic scheduling through the prediction service bus. The data interface layer is responsible for real-time alignment, format conversion, and quality verification of raw data obtained from historical databases, ensuring the timeliness and consistency of input data.

4. Summary

Grid safety dispatch imposes higher requirements on LF accuracy. In response to the challenges posed by the complex multi-scale characteristics, dynamic spatial dependencies, and strong non-stationarity in power LF, this research designed the MS-AGCN-MHA power LF model. This study proposed the AGCN-MHA prediction algorithm, which combined AGCN to characterize dynamic spatial dependencies and the MHA mechanism to capture long-term and short-term time series correlations. It was supplemented by Fourier transform FD analysis and residual units to achieve end-to-end load prediction. The experimental results showed that the model achieved R2 values of 0.854 and 0.843, respectively, on the ISO-NE and BuildingsBench datasets with a 6-hour prediction step length. The model maintained R2 values of 0.769 and 0.775 for 48-hour predictions. The RMSE values under interference scenarios such as Gaussian noise, pulse interference, load changes, and extreme cold weather were 47.96 MW, 49.10 MW, 51.31 MW, and 53.24 MW, respectively, with inference delays of 15.17 ms, 15.52 ms, 15.76 ms, and 15.95 ms, respectively. The Pearson correlation coefficient reached 0.94, significantly outperforming comparison models such as ST-GCN-Attention, MSCNN-MHA, and GCN-GRU. Research indicates that the synergistic effect of dynamic spatial modeling, multi-scale temporal feature extraction, and FD information fusion is a key mechanism for improving prediction accuracy, real-time performance, and inter-

ference resistance. However, there are still shortcomings in the research, such as a large number of model parameters and long training times. Future work should focus on optimizing the model structure while maintaining prediction performance and reducing computational overhead. Additionally, the potential applications of the model in power grid scenarios with higher proportions of new energy sources and more frequent dynamic changes in node topology should be explored.

References

[1] H. Dai, C. Zhang, Z. Zhen, and F. Wang, "Short-term net load forecasting based on temporal-spatial feature clustering and two-layer dynamic graph convolutional network modeling," *High Voltage Engineering*, vol. 50, no. 9, pp. 3914–3923, 2024.

[2] R. Cai, B. Xia, X. M. Zhu, L. Wang, J. R. Gu, and J. G. Tang, "Design of a risk model and analytical decision information system for power operation in the context of smart grid," *International Journal of Power and Energy Systems*, vol. 44, no. 10, 2024.

[3] B. Wei, C. Gao, H. Z. Cao, C. W. Wang, J. K. Wu, and H. L. Li, "An alternative optimisation model for reserve capacity of power system considering coordinative aggregation of large-scale renewable energy," *International Journal of Power and Energy Systems*, vol. 44, no. 10, 2024.

[4] V. Pentsos, S. Tragoudas, J. Wibbenmeyer, and N. Khdeer, "A hybrid lstm-transformer model for power load forecasting," *IEEE Transactions on Smart Grid*, vol. 16, no. 3, pp. 2624–2634, 2025.

[5] Q. Duan, Z. Chao, C. Fu, Y. Zhong, J. Zhuo, and Y. Liao, "Design of short-term power load forecasting model based on deep neural network," *Strategic Planning for Energy and the Environment*, vol. 43, no. 2, pp. 425–452, 2024.

[6] Y. Yang, H. Lou, Z. Wang, and J. Wu, "Pinball-huber boosted extreme learning machine regression: a multiobjective approach to accurate power load forecasting," *Applied Intelligence*, vol. 54, no. 17, pp. 8745–8760, 2024.

[7] R. Banik and A. Biswas, "Enhanced renewable power and load forecasting using rf-xgboost stacked ensemble," *Electrical Engineering*, vol. 106, no. 4, pp. 4947–4967, 2024.

[8] J. Wu, X. Lu, H. Liu, B. Zhang, S. Chai, Y. Liu, and J. Wang, "Ultra-short-term multi-region power load forecasting based on spearman-gcn-gru model," *Zhongguo Dianli*, vol. 57, no. 6, pp. 131–140, 2024.

[9] C. Zhang, Y. Yu, T. Zhang, K. Song, Y. Wang, and S. Gao, "Short-term load forecasting based on graph convolution and dendritic deep learning," *IEEE Transactions on Network Science and Engineering*, vol. 12, no. 4, pp. 3221–3233, 2025.

[10] Y. Yuan, Q. Yang, J. Ren, X. Mu, Z. Wang, Q. Shen, and Y. Li, "Short-term power load forecasting based on skdr hybrid model," *Electrical Engineering*, vol. 107, no. 5, pp. 5769–5785, 2025.

[11] Y. Feng, J. Zhu, P. Qiu, X. Zhang, and C. Shuai, "Short-term power load forecasting based on tcn-bilstm-attention and multi-feature fusion," *Arabian Journal for Science and Engineering*, vol. 50, no. 8, pp. 5475–5486, 2025.

[12] B. Jiang, H. Yang, Y. Wang, Y. Liu, H. Geng, H. Zeng, and J. Ding, "Dynamic temporal dependency model for multiple steps ahead short-term load forecasting of power system," *IEEE Transactions on Industry Applications*, vol. 60, no. 4, pp. 5244–5254, 2024.

[13] X. Pu and M. Zhang, "Short-term power load forecasting under multiple weather scenarios based on dual-channel feature extraction (dcfe)," *Applied Sciences*, vol. 15, no. 21, pp. 11733–11745, 2025.

[14] Z. Song, F. Xiao, Z. Chen, and H. Madsen, "Probabilistic ultra-short-term solar photovoltaic power forecasting using natural gradient boosting with attention-enhanced neural networks," *Energy and AI*, vol. 20, no. 2, pp. 312–325, 2025.

[15] J. Wen and Z. Wang, "Short-term power load forecasting with hybrid tpa-bilstm prediction model based on cssa," *Computer Modeling in Engineering & Sciences*, no. 7, pp. 749–765, 2023.

[16] F. Huang, H. Zhao, P. Yi, P. Li, and J. Peng, "An improved power load forecasting method based on transformer," *Modern Electric Power*, vol. 40, no. 1, pp. 50–58, 2023.

[17] L. Yi, J. Zhu, Y. Wang, J. Liu, S. Wang, and B. Liu, "Short-term power load forecasting based on orthogonal pca-lpp dimension reduction and igwo-bilstm," *Recent Patents on Mechanical Engineering*, vol. 16, no. 1, pp. 72–86, 2023.

[18] S. Hu, W. Cai, J. Liu, H. Shi, and J. Yu, "Refining short-term power load forecasting: an optimized model with long short-term memory network," *Journal of Computing and Information Technology*, vol. 31, no. 3, pp. 151–166, 2023.

[19] P. P. Groumpos, "A critical historic overview of artificial intelligence: issues, challenges, opportunities, and threats," *Artificial Intelligence and Applications*, vol. 1, no. 4, pp. 197–213, 2023.

Biographies

Xufeng Wu earned his Bachelor's degree in Electrical Engineering and Automation from China Three Gorges University in 2010. He has been working as an engineer at Shenzhen Power Supply Bureau Co., Ltd. since 2010. His research interests mainly focus on load forecast.

Min Chen earned his Bachelor's degree in Electrical Engineering and Automation from China Agricultural University in 2002. He subsequently obtained a Master's degree in Business Administration from Southwest Jiaotong University in 2005. He has been working as a Researcher at Shenzhen Power Supply Bureau Co., Ltd. since 2005. His research interests mainly focus on Electric Power Marketing.

Nan Dong obtained her Bachelor's degree in Electrical Engineering and Automation from Northeast Electric Power University in 2009 and her Master's degree in Power System and Automation from North China Electric Power University in 2012, with her research focusing on power supply and demand. She worked as a Researcher at Southern Power Grid Electric Power Research Institute from 2012 to 2017 and has been serving as a Project Manager at Southern Power Grid Energy Development Research Institute since 2018. Her research interests include power planning and load forecasting.

Yuwen Wu received the B.E. degree in electrical engineering from Huazhong University of Science and Technology, Wuhan, China, in 2019, and the M.S. degree in electrical engineering from Wuhan University, Wuhan, China, in 2022, respectively. He is currently a researcher with the Energy Development Research Institute of China Southern Power Grid Co., Ltd. His research interests include power supply & demand, power planning, load forecast, and the Electricity-Economy relationship.

Zhanzhi Liu earned his Master's degree in Electrical Engineering from The Hong Kong Polytechnic University in 2017, specializing in power system analysis. He worked as an Engineer at China Energy Engineering Group, Guangdong Electric Power Design Institute, from 2017 to 2022, and has been serving as a Researcher at Southern Power Grid Energy Development Research Institute since 2022. His research interests include demand response and load forecast.

Buying Su earned her Bachelor's degree in Electrical Engineering and Automation from Tianjin University in 2013 and her Master's degree in Electrical Engineering from the same university in 2016. From 2016 to 2023, she worked as an Engineer at Guangdong Electric Power Design Institute Co., Ltd., and since 2023, she has been serving as a Researcher at Southern Power Grid Energy Development Research Institute Co., Ltd. Her research interests include power system planning and electricity demand forecasting.