International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-33

https://doi.org/10.70917/ijcisim-2026-1104
Article

Interval forecasting of electric load based on an IBKA-Optimized MSTCN–BiLSTM–QR Model: an empirical study on the U.S. residential load data

Zixiang Long [1,*]

[1] Faculty of Electrical and Control Engineering, Liaoning Technical University, Huludao, Liaoning, 125105, China
* Correspondence author: 17821400812@163.com

Abstract: Considering the critical role of short-term power load forecasting in ensuring safe and stable power system operation, this study proposes an interval forecasting model based on an Improved Black-winged Kite Algorithm (IBKA)-optimized MSTCN–BiLSTM–QR framework. Multivariate time series inputs are first constructed, then a Multi-Scale Temporal Convolutional Network (MSTCN) extracts multi-scale temporal features, while BiLSTM captures bidirectional dependencies, enhancing representation of nonlinear load characteristics. Quantile Regression (QR) converts point forecasting into interval forecasting, enabling effective uncertainty characterization, and IBKA optimizes key hyperparameters to improve convergence and accuracy. Empirical analysis uses residential load data from four U.S. regions (U1–U4) and compares the model with CNN-LSTM, CNN-BiLSTM, TCN-LSTM, Transformer–TCN–GRU, TCN-Informer-BiGRU, and TCN-QRNN. Benchmark tests on six functions (F1–F6) show IBKA achieves optimal or near-optimal Best, Mean, and Std values, outperforming GWO and BKA, and demonstrating superior global search and convergence stability. For point forecasting, IBKA-MSTCN-BiLSTM-QR attains an average R^2 of 0.9929, outperforming TCN-Informer-BiGRU (0.9921) and Transformer–TCN–GRU (0.9904). MAPE decreases to 1.943%, ~26.8% lower than CNN-LSTM, and RMSE reaches 0.0372 kW, ~30.6% lower. Interval forecasting at 80% confidence yields PICP 0.798–0.818 and MPICD 0.063–0.075 kW; at 95% confidence, PICP 0.946–0.958 and MPICD 0.095–0.112 kW, indicating balanced coverage and interval width. Additional dataset validation confirms strong cross-dataset generalization. Ablation studies show full model performance (PICP = 0.844, MPICD = 0.0425 kW) declines when removing IBKA, MSTCN, or BiLSTM, highlighting the contribution of each component and the synergy of MSTCN and BiLSTM. Overall, the proposed model achieves high accuracy, robustness, and reliability in complex load scenarios, providing effective support for power system dispatch optimization and risk-aware decision-making.

Keywords: Short-term electric load forecasting; Interval forecasting; Improved Black-winged Kite Algorithm (IBKA); Multi-scale Temporal Convolutional Network (MSTCN); Bidirectional Long Short-Term Memory (BiLSTM); Quantile Regression (QR)

1. Introduction

Driven by the global energy transition and the goals of carbon peaking and carbon neutrality, power systems are rapidly evolving from traditional paradigms toward new systems dominated by renewable energy sources [1, 2]. With the increasing penetration of renewable energy, its inherent variability and uncertainty significantly amplify the randomness of net load in power grids, posing greater challenges to power balance, stable operation, and dispatch decision-making [3]. In this context, high-precision load forecasting, combined with coordinated optimization of generation–grid–load–storage resources,

has become an essential approach to achieving efficient energy utilization and supply–demand balance [4]. Meanwhile, with the deepening of electrification and the advancement of electricity substitution, the structure of the demand side is becoming increasingly diversified. The large-scale integration of electric vehicles, energy storage systems, and intelligent electrical devices, along with the influence of natural disasters and unexpected events [5], has led to more pronounced non-stationary, nonlinear, and stochastic characteristics in power load patterns [6]. The increasing complexity on both the supply and demand sides has significantly reduced the effectiveness of traditional load forecasting models that rely heavily on historical data, making them inadequate for the evolving needs of modern power systems [7]. Therefore, in the context of constructing new-type power systems, load forecasting—serving as a key technology for system planning, operation, and dispatch—urgently requires methodological innovation and enhancement. Developing models with both high accuracy and strong robustness is essential to effectively address the complex and dynamic operating conditions of modern power systems [8].

With the rapid development of artificial intelligence (AI) technologies, their integration with statistical analysis methods and big data processing capabilities has demonstrated significant advantages in complex data modeling, particularly for power load series characterized by non-stationarity and nonlinearity [9]. Consequently, AI-based approaches have been widely applied in the field of power load forecasting and have gradually matured. Existing studies generally classify AI-based load forecasting methods into two main categories: single models and hybrid models [10], as summarized in **Table 1**, which compares the models and output types adopted by different researchers. Single models include both traditional machine learning methods and deep learning models. In comparison, deep learning approaches generally outperform conventional methods in terms of forecasting accuracy and overall performance, owing to their strong nonlinear mapping capabilities and hierarchical feature extraction. In practical applications, Zhang et al. [11] employed a Long Short-Term Memory (LSTM) network for power load forecasting, demonstrating that its gating mechanism effectively captures long-term dependencies in time series data. Visualization of hidden layer features further confirms that LSTM can automatically extract dynamic characteristics of load variations, thereby significantly improving model performance. Meanwhile, Wang et al. [12] developed a load forecasting model based on the Transformer architecture. By incorporating a multi-head attention mechanism, the model effectively captures long-range temporal dependencies and enables efficient global feature representation, thereby improving forecasting accuracy and robustness. In addition, Lu et al. [13] applied the Temporal Convolutional Network (TCN) to load forecasting tasks. By incorporating dilated causal convolutions, the model effectively captures multi-scale temporal dependencies, thereby improving forecasting performance and stability. However, these approaches still exhibit certain limitations. For instance, LSTM and its variants have limited responsiveness when dealing with abrupt and highly volatile load fluctuations. Although Transformer-based models achieve superior performance, their high computational complexity results in increased operational costs. In contrast, TCN can flexibly adjust dilation rates, enabling efficient multi-scale feature extraction while maintaining computational efficiency. Therefore, this study adopts TCN as the foundation for the feature extraction module to more effectively capture multi-scale temporal characteristics in power load data.

As power load characteristics become increasingly complex, the capability of single models to capture multidimensional correlations is gradually limited [14]. In contrast, hybrid forecasting models integrate features extracted by multiple methods and diverse load patterns, enabling multi-perspective modeling and generally achieving superior forecasting performance and stronger generalization ability. Alhussein et al. [15] proposed a power load forecasting framework that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), where CNN is used to extract local features from input data and LSTM is employed to learn temporal dependencies. Experimental results demonstrate that the proposed hybrid model significantly improves forecasting accuracy and overall performance. Pu et al. [16] introduced a GRU-TCN interactive learning method that integrates Gated Recurrent Units (GRU) and Temporal Convolutional Networks (TCN) for power system data forecasting. By leveraging the temporal modeling capability of GRU and the multi-scale feature extraction advantage of TCN, the method effectively captures the coupling relationships within the data. The results demonstrate that, even under incomplete information, the model can accurately capture interactions among multiple energy sources and achieve high forecasting accuracy. Han et al. [17] proposed a forecasting model based on a temporal Transformer architecture, which integrates probabilistic density estimation with the Transformer framework to construct a Gaussian distribution at each time step during forecasting. The results indicate that this approach effectively characterizes data uncertainty and achieves a high forecasting accuracy of up to 0.994, thereby significantly enhancing model performance. Liu et al. [18] developed a high-precision load forecasting model by integrating an improved Whale Optimization Algorithm (WOA) with LSTM. In addition, Complete Ensemble

Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is introduced to decompose the original data, improving data quality and training efficiency. By optimizing model parameters using an improved Whale Optimization Algorithm (WOA), the method achieves a forecasting accuracy of up to 99.05%. Feng et al. [19] proposed a hybrid model that combines TCN, BiLSTM, and an attention mechanism. By incorporating power load data together with multiple external influencing factors to construct multivariate time series, the model effectively captures both short-term and long-term temporal dependencies as well as inter-variable correlations, thereby significantly improving forecasting accuracy and stability compared with traditional methods. In summary, although hybrid models enhance feature extraction capability through multi-structure integration, most existing studies still focus on point forecasting, which limits their ability to comprehensively characterize power load uncertainty and reduces their practical value in real-world dispatching and risk-based decision-making. In addition, some models rely on empirical settings or conventional optimization methods for parameter tuning, making them prone to local optima and adversely affecting overall forecasting performance and stability.

To address the aforementioned issues, this study proposes an IBKA-optimized MSTCN–BiLSTM–QR model for electric load interval forecasting. At the feature extraction level, an improved MSTCN is employed to enhance multi-scale feature fusion, while a BiLSTM network is integrated to capture bidirectional temporal dependencies. For uncertainty modeling, quantile regression is introduced to transform point forecasting into interval forecasting, enabling a more comprehensive representation of load fluctuation ranges. For parameter optimization, the IBKA algorithm is utilized to enhance the model's global search capability and avoid premature convergence to local optima. Through these multi-dimensional improvements and empirical analysis based on residential load data from four regions in the United States, the proposed model demonstrates superior performance in forecasting accuracy, stability, and uncertainty representation, making it well-suited for load forecasting tasks in complex power system environments.

Table 1. Comparison of Research Models and Output Types Used by Different Scholars in Electricity Load Forecasting Tasks

Research scholar	Model used	Single-model /Multi-model	Forecasting type
Zhang et al. [11]	LSTM	Single-model	Point forecasting
Wang et al. [12]	Transformer	Single-model	Point forecasting
Lu et al. [13]	TCN	Single-model	Point forecasting
Alhussein et al. [15]	CNN-LSTM	Multi-model	Point forecasting
Pu et al. [16]	GRU-TCN	Multi-model	Point forecasting
Han et al. [17]	Probability Density Statistics - Transformer	Multi-model	Point forecasting
Liu et al. [18]	CEEMDAN-IWOA-LSTM	Multi-model	Point forecasting
Feng et al [19]	TCN-BiLSTM-Attention	Multi-model	Point forecasting

2. Theoretical Foundation

2.1. Improved Black-winged Kite Algorithm (IBKA)

2.1.1. Basic Black-winged Kite Algorithm (BKA)

The Basic Black-winged Kite Algorithm (BKA) is an intelligent optimization method inspired by the predatory behavior of black-winged kites, primarily designed for solving continuous optimization problems [20, 21]. In nature, black-winged kites exhibit both random exploration and targeted hunting during foraging, and this behavioral characteristic provides important inspiration for algorithm design. BKA mimics the dynamic balance between "global exploration" and "local exploitation" observed during predation, enabling efficient search within the solution space and gradual convergence toward the optimal solution. Similar to most swarm intelligence optimization algorithms, BKA initializes individuals with uniformly distributed positions to ensure diversity and adequate coverage of the search space. The initial positions of individuals in the population are typically generated randomly according to Equation (1).

$$Q_i = BK_{lb} + rand(BK_{ub} - BK_{lb}) \tag{1}$$

where $i \in \{1, 2, \ldots, N\}$; BK_{lb} and BK_u represent the lower and upper bounds, respectively, for the

j-th-dimensional black-winged kite; and *rand* is a random number in the interval [0, 1].

During the predation process, black-winged kites rapidly adjust their flight trajectories according to the position of their prey and initiate an attack, demonstrating strong target-oriented behavior and local exploitation capability. Based on this characteristic, the attack process can be mathematically modeled, as expressed in Equation (2).

$$y_{t+1}^{i,j} = \begin{cases} y_t^{i,j} + n \bullet (2r-1) \bullet y_t^{i,j} & g \geq r \\ y_t^{i,j} + n \bullet [1 + \sin(r) \bullet y_t^{i,j}] & g < r \end{cases} \tag{2}$$

$$n = 0.05 \bullet e^{-2 \bullet (t/T)} \tag{3}$$

where $y_t^{i,j}$ and $y_{t+1}^{i,j}$ represent the positions of the *i*-th black-winged kite in the *j*-th dimension and at the (*t*+1)-th iteration step, respectively; *r* is a random number between 0 and 1; g is typically the constant 0.9; *T* is the total number of iterations; and *t* denotes the number of iterations completed so far.

When food resources are scarce or environmental conditions change, the black-winged kite searches for new potential prey areas through migratory behavior, which reflects strong global exploration capability and environmental adaptability. Based on this behavioral characteristic, the migration process can be mathematically modeled, as expressed in Equation (4).

$$y_{t+1}^{i,j} = \begin{cases} y_t^{i,j} + C(0,1) \bullet (L_t^j - h \cdot y_t^{i,j}) & F_i \geq F_{ri} \\ y_t^{i,j} + C(0,1) \bullet (y_t^{i,j} - L_t^j) & F_i < F_{ri} \end{cases} \tag{4}$$

$$h = 2 \bullet \sin(r + \pi / 2) \tag{5}$$

2.1.2. Multi-strategy Improved Black-Winged Kite Algorithm (IBKA)

2.1.2.1. Tent Chaotic Mapping

In the basic Black-Winged Kite Algorithm (BKA), the initial positions of population individuals are typically generated randomly, which may lead to an uneven distribution in the solution space, thereby affecting the convergence speed and optimization accuracy to a certain extent. To address this issue, this study introduces a chaotic mapping mechanism into BKA to enhance the uniformity of population distribution and improve overall diversity. Among various chaotic mapping methods, the Tent chaotic mapping exhibits notable advantages due to its strong ergodicity, uniform distribution characteristics, and fast convergence performance [22, 23]. Therefore, the Tent mapping is employed to optimize the initialization of the population, and its mathematical expression is given in Equation (6).

$$x_{n+1} = \begin{cases} 2x_n & 0 \leq x_n \leq 0.5 \\ 2(1-x_n) & 0.5 \leq x_n \leq 1 \end{cases} \tag{6}$$

where x_n represents the chaotic value at the current iteration; x_{n+1} represents the chaotic value at the next iteration.

On the basis of introducing the Tent chaotic map for population optimization initialization, it is further integrated with the individual position generation mechanism in Equation (1) to construct an improved population initialization strategy. This approach ensures adequate coverage of the search space while enhancing the uniformity and randomness of individual distribution, thereby strengthening the global search capability of the algorithm. Consequently, the updated population initialization expression is formulated as shown in Equation (7).

$$Q_i' = BK_{lb} + \text{Tent}(BK_{ub} - BK_{lb}) \tag{7}$$

where “Tent” is a value in the interval [0,1] generated by applying the “Tent” mapping.

2.1.2.2. Based on the dynamic lens imaging-based opposition-based learning strategy.

During the predation phase of the BKA, individuals in the population typically converge toward the region of the current best-performing solution. Although this single-direction exploitation mechanism accelerates convergence, it also tends to cause a rapid loss of population diversity, thereby increasing the risk of premature convergence to local optima. To address this issue, an improved opposition-based learning strategy incorporating a dynamic convex lens imaging mechanism is introduced in Ref. 24 [24]. This strategy integrates the concept of opposition-based learning, enabling individuals to not only move toward the current best solution but also obtain corresponding opposite positions in the search

space. In other words, the search process is guided not only by an attraction toward the optimal region but also by a complementary "oppositional guidance" mechanism, which effectively expands the exploration scope. Furthermore, the dynamic convex lens imaging mechanism serves as the core of this approach. By simulating the imaging process of a convex lens, the mapping relationship is adaptively adjusted according to the iteration process, allowing the strength and direction of opposition learning to vary dynamically over time. Under this mechanism, the movement of individuals is no longer restricted to a single direction; instead, it is jointly influenced by both the guidance of the current best solution and multi-directional exploratory information. This enables a more comprehensive and balanced exploration of the solution space. Overall, the proposed method effectively alleviates premature convergence and enhances the ability to escape local optima. The basic principle is shown in **Figure 1.**

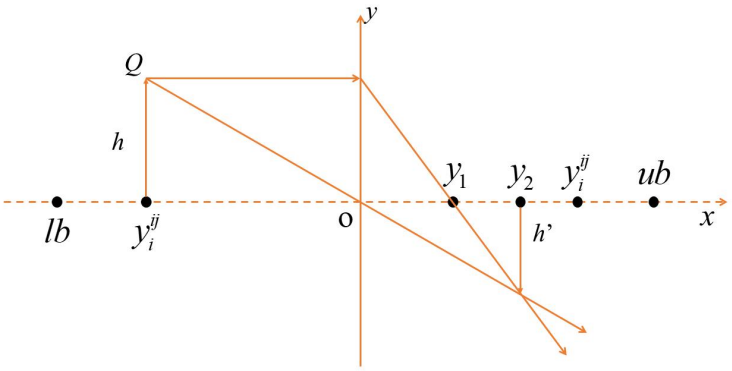


Figure 1. Principles of dynamic convex lens imaging

In the dynamic convex lens imaging learning strategy shown in **Figure 1**, *O* is the midpoint of the interval [*lb*, *ub*], *h* is the height of the current light source *Q*, and *h*' is the height of the image *Q*' of light source *Q* [25]. Based on the aforementioned dynamic convex lens imaging learning principle, a new expression for $y_{t+1}^{i,j}$ is obtained, as shown in Equation (8).

$$y_{t+1}^{i,j} = (ub + lb)/2 + (ub + lb)/2k - y_t^{i,j}/k \tag{8}$$

where k=2r.

2.1.2.3. Based on the Fraunhofer diffraction-based correction strategy.

When the light source and the observation screen are both sufficiently far from the diffraction aperture (e.g., a circular aperture), the incident and diffracted waves can be approximated as parallel rays. Under this condition, the resulting diffraction phenomenon is referred to as Fraunhofer diffraction [26]. In this diffraction regime, the diffraction pattern of a circular aperture exhibits a characteristic concentric ring structure. The position of the first dark ring outside the central bright spot (Airy disk) is determined by the first zero of the first-order Bessel function. The corresponding spatial location can be approximately expressed as shown in Equation (9).

$$x = \frac{\pi a[\sin(\theta)]}{\lambda} \approx 3.3817 \tag{9}$$

Under the above conditions, the position of the dark rings in Fraunhofer diffraction from a circular aperture can be characterized by the diffraction angle. By combining geometric relationships, the radius R of the diffraction pattern (e.g., the Airy disk) on the observation screen can be determined from the relationship between the diffraction angle and the propagation distance. Accordingly, its expression can be derived as follows.

$$R = L \bullet \tan\theta \tag{10}$$

The variable L indicates the distance extending from the circular aperture to the screen.

Since θ is typically very small, this study derives Equation (11) [25] using a small angle $\tan\theta \approx \sin\theta$.

$$R = L\sin\theta = L\frac{3.8317\lambda}{a\pi} \tag{11}$$

To address the issues of large random step-size fluctuations and low convergence efficiency in the migration phase of BKA—particularly the tendency to maintain an excessively wide search range even when approaching the global optimum, which hinders timely convergence—a Fraunhofer diffraction-based correction strategy is introduced for improvement. This strategy is inspired by the characteristic of energy distribution in diffraction phenomena, where the intensity gradually concentrates with increasing propagation distance. Based on this principle, the individual migration step size is adaptively regulated, enabling the search process to transition from coarse-grained global exploration to fine-grained local exploitation. In this way, excessive random perturbations in the later stages are effectively suppressed, while both convergence accuracy and stability near the optimal solution are significantly enhanced. The corresponding mathematical formulation is given in Equation (12).

$$y_{t+1}^{i,j} = \begin{cases} y_t^{i,j} + C(0,1)\bullet\left|\frac{ub-lb}{w}\right|\frac{(y_t^{i,j} - L_t^j)e}{a\pi} & F_i < F_{ri} \\ y_t^{i,j} + C(0,1)\bullet\left|\frac{ub-lb}{w}\right|\frac{(L_t^j - m\bullet y_t^{i,j})e}{a\pi} & F_i \geq F_{ri} \end{cases} \tag{12}$$

where e, a, and w are constants used to adjust the size of the leader's individual influence range.

2.1.2.4. Workflow of the Multi-Strategy Enhanced Black-winged Kite Algorithm

In summary, the overall workflow of the Black-winged Kite Algorithm integrated with multiple improvement strategies is described as follows.

Step 1: Initialize the algorithm parameters, including the population size N, the maximum number of iterations Max, and the chaotic coefficient r. The initial population positions are generated using Equation (7), and their fitness values are calculated. The individual with the best fitness is selected as the current leader.

Step 2: Update the positions of the population members. During the predation phase, an opposition-based learning approach combined with dynamic convex lens imaging is applied to adjust individual positions, as described by Equation (8). In the migration phase, a Fraunhofer diffraction-inspired correction mechanism is used to refine the locations of individuals according to Equation (12).

Step 3: Evaluate the fitness of the newly updated population and identify the best-performing individual, which is designated as the leader for the current iteration.

Step 4: Check whether the stopping criteria are met, such as achieving the maximum iteration count or reaching the desired solution accuracy. If the conditions are satisfied, the algorithm concludes; otherwise, the process returns to Step 2, and the iterative optimization continues until termination conditions are fulfilled.

2.2. Feature extraction network based on MSTCN (Multi-Scale Temporal Convolutional Network).

Traditional TCNs typically expand the receptive field by stacking causal convolutional layers with different dilation rates, thereby effectively modeling long- and short-term dependencies in time series data [27, 28]. However, for complex multi-scale features in multivariate component matrices, simply increasing network depth to enhance feature extraction capability often leads to higher computational cost and reduced training efficiency. To address this limitation, this study proposes an improved design based on the conventional TCN architecture. By optimizing the network structure, the model can more efficiently capture multi-scale information. While maintaining strong feature extraction capability, the proposed approach reduces model complexity, thereby enabling more effective representation and efficient learning of multivariate time-series features.

2.2.1. Adjustment of the TCN network architecture

The original single serial structure of the conventional TCN is optimized into a hybrid serial–parallel architecture to fully exploit the feature representations extracted by causal convolutional layers with different dilation rates. On the one hand, in the serial pathway, the output of the previous TCN layer is fed into the next layer as input, ensuring progressive propagation and accumulation of temporal information across the network, thereby maintaining the continuity and consistency of feature

representations. On the other hand, in the parallel pathway, the outputs of each TCN layer are simultaneously introduced into a multi-head attention gating mechanism, where the contribution of features from different layers is adaptively weighted according to their importance, enabling effective fusion and complementarity of multi-scale features. By introducing this serial–parallel architecture, the model no longer relies solely on the output of the final layer for decision-making. Instead, it integrates feature information from all layers, avoiding the attenuation or loss of fine-grained shallow features during deep propagation. As a result, the representational capacity and information utilization efficiency for multi-scale characteristics of power load data are significantly improved. The network structures before and after the modification are shown in **Figure 2.**

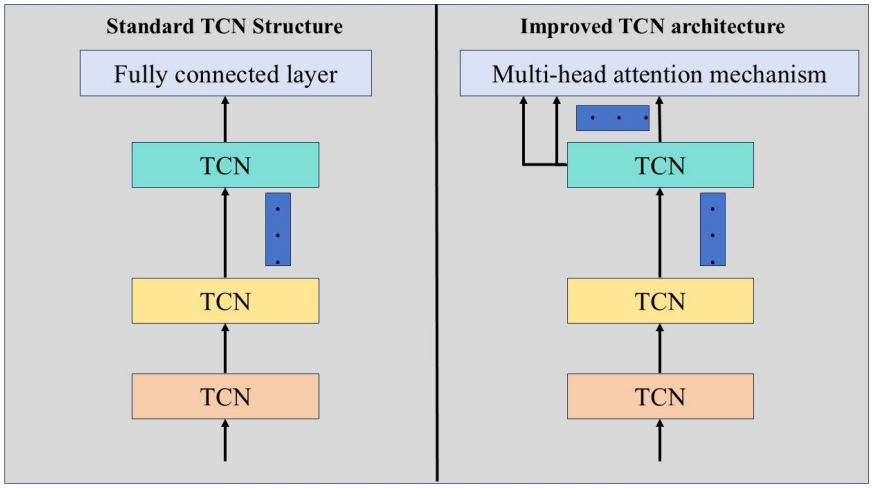


Figure 2. Comparison of TCN Architectures

2.2.2. Introduction of the multi-head attention mechanism

In the improved TCN architecture, a multi-head attention mechanism is introduced to enhance the model's ability to represent critical features. By employing multiple attention heads in parallel, the input features are modeled from multiple perspectives, enabling dynamic reweighting of feature importance. Each attention head independently learns informative patterns within different feature subspaces and adaptively adjusts attention weights according to inter-feature correlations. This mechanism enables the model to selectively emphasize features that contribute more significantly to the forecasting task [29], while suppressing redundant or irrelevant information. As a result, the capability of capturing key features and the overall modeling performance are significantly improved. The detailed implementation process is shown in **Figure 3.**

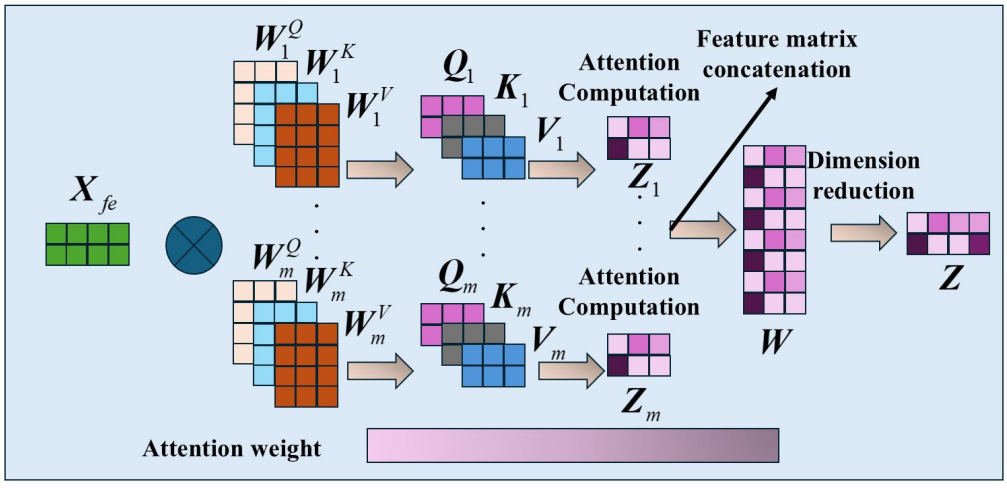


Figure 3. Schematic diagram of the multi-head attention mechanism

Suppose the MSTCN feature extraction network consists of three TCN layers with different dilation

rates. The input matrices for each layer are denoted by $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\} \in \boldsymbol{R}^{C\times D}$, and the feature matrix $\boldsymbol{X}_{fe}$ is generated by concatenating them, as shown in Equation (13).

$$\boldsymbol{X}_{fe} = Contact\left(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\right) \tag{13}$$

Given that the linear transformation matrix for the *m*-th element is $\left\{\boldsymbol{W}_m^Q, \boldsymbol{W}_m^K, \boldsymbol{W}_m^V\right\} \in \boldsymbol{R}^{D\times D}$, map $\boldsymbol{X}_{fe}$ to the query, key, and value matrices $\boldsymbol{Q}_m$, $\boldsymbol{K}_m$, and $\boldsymbol{V}_m$, respectively, using the mapping formula shown in Equation (14).

$$\begin{cases} \boldsymbol{Q}_m = \boldsymbol{X}_{fe}\boldsymbol{W}_m^Q \\ \boldsymbol{K}_m = \boldsymbol{X}_{fe}\boldsymbol{W}_m^K \\ \boldsymbol{V}_m = \boldsymbol{X}_{fe}\boldsymbol{W}_m^V \end{cases} \tag{14}$$

Perform a dot product between $\boldsymbol{Q}_m$ and $\boldsymbol{K}_m$ to obtain the correlation between arbitrary points; simultaneously, scale the result of the dot product along the specified dimension $\sqrt{d_k}$ to prevent the result from becoming too large and causing the gradient to vanish; the scaled result is converted into a probability score via the *softmax* function and multiplied by $\boldsymbol{V}_m$, yielding the output $\boldsymbol{Z}_m$ of the corresponding attention head, as shown in Equation (15).

$$\boldsymbol{Z}_m = Attention\left(\boldsymbol{Q}_m, \boldsymbol{K}_m, \boldsymbol{V}_m\right) = softmax\left(\frac{\boldsymbol{Q}_m \boldsymbol{K}_m^{\ T}}{\sqrt{d_k}}\right)\boldsymbol{V}_m \tag{15}$$

The outputs from each attention head are concatenated to obtain the combined attention feature matrix $\boldsymbol{W}$. This matrix is then subjected to a linear transformation and normalization to obtain the final output feature representation $\boldsymbol{Z}$ [30], whose mathematical form is shown in Equation (16).

$$\boldsymbol{Z} = Contact\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, ..., \boldsymbol{Z}_m\right)\boldsymbol{W}^0 = \boldsymbol{W}\boldsymbol{W}^0 \tag{16}$$

where $\boldsymbol{W}^0$ is the linear transformation matrix.

2.2.3. Replacement of the activation function

In traditional TCN network models, the *ReLU* function is typically chosen as the activation function to enhance the model's ability to represent nonlinearities. This function offers advantages such as simple computation and fast convergence; its specific form is shown in Equation (17).

$$ReLU = \max\left(0, x\right) \tag{17}$$

where x represents the input features to the *ReLU* function. This function sets negative features to 0, resulting in a gradient of 0 during backpropagation. Since the parameters cannot be updated, the neuron dies, thereby reducing the model's expressive power. In this study, we replace *ReLU* with *GeLU*, as shown in Equation (18).

$$GeLU\left(x\right) = x \bullet \phi\left(x\right) \tag{18}$$

where $\phi\left(x\right)$ represents the cumulative distribution function of x, which is expressed as.

$$\phi\left(x\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{\rho^2}{x}} d\rho \tag{19}$$

Compared with traditional activation functions, the *GeLU* function retains a non-zero gradient in the negative region [31], thereby alleviating the "dying neuron" problem to some extent.

2.3. BiLSTM forecasting model

In time series forecasting tasks, LSTM introduces a gating mechanism that effectively regulates information flow across different time steps, thereby alleviating the problems of vanishing and exploding gradients commonly encountered in traditional recurrent neural networks when modeling

long sequences [32]. Based on this concept, the Bidirectional Long Short-Term Memory (BiLSTM) network, as an advanced extension of the standard LSTM architecture, is designed to capture temporal dependencies in both forward and backward directions. Specifically, BiLSTM constructs two separate LSTM branches: one processes the input sequence in the forward direction to learn dependencies from past to future, while the other processes the sequence in the backward direction to capture dependencies from future to past. By simultaneously modeling both directions, BiLSTM can effectively recognize patterns that depend on both preceding and succeeding time steps, which is particularly useful for sequential data where context from both ends influences prediction. The outputs generated by these two branches are subsequently integrated, often through concatenation or summation, to form a unified and more comprehensive feature representation that enhances the model's ability to understand complex temporal dynamics and improve prediction accuracy. This bidirectional modeling strategy enables BiLSTM to demonstrate stronger capability in handling complex time series forecasting tasks with inherent temporal correlations [33, 34]. The network structure is illustrated in **Figure 4.**

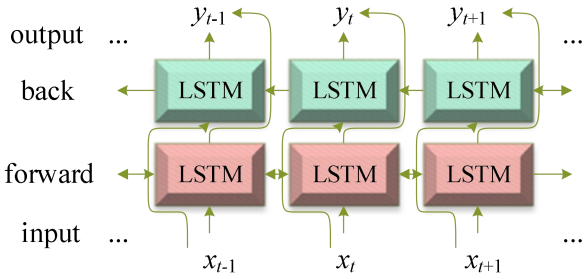


Figure 4. BiLSTM architecture diagram

2.4. Quantile Regression (QR)

The output layer of the model adopts a Quantile Regression (QR) layer, which maps the high-level features extracted by the MSTCN–BiLSTM network into forecasts under different conditional quantiles. As mentioned in the Introduction, electric load is influenced by multiple uncertain factors and exhibits significant randomness and volatility. Traditional forecasting methods are generally based on the assumption of a normal distribution and typically provide only a single point forecast, which is insufficient to fully capture the uncertainty and variability of the forecasting results. Quantile Regression, first proposed by Koenker et al. [35] in 1978, does not require any prior assumption about the data distribution and is capable of characterizing the relationship between explanatory variables and response variables at different quantile levels. Compared with conventional regression methods, QR can model the conditional distribution from multiple quantile perspectives, thereby providing a more comprehensive and robust description of the uncertainty characteristics of the target variable, which is difficult to achieve with classical regression models. Within deep learning frameworks, the integration of quantile regression is straightforward. Specifically, the fully connected layer at the end of the network can be replaced with a quantile regression output layer, and the traditional loss function can be substituted with a quantile loss function (e.g., Pinball Loss). This enables the transition from point forecasting to interval or probabilistic forecasting, thereby enhancing the model's adaptability and representational capability in complex electric load forecasting scenarios.

Let the independent variable be $\boldsymbol{X}=\left[x_1,x_2,...,x_n\right]$ and the dependent variable be $\boldsymbol{Y}=\left[y_1,y_2,...,y_n\right]$. The linear quantile regression model is expressed as follows.

$$Q_{y_i}\left(\tau\middle|\boldsymbol{x_i}\right)=\boldsymbol{\beta}\left(\tau\right)\boldsymbol{x_i},i=1,2,...,n \tag{20}$$

where $Q_{y_i}\left(\tau\middle|\boldsymbol{x_i}\right)$ is the τ -th conditional quantile of the dependent variable y_i, lying between (0, 1); $\boldsymbol{x_i}$ is an (m+1)-dimensional vector; and $\boldsymbol{\beta}\left(\tau\right)$ is the vector of regression coefficients, $\boldsymbol{\beta}\left(\tau\right)=\left[\beta_0\left(\tau\right),\ \beta_1\left(\tau\right),...,\ \beta_m\left(\tau\right)\right]$. When the model's training data is known, the problem of finding the vector of regression coefficients $\boldsymbol{\beta}\left(\tau\right)$ at different quantile points can be reformulated as minimizing the loss function L [36], as shown in Equation (21).

$$\hat{\boldsymbol{\beta}}\left(\tau\right)=\arg\min_{\beta} L=\arg\min_{\beta}\sum_i^n\gamma_\tau\left[y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\left(\tau\right)\right] \tag{21}$$

where γ_τ is the absolute value of the slope, which is calculated as shown in Equation (22).

$$\gamma_\tau(s)=\begin{cases}\tau s, & s \geq 0\\ (\tau-1)s, & s<0\end{cases} \tag{22}$$

where $s = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)$.

In this study, based on the deep features extracted by the MSTCN–BiLSTM network, a quantile regression layer is employed to simultaneously generate electric load forecasts at multiple quantile levels (0.8 and 0.95), thereby constructing forecasting intervals. This approach not only provides conventional point forecasts but also characterizes the uncertainty range of the forecasts. It effectively captures the fluctuation characteristics of electric load under complex operating conditions, thereby significantly improving the model's reliability, robustness, and practical applicability.

2.5. Power load interval forecasting process and steps

The overall process of the power load interval forecasting model based on the IBKA-optimized MSTCN–BiLSTM–QR framework is shown in **Figure 5**. The proposed framework mainly consists of five stages: data preprocessing, feature extraction, sequence modeling, parameter optimization, and interval forecasting output. The detailed procedures are as follows.

First, in the data preprocessing stage, raw power load data and related influencing factors are cleaned, missing values are handled, and normalization is performed to eliminate the interference of abnormal data on model training, thereby constructing a multivariate time series input.

Second, in the feature extraction stage, the preprocessed multivariate time series is fed into the improved MSTCN network. Through a serial–parallel architecture of dilated causal convolutions, features at different temporal scales are extracted. Meanwhile, a multi-head attention mechanism is introduced to achieve adaptive weighting and fusion of multi-scale features, resulting in richer and more discriminative representations.

Third, in the sequence modeling stage, the high-dimensional features extracted by MSTCN are input into the BiLSTM network. By leveraging its bidirectional structure, the model learns both forward and backward temporal dependencies, further enhancing its ability to represent complex temporal dynamics.

In the model optimization stage, the improved Black-winged Kite Algorithm (IBKA) is introduced to globally optimize key hyperparameters of the model, such as learning rate and the number of hidden neurons. By incorporating chaotic initialization, dynamic opposition-based learning, and diffraction-based correction strategies, IBKA effectively improves the accuracy and efficiency of the parameter search, thereby enhancing overall model performance and stability.

Finally, in the forecasting output stage, a Quantile Regression (QR) layer is employed to map the features extracted by the BiLSTM network and simultaneously generate forecasts at multiple quantile levels (e.g., 0.8 and 0.95), thereby constructing electric load forecasting intervals. This approach not only provides point forecasts but also characterizes the uncertainty range of the forecasts, offering a more reliable basis for power system dispatching and risk-informed decision-making.

In summary, the proposed IBKA–MSTCN–BiLSTM–QR model establishes an end-to-end framework ranging from data-driven feature extraction to interval forecasting output through multi-module collaboration, significantly improving forecasting accuracy, robustness, and uncertainty quantification capability.

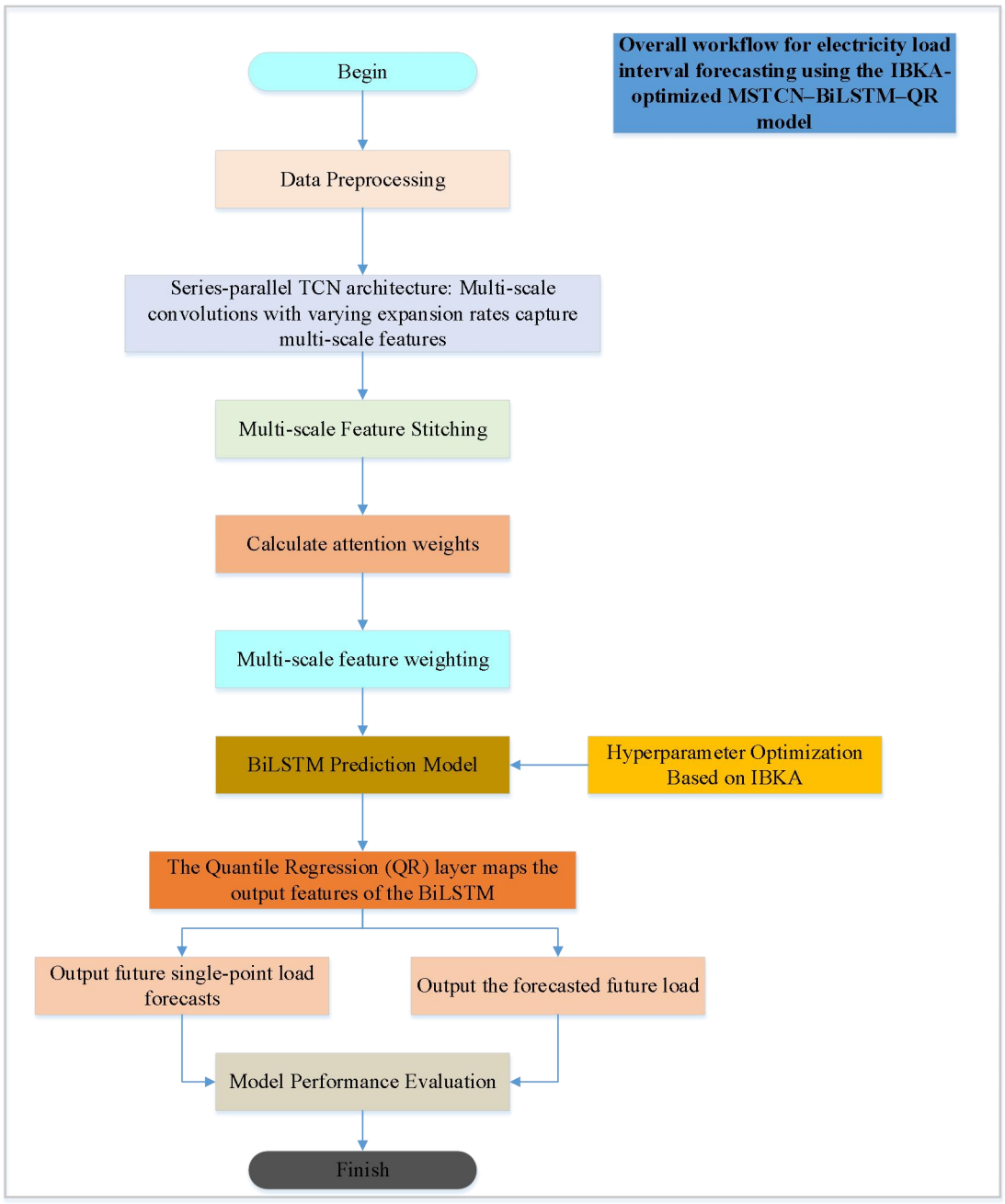


Figure 5. Flowchart of the Electric Load Forecasting Process

3. Simulation experiments and result analysis

3.1. Data acquisition and preprocessing

The dataset used in this study is obtained from a publicly available residential load dataset covering four regions in the United States (http://www.csee.org.cn/). The four regions considered in this study are Napa County, California; Lansing, Michigan; Austin, Texas; and The Dalles, Washington. For the convenience of subsequent modeling, comparative analysis, and figure annotation, these four regions are denoted as U1, U2, U3, and U4, respectively. The datasets for each region are illustrated in **Figure 6**. It can be observed that there are significant differences among regions in terms of load magnitude, peak-to-valley variation, and fluctuation intensity. Therefore, it is necessary to develop region-specific forecasting models and conduct comparative analyses.

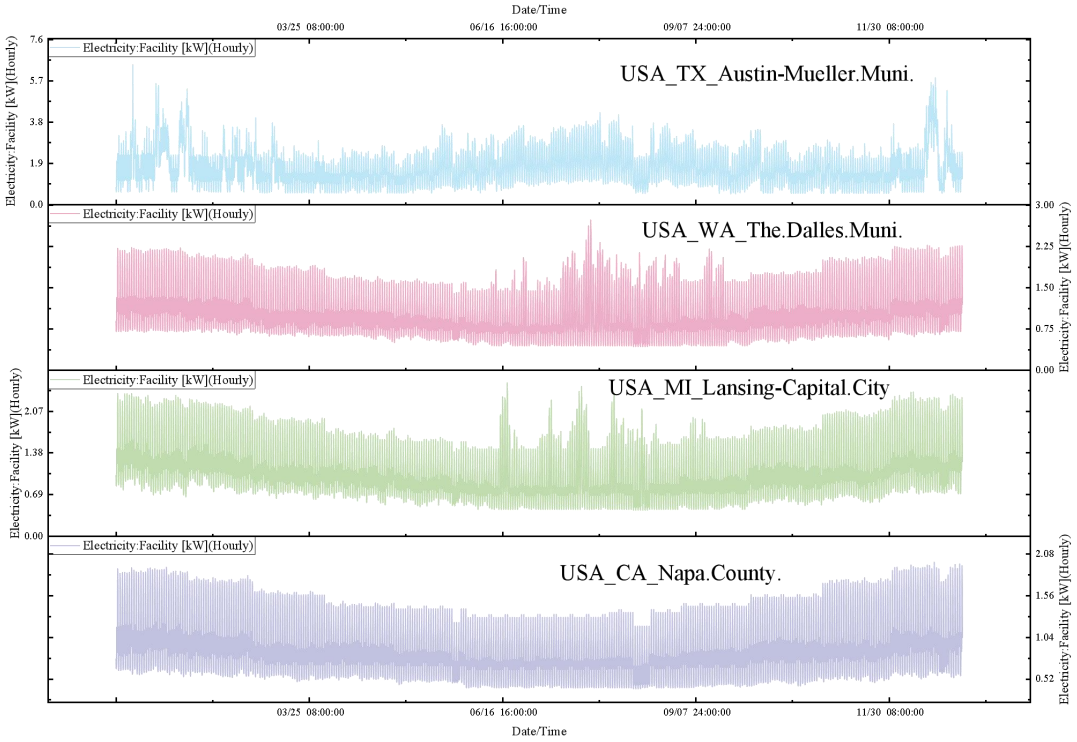


Figure 6. Sources of electricity load data and time series chart

All datasets are sampled at an hourly resolution and include 12-dimensional energy consumption features, such as total building electricity load, gas consumption, heating, cooling, ventilation, lighting, and indoor equipment. In this study, Electricity:Facility [kW](Hourly) is selected as the target for load forecasting, while the remaining numerical energy variables are used as candidate input features to characterize the multi-source influencing factors in residential load variation. Taking the U1 dataset as an example, as shown in **Figure 7**, eight representative energy consumption features in the U1 region exhibit clear seasonal and periodic patterns over the annual scale. Different subsystems not only show coordinated variations but also demonstrate certain heterogeneity. From an overall perspective, gas-related loads (Gas:Facility and Heating:Gas) remain at relatively high levels at the beginning and end of the year, while they decrease significantly during mid-summer, reflecting a typical winter-high and summer-low pattern dominated by heating demand. This trend is strongly correlated with climatic conditions, indicating that heating load is one of the key drivers of overall energy consumption fluctuations. Meanwhile, HVAC fan electricity consumption (HVACFan:Fans) shows a similar pattern, with more frequent fluctuations during winter and transitional seasons and relatively lower levels in summer, suggesting its close relationship with the overall HVAC system operation. Regarding lighting loads, both indoor lighting (General: InteriorLights) and outdoor lighting (General: ExteriorLights) exhibit relatively smooth trends with moderate seasonal variations. Among them, outdoor lighting is more significantly affected by daylight duration, remaining higher in winter and slightly decreasing in summer, while indoor lighting remains relatively stable, reflecting consistent usage patterns. For equipment-related loads, indoor equipment (Appl: InteriorEquipment) and miscellaneous equipment (Misc: InteriorEquipment) remain generally stable but present certain random fluctuations and short-term peaks, indicating the inherent uncertainty of residents' daily electricity usage behavior. Although these loads do not exhibit strong seasonality, they play an important role in short-term load variability. In addition, the water heating system (Water Heater:WaterSystems) maintains a relatively high and stable level throughout the year, with only occasional fluctuations, indicating that it represents a rigid demand load and provides a stable contribution to total energy consumption. Overall, these multi-source energy features present diverse temporal patterns: heating and gas loads exhibit strong seasonality; lighting and equipment loads are relatively stable with short-term fluctuations; and systems such as ventilation and water heating achieve a balance between stability and variability. The coordinated evolution of these multidimensional features provides a rich information basis for characterizing building electricity load and offers important data support for the development of high-accuracy load forecasting models.

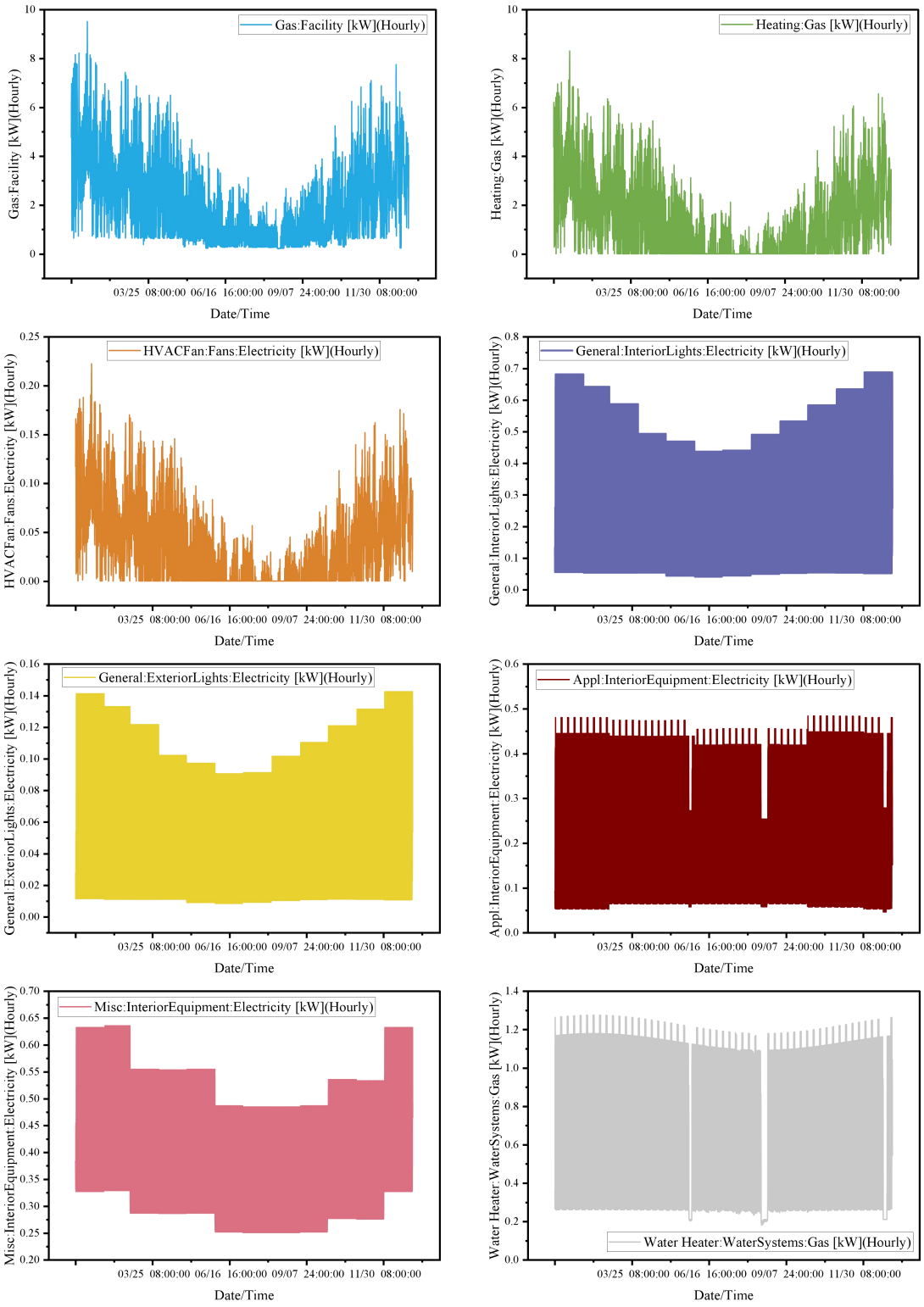


Figure 7. Time-series chart of energy consumption characteristics in the U1 area

During the data preprocessing stage, the original field names are first unified and cleaned, and all numerical variables are extracted as candidate input features. Subsequently, the target column is removed from the input set to avoid information leakage. Missing values are handled using forward filling and backward filling methods. To eliminate the impact of different variable scales on model training, StandardScaler is applied to normalize both the input features and the target load. Considering the strong temporal dependency of electricity load, a sliding time window with a length of 10 hours is further employed to construct supervised learning samples. The dataset is then divided into training and testing sets in an 8:2 ratio according to chronological order. This splitting strategy effectively prevents

future information leakage and ensures that the model training and evaluation process aligns with real-world short-term load forecasting scenarios.

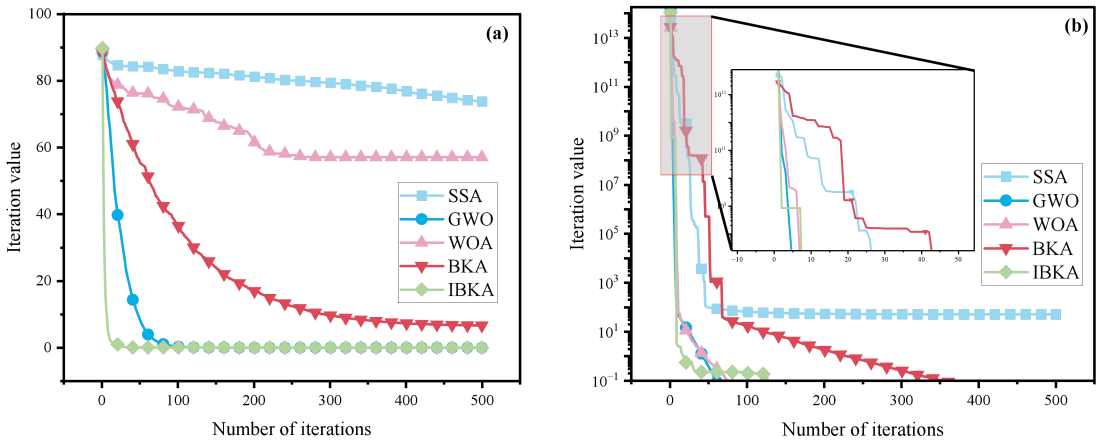
3.2. Experimental Setup Platform

To ensure fairness and reproducibility in model training and comparative experiments, all experiments in this study were conducted on a unified computational platform. The hardware configuration includes an Intel Core i7 multi-core processor (base frequency of approximately 3.0 GHz or higher), 32 GB of RAM, and an NVIDIA GeForce RTX 3060 GPU (12 GB VRAM), which fully meets the computational requirements of multivariate time-series modeling and deep neural network training, thereby improving the efficiency of model training and parameter optimization. In terms of software environment, the operating system is Windows 11 (64-bit). The deep learning framework is implemented in Python 3.8, and the PyTorch library is used for the construction and training of the MSTCN-BiLSTM model. In addition, NumPy and Pandas are employed for data preprocessing and analysis, while Matplotlib is used for result visualization. Furthermore, to ensure the reproducibility of the experimental results, all model training is performed under fixed random seeds.

3.3. Verification of the effectiveness of the IBKA algorithm

To evaluate the effectiveness and superiority of the IBKA algorithm in terms of optimization efficiency and global search capability, six representative benchmark functions widely adopted in Ref. [37] are selected for comparative experiments. These functions consist of both unimodal and multimodal categories. Specifically, the unimodal functions (F1–F3) are used to assess convergence speed and optimization accuracy, while the multimodal functions (F4–F6) are employed to examine the algorithms' ability to escape local optima and their global exploration performance.

For a comprehensive performance assessment, the basic Black-winged Kite Algorithm (BKA) [38], Sparrow Search Algorithm (SSA) [39], Grey Wolf Optimizer (GWO) [40], and Whale Optimization Algorithm (WOA) [41] are selected as benchmark comparison methods. To evaluate the effectiveness of the IBKA algorithm on a variety of optimization problems, a comparative study is performed under consistent experimental settings. To ensure fairness and consistency, all algorithms are configured with identical parameters, including a population size of 20 and a maximum of 500 iterations. Each algorithm is independently run 50 times to reduce the impact of stochastic variability. The performance of the algorithms is quantitatively assessed using three metrics: the best-obtained solution (Best), the average solution across runs (Mean), and the standard deviation (Std). These indicators collectively provide a thorough evaluation of both the accuracy and stability of the optimization results. The comparative results on different benchmark functions are presented in **Figure 8** and **Table 2.**



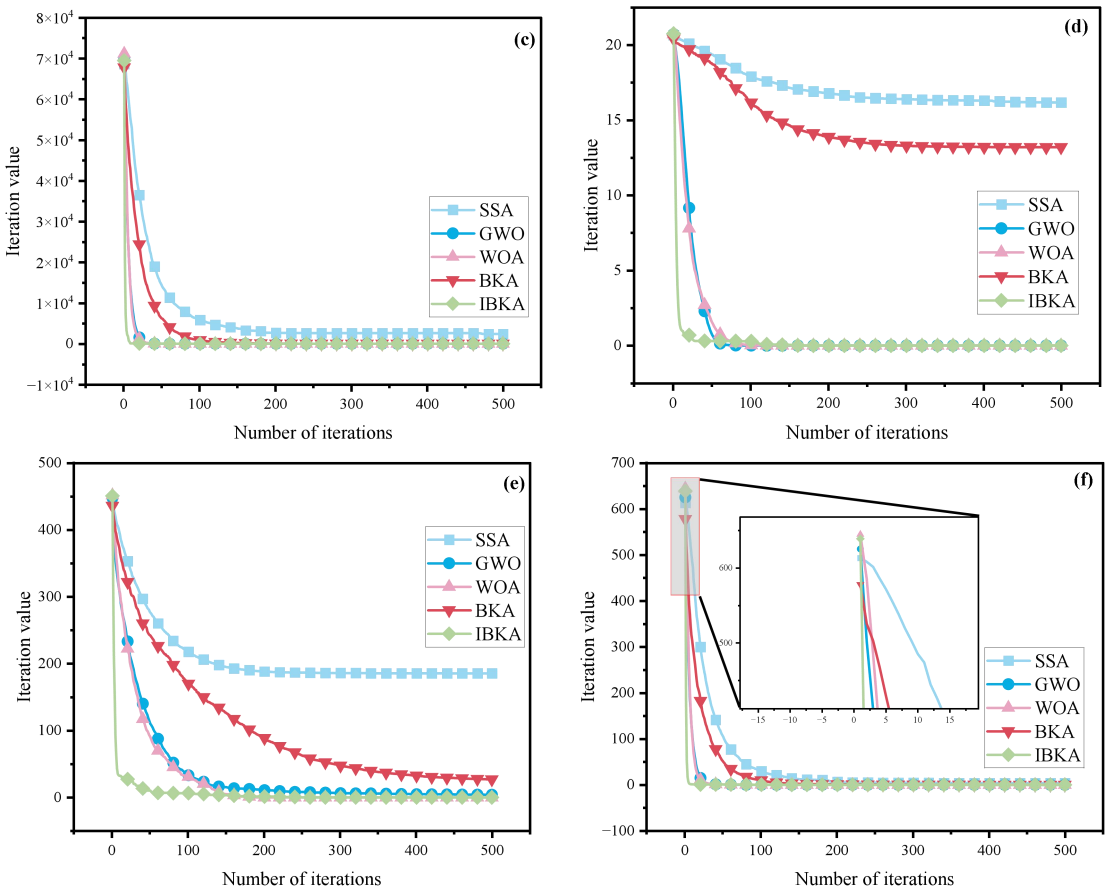


Figure 8. Iteration curves of fitness values for different optimization algorithms (a) F1 (b) F2 (c) F3 (d) F4 (e) F5 (f) F6

Table 2. Comparison of Performance Results for Each Algorithm

Function	Indicator	SSA	GWO	WOA	BKA	IBKA
F1	Best	41.812905	7.95E-07	1.897216482	2.418563902	4.12E-298
	Mean	72.96344182	1.12E-05	56.78421953	6.589104772	5.91E-133
	Std	14.1027736	1.19E-05	26.09133455	2.842317905	4.62E-132
F2	Best	0.000781294	6.45E-15	4.02E-58	0.000347118	6.01E-278
	Mean	49.98211637	3.55E-14	1.18E-47	0.014391562	2.11E-135
	Std	22.01455231	2.58E-14	4.87E-47	0.019874633	1.33E-134
F3	Best	1.72E-05	1.09E-24	2.63E-81	6.85E-05	1.02E-320
	Mean	2387.563442	6.51E-23	1.88E-61	0.042103918	9.76E-255
	Std	5198.274611	1.52E-22	1.27E-60	0.066912344	0
F4	Best	2.391774628	2.61E-13	4.39E-16	0.000451902	4.44E-16
	Mean	15.96311877	1.36E-12	3.92E-15	13.04822984	4.44E-16
	Std	5.742119883	1.29E-12	2.91E-15	9.633118402	0
F5	Best	96.10288451	1.74E-15	0	0.053118447	0
	Mean	184.7765123	4.082113776	0	27.11844693	0
	Std	37.42891562	5.193774621	0	17.60211854	0
F6	Best	0.000291447	0	0	0.000118392	0
	Mean	3.701884552	0.006031228	0.015892441	0.051336772	0
	Std	18.10355762	0.010488112	0.065118903	0.081447216	0

The experimental results in **Table 2** and **Figure 8** demonstrate that IBKA exhibits significant performance advantages across all six benchmark test functions, fully validating its effectiveness and superiority in terms of optimization accuracy and convergence performance. For the unimodal functions (F1–F3), IBKA achieves the best or near-zero results in all three metrics (Best, Mean, and Std). In particular, for F1 and F3, the optimal values obtained by IBKA are very close to the theoretical optimum (approaching zero), and both the mean and standard deviation are markedly better than those

of comparison algorithms such as SSA, WOA, and BKA. This indicates that IBKA has clear advantages in convergence speed and optimization precision. Moreover, the extremely small or nearly zero standard deviation demonstrates strong stability and consistency across multiple independent runs. For the multimodal functions (F4–F6), IBKA also shows strong global search capability. Compared with traditional algorithms that are prone to premature convergence into local optima, IBKA consistently achieves superior or equivalent optimal solutions (e.g., reaching zero-level optimal values in F5 and F6), indicating that the proposed improvement mechanisms effectively enhance its ability to escape local optima. In addition, for the more complex function F4, IBKA outperforms BKA and SSA in both mean value and standard deviation, further confirming its robustness. Overall, by integrating multiple improvement strategies, IBKA significantly enhances global exploration capability and convergence performance. It demonstrates strong adaptability and stability across different types of optimization problems, providing a reliable foundation for subsequent parameter optimization in power load forecasting models.

3.4. Model parameter optimization and selection of evaluation metrics

3.4.1. Model parameter optimization

To improve the generalization capability of the IBKA-MSTCN-BiLSTM-QR model for load data from different regions, the IBKA algorithm is utilized to perform global optimization of critical hyperparameters. The parameters under optimization include the number of layers and channels in the MSTCN component, as well as the number of layers and hidden units in the BiLSTM component. For each of the four regional datasets, the data are temporally split into training and testing sets using an 80:20 ratio, and a portion of the training data is reserved as a validation subset to evaluate the fitness function. The IBKA optimization is conducted with a population size of 20 and a maximum of 100 iterations. The fitness function is formulated based on the quantile loss calculated on the validation set, ensuring that the optimization directly aligns with the interval forecasting objective. Once the optimal hyperparameters are identified, the model is retrained on the full training set using these parameters, with the number of training epochs set to 50 to finalize the model for subsequent evaluation.

Table 3. Model parameter optimization results

Region	TCN Layers	Channel Size	BiLSTM Layers	BiLSTM Neurons	Validation Quantile_Loss
U1	3	125	1	90	0.238377
U2	3	105	1	91	0.218605
U3	4	127	1	79	0.226018
U4	4	104	1	93	0.235834

As shown in **Table 3**, the optimal structural parameters vary across the four regions, indicating that the fluctuation patterns and nonlinear relationships of load series are not identical. The optimal number of MSTCN layers for U1 and U2 is 3, whereas U3 and U4 adopt a 4-layer structure, suggesting that the latter two require deeper multi-scale temporal feature extraction to capture more complex load dynamics. For all regions, the optimal number of BiLSTM layers is 1, implying that after MSTCN extracts local and multi-scale features, a single-layer bidirectional recurrent structure is sufficient to model the main temporal dependencies, and adding more layers does not further improve validation loss. In terms of channel size and hidden units, U1 and U3 have relatively larger optimal channel numbers, 125 and 127 respectively, indicating more complex mappings between input features and load variations in these regions, thus requiring higher-dimensional convolutional representations. In contrast, U2 and U4 have smaller channel sizes of 105 and 104, reflecting more compact model structures. The number of BiLSTM hidden units ranges from 79 to 93, suggesting regional differences in modeling long-term dependencies. Overall, IBKA can adaptively search for suitable model structures based on regional data characteristics, effectively avoiding the limitations of fixed hyperparameter settings on model generalization.

Figure 9 shows the training iteration process under the optimal parameter configurations for the four regions. It can be observed that the loss in all regions decreases progressively with the increase in training epochs and gradually converges, indicating that the model can effectively learn the temporal patterns in load sequences. The training curves of U1, U2, and U4 show relatively smooth downward trends, reflecting more stable temporal patterns in these regions. In contrast, U3 exhibits more noticeable fluctuations during training, which is associated with its stronger peak variations and more complex nonlinear characteristics. Overall, 50 training epochs are sufficient for the model to reach a

stable convergence state, validating the rationality of the selected training epochs and the IBKA-optimized parameters.

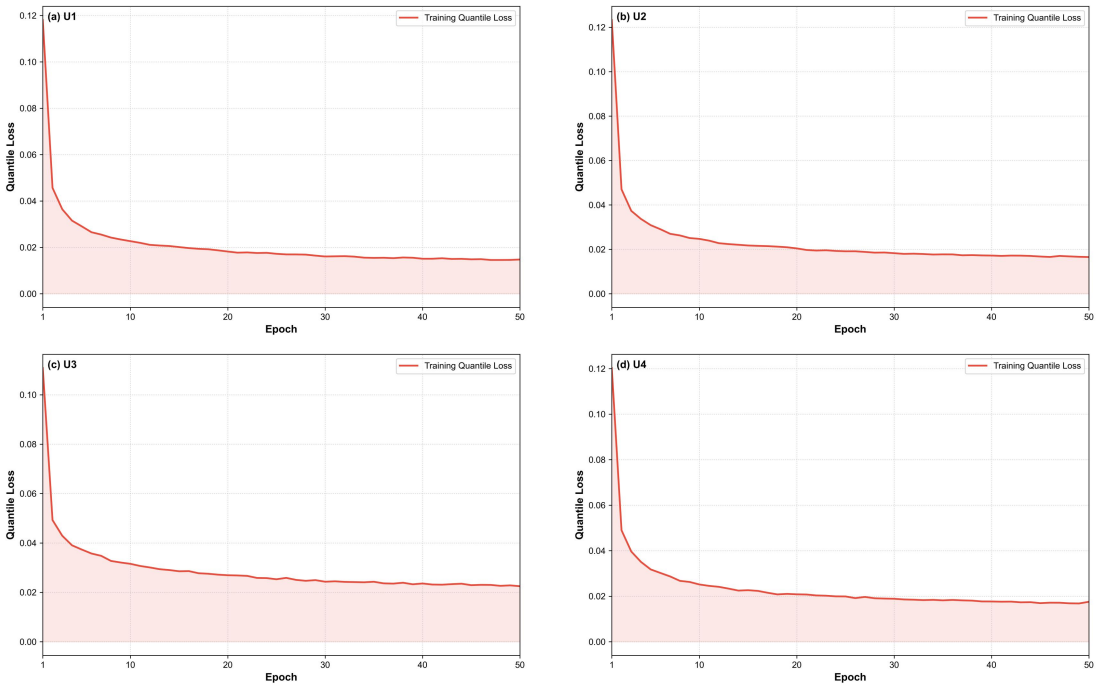


Figure 9. Model training iteration diagram (a) U1 (b) U2 (c) U3 (d) U4

3.4.2. Selection of evaluation metrics

To comprehensively evaluate the performance of the proposed model in the power load forecasting task, multiple evaluation metrics are selected from both point forecasting accuracy and interval forecasting reliability for a holistic analysis, as summarized in **Table 4.** For point forecasting evaluation, the coefficient of determination (R^2), mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean square error (RMSE) are adopted. Among them, R^2 reflects the goodness of fit between the predicted and actual values, with a value closer to 1 indicating better fitting performance. MAPE measures the relative error between predicted and actual values and provides strong interpretability. MAE and RMSE quantify forecasting deviations from the perspectives of absolute and squared errors, respectively, where RMSE is more sensitive to large errors and can effectively reflect model stability and robustness. For interval forecasting evaluation, forecasting interval coverage probability (PICP) and mean forecasting interval width (MPIW) are used as the primary metrics. PICP measures the proportion of true values falling within the predicted interval; a value closer to the predefined confidence level indicates higher reliability of interval forecasting. MPIW describes the average width of the forecasting interval, where a smaller value indicates a more compact and informative interval. Under the premise that PICP satisfies the required confidence level, a smaller MPIW implies higher forecasting precision and greater practical applicability of the model.

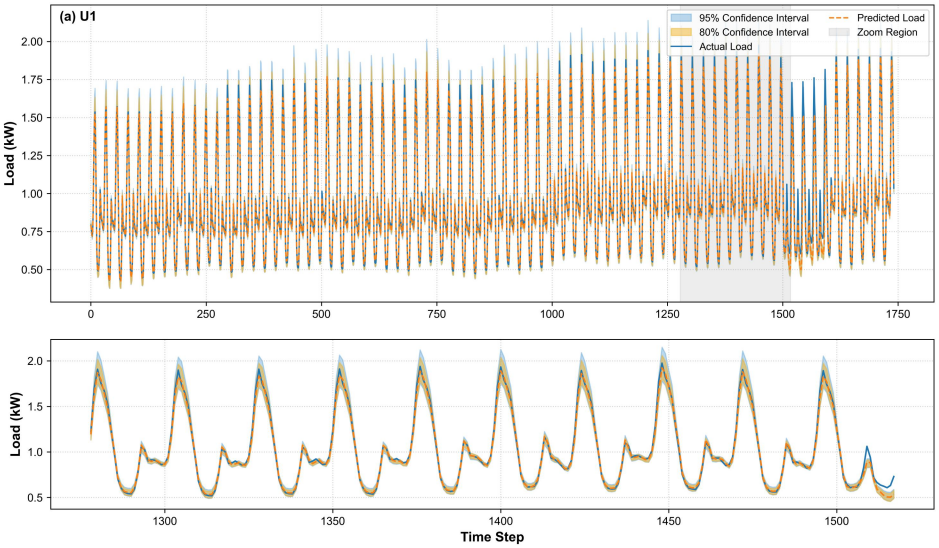
Table 4. Summary of Model Forecasting Evaluation Metrics

Evaluation type	Evaluation metrics	Equation
Point forecasting	R-Square (R^2)[42]	$R^2=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}$ (23)
	Mean Absolute Percentage Error (MAPE/%)[43]	$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left\|(\hat{y}_i-y_i)/y_i\right\|$ (24)
	Mean Absolute Error (MAE)[44]	$MAE=\frac{1}{n}\sum_{i=1}^{n}\left\|y_i-\hat{y}_i\right\|$ (25)
	Root Mean Squard Error (RMSE/ kW)[45]	$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)}$ (26)
Interval forecasting	Prediction Interval Coverage Probability (PICP)[46]	$PICP=\frac{1}{n}\sum_{i=1}^{n}I\left(y_i\in\left[\hat{y}_i^l,\hat{y}_i^u\right]\right)$ (27)
	Mean Prediction Interval Coverage Distance (MPICD/kW)[47]	$MPICD=\frac{1}{n}\sum_{i=1}^{n}\left(\left\|\frac{\hat{y}_i^u-\hat{y}_i^l}{2}-y_i\right\|\right)$ (28)

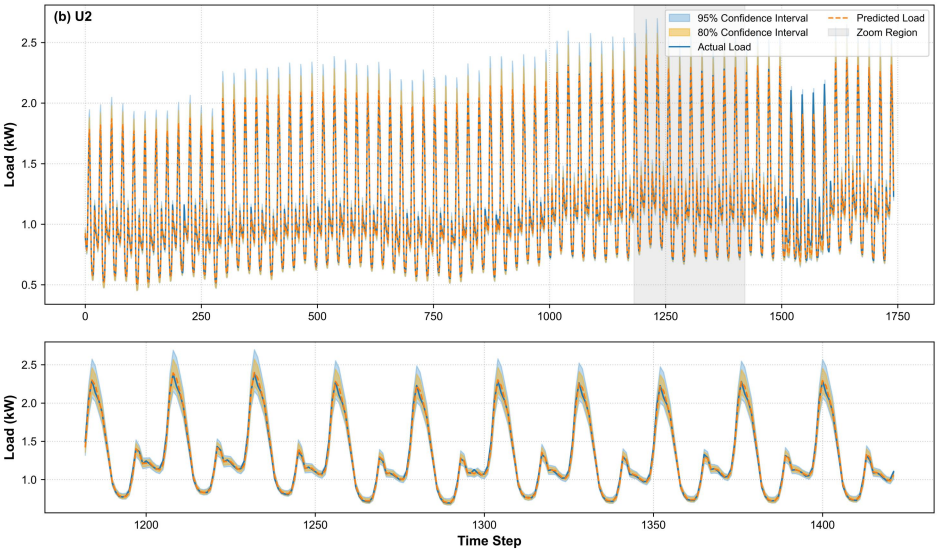
Note: where y_i the true value corresponding to the *i*-th sample; $\hat{y}_i$ the predicted value corresponding to the *i*-th sample; $\bar{y}$ is the mean of the true values; *n* is the total number of samples; y_i is the *i*-th observation; $\hat{y}_i^l$ is the lower limit of the forecasting interval; $\hat{y}_i^u$ is the upper limit of the forecasting interval; $I(\bullet)$ is the indicator function, which takes the value of 1 when the condition inside the parentheses is established, otherwise it is 0; y_{min} is the minimum value of the observed value, y_{max} is the maximum value; α is the set confidence level, for example, 90% confidence level corresponds to α=0.1.

3.5. Output of power load interval forecasting results

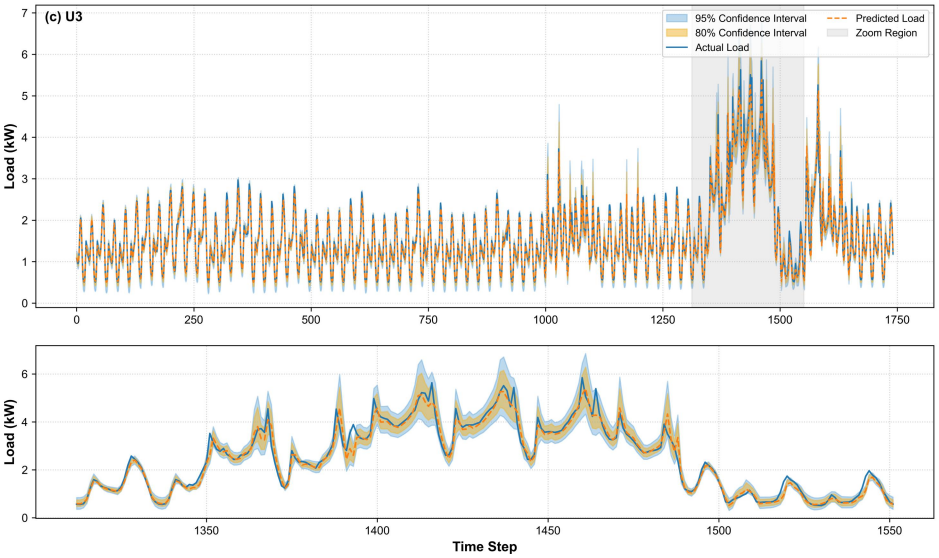
After completing model parameter optimization and final training, the quantile regression layer is employed to output predictions at different quantiles, which are further used to construct prediction intervals for electric load at the 80% and 95% confidence levels. Specifically, the 80% prediction interval is formed by the 0.1 and 0.9 quantiles, while the 95% prediction interval is defined by the 0.025 and 0.975 quantiles. The median quantile is adopted as the point forecast. **Figure 10** presents the load forecasting interval results on the test sets of regions U1–U4. The upper part of each subplot shows the complete test sequence, while the lower part provides a zoomed-in view, enabling simultaneous observation of the overall forecasting trend and detailed interval boundaries. As illustrated in **Figure 10**, the forecasting curves for all four regions closely track the actual load variations, and the prediction intervals can adaptively expand or contract according to the magnitude of load fluctuations. In regions U1, U2, and U4, where load variations are relatively stable, the prediction intervals remain narrow, and most actual values fall within the intervals, indicating that the model achieves high coverage while maintaining compact intervals. In contrast, region U3 exhibits more pronounced peak fluctuations, and the model correspondingly produces wider prediction intervals, demonstrating its capability to effectively capture uncertainty. The zoomed-in results further show that during periods of rapid load increase or decrease, the model can still follow the trend accurately, with interval boundaries covering the main fluctuation range. At the 80% confidence level, the PICP values for U1–U4 are 85.48%, 86.85%, 82.55%, and 80.31%, respectively, all reaching or approaching the nominal confidence level. This indicates that the model maintains reliable coverage even with relatively narrow intervals. The corresponding MPICD values are 0.0240 kW, 0.0262 kW, 0.0768 kW, and 0.0307 kW, respectively, with U3 showing the largest MPICD. This is mainly due to the more significant peak-to-valley variations in the Austin region, which require moderately wider intervals to capture stronger fluctuations. When the confidence level increases to 95%, the PICP values for U1–U4 rise to 91.68%, 94.66%, 92.42%, and 91.39%, respectively, indicating enhanced coverage performance. Meanwhile, the MPICD values are 0.0242 kW, 0.0290 kW, 0.0774 kW, and 0.0303 kW, remaining at relatively low levels overall. This demonstrates that the model does not simply enlarge the intervals to improve coverage, but rather achieves a good balance between coverage capability and interval compactness. Overall, **Figure 10** shows that the IBKA-MSTCN-BiLSTM-QR model can provide stable and reliable load forecasting intervals under different confidence levels, offering richer uncertainty information for practical power dispatching and risk assessment.


(a) U1
95% Confidence Interval
80% Confidence Interval
Actual Load
Predicted Load
Zoom Region
Load (kW)
Time Step

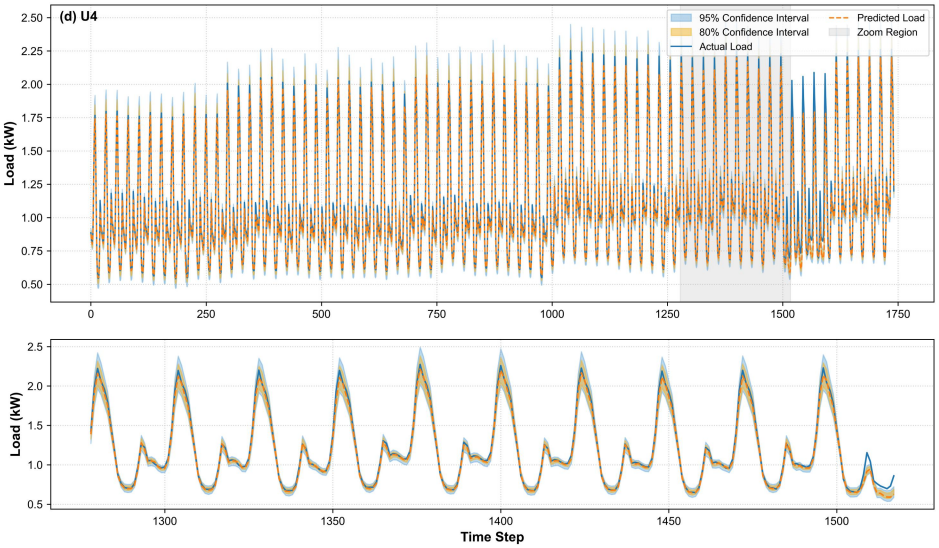
(a) U1


(b) U2
95% Confidence Interval
80% Confidence Interval
Actual Load
Predicted Load
Zoom Region
Load (kW)
Time Step

(b) U2


(c) U3
95% Confidence Interval
80% Confidence Interval
Actual Load
Predicted Load
Zoom Region
Load (kW)
Time Step

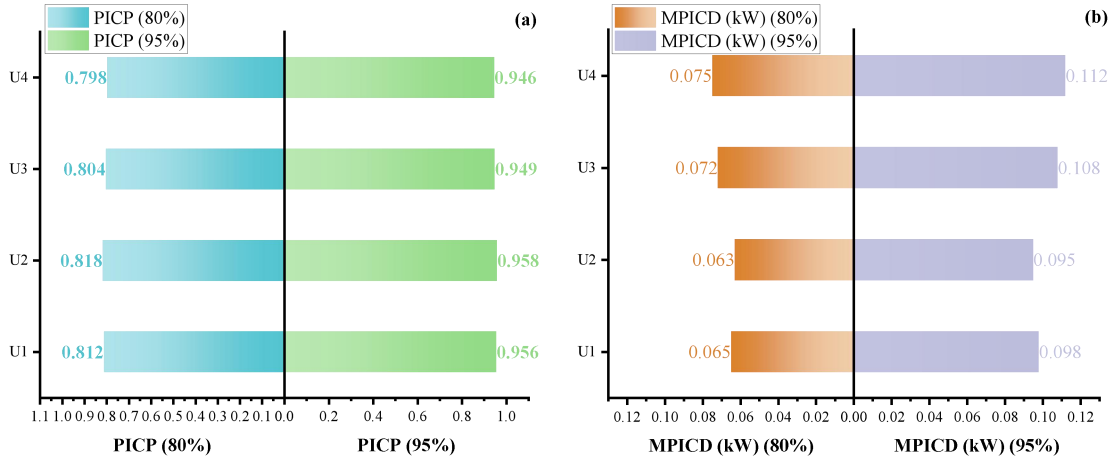
(c) U3



(d) U4

Figure 10. Forecast results for power load zones

Figure 11 presents the comparison results of interval forecasting performance across four datasets under different confidence levels, where (a) shows the forecasting interval coverage probability (PICP) and (b) shows the mean forecasting interval width (MPIW). Overall, the proposed model demonstrates strong interval forecasting capability at both 80% and 95% confidence levels, achieving high coverage while maintaining relatively narrow interval widths. Specifically, under the 80% confidence level, the PICPs of the four datasets are 0.812, 0.818, 0.804, and 0.798, respectively, all of which are close to the nominal confidence level, indicating that the constructed forecasting intervals are highly reliable. Meanwhile, the corresponding MPIWs range from 0.063 to 0.075 kW, reflecting compact intervals and demonstrating that the model achieves a good balance between coverage and sharpness. In contrast, under the 95% confidence level, the PICP further increases to above 0.946, reaching up to 0.958, indicating that the model can effectively capture most of the true values. At the same time, the MPIW increases to 0.095–0.112 kW, reflecting a reasonable expansion of the forecasting intervals with increasing confidence levels. From the perspective of different datasets, U1 and U2 exhibit smaller interval widths and higher coverage rates, indicating more stable data distributions and more accurate forecasting performance. In comparison, U3 and U4 show slightly larger MPIWs, suggesting that the model adaptively widens the forecasting intervals under higher data volatility to maintain satisfactory coverage performance. Overall, the results in **Figure 11** demonstrate that the proposed model achieves a good balance between coverage probability and interval width, showing strong reliability and practical applicability in interval forecasting tasks.



(a) PICP (b) MPICD

Figure 11. Comparison of interval evaluation metrics across four datasets

3.6. Comparison of point forecasting results of different models

To verify the superiority of the proposed IBKA-MSTCN-BiLSTM-QR power load forecasting model in point forecasting, its results are compared with several representative combined forecasting models, including CNN-LSTM [15], CNN-BiLSTM [48], TCN-LSTM [28], Transformer–TCN–GRU [49], TCN-Informer-BiGRU [50], and TCN-QRNN [51]. Based on the U1–U4 U.S. residential electricity load datasets, all models adopt the same data partition strategy, i.e., an 8:2 random split for training and testing sets, to ensure a fair comparison. **Table 5** summarizes the performance of different models on the four datasets (U1–U4), using the coefficient of determination (R^2), mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean square error (RMSE) as evaluation metrics. Overall, the proposed model achieves the best performance across all datasets and evaluation indicators, demonstrating superior fitting capability.

First, in terms of R^2, the proposed model maintains consistently high fitting accuracy on both training and test sets. On the U1 dataset, the R^2 values of the proposed model reach 0.9964 and 0.9938, respectively, significantly outperforming all comparison models. For instance, CNN-LSTM achieves only 0.9875 on the test set, while CNN-BiLSTM slightly improves to 0.9883, both still notably lower than the proposed model. Even stronger baselines such as Transformer–TCN–GRU and TCN-Informer-BiGRU obtain test R^2 values of 0.9912 and 0.9924, respectively, which still show gaps of approximately 0.0014–0.0026. Similar trends are observed across U2–U4 datasets, indicating that the proposed model has stronger capability in capturing complex nonlinear relationships.

Second, the error-based metrics further highlight the advantages of the proposed method. Taking MAPE as an example, on the U2 dataset, the proposed model achieves a test MAPE of 1.7892%, while CNN-LSTM reaches 2.4635%, corresponding to an error increase of over 37%. CNN-BiLSTM and TCN-QRNN also exhibit higher errors of 2.3387% and 2.2763%, respectively. Although Transformer–TCN–GRU performs relatively well (approximately 1.9287%), it still does not match the accuracy of the proposed model. This demonstrates the clear advantage of the proposed method in reducing relative forecasting errors.

Further analysis of MAE and RMSE provides a more intuitive reflection of forecasting deviation. On the U1 dataset, the proposed model achieves an RMSE of only 0.0345 kW, whereas CNN-LSTM and CNN-BiLSTM reach 0.0498 kW and 0.0473 kW, respectively, corresponding to nearly 30% higher errors. Although Transformer–TCN–GRU and TCN-Informer-BiGRU perform better, their RMSE values of 0.0395 kW and 0.0372 kW are still higher than that of the proposed model. On the more complex U4 dataset, the proposed method maintains strong stability, with RMSE controlled at 0.0408 kW, while CNN-LSTM increases to 0.0598 kW, demonstrating superior robustness under complex operating conditions.

From a structural perspective, the performance differences among comparison models are reasonable. CNN-LSTM and CNN-BiLSTM can extract local features but are limited in modeling long-term dependencies [52], leading to suboptimal forecasting accuracy. TCN-based models improve temporal modeling ability but still struggle to capture global information effectively [53]. Hybrid models such as Transformer–TCN–GRU enhance feature representation via attention mechanisms but may suffer from structural complexity and insufficient feature fusion [49, 54]. In contrast, the IBKA-MSTCN-BiLSTM-QR model effectively integrates multi-scale temporal feature extraction with nonlinear modeling capability, significantly improving both forecasting accuracy and generalization performance.

Table 5. Comparison of Forecasting Results Across Different Models

Sample ID	Model	R^2		MAPE (%)		MAE		RMSE (kW)	
		training set	test set	training set	test set	training set	test set	training set	test set
U1	This study	0.9964	0.9938	1.5467	1.7895	0.0191	0.0204	0.0272	0.0345
	CNN-LSTM	0.9912	0.9875	2.1354	2.4863	0.0287	0.0315	0.0412	0.0498
	CNN-BiLSTM	0.9921	0.9883	2.0487	2.3521	0.0275	0.0302	0.0396	0.0473
	TCN-LSTM	0.9933	0.9896	1.9235	2.2148	0.0251	0.0284	0.0368	0.0442
	Transformer–TCN–GRU	0.9945	0.9912	1.7823	2.0317	0.0228	0.0256	0.0334	0.0395
	TCN-Informer-BiGRU	0.9952	0.9924	1.6895	1.9456	0.0213	0.0238	0.0312	0.0372
	TCN-QRNN	0.9927	0.9889	2.0042	2.2984	0.0264	0.0296	0.0382	0.0457
U2	This study	0.9965	0.9939	1.5461	1.7892	0.0188	0.0201	0.0269	0.0339
	CNN-LSTM	0.9915	0.9878	2.1186	2.4635	0.0281	0.031	0.0407	0.0492
	CNN-BiLSTM	0.9924	0.9886	2.0324	2.3387	0.027	0.0298	0.0391	0.0469
	TCN-LSTM	0.9936	0.9899	1.9072	2.1983	0.0247	0.028	0.0363	0.0438
	Transformer–TCN–GRU	0.9947	0.9914	1.7654	2.0125	0.0224	0.0252	0.033	0.0391
	TCN-Informer-BiGRU	0.9954	0.9926	1.6728	1.9287	0.0209	0.0234	0.0308	0.0368
	TCN-QRNN	0.993	0.9891	1.9885	2.2763	0.026	0.0292	0.0378	0.0452
U3	This study	0.9959	0.9937	1.7265	1.9521	0.0201	0.0365	0.0286	0.0396
	CNN-LSTM	0.9908	0.9869	2.3547	2.6851	0.0305	0.0412	0.0438	0.0556
	CNN-BiLSTM	0.9916	0.9876	2.2654	2.5732	0.0292	0.0395	0.0421	0.0532
	TCN-LSTM	0.9929	0.9888	2.1187	2.4126	0.0268	0.0379	0.0394	0.0498
	Transformer–TCN–GRU	0.994	0.9905	1.9653	2.2187	0.0243	0.0358	0.0359	0.0462
	TCN-Informer-BiGRU	0.9948	0.9921	1.8542	2.0853	0.0226	0.0342	0.0335	0.0435
	TCN-QRNN	0.9922	0.9881	2.1436	2.4487	0.0276	0.0386	0.0402	0.0507
U4	This study	0.9951	0.9901	1.9532	2.2414	0.0214	0.0325	0.0375	0.0408
	CNN-LSTM	0.9895	0.9842	2.6543	2.9875	0.0336	0.0441	0.0485	0.0598
	CNN-BiLSTM	0.9903	0.9851	2.5421	2.8653	0.032	0.0423	0.0468	0.0574
	TCN-LSTM	0.9918	0.9867	2.3864	2.6438	0.0295	0.0396	0.0431	0.0528
	Transformer–TCN–GRU	0.9932	0.9883	2.2147	2.4582	0.0267	0.0368	0.0396	0.0483
	TCN-Informer-BiGRU	0.994	0.9892	2.1036	2.3265	0.0251	0.0352	0.0372	0.0459
	TCN-QRNN	0.9912	0.986	2.4315	2.7124	0.0308	0.0405	0.0446	0.0541

Figure 12 presents the average performance comparison of different models across four datasets (U1–U4). Overall, the proposed model achieves the best performance across all evaluation metrics, demonstrating superior forecasting accuracy and stability. In terms of the R^2 metric, the proposed model attains values of 0.9960 and 0.9929 on the training and test sets, respectively, which are significantly higher than those of all comparison models. This indicates a stronger goodness of fit and the ability to effectively capture complex nonlinear relationships in the data. For error-based metrics, the IBKA-MSTCN-BiLSTM-QR model also shows clear advantages. Specifically, the test MAPE is 1.9430%, which is reduced by approximately 26.8% and 23.3% compared with CNN-LSTM (2.6556%) and CNN-BiLSTM (2.5323%), respectively. Meanwhile, the RMSE is only 0.0372 kW, representing a reduction of about 30.6% compared with CNN-LSTM (0.0536 kW). Even when compared with the relatively strong TCN-Informer-BiGRU model (RMSE of 0.0408 kW), the proposed method still shows noticeable improvement. This demonstrates its stronger capability in reducing forecasting errors. In addition, the gap between training and testing performance indicates that the proposed model exhibits smaller performance fluctuations and more stable generalization ability. In contrast, some comparison models show an increases in test errors, suggesting potential overfitting or insufficient robustness in complex scenarios.

In summary, across all datasets and evaluation metrics, the proposed model consistently achieves the best results. It not only maintains leading performance in R^2 but also significantly reduces MAPE, MAE, and RMSE, fully demonstrating its effectiveness and superiority in complex time-series forecasting tasks.

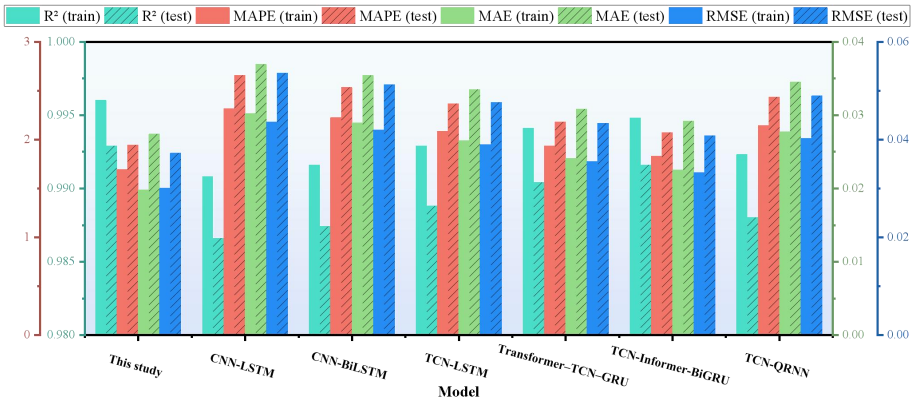
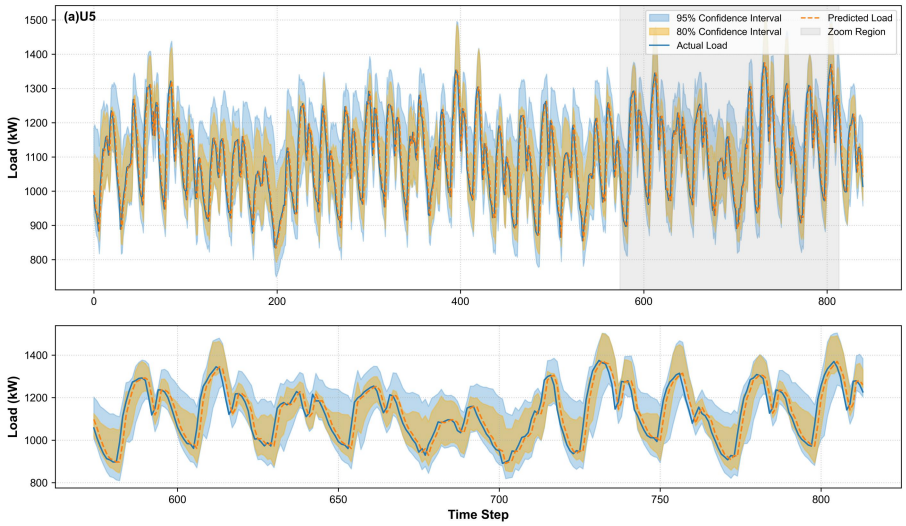


Figure 12. Comparison of average performance metrics across four datasets for different models

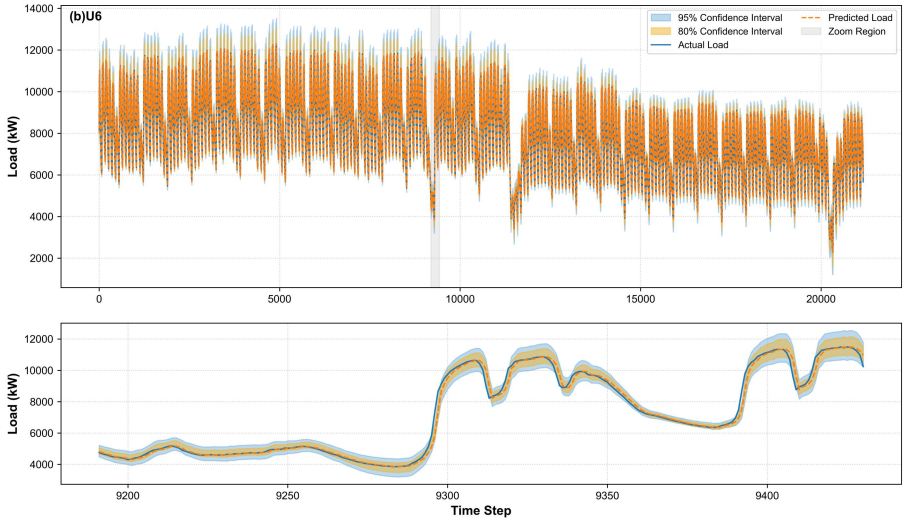
3.7. Model applicability analysis

To further evaluate the applicability of the proposed IKBA-MSTCN-BiLSTM-QR interval forecasting model on different types of electricity load datasets, two additional datasets with significantly different characteristics, denoted as U5 and U6, are introduced based on the original U1–U4 datasets. These datasets differ substantially in terms of regional source, sampling frequency, load scale, and exogenous variable composition, thereby enabling a comprehensive assessment of the model's generalization capability under cross-region, cross-scale, and heterogeneous data structure conditions. Specifically, the U5 dataset is obtained from the publicly available Kaggle dataset *Hourly Load India - Electrical Load Forecasting* (https://www.kaggle.com/datasets/shubhamvashisht/hourly-load-india-electrical-load-forecasting), which contains hourly national electricity load data of India along with corresponding meteorological information. This dataset reflects the coupling relationship between large-scale electricity load and weather factors. The U6 dataset is derived from the Electrician Competition dataset, constructed by integrating load data and meteorological data. The original load data consist of 96 sampling points per day at a 15-minute resolution. In this study, the data are first transformed from a wide format into a continuous time series with corresponding timestamps. The meteorological data include five daily variables: maximum temperature, minimum temperature, average temperature, average relative humidity, and precipitation. These data are then aligned with the load data by date to form a unified load–weather dataset. Due to its higher temporal resolution and longer time span, this dataset is suitable for evaluating the model in high-frequency load forecasting scenarios.

The data preprocessing procedure is consistent with the previous analysis. First, both input features and target load are standardized. Then, supervised learning samples are constructed using a sliding window with a length of 10. Finally, the dataset is divided into training and testing sets in a chronological order with a ratio of 8:2, which effectively avoids future information leakage and ensures that the experimental setup is consistent with real-world short-term load forecasting applications.



(a) U5 India national-level load dataset



(b) U6 load-meteorological fusion dataset

Figure 13. Comparison of interval prediction results across different electricity load datasets

Based on the analysis of **Figure 13**, the results on the U5 dataset indicate that the proposed model can effectively track the overall variation trend of the national-level load series. On the test set, the MAE, RMSE, and MAPE are 35.1259, 44.8474, and 3.1952%, respectively, suggesting that the fo values are close to the actual load with relatively small average errors. The R^2 value reaches 0.8465, demonstrating that the model can explain most of the load variation and exhibits strong point fo capability. From the perspective of interval fo, the PICP of U5 is 84.5238% at the 80% confidence level and 97.5000% at the 95% confidence level, both meeting or exceeding the nominal confidence levels. This indicates that the model not only provides accurate point fo results but also constructs reliable prediction intervals to effectively capture the uncertainty in load fluctuations. The zoomed-in view further shows that the prediction intervals can adaptively adjust with the magnitude of load variations, maintaining good coverage during both rising and falling periods.

For the U6 dataset, which features high-frequency sampling at 15-minute intervals, the intra-day load fluctuations are more pronounced. The experimental results demonstrate that the model also achieves strong performance on this dataset. The MAE, RMSE, and MAPE are 154.4441, 266.2174, and 1.9081%, respectively, indicating that the fo results can effectively capture the main patterns of high-frequency load variations. The R2R^2R2 value reaches 0.9824, showing a high degree of consistency between fo results and actual load values. In terms of interval fo performance, the PICP of U6 is 81.4201% at the 80% confidence level, which is close to the target level, and 91.4427% at the 95% confidence level, covering the vast majority of actual samples. Due to the frequent short-term fluctuations and local spikes in this dataset, the 95% interval coverage is slightly lower than the theoretical level; however, the model still maintains good reliability and stability. The local zoom-in results further indicate that the model can effectively capture peak–valley transitions in 15-minute load series.

Overall, the results on U5 and U6 demonstrate that the proposed model exhibits strong applicability across datasets with different sources, sampling frequencies, and load scales. The performance on U5 confirms that the model can be successfully transferred to national-level load fo tasks while maintaining high accuracy and interval coverage on public datasets. The results on U6 further indicate that the model is also well-suited for high-frequency load fo scenarios, effectively capturing intra-day periodicity and short-term variations. Compared with validation on a single dataset, the additional experiments on U5 and U6 further verify the generalization capability of the model. Specifically, the MSTCN module extracts local features at multiple temporal scales, the BiLSTM captures bidirectional temporal dependencies, and the quantile regression output layer provides prediction intervals at different confidence levels based on point fo. Therefore, the proposed model is applicable not only to building load datasets but also to national-scale load data and high-frequency load–weather integrated datasets.

3.8. Ablation experiment results

To verify the effectiveness of each component in the proposed IBKA-MSTCN-BiLSTM-QR model and its contribution to overall forecasting performance, a systematic ablation study is conducted on the

U1–U4 U.S. residential electricity load datasets. In the experimental design, the complete model (IBKA-MSTCN-BiLSTM-QR) is taken as the baseline, and a series of comparison models are constructed by progressively removing or replacing key modules. For the interval forecasting task, the results generated by the QR layer under the 80% confidence level are used as the primary evaluation target. The specific structures of the compared models are summarized in **Table 6.**

Table 6. Comparison of Ablation Experiment Models

Model Number	Model Structure
Model A (Full Model)	IBKA-MSTCN-BiLSTM-QR
Model B	Without the IBKA optimization algorithm, fixed empirical parameters are used (MSTCN-BiLSTM-QR)
Model C	Replacing MSTCN with the conventional TCN structure (IBKA-TCN-BiLSTM-QR)
Model D	Removal of BiLSTM, retaining only MSTCN with direct output to the QR layer (IBKA-MSTCN-QR)
Model E	Removal of MSTCN, retaining only BiLSTM with direct output connected to the QR layer (IBKA-BiLSTM-QR)

Figure 14 presents the performance comparison of different ablation models in the interval forecasting task. Overall, the complete model (Model A) achieves the best performance across all evaluation metrics, with a PICP of 0.844 and an MPICD of 0.04247 kW, indicating that it maintains a relatively high coverage probability while achieving a well-constrained forecasting interval. When the IBKA optimization algorithm is removed (Model B), the model performance declines to some extent, with PICP decreasing to 0.821 (a reduction of approximately 2.7%) and MPICD increasing to 0.04638 kW. This indicates that the parameter optimization strategy plays a crucial role in improving interval forecasting accuracy. Further replacing MSTCN with the conventional TCN structure (Model C) leads to a continued performance degradation, where PICP drops to 0.803 and MPICD increases to 0.04992 kW. Compared with the complete model, the interval width increases by approximately 17.5%, demonstrating the significant impact of multi-scale feature extraction on model performance. When the BiLSTM module is removed (Model D), the ability to capture temporal dependencies is weakened, resulting in a further decrease in PICP to 0.789 and an increase in MPICD to 0.05215 kW. The interval width expands by approximately 22.8% compared with the full model, indicating the importance of bidirectional temporal modeling. In Model E, which retains only the BiLSTM structure, the worst performance is observed, with PICP decreasing to 0.765 (a reduction of about 9.4% compared with the complete model) and MPICD increasing to 0.05583 kW (an increase of approximately 31.5%). This suggests that single-scale temporal modeling is insufficient to effectively capture complex dynamic characteristics. In summary, as key modules are progressively removed, the model performance shows a clear trend of "decreasing coverage and widening interval," fully verifying the importance of IBKA, MSTCN, and BiLSTM in improving interval forecasting performance. In particular, the synergistic mechanism of multi-scale feature extraction and bidirectional temporal modeling plays a critical role in enhancing overall model effectiveness.

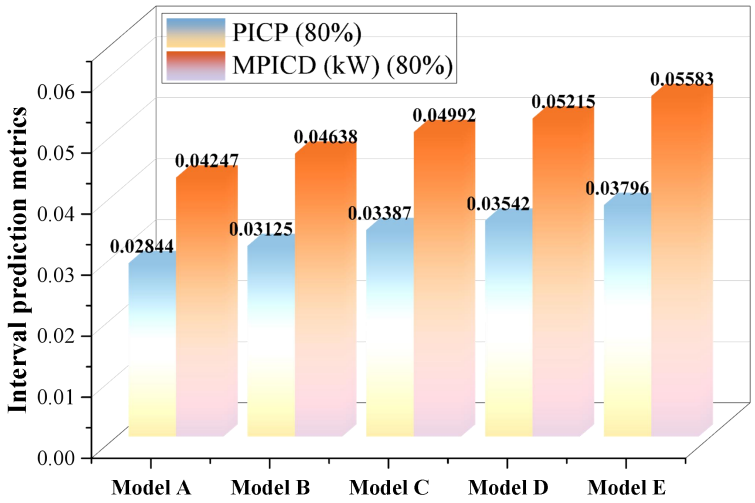


Figure 14. Comparison of model ablation experiment results

3.9. Feature indicator contribution analysis

To further investigate the impact of different input features on electricity load forecasting, this study employs the SHAP (Shapley Additive Explanations) method to analyze feature contributions. Since the trained weights of the deep learning model were not separately preserved in the experimental results, a tree-based surrogate model is constructed using the same input variables, and TreeExplainer is applied to compute the SHAP values for each feature. The surrogate model achieves an R^2=0.9968 on the test set, demonstrating its strong capability to approximate the nonlinear relationship between load and multidimensional energy consumption features. Therefore, it can be used to identify the key influencing factors and their directional effects. It should be emphasized that this SHAP analysis is intended to explain the contribution patterns of the input features to load forecasting and does not directly correspond to the internal weight interpretation of the deep learning model.

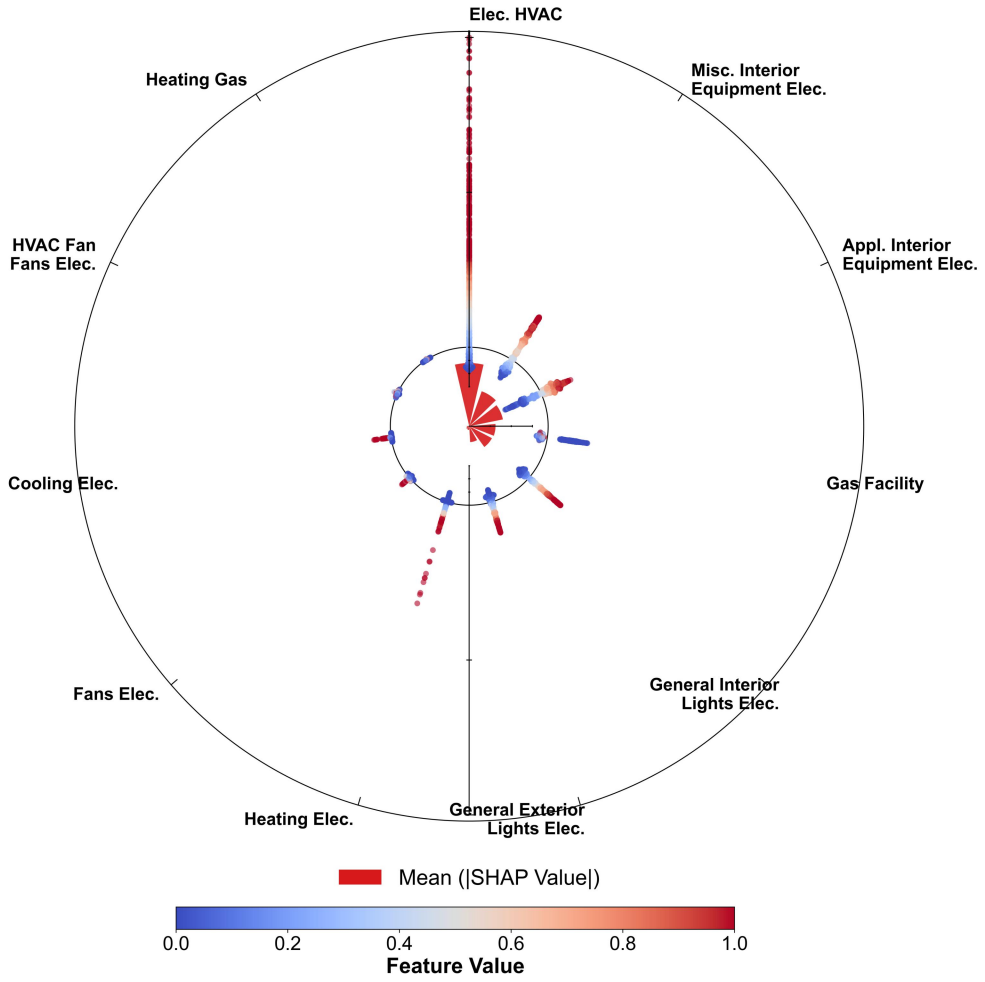


Figure 15. Polar SHAP analysis plot of feature contribution

Figure 15 shows the mean absolute SHAP values of all features and their sample-level contribution distributions. The red bars represent the mean absolute SHAP values, where larger values indicate greater overall importance. The color of the scatter points denotes feature magnitude, while their horizontal distribution reflects the direction and strength of each feature's impact on the model output. The results show that Elec. HVAC has the highest mean absolute SHAP value (0.2200), indicating that it is the most influential factor in electricity load forecasting. This suggests that HVAC-related electricity consumption plays a dominant role in driving residential load variations.

Figure 16 further depicts the relationships between the top four features and their SHAP values. Elec. HVAC exhibits a clear positive correlation with SHAP values, indicating that higher HVAC electricity consumption leads to increased load forecasts and is a key contributor to peak demand. Similarly, Misc. Interior Equipment Elec. and Appl. Interior Equipment Elec. also show strong positive

effects, with their contributions increasing as feature values rise, highlighting the importance of indoor equipment and appliance usage in load growth. In contrast, Gas Facility shows a nonlinear and stratified distribution pattern, suggesting that its relationship with electricity load is more complex and influenced by factors such as seasonal variation, heating modes, and building operation conditions. The contribution of this feature varies across different value ranges, reflecting the coupling relationship between gas and electricity consumption in residential energy systems.

Overall, the SHAP analysis indicates that the model mainly relies on HVAC load, indoor equipment, appliance usage, and energy coupling characteristics for prediction. These findings are consistent with real-world residential energy consumption patterns, demonstrating that the proposed model not only achieves high accuracy but also captures physically meaningful driving factors, providing interpretable support for load management, demand response, and energy optimization.

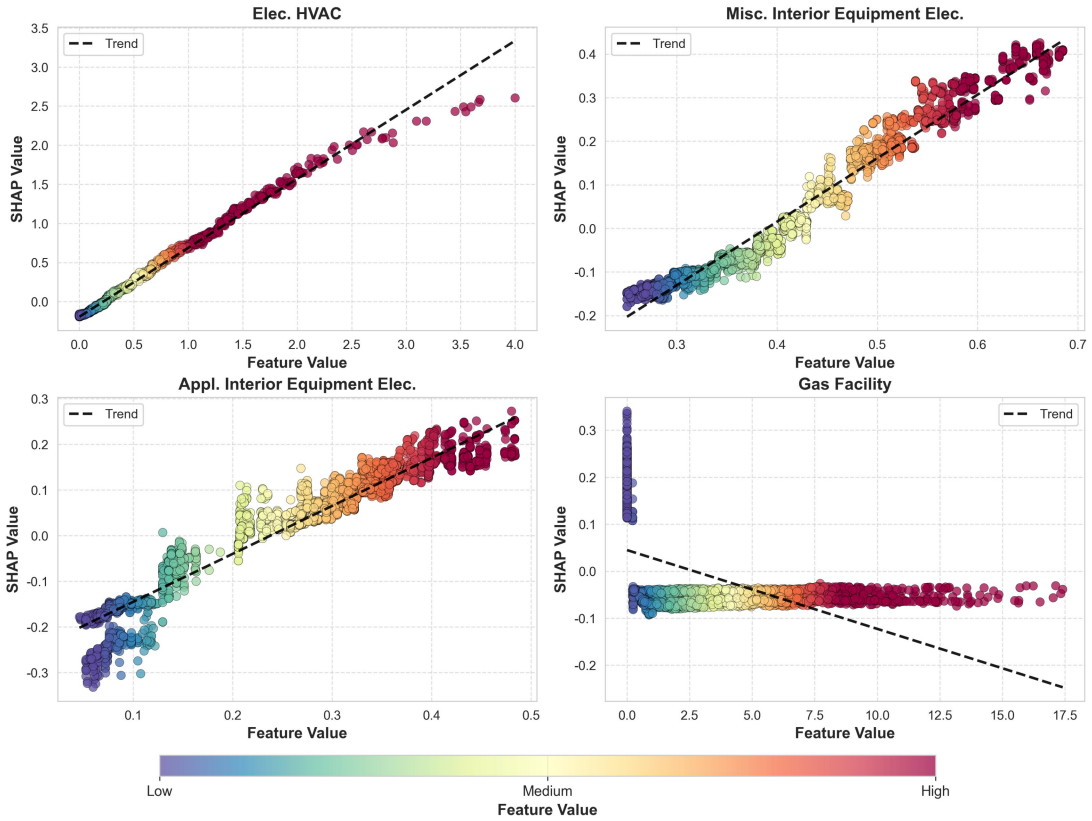


Figure 16. Correlation between the top four features and SHAP values

4. Discussion

From the experimental results, the proposed model demonstrates strong overall advantages in both point forecasting and interval forecasting tasks. On the one hand, MSTCN, through its serial–parallel architecture and multi-head attention mechanism, effectively enhances multi-scale feature extraction capability and compensates for the limitations of conventional TCN in deep information utilization. On the other hand, BiLSTM further strengthens the ability to capture complex load variation patterns via bidirectional temporal modeling, enabling the model to maintain high accuracy even under non-stationary and highly volatile conditions. In addition, the introduction of quantile regression extends the model from deterministic point forecasting to probabilistic interval forecasting, allowing a more comprehensive characterization of load uncertainty, which is of significant importance for practical dispatching and risk assessment. In terms of parameter optimization, IBKA significantly improves global search capability through multiple enhancement strategies, effectively avoiding the local optimum problem commonly encountered in traditional optimization algorithms, thereby enhancing model stability and generalization performance. However, this study still has certain limitations. For example, the model architecture is relatively complex, leading to higher training time and computational resource consumption. Moreover, the dataset used in this study mainly consists of U.S. residential load data, which, although representative, still requires further validation in industrial load or regional power grid scenarios [55]. Future work will focus on lightweight model design [56],

multi-source heterogeneous data fusion [57], and cross-region generalization improvement [58] to further enhance the engineering applicability of the proposed model.

Future research will particularly focus on the application framework for smart grids. This framework takes high-accuracy load forecasting as the core driving force and establishes an end-to-end collaborative system integrating data perception, intelligent decision-making, and system execution, enabling efficient coordination and optimized control of all power grid components. From an overall perspective, the framework follows a hierarchical design logic of "data-driven modeling—model support—dispatch optimization—security assurance—infrastructure support," reflecting a clear systems engineering structure and intelligent evolution trend. At the core level, the forecasting foundation layer integrates meteorological information, grid operation status, and multi-source heterogeneous user-side data to construct high-dimensional time-series inputs, thereby enhancing data representation capability. Based on this, the IBKA algorithm is introduced to achieve global optimization of model hyperparameters, significantly improving convergence performance and generalization ability. Meanwhile, the MSTCN–BiLSTM cooperative structure combines multi-scale feature extraction with bidirectional temporal dependency modeling, enabling accurate characterization of non-stationary and complex dynamic load behaviors. The quantile regression layer further outputs interval forecastings, providing reliable support for uncertainty analysis. At the application level, the forecasting results directly support power system dispatching and optimization decisions. Through unit commitment and demand response mechanisms, dynamic supply–demand balance can be achieved. In the context of high penetration of renewable energy, the forecasting results also provide critical support for wind and solar output fluctuation forecasting and coordinated energy storage scheduling, thereby improving renewable energy integration and system flexibility. At the security assurance layer, forecasting-driven risk assessment and state monitoring mechanisms enable proactive identification of potential operational risks, thereby enhancing grid resilience and stability. At the infrastructure layer, smart meters, IoT devices, and edge computing technologies provide the foundation for real-time data acquisition and rapid processing. Combined with secure communication networks, they ensure efficient and reliable transmission of information and control signals. In summary, the proposed framework establishes a closed-loop intelligent operation mechanism driven by load forecasting, achieving full-process intelligence from data acquisition to decision execution, and providing an important technical pathway for the efficient, secure, and low-carbon operation of future smart power systems.

5. Conclusions

(1) The comparative experiments on six representative benchmark functions (F1–F6) demonstrate that the IBKA algorithm significantly outperforms traditional optimization methods such as SSA, GWO, WOA, and BKA in terms of both optimization accuracy and convergence performance. For unimodal functions, both the best (Best) and mean (Mean) values approach the theoretical optimum (below the order of 10^{-130}), while the standard deviation is nearly zero, indicating extremely high stability. For multimodal functions, IBKA consistently achieves the global optimal solution (e.g., reaching zero for F5 and F6), effectively avoiding premature convergence to local optima. These results confirm that the introduction of Tent chaotic mapping, dynamic lens-based opposition learning, and diffraction correction strategies significantly enhances the global search capability and convergence efficiency of the algorithm.

(2) The proposed IBKA–MSTCN–BiLSTM–QR model achieves excellent forecasting performance across the U1–U4 datasets. For point forecasting, the test R^2 remains consistently above 0.99 (up to 0.9951), with the lowest MAPE of 1.7892% and RMSE ranging from 0.0339 to 0.0408 kW. For interval forecasting, the PICP under the 80% confidence level is approximately 0.798–0.818, increasing to 0.946–0.958 under the 95% confidence level, while MPICD remains within 0.063–0.112 kW. These results indicate that the proposed model not only achieves high forecasting accuracy but also maintains a good balance between coverage probability and interval width, demonstrating strong reliability and practical applicability.

(3) Compared with models such as CNN-LSTM, TCN-LSTM, Transformer–TCN–GRU, and TCN-Informer-BiGRU, the proposed model consistently achieves the best performance across all evaluation metrics. Specifically, the average test R^2 reaches 0.9929, which is approximately 0.001–0.005 higher than competing models. The MAPE is reduced to 1.9430%, representing an improvement of about 26.8% over CNN-LSTM. The RMSE is only 0.0372 kW, showing a reduction of approximately 30% compared with traditional models. Moreover, stable performance across all datasets (U1–U4) further confirms the strong generalization ability and adaptability of the proposed model under different load fluctuation patterns and complex operational environments.

(4) The ablation study results indicate that each key component contributes significantly to overall model performance. The complete model achieves a PICP of 0.844 and an MPICD of 0.04247 kW.

After removing IBKA, PICP decreases to 0.821; replacing MSTCN further reduces it to 0.803; removing BiLSTM leads to a decline to 0.789; and retaining only the BiLSTM structure results in the worst performance with a PICP of 0.765 and an MPICD increased to 0.05583 kW. The overall trend of "decreasing coverage and increasing interval width" demonstrates the critical importance of multi-scale feature extraction and bidirectional temporal modeling. In addition, feature contribution analysis further validates the importance of key input variables, enhancing both model interpretability and engineering applicability.

Nomenclature List

R^2	Coefficient of Determination
MAPE (%)	Mean Absolute Percentage Error
MAE	Mean Absolute Error
RMSE (kW)	Root Mean Square Error
PICP	Prediction Interval Coverage Probability
MPIW	Mean Prediction Interval Width
n	Total number of samples
y_i	True value of the i-th sample
$\hat{y}_i$	Predicted value of the i-th sample
$\bar{y}$	Mean of true values
α	Confidence level parameter
$y_{\min}$	Minimum of observed values
$y_{\max}$	Maximum of observed values
I(condition)	Indicator function
TCN Layers	Number of Temporal Convolutional Network layers
Channel Size	Number of channels in TCN layer
BiLSTM Layers	Number of Bidirectional LSTM layers
BiLSTM Neurons	Number of hidden neurons in BiLSTM layer
Validation Quantile_Loss	Validation loss based on quantile regression
Best, Mean, Std	Evaluation metrics for optimization algorithms
F1–F6	Benchmark test functions

Author Contributions
Zixiang Long: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Software, Data curation, Conceptualization.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest
The authors declare no conflicts of interest.

Data Availability Statement
The data used to support the findings of this study are available from thecorresponding authors upon request.

References

1. Shu Y, Tang Y, Zhang Z, Zhang F, Zhong W, Shi H, et al. Pathway for New-type Power System in China. CSEE Journal of Power and Energy Systems. 2025;11(6):2684-95.https://dx.doi.org/10.17775/CSEEJPES.2025.05700
2. Wang Y. The Goals of Carbon Peaking and Carbon Neutrality and China's New Energy Revolution. Frontiers of Economics in China. 2022;17(2):325.https://dx.doi.org/10.3868/s060-015-022-0012-5
3. Medina C, Ana CRM, González G. Transmission grids to foster high penetration of large-scale variable renewable energy sources–A review of challenges, problems, and solutions. International Journal of Renewable Energy Research (IJRER). 2022;12(1):146-69.https://dx.doi.org/10.20508/ijrer.v12i1.12738.g8400

4. Zhang Y, Shen J, Li J, Yu M, Chen X, Yin Z. Achieving high precision and balanced multi-energy load forecasting with mixed time scales: a multi-task learning model with stacked cross-attention. Energy and AI. 2025:100561.https://dx.doi.org/10.1016/j.egyai.2025.100561
5. Xu N, Han J, Xiong Y, Wang ZL, Sun Q. TENG applications in transportation and surrounding emergency management. Advanced Sustainable Systems. 2022;6(10):2200267.https://dx.doi.org/10.1002/adsu.202200267
6. Lei W, Jiang Y, Zeng X, Fan Z. Research on the transmission law of kurtosis of SDOF system under nonstationary and non-Gaussian random excitations. Mechanical Systems and Signal Processing. 2022;165:108292.https://dx.doi.org/10.1016/j.ymssp.2021.108292
7. Azeem A, Ismail I, Jameel SM, Harindran VR. Electrical load forecasting models for different generation modalities: a review. IEEe Access. 2021;9:142239-63.https://dx.doi.org/10.1109/ACCESS.2021.3120731
8. Kandil M, El-Debeiky SM, Hasanien N. Long-term load forecasting for fast developing utility using a knowledge-based expert system. IEEE transactions on Power Systems. 2002;17(2):491-6.https://dx.doi.org/10.1109/TPWRS.2002.1007923
9. Swain A, Hossain I, Liu C, Pong PW. Modeling Non-Linear and Non-Stationary Magnetic Signals: An Enhanced Signal Processing Strategy for Wind Time Series Analysis. IEEE Transactions on Magnetics. 2025.https://dx.doi.org/10.1109/TMAG.2025.3557261
10. Wang Z, Srinivasan RS. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. Renewable and sustainable energy reviews. 2017;75:796-808.https://dx.doi.org/10.1016/j.rser.2016.10.079
11. Zhang Y, Ai Q, Lin L, Yuan S, Li Z. A very short-term load forecasting method based on deep LSTM RNN at zone level. Power System Technology. 2019;43(6):1884-92.https://dx.doi.org/10.13335/j.1000-3673.pst.2018.2101
12. Wang C, Wang Y, Ding Z, Zheng T, Hu J, Zhang K. A transformer-based method of multienergy load forecasting in integrated energy system. IEEE Transactions on Smart Grid. 2022;13(4):2703-14.https://dx.doi.org/10.1109/TSG.2022.3166600
13. Lu L, Zhisheng Z. Short-term load forecasting of a power system based on multi-scale feature enhanced DHTCN. Power System Protection and Control. 2023;51(10):172-9.https://dx.doi.org/10.19783/j.cnki.pspc.221134
14. Hernandez L, Baladron C, Aguiar JM, Carro B, Sanchez-Esguevillas AJ, Lloret J, et al. A survey on electric power demand forecasting: future trends in smart grids, microgrids and smart buildings. IEEE Communications Surveys & Tutorials. 2014;16(3):1460-95.https://dx.doi.org/10.1109/SURV.2014.032014.00094
15. Alhussein M, Aurangzeb K, Haider SI. Hybrid CNN-LSTM model for short-term individual household load forecasting. Ieee Access. 2020;8:180544-57.https://dx.doi.org/10.1109/ACCESS.2020.3028281
16. Pu X, Xiao H, Wang J, Pei W, Yang J, Zhang J. A novel GRU-TCN network based Interactive Behavior Learning of multi-energy Microgrid under incomplete information. Energy Reports. 2023;9:608-16.https://dx.doi.org/10.1016/j.egyr.2023.04.128
17. Han Y, Tong X, Deng Y. Probabilistic Distribution Estimation and Temporal Transformer-Based Interval Prediction in Day-ahead Wind Power Prediction. Conference Probabilistic Distribution Estimation and Temporal Transformer-Based Interval Prediction in Day-ahead Wind Power Prediction. p. 1-11.
18. Liu H, Li Z, Li C, Shao L, Li J. Research and application of short-term load forecasting based on CEEMDAN-LSTM modeling. Energy Reports. 2024;12:2144-55.https://dx.doi.org/10.1016/j.egyr.2024.08.035

19. Feng Y, Zhu J, Qiu P, Zhang X, Shuai C. Short-term power load forecasting based on TCN-BiLSTM-attention and multi-feature fusion. Arabian Journal for Science and Engineering. 2025;50(8):5475-86.https://dx.doi.org/10.1007/s13369-024-09351-5
20. Du C, Zhang J, Fang J. An innovative complex-valued encoding black-winged kite algorithm for global optimization. Scientific Reports. 2025;15(1):932.https://dx.doi.org/10.1038/s41598-024-83589-9
21. Wang J, Wang W-c, Hu X-x, Qiu L, Zang H-f. Black-winged kite algorithm: a nature-inspired meta-heuristic for solving benchmark functions and engineering problems. Artificial Intelligence Review. 2024;57(4):98. https://dx.doi.org/10.1007/s10462-024-10723-4
22. Fu Y, Liu D, Fu S, Chen J, He L. Enhanced Aquila optimizer based on tent chaotic mapping and new rules. Scientific reports. 2024;14(1):3013.https://dx.doi.org/10.1038/s41598-024-53064-6
23. Qi Y, Jiang A, Gao Y. A Gaussian convolutional optimization algorithm with tent chaotic mapping. Scientific Reports. 2024;14(1):31027.https://dx.doi.org/10.1038/s41598-024-82277-y
24. Wang C, Chen N, Heidrich W. do: A differentiable engine for deep lens design of computational imaging systems. IEEE Transactions on Computational Imaging. 2022;8:905-16.https://dx.doi.org/10.1109/TCI.2022.3212837
25. Qu H, Shao X, Gao H, Chen Q, Guang J, Liu C. A Prediction Model for Methane Concentration in the Buertai Coal Mine Based on Improved Black Kite Algorithm–Informer–Bidirectional Long Short-Term Memory. Processes. 2025;13(1):205.https://dx.doi.org/10.3390/pr13010205
26. Budak VP, Efremenko DS, Smirnov PA. Fraunhofer diffraction description in the approximation of the light field theory. Light and Engineering. 2020;28(5):25-30.https://dx.doi.org/10.33383/2020-021
27. Gong J, Qu Z, Zhu Z, Xu H, Yang Q. Ensemble models of TCN-LSTM-LightGBM based on ensemble learning methods for short-term electrical load forecasting. Energy. 2025;318:134757.https://dx.doi.org/10.1016/ j.energy.2025.134757
28. Tian J, Liu H, Gan W, Zhou Y, Wang N, Ma S. Short-term electric vehicle charging load forecasting based on TCN-LSTM network with comprehensive similar day identification. Applied Energy. 2025;381:125174. https://dx.doi.org/10.1016/j.apenergy.2024.125174
29. Li J, Wang X, Tu Z, Lyu MR. On the diversity of multi-head attention. Neurocomputing. 2021;454:14-24.https://dx.doi.org/10.1016/j.neucom.2021.04.038
30. Wang G, Zhang J, Cui L, Xue S, Zhang B, Zhang P. Substation-level distributed rooftop photovoltaic power day-ahead prediction based on double attention mechanism transformer model. Journal of Global Energy Interconnection. 2024;7(4):393-405.https://dx.doi.org/10.19705/j.cnki.issn2096-5125.2024.04.005
31. Kim J, Choi K, Park I-C. Hardware-Efficient Unified Approximation for Implementing Diverse Smooth Activation Functions. IEEE Transactions on Computers. 2026.https://dx.doi.org/10.1109/TC.2026.3661496
32. Zhang S, Chen R, Cao J, Tan J. A CNN and LSTM-based multi-task learning architecture for short and medium-term electricity load forecasting. Electric power systems research. 2023;222:109507. https://dx.doi.org/10.1016/j.epsr.2023.109507
33. Han J, Zeng P. Residual BiLSTM based hybrid model for short-term load forecasting in buildings. Journal of Building Engineering. 2025;99:111593.https://dx.doi.org/10.1016/j.jobe.2024.111593
34. Zhang W, Ma B, Wang D. Temporal forecasting of electric vehicle charging load using an IVY-VMD-TCN-BiLSTM model with cross-zone evaluation. Scientific Reports. 2026.https://dx.doi.org/10.1038/s41598-026-47962-0
35. Koenker R, Bassett Jr G. Regression quantiles. Econometrica: journal of the Econometric Society. 1978:33-50.https://dx.doi.org/10.2307/1913643

36. Shen L, Bao Y, Hasan N, Huang Y, Zhou X, Deng C. Intelligent crude oil price probability forecasting: Deep learning models and industry applications. Computers in Industry. 2024;163:104150.https://dx.doi.org/10.1016/j.compind.2024.104150
37. Chen Q, Qu H, Liu C, Xu X, Wang Y, Liu J. Spontaneous coal combustion temperature prediction based on an improved grey wolf optimizer-gated recurrent unit model. Energy. 2025;314:133980
38. Luo S, Chen X, Pang X, Wang B, Zheng Z. Mid-term power load forecasting using an ensemble deep learning model with BKA and CWGAN-GP enhancements. Scientific Reports. 2026.https://dx.doi.org/10.1038/ s41598-026-49674-x
39. Zhang X. Short-term electrical load forecasting strategy based on EEMD-SSA-BiLSTM. Signal, Image and Video Processing. 2025;19(9):751.https://dx.doi.org/10.1007/s11760-025-04347-6
40. Li T, Qian Z, He T. Short-Term Load Forecasting With Improved CEEMDAN and GWO-Based Multiple Kernel ELM. Complexity. 2020;2020(1):1209547.https://dx.doi.org/10.1155/2020/1209547
41. Han X, Shi Y, Tong R, Wang S, Zhang Y. Research on short-term load forecasting of power system based on IWOA-KELM. Energy Reports. 2023;9:238-46.https://dx.doi.org/10.1016/j.egyr.2023.05.162
42. Saeed F, Rehman A, Shah HA, Diyan M, Chen J, Kang J-M. SmartFormer: Graph-based transformer model for energy load forecasting. Sustainable Energy Technologies and Assessments. 2025;73:104133.https://dx.doi. org/10.1016/j.seta.2024.104133
43. Hu J, Duan P, Cao X, Xue Q, Zhao B, Zhao X, et al. A multi-energy load forecasting method based on the Mixture-of-Experts model and dynamic multilevel attention mechanism. Energy. 2025;324:135947. https://dx.doi.org/10.1016/j.energy.2025.135947
44. Memarzadeh G, Amirteimoury F, Noori H, Keynia F. An electrical load forecasting model based on a novel closed loop neural networks and interaction gain feature selection. Results in Engineering. 2025; 27:105800.https://dx.doi.org/10.1016/j.rineng.2025.105800
45. Royal E, Bandyopadhyay S, Newman A, Huang Q, Tabares-Velasco PC. A statistical framework for district energy long-term electric load forecasting. Applied Energy. 2025;384:125445.https://dx.doi.org/10.1016/j.apenergy.2025.125445
46. Feng Z-k, Liu P, Niu W-j, Fu X-y, Xiao Y, Yang T, et al. Twin extreme learning machine model and cooperation search algorithm for multi-step-ahead point and interval runoff prediction. Journal of Hydrology. 2025;653:132778.https://dx.doi.org/10.1016/j.jhydrol.2025.132778
47. Chen Q, Kang J, Peng C, Zhang R, Tian J, Yang C, et al. Prediction model of spontaneous combustion temperature intervals of coal samples from multiple mining regions. Fuel. 2026;425:139383. https://dx.doi.org/10.1016/j.fuel.2026.139383
48. Wang B, Wang L, Ma Y, Hou D, Sun W, Li S. A short-term load forecasting method considering multiple factors based on VAR and CEEMDAN-CNN-BILSTM. Energies. 2025;18(7):1855.https://dx.doi.org/10.3390/en18071855
49. Alsamraee SA, Khanna S. A Hybrid Transformer–TCN–GRU Based Model for Thermal Load Forecasting of a Large University Campus. Energy. 2026:140114.https://dx.doi.org/10.1016/j.energy.2026.140114
50. Zhang H, Zhou M, Chen Y, Kong W. Short-term power load forecasting for industrial buildings based on decomposition reconstruction and TCN-Informer-BiGRU. Energy and Buildings. 2025:116317.https://dx.doi.org/10.1016/j.enbuild.2025.116317
51. Mochurad L, Levkovych R. TCN-QRNN model for short term energy consumption forecasting with increased accuracy and optimized computational efficiency. Scientific Reports. 2025;15(1):28488.https://dx.doi.org/10. 1038/s41598-025-14423-z

52. Wu K, Wu J, Feng L, Yang B, Liang R, Yang S, et al. An attention-based CNN-LSTM-BiLSTM model for short-term electric load forecasting in integrated energy system. International Transactions on Electrical Energy Systems. 2021;31(1):e12637.https://dx.doi.org/10.1002/2050-7038.12637
53. Mao J, Xu W, Li D, Zhu H, Yang F. A TCN-Attention fusion model for fault prediction and remaining useful life estimation of large-scale mining equipment. Scientific Reports. 2026.https://dx.doi.org/10.1038/s41598-026-43145-z
54. Mahmoud A, Mohammed A. Enhancing time series forecasting: a hybrid TCN-transformer approach. Neural Computing and Applications. 2026;38(9):342.https://dx.doi.org/10.1007/s00521-026-12048-5
55. Onu UG, de Lorenci EVN, de Souza AZ, Balestrassi PP, Eicker U. Enhancing industrial load hosting capacity in rural areas of developing countries through distributed energy resource integration. Renewable Energy. 2025;255:123816.https://dx.doi.org/10.1016/j.renene.2025.123816
56. Meng Z, Wang D, Han X, Zhang Z, Zhang X, Ni Y, et al. Fatigue-constrained lightweight design of electric truck frame considering power battery system layout: Meng et al. Structural and Multidisciplinary Optimization. 2025;68(7):133.https://dx.doi.org/10.1007/s00158-025-04077-w
57. Ye H, Teng X, Song B, Zou K, Zhu M, He G. Multi-source data fusion-based grid-level load forecasting. Applied Sciences. 2025;15(9):4820.https://dx.doi.org/10.3390/app15094820
58. Jiongwei C, Xiang L, Jiahua W, Huimin Z, Dongqin Y, Juan B, et al. Improving the utilization of PV-wind power by thermal, hydro and pumped storage considering the local and cross-regional power demand. Energy. 2026:139919.https://dx.doi.org/10.1016/j.energy.2026.139919

About the Author

Zixiang Long was born in Xuzhou, Jiangsu, P.R. China, in 2005.Now, he studies at the College of Faculty of Electrical and Control Engineering, Liaoning Technical University. His research interests include smart grid information engineering.